From simulations to real data, on the relevance of deep learning models and domain adaptation: Application to astrophysics with CTAO and LST-1

Michaël Dell'Aiera

UNIVERSITÉ SAVOIE MONT BLANC

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ SAVOIE MONT BLANC

Spécialité : **STIC Informatique**

Arrêté ministériel : 25 Mai 2016

Présentée par

Michaël Dell'aiera

Thèse dirigée par **Alexandre Benoit** et
codirigée par **Thomas Vuillaume**

préparée au sein du **LAPP et LISTIC**
dans **l'École Doctorale SIE**

From simulations to real data, on the relevance of deep learning models and domain adaptation: Application to astrophysics with CTAO and LST-1

Thèse soutenue publiquement le **14 Octobre 2024**,
devant le jury composé de :

M. Giovanni Lamanna
Directeur de recherche, Université Savoie Mont Blanc, Président
Mme Marie Chabert
Professeure des universités, Université de Toulouse, Rapporteuse
M. Karl Kosack
Ingénieur de recherche, Université Paris Saclay, Rapporteur
Mme Alice Caplier
Professeure des universités, Université Grenoble Alpes, Examinatrice
M. David Rousseau
Directeur de recherche, Université Paris Saclay, Examinateur
M. Alexandre Benoit
Professeur, Université Savoie Mont Blanc, Directeur de thèse
M. Thomas Vuillaume
Ingénieur de recherche, Université Savoie Mont Blanc, Co-Directeur


UNIVERSITÉ
SAVOIE
MONT BLANC

Tout est éphémère, et le fait de se souvenir, et l'objet dont on se souvient.

Remerciements

S'il y a bien une façon de modéliser physiquement mon parcours jusqu'à cette thèse, ce serait probablement par un mouvement brownien. Jamais, par le passé, je n'aurais imaginé qu'un tel enchevêtrement d'événements aléatoires m'aurait conduit à soutenir une thèse dans le domaine florissant de l'apprentissage profond. Ce goût pour la recherche n'est peut-être pas inné. Mais je tiens à remercier, en tout premier lieu, Marc et Erwann de me l'avoir transmis durant mes études d'ingénieur. Vous avez su éveiller ma curiosité pendant mes études d'ingénieur et m'initier à cette passion qui m'anime aujourd'hui.

Naturellement, un tel travail ne peut être l'œuvre d'une seule personne. La recherche est avant tout un effort collectif, et cette thèse n'aurait pas vu le jour sans le soutien et l'implication de nombreuses personnes. Je tiens d'abord à exprimer ma profonde gratitude aux membres du jury qui ont accepté d'évaluer mes travaux. Merci pour le temps et l'attention que vous avez consacrés à analyser et commenter mon travail. Je souhaite également adresser un immense merci à mes encadrants, Alexandre et Thomas, pour leur soutien indéfectible tout au long de ces années. Vous m'avez guidé avec bienveillance et rigueur, en mettant à ma disposition toutes les clés pour que cette thèse se déroule dans les meilleures conditions. Ce fut un véritable plaisir de travailler à vos côtés. Merci également à Mikael pour ton précieux accompagnement au cours de ma première année, j'ai beaucoup appris grâce à toi. Je remercie l'équipe du LAPP, du LISTIC pour ces merveilleux moments passés à vos côté. Merci à tous les opérateurs du LST-1 qui m'ont accompagné et avec qui j'ai partagé cette expérience inoubliable. Merci enfin à toute l'équipe de MUST pour votre disponibilité et votre travail remarquable.

Bien entendu, je remercie ma famille, pour votre soutien constant et vos paroles réconfortantes, qui m'ont accompagné à chaque étape. Merci pour tout ce que vous avez fait pour moi, souvent sans que j'aie besoin de demander. Je vous dois une part de cette réussite.

Pour finir, j'aimerais adresser une pensée toute particulière à ma chère Gelin. Ta présence, constante et inestimable, m'a apporté la sérénité nécessaire pour traverser ces trois années. Tu as été ma plus grande force tout au long de ce parcours, qui n'aurait pas été le même sans toi à mes côtés. Maintenant, une nouvelle page s'écrit pour nous, à l'autre bout du monde !

Contents

List of acronyms

- AE: Auto-Encoder
- CE: Cross-Entropy
- CNN: Convolutional Neural Network
- CTAO: Cherenkov Telescope Array Observatory
- CV: Computer Vision
- DNN: Dense Neural Network
- EAS: Extensive Atmospheric Shower
- GPU: Graphics Processing Unit
- HEGRA: High-Energy Gamma-Ray Astronomy
- H.E.S.S.: High Energy Stereoscopic System
- IACT: Imaging Atmospheric Cherenkov Telescopes / Imaging Atmospheric Cherenkov Techniques
- LST: Large-Sized Telescope
- MAE: Masked Auto-Encoder / Mean Absolute Error
- MAGIC: Major Atmospheric Gamma-ray Imaging Cherenkov Telescopes
- MLP: Multi-Layer Perceptron
- NLP: Natural Language Processing
- NSB: Night Sky Background
- PMT: Photo-Multiplier Tube
- PSF: Point Spread Function
- RF: Random Forest
- SNR: Signal-to-Noise Ratio
- VERITAS: Very Energetic Radiation Imaging Telescope Array System

1 The science of gamma-ray astronomy

This chapter gives an introduction to the project that encompasses this thesis and provides a brief history of gamma-ray astronomy, along with the fundamental scientific questions that result from this discipline. Subsequently, it concisely describes the instruments that are used to complete these objectives and the workflow that have been currently selected. Finally, we present the underlying challenges and research directions that we address in this thesis.

1.1 Introduction

The birth of gamma ray astronomy

The early history. The late 19th century witnessed for the first time the materialization of the concept of cosmic radiation. Following the work of Charles T. R. Wilson on the ionization of the atmosphere, this radiation was categorized as extraterrestrial gamma rays [Mar91]. Thanks to the active research on radioactivity and the production of ionization chambers during this era, measures of radiation revealed that a residual amount could persist even in the absence of any radioactive substances. The question of its source arose and the prevailing theories attributed its origin either to the surface of the Earth or the atmosphere. Yet, anomalies have been observed both at the top of the Eiffel Tower and in undersea recordings. In order to validate or refute the current hypotheses, Victor Hess conducted, between 1911 and 1912, a series of hot-air balloon ascensions to measure the rate of radiation in high altitudes (Figure 1.1). Against all odds, he found out that the rate increased fourfold during the ascensions. He concluded that radiation is emitted from space, describing it as the "radiation from above" in his published results. The Austrian physicist finally clinched the Nobel prize in 1936 for his discovery. The nature of these rays were at this moment completely attributed to gammas, until the analysis of their flux in 1927 revealed their true identity.

Figure 1.1: Victor Hess taking place in a hot-air balloon to conduct a series of experiments.

Current situation. Nowadays, modern physics teaches us about the nature of cosmic rays. Illustrated on Figure 1.2, Hadrons compose around 90% of its the flux, and are mostly protons, but also neutrons to a lesser extent, plus their corresponding antiparticles. Leptons represent the second most common particles and are constituted primarily of electrons and positrons. Finally, photons are the rarest entity and have a ratio to protons that is usually below 10^{-4}. Due to the influence of extra-galactic magnetic fields on charged particles, leptonic and hadronic rays deviate from their trajectory towards Earth. Unless the source is positioned very close to Earth, or the energy of the particle is extremely high, these deviations makes the tracking of their source impossible. However, photons are uncharged particles travelling in straight lines and the study of their high-energy representatives is called gamma-ray astronomy. One hundred years after their discovery, gamma rays appear to be an essential messenger in investigating space to test the current understanding of astrophysics.

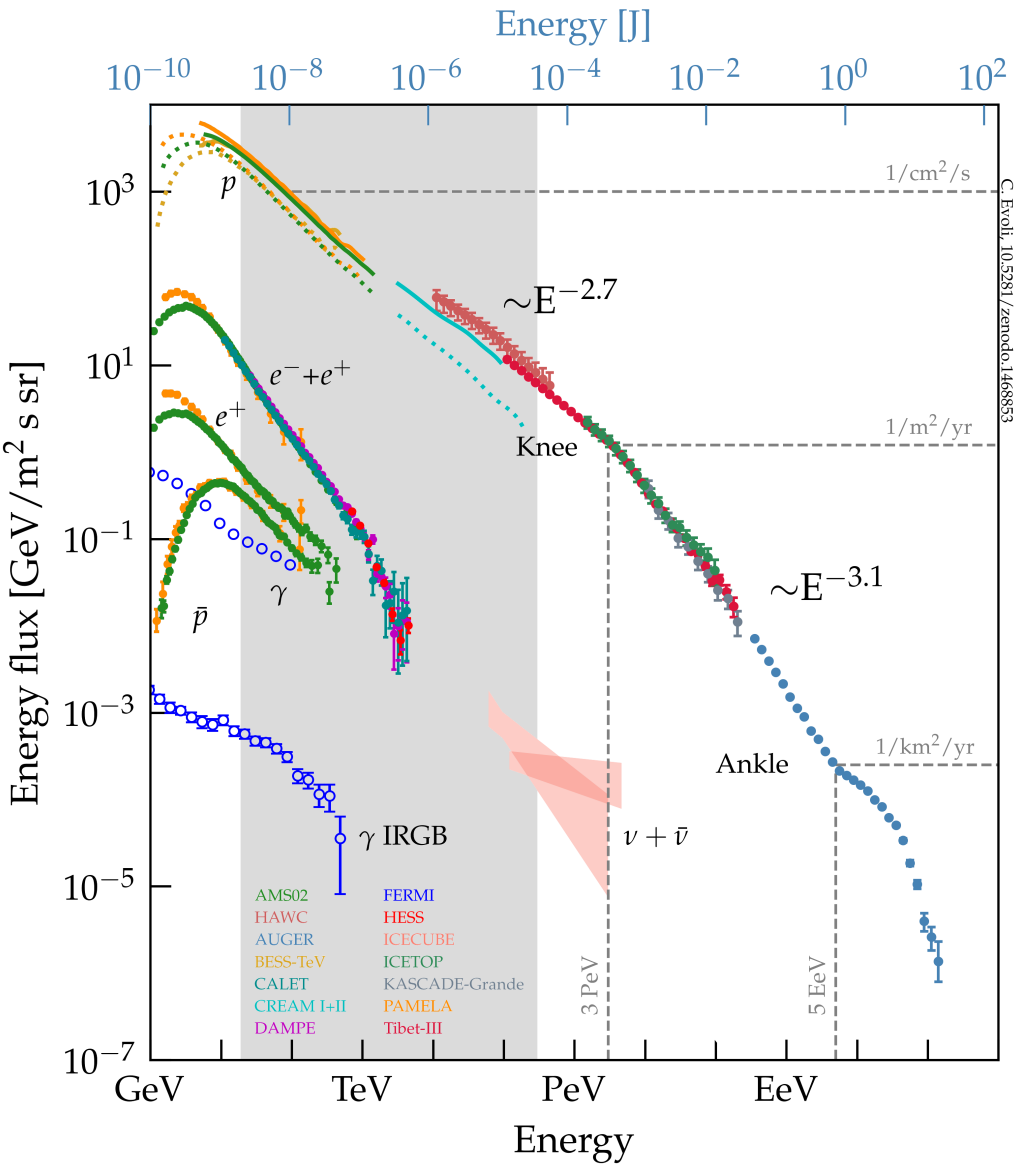


Figure 1.2: The energy spectrum of the all-particle cosmic rays (electron e^-, positron e^+, gamma γ, proton p) as detected from the main instruments. As the energy E increases, the energy flux decreases proportionally to $E^{-\Gamma}$, where the spectral index Γ depends on the energy level considered. The complete Cherenkov Telescope Array Observatory energy range of detection is highlighted with the gray surface and spreads out from 20 GeV to 300 TeV. The top x-axis refers to the energy in the Joule (J) system but it is more common in astrophysics to refer to it in the electron-volt (eV) system. Knowing the distribution of the cosmic particles gives a feedback on the realism of the reconstructed distribution of detection algorithms. It is noticeable that the greater the energy, the lesser the flux. Two factors allow to increase the capability detection, which are the surface area of the telescope, and the exposure time. Source [Jac20].

Revealing the invisible

Over the last centuries, physicists racked their brains over the understanding of the true nature of electromagnetic radiations. Remarkably, visible light, what the humans can see, coincide by essence to any other type of electromagnetic waves, and their distribution can be represented through a mathematical tool according to their frequency or wavelength, namely the electromagnetic spectrum. In fact, the composition of the atmosphere surrounding the Earth makes it partially or totally opaque to some range of light frequencies. Consequently, the human eye evolved to be sensitive to the specific wavelengths at which the atmosphere exhibits the highest degree of transparency. But the comprehension and analysis of the other segments that extend beyond this specific range, that is thus inaccessible to us, can be very informative on the processes at stake in many physics phenomena. Depicted in Figure 1.3, the electromagnetic spectrum expresses the idea that only differs the wavelength - or inversely the frequency - of the entity, and thus the energy. In order to reveal the invisible, the late decade has witness the emergence of multiple generation of "augmented eyes", or observatories, such as the Cherenkov Telescope Array Observatory (CTAO). In particular, this project aims to detect the highest energy photons. In short, gamma-rays are photons, such as the visible light, but with a much, much greater energy.

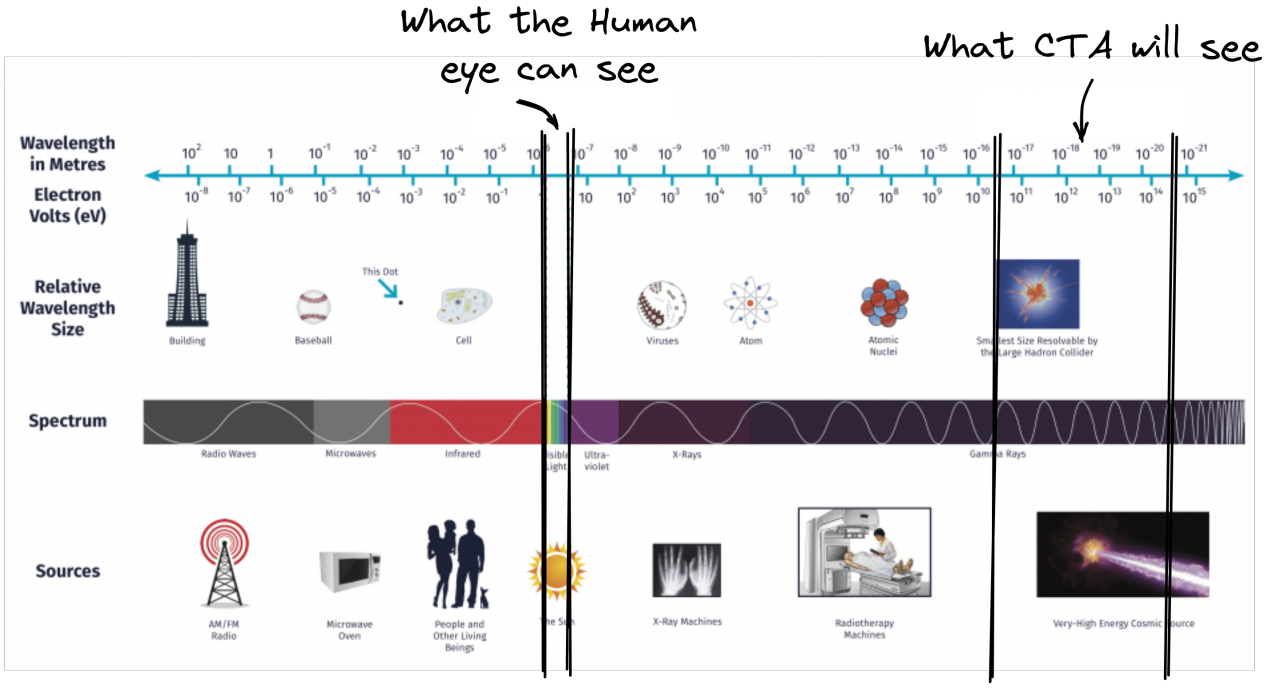


Figure 1.3: The electromagnetic spectrum reveals a great variety of radiation, ranging from the lowest frequencies corresponding to the radio and microwaves, to the open-bound highest frequencies referring to X-rays and gamma-rays. In between are presented the most commonly known wavelengths, namely the infrared, the visible light and the ultraviolet. They can be equivalently quantified using their energy: an electron-volt (eV) is a unit energy corresponding roughly to the visible light (from 2 to 3 eV), and humans can sense wavelengths generally between 400 to 800nm. On the contrary, the newly built CTAO telescopes will detect very high energy gamma rays above 20 GeV, that is to say energies more than billion times greater than the visible radiations. Extracted from the CTAO website[1], modified.

The electromagnetic spectrum offers the opportunity to observe a single phenomenon with multiple points of view, as collecting light of different frequencies from a celestial object allows to gather much more information than to only consider a limited range of energy. This technique is known as multi-modality, or multi-wavelength analysis in astrophysics, and it has been already widely used for many terrestrial purposes using satellites, e.g. automatic land classification, land-cover segmentation, disaster damage assessment, change detection, disaster monitoring and early-warning systems. For example, Figure 1.4 shows the Crab nebula measured at six different wavelengths.

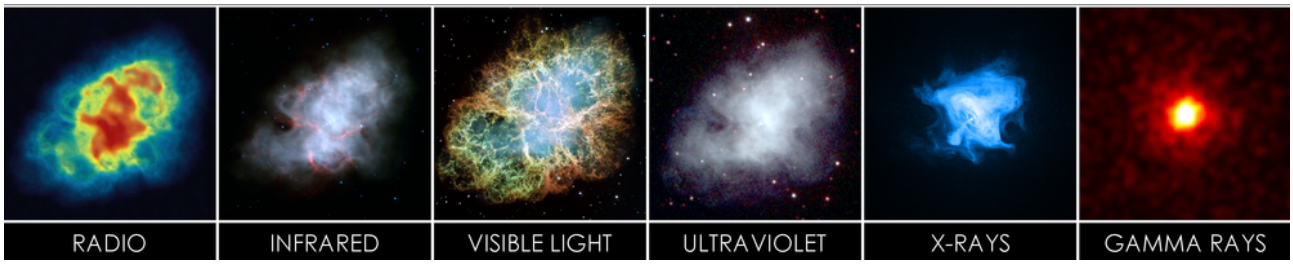


Figure 1.4: The Crab nebula observed from different frequencies. The detection of multiple wavelengths shed light on a multitude of information that can be analysed in order to understand and update the scientific knowledge. Some channels might reveal structural information, while others are more suitable to highlight processes at stake within the object. Extracted from the CTAO website[2].

Although the current generation of gamma-ray telescopes has led to many discoveries, their limited sensitivity, defined as their ability to detect a gamma-ray flux over comparable exposure times and surface areas, suggests that they are approaching the limits of their discovery potential. With a larger ground footprint and with the incorporation of next-generation detectors, the CTAO will enable to achieve nearly ten times the sensitivity of existing instruments, in addition to offering enhanced spatial resolution.

Gamma-ray astronomy

Gamma-ray astronomy refers to the observation of the universe in the gamma-ray segment of the electromagnetic spectrum, encompassing photons with energies typically greater than 100keV, as defined in [Mar91]. Analogously, Very High Energy (VHE) gammas corresponds to particles with an energy generally greater than a few hundred TeV or a few GeV [VB09a]. It is perhaps the least studied segment compared to the radio and x-rays, thus the understanding and detection of known sources was rather shallow until recently. Nowadays, more than 250 very high gamma-ray sources of both galactic and extra-galactic origins have been detected [WH08]. The study of gamma-ray astronomy holds enormous potential because these particles are the witness of the most violent mechanisms occurring in the cosmos, produced in region of the universe with high temperature, density, and electromagnetic fields. Their neutral nature - uncharged particles - facilitates their direct observations as their trajectory is not perturbed by magnetic fields, but the difficulty lies in the low flux levels of the sources, with photon counts ranging from 10^{-4} to 10^{-6} per cm^{-2} s^{-1}, requiring

[1] https://www.ctao.org/

the detectors to have substantial sensitive areas and extended exposure times. Gamma rays spring when cosmic rays interact with ambient matter, magnetic fields, or photons.

Since the beginning of the observation of the universe through the eyes of spatial and terrestrial instruments, various types of gamma ray sources have been discovered and categorised by where they are located in the universe: within our galaxy, or in external galaxies. Figure 1.5 is a non-exhaustive list of plausible or detected gamma-ray sources.

Although gamma-ray astronomy is a historical science, the fostering of the theory and new investigations brought more unsolved interrogations. CTAO will bring a fresh perspective with the hope of answering some of the main current astronomical questions at stake among the scientific community. Author of [Hof18] classifies the contribution into a couple of broad categories.

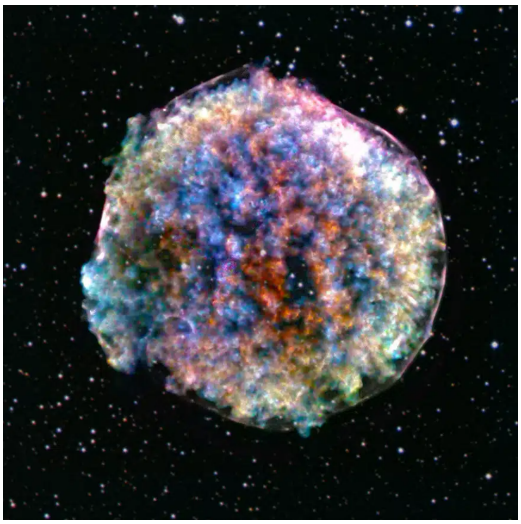
- Understanding the origin and role of relativistic cosmic particles. Cosmic rays simply refer to atomic particles such as proton or electron, but accelerated to tremendous energies that make them travel near the speed of light. Even though they continuously bombard the Earth's atmosphere since the dawn of time, they only recently drawn the attention of the physicists. The origins of cosmic rays, along with the phenomena triggering their acceleration or also their role in the birth of stars and in the evolution of galaxies are still debated topics.

- Probing extreme environments: Black holes are mysterious sphere of darkness that engender an unimaginable force of gravitation. Although nothing - not even light - can escape their event horizon, the accumulation of matter around them is a formidable source of very high energy photons. Besides, as a lighthouse guiding the physicists towards a unified field theory, they may play a key role in the understanding of our universe. Researchers are investigating the possibility that gamma rays may give clues on the physics surrounding black holes or neutron stars.

- Multi-messenger analysis: Multi-messenger astronomy refers to the studying of astrophysical objects and phenomena using a combination of multiple messengers such as electromagnetic radiation, neutrinos, cosmic rays, and gravitational waves. Recently, simultaneous observations of gravitational waves and electromagnetic signals from neutron stars merger [17], as well as observations of gamma rays and hypothetically neutrinos emitted by a flaring blazar [18], encouraged multi-messenger astronomy to become a prominent field in astronomy.

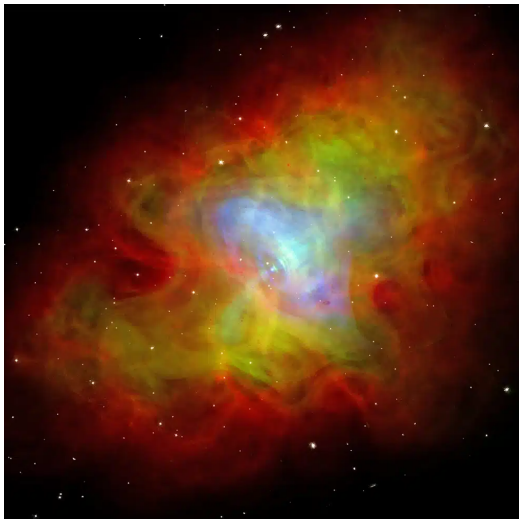
- Exploring frontiers in physics: The nature and the distribution of dark matter are a mystery that may partially be unveiled through the eyes of gamma-ray astronomy. Additionally, quantum gravity effects might cause time delays between photons of different energies traveling across cosmological distances, suggesting a deviation from Einstein's Special Relativity known as Lorentz Invariance Violation. By measuring gamma rays from exceptionally bright sources, the CTAO will be able to investigate these effects.

[3] https://www.ctao.org/

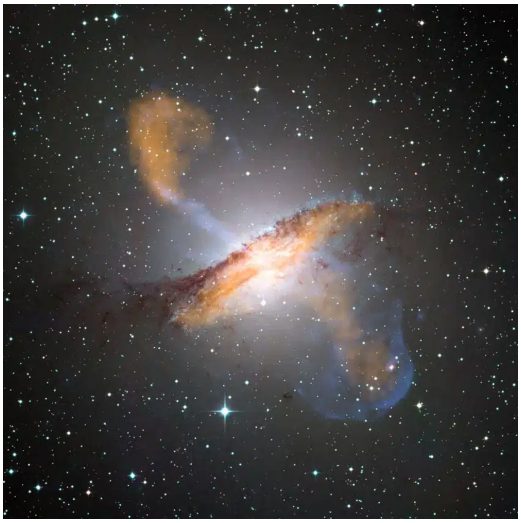


Supernova remnants (SNR) designate the expanding and evolving gas clouds observed after the death of a star. When it reaches the end of its life cycle, stars with enough mass can undergo catastrophic explosions, releasing an immense amount of energy and matter into space. Located in the Perseus Arm of the Milky Way, the Crab Nebula is one of the most famous supernova remnant, and is a widely studied astronomical object with a well-known spectrum. RX J1713.7-3946 is another type of SNR that exhibits the characteristic shell morphology. It was also the first SNR to be detected in very high-energy gamma rays [Abd+18].

When a star collapses at the end of its life, the remnant core may turn into a neutron star. A **pulsating radio source**, or pulsar, refers to a highly magnetized, rotating neutron star that emits beams of electromagnetic radiation from its magnetic poles. Their observation occurs exclusively when a beam is oriented toward Earth. From this directional emission arises a pulsed appearance. Neutron stars, characterized by high density and rapid, regular rotational periods, generate precise intervals between pulses, spanning from milliseconds to seconds. When the ultra-fast wind of particles expelled by pulsars collides with the surrounding medium, it creates a shock wave that accelerates particles. These particles spread out, forming a pulsar wind nebula (PWN), and emit radiation that can reach gamma-ray energies. The Crab Nebula, a supernova remnant, is more precisely classified as a PWN.

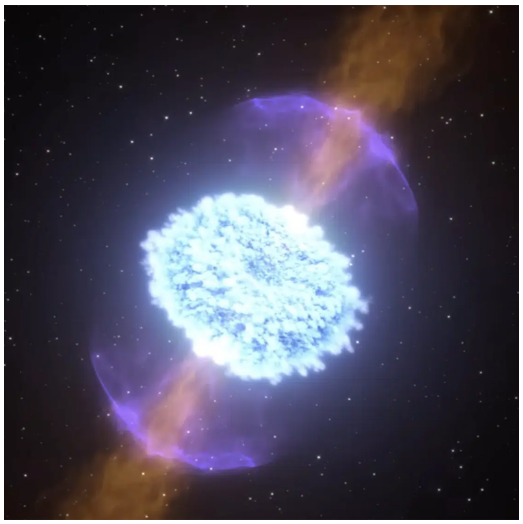


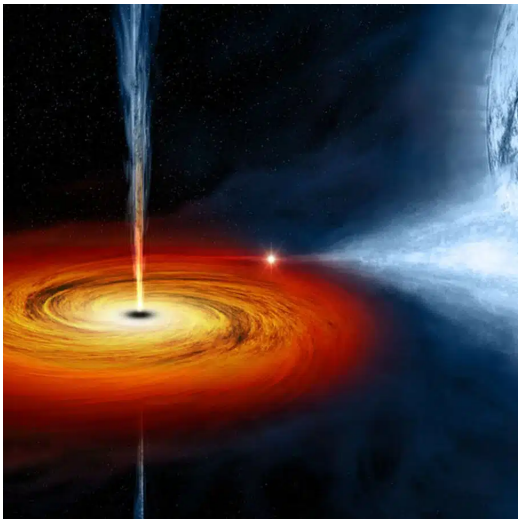
Like their death, the birthplaces of stars are also interesting for gamma-ray astronomy. In fact, their birth conditions largely dictate their lifespan and behavior. This is why researchers focus on studying star-forming regions, described as **dense molecular clouds**, where the conditions are conducive to gravitational collapse and protostars formation. Cosmic rays accelerated locally in or near the clouds, by for instance SNRs, collide with the material in the clouds and produce secondary particles that decay into gamma rays. Therefore, CTAO observations of these systems are expected to provide insights into the relationship between high-energy particles and star formation.



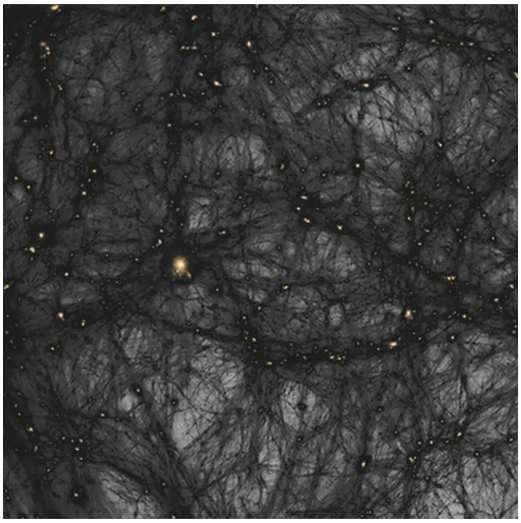
An **active galactic nucleus** (AGN) is a super-massive black hole (more than million solar masses) with an accretion disc and jets. They emit particles with energies covering the entire electromagnetic spectrum from radio to VHE gamma-ray. A galaxy harboring an AGN is referred to as an active galaxy. If the emitted relativistic jet is pointing towards Earth, they are referred to as **blazar**. Their flux is highly variable, frequently undergoing rapid and substantial fluctuations in brightness over short timescales (from hours to days). Markarian 501 and BLLac are two famous and bright blazars radiating at the TeV energy range and are great school cases to test and compare new detection baselines.

Gamma-ray bursts (GRB) are characterized by the emission of a brief yet intense flux of gamma-ray particles (tens of milliseconds to a few minute). They were accidentally detected for the first time in 1967 with the launch of the Vela satellites designed to identify artificial nuclear explosions from space [Mar91]. Their origin remained elusive for a long time, and the difficulty of their observation resides from the brevity of their emission. Nowadays, they can be attributed to events like hypernovas or mergers of compact binaries. Initially believed to be confined to an energy range ranging from tens of keV to a few MeV [Pir99], recent observations have revealed the existence of higher-energy GRBs [19b; 19c; Abd+21]. They are uniformly distributed in the sky, within or outside of the galactic plane.



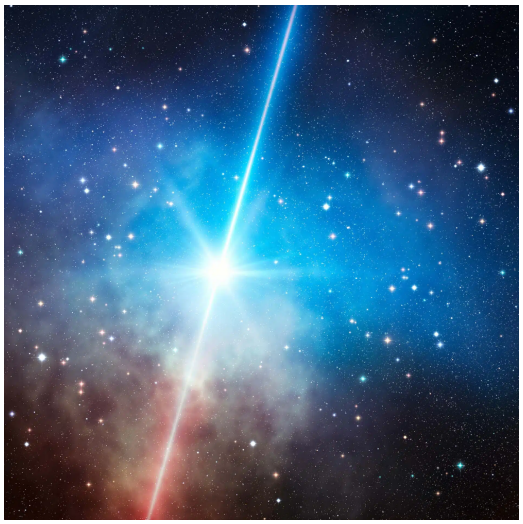


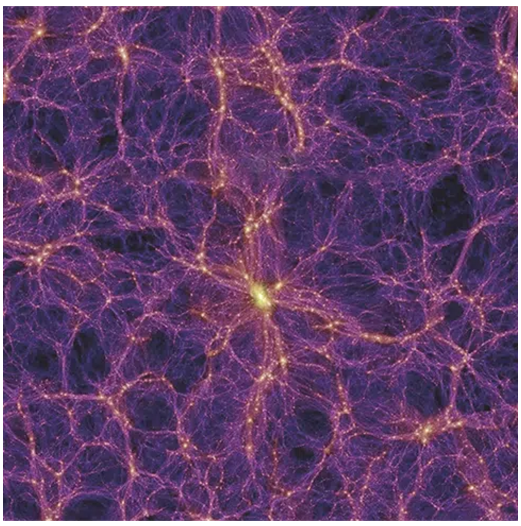
Binary systems are singular structures composed of two objects orbiting each other closely. Star-based organizations produce powerful stellar winds that can collide and accelerate particles, resulting in gamma-ray emissions. If one companion in the binary system is a compact object, such as a black hole or neutron star, it can pull in material from the stellar companion via an accretion disk, emitting jets that accelerate particles and produce gamma rays.



Dark matter is a form of matter that makes up about 27% of the universe, yet it does not emit, absorb, or reflect light, making it invisible and detectable only through its gravitational effects. The CTAO aims to detect dark matter by searching for gamma rays emitted when its particles, presumably weakly interacting massive particles (WIMPs), annihilate each other. While current instruments lack the sensitivity to detect the signals predicted by these models, the CTAO may achieve this crucial level of sensitivity [Abe+24].

Quantum gravity is a field of theoretical physics that seeks to describe gravity according to the principles of quantum mechanics. One remarkable effect is the creation of time delays between photons of different energies traveling across vast distances, suggesting a deviation from Einstein's Special Relativity known as Lorentz Invariance Violation [Bol+22; Col+24]. By measuring gamma rays from exceptionally bright sources, the CTAO will be able to investigate these effects with unprecedented statistical precision.





The vast majority of the universe seems empty. Matter is concentrated in distinct structures, like galaxy clusters or super-clusters, located in filaments. There is currently no consensus about the extent of emptiness in the voids separating filaments, but they may contain relics from the universe's earliest moments. For this purpose, the CTAO will conduct indirect studies of the extragalactic background light (EBL), which fills the space between galaxy clusters. EBL photons causes gammas to be absorbed over long distances, leading to a softening of the spectra of distant objects like AGNs or GRBs. This process can be measured to enhance our understanding of the universe's formation.

Figure 1.5: Gamma-ray sources catalog. The images depicted are not gamma-ray images but artistic representations or observations from X-ray and optical telescopes. Images are extracted for the CTAO website.[3]

1.2 Imaging Atmospheric Cherenkov Technique

From an experimental point of view, the localization of the gamma-ray detector - ground-based or in space - and the technology deployed dictate the detection strategy of the incident particles. In the case of the CTAO, telescopes are located on the Earth surface. In fact, the construction of space-based observatories capable of directly detecting VHE gamma rays is extremely challenging. The limitation mainly comes from the size of the embedded detectors and the exponentially decreasing energy flux as the energy grows, which would require impractically long exposure times. Furthermore, ground-based telescopes take advantage of the atmosphere to indirectly detect these VHE gamma rays over a large area, making them more efficient and practical for energies greater than a few GeV; thus the acquisition system relies on indirect observations through a particle shower mechanism.

The particle shower mechanism

Direct observation of gamma rays is highly compromised when conducted on the Earth surface because the atmosphere is opaque to this portion of the electromagnetic spectrum. The detection trick consists of using the atmosphere as a calorimeter. When they start penetrating the atmosphere, the cosmic radiation interacts with it. This leads to the development of showers of secondary particles within nanoseconds. Purely electromagnetic shower, illustrated in Figure 1.6, can be determined by two fundamental processes, namely Bremsstrahlung ([BH34]) and pair production. The former corresponds to the conversion of a gamma into an electron and a positron. The latter is a process that occurs when a charged particle, like an electron, is slowed down by another charged particle, and this lost energy is emitted as electromagnetic radiation, like gamma rays. Because the resulting secondary particles are travelling faster than the speed of light in the air, resulting in a blue flash equivalent to the supersonic boom when an object moves faster than the speed of sound. This emitted light is called Cherenkov light, in tribute to Russian physicist Pavel Cherenkov, who discovered this phenomenon in 1934. A few years after his finding, Frank and Tamm reported in [FT37] the eponym equation describing the amount of Cherenkov radiation emitted from this mechanism. Depending on the nature of the particle entering the atmosphere, cascades may vary in shape, and generally covers a patch of around 100m^2 of ground surface. Noticeably, the division process of a gamma-ray into a pair of electrons and vice-versa is continuously repeated, underscoring that the only difference between a shower produced by a gamma and an electron is the first interaction, which makes them very difficult to distinguish. On the contrary, as depicted in Figures 1.7 and 1.8, gamma-induced and proton-induced showers can more easily be differentiated with the emitted sub-particles and the shape of the cascades, although the incident energy of the incident ray is the same. Furthermore, the energy of the particle is directly related to the spread of the shower and the amount of Cherenkov light that will be emitted. This shows that there is an intrinsic limit of the energy of the particle that can be detected, as lower energy incident particles may not trigger the telescope detection systems.

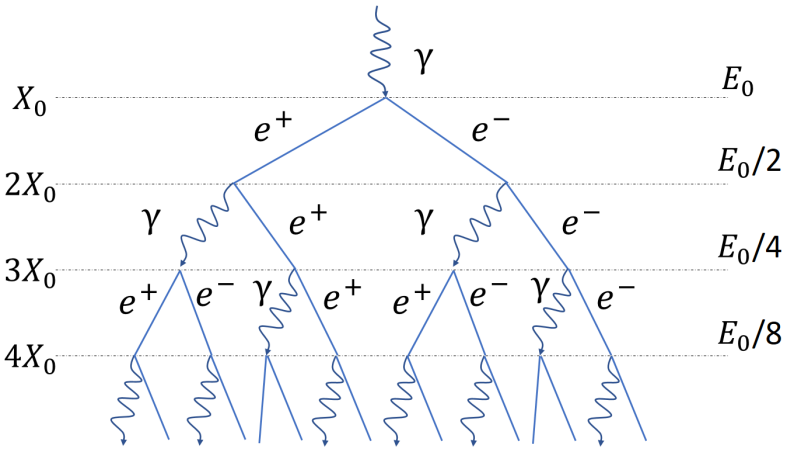


Figure 1.6: Simplified model of an electromagnetic particle shower for a gamma particle interacting with the atmosphere. X_0 corresponds to the altitude of first interaction, where the incident gamma is sub-divided into two electrons and the overall energy is halved. Source [Lem04].

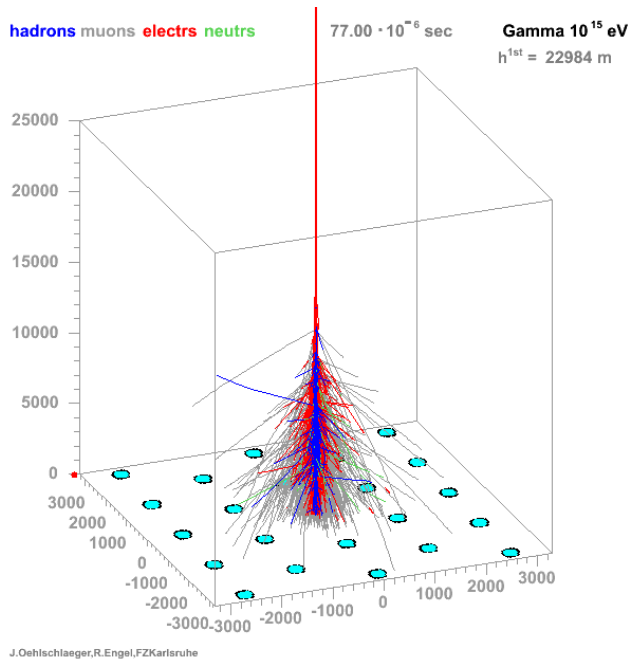


Figure 1.7: Simulation of a gamma shower. Image extracted from the CORSIKA website[4].

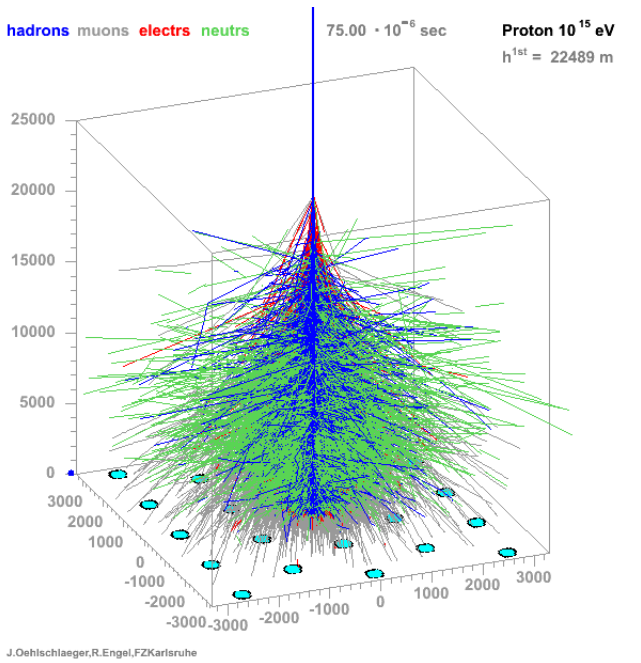


Figure 1.8: Simulation of a proton shower. Image extracted from the CORSIKA website[5].

Modern space-based and ground-based instruments

The gamma-ray astronomy theory preceded the experimentation, and the pioneering work that laid the foundation for the Cherenkov light detection mechanism was carried out by Galbraith and Jelley in 1953 [GJ53]. As a proof-of-concept, their rudimentary setup consisted simply in a 25cm diameter parabolic mirror of a short focal length placed inside a

[4] https://www.iap.kit.edu/corsika/index.php

dustbin, and a 2-inch photo-multiplier tube positioned at the mirror's focal point. With this installation, they managed to detect Cherenkov light flashes emanating from air showers. Subsequent advancements in optical systems, hardware and detection strategies have led to the establishment of the first ground-based observatories dedicated to very-high-energy gamma-ray observations [Mir22]. The first space-based instrument investigating high-energy gamma rays was carried abroad NASA's first fully dedicated gamma-ray satellite, Explorer 11, in 1961. Nowadays, the last generation of cutting-edge gamma-ray detectors is paving the way to a new vision of the universe at energies that have never been seen before. We give an overview of the functioning of each technology along with their strength and drawbacks.

Space-based telescopes

Space-based observatories have the advantage of being able to observe regions of the electromagnetic spectrum that are otherwise absorbed by the Earth's atmosphere. As they are located above the clouds, they are not affected by weather conditions, neither by the day and night cycles. In addition, they dispose of a wide field of view. However, because it is expensive and difficult to transport materials in space, they have restricted detection areas ($\sim 1\text{m}^2$), and an angular resolution of a few degrees. Given the scarcity of gamma-ray particles, the small size of the detector reduces the number of possible observations.

Between 1990 and 2003, NASA began a remarkable expedition by launching a series of four cutting-edge space-based telescopes known as the Great Observatories [79]. Each satellite embarked a unique technology, allowing their combination to cover the electromagnetic spectrum from infrared red to gamma ray. As one of the main component of the program, the Compton Gamma Ray Observatory (CGRO) aimed to study the gamma-ray segment in the 20 MeV to 30 GeV range using the EGRET instrument. Nowadays, the most notorious gamma-ray satellite is undoubtedly the Fermi spacecraft, that carries on board the Large Area Telescope (LAT) that aims to detect gamma rays from space [Abd+10]. The anticoincidence module is the first layer of the detector and filters charged particles, also designated as the background noise, thus only uncharged particles will enter the following components. At the next level, the principle of detection is based on the production of a pair of electron-positron through the interaction of the incoming particle and one of the tungsten sheets present onboard. The energy is measured using a calorimeter and produces a flash of light proportional to the particle energy, whereas the arrival of direction is measured with the tracker instrument. Fermi-LAT can detect gamma-rays ranging from 8 keV and 300 GeV but its effective collection area corresponds at most to the detector surface, which approximately worth 0.65m^2.

Ground-based telescopes

There exists multiple detection techniques for ground-based telescopes in the context of gamma-ray astronomy. In this thesis, we are interested in Imaging Atmospheric Cherenkov Telescopes (IACT), instruments equipped with camera able to detect VHE gamma ray photons - above tens of GeV - through imaging their Cherenkov light produced by their corresponding particle shower at the nanosecond scale. IACT celebrated its first success in 1989 with the detection of the Crab nebula from the Whipple observatory in Arizona [Wee+89]. Compared to space-based detector, these instruments offer a detection area above $10^5 m^2$, thus the possibility of observing very faint flux of particles within a reduced observation window. Yet, technological, financial and physics (irreductible background) limitations constrain the sensitivity range of energy of the detectors. Throughout the history of their constructions and the technology employed, they can be divided into four generations [Mir22].

Figure 1.9: The Whipple observatory. Image credit[6].

Among the many early generation of telescopes that have seen the light of day, the first to make a positive detection of a gamma-ray source is the Whipple observatory (Figure 1.9). In 1967 began on Mount Hopkins, in the United States, the construction of its first detector, consisting in a 10m diameter telescope. In 1968, the first pioneer publication of the observatory stated the observations of 13 gamma-ray source candidates. At this time, the Crab Nebula was already known in the optical, radio, and X-ray segments, but the observations did not lead to any detection in the gamma portion. Technological progresses improved the instrument over time, characterized by fast PhotoMultiplier Tubes (PMTs), a large reflector diameter providing a relatively low detection threshold, and a 37 pixel camera producing images on an hexagonal grid of pixels. Finally, the Whipple 10-meter gamma-ray telescope detected the Crab nebula for the first time in the gamma-ray segment with a significance of 9σ in 1989, and had a momentous impact on gamma-ray astronomy. However, it wasn't just improved electronics that made this possible, but also the development of new analysis techniques that relied on computational power and algorithm engineering. Thenceforth, the field started to become computer-dominated.

From the basis of this well-known and proven technologies, arose the goal of conducting detection in stereoscopy [Mir22]. For this purpose started the construction of the High-Energy Gamma-Ray Astronomy (HEGRA), the pioneering atmospheric Cherenkov telescope dedicated to exploring the gamma-ray universe with multiple telescopes [Koh+96], it began its mission in 1987, collecting invaluable data over the course of 15 years until 2002. Constituted by 5 small-sized telescopes, the observatory was limited in low-energy sensitivity. As

[6] https://www.mmto.org/the-story-of-the-observatory/

a direct successor, and with the wish to increase sensitivity at this energy segment, the High Energy Stereoscopic System (H.E.S.S) was built in the Namibian desert [Hin04]. Composed of four 12m-diameter and one 28m-diameter telescopes, it is fully operational since 2003, and has detected dozen of gamma-ray sources and allowed to confirm the hypothesis that supernovae may accelerate particles up to 10^{14} eV. Similarly, with the will of exploring the unknown sub-100 GeV part of the gamma-ray spectrum, the Major Atmospheric Gamma-ray Imaging Cherenkov Telescope (MAGIC) [Bai03], located on the island of La Palma on the same site as HEGRA, has been designed for this complex task. With 2 telescopes working in stereoscopy, it is operational since 2004, and in pairs since 2009. In addition, the Very Energetic Radiation Imaging Telescope Array System (VERITAS) was constructed in Arizona at the Whipple observatory and the four detectors are fully operational since 2007 [Bra+99].

Motivated by the success of the previous observatories, CTAO project was initiated in 2006 with the aim of developing the next generation of IACT, and greatly improves the sensitivity in comparison to its counterparts [19a]. After completion, over 60 telescopes will be assembled and dispatched on both hemispheres: the northern site on La Palma island (Figure 1.10) and on the southern site in Chile. This combination will allow to cover almost the entire sky (Figure 1.11).

The design of the telescope array layouts is a trade-off. The greater the distance, the better the stereoscopic reconstruction, but the lower the probability of detection of a low energy particle. Last but not least, although the ground footprint of the particle shower doesn't significantly differ depending on the incoming particle energy, low energy events will be too faint to be detected if the telescopes are distant from the impact point. Thus, the LSTs are located in the center so they can work conjointly for the detection of low energy particles. It is worth noticing that the northern site will not contain any SSTs, thus it will mainly focus on the detection of low to mid-energy gamma rays, ranging from 20 GeV to 5 TeV. In fact, the north site is optimized for observing extragalactic sources, which generally emit lower-energy gamma rays. Since SSTs are specialized for detecting VHE particles, which are more commonly produced by Galactic sources, they are more suitable for the south site, which offers a better view of the Milky Way.

The expected energy coverage of the CTAO ranges from 20 GeV to 300 TeV, which will be achieved using three different sized telescopes:

- The Large-Size Telescope (LST) is the largest instrument of the project with a aperture of 24m. It is designed to detect low-energy particle showers, as low as around 20 GeV. Each images produced have 1855 pixels. Their will be 4 LSTs deployed in total, all built on the north site, each covering a effective collection area of 370m^2. The first LST (LST-1) has been inaugurated in October 2018, and has passed the commissioning stage since the end of 2023. It is currently the only instrument in operation.

- The MST, with a 12-meter aperture, operates in the 150 GeV to 5 TeV range. Various versions of the MST exist and they will be equipped with cameras having 1764 or 1855 pixels. The northern and southern sites will host 9 and 14 MSTs, respectively, each system corresponding to an effective collection area of 88m^2.

- The SST is the smallest detector, with diameters ranging from 4 to 7m, optimized for high energies up to 300 TeV. The cameras will produce images with 2048 pixels. A

total of 37 SSTs will be deployed at the southern site, integrating an effective collection area of approximately 5m^2 per unit. They are specifically designed to detect the highest energy EAS that occur less frequently than their lower-energy counterparts. To boost the chances of capturing these rare events, a greater number of SSTs will be spread across several square kilometers at the CTAO-South array site.

Showers generated by the highest energy gamma rays are rare, as depicted in Figure 1.2, but they produce a significant amount Cherenkov light. Consequently, deploying a large number of SSTs covering a large area increases the likelihood of their detection. Conversely, the LSTs, fewer in number, feature large reflectors to capture the fainter but more frequent lower-energy flashes. The MSTs are designed to cover the intermediate energy range. Furthermore, satellite and ground-based observatories are complementary. While satellites use the great field of view and their high duty time to provide alerts of an imminent event, IACT observatories react promptly and follow if possible the observation to more precisely characterise the spatial phenomenon. Considering the energy ranges, Fermi-LAT overlaps at lowest energies covered by the LSTs, between a few dozen to a few hundred of GeV.

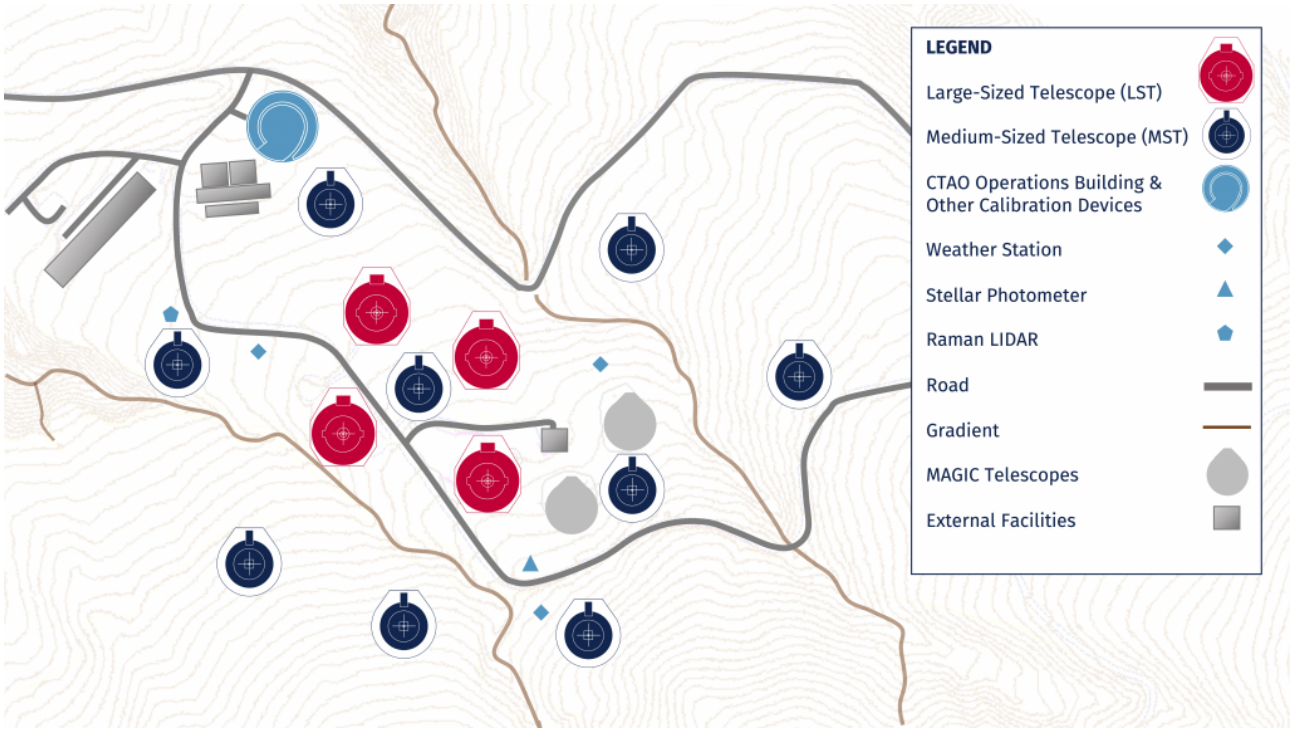


Figure 1.10: Located on the Spanish island of La Palma at an altitude of 2200m, juxtaposed to the outer rim of the caldera, CTAO-North benefits from the exceptional sky conditions. Extracted from the CTAO website[7].

From the photon to the electron

The telescope mirrors reflect and focus the emitted Cherenkov light on a camera, which can be seen as a collection of photo-multiplier tubes that are sensitive to a single photon. The fleeting emission of Cherenkov flashes - a particle shower fully deploys in nanoseconds -

[7] https://www.ctao.org/

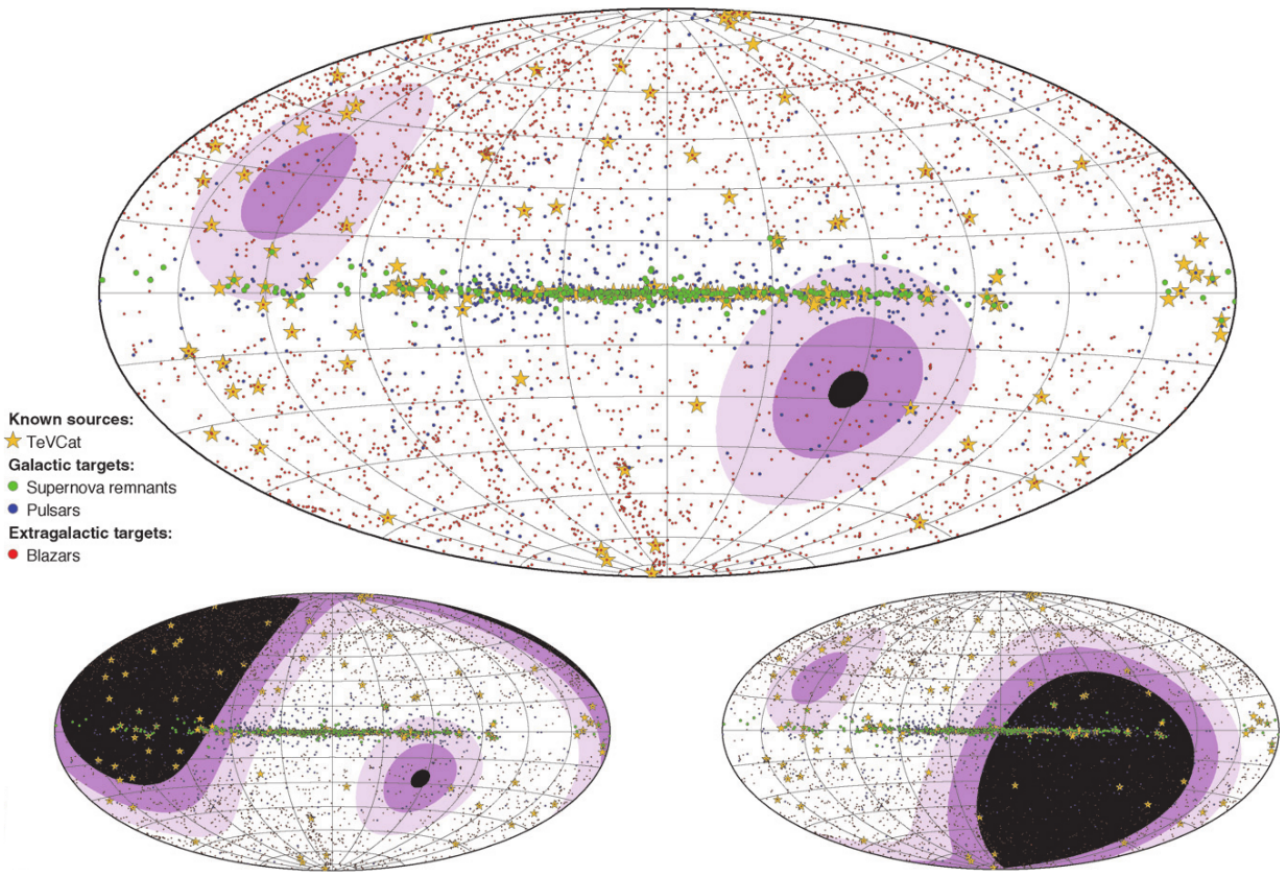


Figure 1.11: The sky coverage map in Galactic coordinates, with the galactic plane placed at the equator. CTAO south and north are respectively shown at the bottom left and right, and their combination corresponds to the top figure. The color scale indicates the minimum zenith angle under which a target is visible. CTAO works best at small zenith angles (below 30 degrees), while 60 degrees represent the practical limit. Source [Hof18].

imposes the cameras to capture images about a million times faster than a standard camera. This will be achieved using high-speed digitization and triggering technology capable of capturing shower images at a rate of one billion frames per second, with the sensitivity to resolve single photons, thanks to cutting-edge photo-multipliers that convert them into electrical signals. In such systems, the incoming light produces a cascade of electrons that are amplified using a high voltage, triggering a chain reaction. This sensitivity is remarkable but the quality of the observation highly depends on the meteorological conditions such as the presence of the moon in the sky. Two electrical signals are extracted from this amplification process, namely the low gain and the high gain. When the high gain is saturated, the low gain is used instead. The overall process of detection is illustrated by the Figure 1.12.

1.3 Detection workflow of the LST-1

The LST-1 is currently the only operational telescope of the CTAO. Because this thesis focuses on real observational data, we describe in the following the workflow associated with this type of telescope in particular.

The CTAO image analysis aims at synthesizing an event map that highlights the presence

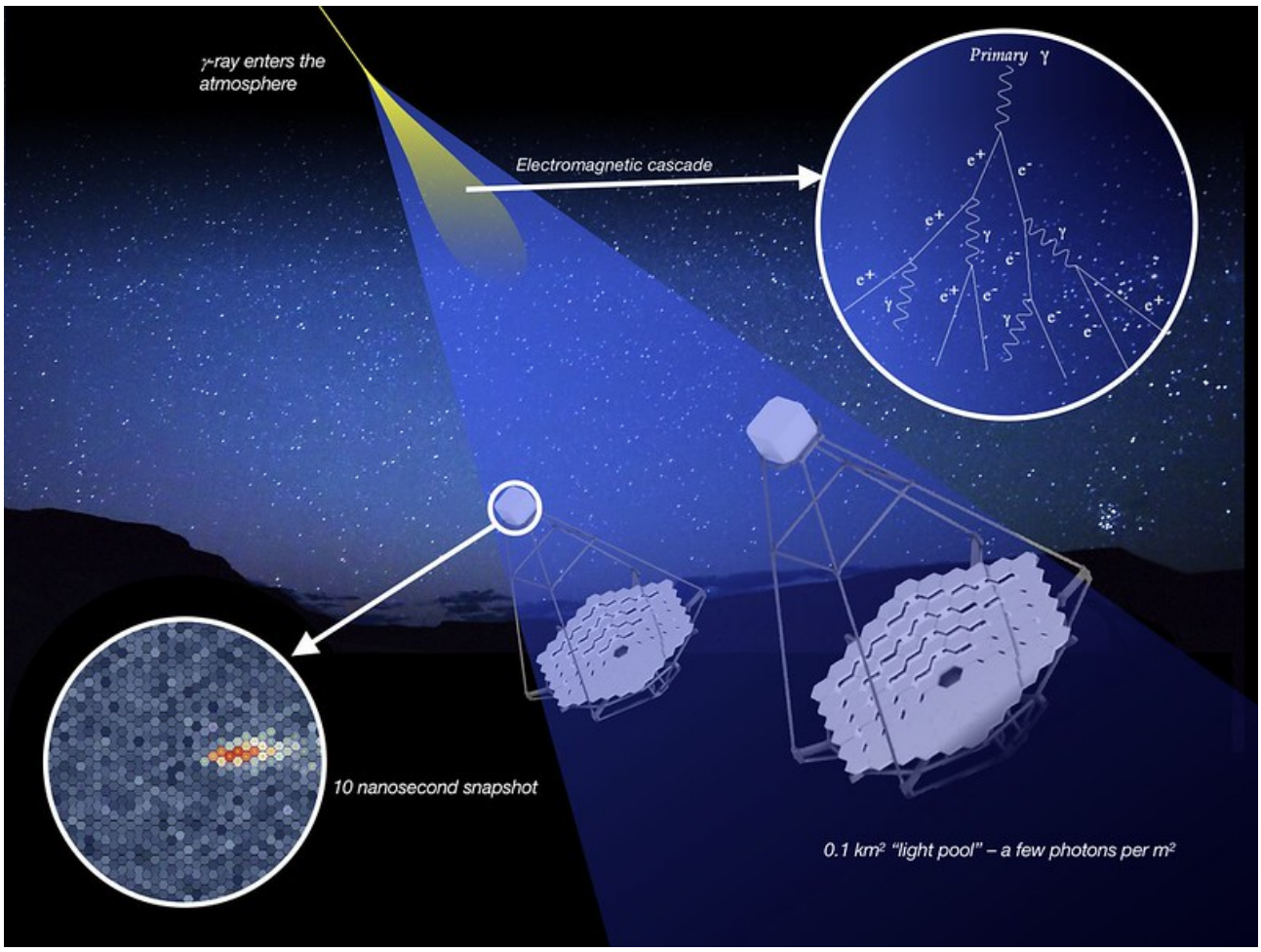


Figure 1.12: Principle of gamma detection. The incident photon produces a particle shower emitting Cherenkov radiation triggering the telescope detection which captures one snapshot every 10 nanosecond. This video sequence is integrated into a single image that will be further processed. Source: CTAO Flickr.

of gamma-ray sources in the observed portion of the sky. From the raw sequences to the exploitable data to be analysed, the acquisitions have to go through a multi-step procedure. The corresponding gamma reconstruction workflow is illustrated in Figure 1.13. On a broad scale, it can be categorized as a three-step procedure, respectively calibration, integration and reconstruction. The latter can be seen as a inverse problem. Using the images produced through the particle shower mechanism and the Cherenkov light emission, the goal is to retrieve the physical properties of the incident particle, that are the energy, direction of arrival and particle type. This means we are trying to infer the unknown cause (the incident particle's properties) from known effects (the observed images), which involves dealing with complex, high-dimensional data where traditional analytical methods may struggle. Machine learning, and particularly deep learning, excels at handling such complex relationships, learning patterns that are not easily captured by hand-crafted features.

As understanding the mechanisms occurring upstream any machine learning system training is crucial to highlight any potential bias, or to justify the choice of hyperparameters such as the datatype (bfloat, tensorfloat32, etc), we provide in the following an overview of each step. Yet, in our study, we do not have control on the calibration and integration

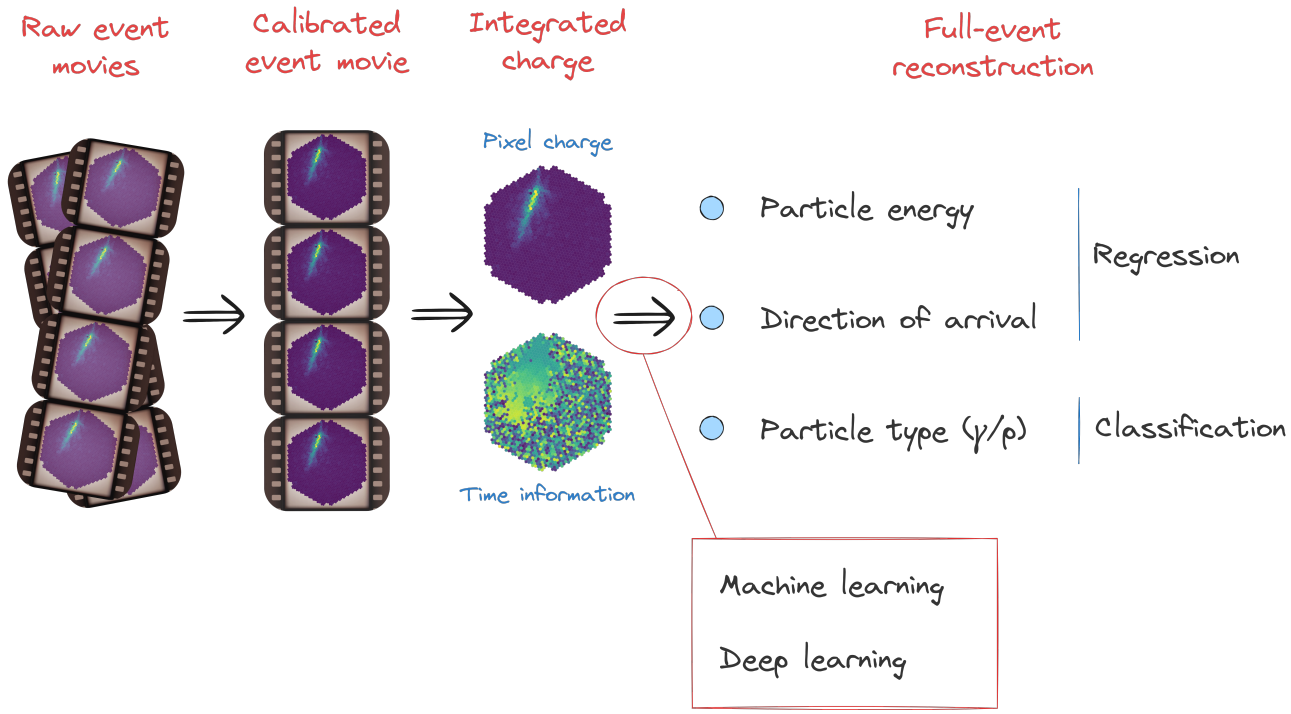


Figure 1.13: Workflow of the particle reconstruction based on IACT. The raw sequences are calibrated and integrated into a two images. Depending on the nature of the reconstruction algorithm, a cleaning and filtering procedure can be applied.

standardized processes.

Image calibration

When the Cherenkov image intensity produced by the incident particle surpasses the telescope trigger level, a sequence of about 40 images is produced at the rate of 1 image every nanosecond. These digitized waveforms are obtained for each pixel, with a memory consumption of 12 bit per sample. The camera requires to cover a large range of expected shower image intensities, which imposes the telescope to have two channels with different amplifications, a high and low gain channel, offering a wide dynamic interval from 1pe to 3000pe [al21].

Because of the diversity of observation and hardware conditions, a calibration has to be carried out. For this purpose, systematic processes are performed before data acquisitions and calibration images are taken throughout the night. Because they are subject to variations between themselves, each pixel of the camera is calibrated.

- Electronic systems come with electronic noise. This noise is measured once per night before the acquisitions with the camera's shutter closed, as the influence of any external factor should be avoided.

- Although telescope acquisitions are performed during the night, a residual amount of light from night sky persists due to external factors like stars, the moon, the zodiacal light, airglow, etc. Ultimately, these photons reach the telescope optical systems and mix with the Cherenkov signal of interest to produce a noise, that is referred to the Night Sky Background (NSB) in the astrophysics community. When observing the sky,

the NSB level varies significantly over time and depends on the coordinates of the targets. The galactic or extra-galactic nature of the source plays a significant role as many stars are visible on the Milky Way plan. In order to track the NSB fluctuations, interleaved pedestals are taken at the rate of roughly 100Hz and record events with no signal, containing only background noise.

- Flat-field correction is a standard calibration procedure for optical devices in order to compensate pixel-to-pixel sensitivity and distortions. Flat-field (FF) events are events produced by the Calibration Box (Calibox), uniformly illuminating the camera with the diffused light of a laser set to a specific wavelength of $\lambda = 355$nm. The interleaved flat-field events are also acquired at a rate of about 100 images per second.

- Telescopes cannot be approximated as flawless instruments and suffer from technical limitations. The point spread function (PSF) corresponds to the spatial distribution of light from a point source as it appears on the telescope's detector or image plane. It characterizes how a perfect point source of light is spread out due to various optical and instrumental factors, such as diffraction, aberrations, and imperfections in the optical system of the telescope. It provides essential information about the resolution, and helps to determine how closely two adjacent objects can be resolved in the observed image. The PSF is calculated at the beginning of the observation night.

Each observation site will host a diversity of telescopes that will simultaneously trigger on a signal of interest. Thus, the multiplicity of telescopes and acquisition cameras require to calibrate the inputs in order to make possible a data comparison. Procedures have to result in outputs that are as standardized as possible between cameras and hardware systems. Typically, the signal must be converted from ADC (Analog-to-Digital Converter) to the number of photo-electrons detected by the photomultiplier tubes. The calibration step consists in converting the electrical signal from the photo-multipliers into a common metric across all telescopes which is defined as the number of detected photons.

Image integration

Classically, the direct utilization of waveforms can be challenging because the Cherenkov light emitted from the incident EAS creates a sequence of footprints that can be highly dissimulated within the background noise when the energies are rather small.

Waveforms are integrated to produce an amplitude map that corresponds to the amount of charges that has been accumulated from the event in each corresponding pixel. Although this non-bijective procedure eventually leads to information loss, it allows reducing the level of noise. In fact, the integration procedure improves the signal-to-noise ratio (SNR) by averaging out random noise. Over time, noise tends to cancel itself out, while the underlying signal accumulates. This leads to an improvement in SNR. Furthermore, it simplifies the data and the complexity of feature extraction algorithms.

Individually applied to each pixel, the local peak integrator algorithm is a straightforward technique that aims to find the local maximum signal and integrate it within a window of 12ns (that is to say 12 images) around the peak. The response time of the PMs must be taken into account, thus the window is not centered on the maximum value but rather

shifted by 5ns. The major drawback of this integration technique is that high NSB can produce fluctuations that could be picked by the sliding window pulse integrator instead of the actual Cherenkov pulse and act erratically.

Using the full waveforms is advantageous because it preserves more information from the signal. While traditional machine learning techniques may struggle to process this level of complexity, deep learning has emerged as a promising tool capable of extracting valuable features from the full data. With its ability to automatically identify patterns in raw waveforms, it offers the potential to significantly enhance the accuracy and effectiveness of analysis in the context of IACT.

Image cleaning or rejection

In addition to the integration procedure, images must be filtered to further reduce the noise level such that quantities of interest (e.g. moments in the case of Hillas parameters) can be computed. Therefore, from the resulting charge maps and temporal information is applied a cleaning or rejection procedure to filter unexploitable images. In more detail, cleaning determines if a pixel is retained, whereas leakage and parameter cuts determine if an image is removed from further processing. Different algorithms exist:

- The *Cleaning* algorithm removes noise of images and keeps the pixels containing the shower signal. The most common procedure is the *Tail Cut cleaning*, which leverages a two-threshold strategy. The picture threshold defines the minimum intensity of the retained pixels, while the boundary threshold, defines the minimum intensity to retain pixels neighboring the picture pixels [Les+01a]. Both thresholds are defined by an expert, but the Crab Nebula analysis presented in [Abe+23] contains respectively the values 8 and 4.

- The *Intensity cuts* algorithm discards images where the sum of the intensity is lower that a defined threshold.

- The *Leakage cuts* algorithm aims to discard truncated showers and filters images based on leakage threshold. The leakage corresponds to the fraction of the ellipsoid located at the border of the image.

Although this cleaning or rejection step is necessary to ensure high-quality reconstruction of traditional algorithms, deep learning techniques offer the possibility of bypassing this process while retaining more information from the data. By allowing the model to automatically handle noise and extract relevant features, it can enhance the overall performance of the reconstruction.

The standard reconstruction for the LST-1 observations: the Hillas+RF methodology

The full-event reconstruction of the incident particle aims at retrieving the main physical features that are its energy, direction, and type. These estimated characteristics form the basis of the analysis, from which are derived more sophisticated figures of merit (see Section 4.3 of Chapter 4 and Section 5.3 of Chapter 5).

In many image analysis techniques, feature extraction is a fundamental step where complex data is simplified into a set of meaningful, usually hand-designed, parameters that can be further processed by machine learning algorithms. In the particular case of IACT, when a gamma-ray enters the atmosphere, it produces a particle shower that emits Cherenkov light. This light is captured by the telescope optical system as an image, which typically appears as an elongated ellipse. Based on this morphological assumption, in the current CTAO workflow, the feature extraction is performed with the Hillas algorithm, which analyzes the camera image by calculating its geometric properties [Hil85]. These parameters are referred to as Hillas parameters.

Typically, multivariate analysis techniques must be applied, such as classical machine learning algorithms or simple lookup tables, to the extracted features in order to reconstruct the incident particle properties. The strong interdependence between each parameter makes the IACT data analysis multi-task by nature. Classically, the Hillas parameters are fed to Random Forests (RF) [al08], and the combination of both techniques is referred to as *Hillas+RF*. In the case of the Crab Nebula analysis presented in [Abe+23], the quantities that are used as inputs of the RF are computed on the cleaned set of pixels and are defined as the following:

- The base-10 logarithm of the total intensity, calculated as the sum of the pixel charges.
- The coordinates of the centroid in the camera's reference frame.
- The dimensions of the ellipse representing the image signal, determined by the second-order moments of the charge distribution along the minor (width) and major (length) axes. The ratio of the signal's width to its length is also computed.
- The skewness and kurtosis, which are respectively measures of the asymmetry and flatness of the charge distribution.
- The gradient of signal arrival times across the image, calculated along its major axis.
- The proportion of the image's total charge recorded by the pixels located at the edge of the camera or by their immediate neighbors.
- The azimuth and altitude, representing the telescope's pointing direction.
- The distance between the centroid and the point along the image's major axis that is closest to the true source position. Furthermore, the side of the centroid where the true source lies on along the major axis is also considered, as well as the distance between the image centroid and the true source position in the camera's frame.
- The absolute difference between the calculated distance to the source and the reconstructed distance.

Depending on which quantity to regress or classify, a subset of these parameters are used as inputs of the RFs, and a detailed description can be found in [Abe+23].

Gamma / hadron separation

Because the flux of particles is supplied mainly by protons, the detection of gamma-ray sources highly relies on the classification ability of the considered reconstruction algorithm. In fact, although gamma-ray-induced showers result from pure electromagnetic processes, cosmic-ray-induced showers also incorporate hadronic components. This contrast is reflected in the morphology of the shower image captured by the IACT camera and can serve as a means to categorize events detected by IACT based on the primary particle type.

Direction reconstruction

In parallel, the uncharged nature of gamma-ray is a convenient asset to detect their source. Because they are not deviated from their trajectory by magnetic fields, finding their direction directly allows to find the coordinates of the objects that emitted them.

The determination of the direction reconstruction with a single telescope is challenging. It is currently achieved using the monoscopic displacement methods described in [Les+01b]. In fact, the Cherenkov light produced by the shower forms an image that looks like an elongated ellipse. The long axis of this ellipse points along the path of the particle shower, and the location of the gamma-ray source lies somewhere along this axis, near the tip of the light distribution, which represents the starting point of the shower. The direction of the source can also be identified based on the asymmetry of the image, as it tends to be lopsided towards the source.

In the recent Crab analysis paper of the CTAO [Abe+23], two RF models are trained using simulated gamma-ray data to predict the distance between the center of the image (the centroid) and the point along the shower's main axis that is closest to the true gamma-ray direction, and the side of the image's major axis the true direction of the gamma-ray lies relative to the centroid. When combined, these RF models help pinpoint the gamma-ray's original direction.

Energy reconstruction

Regarding the energy component, the initial energy of the particle determines the amount of interactions it undergoes within the atmosphere, which in turn shapes the development of the atmospheric cascade that emits the Cherenkov light. Therefore, the intensity and morphology of the structured and integrated images carry relevant information about the energy of the particle that must be estimated.

For instance, a 1 TeV particle arriving directly along the telescope's axis will produce a certain brightness, but a particle of the same energy impacting further from the telescope will appear dimmer due to geometric factors. Hence, to accurately estimate the energy, one needs to consider both the image intensity and the impact position of the shower. The impact position can be inferred by extracting specific image features such as width, length, and the centroid of the Cherenkov light distribution.

The need for more sophisticated methods

Previous studies have highlighted the intrinsic limitations that the standard analysis is suffering from, especially at lower energy level [Vui+21]. A more detailed description is given in the Section 2.5 of the next chapter. This suggests that more advanced techniques must be explored in order to break free from the constraints imposed the morphological assumption, the Hillas parameters and the cleaning procedure. Deep learning appears as a natural extension of the standard analysis. In fact, the γ-PhysNet neural network has been introduced in [Jac20] as a deep learning alternative to Hillas+RF and has shown an increase in performance by leveraging the power of a more complete information from integrated images.

1.4 Conclusion

Problematic

The story of astronomy is a fascinating and ongoing tale of humanity's conceptual exploration of the Universe and our place within it. Since ancient times, humans have been captivated by the stars, planets, comets, singularities within the observations, and have sought to understand their movements and origins. Thus, over the last centuries, the oldest science has evolved from a combination of observations and superstitions to a highly sophisticated and technologically advanced discipline. Nowadays, through the eyes of cutting-edge telescopes and satellites, most of the mysteries surrounding the Universe have been unveiled and allowed astrophysicists to test the current hypotheses, question established knowledge, and discovering new, unexpected phenomena.

The CTAO, once completed and operational, will be one chess piece of this experimental science. Numerous challenges will be faced, among which we can define managing the tremendous amount of data generated by the array of telescopes, conducting real-time analyses, employing stereoscopic techniques, and developing robust and interpretable results. These challenges are not unique to CTAO but are shared by other particle physics projects such as neutrino detection, as evidenced by the IceCube project's advancements [18].

This thesis aims to enhance the image analysis methodologies used in CTAO through the application of deep learning models. Building upon the pioneering work of [Jac20], which developed the first full-event deep learning reconstruction network, we address some of the limitations identified in their research. These limitations and the strategies to overcome them will be discussed in detail in the subsequent chapters. By leveraging advanced deep learning techniques, this work aspires to contribute significantly to the field, improving both the accuracy and the efficiency of data analysis in high-energy astrophysics.

Contributions

In order to tackle the issues encountered by CTAO, we propose in this thesis the following contributions:

- We introduce deep unsupervised domain adaptation in the context of IACT image analysis to account for the inherent discrepancies between the training simulations and the real telescope acquisitions, for the detection of known gamma-ray sources. We have selected, implemented and validated three relevant approaches. We tackle the

shift in particle ratio between both distributions with a reformulated training strategy and is referred to conditional domain adaptation.

- We introduce and evaluate information fusion to tackle the NSB affecting night observations. Relying on Conditional Batch Normalization [Vri+17], this novel approach allows to integrate prior information within the model to make it more robust. It will be put in perspective with domain adaptation.

- As a first step to a global neural network for the all-sky analysis, we implement high capacity vision Transformers. Based on the idea that pre-training allows to improve generalization and performance with fine tuning, we propose two models, γ-PhysNet-Prime and γ-PhysNet-Megatron.

Thesis outline

This thesis has been structured hierarchically and is adopting a coarse to fine point of view. The foundation is set for understanding the most recent methodologies applied throughout the research, before integrating the selected advanced deep learning techniques with simulated and practical applications, aiming to push the boundaries of current methodologies and improve the accuracy and efficiency of IACT image analysis.

Chapter 2 introduces the fascinating story of deep learning along with modern techniques, providing a comprehensive review of state-of-the-art approaches in model architecture, domain adaptation, information fusion multi-task balancing, and the application of machine learning to IACT image analysis. We give a detailed description of the limitations that encounter current machine and deep learning methods and set the mathematical framework to help understand and implement the improved techniques to be applied.

Chapter 3 presents the core contributions of this thesis that will be validated in the subsequent chapters. We introduce a set of complementary strategies that aims to increase model generalization behaviours on new, real telescope acquisitions. Firstly, we introduce CBN modules as a conditioning strategy to modulate the model with respect to known relevant physical parameters. This approach allows the model to incorporate multiple sources of information, improving its ability to generalize. Secondly, unsupervised domain adaptation is applied within a multitask framework. This methodology enhances the model's adaptability to unknown variables that contribute to data domain shifts. In a realistic scenario, we investigate domain adaptation in the context of strong class imbalance, an area that has not been extensively studied to the best of our knowledge. Finally, as a more general approach, we introduce the Transformer architecture. This high-capacity model represents a preliminary step towards developing a comprehensive framework, showcasing its potential to advance the field further.

Then, Chapter 4 marks the first application of our contributions. It explores the application of unsupervised domain adaptation, input conditioning and Transformers within the controlled environment defined by LST simulations. This chapter demonstrates how these techniques can enhance the accuracy and robustness of the models when applied to simulated data. To this end, we evaluate the impact of the NSB and label shift in preparation to the application to real data. We investigate multi-task and domain adaptation is an extensive study which allows us to select the most promising methods for the application to the real case scenario in the next chapter.

Furthermore, Chapter 5 focuses on the practical application of the selected, implemented and validated techniques to the Crab Nebula. It evaluates the performance of the new models against standard analysis methods and the γ-PhysNet neural network. This comparative analysis highlights the improvements and potential benefits of the proposed approaches in real-world scenarios.

Finally, Chapter 6.1 offers a detailed summary of the impact of our contributions in both scenarios and paves the way to new short- and long-term perspectives.

Appendices are available in 8 to dive deeper into the details of our analysis.

Papers and software contributions

During this thesis, the following papers have been published, submitted or are in preparation. In this specific context of the CTAO project, each publication is subject beforehand to a pear review and validation process within the international collaboration before its proposal to a conference of journal. Our contributions are the following:

- A proof-of-concept paper has been published at the CBMI 2023 international in [Del+23] (Michaël Dell'aiera, Mikaël Jacquemont, Thomas Vuillaume and Alexandre Benoit, *Deep unsupervised domain adaptation applied to the Cherenkov Telescope Array Large-Sized Telescope* in 20th International Conference on Content-based Multimedia Indexing 2023). It focuses on the application on diverse unsupervised domain adaptation methods in the case of simulations.

- A second proceeding has been accepted at the ADASS 2023 international in [Del+24] (Michael Dellaiera, Cyann Plard, Thomas Vuillaume, Alexandre Benoit and Sami Caroff, *Deep Learning and IACT: Bridging the gap between Monte-Carlo simulations and LST-1 data using domain adaptation* in Astronomical Data Analysis Software & Systems 2023). It applies the previously discussed methods to real data, comparing them with standard analytical approaches.

- A third paper has been submitted to the CVIU journal. It presents an extensive study on digit and LST simulated datasets to identify the optimal combination of multi-task balancing and unsupervised domain adaptation techniques. In addition, it introduces the conditional versions of our models. This paper integrates the contributions of Chapters 3 and 4.

- A fourth paper is in preparation, providing an in-depth analysis of the real dataset of the Crab Nebula. Furthermore, it introduces our CBN-based approach. As a result, it relies on Chapters 3 and 5.

- A fifth paper is in preparation, exploring the application of the Transformer model to both simulated and real data.

Along with the continuous improvement of the GammaLearn software[8], the analysis efforts have also resulted in the development of the GammaScan software[9] in cooperation with Cyann Plard.

[8] https://gitlab.in2p3.fr/gammalearn/gammalearn
[9] https://gitlab.in2p3.fr/gammalearn/gammana/crabdann

1.5 GammaLearn: An Open Science approach

The GammaLearn[10] project [Jac20] was born in 2017 from a collaboration between the Laboratoire d'Annecy de Physique des Particules[11] (LAPP), the Laboratoire d'Informatique, Systèmes, Traitement de l'Information et de la Connaissance (LISTIC)[12] and Orobix[13] and aims at fostering innovating methods in artificial intelligence for the Cherenkov Telescope Array Observatory image processing. On a larger scale, this project has been funded by the European Union commission through multiple EOSC programs (ASTERICS, ESCAPE[14], EOSC Future[15]. It aspires to share and develop solutions to commonly faced issues for both astronomical and particle physics experiments such as high data volume transfer, storage, political and organizational decisions or calculus. In this context, GammaLearn is part of the EOSC Future Work Package 6 (WP6) and aims to deliver novel analysis pipelines for the CTA data. Although GammaLearn is an open-source software, it is important to notice that the data used throughout this report is acquired by the first Large-Sized Telescope (LST-1) and it cannot be publicly communicated. Nevertheless, this work is the continuity of the long journey of the gamma-ray astronomy that started a century ago.

[10] https://gitlab.in2p3.fr/gammalearn/gammalearn
[11] https://lapp.in2p3.fr/?lang=en
[12] https://www.univ-smb.fr/listic/en/
[13] https://orobix.com/en/
[14] https://projectescape.eu/
[15] https://eoscfuture.eu/

2 Application of Deep Learning to IACT

The previous chapter introduced gamma-ray astronomy and outlined the physics questions at stake of the CTAO. It discussed how IACT telescopes collect a prodigious number of images through the particle shower mechanism that must be analysed. In this chapter, we first present the historical context of deep learning and then explore state-of-the-art neural network architectures. We summarize the application of machine and deep learning methods for reconstructing incident particle parameters and highlight their limitations. To address these challenges, we delve into the concepts of input conditioning and domain adaptation. Given the multi-task nature of the problem, we also describe the methodology for multi-objective balancing. Finally, we justify our selection of the most promising techniques for our astronomical applications.

2.1 Introduction

Humans exhibit an innate propensity to categorize everything. This inclination can manifest in daily activities, for example, sorting emails into distinct folders before starting work in the morning, or in the organization of bookstores by ordering the authors alphabetically on the shelves. It also manifests in scientific projects, resulting in remarkable works, such as the meticulous classification and examination of living species and fossils engaged by Charles Darwin in his pioneered seminal work, "On the Origin of Species", published in 1859. Through this conceptual experience, commonly referred to as categorization, ideas and objects are organized, allowing us to efficiently store information, facilitate retrieval, and infer characteristics. This ability is considered to be a fundamental cognitive principle. Thus, it becomes natural for our species to also classify the concept of science over time. Science can be described as a set of guiding principles and methodologies that are identified as paradigms [Hey+09]. Illustrated on Figure 2.1, they shed light on our dynamic and conflicted relationship with this discipline, shaping the way scientific inquiry is conducted and knowledge is acquired, providing a structured framework for researchers to approach questions, formulate hypotheses, and interpret findings.

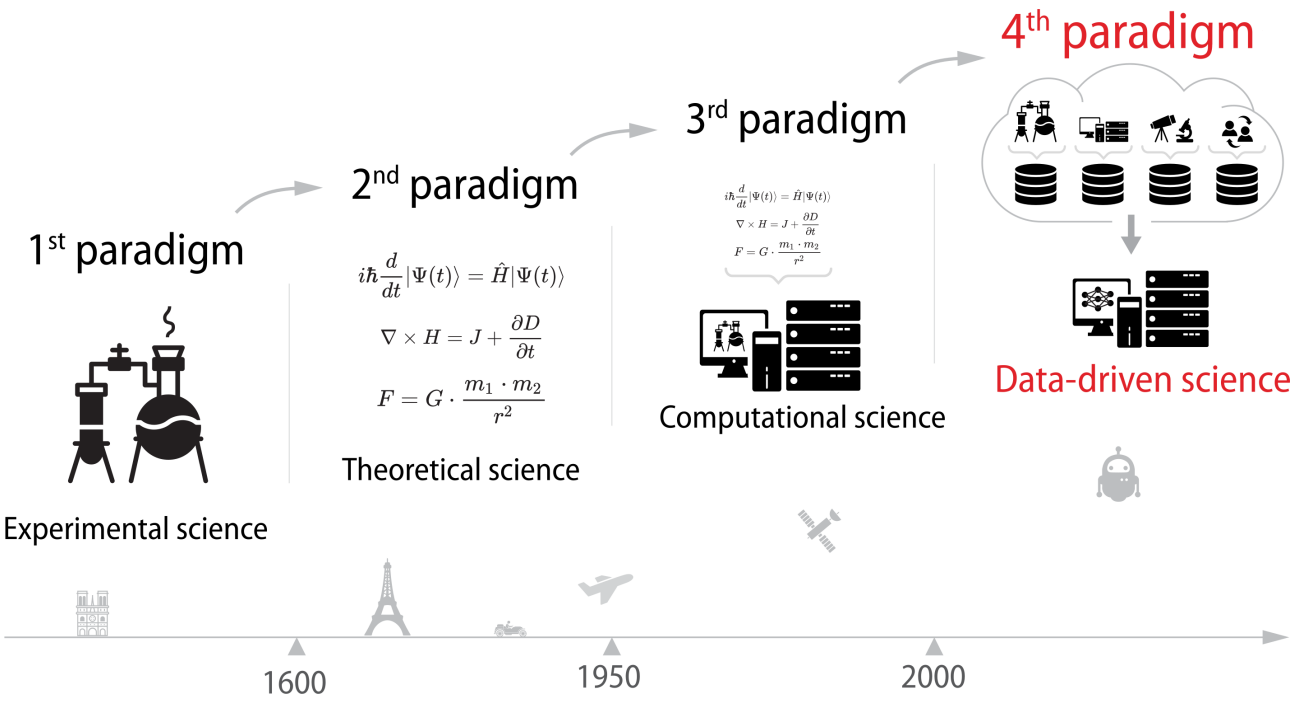


Figure 2.1: The four paradigms of science. Source: Fidle[1].

The first paradigm. Eons ago, when the first humans started to inhabit the Earth, the understanding of the world was limited to direct sensory experiences. Consequently, for thousands of years, ere the beginning of language and the invention of writing, and persisting until the 17th century, science relied mostly on empiricism, that is to say derived from observable phenomena. The application of mathematics to describe the world was still at its early stages. Yet, the first paradigm formed the bedrock of the scientific method, where results are deduced from observation and experimentation, defining rudimentary scientific

[1] https://fidle.cnrs.fr/

understanding. However, the lack of a proper structuring of thoughts, the inability to infer predictions and probably the dominant cultural dogma hindered the development of science.

The second paradigm. The 17^{th} marks a pivot towards a more theoretical vision of the world, placing a strong emphasis on reasoning and deductive logic. Initiated by Isaac Newton, his revolutionary laws of motion and universal gravitation exemplified the power of mathematics as a language to express the underlying principles governing the physical universe, and unravelling the mysteries of nature. The deliberate efforts perpetuated by his contemporary scientists finally demonstrated the profound interconnection between mathematics and the sciences.

The third paradigm. Despite the development of the main physics theories (electromagnetism, relativity, quantum mechanics) in the last centuries, some challenges remained frustratingly beyond reach due to their inherent complexity. Notable examples include the Navier-Stokes equations, the N-body problem, and others. While some insights and approximations could be obtained from simplified models, the true essence of these problems remained elusive. Fortunately, the rise of computers, and thus computational power, allowed to lift the veil on these once-inaccessible mysteries, and the computation of simulations made possible scientific advancements. This is by nature a popular paradigm among physics problems. As science tells us about the phenomena through a theoretical description, simulations are easily available.

The forth paradigm. The fourth paradigm, initially proposed by Jim Gray in [Hey+09], is a fundamental shift in approach that emerged from a very recent discipline called deep learning. In previous methodologies, hypotheses were considered a priori. To illustrate this statement, consider the conventional analysis approach in the context of CTAO. Traditionally, the measured signal of interest is assumed to have an ellipsoidal shape, a hypothesis introduced based on human intuition, from which a feature-extracting algorithm is implemented. In contrast, deep learning takes a divergent path by working directly with raw data, devoid of prior assumptions. In this paradigm, hypotheses are derived a posteriori, meaning they are formulated after an analysis of the data.

2.2 The long journey of artificial intelligence

Data science is a recent discipline that aims to extract meaningful information from the data. The nature of the data and the questions they address can vary widely. Although it includes many techniques, artificial intelligence (AI) stands out as one of its most prominent components. Artificial intelligence refers to the broad concept of creating machines that can perform tasks that typically require human intelligence, such as reasoning, problem-solving, decision-making, creativity or innovation [GBC16]. This can be achieved through various techniques, including rule-based systems, symbolic logic, expert systems, machine learning and more recently deep learning. Fundamentally, it is a mathematical and computer science discipline, which aims at solving a great variety of different and complex tasks, generally related to computer vision, natural language processing or audio recognition. Recently, deep

learning has been successfully applied in many diverse signal and image processing cases and multimedia image analysis led to foster and extend this technique to other domains such as astronomy. Although it has become an unavoidable topic of research, its journey started almost one century ago, and has been marked by periods of disinterest and phases of regain from the scientific community [CCM18].

Cybernetics

Scientists from diverse backgrounds (mathematicians, logicians, engineers, physiologists, anthropologists, psychologists, and more), and among the brightest of the time, participated in a series of conferences between 1942 and 1953. Known as the Macy conferences, their goal was to provide a unified vision for the emerging fields of automation, electronics, and information theory. Aiming to present a "complete theory of control and communication, both in animals and machines", the research was driven by the idea of building a general science of the functioning of the mind. Although the outlines of this research endeavour might appear undefined, Cybernetics revolves around the key concept of feedback. Feedback refers to the process of incorporating the observed outputs, usually in the form of a measurement error, as inputs to decide for further action, creating a circular causal relationship. Deriving from the Greek, "the art of steering" a boat, it conveys the idea that to reach its destination, sailors must correct the perturbations, such as wind, or waves from the ocean, to retrieve the right path. Steering and getting to the goal is a concept shared by all intelligent systems. As an example, humans self correct their body temperature, their balance while walking in the street, their conversation so that words and ideas are comprehensible. The physical, biological, and social worlds become a link of the chain. Popularized in 1948 by the mathematician Norbert Wiener in his pioneering book "Cybernetics, or Control and Communication in the Animal and the Machine", as a summary of the main principles emerged from the Macy conferences, Cybernetics is conceptualized as an autocorrective machine capable of modifying its behaviour and goals based on probabilistic operators, adapting from its errors rather than relying on internal predetermined actions. From these concepts emerged two competing schools of thoughts for artificial intelligence. On the one hand, connectivism refers to the massive parallel calculus of elementary functions, distributed within the network, with the meaningful outcomes due to emerging effect induce by those elementary operations. On the other hand, symbolism aims to calculate symbols that have a material reality and a semantic value of representation, and to implement high-level rules within the machine to break the link with the outer world and open an autonomous reasoning world within the calculator.

The saga of AI

Figure 2.2: The Dartmouth workshop (1956). From left to right: Selfridge, Rochester, Solomonoff, Minsky, More, McCarthy, Shannon[2].

Although the concept of AI finds its roots in cybernetics, its first chapter started when McCulloch and Pitts introduced in 1943 a pioneering work that became the inspirational foundation for future research, unveiling the first model of the artificial neuron [MP43]. In 1956 took place the historic Dartmouth Conference [McC+55], orchestrated by visionaries like Minsky and McCarthy, often hailed as the birth of this new computer science domain, with the term "artificial intelligence" itself coined by the main actors of the workshop. However, nowadays, this neologism sparks controversy on two fronts: the elusive definition of intelligence and the presumption that, once defined, it could be replicated artificially. This terminology, yet currently the most adopted, was created to oppose symbolist AI to connectivist AI and its adaptive adjustment of the inputs and outputs.

Vol. VI, No. 2, Summer 1958

research trends

CORNELL AERONAUTICAL LABORATORY, INC., BUFFALO 21, NEW YORK

The Design of an

Intelligent AUTOMATON

by FRANK ROSENBLATT

Introducing the perceptron — A machine which senses, recognizes, remembers, and responds like the human mind.

STORIES about the creation of machines having human qualities have long been a fascinating province in the realm of science fiction. Yet we are now about to witness the birth of such a machine — a machine capable of perceiving, recognizing, and identifying its surroundings without any human training or control.

Development of that machine has stemmed from a search for an understanding of the physical mechanisms which underlie human experience and intelligence. The question of the nature of these processes is at least as ancient as any other question in western science and philosophy, and, indeed, ranks as one of the greatest scientific challenges of our time.

Our understanding of this problem has gone perhaps as far as had the development of physics before Newton. We have some excellent descriptions of the phenomena to be explained, a number of interesting hypotheses, and a little detailed knowledge about events in the nervous system. But we lack agreement on any integrated set of principles by which the functioning of the nervous system can be understood.

We believe now that this ancient problem is about to yield to our theoretical investigation for three reasons:

First, in recent years our knowledge of the functioning of individual cells in the central nervous system has vastly increased.

Second, large numbers of engineers and mathematicians are, for the first time, undertaking serious study of the mathematical basis for thinking, perception, and the handling of information by the central nervous system, thus providing the hope that these problems may be within our intellectual grasp.

Third, recent developments in probability theory and in the mathematics of random processes provide new tools for the study of events in the nervous system, where only the gross statistical organization is known and the precise cell-by-cell "wiring diagram" may never be obtained.

Receives Navy Support

In July, 1957, Project PARA (Perceiving and Recognizing Automaton), an internal research program which had been in progress for over a year at Cornell Aeronautical Laboratory, received the support of the Office of Naval Research. The program had been concerned primarily with the application of probability theory to

Figure 2.3: Rosenblatt's article first page[3]

In 1957, inspired by the artificial neuron of McCulloch and Pitts, but with the addition of a learning mechanism, emerged from the mind of Rosenblatt the Perceptron, acclaimed as "the first machine capable of having an original idea" [Ros58]. Embodying the first connectivist machine, it is conceived for image recognition, and received funding from the American Office of Naval Research (ONR). Yet, overly ambitious expectations and unattainable goals led to a decline in investments and research, and the 1970s cast a shadow on AI, labeled as its first winter. They arise in an era affected by a strong will of programming a "conscious brain", logical, characterized by explicit reasoning of underlying systems, modelling the human expert's approach. During the two decades of expert systems domination, some fundamental advancements have been made on the other side.

[2] https://aiws.net/the-history-of-ai/this-week-in-the-history-of-ai-at-aiws-net-the-dartmouth-conference-ended-on-august-17th-1956/

[3] https://news.cornell.edu/stories/2019/09/professors-perceptron-paved-way-ai-60-years-too-soon

A pivotal moment arrived in 1986 when Rumelhart, Hinton, and Williams presented the backpropagation algorithm, unlocking the potential for deeper neural networks [RHW86]. Renew interest for the "unconscious brain", revived interest for connectivist AI and yield in an unstoppable growing momentum. The landscape shifted again in 2012 during the annual and well-known ImageNet competition - a database of 1 million images and 1000 object classes created by researchers at Princeton -, heralding a new era in computer vision, with the reborn of a technology that has been marginalized for a long time. At that time, the prevalent algorithms were Kernel methods (Kernel PCA, Kernel SVM, etc.), leveraging manual feature extraction, and yielding honourable results. However, the ImageNet problem remained highly challenging, until a Canadian team led by Geoffrey Hinton, Alex Krizhevsky and Ilya Sutsever used a neural network architecture and over-dominated the battle, with a clear improvement in classification accuracy compared to the concurrence [KSH12]. Nowadays, AI imposed everywhere, and as the story continues, its trajectory unfolds with each innovation. From the single artificial neuron humble beginnings to billions of parameters networks, we assist to a relentless pursuit of AI and its transformative potential.

The data, computing power and high capacity trilogy

Among the factors reponsible for to the first winter of AI, the limitations in computational power, the lack of annotated data, and the modest size of the models played a crucial role [Sun+17]. Led by the development of video games, graphics processing units (GPUs) computation capacities rapidly increased over the year, to a factor of $\times 10^6$ flops over the last 25 years [Sev+22]. The potential of GPUs for massive parallelization has been actively explored starting from 2009 [RMN09], and has opened up new possibilities with the current trends in large dataset sizes and model capacities. The second factor that facilitated the growth of deep learning is the accessibility of data, especially with the deployment of the Internet. The evolution of the size of training datasets highlights a deluge of data with a raise of $\times 10^4$ size in 25 years [VH22]. In fact, performance scales with dataset size [Sun+17]. As the last part of our trilogy, model capacity, that is to say the number of parameters of the models, has increased by a factor of $\times 10^6$ in the last 25 years [Sev+22]. This confluence of enhanced computational power and the exponential growth of accessible data laid the foundation for the resurgence and flourishing of AI.

2.3 Introduction to deep learning

Throughout the exploration of the history of artificial intelligence, it appears that the primary objective of artificial intelligence is the automation of repetitive tasks traditionally attributed to humans. Drawing inspiration from nature, mathematicians and computer scientists devised tools that aimed to mimic biological processes. In the same way that humans are learning, AI algorithms must also do. Before the Internet, data availability was limited, and datasets were rather small. In this context, it became necessary for data scientists to engineer handmade general features, introducing some biases inherent in our a priori knowledge. This area evolved into what is nowadays referred to as machine learning, and usually involves the tuning of a few parameters. However, the advent of abundant data and increased computational power has revolutionized this landscape of data science and eliminated the ultimate need of the manual feature engineering procedure. Instead, raw data is

directly fed to the algorithms, allowing them to autonomously construct their own representations. This new modus operandi is known as deep learning, and operate on large-scale, complex datasets, automatically extracting intricate features. However, deep learning introduces challenges, with the management of a many parameters, requiring a careful tuning for optimal performance.

The key role of supervision

Machine learning and deep learning can both be classified in four distinct categories, depending on the goal and the availability of the data, as depicted in Figure 2.4. In the case of labelled data, it is known as supervised learning methodology. In this context, we consider either a regression problem, for example estimating the energy from the images resulting from telescope acquisitions, or a classification problem, such as distinguishing between different classes of object (a gamma / proton detector for example). In unsupervised leaning, we assume that there are no labels applied to the training data, thus the strategy is different. It can be subdivided into clustering (creation groups, like the k-means algorithm) and reducing the dimensionality of the data (like the Principal Component Analysis). Reinforcement learning is another type of machine learning that consists in teaching an agent a policy, or strategy, to make decisions by interacting with an environment. It has been successfully applied to various domains such as robotics, game playing, and autonomous systems. Finally, the last category is transfer learning. The contextual framework of this research is the application of supervised learning in the context of IACT.

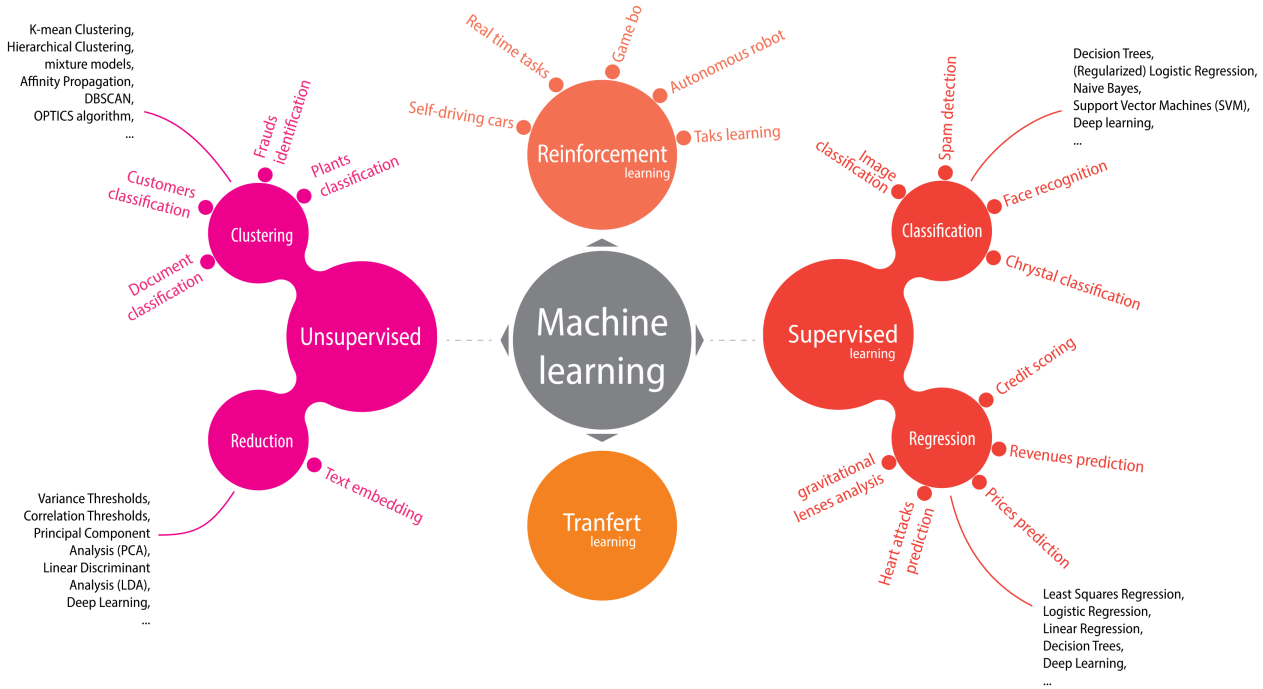


Figure 2.4: Machine learning and deep learning can be subdivided in four categories: Supervised, unsupervised, reinforcement and transfer learning. Image credit: Fidle[4].

[4] https://fidle.cnrs.fr/

The process of learning

Supervised learning can be interpreted as an interpolation problem in a high-dimensional space [Mal16]. It consists in approximating a parametric function f^θ from labelled training examples. This is a challenging optimization problem because of the plethora of local minima that one must navigate to find the best solution or at least approach it. This process is carried out by optimizing a specific metric known as the objective function. From a probabilistic perspective, the estimation of the optimal parameters can be achieved from Maximum Likelihood Estimation (MLE) or Maximum A Posteriori (MAP). In the case of deep learning models, the number of parameters can reach millions, or billions of parameters. Considering the posterior distribution of the parameter θ given the observed data X, the MAP estimate $\hat{\theta}_{MAP}$ and can be computed using the Bayes theorem.

$$\begin{aligned}\hat{\theta}_{MAP} &= \arg\max_\theta p(\theta|X) \\ &= \arg\max_\theta \frac{p(X|\theta)p(\theta)}{p(X)} \\ &= \arg\max_\theta p(X|\theta)p(\theta) \\ &= \arg\max_\theta (\log p(X|\theta) + \log p(\theta))\end{aligned} \tag{2.1}$$

The prior $p(\theta)$ can act as a regularizer. In the case of a centred Gaussian prior with a small variance, the parameter estimates will shrink towards zero, preventing overfitting.

It is commonly assumed that the true label corresponds to the output of the neural network polluted by a Gaussian noise. This assumption is justified by invoking the Central Limit Theorem, which states that the sum of a large number of independent and identically distributed random variables will be approximately normally distributed, regardless of the original distribution. In the context of neural networks, the Gaussian noise offers a way to capture uncertainties in the predictions that may arise due to different factors, such as limited data, variations in input patterns, or inherent complexity in the learned relationships. Considering the scenario where we want to estimate a parametric function f^θ under a Gaussian distribution given N labelled observed data $(X, Y) = \{x_i, y_i\}_{i=1}^N$, the likelihood becomes the product of each individual likelihood.

$$p(Y|X,\theta) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - f^\theta(x_i))^2}{2\sigma^2}\right) \tag{2.2}$$

In the case of a classification problem, the likelihood is based on the probability of the K possible observed class labels given the predicted probabilities $f_k^\theta(x)$.

$$p(Y|X,\theta) = \prod_{i=1}^{N} \prod_{k=1}^{K} [f_k^\theta(x_i)]^{\mathbb{I}(y_i=k)} \tag{2.3}$$

where $\mathbb{I}(y_i = k)$ is an indicator function that is 1 if $y_i = k$ and 0 otherwise.

The optimization of the objective function is performed using the Gradient Descent (GD) algorithm. Mini-batch (or Stochastic) Gradient Descent (SGD) is a stochastic alternative to GD and allows computing an estimate of the gradients from subsets of the training dataset

[LeC+12]. Although these estimates may be noisy, they offer computational efficiency and can lead to better convergence. Conversely, batch learning tends to converge to the minimum of the basin in which the initialized weights are placed. Stochastic gradients allow jumping from a basin to another, with maybe a deeper minimum. The weight update rule in the optimization process involves the current state θ_t from which is subtracted the gradients $\nabla_\theta f(\theta_t)$ proportionally to the learning rate μ.

$$\theta_{t+1} = \theta_t - \mu \nabla_\theta f(\theta_t) \tag{2.4}$$

The variance of the fluctuation around the local minimum is proportional to the learning rate, thus it can be interesting to schedule the learning rate using a learning rate scheduler. A detailed overview of GD algorithms is given in [Rud17]. Traditionally, weights are optimized across the model using the chain rule [RHW86].

2.4 The architecture of neural networks

Designing the architecture of a neural network plays an important role as it is a vector of the prior information that we can include to our problem. The first step in understanding the composition of neural networks is to describe the elementary units that they are made of, namely the artificial neuron. Then, we will see how the assembling of many neurons forms layers, and the association of many layers generate a network.

The artificial neuron

Inspired by the functioning of the brain, McCulloch and Pitts derived a simple mathematical model that became the basis of modern research in the creation of the artificial neuron. Figure 2.5 is a simplified schema of the biological neuron, which consists in inputs (dendrite) that process the information and provide outputs (axon terminal) that are either in an *ON* mode (trigger mode), or in an *OFF* mode. The trigger mode is reached when the inputs pass a threshold value.

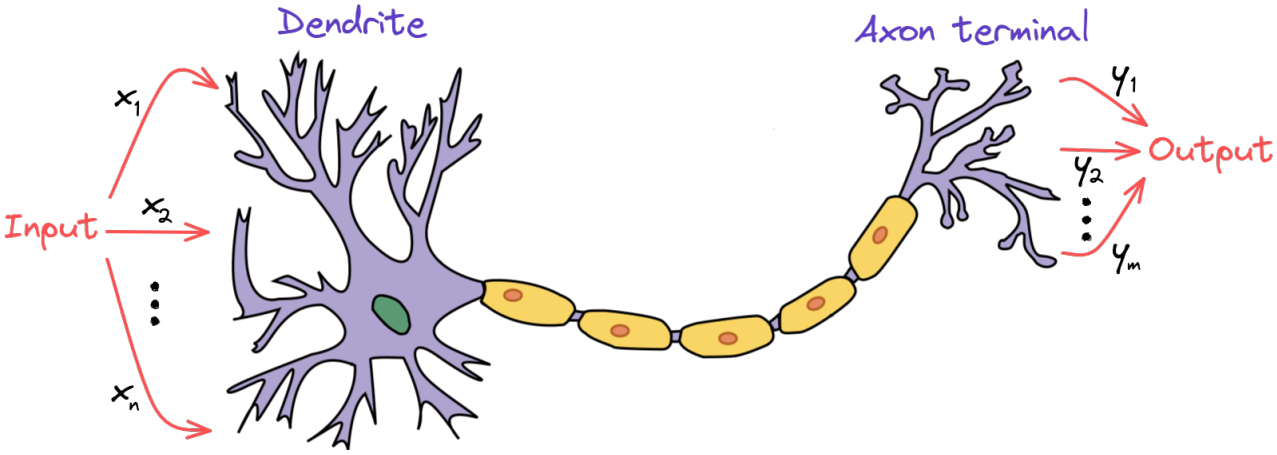


Figure 2.5: A simplified representation of a real neuron. Image credit: Wikipedia.

The artificial neuron, or node, depicted in Figure 2.6, is the fundamental parametric unit of any neural network. It performs a non-linear operation with a linear mapping and

a non-linear activation function σ. The linear mapping is simply a weighting sum over the neuron N entries x_i multiplied by the trainable weights w_i, and a trainable bias b is added. In biological terms, the bias is a measure of how easy it is to get the neuron to fire. A neuron produces one output, also named feature for hidden layers. The set of parameters $(w_i, b) \in \mathbb{R}^{N+1}$ provides the neuron with the ability to learn from the data. The output $y \in \mathbb{R}$ can be written as:

$$y = \sigma\left(\sum_{i=1}^{N} w_i x_i + b\right) \tag{2.5}$$

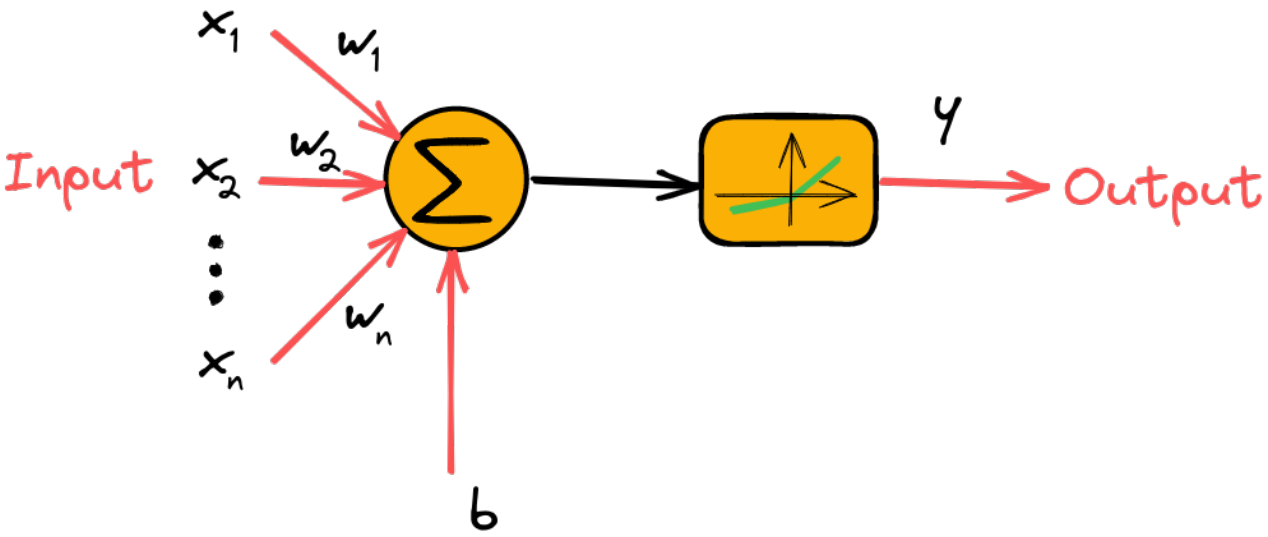


Figure 2.6: The artificial neuron.

Dense neural networks

Fully connected layers, also named dense layers, refers to the concatenation of neurons that are connected to all the neurons of the previous layer. The output of such a layer is called a feature map. Then, Dense Neural Networks (DNN) correspond to the concatenation of multiple dense layers. Also called Multi-Layer Perceptron (MLP), DNNs are universal function approximators [Cyb88]. Considering the layer l and the corresponding weights and biases $W^l \in \mathbb{R}^{N\times M}$ and $b^l \in \mathbb{R}^M$, the output denotes as $y^l \in \mathbb{R}^M$ becomes a vector computed from the layer's input $X \in \mathbb{R}^N$.

$$y^l = \sigma\left(W^l X + b^l\right) \tag{2.6}$$

Convolutional neural networks

DNNs have the particularity that it does not take into account the inherent structure of the input. However, the world is organized into a system of hierarchical structures for the reason that this kind of anatomy is closely related to the notion of stability [Sim91]. One may for example consider the following observations :

- Bodies are made of organs, that are made of cells, that are made of molecules that are made of atoms, that are made of quarks.
- Libraries are made of books, that are made of chapter, that are made of sentences, that are made of words, that are made of letters.
- Images are made of objects that are made of voxels, that are made of pixels.

The introduction of convolutional neural networks (CNN) [LeC89] made possible the exploitation of this intrinsic structures of objects in natural images by leveraging the idea that adjacent pixels are related to each other. CNNs follow a pyramidal decomposition (in reference to Gaussian, Laplacian, steerable or wavelet decomposition), allowing to greatly reduce the number of parameters compared to DNN.

Filter convolutions constitute the building blocks and rely on the convolution operation, which finds its roots in the signal and image processing science. They play a key role in designing filters to separate images into different frequency ranges. On the contrary to general filters or wavelets, which extract specific and hand-designed features from the images, such as directional edges, the features extracted by the convolutional layers are not necessarily human readable, but it is observed that the deeper the filter, the more abstract and specialized they become. The strength of CNNs are the notions of translation equivariance and locality, which can be achieved with translation-invariant filters, that is to say filters sharing the same weights across the image. The discrete 2-dimensional convolution operation involves an image or feature $I \in \mathbb{R}^{H \times W}$, a filter or kernel $K \in \mathbb{R}^{M \times N}$ and an output $(I * K)$.

$$\forall (i,j), (I * K)(i,j) = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} I(i+m, j+n) \cdot K(m,n) \tag{2.7}$$

One of the main tools classically incorporated within CNNs as part of the pyramidal process is pooling layers. They perform a downsampling operation on feature maps, reducing spatial dimensions while preserving key features. This drastically decreases computational complexity and helps prevent overfitting. The most common pooling method is max pooling, which selects the maximum value from a small neighbourhood of pixels.

$$\forall (i,j), P(i,j) = \max_{0 \leq m < M, 0 \leq n < N} I(i+m, j+n) \tag{2.8}$$

From the notion of locality introduced by filters emerges the concept of receptive field that refers to the region of the input image that influences the output of a neuron in a given layer. It essentially determines the amount of context or the area of the input image that the network considers when making decisions. Pooling and stride operations increase the field by reducing the size of the latent space.

Residual neural network

Early efforts in the development of deep learning models were often stymied at the training step by the gradient vanishing problem. As neural networks became deeper and more complex, the gradients required for training would diminish exponentially while propagated backward through the layers as a consequence of the chain rule. Causing the gradients to approach zero, parameters would remain unmodified, posing a severe obstacle to the training of deep networks and limiting the depth of architectures that could be practically employed.

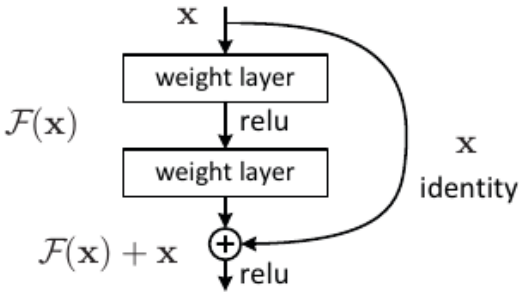


Figure 2.7: The ResNet block. Source [He+15a].

Historically, this bottleneck forced designers to confine neural network architectures to shallower designs, thereby constraining their potential to capture complex patterns and preserve their generalization capabilities. The gradient vanishing problem not only impeded the progress of neural networks but also spurred extensive research into potential solutions.

Recently, residual neural networks (ResNet) [He+15a] have revolutionized neural network design by introducing residual connections. Illustrated on Figure 2.7, the input x flows within two branches, respectively the CNN branch defined by $\mathcal{F}$ and the identity branch. The key difference in the calculation of the gradients is the identity matrix Id, which ensures that the gradient $\frac{\partial L}{\partial x}$ with respect to the input includes a direct path for the gradient flow, preventing it from vanishing as easily.

$$\frac{\partial L}{\partial x} = \frac{\partial L}{\partial y} \cdot \left(\frac{\partial \mathcal{F}(x)}{\partial x} + Id \right) \tag{2.9}$$

These bypass connections help maintain sufficient gradient values during propagation, thereby alleviating the gradient vanishing problem. This architectural innovation has been instrumental in achieving state-of-the-art results across a variety of tasks and domains. The ResNet block offers a versatile framework with numerous possible implementations [He+16].

The ascension of Transformers

While CNNs are a tried-and-tested class of neural network architecture that has been widely used for computer vision tasks, they are not well suited for processing sequential data such as text or speech. This is because CNNs are designed to operate on fixed-size inputs and are limited by the size of their receptive fields, which can make it difficult to capture long-range dependencies in the data. The number of operations that are necessary to relate two signals at different positions in the sequence increase with the distance between the positions, meaning that distant positions are more difficult to relate.

Introduced initially in the Natural Language Processing (NLP) paradigm, Transformers [Vas+17] address these limitations by using self-attention modules to model dependencies between different parts of a sequence, allowing the model to selectively focus on relevant

information and capture long-range dependencies in the input data. This makes them particularly effective for natural language processing tasks, such as machine translation and text generation, where the input sequences can be very long and the relationships between different parts of the sequence are important [Ope23]. In addition, Transformers are highly parallelizable, making them a faster and more scalable solution than Recurrent Neural Networks (RNNs), which were traditionally used for sequential data processing. This allows them to process large amounts of data quickly, making them well suited for use in real-time applications. However, this formulation relies on a large number of parameters, necessitating a large and diverse training dataset to learn meaningful features and relationship. As a result, the current mainstream approach consists in training the Transformers on a large text corpus and to fine-tune on smaller task-specific datasets.

Although Transformers have been originally designed to process time series, some transformations converting bidimensional signals into sequential data can be applied to images or higher spatial dimension inputs, and are referred to as Vision Transformers (ViT) [Dos+20]. This new architecture design offers efficiency in computation and scalability in the Computer Vision field, paving the way to the large models era ($> 10^{10}$ parameters [Sev+22]).

In the context of NLP, raw text data isn't inherently interpretable by models and requires conversion into a numerical format that models can process effectively. Moreover, natural language inputs vary in length, necessitating a method to break down text into sequential units. This allows networks to handle inputs of varying lengths by either padding shorter sequences or truncating longer ones to a uniform size. Tokens serve this purpose as fundamental units processed by the model. Tokenization refers to the process of converting raw data into a sequence of tokens. In the case of ViTs, they correspond to fixed-size image patches.

Inductive bias

Learning aims to draw a general rule for a whole population based on a limited subset. From the set of observations that is available, it can be induced a set of hypotheses that corresponds to all the possible parameters of the model. Inductive bias refers to the restriction to a specific group of parameters, corresponding to a restriction of the hypothesis space. Authors of [Mor+21] demonstrate that ViT models suffer from inductive biases more than their CNN counterparts. In fact, the latter are equipped with two fundamental concepts that are locality and translation equivariance. However, ViTs mostly rely on self-attention layers, which are global operators. Nevertheless, it has been demonstrated that training ViTs on large datasets trumps inductive biases [Dos+20].

ViTs have become the new state-of-the art approaches and supersede CNNs for image classification tasks, but face difficulties for object detection and semantic segmentation, that is to say tasks for which inductive biases such as translation equivariance are important. Hierarchical Transformers introduce CNN priors and have demonstrated that they can be used a backbone instead of CNNs as a generic vision feature extractor beyond image classification. Whereas previous ViTs produce single resolution feature maps, maintaining a quadratic complexity which makes it impossible for semantic segmentation and object detection, Shifted WINdows Transformers (SwinTransformers) [Liu+21b] use a multi-stage design, with each stage creating a different grid size. With the introduction of shifted win-

dows, they allow to limit the computation of self-attention to non-overlapping local windows while allowing cross-window connection. The computation complexity becomes linear with respect to the input size.

Attention: The Transformers mechanism

Before the advent of Transformers, models were limited in how they integrated image features based on their positions within images. These interactions were constrained by the receptive field [Luo+17], which determined the minimum depth of the model. Additionally, convolution operations inherently create a Gaussian field that reduces the importance of features located towards the periphery of the grid. Conversely, attention is the core of the Transformers architecture, allowing tokens to be weighted based on their importance, independently of their position in the input. Many types of attention have emerged since the beginning of Transformers:

- Self-Attention (SA) is the core mechanism in the original Transformer model introduced by [Vas+17] for NLP tasks. It allows each word to attend to every other word in the sentence, making it possible to capture long-range dependencies efficiently. In ViTs, self-attention is applied to image patches to estimate their relationships, enabling the model to understand the spatial and contextual information of the entire image [Dos+20]. Given an input sequence represented by the matrix $X \in \mathbb{R}^{N \times d}$, where N is the number of tokens and d is the dimensionality of each token's embedding, the input matrix X is first linearly projected to obtain the queries Q, keys K, and values V using the weight matrices $W_Q, W_K, W_V \in \mathbb{R}^{d \times d_k}$, with d_k the dimensionality of the queries and keys.

$$\text{Self-Attention}(X) = \text{Softmax}\left(\frac{XW_Q(XW_K)^T}{\sqrt{d_k}}\right)XW_V \tag{2.10}$$

  Such operation can also be applied to traditional CNN models. In that case, the input is defined as $X \in \mathbb{R}^{C,N}$, where C is the number of channels and N the number of pixels. Besides, a trainable scaling parameter γ is introduced and initialized to the value 0 in order to let the network learn local dependencies at the beginning of the training focusing on long-range ones.

$$X^{\text{SA}} = \gamma\ \text{Self-Attention}(X) + X \tag{2.11}$$

- Channel-wise Attention (CA) focuses on the interdependencies between the feature channels. It emphasizes the most important ones and suppresses less them if considered less relevant, allowing the model to enhance informative features [Hu+19]. Given an input feature map $X \in \mathbb{R}^{C \times H \times W}$, where H is the height, W is the width, and C is the number of channels, the first step consists in applying global average pooling to each channel to obtain a vector $\mathbf{z} \in \mathbb{R}^C$, where z_c is the pooled value for the c-th channel.

$$z_c = \frac{1}{H \times W}\sum_{i=1}^{H}\sum_{j=1}^{W} X_{i,j,c} \tag{2.12}$$

The pooled vector $\mathbf{z}$ then passes through two fully connected layers to learn the channel-wise attention weights. The first FC layer, defined by $W_1 \in \mathbb{R}^{\frac{C}{r} \times C}$, reduces the dimensionality, while the second FC layer, defined by $W_2 \in \mathbb{R}^{C \times \frac{C}{r}}$, restores it. r is a hyper-parameter corresponding to the reduction ratio, and σ refers to the sigmoid activation function. The attention weights $a \in \mathbb{R}^C$ are computed as:

$$a = \sigma(W_2 \cdot \text{ReLU}(W_1 \cdot z)) \tag{2.13}$$

The channels are finally rescaled to the original input feature map X by multiplying it with the attention weights:

$$X^{\text{CA}}_{i,j,c} = a_c \cdot X_{i,j,c} \tag{2.14}$$

- Spatial Attention (sA) allows the network to focus on the most relevant parts of the image by emphasizing important spatial regions. If $X \in \mathbb{R}^{C \times H \times W}$, where H is the height, W is the width, and C is the number of channels, the first step reduces the current feature maps to a mean image over the channel dimension.

$$y_{i,j} = \sum_{c=1}^{C} X_{i,j,c} \tag{2.15}$$

The resulting mono-channel feature goes through a series of convolutions and non-linearity, followed by a sigmoid function that rescales the attention map between 0 and 1.

$$S = \sigma(\text{Conv}_{1\times1}(\text{ReLU}(\text{Conv}_{1\times1}(y)))) \tag{2.16}$$

Finally, the input X is rescaled based on the computed spatial attention map.

$$X^{\text{sA}}_{i,j,c} = S_{i,j} \cdot X_{i,j,c} \tag{2.17}$$

- Dual Attention (DA) [Fu+19] is the integration of both spatial and channel attention within the same attention module. It corresponds to the Hadamard product between the output of the spatial attention X^{sA} and the output of the channel-wise attention X^{CA}

$$X^{\text{DA}} = X^{\text{sA}} \cdot X^{\text{CA}} \tag{2.18}$$

Multi-head attention

Instead of applying the attention mechanisms directly on the whole vector, it has been experimentally proven that decomposing the vector into sub-vectors is not only computationally more efficient but also yields better performances [Vas+17]. Each subpart is referred to as a *head*, consisting in the replication and the concatenation of the attention mechanism. Although it has been demonstrated that the use of multi-head attention improves the model performance, the interactions between the heads and their contribution to the network is

not fully understood. [CLJ20] proposes to promote collaboration between the heads instead of concatenation since they aim to solve the same tasks but were originally designed to accomplish their goal independently.

Pre-training

ViTs outperform CNN-based approaches when trained on a large dataset, but they don't compete favourably when trained from scratch on small datasets. However, it is possible to retrieve similar performances when adapting the training recipe [TCJ22].

- Stochastic depth [Hua+16] is a regularization procedure that aims to fight overfitting and helps learn robust features. During training, entire layers are randomly dropped with a certain probability, and are effectively skipped during forward and backward pass computations. Thus, the model is trained to use a reduced set of layers for that particular iteration. It is especially effective for deep neural networks, as it alleviates gradient vanishing and encourages learning redundant features.
- LayerScale aims to normalize the output of each Transformer layer and rescale it by a learned parameter, which is specific to that layer.
- Compared to the cross-entropy loss function, the use of the binary cross-entropy has shown a significant gain in performance for the largest models.
- 3-Augment is a data augmentation procedure consisting in three transformations. Firstly, grayscaling allows colour invariance and teach the network to focus on shapes. Secondly, solarization adds noise to the colour channels. Lastly, a Gaussian blur is applied to moderately degrade the high frequencies within the images.
- Simple Random Crop (SRC) and Random Resized Crop (RRC) are two other types of data augmentation that allow increasing model generalization. However, while SRC keeps the image aspect ratio, RRC may introduce data discrepancy.

Some of the proposed suggestions are not applicable to our IACT analysis. In the case of the CTAO images, pixel charges are mono-channel and the cropping procedure may not be adequate as it could crop a part of the signal and alter the shape of the signal, which is the main means of differentiation between gammas and protons. Furthermore, it is possible to have multiple islands within the same image, which could give more clues about the properties of the incident particle. Consequently, 3-Augment and SRC/RRC cannot be integrated in our approaches.

ConvNeXt: The return of the CNNs

The strength of hierarchical Transformers is the integration of a hybrid CNN-Transformer methodology through the sliding window strategy, which involves attention within local windows rather than global attention mechanisms. Swin Transformer [Liu+21b] represents a significant advancement in this domain, demonstrating for the first time that Transformers models can serve as a general vision backbone and achieve state-of-the-art performance across various computer vision tasks beyond image classification. This success underscores the key insight that the fundamental principles of convolution remain relevant and are still

highly valued. With the aim to re-explore the pre-ViT dominant approach, the ConvNeXt has been introduced in [Liu+22] as a family of pure CNN models that compete favourably with Transformers in terms of accuracy and scalability. The recipe to success holds in modernizing the primitive architectures towards the design of a state-of-the-art hierarchical ViTs, but without the addition of any attention modules.

2.5 Application to IACT

The conventional standard analysis for IACTs is based on classical CV algorithms designed to extract relevant features from the images, and the application of deep learning is still in its early stages. As part of the third paradigm of science, and depending on the selected hypothesis, the standard analysis can further be classified in two categories, moment-based and model-based (also referred to as template-based).

Machine learning approaches

Traditionally, machine learning follows a two-step procedure. General parameters are firstly derived from the input images, and then these parameters are fed into a trainable model to compute the quantity of interest. The pioneer neural network applied to the background rejection using IACT was a simple yet effective MLP composed of 109 neurons distributed onto 10 hidden layers [Rey93]. The inputs consisted in 8 parameters derived from the ellipsoid morphology to the signal. Nowadays, with the development of computational resources and simulation power, it is possible to increase the complexity of the model and the size of the dataset. Yet, the strategy remains the same.

Relying on the Hillas algorithm proposed in [Hil85], the CTAO standard analysis is a moment-based algorithm that aims at extracting image parameters based on their inherent properties up to their third moment. These parameters are then fed to Random Forests (RF) [al08] or boosted decision trees [OEE09], and even MLP [MGP15] to predict the physical properties of the primary particles. As described in Section 1.3 of the previous chapter, it is referred to as *Hillas+RF* when applied in coordination with RF. Because Hillas+RF leverages a background cleaning procedure that removes most of the noise from the data, the performance of this algorithm diminishes notably at lower energy levels. This decline is primarily attributed to the increased difficulty in reconstructing faint images, where the Cherenkov signal is confined to only a few pixels. Finally, the event reconstruction is addressed using a specific RF for each task. However, utilizing multiple single-task models can lead to prediction degeneracy when tasks are strongly interconnected. In contrast, template-based methods, such as Impact [PH14] and Model++ [NR09], employ likelihood functions to match the recorded signal with a bank of images. Although these techniques exhibit superior results compared to moment-based methods, especially at lower energies, the considerable computational time and resources they require will pose major challenges to be able to cope with the tremendous amount of data that CTAO will provide.

Generalization to the real scenario and data adaptation

A major drawback of the application of moment-methods is the lack of generalization capability, which turns them sensitive to distribution discrepancies, especially between the

simulations used for training and the real telescope acquisitions. This domain shift can nevertheless be reduced by taking into account measures of NSB during data acquisition. As demonstrated in [PMO22], the NSB is one of the most contributing factors of domain discrepancy and is usually approximated with a Poisson distribution. This correction, or tuning, consists in modifying the training data through the addition of a Poisson noise to match the real background. This noise padding procedure is hereby named data adaptation. Ideally, the noise should be added at the waveform level, before the pulse integration, but it is a costly process. A simpler method consists in adding the noise directly at the integrated pixel charge level. Although NSB tuning yields to an increase in performance [Jac+21], other unknown differences persist and limit the possible gains. However, detecting gamma rays at lower energies is crucial as they carry the majority of the gamma flux [Abe+23]. Neglecting this portion of the electromagnetic spectrum ultimately decreases the detection capabilities, limiting the quantity of data available for the analysis. In such scenarios, the use of deep neural networks can potentially greatly improve the performance at lower energies. On the other hand, template-based methods are intrinsically less sensitive to domain discrepancies, as NSB modelling is part of the fitting procedure. Nevertheless, unknown discrepancies cannot by definition be integrated into the process.

Stereo analysis, the future of the CTAO

The stereoscopic analysis aims to combine different images from the same events with neighbouring telescopes, as illustrated in Figure 2.8. The simultaneous detection of the air showers facilitates the determination of the direction of arrival with the identification of the intersection point of the major axis extensions of the Hillas ellipses corresponding to images captured by each camera. However, in the case of a single telescope application, determining the direction becomes significantly more complex. To remove the direction ambiguity in such cases, the asymmetry of the shower image along its major axis can be utilized, which can be quantified by considering the third moment of the intensity distribution, known as the asymmetry coefficient or skewness. Nevertheless, it's important to acknowledge that due to statistical fluctuations in shower development and potential irregularities in cleaned images, the sign of asymmetry may not always lead to the correct solution.

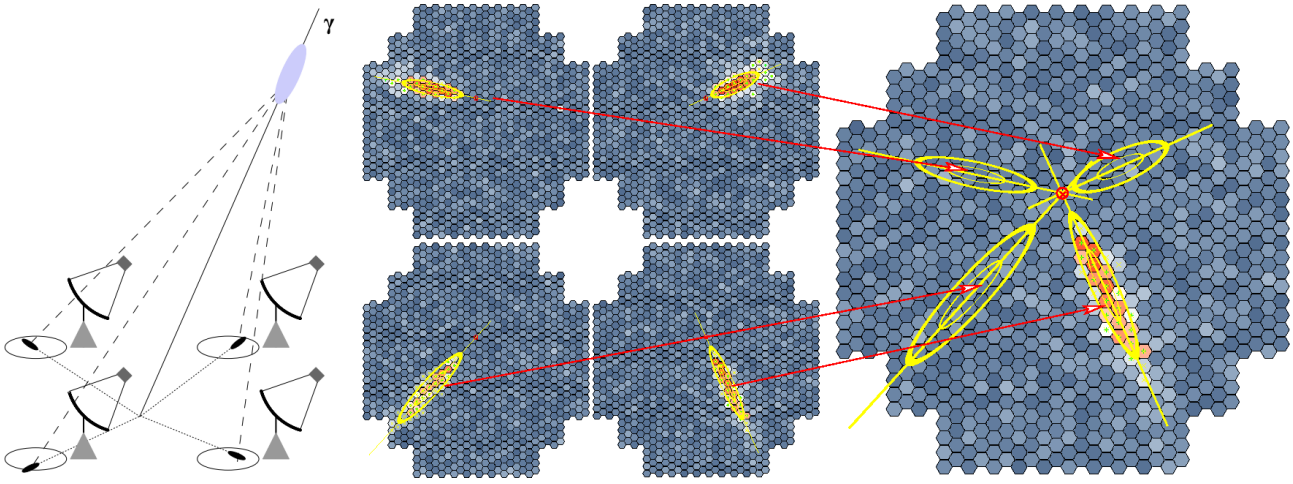


Figure 2.8: Hillas algorithm. The incident particle will produce different images on each telescope and can be combined to determine its parameters. Source [Mas10] and [VB09b].

Deep learning applied to IACTs

The application of deep learning for the event reconstruction endeavours to address the issues encountered by classical machine learning approaches. It does not require preprocessing the inputs using cleaning procedures, and utilizes the complete image information, holding the potential to enhance performance metrics, especially at low energies. A review of the existing methods has been established by [DK24], and the next section summarizes the state-of-the-art deep learning approaches applied to the main IACT collaborations.

Located on the Namibian high plateaus and named in honour of Victor Hess, the H.E.S.S. observatory consists of an array of five telescopes, including four 12-meter diameter and one 28-meter diameter instruments. In the works of [Shi+19] and [PO20], the particle classification and the direction reconstruction of the primary events are performed combining of a CNN with a Recurrent Neural Network (RNN) in the same architecture, in a stereoscopic reconstruction procedure. This aims to account for the detection sequences with the time ordering dictated by the triggers of each of the telescopes. Moreover, [PO20] has included the Hillas parameters as an auxiliary input of the neural network. Although authors shed light on an improvement for the particle classification, they measured similar results compared to Hillas+RF for the direction regression. In addition, they apply their model on real observations, and concludes that this approach is sensitive to the sky brightness in the region of the observed source. In the work of [De+22], three distinct tasks are handled using deep learning models: background rejection, multi-category classification for the specific particle class categorization, and anomaly detection to classify whether the incident particle falls into the standard model particles. They designed a CNN for the supervised task and an auto-encoder for the unsupervised one, under the assumption that the reconstruction error associated to an anomaly must increase compared to the mostly represented images in the training set. They concluded that their classifiers obtained state-of-the-art accuracy for background rejection, and encouraging results in the case of the multi-category classification.

In the case of the MAGIC observatory, authors from [Mie+21] implemented three single-task CNN-based models to regress the energy, the direction of arrival or the particle type separately. Furthermore, stereoscopy is introduced by concatenating the images into a unique multi-channel image. The domain confusion problem is tackled by cleaning the images before training the models, but this approach relies on the same image cleaning routines as the standard analysis. Besides, authors obtained a similar sensitivity of detection regarding real acquisitions compared to the standard analysis.

The CTAO is the new generation of IACTs located in La Palma and Chile. As one of the first applications of deep learning to the CTAO, [Nie+17] explored the single-image classification framework using CNNs, and highlighted the feasibility of the approach. [PST18] focuses on the full-event reconstruction problem (classification of the particle type and regression of the physical parameters) with multiple distinct CNNs in a stereoscopic scenario, demonstrating great improvements in the reconstruction capabilities. [Nie+21] define a monoscopic *TRN-single-tel* model, embedding a shallow CNN with residual connections. Although they achieve a full-event reconstruction, each task is addressed separately with a specific network. Currently, most of the methods utilize integrated pixel charges as input of their models. Yet, in contrast, most IACT cameras can read out the entire photosensor waveform. Since the arrival time of Cherenkov photons from an EAS to the camera plane

depends on the height of their radiation and the distance to the telescope, these waveforms contain additional information that could be valuable for event classification. The study by [Spe+21] investigates the potential of leveraging this waveform information using deep learning methods to suppress the air shower background caused by both protons and electrons. Although achieving effective background rejection of electrons is challenging, the addition of this information appears to enhance the performance of gamma/hadron separation. [Riq+23] explores semi-supervised learning associated to observation-like simulations with a pseudo-labelling strategy consisting of assigning the classifier's prediction as the true label of the unlabelled data. As the classifier cannot correctly classify in the early stages of the training, an epoch-dependent weighted loss allows shifting from a training dominated by labelled data to the exclusive use of unlabelled data. The feature representation is extracted using a CNN encoder and they illustrate the correlation between the Hillas parameters and the latent features. Remarkably, authors of [Jac+21] introduce the γ-PhysNet, the first full-event reconstruction neural network of the CTAO for the detection of gamma-ray sources. Oppositely to the Hillas+RF mono-task standard analysis, the neural network reconstructs each parameter simultaneously. It has been successfully applied in inference on both simulated and real data in [Vui+21], and has been compared to the standard analysis method, described in Section 2.5. The most remarkable result on simulations concerns the performance of the neural network model at low energies, where it clearly outperforms the Hillas+RF on all metrics. In addition, the performance of the model has been evaluated on real sources, and it is noticeable that the γ-PhysNet detects on average more gamma events. Yet, this study also highlights the impact of the NSB on the model results, and concludes on the necessity to implement complementary strategies in order to retrieve the loss of performance.

Addressing ML generalization issues

An important work has been initiated by [PMO22] to investigate the generalization capabilities of CNNs to various simulated observation conditions. It presents a comprehensive overview of credible sources of discrepancies affecting training and test images. Among the possible causes, dead pixels in the camera significantly degrade the particle classification accuracy, yet their impacts can be mitigated by interpolating their values with the intensity average of neighbouring pixels. Moreover, this study examines the variations in the NSB light and the dependence on the zenith angle and evaluates their effects on the background rejection performance. As gammas and protons EAS can produce similar integrated images, and the classification strategy must rely on subtle intensity and contour differences. With the help of attention, the γ-PhysNet neural network presented in [Jac20] increases its robustness in the gamma/hadron separation task and manages to focus on the most relevant discriminative features. Furthermore, as illustrated in [Vui+21], the direct application of deep learning models to real data analysis engenders measurable biases. In more detail, regarding the position reconstruction of Markarian 501, the spatial coordinates of the source are shifted compared to its real location. The privileged hypothesis points out the slight difference between the pointing directions of the telescope during the observation and of the simulations used for training the model. Secondly, the Crab nebula analysis sheds light on the role of NSB as a major factor of domain discrepancy, in concordance with [PMO22]. The current solution consists in generating new training data with the addition of a Poisson

noise within the images to match the observation background. This strategy is referred to as data adaptation in this thesis. The γ-PhysNet and Hillas+RF have been compared on both training datasets, and it has been evaluated that the addition of a Poisson noise boost the performance of both methods.

Focus on the γ-PhysNet

The γ-PhysNet neural network forms the groundwork of our research, and we provide in the section a more detailed overview of the model. Illustrated in Figure 2.9, it is composed of two entities that are respectively a ResNet feature extractor [He+15a] augmented with attention mechanisms [Fu+19], and a multi-task architecture for the reconstruction of an incident particle physical parameters. The former is denoted by the parametric function $G_f^{\theta_f}$, while the multi-task component contains an energy, direction, impact and particle classification branch respectively defined by the parametric functions $G_\epsilon^{\theta_\epsilon}$, $G_\alpha^{\theta_\alpha}$, $G_\delta^{\theta_\delta}$ and $G_c^{\theta_c}$.

The model must be able to discriminate between a gamma and a proton; hence the particle classifier must be trained on both classes. However, as the main goal of the neural network is the full-event reconstruction of incident gamma particles, the parameters associated with the energy, direction and impact reconstructions are only optimized on gammas. This can be achieved considering a particle mask, defined as the following characteristic function:

$$\mathbb{1}_\gamma(y) = \begin{cases} 1 & \text{if } y = y_\gamma \\ 0 & \text{otherwise} \end{cases} \tag{2.19}$$

where y_γ is the gamma label. Then, the regression loss functions are computed as the multiplication of the selected criterion and the particle mask. Overall, the global objective function is the sum of each mono-objective. Regarding the regression tasks, the Mean Absolute Error (MAE) is preferred, and the classification is performed with Cross-Entropy (CE).

$$\begin{aligned} \mathcal{L}(\theta_f, \theta_\epsilon, \theta_\alpha, \theta_\delta, \theta_c) \quad &= \lambda_{\text{energy}} \textstyle\sum_i MAE\left(G_\epsilon^{\theta_\epsilon}(G_f^{\theta_f}(x_i)), y_i\right) \\ &+ \lambda_{\text{direction}} \textstyle\sum_i MAE\left(G_\alpha^{\theta_\alpha}(G_f^{\theta_f}(x_i)), y_i\right) \\ &+ \lambda_{\text{impact}} \textstyle\sum_i MAE\left(G_\delta^{\theta_\delta}(G_f^{\theta_f}(x_i)), y_i\right) \\ &+ \lambda_{\text{class}} \textstyle\sum_i CE\left(G_c^{\theta_c}(G_f^{\theta_f}(x_i)), y_i\right) \end{aligned} \tag{2.20}$$

where the coefficients λ_{energy}, $\lambda_{\text{direction}}$, λ_{impact} and λ_{class} are the weighting coefficients that the contribution of each loss. In the case of the γ-PhysNet, they are dynamically determined through an automatic estimation procedure. Following our ablation study available in Appendix 8.11, the Uncertainty Weighting algorithm introduced in [KGC18] provides the best performance over the selected methods described in Section 2.8.

Limitations and perspective in the generalization to real observations

The application of γ-PhysNet to real data reveals both the strengths and limitations of current deep learning approaches. On one hand, the exploitation of image background noise,

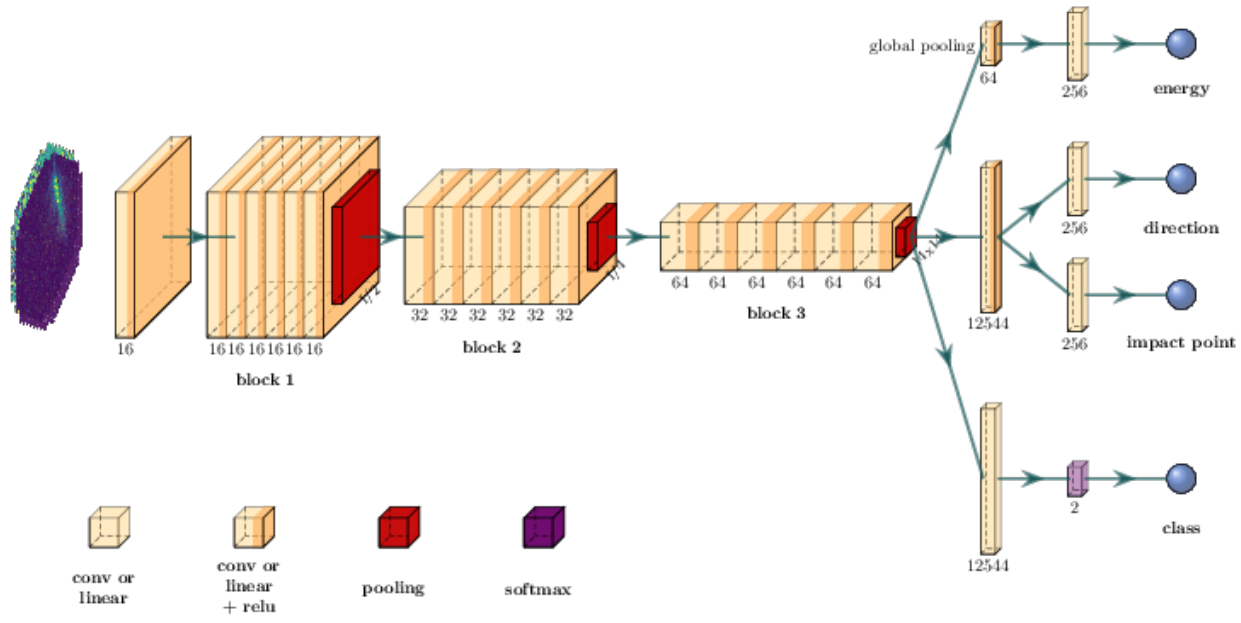


Figure 2.9: The γ-PhysNet architecture.

combined with the utilization of data-driven optimization of tailored filters through the deep learning paradigm enable the detection of more gamma events. Thereby this allows the computation of high-level figures of merit with greater statistical power compared to the standard analysis. On the other hand, these tailored filters have a significant drawback: training the network on simulated data lacks generalization and introduces biases when it is applied to real acquisitions. To address this, data adaptation, which consists in NSB matching, has proven essential for enhancing network performance. This necessity underscores the importance of research into generalization techniques, leading to deep learning methods that address domain discrepancy, commonly referred to as deep domain adaptation and information fusion.

2.6 Information fusion as a mechanism to inject additional knowledge to the network

Machine learning models often struggle with robustness when the test distribution differs from the training one. Although the representativity of the data may be sufficient, limitations can be expected because the neural network has no access to external yet crucial parameters. Thus, enhancing the visibility of additional information to the network is critical. These additional determining factors can take various forms. For example, considering the land classification problem, satellites embed different sensors gathering multi-spectral images (from different sources), and inputs become naturally multi-channel. Oppositely, the integration of different modalities in the form of scalars, images, or texts, can significantly enhance the capabilities of vision deep learning systems by providing a richer context. Analogously, humans naturally integrate data through our five senses, enabling the brain to perceive and process a diversity of sensations from various sensors. This learning approach, referred to as multi-modal, aims to leverage complementary information from

different aspects to improve the performance of machine learning models, making them understand and interpret data more humanlike [Ion+14]. Key techniques for multi-modal learning in neural networks are listed here:

- Attention mechanisms [Vas+17] allow neural networks to focus on the most relevant parts of the input data, improving the handling of complex tasks by dynamically weighting the importance of different input features. Their designs are well suited for information fusion. In Transformers, external variables can be projected using a trainable MLP (or DNN) into tokens, that can be integrated into the image token representation. They are then put into perspective with the other projected patches within the mechanism.

- Combining features from different modalities can be done at various stages of the neural network, including early fusion, late fusion, and intermediate fusion. In that case, data from different modalities can be either concatenated or merged.

- The additional variables can be used as a conditioning input through the design of the Conditional Batch Normalization (CBN) modules, as introduced in [Vri+17]. Batch Normalization layers are extended to learn not only scaling parameters, but also shifts that are conditioning bu the new inputs. A more detailed overview of this technique is given in the next chapter.

Information fusion in deep learning is an interesting choice to add external knowledge to the network. The selection of the most promising technique is rather experimental, and highly depends on the study case. Yet, they are a compelling alternative to data adaptation, which consists in modifying the data preferably, while we are interested in that case to modify the model structure.

2.7 An overview of the domain adaptation approaches

Current applications of deep learning in IACT data analysis highlight the specialization problem that occurs when training and test data are not drawn from the same population. In fact, this issue is common in many real-world scenarios where high-quality datasets are essential for training effective and robust deep learning models. Biases and representational issues can significantly affect performance and generalization ability, and prior works shed light on this systematic effect, showing that datasets are inherently prone to biases [TE11]. In this paper, a simple demonstration involves considering several distinct datasets representing the same scenes and classes, and training a classifier to identify the dataset from which an input image is sampled. Despite the expectation that this would be a difficult task due to the varied and rich nature of the datasets, the model performs significantly better than random guessing, achieving 39% accuracy instead of 8%. This outcome indicates that each dataset has a unique, strong signature or built-in bias. Rather than representing the real world, these datasets become closed systems. As a result, training a model on one dataset and testing it on another leads to significantly degraded performance, revealing a lack of generalization across different but related domains. This situation underscores the need for domain adaptation techniques to improve model robustness and generalization.

Many overviews of domain adaptation have been proposed in the literature [WD18], [KL21], [Csu17], [Zha19], [WC18], [Zha+20]. This thesis will rely on [WD18] and [Zha+20] to introduce the notations and definitions. Overall, domain adaptation takes part of a broader context that is referred to as transfer learning. Figure 2.10 illustrates the taxonomy of transfer learning with a focus of domain adaptation methods that we cover in this work.

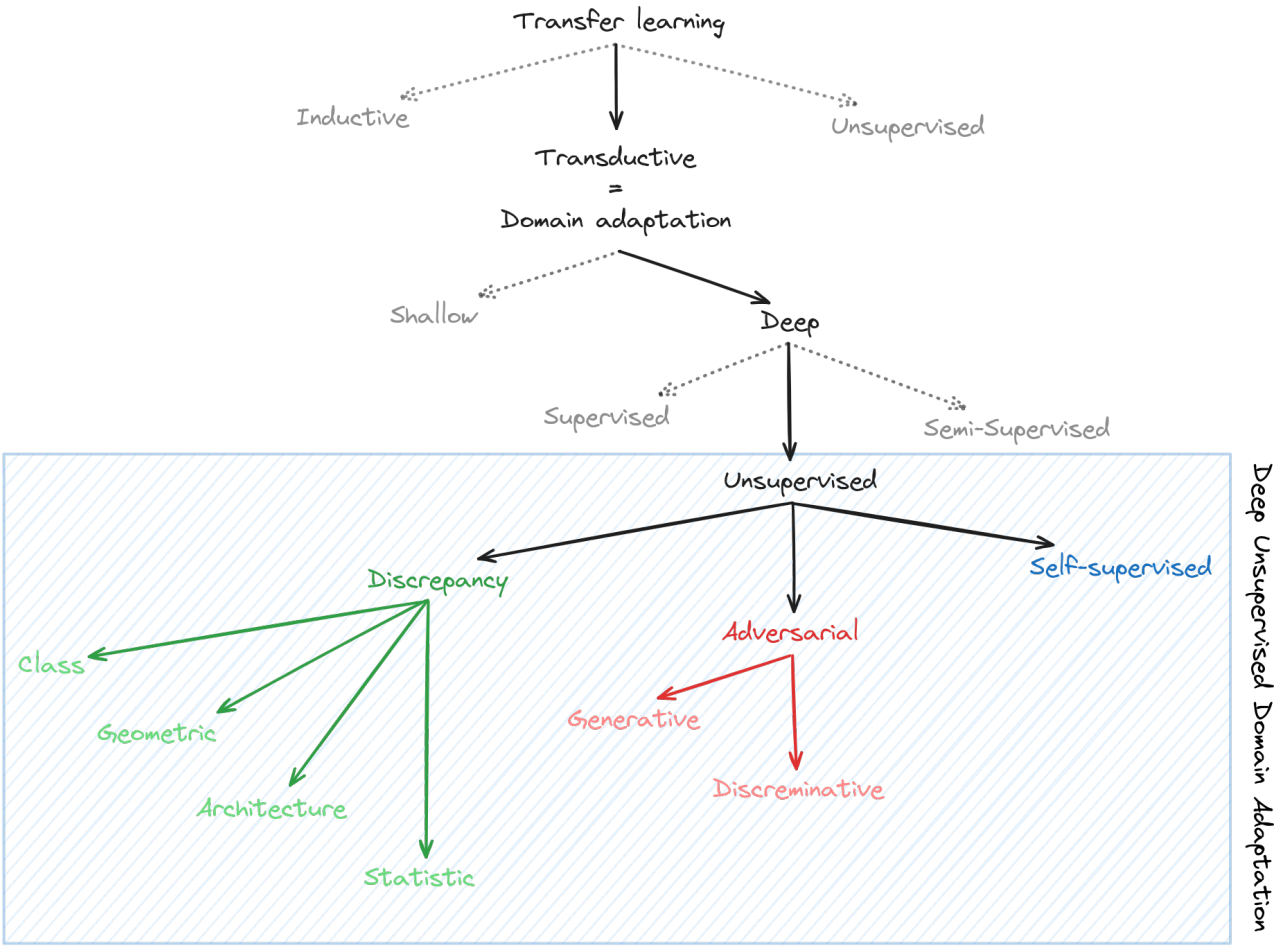


Figure 2.10: Domain adaptation overview.

Domain adaptation as a subset of transfer learning

Transfer learning [Zhu+19] is considered when the training dataset lacks sufficient data to conduct a training from scratch, or when a model leverages the knowledge from another network as a starting point. In this approach, a pre-trained model is set as the starting point and provides useful features for the new task. This method relies on the observation that most neural networks exhibit a common behaviour, where the first layers function as general Gabor-like filters or colour blobs, which are not specific to any particular dataset or task [Yos+14]. Conversely, the features in the final layers are tailored to particular tasks. It turns out that initializing a model with features from a distant task yields better generalization performances than a random initialization [Yos+14].

Notations and definitions

We introduce in the following the taxonomy that is used in this thesis. A domain $\mathcal{D}$ consists in a feature space $\mathcal{X}$ and a marginal probability distribution $P(X)$, with $X = \{x_1, ..., x_n\} \in \mathcal{X}$ corresponding to a subset of samples. In computer vision, the feature space is usually defined as the pixel space, such that $\mathcal{X} = \mathbb{R}^d$, where d represents the number of pixels. Given a domain $\mathcal{D} = \{\mathcal{X}, P(x)\}$, a task consists in another feature space $\mathcal{Y}$ and an objective predictive function $f(.)$. From a probabilistic perspective, this function can also be interpreted as a conditional probability distribution $P(y|x)$, which can be learnt from the labelled data $x_i \in \mathcal{X}$ and $y_i \in \mathcal{Y}$. We consider the case study of unsupervised domain adaptation with an annotated source domain $\mathcal{D}^s = \{\mathcal{X}^s, P^s(x)\}$ and a target collection that contains no label information $\mathcal{D}^t = \{\mathcal{X}^t, P^t(x)\}$. They are respectively associated with the tasks $\mathcal{T}^s = \{\mathcal{Y}^s, P^s(y|x)\}$ and $\mathcal{T}^t = \{\mathcal{Y}^t, P^t(y|x)\}$. Source and target samples are denoted respectively as x_i^s and x_i^t.

The traditional machine learning scenario sets $\mathcal{D}^s = \mathcal{D}^t$ and $\mathcal{T}^s = \mathcal{T}^t$, in other words the source and target datasets are sampled from the same underlying distribution, and the source and targets tasks are identical. On the other hand, transfer learning defines three environments where either $\mathcal{D}^s \neq \mathcal{D}^t$, or $\mathcal{T}^s \neq \mathcal{T}^t$, or both, and there are referred to as respectively inductive, transductive and unsupervised. Using the following definition, domain adaptation appears to be a subset of transfer learning, and belongs to transductive learning, that is to say subject to similar tasks but different domains. Furthermore, the literature defines three types of domain shifts that are the covariance shift, the label shift and the concept shift defined below.

Covariate shift

Covariate shift, or feature shift, appears when the conditional source and target distributions of y under the condition x are similar, but the marginals according to X are different:

$$\forall x, P^s(y|x) = P^t(y|x) \text{ but } P^s(x) \neq P^t(x) \tag{2.21}$$

In the case of IACTs, simulations are abundant and derived from equations of the underlying theory, but simplifications are intrinsically present and does not perfectly reflect the reality. In the following, simulations are considered as the source dataset, while real data corresponds to the target dataset.

Label shifts

The originality of domain adaptation lies in incorporating the unlabelled target data in the training. However, there are also no guarantees that the source and target label distributions, respectively $P^s(y)$ and $P^t(y)$ are similar for all labels y. Formally, label shifts thus exist when the conditional source and target distributions x under the condition y are similar, but the marginals according to Y are different:

$$\forall y, P^s(y|x) = P^t(y|x) \text{ but } P^s(y) \neq P^t(y) \tag{2.22}$$

Label shift is strongly present in IACTs. Simulated datasets are usually well balanced, with an equivalent amount of gammas and protons. In real acquisitions, as illustrated in Figure 1.2, the proportion of particles varies depending on the energy. Most of the approaches of the literature rely on importance weighting, which consists in estimating the

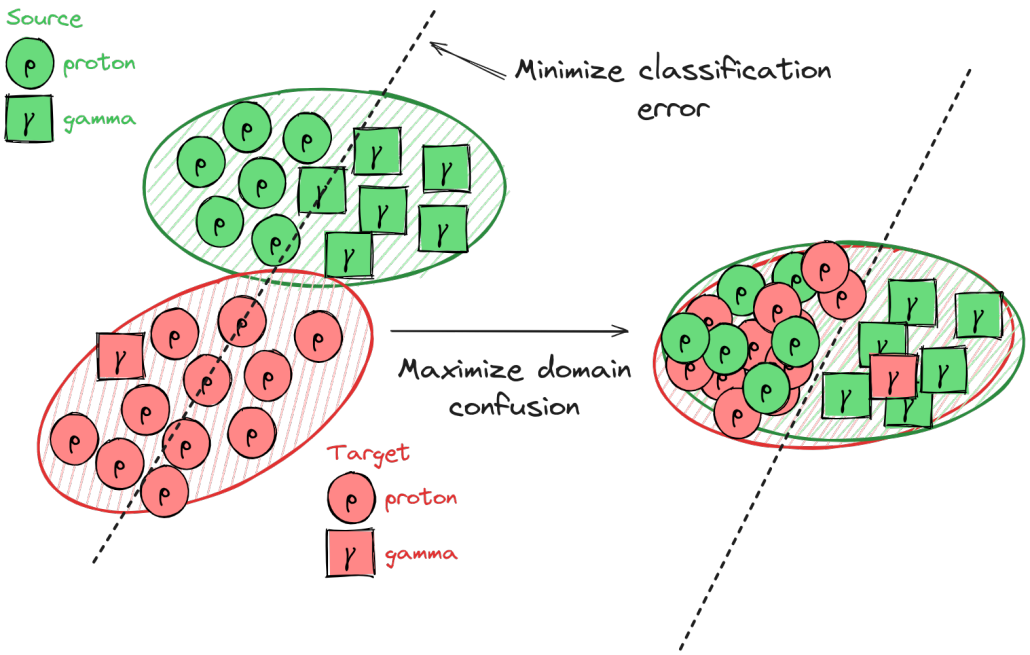


Figure 2.11: Illustration of domain confusion in the case of the CTAO image analysis. The source (simulations) and target (real acquisitions) datasets are dissimilar, which leads to distinct supports. Training on the source dataset to minimize the classification error finds a separating hyperplane that will successfully split the two represented classes in the source domain, but not in the target domain. Hence, extending the loss function with a domain confusion objective to maximize will result in finding a hyperplane that will correctly separate the two classes in both domains. Furthermore, the ratio of particles in both domains are different.

ratio of the source and target classes $\omega(y)$, and re-weight the target distribution accordingly [Liu+21a; Zha+13; Azi+19; LWS18], so that:

$$P^s(y) = \omega(y)P^t(y) \tag{2.23}$$

Although this approach yields increasing performance, it is only applicable when such ratio is still acceptable, that is to say each mini-batch contain enough statistics.

Concept shifts

Concept shifts describes the case where the joint probability is different for the source and the target domain.

$$P^s(x,y) \neq P^t(x,y) \tag{2.24}$$

A measure of domain discrepancy

We can define two distributions on the space of probability measures $P^s, P^t \in \text{Prob}(\mathcal{X})$ with $\mathcal{X}$ a compact metric set, such as $\mathcal{X} = [0,1]^d$ for the space of images. Let $\Pi(P^s, P^t)$ denotes the set of all joint distributions $\gamma(x,y)$ with P^s and P^t the respective marginal distributions. Theoretically, domain discrepancy can be measured according to [ACB17] using multiple metrics:

- The total variation (TV) distance is calculated with the subset of points that maximizes the divergence.

$$\delta(P^s, P^t) = \sup_{A \in \Sigma} |P^s(A) - P^t(A)|$$

- The Kullback-Leibler divergence

$$KL(P^s \| P^t) = \int \log\left(\frac{P^s(x)}{P^t(x)}\right) P^s(x) d\mu(x)$$

- The Jenson-Shannon distance, that aims at solving the asymmetry of the KL-divergence

$$JS(P^s \| P^t) = KL(P^s \| P^m) + KL(P^t \| P^m)$$

- The Wasserstein distance, also referred to as the Earth Mover distance, can be defined as :

$$W(P^s, P^t) = \inf_{\gamma \in \Pi(P^s, P^t)} \mathbb{E}_{(x,y)\sim\gamma}[\|x - y\|]$$

$$= \sup_{\|f\|_L \leq 1} \mathbb{E}_{x\sim P^s}[f(x)] - \mathbb{E}_{x\sim P^t}[f(x)]$$

  where the supremum is applied on all the functions $f : \mathcal{X} \to \mathbb{R}$ which respect the 1-Lipschitz constraint.

- The Proxy A-distance (PAD), introduced in [Ben+06], can be implemented as an error measure derived from the accuracy of a domain classifier that aims to classify whether the input data is sampled for the source or the target distribution. If $\mathcal{H}$ defines the hypothesis class, PAD is computed as:

$$d_A(P^s, P^t) = 2\left(1 - 2 \min_{h' \in \mathcal{H}} \text{err}\,(h')\right)$$

Overview of domain adaptation

A possible multi-step adaptation procedure

At the broadest scale, domain adaptation can be subdivided into two groups [Zha+20]. When the source and the target distributions are assumed to be closely related, knowledge transfer can supposedly be executed in one-step, referring to one-step domain adaptation techniques. Oppositely, when the two datasets are distant, intermediate domains can be computed to help gradually reduce the discrepancy, leading to multi-step domain adaptation methods.

Besides, the source and target feature spaces $\mathcal{X}^s$ and $\mathcal{X}^t$ may be either identical or different. In the first case, we refer to homogeneous domain adaptation, whereas in the second case we refer to heterogeneous domain adaptation. Hence, the difference comes from the representativeness of the elements constituting the distributions.

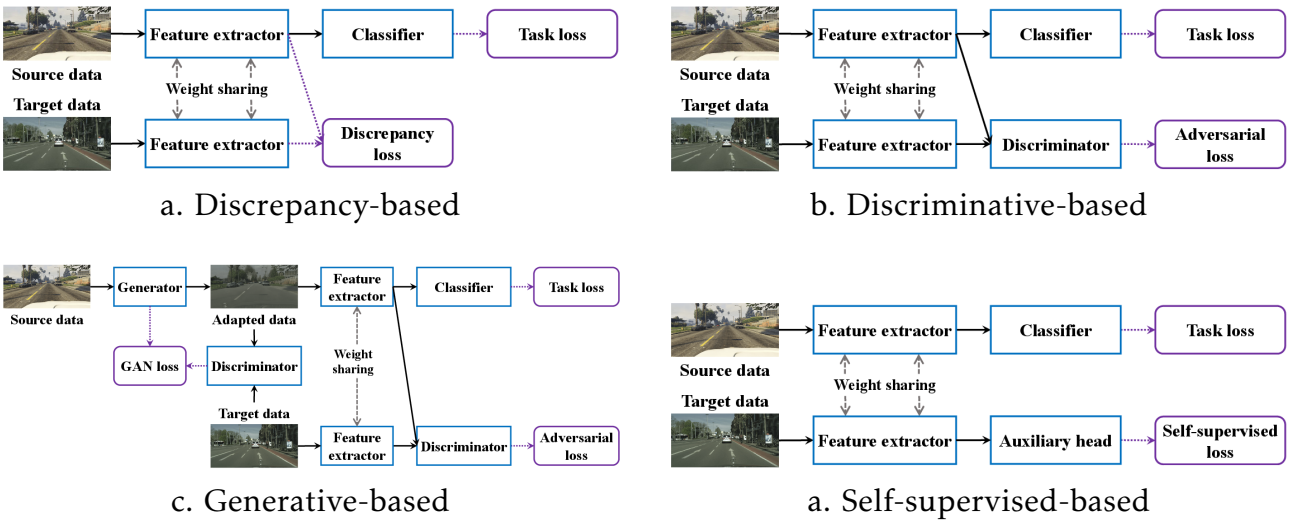


Figure 2.12: Unsupervised domain adaptation overview. Source [Zha+20].

Shallow and deep domain adaptation

Shallow and deep domain adaptation refer to different approaches to adapting the model considering a target domain. The former focuses on adapting the model at the input - e.g. the pixel space for an image - or at the feature level, while the latter modifies the model's internal structure. In general, deep domain adaptation tends to be more effective than its shallow variation, because it allows the model to learn a more nuanced and domain-specific or invariant representation. Subsequently, the succeeding chapters of this thesis will solely discuss the use of deep domain adaptation.

The context of supervision

Each of these categories can be subdivided into three cases depending on the availability of the labels. Firstly, in supervised domain adaptation there is access to labelled target data, but it may not be sufficient in number. Secondly, in semi-supervised domain adaptation one has access to a very limited amount of labelled target data which can nevertheless provide some information to the network. Thirdly and lastly, in unsupervised domain adaptation there is no access to any labelled target data. In the context of this thesis simulated and real data have different feature spaces and no labelled are available, hence the configuration falls into unsupervised domain adaptation.

Deep unsupervised domain adaptation

We progressively reduced the span of domain adaptation to find the most appropriate set of algorithms that fit our CTAO use case and we found out that deep unsupervised one-step domain adaptation reflects the most of the requirements. This category of algorithms can further be subdivided into four types of models, as illustrated in Figure 2.12.

Discrepancy-based methods

Firstly, discrepancy-based methods aim to minimize a measure of disparity between distributions. Although the computation of such distance or divergence theoretically requires the

knowledge of the underlying distributions, neural networks offer the flexibility to estimate them stochastically. Various standard metrics exist, involving for example the computation of moments. In the case of the first-order statistics, it is referred to as Maximum Mean Discrepancy (MMD) [Tze+14], and its empirical approximation can be calculating on the mean feature maps of the neural networks. [Lon+15] introduces characteristic kernels to calculate the mean embedding of the feature vectors, highlighting that the application of multiple kernels contribute in enhancing the classification performance. Deep Correlation Alignment (DeepCORAL) [SS16] aims to minimize the distance of the second-order statistics of the source and target latent features of size d, the correlation alignment quantifying the misalignment between the covariance matrices C_S and C_T through the optimization of the following objective function:

$$\mathcal{L}_{DeepCORAL} = \frac{1}{4d^2}\|C_S - C_T\|^2_{Frobenius} \tag{2.25}$$

As reducing the first- and second-order moments are not adapted for non-Gaussian distributions, Higher-order Moment Matching (HoMM) [Che+19] unifies the framework by extending these approaches to higher-level statistics coupled with characteristic kernels. Other popular approaches leverage Optimal Transport (OT) theory to calculate the Wasserstein distance. OT [Vil08] is a field of mathematics studying the optimal transportation of one measure onto another, that is to say with a minimum cost. It has been recently introduced in domain adaptation [Cou+15] for several reasons:

- It allows to find a transformation that maps the input data into the source and target distributions following the principle of least effort.
- It describes a framework to compute distances between empirical probability distributions.
- New computation strategies of the optimal transportation plan, especially stochastic methods, have made this discipline suitable for real-world applications and large datasets.

Following this approach, Deep Joint Distribution Optimal Transport (DeepJDOT) [Dam+18] relies on optimal transport to estimate the Wasserstein distance between both domains on the deepest latent features which, when minimized, ensures the overlapping of both distributions. The corresponding objective function is usually defined as:

$$\mathcal{L}_{DeepJDOT} = \sum_{i,j} \pi_{i,j} C_{i,j} \tag{2.26}$$

where π corresponds to the optimal transport plan and C to the cost matrix computed from the encoded source and target latent representations $G(x_i)$ and $G(x_j)$ as:

$$C_{i,j} = \|G(x_i) - G(x_j)\|^2 \tag{2.27}$$

. Domain-Transformer (DoT) [Chu+22] is an optimal transport related Transformer model that focuses on learning semantic consistency across both the source and the target domains to improve generalization of the classifier. Cross-Domain Transformers (CDTrans) [Xu+21] relies on pseudo-labels and uses cross-attention to reduce the impact of label noise when labels are not correctly assigned to the target samples.

Discriminative-based methods

Secondly, adversarial discriminative models implement a domain discriminator along with a dual min-max objective function to learn domain-invariant features, such as Domain Adversarial Neural Network (DANN) [Gan+16]. The feature extractor G is complemented by an additional objective that is to mislead the domain classifier D, and therefore to ensure domain invariance while providing good performances on the other classification or regression task. As an adversarial optimization problem, it can be formulated using a Gradient Reversal Layer (GRL) $\mathcal{R}$, a pseudo-function that reverses the sign of the gradient during the backward pass. In practice, the propagated gradient returned from the domain classifier is weighted by an epoch-dependent function, ensuring the training of the feature extractor on the supervised task in the first place before incorporating the adaptation process smoothly over time. DANN objective function can thus be defined, using the Cross-Entropy (CE) metric, as:

$$\mathcal{L}_{DANN} = \sum_i CE\left(D(\mathcal{R}[G(x_i)]), d_i\right) \tag{2.28}$$

Furthermore, the domain classifier can be replaced with a domain metric. As the Wasserstein distance indeed has nicer properties regarding the gradients during backpropagation, it is considered as a promising tool to update and improve the model's performances [She+18]. Such an approach necessitates imposing a Lipschitz constraint on the domain metric, which can be easily implemented with a penalization on the derivative of the metric gradients [Gul+17].

Although DANN manages to find domain-invariant representations through the domain classifier, the generated features may be ambiguous when the data lie close to the class boundaries and therefore may not be classified or regressed accurately. Besides, if the distributions are very different, finding a good shared representation may be difficult. Those issues can be solved with the maximum classifier discrepancy method [Sai+18] which replaces the discriminator or critic with a pair of task-specific classifiers. In more detail, target samples associated with two different labels are more likely to belong to the disagreement region. It turns out that minimizing the disagreement between the classifiers is equivalent to forcing the feature extractor to not generate features that are outside the source support, hence leading to a domain alignment. [Lee+19] is the natural extension of [Sai+18] with the use of the Wasserstein distance. This article proposes to replace the discrepancy loss function with a sliced 1-Wasserstein discrepancy, a 1-dimension variational formulation of the 1-Wasserstein distance [KMR18].

The previously introduced algorithms share the fact that the adversarial domain task and the classification or regression task do not cooperate while getting optimized, that is to say they are considered independently and thus can lead to poorer performances. [Wei+21] proposes to include domain adaptation into a meta-learning scheme. Meta-learning usually refers to the *learn how to learn* paradigm and has been developed in domain generalization [Li+17], which is a more general context than unsupervised domain adaptation in the sense that we do not have access to any target data at all. In such approaches, one searches to generalize by creating virtual test domains. [Wei+21] leverages meta-optimization by considering the domain alignment as the meta-train and the classification task as the meta-test.

In an adversarial context, balancing the two cooperative entities is crucial but challenging. Moreover, strong domain-specific features may impact the quality of the adaptation. In this context, [Cui+20] defines the concept of bridges, aiming to model the shift between the domain-specific and domain-invariant features, where gradually vanishing bridges leads to reduce domain discrepancy.

Generative models. Oppositely, adversarial generative methods are based on Generative Adversarial Networks (GANs) [Goo+14], and aim to proceed image-to-image translation [Iso+16]. GANs have been a great tool to generate rich contents such as images. They rely on a cooperative game between two entities that are trained simultaneously, namely a generator and a discriminator. The former aims at synthesizing images resembling real data, whereas the latter tries to distinguish real data from synthesized ones. Although they have been widely used in many different fields, they also suffer from some drawbacks such as vanishing gradients, poor gradient quality or mode collapse [Sal+16]. As a result, the training procedure may be a difficult task. Wasserstein GANs [ACB17] (WGANs) aims at improving the quality of convergence with the use of the Wasserstein distance function, which has better properties regarding the gradients compared to the Jensen-Shannon divergence that is classically used in GANs. WGANs allow the discriminator to be trained optimally which will result in high quality gradients, allows more robustness towards the generator architecture and solve the mode collapse issue. However, in order to enforce the necessary Lipschitz constraint on the discriminator, hereafter called a critic, its weights must lie in a compact space, which can be achieved by clipping the weight values after each gradient updates. Some improvements have been proposed [Gul+17] with the addition of a gradient penalty in the objective function. However, if the training set is composed of both the simulated and the real telescope images, these mentioned techniques will focus on learning the union of the distributions, and samples will be drawn randomly from the generator. Conditional GANs [MO14] (cGANs) allow conditioning the basic GAN architecture by adding extra information similarly to both the generator and the discriminator. In that case, the added condition could either be related to physical parameters, permitting synthesizing images with the desired features, or $c \in \{\text{real}, \text{simulation}\}$ allowing generating data from either one or the other distribution. cGANs can be improved by the addition of an L1 penalty is the loss function [Iso+16]. The application of GANs for unsupervised domain adaptation is not immediate, but they can be used as building blocks for more advanced techniques, such as image-to-image translation methods which consist in transferring the representation of one domain into another. For example, COGAN [LT16] is one type of such algorithm combining unpaired images with the use of two GANs synthesizing images from each domain. CycleGAN [Zhu+20] introduces the cycle-consistency constraint to perform unpaired image-to-image translation and ensure that the mappings are bijective. While discriminative models aims to obtain domain-invariant features, these approaches aim to project one distribution into another.

Nowadays, Transformers have largely replaced GANs in image synthesis thanks to the global context understanding they can capture better than CNNs, which are typically used in the latter. This allows for more coherent and detailed image synthesis, especially in complex scenes. Moreover, Transformers can be more easily parallelized during training compared to GANs, which often requires careful balancing between the generator and discriminator, making training large-scale models faster and more efficient. Furthermore, they

are highly versatile and can be used in multi-modal tasks where image synthesis is conditioned on other types of data, such as text, thanks to their performing embedding system. For these reasons, Image-to-image Translation with Transformers [Zhe+22] has been proposed as an unpaired image-to-image translation network following an encoder-decoder architecture and combines depth-wise convolutions along with self-attention into a hybrid perception block to highlight short- and long-range dependencies.

Self-supervised

Lastly, self-supervised methods integrate auxiliary self-supervised learning tasks into the primary task network, as co-training proves beneficial in reducing the dissimilarity between the source and target domains. Such methods, for example Multi-Task Auto-Encoder [Ghi+15], learns to transform the initial distribution into analogues, so that the features that are extracted from the encoder are robust to domain variations. In a second place, these invariant features are used as the input of a classifier or regressor.

Connections with multi-task balancing

Most of the domain adaptation papers solely rely on evaluating their methods with a single classification or segmentation task, although real-world scenario requires a combination of tasks that must be optimized simultaneously. Yet, this context is marginally represented in the literature.

Cross-Domain Self-supervised Multi-task Feature Learning [RL18] aims at solving simultaneously three different tasks from a synthetic dataset. Firstly, instance contour detection consists in finding an edge map, that allows the network to learn continuous perimeters around objects. Secondly, the depth prediction requires the network to understand the shape and the relative position of the observed objects in the scene. Finally, the surface normal estimation is the last task and is highly related to the depth estimation. Combining the last two tasks improves their solving and it turns out that combining in our case the direction with the impact point helps improve the physical parameter reconstructions. The domain adaptation approach that has been used is similar to DANN [Gan+16] and refers to the implementation of a domain classifier competing with the feature generator in an adversarial way. Another interesting remark is that minimizing the domain discrepancy in an intermediate latent space layer rather than the last one produces better results. Unsupervised Multi-Task Domain Adaptation [YY21] is the extension of the Image-to-Image Translation for Domain Adaptation [Mur+18] to the multi-task paradigm. [ZLO19] identified that the domain-invariant feature extraction, the domain-specific reconstruction and the cycle consistency were the three dominant fundamentals of a performing unsupervised domain adaptation.

However, these articles share that the multi-task attributed weights are user-defined hyperparameters. To the best of our knowledge, there is no work in the literature aiming to integrate domain adaptation with automatic task-balancing.

2.8 Multi-task learning

Multi-task balancing (MTB) [Car04] refers to the simultaneous optimization of distinct but generally related tasks during the training step. Analogically, learning a new language is easier when confronting speaking, writing and listening together. The classical approach for determining the loss weights w_t that are in competition during the training process is conducted through a grid search optimization, and the choice of this hyper-parametrization has a critical impact on the performance of the model, even on a low-complexity dataset. The global loss function thus becomes a weighted sum (using coefficients w_t) of the T mono-objectives:

$$\mathcal{L} = \frac{1}{T}\sum_{t=1}^{T} w_t \mathcal{L}_t \tag{2.29}$$

The computational complexity thus increases exponentially as the number of tasks grows, shedding light on the need for task auto-balancing.

In addition to reducing computation costs, the advantage of multi-task learning is twofold. Not only designing an architecture performing on multiple and complementary problems at the same time allows each task to benefit from each other [He+17], but also auxiliary tasks can be included to help constrain the problem and improve the overall performance [Jac+21].

Many automatic loss balancing methods have been proposed in the literature [ZY17]. Usually defined as baselines, Equal Weighting (EW) and Random Loss Weighting (RWL) [Lin+22] are the most straightforward balancing procedure and respectively consist in applying the same coefficient to each task, or to sample them from a distribution, typically Normal. When the weights are tuned using a grid search procedure or with the expert's knowledge, they fall into the scope of Manual Weighting (MW).

Automatic task balancing method overview

The contribution of a task to the optimization of the model parameters can be quantified by the norm of its gradients. Depending on their implementations, certain tasks may have strong gradients at the beginning of the training, thereby driving parameters into a global minimum where the other tasks may potentially not optimize properly. In order to introduce task balancing, [Che+18] instantiates the GradNorm (GN) algorithm. This method aims to weight the task by a measure of their contribution through the determination of the inverse training rate r_t and its projection on a common scale $\overline{G}_W$ by multiplying it with the norm of the gradients $G_W^{(t)}$ calculated at the last shared layer W. It corresponds to finding the weights that minimize the following GradNorm objective function:

$$\mathcal{L}_{grad}(w_1, ..., w_T) = \sum_{t=1}^{T} \left| G_W^{(t)}(i) - \overline{G}_W r_t(i)^{\alpha} \right|_1 \tag{2.30}$$

where α remains the last hyperparameter that tunes the power of the multitasking effect.

Although GN only considers the norm of the gradients, their orientations appear to play a role in the quality of training. In the very high dimensional landscape of the weights, tasks can be in opposition if the gradients are not co-linear, thus their association can be counterproductive. A simple measure of conflict is the cosine similarity. [Yu+20] proposes to

solve conflicting gradients by projecting them into non-conflicting basis. [Gua+22] identifies the shared layers that are conflicting the most frequently to change them to task-specific parameters.

Inspired by GN, the Dynamic weight averaging (DWA) [LJD19] algorithm self-updates over time by assessing the rate of loss change for each task. However, in contrast to GN, which necessitates access to the model's internal gradients, DWA only relies on the numerical task losses, which ease significantly the implementation of the method. The coefficients at the time $n \in \{1,...,N = \text{max epoch}\}$ can be computed as follows:

$$\lambda_t(n) \quad = T \times Softmax(w_t(n-1)) = \frac{T \times exp\left(\frac{w_t(n-1)}{Temp}\right)}{\sum_{k=1}^{T} exp\left(\frac{w_k(n-1)}{Temp}\right)} \tag{2.31}$$

where $Temp$ is a hyperparameter that defines the temperature monitoring the softness of the task weighting. If $Temp \to +\infty$, then $\lambda_t \to 1$ and we fall back into the EW procedure. The relative descending rate w_t is defined as:

$$w_t(n-1) \quad = \frac{L_{t-1}}{L_{t-2}} \tag{2.32}$$

where have $\sum_{t=1}^{T} \lambda_t = T$.

Another relevant task balancing approach is Uncertainty Weighting (UW) [KGC18] that relies on the task-dependent homoscedastic uncertainty to estimate the loss coefficients. This uncertainty can be defined as the standard deviation of the error distribution under the assumption that the associated likelihood follows either a Normal, Laplacian or Softmax distribution, which intrinsically limits the use case of this approach to a small selection of loss functions. The global objective loss can be derived as:

$$\mathcal{L}(s_1,...,s_T) = \sum_{t=1}^{T} K e^{-s_t} \mathcal{L}_t + s_t \tag{2.33}$$

where $s_t = log(\sigma_t^2)$ defines the log-variance. The constant K depends on the likelihood function, and $K = 2$ for Normal or Laplacian distributions and $K = 0.5$ for the SoftMax distribution.

Synthesis on multi-task balancing

Current applications of multi-task balancing within the γ-PhysNet framework involves optimizing the tasks through UW. While this technique enhances performance, it mathematically restricts the use of a subset of loss functions. Some concerns arise because, in the context of domain adaptation, many approaches aim to reduce domain discrepancies using mathematical tools that are not strictly likelihood functions, and thus cannot be directly integrated into the current framework. Therefore, there is a need to implement and evaluate alternative balancing procedures that are agnostic to the domain adaptation objective function. Evaluation of the selected methods are displayed in the Appendices 8.11 and 8.6.

2.9 Conclusion

IACT image analysis, through the reconstruction of the incident particle physical attributes, is inherently a multi-task process where each task is correlated. A common solution in standard analysis involves processing these tasks sequentially using multiple RF models, with the output of one model serving as the input for the next one. Leveraging the strengths of multi-task learning, it has been demonstrated that simultaneously reconstructing the attributes can lead to better performance [Jac20], yet the challenge resides in the complexity of optimizing the objective function. Balancing the weights can be conducted through automatic and dynamic procedures, which mitigates the necessity of extensive grid searches.

The entirety of the machine learning methods for full-event reconstruction relies on training models with simulations exclusively. Yet, each work highlights the difficulty of addressing variations in observation conditions, which ultimately impact the performance. The current method involves adapting data to match the NSB. Although this approach shows significant improvements [Vui+21], it has limitations due to the assumption that the NSB follows a Poisson distribution. Furthermore, it does not address other types of discrepancies. Leveraging more information from real data could greatly improve gamma-ray source detection. To the best of our knowledge, there is no current work in IACT involving the inclusion of observation images in the training procedure.

The limitations of the standard analysis and the γ-PhysNet enunciated in Section 2.5 induced the need for data and domain adaptation, and information fusion as the strategies to adopt to tackle domain discrepancies. In addition, the panacea incorporates the removal of any run-wise procedure by proposing a global and general model. Therefore, it has led to consider the four following research tracks as the guideline of this thesis:

- If the inputs and discrepancies are fully controlled, the integration of fusion strategies appears to be the most promising approach.
- In the case of strong and unknown shifts between domains, unsupervised domain adaptation is the only applicable technique.
- A combination of both approaches is conceivable to leverage their strength simultaneously.
- High capacity models, like Transformers, offer the possibility to combine generalization, domain adaptation and information fusion through their training strategy and their flexible token representation.

3 Methodological contributions for model generalisation in IACT

The previous chapter has paved the way to novel techniques aiming to improve the detection capabilities of the γ-PhysNet. We have carefully selected the most promising methods among two categories of methodologies to be applied for IACT image analysis: unsupervised domain adaptation and input conditioning. In this chapter, we provide a detailed description of the contributions of this thesis. Firstly, we provide the necessary materials for the implementation of input conditioning with the use of CBN modules. Secondly, we integrate unsupervised domain adaptation, through DANN, DeepCORAL and DeepJDOT models, into the γ-PhysNet architecture. Furthermore, we take into account the strong inherent label shift between simulations and real acquisitions as an extreme limit of current work on importance weighting to the case of IACT. Thirdly, we combine and evaluate multi-task balancing and domain adaptation and instantiate the Gradient Layer as an extension of the Reversal Layer of DANN to any domain adaptation model. Finally, we introduce the Transformer architectures, the γ-PhysNet-Prime and γ-PhysNet-Megatron, as plausible global models for IACT.

3.1 Introduction

The problematics of domain discrepancies

In Chapter 2, we delved into a deep understanding of the recurrent limitations affecting both the standard analysis and the deep learning models. It has principally shed light on a classical issue encountered in many particle physics applications: Simulations for model training are abundant, but differ from real observations to some extent, leading to domain shifts. These discrepancies can be categorized into two distinct groups. On the one hand, known differences between simulations and real data. On the other hand, unknown differences, that cannot be theorized or simulated. Therefore, the extensive overview of domain adaptation and information fusion paved the way to these specific guidelines:

- Addressing known discrepancies: The main discrepancies influencing the reconstruction performance of the models can be named and are well understood. In that case, the inclusion of relevant external information, by redesigning the model accordingly, has an interesting role to play to tackle them. In this work, we propose to involve such additional information as conditioning inputs through the use of CBN modules. However, they are intrinsically limited to these known differences.

- Tackling unknown discrepancies: Nevertheless, unknown differences remain, and there are no precise bounds on how much they contribute to the degradation of performance. Then, unsupervised domain adaptation can be adopted to surpass both known and unknown domain shifts, regardless of their nature.

- Combining both approaches: The key may lie in merging both approaches to comprehensively address all types of domain shifts successfully.

- Global model: Through the use of Transformers, information fusion, domain adaptation and generalization becomes within reach thanks to their powerful and flexible architectures.

In the LST standard analysis based on IACT, a significant drawback can be attributed to the current methodology. Either a specific model is selected depending on the considered observations, or, in the case of GammaLearn, a new training is required for each new observation. To the best of our knowledge, there is no prior work aiming to conduct all-sky analysis within a single framework. Yet, solutions exist, but are difficult to implement and remain costly. For example, it is possible to train a general model and fine-tune it for specific conditions. Conversely, we can develop a global network with very high capacity. The ideal solution involves implementing a model that can be applied universally, without being limited to specific acquisitions. Nevertheless, this approach imposes certain constraints: access to a diverse range of input data covering all possible observation conditions, and a neural network containing enough parameters to incorporate various types of additional information. By focusing on these declarations and constraints, we aim to develop a robust model capable of accurate performance across different scenarios, it becomes relevant to study large capacity models like Transformers.

The specific and simplified case of the LST image analysis

We precisely described in Section 2.5 that, in the case of IACT, known differences are led by NSB variations. In more detail, simulations intrinsically contain a noise approximated by a Poisson distribution of rate λ_{MC} that aims to simulate the real background of the telescope images, but this amount is usually set as a constant rate across the whole dataset. Nevertheless, during acquisitions of real images, external factors like moonlight, starlight, afterglow, and zodiacal light among others cause the amount of light pollution to fluctuate over time. Still approximated by a Poisson distribution, the noise rate becomes a time-dependent function, $\lambda_{real} = \lambda_{real}(t)$. However, training different models across various $\delta\lambda(t) = \lambda_{real}(t) - \lambda_{MC}$ is a very costly process. A typical solution referred to as data adaptation presented in Section 2.5, consists in tuning the dataset to match the real background level by injecting Poisson noise of parameter $\delta\lambda$. This strategy yields ultimately to an increase in performance, as illustrated in [Vui+21]. This optimization is usually performed on each so-called run, a continuous sequence of telescope acquisitions of roughly 20-minute. Yet, this time frame may also contain great variations of light pollution, which is equivalent to test with $\delta\lambda > 0$, leading to degraded outcomes. Such scenario is presented in Figure 5.2 of Chapter 5. This underlines the necessity for more robust approaches.

3.2 Method selection for the application to IACTs

Input conditioning

The integration of supplementary information, as detailed in Section 2.6, is a promising methodology to address the NSB discrepancy. This yields to multimodal CNNs that can incorporate external parameters through early, intermediate, and late fusion techniques. These additional variables can take many different forms, like scalars, images, and others. While these methods effectively integrate extra information, they require substantial trial and error to optimally design the network. Furthermore, fine-tuning a pre-trained model for different scenarios is more challenging with this approach due to the need to re-optimize many parameters depending on the chosen fusion stage.

Conversely, CBN modules offer several advantages over traditional layer-based fusion methods. Firstly, they involve a significantly lesser hyperparametrization, and the optimization process can guide them into BN equivalents if the conditioning is unnecessary. Secondly, CBN modules are particularly advantageous for fine-tuning. This process involves simply replacing the BN layer of the optimized model with CBN modules. Thus, it is theoretically possible to have a global pre-trained model, which is fine-tuned to a specific observation (with different pointings and NSB), thereby substantially reducing computational costs. Consequently, we have selected CBN as the preferred methodology for information fusion in this thesis.

Unsupervised domain adaptation

The core objective of this work is to propose novel approaches to bridge the gap between synthetic and real data, typically through methods known as domain adaptation. As illustrated in the previous Section 2.7 of Chapter 2, domain adaptation is a very wide topic of

research with a great variety of methods. However the context of this thesis takes place in deep unsupervised domain adaptation and de facto reduces the span of the possibly applicable algorithms. The most promising methods in the context of IACT data analysis are summarized hereafter.

In comparison to the current methodology of CTAO data analysis described in Section 2.5 and Section 2.5 of the last chapter, the originality of the deep unsupervised domain adaptation is the inclusion of real observations into the training procedure along with the simulations. Therefore, two opposite strategies emerge from the overview proposed in Section 2.7: learning a domain mapping or learning domain-invariant features. We discuss these directions in the following.

Domain mapping: Limitations of adversarial-generative methods

Adversarial generative methods hinge on the concept that a mapping can serve as a bridge between simulations and real data. This function heavily relies on optimizing GAN-based neural networks. While they offer powerful tools for handling complex domain shifts in unsupervised domain adaptation, they come with several drawbacks that must be carefully managed.

GAN-based models require high-quality and diverse datasets to train effectively. If the source or target domain data is noisy or limited in diversity, they might struggle to learn meaningful mappings. Additionally, various factors influence telescope acquisitions, such as variations in light pollution from the moon, changes in altitude and zenith angles. However, the datasets available, fully described in the Appendix 8.3, are biased towards simulated pointing positions, as mentioned in 2.5 of Chapter 2. Furthermore, label shift, illustrated in Figure 1.2, makes challenging the estimation of a mapping when the target data contain an unequal ratio of particles in the synthetic and real datasets, and precludes cycle-consistency.

Finally, GAN-based training process can be lengthy due to the need for alternating updates between the generator and discriminator networks, along with the fine-tuning required for stability. Yet, the substantial amount of data that will become available after the project's completion necessitates the model to be retrained periodically. In addition, the inference time makes it less suitable for real-time applications because they require a two-step procedure: firstly convert the acquisitions x_{real} to simulations x_{sim} with the GAN-based mapping function $x_{sim} = f(x_{real})$, then reconstruction of the particle parameters y with multi-task architecture $y = g(x_{sim})$.

Domain-invariant features: Advantages of adversarial-discriminative and discrepancy-based methods

Discriminative methods are often simpler to implement compared to generative models. They typically involve adding a domain classifier and modifying the loss function. These methods usually integrate seamlessly with existing neural network architectures and training pipelines, allowing for efficient and stable training without the need for the complex adversarial training loop. Discriminative methods can be more robust to noise in the data compared to generative approaches. They focus on learning domain-invariant features that are useful for the task at hand, which can make them less prone to overfitting noise or artefacts in the data. Moreover, discrepancy measures can be computed efficiently with

stochastic methods, making them suitable for large-scale applications where computational resources are a concern.

We selected three relevant domain adaptation methods that are DANN [Gan+16], DeepJDOT [Dam+18] and DeepCORAL [SS16], to tackle deep unsupervised domain adaptation in the context of multi-task learning. These selected methods fall into the adversarial discriminative and discrepancy-based frameworks. Firstly, DANN is one of the most popular approaches for deep unsupervised domain adaptation and has been applied in a wide variety of contexts. It computes an estimation of the $\mathcal{A}$-distance between domains and has a strong theoretical support [Ben+06]. Secondly, the Wasserstein distance exhibits excellent convergence properties compared to other distance between distributions, as demonstrated in [ACB17], such as Jensen-Shannon or Kullback-Leibler divergence; thus DeepJDOT is a relevant choice to reduce domain discrepancy. Lastly, as the NSB is known to be one of the main contributors to the discrepancies between simulations and real data, minimizing the difference between the first- and second-order statistics seems to be a relevant strategy, making DeepCORAL an interesting choice for our application. In particular, the NSB is usually approximated with a Poisson noise, which rate corresponds either to the mean or to the variance of the distribution. Overall, these three methods constitute a relevant baseline to start building more sophisticated algorithms.

Exploration of self-supervised methods

Self-supervised learning is a subset of unsupervised learning where models are explicitly trained using automatically generated labels. These methods leverage auxiliary tasks, such as auto-encoding and in-painting, to learn useful feature representations from the data itself, without requiring labelled samples from either the source or target domains. When combined with Transformers, they offer a powerful and flexible framework [Ghi+15]. Unlike traditional CNNs, which require careful design to incorporate conditioning inputs (as explained previously in Section 2.6), ViTs use internal token-based representations that easily integrate extra parameters by simply tokenizing and concatenating them with the current representation.

One of the greatest advantages of self-supervised methods is their computational efficiency. Although training on intermediate tasks, which need to be as general as possible, can be resource-intensive, the resulting representations can be reused for fine-tuning on specific observations. Fine-tuning requires significantly fewer resources compared to training models from scratch.

Transformers are high-capacity models that can achieve state-of-the-art performance when trained on large datasets, which is appropriate in the case of CTAO. Firstly, the amount of simulated data is theoretically unlimited, providing access to a vast dataset from various pointing directions. Secondly, unlike other methods that are run-wise — meaning each observation requires distinct training to account for specific discrepancies — ViTs offer the potential to develop a global model. Conversely to fine-tuning, a global model could, in theory, be applied to any observation, eliminating the need for separate training sessions for each new observation and thus significantly simplifying the training process.

Contributions

Based in the selection above, we introduce in this chapter the combination of γ-PhysNet with input conditioning, referred to γ-PhysNet-CBN and the selected domain adaptation techniques. Moreover, we propose original self-supervised methods based on Transformers as plausible candidates for global models. Therefore, the contributions are explicitly the following:

- Input conditioning is applied to the γ-PhysNet to ensure robustness toward NSB variations. This is obtained with the incorporation of the CBN module in replacement with the traditional BN layer.

- Unsupervised domain adaptation is implemented within the γ-PhysNet and evaluated on simulated data in different scenarios. More precisely, an extensive study will be conducted on the γ-PhysNet-DANN. Firstly, the amount of NSB in the target domain is set as a variable drawn from a specific distribution. Secondly, the impact of label shift is measured through its progressive reduction in the target data. We highlight the benefits of the importance weighting in presence of the strong label shift present in our case study.

- Domain adaptation ultimately leads to the inclusion of an additional task to optimize, and we integrate domain adaptation in the multi-task paradigm through a comprehensive study on automatic task balancing. We introduce the Gradient Layer as an extension of the Reversal Gradient Layer in DANN to alleviate the difficult optimization process of DeepCORAL and DeepJDOT.

- We inaugurate the γ-PhysNet-Prime and γ-PhysNet-Megatron, the first Transformers models for the reconstruction of the particle parameters for the CTAO. Both approaches leverage a pre-training following the Masked Auto-Encoder strategy.

3.3 Application of input conditioning for IACT

Contextualisation with IACT and the γ-PhysNet

In the case of IACT, the γ-PhysNet is by design a multimodal approach combining the strength of a pixel charge image simultaneously with a temporal map - the activation time of the photomultipliers - as inputs. Yet, other types of information are available from the observations, like the calibration images, background acquisitions, the pointing directions, amongst others. Such information can theoretically be coherently integrated within the network. As described in the previous chapter, in Section 2.6, there is a plethora of methods for information fusion. In particular, specifically increasing the robustness to some selected quantity of interest can be obtained using CBN modules. For example, it is possible to account for the variations of NSB and make to model resilient to light pollution, which requires to inject the notion of noise as an extra parameter within the network. As another example, the position of the observed source follows the Earth's rotation, and the zenith and altitude angles continuously change as the observation is conducted. Thus, injecting to the pointing direction as an additional input of the model becomes a relevant choice. This pos-

sible improvement is, however, limited to controlled perturbations, and domain adaptation has a complementary role to play when this is not the case.

Conditional Batch Norm

The fusion of information from different modalities is a challenging process, but CBN layers [Vri+17] offer an elegant fusion approach acting on the model feature normalization. In fact, because the NSB and pointing directions directly affects the distribution of the intensity of the images, it is relevant to counter these effects using a distribution-oriented technique.

Batch Normalization

Historically, normalization has been introduced in the pre-GPUs era as a trick to stabilize and accelerate the speed of convergence of former models [Lec+98]. Nowadays, the most wildly used feature normalization approach is Batch Normalization (BN) [IS15], and aims at maintaining stable gradients for the optimization of the model's parameters. Therefore, it increases the speed of reaching the training regimen, and diminishes the influence of parameter initialization, allowing the use of greater learning rates. If the input and the output tensors of the BN algorithm are respectively denoted as $F_{i,c,x,y}$ and $O_{i,c,x,y}$, where i,c,x,y define the indices of the batch ($i \in \{1,...,N\}$), the channel ($c \in \{1,...,C\}$), and the pixel coordinates ($x,y \in \{1,...,X\} \times \{1,...,Y\}$), the normalization procedure can be described as [Bjo+18]:

$$O_{i,c,x,y} = \gamma_c \frac{F_{i,c,x,y} - \mu_c}{\sqrt{\sigma_c + \epsilon}} + \beta_c, \forall i,c,x,y \tag{3.1}$$

where γ_c and β_c are the training parameters. Defining $K = (N \times X \times Y)^{-1}$, the mean $\mu_c = K \sum_{i,x,y} F_{i,c,x,y}$ and variance $\sigma_c = \sqrt{K \sum_{i,x,y} (F_{i,c,x,y} - \mu_c)^2}$ are computed per channel components to standardize the feature distribution $F_{i,c,x,y}$ in the c-axis. Because of possible numerical instability, ϵ is used in the denominator term. The trainable parameters offer the flexibility to cancel the effect of the BN if necessary by setting $\gamma_c = \sqrt{\sigma_c}$ and $\beta_c = \mu_c$.

Conditional Batch Normalization

Originally introduced as a trick to modulate pre-trained ResNet models, CBN aims at replacing BN layers to allow fine-tuning networks with the possibility of adding extra conditioning variables. Furthermore, the strategy can also be applied in an end-to-end context. Classically, it offers the flexibility of computing extra parameters as $\Delta\gamma_c$ and $\Delta\beta_c$, and to add them to the traditional scaling and shifting parameters γ_c and β_c from the BN module, as illustrated in Figure 3.1. The shifts are obtained using the projection of conditioning variables c with an additional encoder G_c into a common latent feature space, denoted as $z = G_c(c)$. Each CBN module, indexed by the integer k, has its own parameters that compute the shifts. In summary, $\Delta\gamma_c^k = f_\gamma^k(z)$ and $\Delta\beta_c^k = f_\beta^k(z)$, where the functions f_γ^k and f_β^k are MLP that computes the shifts of the k^{th} CBN modules from the conditioning latent space. In the following, we omit the index k for simplicity. Overall, the general formulation can easily be extended from the BN equation.

$$O_{i,c,x,y} = (\gamma_c + \Delta\gamma_c) \frac{F_{i,c,x,y} - \mu_c}{\sqrt{\sigma_c + \epsilon}} + (\beta_c + \Delta\beta_c), \forall i,c,x,y \tag{3.2}$$

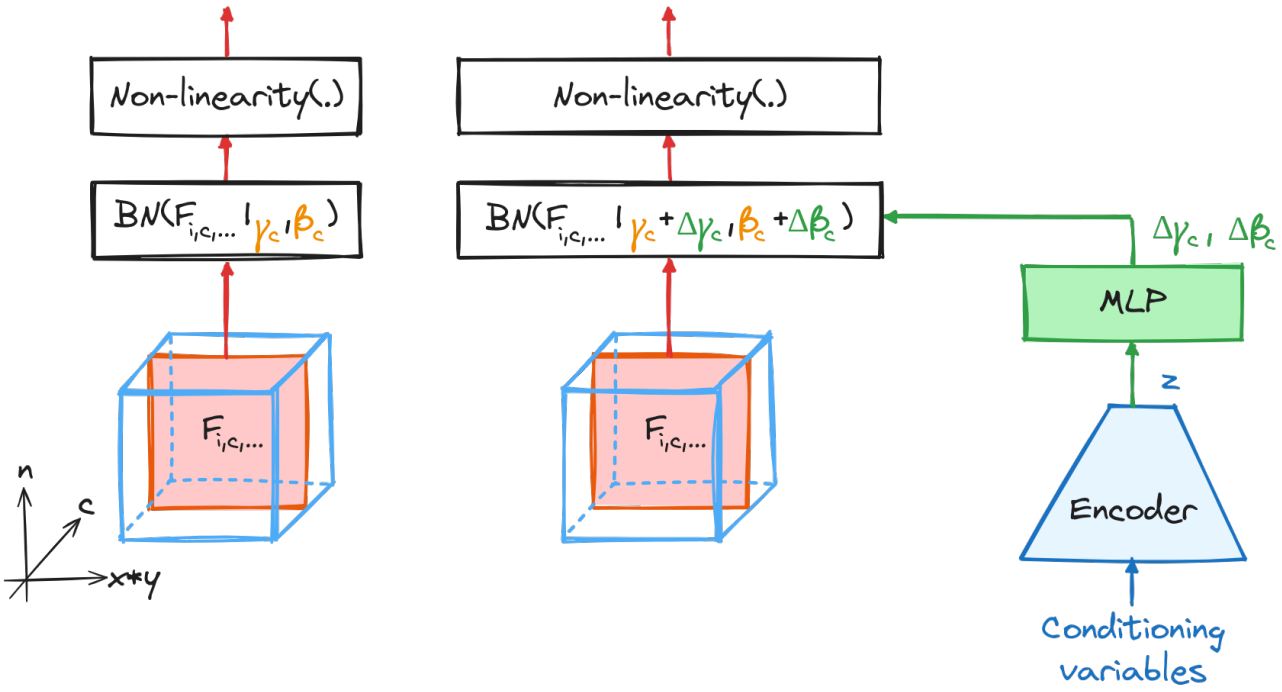


Figure 3.1: The BN mechanism (left) and the CBN mechanism (right). Illustration inspired from [Vri+17].

We refer to the γ-PhysNet-CBN when CBN modules are integrated into the γ-PhysNet. External judiciously chosen parameters are injected in the network through an encoder that produces a latent representation of the conditioning variables. These features are then flowing in the different CBN modules from which the shifts are computed using specific dense layers. The evaluation of the γ-PhysNet-CBN on the digit datasets is available in Appendix 8.5.

Scalar conditioning variables

We introduced CBN modules in Section 2.6 as a way to provide robustness towards NSB. In its simplest form, the simulated background injected in the synthetic dataset corresponds to a Poisson noise of rate λ_{MC} and is constant across all pixels. As a reminder, it is based on this idea that data adaptation, as described previously in Section 2.5, aims to modify the training dataset to match the real data background variations. In that case, this specific distribution can thus be characterized by a single constant scalar, its rate. However, the utilization of CBN modules becomes limited because the shift is constant, and it is possible to train the model on the noisy data directly. Then, taking into account the fact that NSB is subjected to change along runs, we propose to consider more realistic perturbation by sampling the rate λ_{MC} from other distributions. We consider the case where the rate is sampled from a Uniform distrubtion, so that $\lambda_{MC} \sim \mathcal{U}(a, b)$ where a and b are the lower and upper bounds. Therefore, it is possible to consider this sampled rate as the conditioning input. This scenario typically corresponds to data augmentation.

Yet, the NSB is not the only contributing factor of discrepancy, and the pointing angles can play a significant role, as mentioned in the current deep learning limitations in Section 2.5. In fact, any scalar feature could be used as input as long as they are simulated. Thus, the addition of the zenith and azimuth angles are a promising to ensure robustness of the

model against observation pointing variations. It translates by utilizing synthetic images from different simulated pointings, and to inject their corresponding angles as conditioning inputs. This extension can be considered as a short term objective of this work.

Input conditioning, in the form of multiple scalars, can be easily integrated by concatenating them into an input vector. In that case, a linear encoder, in the configuration of an MLP (or DNN), has enough capacity to create an intermediate projection of the variables. An illustration of the γ-PhysNet-CBN with scalar input conditioning is given in Figure 3.2.

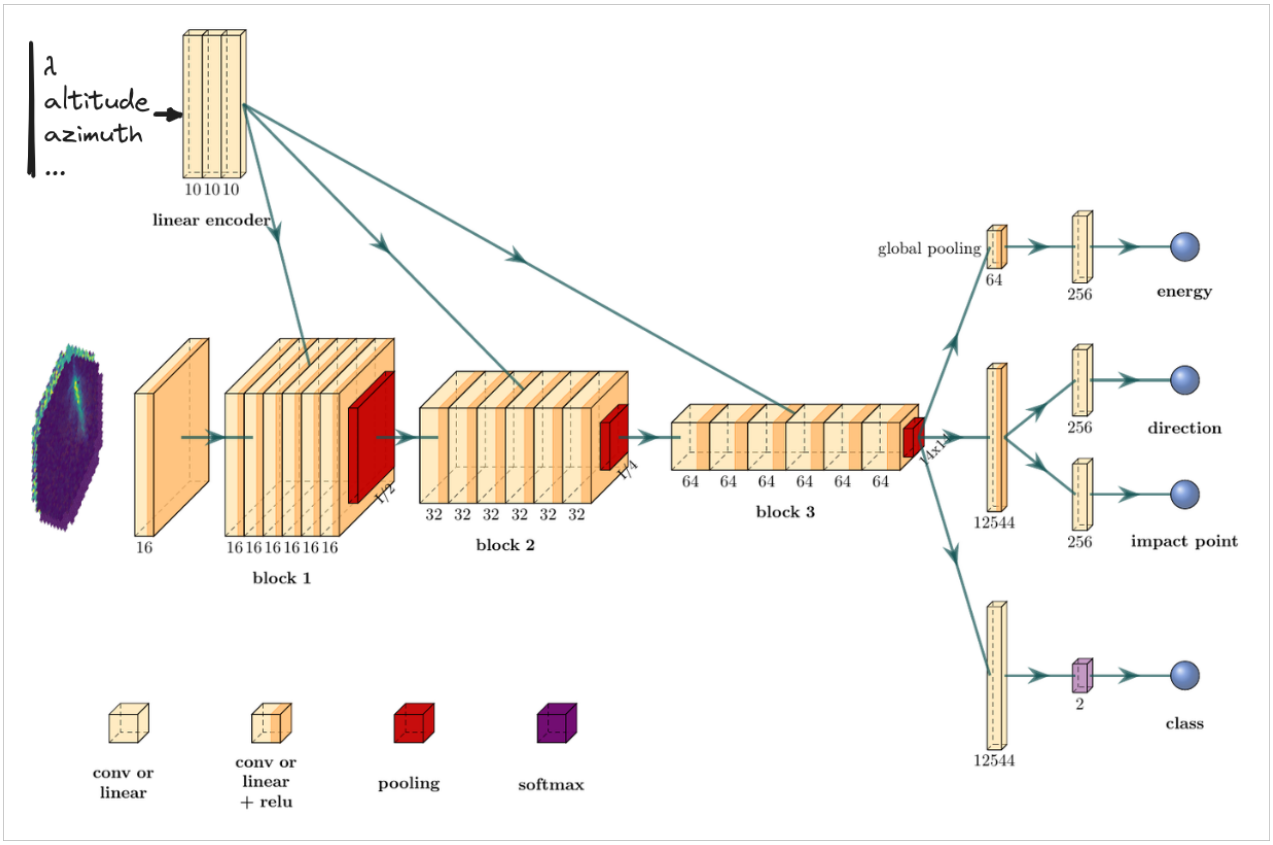


Figure 3.2: The γ-PhysNet-CBN architecture with scalar conditioning.

Image conditioning variables

Although the Poisson hypothesis, combined with pixel-independence, simplifies the parametrization of the NSB, it does not accurately reflect reality in most of the scenarios. More information is thus necessary, in particular when processing real data.

Fortunately, it is possible to use pedestal images to characterize the actual noise. Pedestals are background noise acquisitions sampled at 100 Hz. Captured simultaneously with the source observations, they provide valuable information to estimate the true distribution of the NSB. These images will be carefully studied in the final chapter of this thesis. In summary, they can reveal, among others, the presence of stars in the fields of view, corresponding to groups of correlated pixels with higher intensities compared to the background.

In order to utilize the pedestal information, it is important to understand how data is stored in the LST data storage framework. In fact, events are stored in files, also called sub-run, containing a fix number of $53k$ events each. From a single file, roughly 600 to 700

pedestal images can be obtained, although this amount depends on the telescope trigger rate, and some events, like car flashes, may disrupt this consistency. Nonetheless, a sub-run corresponds to a few seconds of observations, a period too short to contain drastic variations in observation conditions, yet sufficient to include a decent number of events for statistical analysis.

Therefore, our strategy consists in utilizing image conditioning as the mean pedestal image, averaged over a sub-run, as input. Such an image is referred to as a λ-map. In this context, a linear encoder must be replaced by a CNN to exploit pixel correlations, such as those caused by stars, as depicted in Figure 3.3. The methodology is slightly more complex than the scalar input defined earlier, and it can be implemented as follows:

- The first step corresponds to the selection of noise-free synthetic data, without the application of the Poisson noise that simulates the background. They correspond to a pure signal of the Cherenkov light.
- Then, select pedestal images from the observation at stake.
- Thirdly, a continuous λ-map function can be created by averaging sub-runs over time.
- Fourthly, the training images can be created by adding signals to pedestal images.
- Finally, create the image conditioning by interpolating or selecting the closest (temporally) λ-map to the pedestal image.

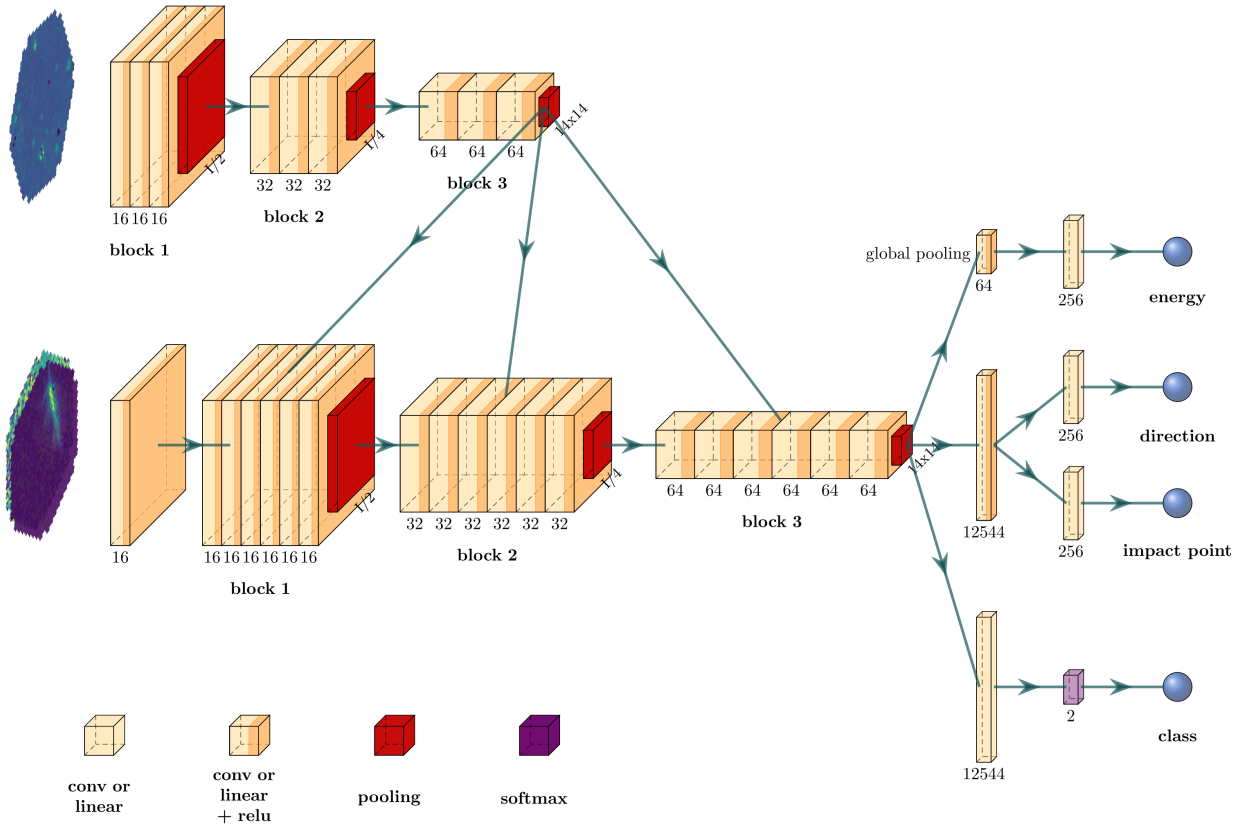


Figure 3.3: The γ-PhysNet-CBN architecture with image conditioning.

Although this methodology can be considered as a short-term objective of this work, it suffers from some drawbacks. Because the λ-maps are created from averaging the pixel values of the pedestals within an entire sub-run, the assumption of a Poisson distribution is still considered. In addition, although this methodology extends the scalar input conditioning by considering more information within the λ-map, it is tailored to specific observations (run-wise), and are not generalizable to multiple telescope acquisitions. Finally, robustness can solely be ensured for known variables. In that case, unsupervised domain adaptation naturally addresses the latter two issues.

3.4 Application of unsupervised domain adaptation for IACT

The γ-PhysNet extended with unsupervised domain adaptation

Selected in Section 3.2 of the last chapter, DANN, DeepCORAL and DeepJDOT appear to be the most relevant approaches to integration into the γ-PhysNet neural network. Regarding DeepCORAL and DeepJDOT, there is no model architecture modification to proceed. The inclusion solely consists in adding a new component in the global objective function. Concerning DANN, there is contrariwise some modifications to implement. As illustrated in Figure 3.4, the adjustment consists in integrating a domain classifier in parallel to the other reconstruction task.

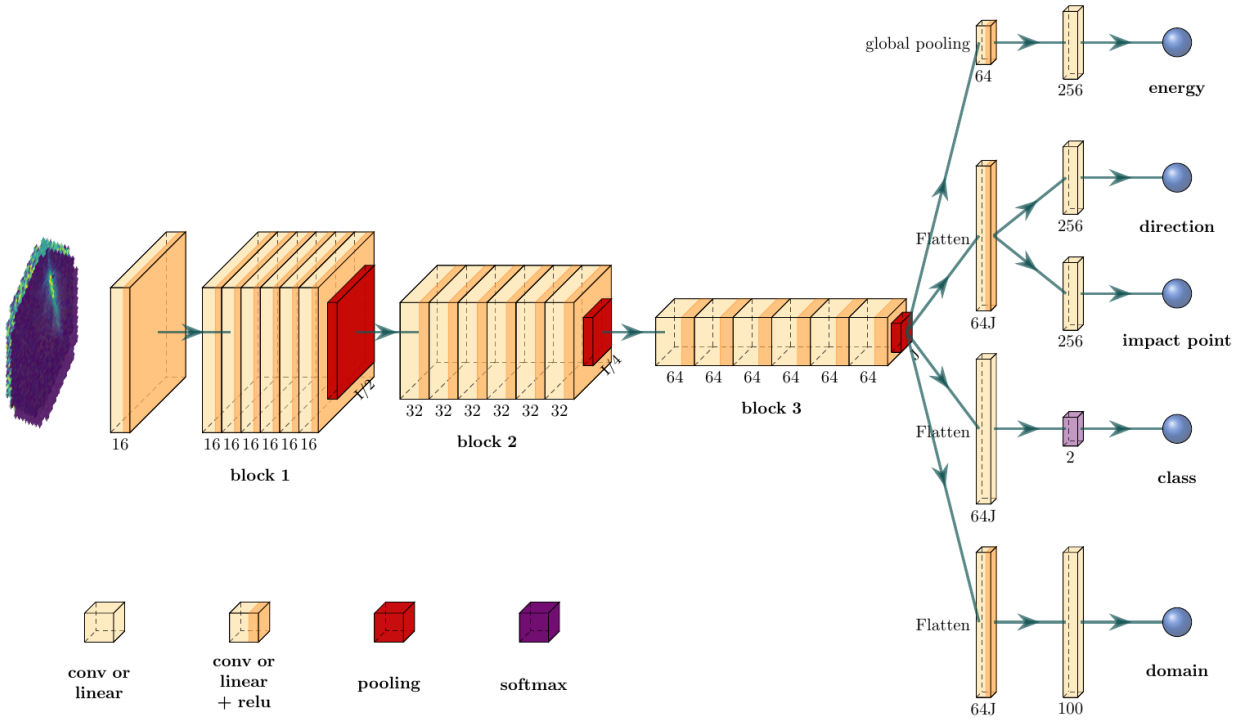


Figure 3.4: The γ-PhysNet-DANN architecture.

Conditional domain adaptation

Importance weighting

One of the most challenging issues to transition from simulations to real data while applying unsupervised domain adaptation is the label shift, that is to say the difference of the ratio between the gamma and proton classes captured by the telescopes. Although they are equally represented in the training simulations, real data contains less than 1 gamma for 10^4 protons. As an example, for mini-batches of a typical size of $\sim 10^2$ samples, there is a very low probability that they contain a single gamma event. Thus, it becomes crucial to readapt the algorithms to take into account such label shift. As mentioned in Section 2.7, state-of-the-art approaches rely on importance weighting. The importance weight ω for a sample with label y is defined as the ratio of the target and source label distribution, respectively P^T and P^S. As described in Equation 2.23, it is classically be computed as [LWS18]:

$$\omega(y) = \frac{P^T(Y = y)}{P^S(Y = y)} \tag{3.3}$$

These weights adjust for the difference in label distributions between the both domains. The source label distribution can be directly estimated from the labelled source mini-batch:

$$\widehat{P^S}(Y = y) = \frac{1}{n_S} \sum_{i=1}^{n_S} \mathbb{1}(y_i = y) \tag{3.4}$$

where n_S is the number of labelled source samples and $\mathbb{1}$ is the indicator function. However, since we do not have labelled target data, P^T must be estimated differently.

Methodology

Following the work of [Liu+21a] on importance weighting, we introduce the conditional versions of DANN (CDANN), DeepJDOT (CDeepJDOT) and DeepCORAL (CDeepCORAL) to tackle the strong label shift encountered with the LST dataset. As the ratio of gammas and protons is approximately estimated from observations, as illustrated in Figure 1.2, it is possible to evaluate the value of the class-wise balancing parameter $\omega \in \mathbb{R}^2$. In fact, if $\omega = \begin{pmatrix} \omega(\gamma) \\ \\ \omega(p) \end{pmatrix}$, the weighting coefficient can be determined as:

$$\begin{aligned} \omega(\gamma) &= \frac{P^T(\gamma)}{P^S(\gamma)} = \frac{n_T^\gamma}{n_S^\gamma} = \frac{\epsilon}{n_S^\gamma} \approx 0 \\ \omega(p) &= \frac{P^T(p)}{P^S(p)} = \frac{n_T^p}{n_S^p} = \frac{n_S^p + n_S^\gamma - \epsilon}{n_S^p} > 1 \end{aligned} \tag{3.5}$$

where $\epsilon = n_T^\gamma$ refers to the residual amount of gammas in the target set, and the composition of the source or target mini-batch respects the following equation:

$$n_S = n_S^\gamma + n_S^p = n_T^\gamma + n_T^p = n_T \tag{3.6}$$

From a training point of view, this is equivalent to minimize the discrepancy between the simulated protons and the whole target dataset. In the following, we denote $\mathcal{D}_S = \{p_S^{(i)}, \gamma_S^{(j)}\}_{i,j}$

the source data containing the source protons $\{p_S^{(i)}\}_i$ and the source gammas $\{\gamma_S^{(j)}\}_j$. Similarly, the target data is defined as $\mathcal{D}_T = \{p_T^{(i)}, \gamma_T^{(j)}\}_{i,j} = \{x_T^{(k)}\}_k$, as they are undifferentiated by nature in the training target dataset. Then, we introduce conditioning strategies on the following optimization methods.

Application to DANN

Specifically, for DANN, the implementation of our tailored importance weighting can be achieved through the rewriting of the DANN loss function established in Equation 2.28. The domain classifier loss is then masked based on the source particle labels, and the loss function becomes:

$$\mathcal{L}_{DANN} = \sum_{i|x_i \in \{p_S, x_T\}} CE\left(D(\mathcal{R}[G(x_i)]), d_i\right) \tag{3.7}$$

Application to DeepJDOT

In the case of DeepJDOT, the conditioning is considered at the computation of the cost matrix level. It is computed only between source protons and uniformly sampled target mini-batches to maintain square arrays. The optimal transport plan is then deduced afterwards from these selected samples, following the two-step procedure proposed in [Dam+18]. If G defines the feature extractor of the neural network, then the cost function is described as:

$$C_{i,j} = \|G(x_i) - G(x_j)\|^2, x_i \in \{p_S^{(i)}\}_i, x_j \in \{x_T^{(j)}\}_{j==i} \tag{3.8}$$

Finally, the loss function can be calculated with the standard optimal transport defined at Equation 2.26.

Application to DeepCORAL

Finally, DeepCORAL computes statistics using source protons and the entire target mini-batch. The source covariance matrix is then computed as:

$$C_S = \frac{1}{m-1}\left(D_S^T D_S - \frac{1}{m}(1^T D_S)^T (1^T D_S)\right) \tag{3.9}$$

where $D_S = \{p_S^{(i)}\}_{i=1}^m$ refers to the mini-batch containing m source protons. The calculation of the target covariance matrix is left unchanged. Then, the loss function can be computed with Equation 2.25. In practise, m is mini-batch-dependent for both CDeepCORAL and CDeepJDOT, as the number of source protons changes at each iteration.

Modification of the target set

Compensating for the label shift can also be effectuated by integrating simulated gammas into the target dataset. Despite the fact that it ultimately introduces bias, this could be considered as promising because gamma simulations accurately represent observed data, and thus the discrepancy with real acquisitions mostly relies on the NSB. However, the gamma/proton ratio in a gamma-ray emitter observation is never known in advance, as it is source-dependent. Finally, it is also not possible to remove gammas from the source

domain to equalize the amount of the target, as they are used in the optimization process of the physical attribute reconstruction branches of the neural network. As a result, those two approaches are not considered in the following.

3.5 Association of multi-task balancing with domain adaptation

A difficult optimization process

Including domain adaptation into the multi-task balancing framework aims at integrating the associated loss coefficients within the automatic determination of the weights of each task. This is nevertheless an ambitious goal because integrating multiple tasks increase the risks of optimization conflicts. A preliminary study is conducted to evaluate the compatibility of the propagated gradients of each task at the last shared layer of the backbone, including domain adaptation and following the architecture presented in Figure 3.4. The analysis of the amplitude and cosine similarity of the gradients at the last shared layer highlights conflicting gradients between the domain adaptation task and the reconstruction of the physical attributes, as depicted in Figure 3.6, turning the optimization into a complex process. As illustrated in Figure 3.5, the difficulty of optimizing different tasks resides in the fact that their associated gradients don't share the same scale (the green dots), and/or can point to different directions in the parameters space (the red arrows). In that case, co-linearity is measured using the cosine similarity metric. Additionally, as defined in [KGC18], their respective task uncertainty (the purple σ) can also be used as a measure of task difficulty. The nature of the standard deviation σ is given in Section 2.8. In summary, strong gradient amplitudes from the domain task can be detrimental for the performance on the source dataset when tasks are in opposition.

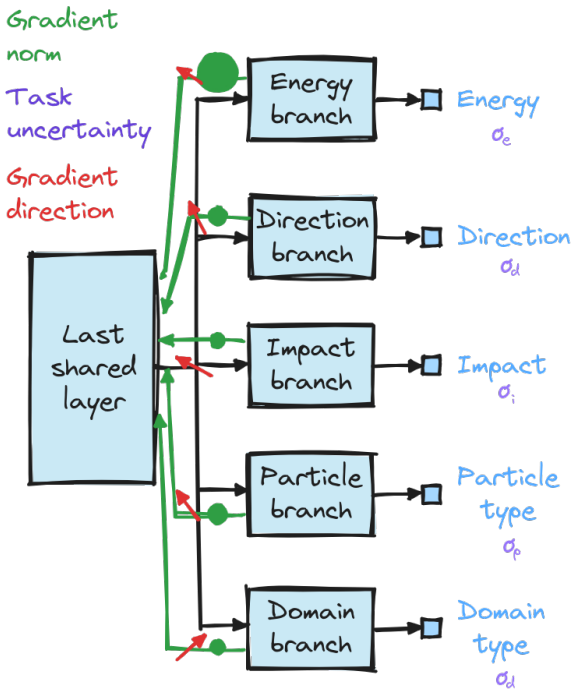


Figure 3.5: Illustration of multi-task learning.

The Gradient Layer as a task-specific learning rate scheduler

The domain task may take the lead of the training process, yet it is crucial to ensure the best performance on the source data. A possibility resides in training a first network specifically on the source data, and to fine-tune it using both domains. Inspired by DANN, a similar end-to-end strategy consists in progressively including the contribution of domain adaptation to the training with the use of a gradient weighting procedure. It can be extended to DeepJDOT

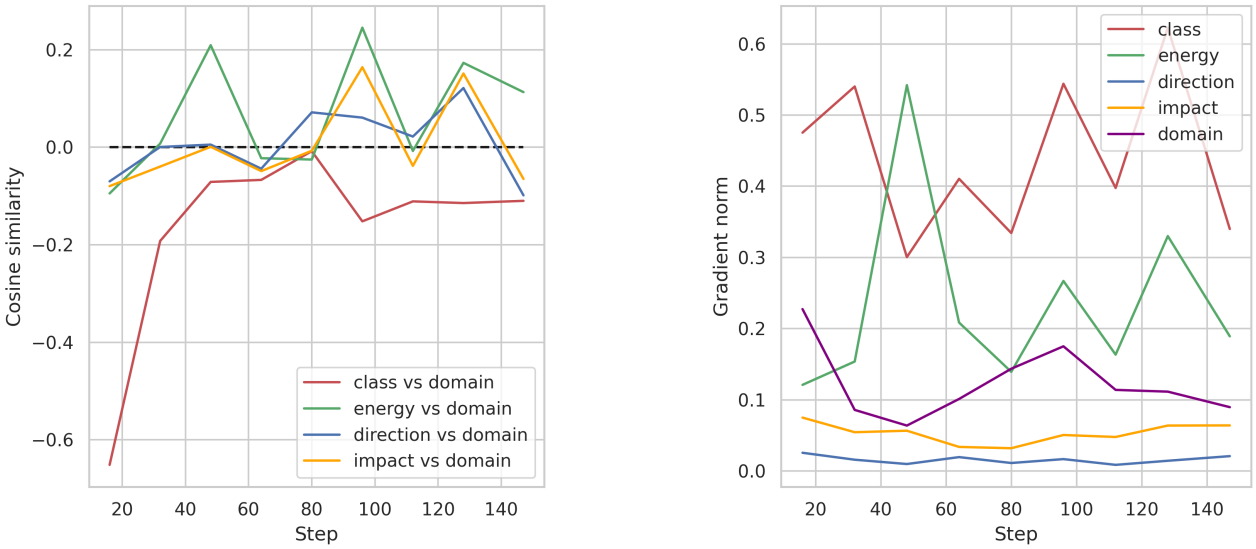


Figure 3.6: The cosine similarity between the domain task and the full-event reconstruction tasks is shown on the left panel. The gradient norms of each task are plotted on the right panel. In both cases, the domain adaptation method is DANN. Conflicting tasks, for example *class* versus *domain* (depicted in red) exhibit a negative similarity which, associated with strong gradients, may affect the quality of convergence of the model.

and DeepCORAL through the definition of a gradient layer, which aims at weighting the domain adaptation gradient during the backward pass. To this end, we consider a Gradient Layer (GL) $\mathcal{G}$ that performs such an operation, and can be reversed in the case of DANN.

$$\mathcal{G}(x) = x \quad \text{and} \quad \frac{d\mathcal{G}}{dx} = (-1)^R f(i) \times I \tag{3.10}$$

where $R = 1$ for reversal else $R = 0$, f is a function of the current iteration i defined by the user, and I is the identity matrix. Commonly, f is chosen as the function that progressively evolves from 0 to 1, for example:

$$f(i) = \frac{2}{1 + e^{-\gamma p(i)}} - 1 \in [0,1] \quad \text{and} \quad p(i) = \frac{i}{\max \text{ epoch}} \in [0,1] \tag{3.11}$$

3.6 Application of Transformers for IACT

The previous contributions aim at upgrading the current γ-PhysNet with the use of CBN modules or the integration of unsupervised domain adaptation in order to reduce the discrepancies between simulations and real acquisitions. Yet, these approaches solely rely on CNN-based architectures. As mentioned in our state-of-the-art Section 2.4, new methods, built around Transformer models, have become the best performing approaches in many applications.

As a concurrent approach for the full-event reconstruction of incident particle physics parameters, we propose a Transformer version of the γ-PhysNet, referred to as γ-PhysNet-Prime. Implementing such architectures come with constraints. Firstly, Transformers need a lot of training data to overcome the inductive bias issue, as discussed in Section 2.4. Secondly, they are usually very high capacity models, with a lot of parameters.

Regarding the inductive bias problem, a lot of simulated data are accessible in the CTAO program. In addition, this bias arises because the attention modules works with token projected from large patches, and loses the notion of locality. Thus, it is possible to mitigate the issue by considering smaller patch size.

Moreover, in this work, as a proof of concept, we will consider small-sized models trained one a single pointing direction. This is feasible in practise because the size of the input images is quite small compared to standard Transformer applications (224×224 in the literature and 1855 pixels in our case).

Finally, pre-training with the Masked Auto-Encoder strategy and incorporating both simulations and real acquisitions in the training process might help generalize and obtain good performance on real data.

Construction of the architecture

Auto-Encoder

Wildly used in signal and image processing, auto-encoding finds its root in data reduction. Before the era of deep learning, techniques relied on mathematical tools, like Singular Value Decomposition (SVD) [KL80], to discard noise from data of interest. In particular, SVD is used to decompose a matrix A into three components $A = U\Sigma V^T$. The singular values in Σ represent the magnitude of each corresponding dimension, and they are typically ordered from largest to smallest. It provides a framework to approximate a matrix with fewer dimensions by keeping only the largest singular values, corresponding to the most meaningful features. Therefore, it corresponds to the simplest form of a (linear) Auto-Encoder (AE). Noise can be rejected by ignoring the components corresponding to the smallest eigenvalues, often determined using the elbow method. However, a linear decomposition like SVD fails to capture complex relationships between variables.

Nowadays, AE are designed with neural networks aiming to learn an identity mapping for the training data while compressing the feature space into a generally smaller representation. The primary objective is to reproduce the inputs at the output layer. It consists of two main components: an encoder and a decoder. The encoder f_θ maps the input x to a hidden representation z, while the decoder g_θ reconstructs an estimate $\hat{x}$ the original input from this hidden representation, effectively reversing the transformation performed by the encoder.

$$\begin{aligned} z &= f_\theta(x) \\ \hat{x} &= g_\theta(z) \end{aligned} \tag{3.12}$$

The last hidden representation z, known as the latent space or bottleneck, captures the most essential features of the data.

Masked Auto-Encoder

The success of SVD decomposition, AE or other data reduction algorithms in Computer Vision highlights the fact that images contain a lot of redundancy. Based on this idea, Masked Auto-Encoder (MAE) [He+21] aims at randomly discarding 75% of the encoded tokens, and to reconstruct them as an in-painting objective. In fact, the power of MAE can be attributed

to its ability to reconstruct missing parts, which helps the model generalize better to unseen data. By learning features that are not specific to the training set but applicable to a wide range of scenarios, the model captures the underlying structure of the signal and background.

Specifically, the MAE architecture is composed of two entities: an encoder and a decoder. Input data is firstly embedded using a linear projection on which are added positional encodings. Then, a fixed amount of patches are discarded through a masking procedure. The encoder is applied exclusively to unmasked patches. This strategy allows training very large encoders while utilizing only a fraction of the compute and memory resources and significantly improves the efficiency, generability and scalability of the model.

In the other hand, the decoder aims to retrieve the missing parts of the puzzle. This can be achieved by re-introducing all tokens as inputs, including the masked ones. A mask token indicates the presence of the tokens that must be predicted.

Application of MAE to LST images

Interpolated images

In the pioneer paper introducing the ViTs [Dos+20], the input images I and patches P are respectively of size $I \in \mathbb{R}^{244\times244\times3}$ and $P \in \mathbb{R}^{16\times16\times3}$, leading to 196 non-overlapping patches. However, in the context of LST simulations and real data, acquisitions are arranged on a hexagonal grid. Using Transformers directly requires interpolating them onto a regular grid, a process implemented in the γ-PhysNet for current analyses, as detailed in [Jac20]. Typically, bilinear interpolation is used to generate an input of size $I \in \mathbb{R}^{55\times55\times2}$, where the channels represent pixel charge and time information. Nevertheless, it is preferable to maintain a smaller size to save computation time, rather than resizing to the same dimensions as in the ViT paper. Both channels are preserved for the pre-training. In this case, we select a patch size of size $P \in \mathbb{R}^{5\times5\times2}$, as illustrated in Figure 3.7, leading to a total of 121 non-overlapping patches. This choice of patch size is arbitrary, but it balances the need for computational efficiency with the granularity required for effective analysis. The size patches corresponding to black areas introduced by the interpolation can also be discarded as they convey no information.

From the patchification and projection steps, a token representation of size 256×265 is produced, corresponding to 265 tokens sharing an embedding size of 256.

Hexagonal images

The patchification process. Although interpolation is beneficial for the γ-PhysNet, as it allows faster training without degradation in performance, working on the hexagonal grid become an interesting choice while working with Transformers. In fact, the structure of images is directly related to the underlying electronic system supporting the PMTs. Practically, groups of 7 pixels share the same hardware module. There are 265 modules of 7 pixels in total forming the 1855 camera pixels. We propose to use this organization to patchify the input data, and the process is illustrated in Figure 3.8. From a system perspective, pixels of the same lot are consecutively stored and loaded in memory. The input projection and the patchification process can then simultaneously be easily implemented with a convolution layer of stride 7. Similarly to the interpolated case, a token represnetation of size 256×265

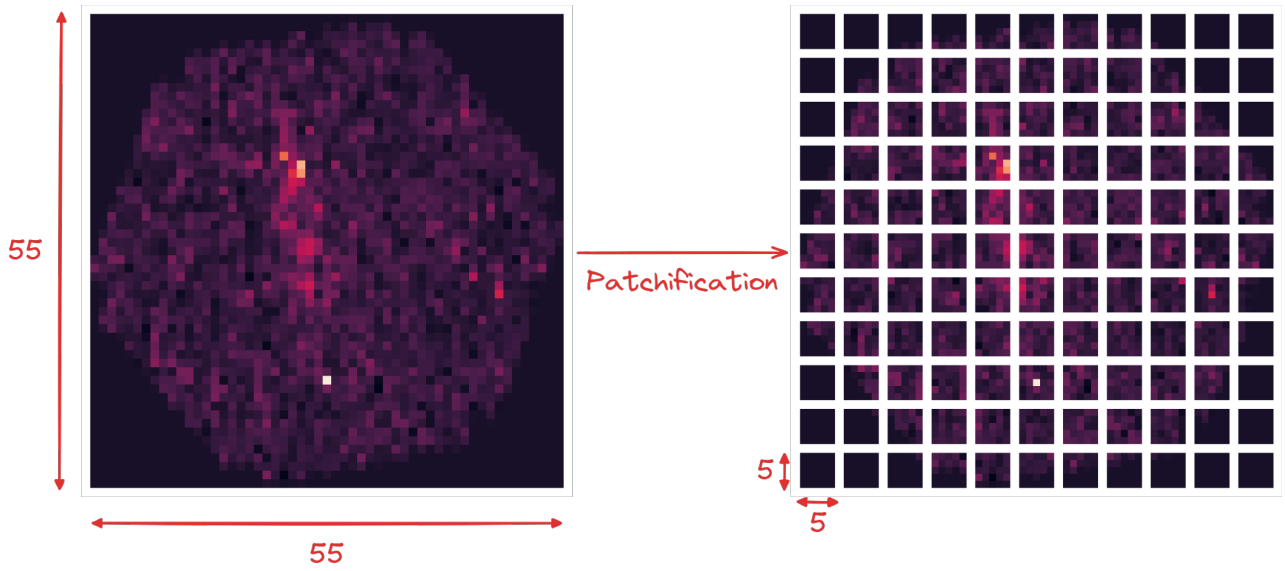


Figure 3.7: The patchification process after bilinear interpolation on a regular grid.

is produced. Among the 265 tokens, 75% is discarded using a random mask, resulting in 66 tokens used for encoding.

Positional encoding of the modules. Additionally, an important step during patch projection is positional encoding. This technique is used to provide spatial information about the token positions within the initial image. In fact, Transformers process inputs simultaneously rather than sequentially, unlike RNN, through self-attention mechanisms. Although this allows the model to weigh the importance of different tokens relative to each other, it is inherently position-agnostic. By injecting information about the position of each token directly into the embeddings, the spatial coherency can be retrieved.

Positional encoding typically adds a positional signal to the input embeddings. This can be done in several ways, with one common method being the use of sine and cosine functions of different frequencies. The idea is to generate unique encodings for each position that the model can learn to interpret [Dos+20]. For a patch located at position (p_x, p_y) in the 2D grid:

$$PE_{(p_x,p_y,2i)} = \sin\left(\frac{p_x}{10000^{\frac{2i}{d}}}\right) \tag{3.13}$$

$$PE_{(p_x,p_y,2i+1)} = \cos\left(\frac{p_x}{10000^{\frac{2i}{d}}}\right) \tag{3.14}$$

Similarly, for the y-coordinate:

$$PE_{(p_x,p_y,2j)} = \sin\left(\frac{p_y}{10000^{\frac{2j}{d}}}\right) \tag{3.15}$$

$$PE_{(p_x,p_y,2j+1)} = \cos\left(\frac{p_y}{10000^{\frac{2j}{d}}}\right) \tag{3.16}$$

Here, i and j denote different dimensions within the embedding vector, and d represents the embedding size. These encodings are then added to form the final positional encoding for each patch. In the case of LST acquisitions, we are using the module centroid for the

values of p_x and p_y to compute the position embedding. Centroids are computed as the mean of the x and y coordinates of each module.

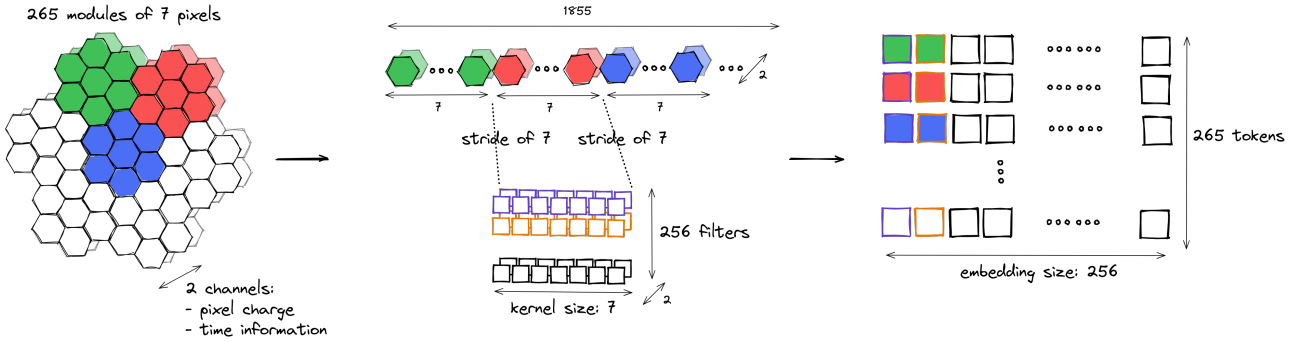


Figure 3.8: Tokenization of the hexagonal LST1 images.

Inclusion of external variables

One strength of Transformers resides in the tokenisation process. While traditional CNN models must incorporate an information fusion strategy, as described in Section 2.6, Transformers leverage a simple projection of the input through a trainable MLP and a concatenation to the patch representation. However, as explained previously in Section 2.4, the benefits of Transformers results in the flexibility of combining different kind of information within the embedded token representation, that are compared within attention modules. In order to inject this information, these external inputs must have their own positional encoding, although they don't belong to the initial image. A common method to achieve this involves assigning them a positional vector that ensures a sufficiently large distance from the actual image tokens. For M external input, their corresponding positional encoding is expressed a vector $v \in \mathbb{R}^{1\times d}$ filled with a constant value $v_i = m$, where m represents the m^{th} external input.

Physics-based pre-training strategy

Self-supervised approaches aims to foster robust and transferable feature representations. Through the optimization of an auxiliary task, Transformer models can be pre-trained to find general features from the input images. In the case of MAE, the masking procedure forces the weights to not rely on specific attributes and to specialize on them. Although this idea allows to improve the performance on natural images, this general strategy misses physics content that can be attribute to IACT. In fact, it could be judicious to help the model learn some physical aspect of the system, through the pre-training. For example, a modality-translation, that would help the model learn to infer the temporal map from the associated pixel charges, could be beneficial for at least two reasons:

- The temporal evolution of the signal in each pixel is closely related to the underlying physical processes, such as the development of particle showers and the propagation of Cherenkov light. By learning to reconstruct the temporal map, the model implicitly learns the physical dynamics of the system.

- Reconstructing the temporal map from the pixel charge encourages the model to learn meaningful representations that capture the relationship between spatial and temporal aspects of the data. This can improve its ability to extract relevant features for downstream tasks such as energy reconstruction, particle identification, and direction estimation.

In this work, we mainly focus on MAE pre-training, and consider this approach as a perspective for future work.

The γ-PhysNet-Prime

Once the model is pre-trained with a MAE, it can finally be fine-tuned to the reconstruction of the physics parameters. For this purpose, the encoder is preserved to allow projecting the images into the new representation, while the decoder is no longer useful and discarded. As it is classically performed in the Transformers paradigm, additional tokens are concatenated to the embedded patches, corresponding to the task tokens. Their particularity is that they are set as trainable parameters. Furthermore, the patch masking procedure inherited from the MAE pre-training becomes obsolete and the entire representation inputs the encoder. The process is illustrated in Figure 3.9.

Once the embedded patches have been encoded, the resulting latent space still contains three types of tokens. The one created from the image patchification, the external variables, and the task tokens. Only the latter are kept. From the task latent space are connected four branches, each of them performing its assigned reconstruction, and are made of a simple MLP.

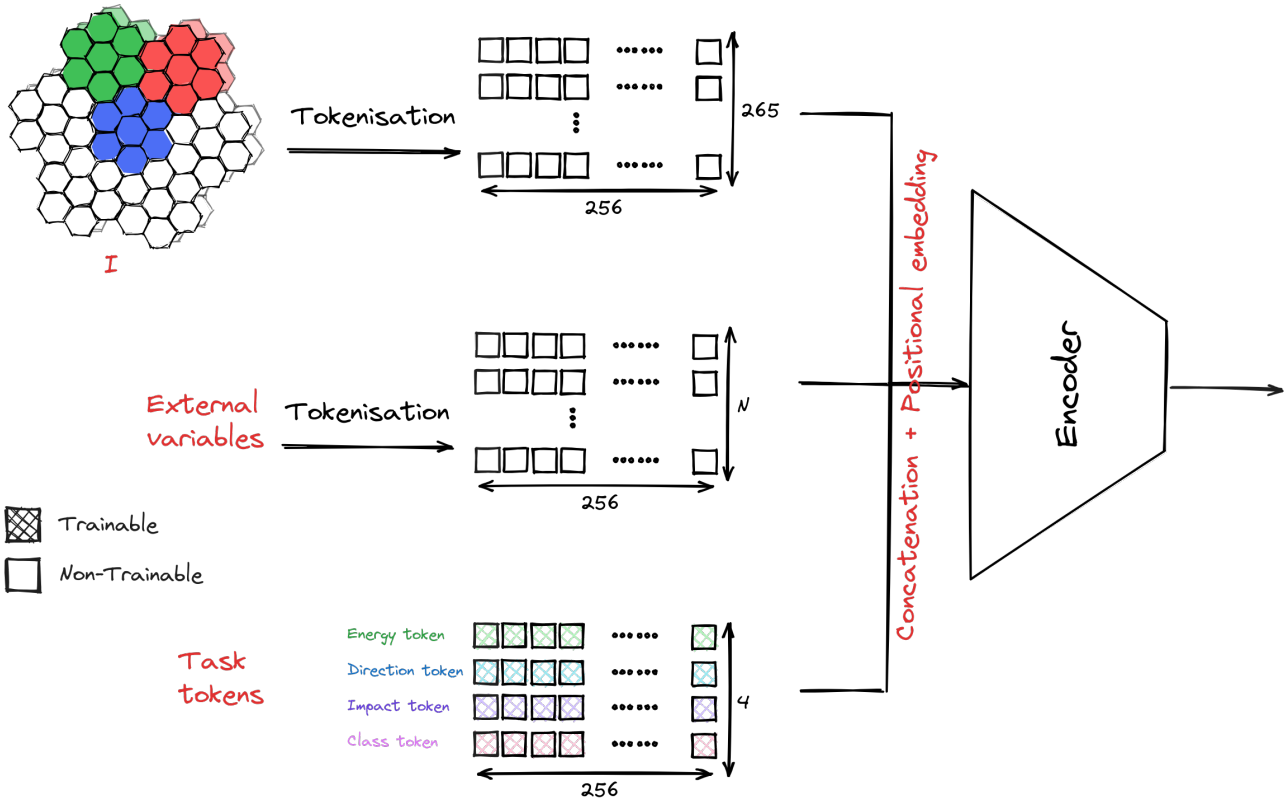


Figure 3.9: Illustration of the encoding process of the γ-PhysNet-Prime.

The γ-PhysNet-Megatron

Although training a MAE can help generalization considering reconstructing inputs from both simulations and real acquisitions, the fine-tuning process is still exclusively performed on labelled synthetic data. Thus, the domain shift is still present within the γ-PhysNet-Prime. As a result, we propose to integrate the DANN strategy into our Transformer model, referred hereafter to as the γ-PhysNet-Megatron.

As discussed in the state-of-the-art chapter, particularly in Section 2.4, there exists many approaches considering domain adaptation with Transformer models. Based on our current knowledge with the standard CNN-based γ-PhysNet, we propose to tackle label shift with DANN as it has provided the best results currently. It thus becomes our baseline from which more sophisticated methods can be further implemented. We did not see any similar approach in the literature at the time our work has been produced.

The implementation of the γ-PhysNet-Megatron is very similar to the γ-PhysNet-Prime. The main difference is the addition of the domain class token in the task representation. The inclusion of domain adaptation is performed comparably to the DANN architecture, with the incorporation of a Gradient Reversal Layer to inverse the sign of the domain gradients.

3.7 Conclusion

In consideration to the limitations of the current standard analysis and deep learning approaches for the detection of gamma-ray source, we have presented in this chapter novel contributions to tackle domain shifts between LST simulations and real observations.

On the one hand, the knowledge of the most impacting factors on domain shifts led us to consider information fusion, and more precisely CBN modules, to extend the work of [Jac20] and to propose the γ-PhysNet-CBN. On the other hand, unsupervised domain adaptation, with the selection of DANN, DeepCORAL and DeepJDOT, aims to reduce both known and unknown domain discrepancies. These two concepts are complementary and explicitly focus on the architecture of the neural network to target the issues at stake.

Vision Transformers are concurrent approaches, more complex but potentially more generalizable and abstract. They aim to replace the γ-PhysNet with the introduction of another architecture based on Transformer modules. This new network can, furthermore, be extended to domain adaptation with the inclusion of domain tasks.

The next two chapters will qualify these propositions, firstly in a controlled environment of simulations (Chapter 4) and on the detection of real data, the Crab Nebula (Chapter 5).

4 Application of input conditioning, domain adaptation and Transformers to the case of simulations

The previous chapters lays the groundwork for understanding machine learning tools essential for the full-event reconstruction of incident particle events within the IACT framework, and has demonstrated the advantages of deep learning over traditional analysis methods. Building on these findings, we have introduced the first contributions of this thesis: the implementation of unsupervised domain adaptation, input conditioning and Vision Transformers into the current deep learning framework for IACT image analysis. In this chapter, we assess their benefits within a controlled environment provided by simulations. The training dataset is succinctly described and offer insights into commonly used figures of merit in the field. Additionally, we address the challenges of integrating domain adaptation into multi-task balancing. This chapter concludes by setting the stage for selecting the methods to be applied to real-world data in the last chapter.

4.1 Introduction

In many physics application, obtaining error-free ground truth labelled data is generally impossible because many indescribable random processes occur during the acquisitions, like weather condition, electronic noise, and others, affecting the quality of the observations. Moreover, the primary objective of the study is often to understand and characterize these observations, which means that we cannot have prior knowledge of them. Fortunately, physics theories provide a description about the phenomena at stake, and the implementation of the underlying equations into modern computing systems allows to access a large number of simulations. Synthetic data has many advantages. As computer graphics imagery gets better with time, more realistic images can be produced, and it can be easier in both time and resources to produce simulations than to collect and label real data. Also, it allows to have a full control of the image parameters, such as lighting, physics, objects, or scenes. Historically, the first production of IACT-based simulations dates back to 1964 [Mir22], and modern synthetic data are nowadays very reliable, as demonstrated by the recent performance in source detection [Abe+23].

The previous chapter has shed light on the necessity to develop new approaches to account for the reconstruction weakness inherent to the standard analysis, as discussed in Section 2.5. This aspect is critical especially at the lower energy range of the gamma-ray spectrum as it contains the majority of the flux. This has been addressed in the previous work of [Jac20] with the creation of the γ-PhysNet neural network, and has shown performance improvements compared to the Hillas+RF method, especially when evaluated on synthetic data. However, the sensitivity of deep learning models to variations in data distributions has resulted in reconstruction biases when models trained on simulations are applied to real acquisitions [Vui+21].

Chapter contributions

Introduced previously in Section 2.7 of Chapter 2, domain adaptation has been defined as a set of techniques to mitigate domain shifts. Its integration into the γ-PhysNet is a promising approach to enhance its detection capabilities of gamma-ray sources. In parallel, a better understanding on its benefits can be achieve through putting the results in perspective with the data adaptation and model robustness, described in 2.6. In the former, NSB matching is performed to equalize the amount of noise in the image background. On the other hand, the latter considers training a model to obtain features that are NSB agnostic. For these purposes, in chapter 2, and more precisely in Section 3.2, we have selected CBN modules, DANN, DeepCORAL, DeepJDOT and ViT as the most favourable approaches for our work.

Furthermore, we detailed the application procedures and functioning of our selected methods in the last chapter. We provided the formulation of input conditioning to tackle NSB variations and ensure model robustness through the integration of conditional batch normalization modules [Vri+17] in place of traditional batch normalization layers [IS15]. Two methodologies were proposed in Section 3.3: a simple approach using scalar conditioning, and a more precise one using image conditioning. We then explored unsupervised domain adaptation and its extension in the context of strong label shift in Section **??**. This led to the development of γ-PhysNet-CDANN, γ-PhysNet-CDeepJDOT, and γ-PhysNet-CDeepCORAL. Moreover, we highlighted the complexity of optimizing the domain task

when associated in a multi-objective context with the reconstruction branches. To this end, we design the Gradient Layer as an extension to any domain adaptation model to gradually incorporate the domain contribution during the training process. Lastly, Section 3.6 introduced γ-PhysNet-Prime and γ-PhysNet-Megatron, the latter being its domain adaptation counterpart, as potential global models to move beyond the current run-wise procedures.

In this chapter, the contributions are threefold:

- Firstly, we exploit the strength of the γ-PhysNet-CBN in the creation of a NSB-robust model in the context of simulations.
- Secondly, we introduce deep unsupervised domain adaptation in the context of IACT image analysis to tackle domain discrepancies. We evaluate the performance of our selected methods to face background variations and their extension in the label shift scenario. We have conducted a extensive experimental study to evaluate the potential of each method and select the most promising one for the application to the real source detection presented in the next chapter. To this end, we assess the performance of the selected multi-task balancing approaches and the contribution of the Gradient Layer.
- Finally, as an exploratory work, we evaluate the MAE reconstruction and the performance of both the γ-PhysNet-Prime and γ-PhysNet-Megatron on synthetic data.

In this chapter, our novel methods are applied to perturbed simulations, as they provide a controlled environment for models performance estimation. There are many factors that can be manipulated to gain a comprehensive understanding of the most critical factors for the success or failure of these techniques, especially the amount of NSB and the shift in target labels. Our approaches will therefore be evaluated on real acquisitions for the detection of the Crab Nebula in chapter 5.

4.2 Data workflow

In the case of CTAO, the private LST consortium dataset of Monte-Carlo simulations is generated with the CORSIKA (COsmic Ray Simulations for KAscade) software [Hec+98]. This software package is used for both the simulations of electromagnetic and hadronic cascades within a simulated atmosphere, allowing for the tracking of secondary particle trajectories down to relatively low energies of around 1 MeV. The CORSIKA package relies on Monte Carlo (MC) modeling to simulate the formation of EAS in the atmosphere, initiated by various particles. Because hadronic particles are sensitive to magnetic field, this software also consider the Earth's magnetic field to modify their trajectories. The processes at stake in CORSIKA are supported by strong theoretical studies and experiments conducted at accelerators and colliders. Besides, a Poisson diffuse light is added and evenly distributed across camera pixels to simulate the Night Sky Background (NSB). If necessary, noise-free synthetic data is available and NSB could in theory be implemented with background images. Nevertheless, the simulation outcomes produced by this package align well with existing experimental acquisitions.

The telescope response to the EAS is obtained with the `sim_telarray` software [Ber08] at a pointing zenith angle of 20° and azimuth 180°, pre-processed with cta-lstchain [Lop+22] using pixel-wise charge integration as described in [Abe+23]. The dataset is here referred to

	$N_{\text{gamma-diffuse}}$	N_{gamma}	N_{proton}	N_{electron}
#Training	1116578	-	757561	-
#Test	-	1057555	756954	1057555

Table 4.1: Number of events in each subset.

as Prod5 LST-1 mono-trigger. This results, for each event, in pairs of images corresponding to the integrated charge amplitude and the photons average arrival time for each pixel. Such bi-modal samples are used as input for our network. As the images lie at the sensor level on a non-conventional hexagonal grid of pixels, we interpolate them on a regular grid using the dl1-data-handler library [Kim+22]. We thus benefit from the high-performing implementations of the PyTorch convolution, reducing the training time of our models, while obtaining the same performance compared to a more sensor tailored approach that relies on indexed convolutions [Jac+19] for hexagonal pixel images.

The standard dataset

Configuration of the standard dataset

The Prod5 LST-1 mono trigger dataset, referred to in this thesis as the standard dataset, is separated into a train and a test set as it is classically performed in machine learning strategy. The number of events in each subset is given in Table 4.1. The training set contains diffused gammas - coming from each direction of the sky -, and protons. It does not contain electrons, as the showers engendered by a gamma or an electron are very similar and only differ with its first interaction, which makes it very complicated to distinguish. On the other hand, the test set contains point source gammas, as it is closer to the real acquisition process, and protons.

An example of pixel charge and temporal information synthetic images resulting from the pre-processing chain is depicted in Figure 4.1. Firstly, the image intensity is correlated with the number of photons that has reached the telescope optical system and been converted into an electrical signal. On the other hand, the background contains fluctuations, that can be attributed to electronics noise, hadronic showers, observational conditions, and other distortions. In the case of simulated data, this background is computed as a Poisson noise of rate λ_{MC}, as real images might exhibit a noise distribution that is usually approximated by such a distribution. Secondly, the time gradient from the temporal map allows to clarify the directional ambiguity of the shower deployment introduced by the waveform integration. It corresponds to the trigger time of the underlying photo-multipliers. In more details, inferring the shower development direction from a Gaussian-like morphology can possibly be estimated from the third-order moments, the skewness and kurtosis, as highlighted in [Abe+23] from the contribution of the Hillas parameters within the Random Forests. However, the frugality of the information and statistical fluctuations complexify greatly this task, and the time gradient remains the most determining factor to lift the veil on the temporal expansion of the signal.

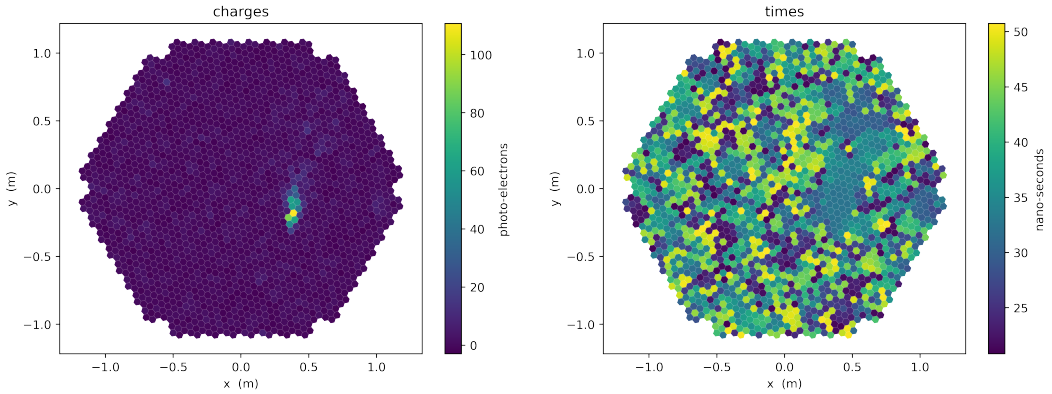


Figure 4.1: Pixel charge (left) and temporal information (right).

Optimization of the model and inherent biases

Model training aims to compute an optimal parametric function between the labels and the high-dimensional input images using a subset of the whole population. This typically requires unbiased datasets, which can be achieved by considering uniformly distributed categories such as energy, intensity, direction and class, among others. However, considering the LST dataset, there are several factors that can lead to unbalanced and biases characteristics.

- The energy follows a specific distribution, depicted in Figure 4.2. In particular, there is a natural limit at lower energies that corresponds to the minimum energy required for this particle shower to propagate through the medium. Additionally, hardware constraints of the telescope's trigger threshold also requires a minimum energy. Besides, at higher energy levels, incident particles generate exponentially more interactions, imposing another hardware limit dictated by available resources to compute the EAS. In such scenario, showers are usually reused multiple times but at different positions to generate output events, meaning that high-energy events are not independent and identically distributed per se. Appendix 8.12 aims to understand the impact of rebalancing the intensities during the training.

- The image intensity resulting from the incident particle energy interactions with the atmosphere is directly correlated with the amount of reactions in that medium, and thus with the incident energy. Therefore, unbalanced energies ultimately lead to unbalanced intensities.

- The training set aims to represent the field of view of the telescope at a specific pointing location. This is refereed to as diffused particles, in opposition to point source which corresponds to a single position in the sky. However, through the simulation of the optical subsystem, although the direction is uniformly distributed before considering the instruments response, the resulting distribution follows a Gaussian law. As a result, the further the particle coordinate from the camera center, the less represented it becomes.

- Among the factors that modifies the energy distribution, the atmosphere thickness that the particles has to penetrate plays a determinant role. In particular, because it

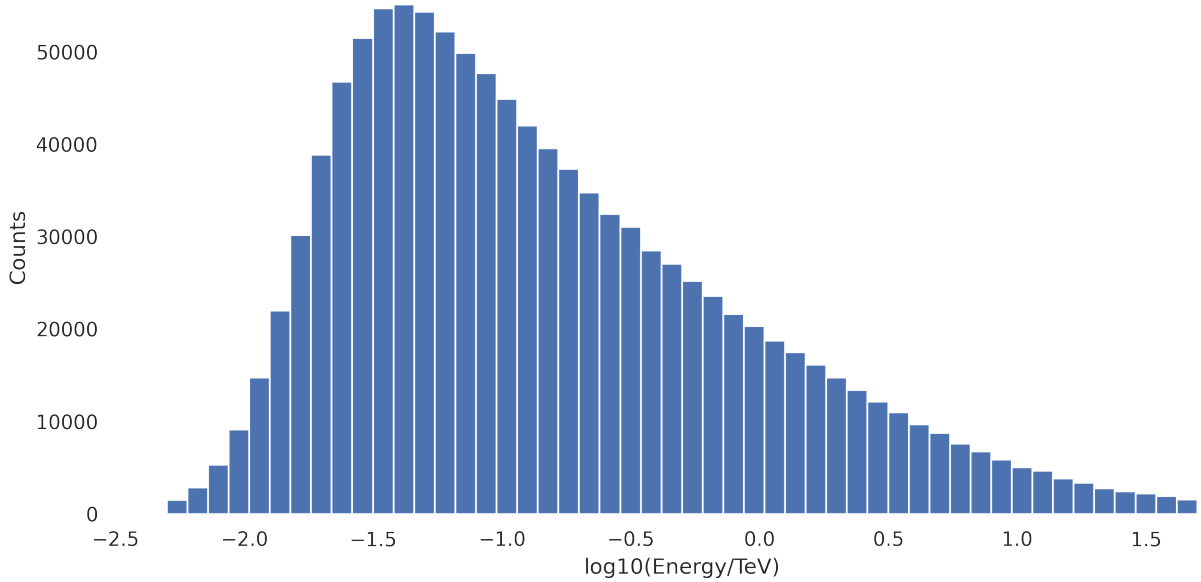


Figure 4.2: The log-energy distribution of training diffuse gamma of the standard dataset.

is computationally not possible to generate simulations covering the whole sky, each training dataset is spatially localized to a specific pointing direction. The confinement of the zenith parameter to a single value ultimately affects the model performance if the test data do not follows the same zenith angle. If the expected gamma-ray source coordinates are far from the training dataset, source position errors must be expected [Vui+21]. A possible corrective approach is to combine images with multiple pointing directions in the training dataset. However, this strategy requires to inject the pointing information in the network in some ways. In addition, the capability of the network to interpolate between pointings, depending on their amounts and values, is a question to be studied.

The complete representation of the training and test simulated datasets are given in the Appendix 8.3.

To conclude this analysis of the standard dataset, some biases have already been identified and may impact model optimization. Although a direct and uncorrected application of deep learning has evinced promising results, their is a need for data adaptation as a rudimentary performance restoring approach and to increase the detection capabilities especially on fainter source.

The tuned dataset

Synthetic data are produced with a small NSB amount corresponding to a dark sky patch. As mentioned in Section 2.5, it is approximated with a Poisson distribution of rate λ_{MC} that can be retrieved either using the image cleaning mask described in Section 1.3 to discard signal or keeping only the background noise through a cautious intensity threshold operation. In our simulated dataset, the rate has been measured at $\lambda_{MC} = 1.77$.

In real scenarios, the NSB distribution varies constantly between observations. The standard procedure then proposes to compensate for this discrepancy with background match-

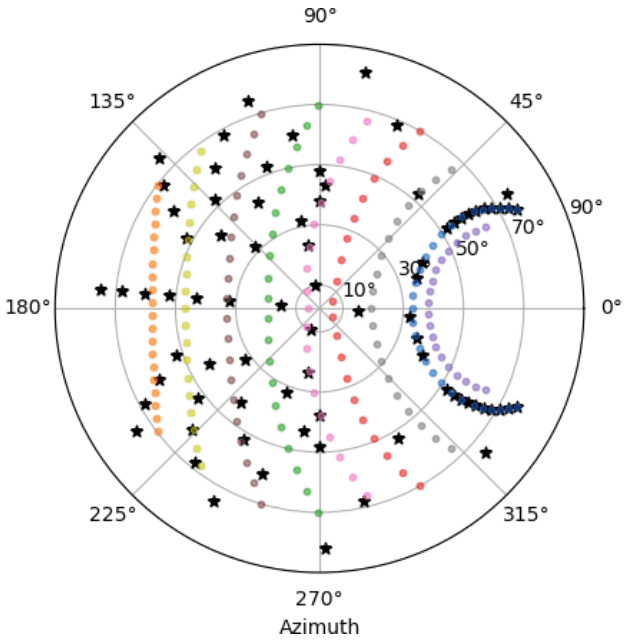


Figure 4.3: Simulated pointing directions. Source [Abe+23].

ing, which consist in adding a sufficient amount of Poisson noise $\delta\lambda$ to match the NSB distribution of the real data λ_{real}. This new dataset is referred to as the tuned dataset. A precise description of the real dataset is given in the next chapter. As a macro value computed from the entire sequence of images, we compute a real rate $\lambda_{real} = 2.23$. Both standard and tuned intensity histograms are plotted in Figure 4.4, and the rate difference is determined as $\delta\lambda = \lambda_{real} - \lambda_{MC} = 0.46$.

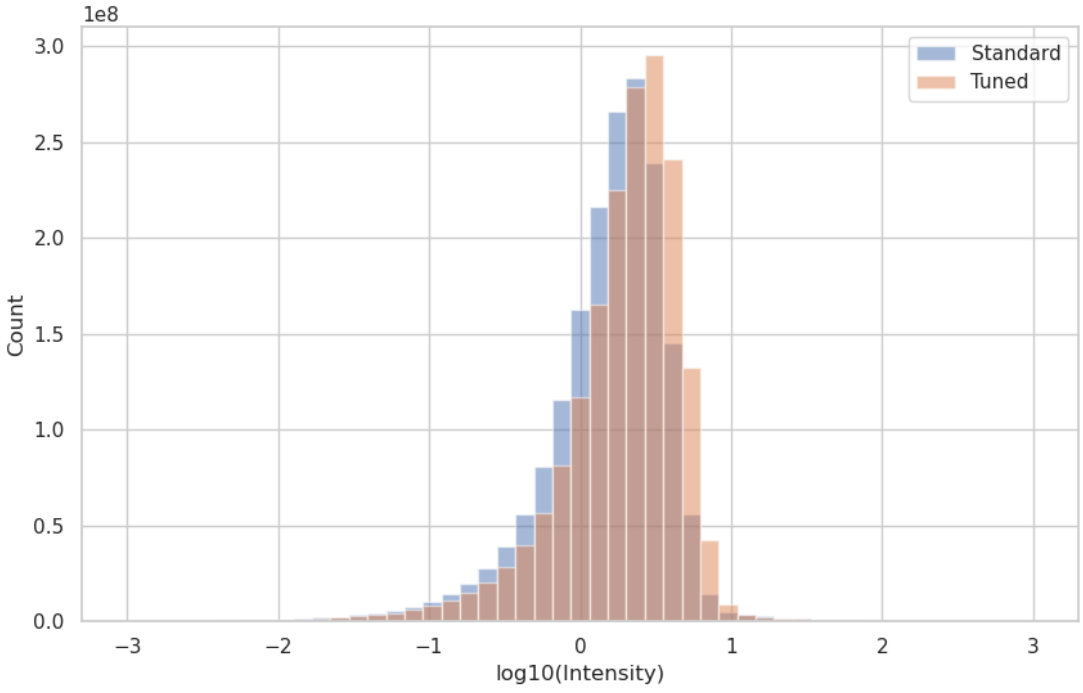
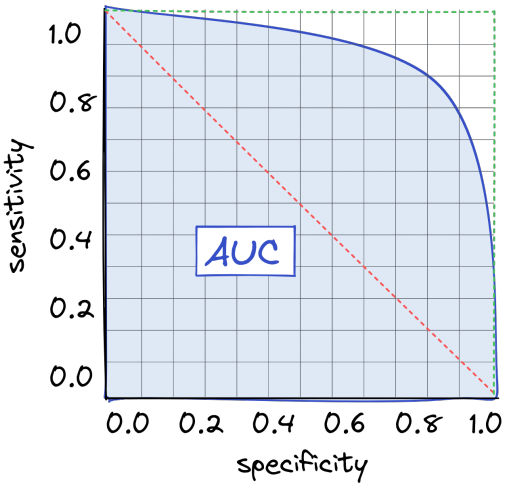


Figure 4.4: Standard and tuned gamma distributions.

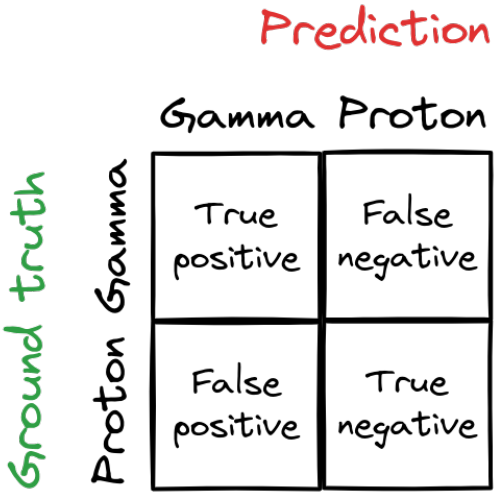
4.3 Figures of merit

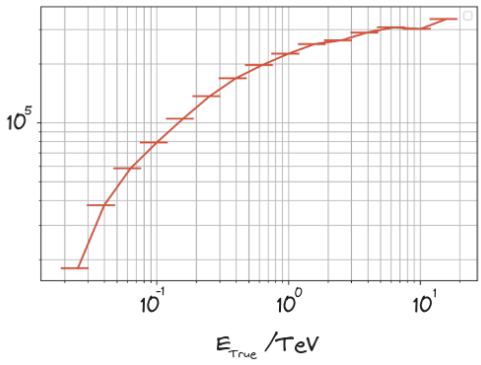
Figures of merit refer to metrics that evaluate the performances of a model. In the case of IACT, physical models involves physics-based metrics in order to measure the impact of the reconstructed events on the telescope efficiency [OC21]. The evaluation of the performance allows comparing different methods with the use of astronomical and classification indicators. Because of the performance dependence in the energy level, they can be computed per bin of energy and thus provide discrete curves. In the following are reported metrics used by astrophysicists to qualify models along with some necessary physics terminology definitions. Table 4.5 summarizes the figures of merit considered in this study.



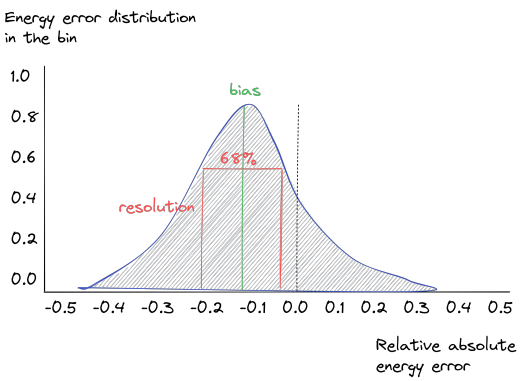
The **Receiver operating characteristic (ROC)** [Faw06] is a graphical representation used to evaluate the performance of a binary classification model. It illustrates the trade-off between the sensitivity and the specificity at various threshold settings. The **Area under the curve (AUC)** is defined as the integral of the Receiver Operating Characteristic (ROC) function. Varying a threshold, we can determine the true positive and the false positive rates. The probability that an incident particle is a gamma instead of a proton is called the **gammaness**.

A **confusion matrix** is a common tool for evaluating the classification performance of a model, by confronting the ground truth with the model predictions. Using this figure of merit, several important normalized performance metrics can be derived, such as the accuracy, the precision, the recall, the specificity and the F1 score. In the case of CTAO, we are interested in background rejection, hence the consideration of two classes: $\{gamma, \overline{gamma}\}$. The sensitivity and specificity are often used to compute the The Receiver operating characteristic (ROC) curve.



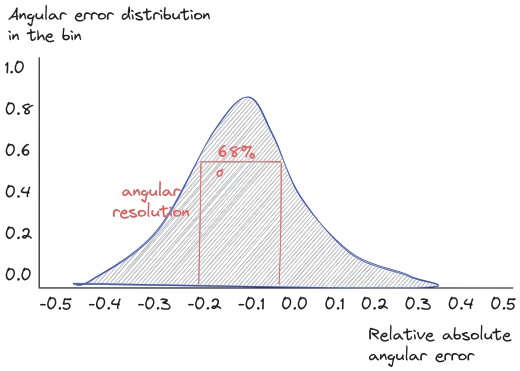


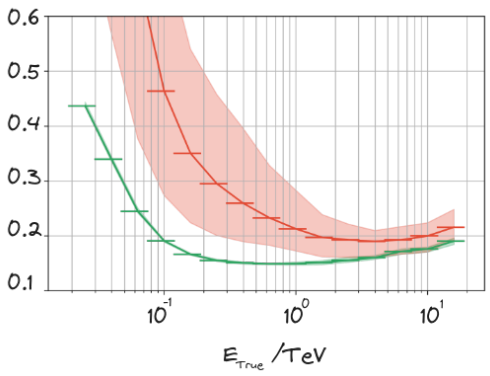
The **effective collection area** corresponds to the ground virtual surface of the telescope measured a the number of gamma particles correctly reconstructed over the total number of simulated gamma particles. It is directly linked to the the classification power of the reconstruction.



The energy resolution is a measure of performance of the energy reconstruction task. It is commonly admitted that the estimated value $\tilde{y}$ follows a Normal distribution of variance σ. It represents the centered interval containing 68% of the relative prediction error $\frac{\Delta E}{E}$ and it is computed per bin of energy. The lower the energy resolution, the less the error spreads around the mean and thus the better the model. Each bin may have a mean relative prediction error known as the energy bias.

The angular resolution, or spatial resolution, is a measure of performance of the direction reconstruction task. Similarly to the energy resolution, this indicator represents the interval containing 68% of the angular separation between the predictions and the true directions. The lower the angular resolution, the more capable the telescope is to distinguish spatial details and thus the better the model. Usually, the altitude and azimuth values (alt, az) are predicted as the direction coordinates.





The telescope responses to the incident particles can be calculated and represented with their **Instrument Response Functions (IRFs)**. In this work, it corresponds to all the metrics described above. The performance variations are typically reported by bins of energy. We propose to summarize the IRFs by integrating each figure over the energy. Although this procedure hides the energy-dependent performance of the model, it allows to illustrate and quantify the difference between each method with a more synthetic metric. In the presented results, the IRFs calculation is similar to [Abe+23].

Figure 4.5: Figures of merit for the analysis of the model performance when applied to simulations.

Lower and upper bounds

It is good practise to constrain the performance of a model within lower and upper bounds, representing the best and worst scenario respectively. We define the best scenario as the performance obtained when training and testing the γ-PhysNet model on the same NSB distribution, thus $\lambda_{MC}^{train} = \lambda_{MC}^{test}$. On the other hand, the worst scenario refers to the case where the test distribution doesn't follow the training data, but in perturbed with a Poisson noise of rate $\delta\lambda$. In that case, $\lambda_{MC}^{test} = \lambda_{MC}^{train} + \delta\lambda$. The best and worst scenario are usually showcased with the green and red curves. It is important to note that in both cases, the

training parameters remain identical to [Jac20].

Influence of parameter initialization

Each experiment is conducted N times when possible to account for the parameter initialization and data shuffling uncertainties, and are represented as the surface area defined by the min and max values. Because of the computation cost, we usually train our models $N = 5$ times. The use of min and max values is more suitable over the standard deviation when the statistic is low. On the other hand, the solid lines corresponds to the mean value across all seeds.

4.4 Application of the current γ-PhysNet to simulations

The results presented in the section constitute the baseline to further evaluate and validate the contributions highlighted in the introduction. In the following, we aim to assess the impact of NSB to the γ-PhysNet performance.

Training parameters

For each of the following experiments, the convergence of the γ-PhysNet is ensured within 30 epochs. Following the recommendation of [Jac20], the weights of the model are updated using the Adam optimizer with a learning rate of $1e^{-3}$, and the weights associated loss balancing function are optimized with the Adam optimizer with a learning rate of 0.025. Both optimizers have a weight decay of $1e^{-4}$. We have conducted experimental studies to identify relevant model, and we provide hereafter the main conclusions. Relying on the multi-task analysis available in Appendix 8.11, Uncertainty Weighting, proposed in [KGC18], offers the best performance. Furthermore, our normalization layer comparison in Appendix 8.8 suggests the use of Batch Normalization [IS15]. Regarding the activation functions, although our recent study in Appendix 8.9 points out small benefits of GELU compared to the traditional ReLU, the latter has been kept in this configuration. A learning rate step scheduler is applied, and the rate is divided by 10 every 10 epochs. In total, the baseline model contains $3.5M$ parameters.

During the telescope observations, the acquired waveforms are encoded on 12 bits. Although the information is cast afterwards on 16 bits, the last bits contain no relevant information. Model training is performed on A100 GPUs hosted on the MUST datacenter[1]. The allocated resources on the A100 GPUs are different depending on the considered datatype. We currently use the default Tensorfloat (float19 type) for the synthetic data.

The value assigned to the batch size usually depends on the GPU memory and the training data size. Although increasing the batch size may allow to faster converge, as experienced in [Smi+18], we fixed a constant value of 256 across all the experiments. A refinement can be performed in the future to explore the contribution of the size of the batch to the performance or the speed of convergence of the training.

[1] https://www.must-datacentre.fr/

Impact of the NSB on the γ-PhysNet

In this section, our objective is to assess the influence of the NSB on the γ-PhysNet neural network. Simulations intrinsically contain a fixed amount of noise, following a Poisson distribution of rates λ_{MC}. An additional Poisson noise is applied with a parameter $\delta\lambda \in \{0.2, 0.4, 0.6, 0.8, 1.0\}$, so that it is equivalent to follow a Poisson distribution of parameter $\lambda = \lambda_{MC} + \delta\lambda$. This illustrates a simple case of real observation behaviours with a variation of light pollution. We first train the γ-PhysNet using $\delta\lambda = 0$, and test it on the specified range of noise rates.

The results are presented in Figure 4.6. As expected, the γ-PhysNet is very sensitive to the difference in training and test distribution. The degradation of the performance depends on the energy level. Lower energy suffers the most from the perturbations, because the signal-to-noise ratio is less favourable. It is remarkable that the energy resolution decreases by more than 50% with a small amount of additional noise $\delta\lambda = 0.2$ at the first bin of energy. However, as it will be highlighted in the next chapter, this scenario is quite common in real observation analysis.

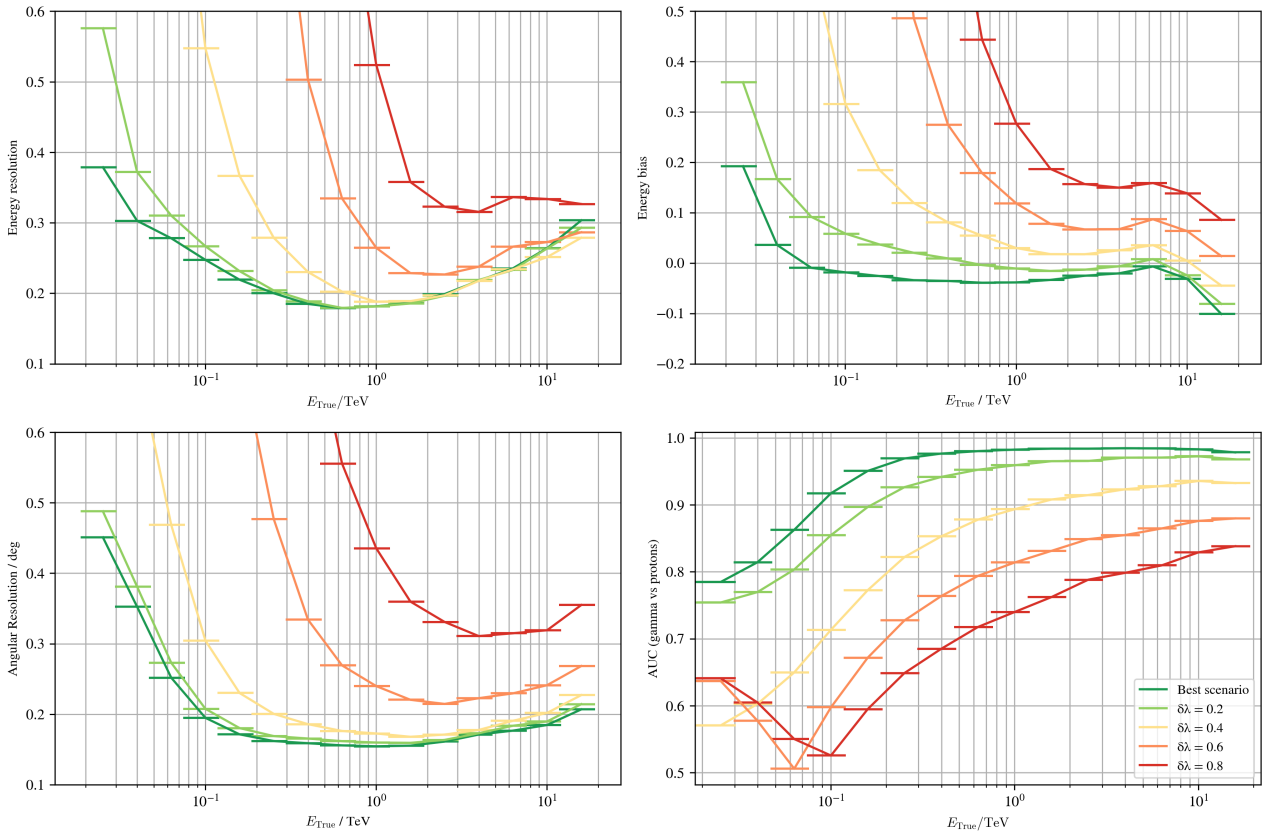


Figure 4.6: Impact of NSB on the baseline γ-PhysNet. The model is trained on a simulated dataset containing a intrinsic Poisson noise of rate $\lambda_{MC} = 1.77$, and evaluated on test datasets with a Poisson noise of rate $\lambda = \lambda_{MC} + \delta\lambda$. Only one seed per experiment has been produced. The training parameters are identical to [Jac20].

In conclusion, variations of NSB decreases the performance of the γ-PhysNet model, especially at the lower energy levels. However, as explained previously in Section 2.5, this part of the spectrum concentrates most of the gamma flux. Therefore, this loss of informa-

tion becomes a bottleneck in source detection, highlighting the need for robustness towards this entity.

4.5 Application of input conditioning to simulations

Preliminary study on the digit datasets

In order to validate and evaluate the effectiveness of the γ-PhysNet-CBN, we conducted a preliminary study on the multimedia digit datasets. In this investigation, data augmentation is performed through the application of a Gaussian noise. The key idea is to make the model robust to this perturbation. A Gaussian distribution can be fully characterized with the first- and second-order moments, namely the mean and standard deviation. In that case, we chose a centred Gaussian (mean is zero), and a varying standard deviation, sampled from a Uniform distribution, $\sigma \sim \mathcal{U}(0,1)$. More details can be found in the corresponding Appendix 8.5, and the conclusions of this analysis are presented here:

- The inclusion of the noise information within CBN modules allows to recover for the loss for $\sigma < 0.4$. Above this value, the performance decreases linearly with the augmentation of the perturbations.
- To retrieve the initial baseline accuracy the γ-PhysNet-CBN must be trained much longer, and the number of necessary epochs increased fivefold. This is yet expected because noise robustness must be learnt compared to solely training on the original data.

Based on this preliminary study, we can therefore expect such behaviour in the LST dataset scenario.

Training parameters

The training parameters are identical to the γ-PhysNet described in Section 4.4. However, relying on our experimental study using the digit datasets in the Appendix 8, we extended the training time up to 150 epochs. This is necessary, because the model needs more time to converge while using data augmentation. The learning rate scheduler divides the rate by 10 every 25 epochs, and the remaining hyperparameters are set identical to previously. Overall, the size of the model is almost identical to the γ-PhysNet, because the linear encoder and CBN modules contain few parameters.

Impact of the NSB on the γ-PhysNet-CBN

We study now the benefits of integrating CBN modules into the γ-PhysNet to tackle the NSB variations. Results are depicted in Figure 4.7. Remarkably, performance on resolution metrics reaches the best scenario regardless the amount of noise applied to the data, ensuring that the model has successfully learnt robust features from noisy images. Energy bias is slightly deteriorated between 1 – 10 TeV, with greater variation at higher noise level, as measured by the surface area. Although a minor loss in classification power is observed at lower energies, the AUC remains almost aligned to the best case. In comparison to the previous

methodology, γ-PhysNet-CBN surpasses domain adaptation in the specific and limited case of synthetic data considering exclusively NSB variations.

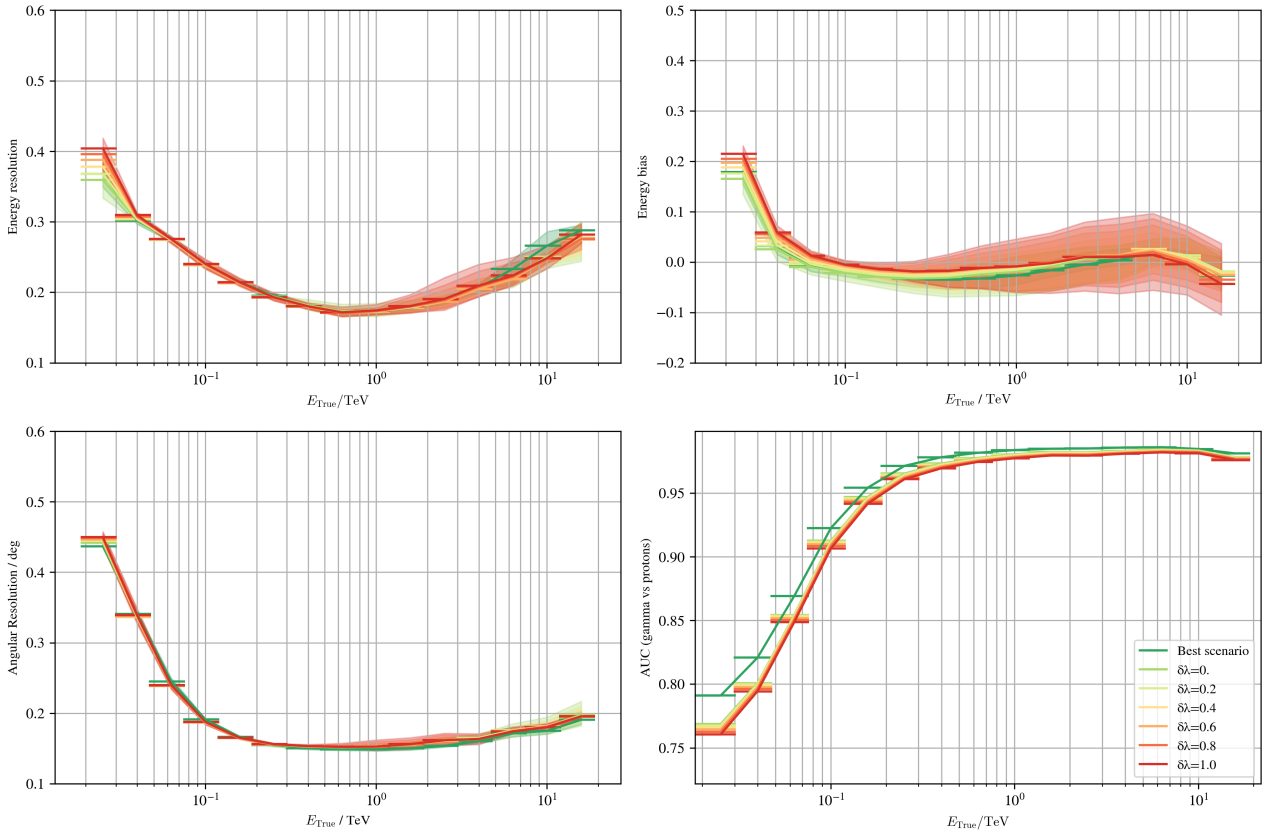


Figure 4.7: Impact of NSB on the γ-PhysNet-CBN. Five seeds per training have been produced. The solid lines show averaged values over the different seeds and the envelops depict the min and max values.

In conclusion, the replacement of BN layers with CBN modules within the γ-PhysNet architecture is relevant to provide NSB robustness on synthetic data. Yet, domain adaptation should still be relevant to help gain robustness on other unknown factors.

4.6 Application of unsupervised domain adaptation with the γ-PhysNet-DANN to simulations

In this section, we conduct a precise evaluation of the integration of the unsupervised domain adaptation technique DANN within the γ-PhysNet model. To this end, the impact of NSB and of label shift are highlighted in detail, and we provide the results of conditioning domain adaptation with the specific IRFs.

Training parameters

In this context, the convergence of the γ-PhysNet-DANN is guaranteed within 50 epochs. Following DANN original paper [Gan+16], the domain classifier is composed of two fully-connected layers of 100 features with a ReLU activation function. In this section, the task

balancing strategy is Uncertainty Weighting. Regarding the hyperparametrization, it is identical to the γ-PhysNet previously described. In total, the resulting γ-PhysNet-DANN based approach has $4.8M$ parameters, an increase of 37% compared to the baseline γ-PhysNet reported in Section 4.4.

Impact of the NSB on the γ-PhysNet-DANN

In order to evaluate the contribution of domain adaptation in real acquisition conditions, targets approximate telescope observations and contain a noise level which rate fluctuates according to a uniform distribution, meaning that $\delta\lambda \sim \mathcal{U}(0,1)$. On the other hand, sources are not modified with $\delta\lambda = 0$. Then, we chose to evaluate γ-PhysNet-DANN on different levels of $\delta\lambda \in \{i \times 0.2\}_{i=0}^{5}$. If the adaptation run smoothly, the best scenario should be reached regardless the value of $\delta\lambda$. Results are presented in Figure 4.8. It turns out that all the IRFs exhibit the same pattern, and the performance of the model decreases as the rate difference increases. In addition, the depicted surface area indicates that parameter initialization has more influence with higher noise levels. The drop in reconstruction quality is especially visible on the energy resolution and bias, but also on the classification power, as suggested by the AUC, with a loss of roughly 10% on most of the spectrum at worst. The angular resolution appears less affected, yet still suffers from the degradation of the signal-to-noise ratio. However, put in perspective with Figure 4.6, the integration of domain adaptation into the framework still helps reducing the impact of the NSB on all metrics. Interestingly, in the case of $\delta\lambda = 0.$, the energy resolution is better at the lower range, while it should be equal to the best scenario. In fact, domain adaptation uses both source and target data, then the model is intrinsically trained with more images. Furthermore, we set in this case a number of epochs to 150, in comparison to the 30 training epochs of the classical network.

To conclude, in comparison to the baseline γ-PhysNet and the results depicted on Figure 4.6, domain adaptation improves the performance but is still sensitive to the amount of NSB. Furthermore, when the rate difference $\delta\lambda$ is measurable, the CBN-based γ-PhysNet presented in Section 4.5 is more robust to variations of NSB.

Impact of the label shift on the γ-PhysNet-DANN

To illustrate the impact of label shift on domain adaptation, we train the γ-PhysNet-DANN model using multiple amount of gamma/proton ratios, denoted as $r \in \{10^{-i}\}_{i=1}^{5}$, in the target dataset. A fixed amount of Poisson noise of rate $\delta\lambda = 0.46$ is also applied to the target, which is representative to a plausible real scenario, as it is demonstrated in the next chapter.

The results are presented in Figure 4.9. The performance metrics are noticeably affected by the label shift, and the consequences are twofold. Firstly, the resolution and bias of energy and angle increase as the ratio decrease, highlighting a loss in performance (a drop of 0.1 to 0.15 points in the energy range 0.01 to 0.1 TeV in both cases). Concurrently, the AUC on the target data increases, which signifies that the particle classifier performs better on noisy events when gammas are less represented in the target dataset. There are several possible explanations to this peculiarity. As evinced in Figure 3.6 in the previous chapter, the particle and domain classification tasks are conflicting. It is possible that reducing the ratio of gammas in the target dataset allows the domain objective to take the lead on the particle task.

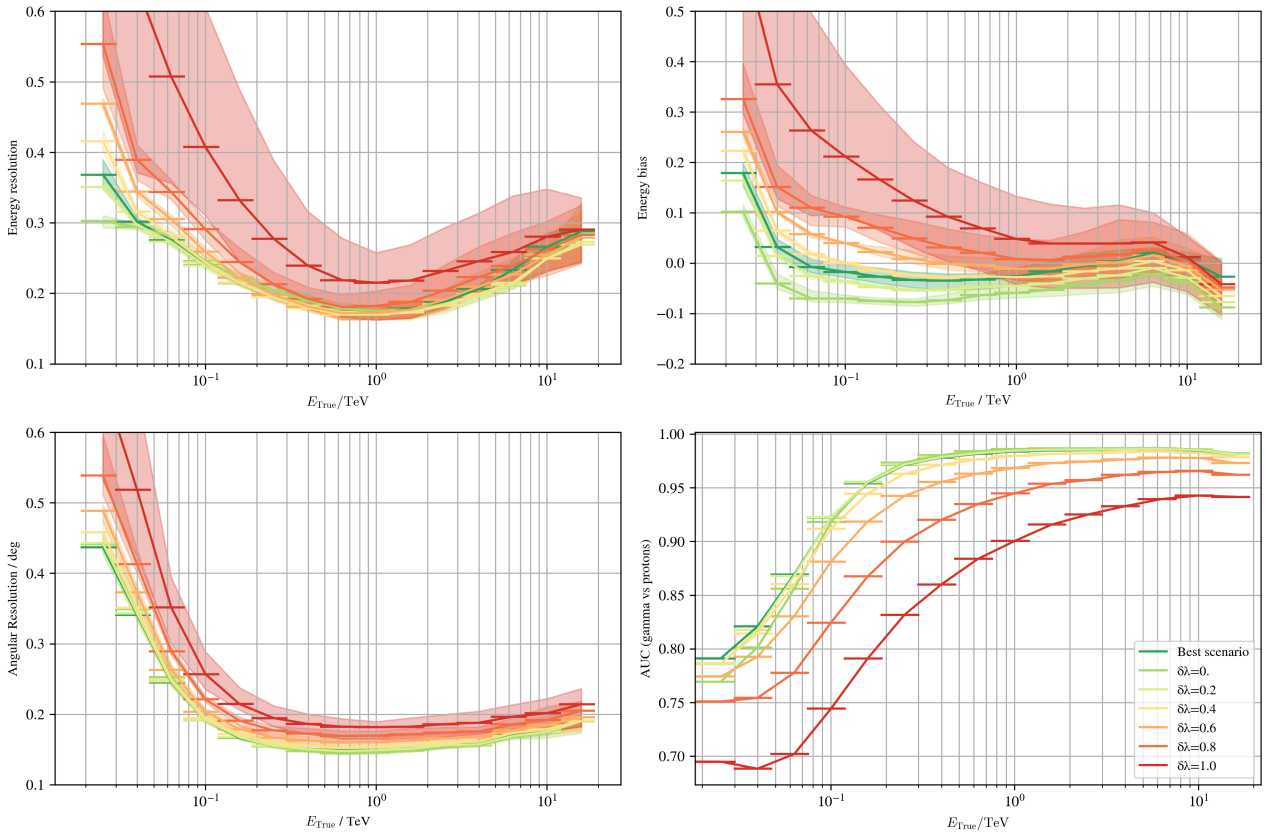


Figure 4.8: Impact of NSB on the γ-PhysNet-DANN. The model is trained on a simulated dataset containing a range of Poisson noise of rates $\delta\lambda$. Only the results on targets are plotted. Each training has been conducted on five different seeds. The solid lines show averaged values over the different seeds and the envelops depict the min and max values.

In summary, the impact of label shift on domain adaptation, as demonstrated by the performance of the γ-PhysNet-DANN across different gamma/proton ratios, underscores the importance of addressing this problem to ensure robust and effective domain adaptation.

Tackling the label shift problem inherent to domain adaptation with conditioning

The previous section has highlighted the degradation brought by label shift on the specific case of the γ-PhysNet-DANN. Based on importance weighting, as discussed in Section 3.4 of the last chapter, our focus in this section lies in quantifying the benefits of conditional domain adaptation. Firstly, the extension of the γ-PhysNet with CDANN shows in Table 4.2 that, while GradNorm (GN) initially outperformed Uncertainty Weighting (UW) without conditioning, we finally observe a change in performance dynamics. CDANN combined with UW yields the best overall results for the energy bias and resolution, demonstrating that the introduction of conditioning allows to recover from label shift, ultimately leading to a restoration of performance levels comparable to those observed previously. Furthermore, CDeepCORAL combined with GN results in a significant improvements in energy bias and resolution at lower energy levels compared to DeepCORAL, although a slight reduction in resolution is observed at intermediate levels. These improvements are consistent

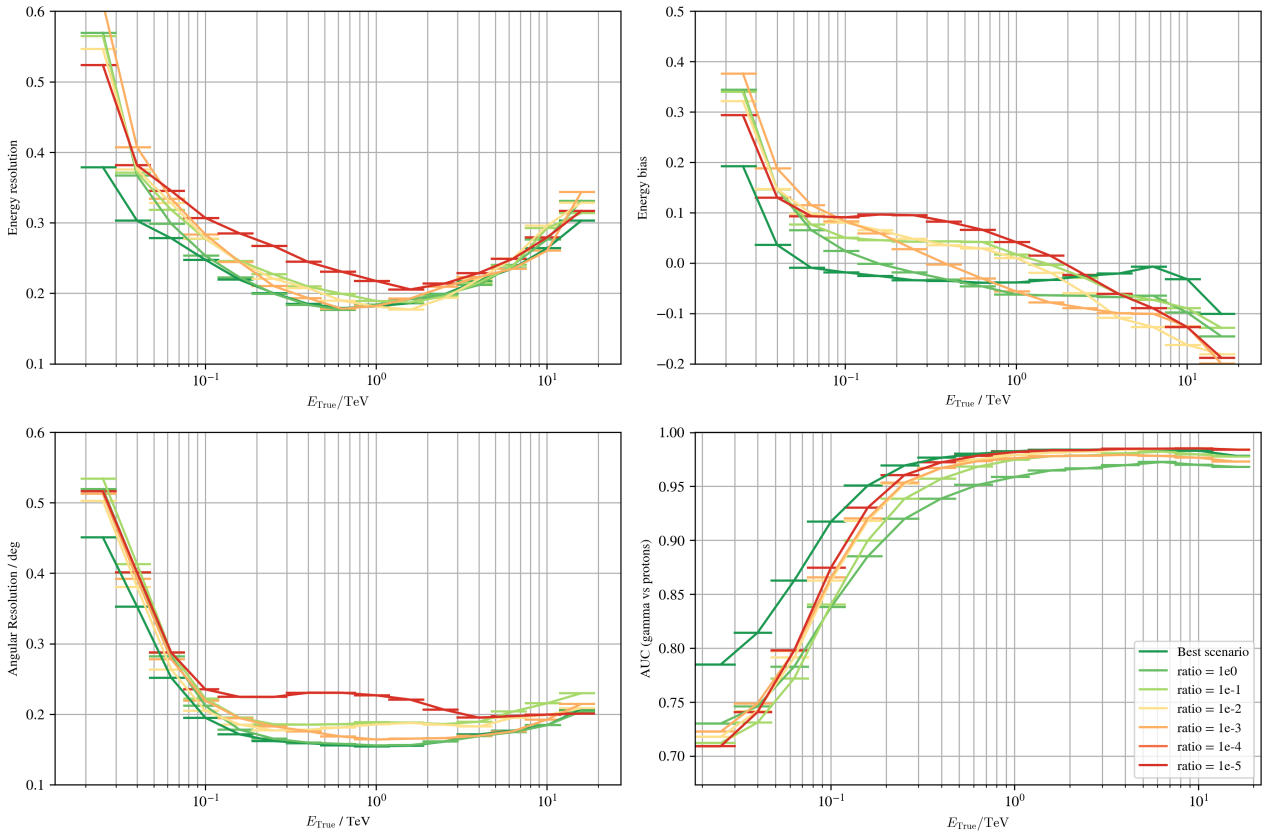


Figure 4.9: Impact of the label shift on on the γ-PhysNet-DANN. The model is trained with target data containing ratios r of gammas and protons, and a fixed amount of noise is applied. Only the results on targets are plotted. One seed has been computed for each case.

across both the source and target domains. However, the conditioning with UW degrades the energy resolution. Finally, CDeepJDOT paired with GN reveals a slight increase in performance on the target energy bias, and the overall metrics exhibit a deterioration while paired with UW. The addition of conditional domain adaptation doesn't allow for DeepJDOT and DeepCORAL to converge while using UW.

For more details, IRFs of γ-PhysNet-DANN and its conditional version are presented in 4.10. The consequences of label shift are mainly visible on the intermediate range of energies.

Preliminary conclusions

Synthetic data allows us to exploit labels in order to evaluate the detrimental impact of specific variables of interest. In this section, we were interested in quantifying the consequences of NSB and label shift on the γ-PhysNet-DANN. As a result, when a complex noise of a variety of Poisson rates is applied to the inputs, this approach fails at recovering the loss across all rates. In fact, the stronger the noise, the less performing the neural network becomes. On the other hand, the introduction of conditioning, translating into a new training procedure, alleviate the label shift problem.

Naturally, real acquisitions are prone many unparametrizable or unknown factors influencing the model's performance. This comprehensive study is therefore limited to the

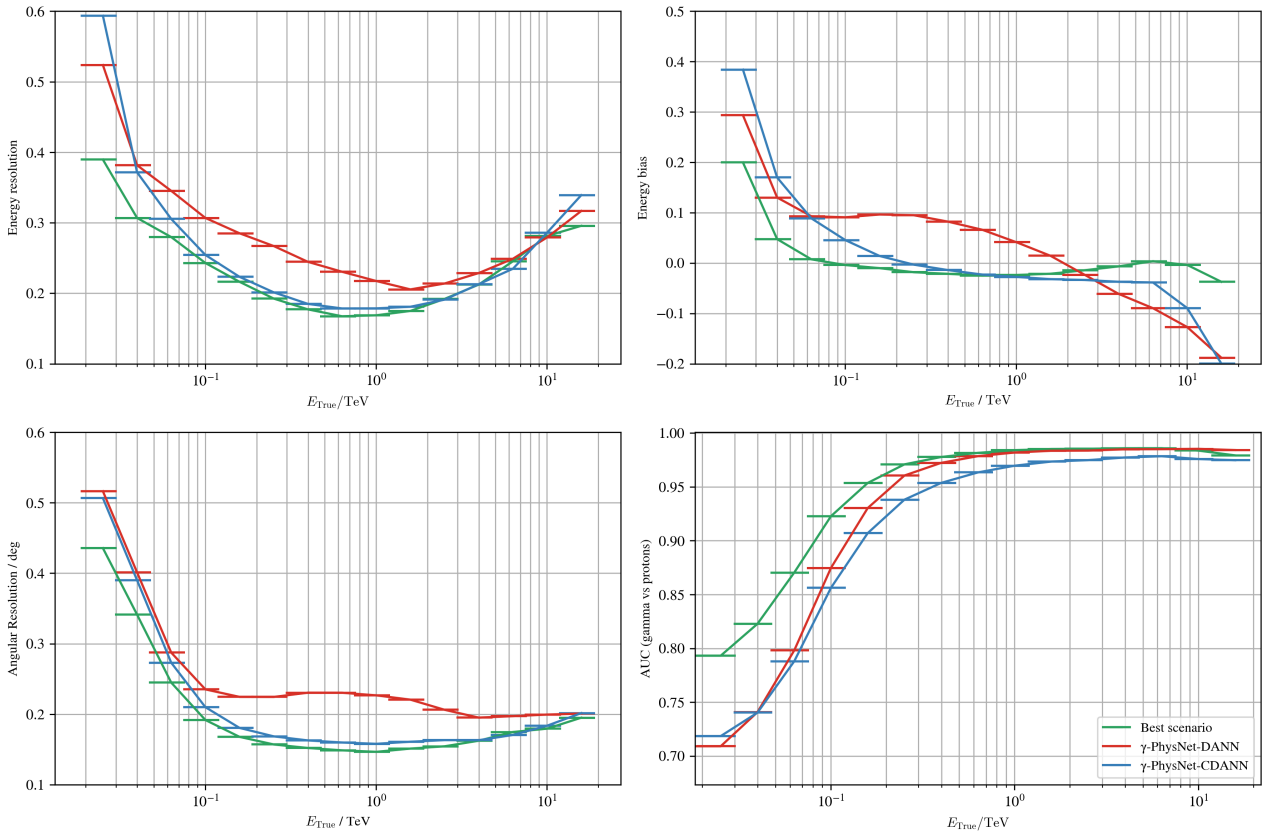


Figure 4.10: Comparison of γ-PhysNet-DANN and γ-PhysNet-CDANN, with the addition of a NSB of rate $\delta\lambda = 0.46$ and a label shift of $r = 10^{-4}$ in the target data. Only one seed as been computed.

available degrees of freedom, yet NSB is demonstrated as one of the main discrepancies between simulations and real data. If NSB variations within the telescope acquisitions are limited, that is to say the range of accessible values for $\delta\lambda$ is narrow enough ($\delta\lambda < 0.6$), the γ-PhysNet-DANN will be a promising method for gamma-ray source detection. For wider ranges, of the rate difference $\delta\lambda$ is measurable, the CBN-based approach seems more relevant.

4.7 On the search of the best unsupervised domain adaptation model: An ablation study

Introduction

In this section, we assess the performance of the γ-PhysNet-DANN, γ-PhysNet-DeepCORAL and γ-PhysNet-DeepJDOT in different scenarios. Firstly, we compare the multi-task balancing methods UW and GN, with different values of the hyperparameter α, in each case. Secondly, we evaluate the contribution of the conditioning to tackle label shift. Lastly, the impact of the gradient layer on the performance of the model is investigated. The comparison of the methods aim to determine the one that suits the most for the real data analysis. Because of the extensive results obtained, the comparison is done on the integrated metrics,

although IRFs provide a better overview of the results. When necessary, we refer to the IRFs and explain which part of the energy levels are the most affected.

Preliminary study on the digit datasets

In order to explore the challenges of integrating domain adaptation into the multi-task paradigm, a preliminary study on the digit datasets is conducted and made available in Appendix 8.6. The reduced size of both the images and the datasets allows us to train many models in many different test conditions, and anticipate the difficulties that can arise for our work on the LST dataset. The conclusions are the following:

- Including multi-task automatic balancing with GN and UW allows to retrieve or outperform manual weighting in almost all cases, without the need to perform extensive grid searches.
- Depending on the considered case, GN or UW is the best algorithm.
- GN is not sensitive to the hyperparameter α on the digit experiments with these selected domain adaptation methods.
- The Gradient Layer is the key to help DeepJDOT converge. Without it, the gradients resulting from the Wasserstein distance are too strong.

Aware of the potential issues that we might face, the following presents the results on the LST dataset.

Training parameters

Similarly, the training parameters applied to our domain adaptation methods are identical to the γ-PhysNet described in Section 4.4. The only difference concerns the training time, and the convergence is guaranteed within 50 epochs. Training batches are made of 512 samples equally distributed between the source and the target sets, each containing 256 samples. The implementation of the domain classifier follows the original DANN paper [Gan+16] and is composed of two fully-connected layers of 100 features with a ReLU activation function, the last connections representing the decision layer discriminating between protons and gammas. Globally, the γ-PhysNet-DeepJDOT and the γ-PhysNet-DeepCORAL have $3.5M$ parameters, whereas the γ-PhysNet-DANN based approach has $4.8M$ parameters. The same set of hyperparameters is applied to each following experiment.

Integration of domain adaptation and multi-task balancing

We evaluate our domain adaptation methods combined with the selected multi-task balancing algorithms in the context of a NSB of constant rate $\delta\lambda = 0.46$ and label shift of ratio $r = 10^{-4}$. The selection is based on the integrated metrics. Yet, a finer energy-dependent analysis is given, built on the study of the constructed IRFs. The complementary results are presented in the Appendix 8, but the main outcomes are depicted in Tables 4.2, 4.3 and 4.4, and respectively describe γ-PhysNet-DANN, γ-PhysNet-DeepCORAL and γ-PhysNet-DeepJDOT in 3 contexts: the initial algorithm, the integration of conditioning (+C), and the inclusion of the gradient layer (+GL).

The comparison is performed using the most performing balancing methods. Thus a preliminary study on the impact of the GN hyperparameter α is produced and we only consider GN with the best hyperparametrization for each domain adaptation algorithms. Secondly, we select the best multi-task techniques through a second review of the methods presented in Section 2.8.

Impact of the GradNorm hyperparameter α

The comparative study of our selected domain adaptation techniques combined with GN across various values of the hyperparameter $\alpha \in \{0.1, 0.5, 1.5, 3.0\}$, available in the Appendix 8, reveals that all models exhibit a better performance with smaller values of α, particularly when $\alpha = 0.1$. Specifically, γ-PhysNet-DANN efficacy noticeably diminishes when $\alpha > 0.1$, or sometimes fails to converge, as evidenced by outcomes on the source data. While γ-PhysNet-DeepCORAL demonstrates less sensitivity to variations in this hyperparameter, examination of the IRFs indicates that the main distinction arises from the resolution at higher energy levels as α increases. Similarly, γ-PhysNet-DeepJDOT corroborates the same conclusion, but higher values of α also deteriorate the performance on the source data. In the following studies, the value of $\alpha = 0.1$ is adopted for all methods.

Comparison of the multi-task balancing methods

Following the ablation study detailed in Appendix 8.11 comparing the relevance of EW, RLW, DWA, GN and UW to the γ-PhysNet neural network, only GN and UW are picked for comparison as they provided the best results.

In our investigations, depicted in the two first columns in Tables 4.2, 4.3 and 4.4, the comparison of UW and GN reveals that all our methods have a better performance when paired with GN while γ-PhysNet-DeepCORAL. However, γ-PhysNet-DeepJDOT fails at converging with UW. This is potentially due to the lack of guarantees of both methods to meet the requirements of UW, and the uncertainty measure of the domain task cannot be compared to the variances of the physical reconstruction branches. As a reminder, the mathematical framework proposed in [KGC18] imposes exclusively the use of Mean Absolute Error, Mean Squared Error and Cross-Entropy. However, the loss objective introduced in DeepCORAL and DeepJDOT are not real likelihood functions and there are no possibilities to connect them directly with such criteria. In conclusion, GN has the advantage of being the most generic compared to UW, but such an algorithm is difficult to implement and requires more compute resources for the calculation of the gradients for each task. Moreover, as will be illustrated shortly, UW has still an important role to play in the convergence of the DANN method.

Impact of conditioning

In this section, we evaluate the impact of the conditioning, and results are depicted in the column named +C of Tables 4.2, 4.3 and 4.4. As highlighted previously in Figure 4.10, this extension has a positive impact of the γ-PhysNet-DANN when combined with UW (-0.4 for the energy resolution and the energy bias, and -0.3 for the angular resolution). Comparing both MTB methods in that context, results are still slightly at the advantage of UW (-0.02 for the energy resolution). Regarding the γ-PhysNet-DeepCORAL, the inclusion of condi-

	DANN		+C		+C+GL	
MTB	UW	GN	UW	GN	UW	GN
E_μ	0.12	0.09	0.08	0.09	**-0.06**	-0.12
E_σ	0.35	0.33	**0.31**	0.33	**0.31**	0.32
θ_σ	0.29	0.27	**0.26**	0.27	0.31	0.31
AUC	**0.83**	0.82	0.82	0.82	0.72	0.77

Table 4.2: Ablation study of γ-PhysNet-DANN. The values are reported for the GN hyper-parameter $\alpha = 0.1$. Only results on target data are provided. +C specifies the conditioning, and +GL designates the use of the Gradient Layer.

tioning improves the energy resolution (-0.6) and the energy bias (-0.7) when used with GN, but drastically deteriorates the metrics with UW (+0.68 on the energy bias). Similar conclusion can be drawn for the γ-PhysNet-DeepJDOT, but the negative impact of this extension is visible on the angular resolution instead (+0.12).

In conclusion, the conditioning has a positive impact for the γ-PhysNet-DeepCORAL and γ-PhysNet-DeepJDOT when coupled with GN, but not with UW. On the contrary, the γ-PhysNet-CDANN has the best results overall compared to the other methods in conjunction with UW.

Impact of the Gradient Layer

We evaluate the benefits of the GL in the domain adaptation models. Results are accessible in the column named *+C+GL* of Tables 4.2, 4.3 and 4.4. It appears that in all cases, weighting the gradients deteriorate the performance, whether it is associated with UW or GN. Moreover, pairing GL with UW in the cases of the γ-PhysNet-CDeepCORAL and γ-PhysNet-CDeepJDOT fails to rectify the convergence issues.

In conclusion, although the GL helps in the converge of DeepJDOT when applied to the digit datasets, as illustrated in Appendix 8.6, is detrimental in the case of LST simulations and can not be considered.

Conclusion

Integrating domain adaptation into multi-task balancing turns out to be a difficult challenge, because of conflicting gradients. We proposed the implementation of the Gradient Layer as a learning rate scheduler to delay the arrival and impact of the considered task, letting the model to train on source before gradually incorporating the contribution of domain adaptation. From our comprehensive study, we have selected the γ-PhysNet-CDANN combined with UW and without Gradient Layer as the most favourable method to be applied to real data.

	DeepCORAL		+C		+C+GL	
MTB	UW	GN	UW	GN	UW	GN
E_μ	0.20	0.19	0.18	**0.13**	0.18	0.17
E_σ	0.44	0.40	1.12	**0.33**	1.13	0.37
θ_σ	0.33	**0.25**	0.45	0.26	0.45	0.27
AUC	**0.82**	**0.82**	0.54	**0.82**	0.59	**0.82**

Table 4.3: Ablation study of γ-PhysNet-DeepCORAL. The values are reported for the GN hyperparameter $\alpha = 0.1$. Only results on target data are provided. +C specifies the conditioning, and +GL designates the use of the Gradient Layer.

	DeepJDOT		+C		+C+GL	
MTB	UW	GN	UW	GN	UW	GN
E_μ	0.55	0.13	-0.14	**0.11**	-0.23	0.19
E_σ	0.86	**0.37**	0.89	**0.37**	0.74	0.41
θ_σ	0.34	0.29	0.45	0.29	0.46	**0.28**
AUC	0.54	**0.81**	0.57	0.80	0.51	**0.81**

Table 4.4: Ablation study of γ-PhysNet-DeepJDOT. The values are reported for the GN hyperparameter $\alpha = 0.1$. Only results on target data are provided. +C specifies the conditioning, and +GL designates the use of the Gradient Layer.

4.8 Application of Vision Transformers to simulations

Presented in the literature as an alternative to traditional CNNs, Transformers have become the new state-of-the-art approaches and general feature extraction encoder for Computer Vision tasks [Liu+21b].

The gamma-ray source analyses in CTAO, whether conducted with the Hillas+RF algorithm or deep learning, are strictly run-wise. In other words, each observation necessitates a specific training and inference. Popular for the scalability they offer, Transformers appear as prominent approach in the search of a global model.

In this section, we aim to compare the reconstruction performance of the γ-PhysNet and its Transformer counterpart. To ensure a fair comparison, the training and test datasets are identical in both cases.

	$N_{\text{gamma-diffuse}}$	N_{gamma}	N_{proton}	N_{electron}	N_{real}
#Training	1116578	-	757561	-	1797361
#Test	-	1057555	756954	1057555	-

Table 4.5: Number of events in each subset.

Pre-training: Application of Masked Auto-Encoder

Training parameters

In our application of MAE, we leverage the hexagonal structure of the data. In that case, no interpolation is applied and inputs correspond directly to the integrated and calibrated images. However, we experimentally observed that pixel charges must be normalized before injection into the network. In fact, due to the great dynamic of the charge values, applying such transformation is crucial to ensure the background is properly reconstructed. Furthermore, as a proof of concept, we consider in this work a small version of the model, and a reduced training dataset composed of a single pointing direction, identical to Sections 4.5 and 4.6. The embedded representation is set to a size of 256, which is half of the size of Transformers mentioned in [Dos+20]. The encoder is made of 4 Transformer blocks, while the decoder is made of 2. In total, the number of heads is set to 8. In that case, our MAE has in total $4.9M$ parameters. The hyperparameters are identical to the implementation of the original MAE paper [He+21]. In summary, the network benefits from two learning rate scheduler: a linear and a cosine annealing. This allows us to have a warm-up rate on the 40 first epochs followed by a decrease. Moreover, we use the AdamW optimizer [LH19], coupled with a learning rate of 5×10^{-5} and a weight decay of 8×10^{-3}. The running average coefficients of the gradients are set to 0.90 and 0.95. The convergence stabilizes within 1000 epochs. Therefore, because training such neural network is time and resource consuming, only one seed is produced.

In order to further apply our Transformer model to the detection of the Crab Nebula, as conducted in the last chapter of this thesis (5), the reconstruction of the inputs is performed on both synthetic and real unlabelled data. In total, the training and validation sets contain respectively $3.6M$ and $0.4M$ of images, which Table 4.5 gives a more precise composition.

Results

In this section, we provide a qualitative analysis of the results obtained by the application of MAE to the LST images. The initial, the masked and the reconstructed synthetic gamma images are depicted adjacently on the same row in Figure 4.11.

The first row displays a gamma, which fingerprint is masked in either side, except for the centre of the ellipsoid. Only one module of signal remains active. It is interesting to observe that, despite the frugality of information, the network managed with good precision to retrieve the shape and direction of the signal. Yet, two comments can be drawn. The elongation seems greater than the true shape, and the bright pixel has been averaged.

The second row exhibits an expected behaviour. The entire gamma has been occluded by the mask, thus the model cannot reconstruct any signal.

We can attribute to the third and fourth rows the same observation as the first row. The brighter pixel has been diminished in intensity and the network managed to recover from the morphology rather precisely.

Lastly, the fifth row presents some encouraging results. Although most of the modules are switched off, the tail of the ellipsoid remains available, which allows the model for understanding the presence of a gamma fingerprint. The direction is recovered rather accurately.

In conclusion, we can draw these following remarks from the reconstructed results:

- The background is mostly averaged, and the variations are not as strong as the labels. This is actually a common issue with auto-encoders. In the case of synthetic images, as described precisely previously in Section 3.1, the NSB is simulated as a Poisson noise, and each pixel is considered as an independent variable. Such an approach thus tends to reconstruct the mean value rather than pixel-level variations.
- If the signal contains bright pixels, their values are reduced.
- If the entirety of the signal is occluded, there is no recovery.

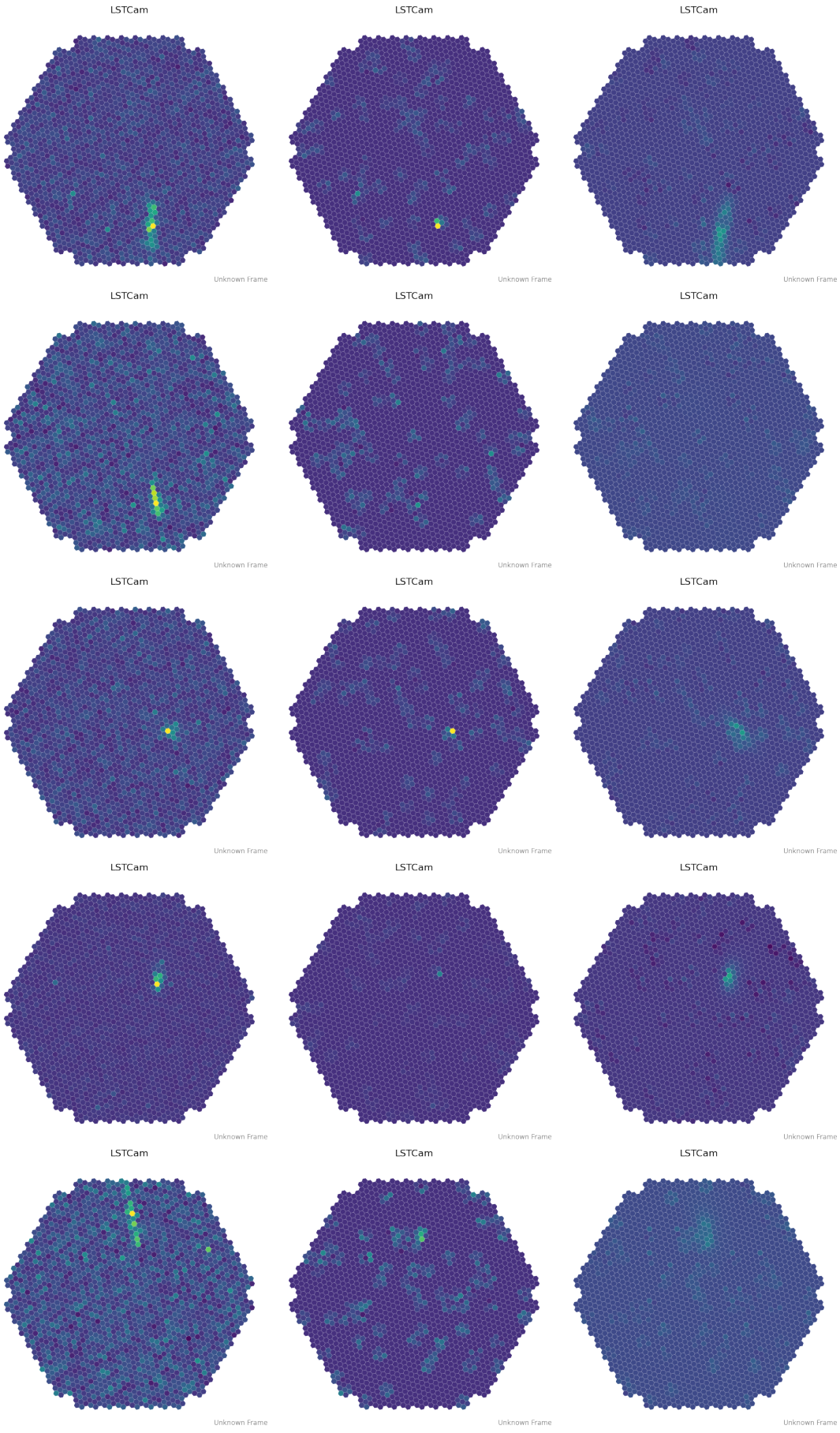


Figure 4.11: Application of the MAE for the reconstruction of the gamma events. From left to right: the initial images, the masked images and the reconstructed images (the scale of the images is identical).

Fine-tuning: Application of the γ-PhysNet-Prime

Once the MAE unsupervised pre-training is completed, the model can be specialized on a defined set of targets. The MAE optimized weights then become the initial starting point for the following optimization. In the case of the particle reconstruction, these new tasks correspond to the one expressed in Figure 2.9, that are the incident energy, direction, impact point and particle type. As a consequence of the difficult pre-training auto-encoding process, the γ-PhysNet-Prime possesses a strong ability to recover missing information from a few remaining hardware modules. The next step, which consists in fine-tuning to the desired objectives, can therefore leverage this designed property.

Training parameters

Because we are no longer interested in reconstructing the input images, the decoder is removed from the architecture, and four branches are integrated, connected to the latent space, in order to retrieve the incident particle attributes. Overall, it contains $3.2M$ parameters. For this training procedure, most of the hyperparameters are similar to the γ-PhysNet. In particular, we use the UW task balancing algorithm, and the optimizers are Adam for both the main model and the log-variances. In the first case, the learning is set to 10^{-3}, whereas it is set to 0.025 in the second. Both have a weight decay of 10^{-4}. The number of epochs is set to 500 as the model tends to overfit after such amount, and the batch size to 256. The learning rate scheduling is identical to MAE, with the only difference that the warm-up is performed on the 5 first epochs only.

Results

In the section, the γ-PhysNet-Prime is applied to both perturbed and unperturbed simulations. In the former, the noise corresponds to the worst scenario, with a rate of $\delta\lambda = 0.46$. This allows us to compare the model to the vanilla γ-PhysNet, and evaluate if the integration of real images in the training set of the MAE somehow improves the results in a degraded scenario. The results are depicted in Figure 4.12. We refer to the *γ-PhysNet (best)* and the *γ-PhysNet-Prime (best)* when no domain shift is set, and to the γ-PhysNet-Prime (worst) and the γ-PhysNet-Prime (worst) when a domain shift is applied.

Comparing both best cases, it appears that the γ-PhysNet-Prime remarkably outperforms our baseline on the energy resolution metric for energies greater than $5 \times 10^{-}2$ TeV. At the highest energy level, the gain is almost of 0.1 of the relative error, that is to say a considerable improvement of 33%. Similarly, the angular resolution is greatly improved for energy greater than 0.1 TeV. In that case, the enhancement reaches around 25%, from a value of 0.2 at the highest for the baseline, and 0.15 for the Transformer. Finally, the same performance is observed on the other figures of merit.

Regarding the worst scenario, we observed ultimately a decrease in performance compared to the best one. However, the model remains exceptionally good on the angular resolution and classification power, although a slight loss is observed at the lowest energy level. However, it suffers from a strong energy bias, and the energy resolution decreases strongly for energy levels greater than 0.1 TeV. Yet, in this scenario, the Transformer remains either in the error bars or is much better than the baseline in degraded condition.

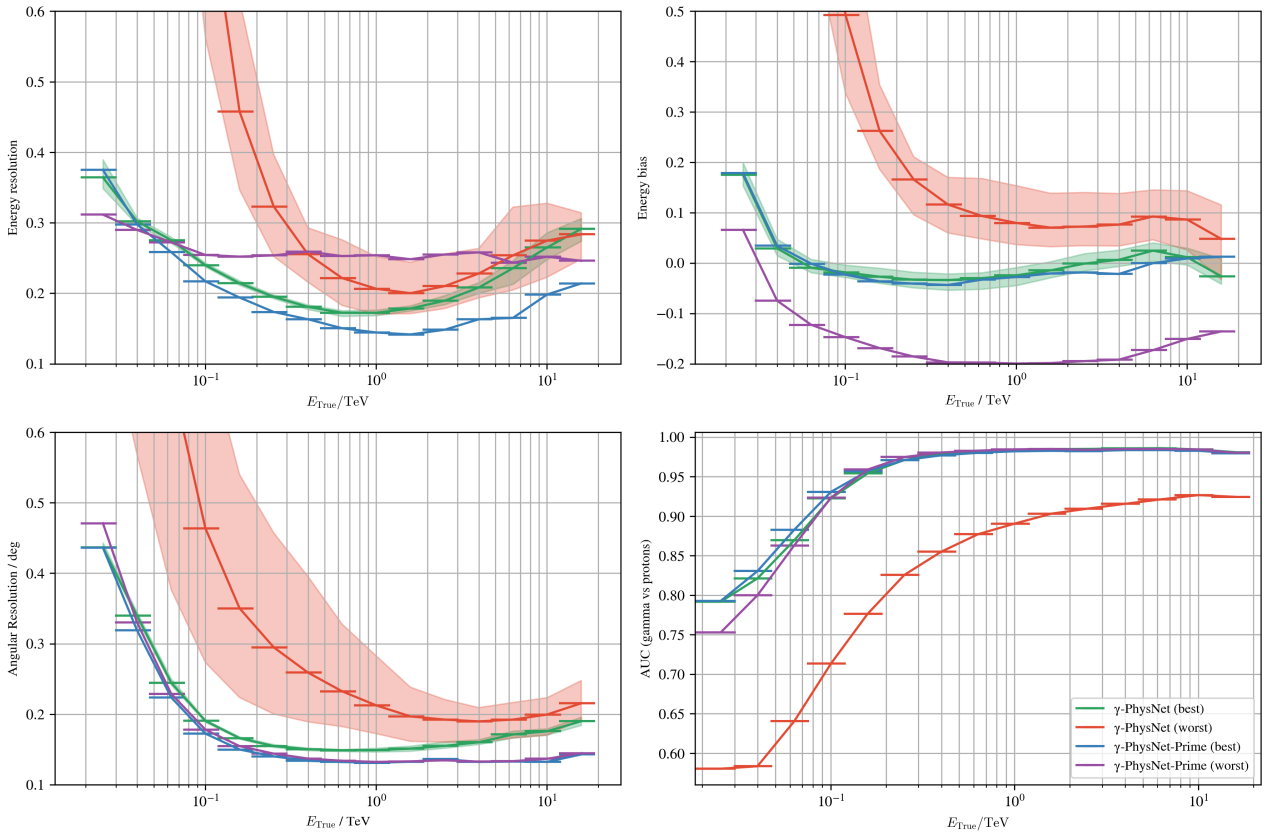


Figure 4.12: Application of the γ-PhysNet-Prime to simulations.

In conclusion, the γ-PhysNet-Prime considerably improves the current state-of-the-art on simulated data. As expected, degraded data introduces a loss in performance, but is remarkably mitigated compared to the γ-PhysNet in similar conditions. Thus, unsupervised domain adaptation has still a role to play in enhancing the performance of the Transformer model, and its results are introduced in the next section. Finally, more experiments must be conducted in order to qualify the model sensitivity to initial parametrization, in agreement to our previous contributions with CBN modules and domain adaptation.

Fine-tuning: Application of the γ-PhysNet-Megatron

Previous section has highlighted that, although ViT-based models allow to considerably increase the performance in both the ideal and degraded conditions, domain shift still introduces a gap in performance which we aim to resolve. In order to improve model adaptation capability, we consider the γ-PhysNet-Megatron, that explicitly adds a domain adaptation task based on DANN.

Training parameters

The hyperparametrization of the γ-PhysNet-Megatron is identical to the γ-PhysNet-Prime. Also, because of the addition of the domain task, the model has roughly 1k extra parameters, coming from the domain token, the domain branch in the multitask architecture, and the encoder. As a first step, the evaluation framework consists in applying a Poisson noise of rate $\delta\lambda = 0.46$ in the target, without label shift.

Results

Because this is the most recent part of the work introduced in this thesis, only preliminary results can be discussed and further experiments are planned in short perspectives. Preliminary results display a peculiar behaviour, highlighting some convergence issue of the domain task. In fact, we noticed that the domain classifier constantly managed to fully discriminate the source and target, leading to an accuracy of 100%, despite the inclusion of the Gradient Reversal Layer. However, there are no signs of improvements in the training metrics. Future works should analyse the gradient norms and directions of each task to possibly highlight conflicting tasks. Further investigations are ongoing to solve this problem, but the application of the γ-PhysNet-Megatron to the real telescope acquisitions will unfortunately not materialize in this thesis. Finally, because the training time of our Transformer models is a strong limitation, more frugal approaches could be considered.

Conclusions on the Transformer-based approaches

We introduced the first Transformer models for CTAO event reconstruction. For this purpose, we leverage a pre-training based on MAE with real and simulated data. With such strategy, the network learns robust features and relationship, and serves as a performing starting point for a following fine-tuning. In the application to the synthetic dataset, the γ-PhysNet-Prime remarkably outperforms the current CNN-based γ-PhysNet, especially at high energy. This lays the foundation for future work to improve the reconstruction capabilities for two main reasons. Firstly, there exists a plethora of techniques to enhance our first draft, with other more physics-based pre-training as described in Section 3.6 of the previous chapter, for example. Lastly, our Transformer model has the capacity to become the first all-sky neural network. Considering that simulations and telescope observations are available at many distinct pointings, a global approach can leverage any possible dataset in a single training, although the computational time will drastically increase if no further refinements in the model architecture are proposed in the current context.

4.9 Comparison of γ-PhysNet and γ-PhysNet-Prime with the standard analysis Hillas+RF

We propose to compare our contributions with Hillas+RF in the controlled environment of simulations. As explained in Section 2.5 of Chapter 2, the standard analysis necessitates a cleaning procedure that is reminded in the following:

- Each pixel is sequentially analysed. If the current pixel value is greater than a pre-defined threshold, then its neighbours are evaluated.
- If at least one neighbour has a value greater than a second pre-defined threshold, both the current central pixel and its neighbour are considered as part of the signal.
- The pixels that don't pass the two-threshold test are considered as background and are discarded.

In the case of neural networks, the entire input data is usually considered, and it is left to the model to consider if the background contains useful information or not. Especially, the

thesis work in [Jac20] analysed the results of the GradCam algorithm in the reconstruction of simulated particle parameters. It appears that, although most of the attention is inclined towards the signal, in both the pixel charge and the temporal map, a part of the answer is also focused on the background noise.

Training parameters

In the case of the standard analysis, Hillas+RF, the hyperparameters have been optimized by the LST collaboration to obtain the best performance in a reasonable training time[2]. Moreover, only one seed has been produced. On the other hand, regarding the γ-PhysNet and the γ-PhysNet-Prime, the hyperparameters are identical to the previous experiments, respectively mentioned in Sections 4.4 and 4.8. The training dataset is identical for all methods. However, because the standard analysis leverage a cleaning procedure, the events that do not pass the cleaning steps are not considered in the training. Consequently, the RFs are optimized on a slightly smaller dataset compared to neural networks.

Results

The results of the comparison are depicted in Figure 4.13. As expected, and in accordance to the previous results published in [Vui+21], the domination of the deep learning architectures is particularly pronounced at the lower energy levels on all metrics. Regarding the bias and resolutions at the first energy bin, they are divided by a factor 2 in favour of our neural networks. Concerning the energy and angular resolution, the γ-PhysNet and its Prime version exhibit superior performance up to 1 TeV. The classification power is totally favorable to our neural network, with a noticeable gain of 10 points at the lowest bin. However, Hillas+RF demonstrates its efficiency on the energy resolution above 1 TeV and outperforms the γ-PhysNet with almost 0.1 of energy resolution error. Nevertheless, this performance is retrieved our Transformer model, which overall exhibits the best results across all metrics and all energy ranges.

4.10 Conclusion

In this chapter, we proposed two novel approaches to take into account domain discrepancies. On the one hand, input conditioning, which aims to ensure robustness towards specifically selected quantities of interest. On the other hand, domain adaptation, which takes into account known and unknown domain shifts to engender domain-invariant features. In addition, the new training procedure designated as conditional domain adaptation allows to compensate for the inherent label shift present between synthetic sources and real targets.

Considering explicitly the variation of NSB in the case of synthetic data, γ-PhysNet-CBN offers the best results on all metrics compared to domain adaptation. Unsupervised domain adaptation theoretically has better properties than information fusion, because unknown variation are also considered. Yet, the training of such models is difficult, and an in-depth

[2]The hyperparameters can be found here: https://github.com/cta-observatory/cta-lstchain/blob/main/lstchain/data/lstchain_standard_config.json

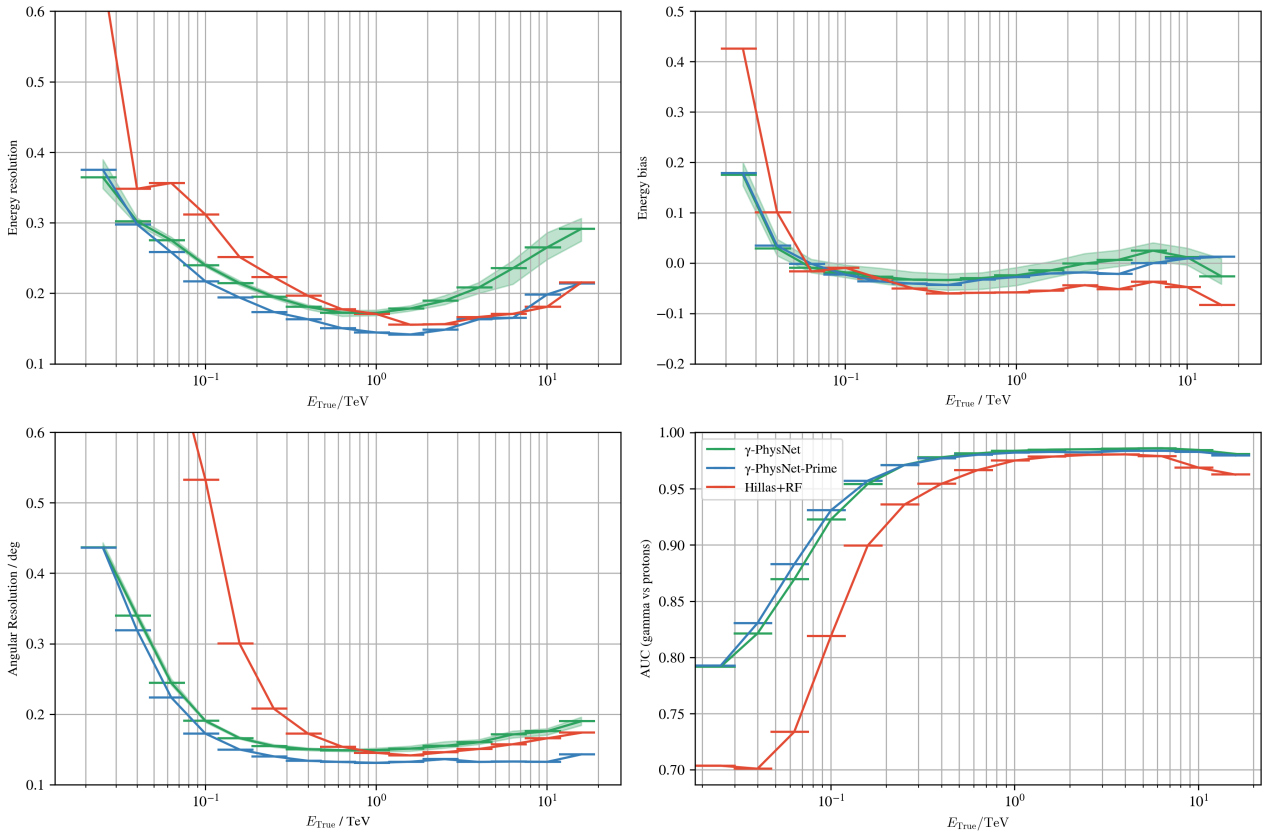


Figure 4.13: Caption

optimization study must be conducted to help domain adaptation provide its best potential. However, as NSB has been designated as one of the main source of discrepancies, precisely tackling this problem through CBN modules may provide acceptable improvements to the astrophysicist community when applied on real data.

Furthermore, our proposed γ-PhysNet-Prime Transformer exhibits the best results compared to each studied methods in the specific case of simulated data within the context of the best scenario. Especially, the integration of real data into the reconstruction task of the pre-training based on MAE has surely a role to play in the performance on the target set. Remarkably, comparing the worst scenario of our Transformer and the γ-PhysNet, the γ-PhysNet-Prime has shown a significant improvement at the lowest energies. Without further refinement, it may already be a very promising approach for the detection of the Crab Nebula, without the need for data adaptation.

Finally, from our ablation study, we have selected γ-PhysNet-CDANN as the most promising approach for the detection of the Crab gamma-ray source in the next chapter.

5 Detection of the Crab nebula

In the previous chapters, we introduced gamma-ray astronomy and the fundamental questions that CTAO will try to answer. Then, we talked about how deep learning can solve the current limitations of the standard analysis and how it is possible to integrate real data into the training process in the domain adaptation framework to minimize the domain shift between the training and test distributions. Finally, we selected, implemented and validated domain adaptation and information fusion methods and evaluated then on a controlled environments offered by simulations. Henceforth, this chapter aims to assess our techniques in real situations. Our goal is the detection of the Crab Nebula, a standard candle with a well-known spectrum, and compare them to the standard analysis and the current γ-PhysNet algorithm.

5.1 Introduction

The deployment of deep learning methods for the reconstruction of physical attributes of incident particles has evinced promising outcomes when conducted on simulations. However, the transition of these approaches to observational data is accompanied by challenges, as deep learning-based models are susceptible to domain shifts. In this chapter, we integrate the γ-PhysNet-CBN, the γ-PhysNet-DANN, its conditional version and the γ-PhysNet-Prime in the physics-based context of the CTAO for the detection of the Crab Nebula. We compare their detection performance with the baselines, the γ-PhysNet and standard analysis Hillas+RF methods.

Conjointly observed for the first time in 1054 by Chinese and Japanese astronomers, the Crab Nebula is a supernova remnant located in the constellation Taurus at approximately 6,500 light years from Earth. This "guest star", as were denominated ephemeral apparitions of bright celestial events in the former Chinese culture, was described as "visible in the daytime like Venus" and direct observations from the naked eyes was possible for two consecutive years [SG03]. It is one of the most extensively studied sources of cosmic rays and was historically the first source detected with IACT. At its heart lies the Crab Pulsar emitting radiation across a broad spectral range from radio waves to gamma rays. Nowadays, the Crab Nebula serves as a space laboratory for astrophysicists, offering valuable insights into the processes governing supernova explosions, pulsar mechanics, and particle acceleration. Its study has significantly contributed to our understanding of high-energy astrophysics and the lifecycle of massive stars.

5.2 The Crab dataset

We consider the Crab dataset that consists in four consecutive runs of observations numbered as 6892, 6893, 6894 and 6895 and processed with lstchain $v0.9.1$. The selection of this dataset is based on the diversity of conditions they offer, allowing to evaluate our techniques with different level of complexity. They respectively contains $9,5$M, $9,2$M, 8.6M and 8.5M of integrated and calibrated images. In more detail, Table 5.1 gives a macro vision of two important attributes to characterize the dataset that are the NSB and the zenith angle of each run. Furthermore, we provide the corresponding Poisson rate shift $\delta\lambda = \lambda_{real} - \lambda_{MC}$, between the rate calculated from pedestal images λ_{real}, and the one estimated from the training simulations λ_{MC}, described in Section 4.2 of Chapter 4. The NSB rate is calculated in this case as the mean value across all pixels from the pedestal images contained within each run.

Pedestals

Pedestals are conjointly acquired with source data at a sample 100 Hz from a signal-free region and contain valuable information to characterize the background during observations. An example of pedestal is presented in Figure 5.1, along with a λ-map as described in Section 3.3 of Chapter 3. Important metrics can be computed from such data. The mean image reveals for example the presence of stars in the field of views, which correspond to correlated pixels of higher values compared to the background. In that case, the assumption that each pixel of the camera is independent becomes inadequate.

Run number	6892	6893	6894	6895
Zenith angle (deg)	14.0	18.1	25.7	30.2
NSB λ_{real} (pe)	2.46	2.35	2.07	2.07
Poisson rate difference $\delta\lambda$	0.70	0.59	0.31	0.31
# Sub-runs	180	174	163	161
# Images (M)	9.5	9.2	8.6	8.5

Table 5.1: Zenith angle and NSB value of each run of the Crab observation.

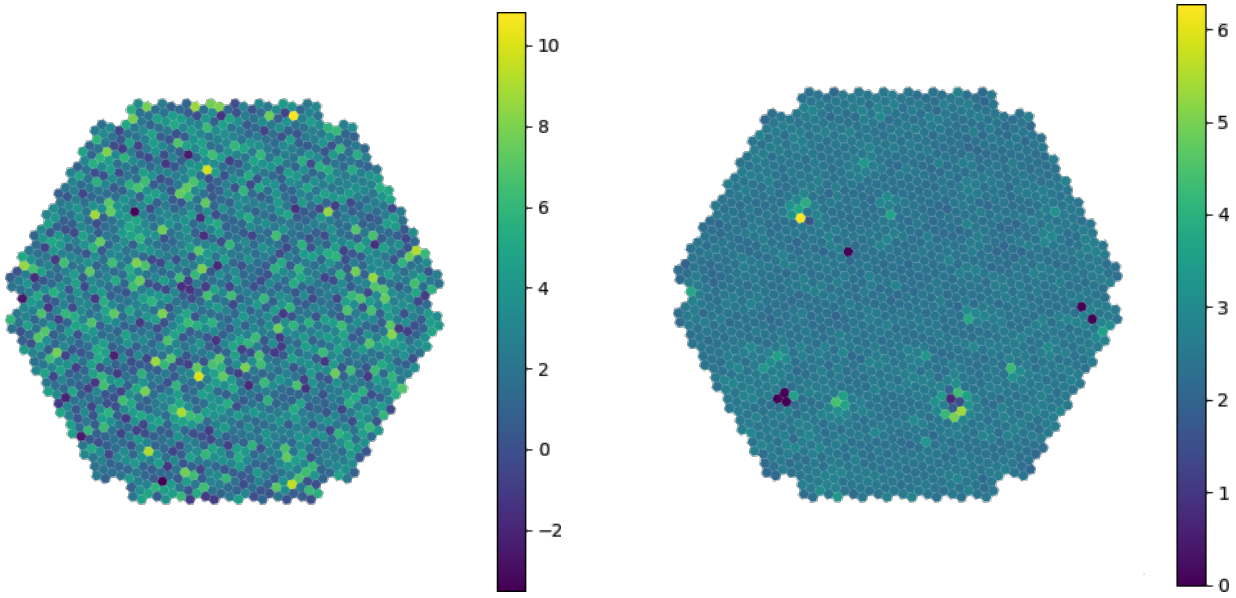


Figure 5.1: A single pedestal image at time t (left), and a pedestal image averaged on 1 second, corresponding to averaging 627 pedestal images contained within one sub-run (right). Under the Poisson hypothesis, each pixel corresponds to an estimation of Poisson rate.

Night Sky Background characterization

The NSB, as presented in the macro Table 5.1 is calculated from averaging spatially and temporally (within a sub-run) each pixel of the pedestals extracted from each run.

The decreasing evolution of the mean NSB reveals the impact of the moonlight during data acquisitions. The first two runs (6892 and 6893) are categorized as presenting moonlight, whereas the two last runs (6894 and 6895) are classified as not affected by this factor. In the former, the level of complexity for the particle reconstruction is greater as the signal-to-noise ratio is less favorable. The mean of the Poisson rate difference $\delta\lambda$ across the whole dataset gives the rate that has been used in the previous chapter $\delta\lambda = \frac{1}{4}\sum_{\text{run}} \delta\lambda_{\text{run}} \approx 0.46$.

A more detailed overview of the evolution of the NSB is given is Figure 5.2. This time, the plot shows the estimated Poisson rate computed from the pedestals of each sub-run as of function of the time. A sub-run corresponds to 53k events, roughly 6 or 7 seconds of data depending on the trigger rate of the telescope. Interestingly, the second run denoted as 6893 is affected twice by light spikes, lasting a couple of seconds, engendering a great value

of NSB rate. From the daily check logs, the spikes corresponds to a trigger rate of 14000 Hz, while the rest of the observations is around 6 – 8000 Hz. They can be attributed to car flashes of vehicles passing by during the observations.

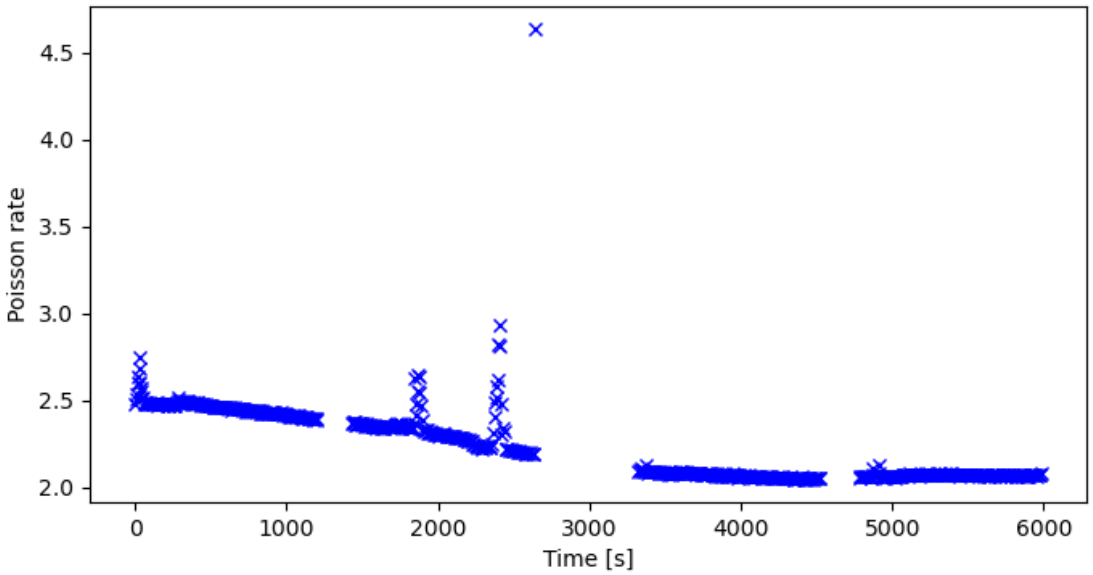


Figure 5.2: Evolution of the NSB rate during the Crab observation under the Poissonian assumption. The rate is computed as the mean value across all the pedestal pixels of a sub-run. Car flashes during the data acquisition are probably responsible for the spikes. The four clusters corresponds to each run.

Evolution of the zenith angle

On the other hand, the zenith angle changes over time as the telescope follows the source, as illustrated in Figure 5.3. Its value must be put in perspective with the value of the training simulations that only contains images synthesized at an angle of 20 degrees. In our case, the angle is increasing twofold from the beginning to the end of the acquisitions. In a nutshell, it translates into the telescope pointing more and more towards the horizon.

5.3 Figures of merit

In the section, we describe the figures of merit that are classically used for the analysis of real data. They differ from the one defined in Section 4.3 of the last chapter which are based on synthetic images and leverage labels that are not available in this context. Therefore, Figure 5.4 gives an overview of the necessary metrics.

Background and number of excess estimations

The estimation of the background noise is a crucial step in detecting gamma-ray sources. In fact, except for the brightest gamma-ray sources, these noise fluctuations completely dominate the signal of interest.

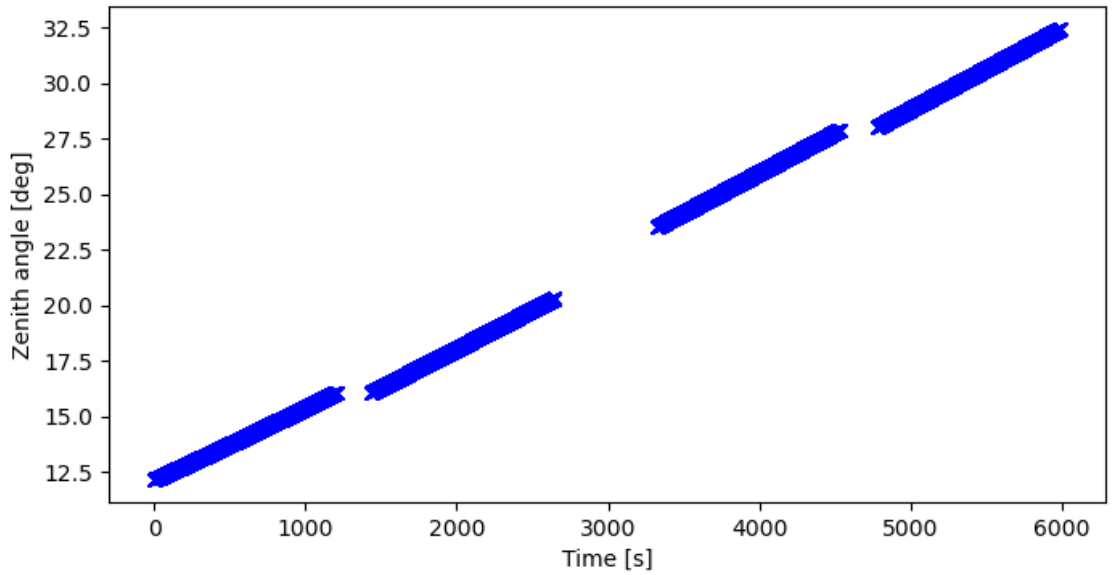


Figure 5.3: Evolution of the zenith angle during the considered observations. Each cross corresponds to the average zenith value of a sub-run. They are divided into four clusters, in agreement with the four runs that are used.

In this work, we use the multiple *OFF* method for the estimation of the number of gamma excess. As illustrated in Figure 5.5, this method involves observing the single *ON* region (the region of the sky where the source(s) may be located) and multiple *OFF* regions (dark patches that contain only background). Observations are conducted in wobble mode, meaning the telescope is offset from the center and not pointing directly at the source.

The number of particles detected in the *ON* and *OFF* regions are respectively denoted N_{ON} and N_{OFF}. The number of background events is supposed to be constant across all regions, then the excess of gamma detected in the *ON* region can be computed as:

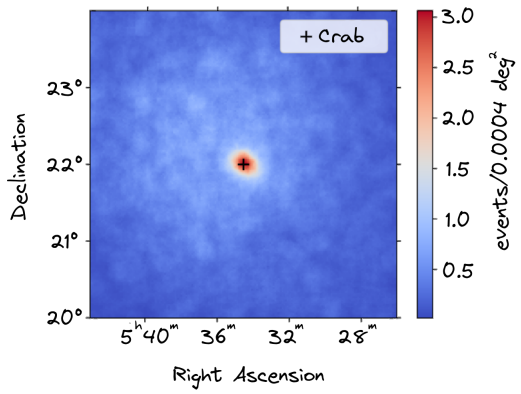
$$N_{excess} = N_{ON} - \frac{N_{OFF}}{\alpha} \tag{5.1}$$

where α is the number of *OFF* regions.

Although the multiple OFF technique is remarkably efficient and simple, it cannot be applied uniformly on the entire field of view of the camera. In that case, the ring method is preferred, which illustration is given in Figure 5.5. Nevertheless, due to the heterogeneity of the sensitivity, there is a need to integrate the notion of correction factor, also called the acceptance, between the ring and the ON region. Its calculation necessitates to create a specific background model. The creation of the significance map, described in Figure 5.4, makes use of ring technique. More details are available in the thesis work of [Bon23].

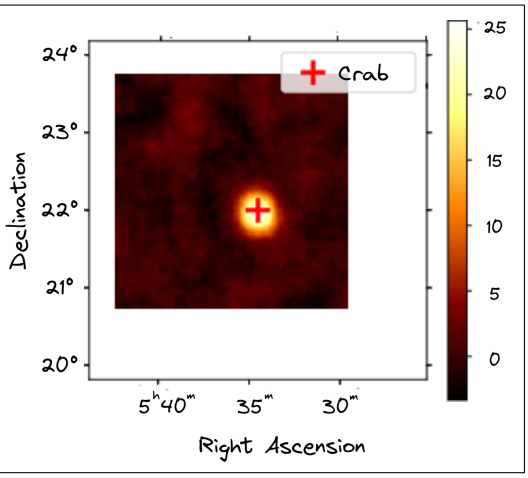
Event selection

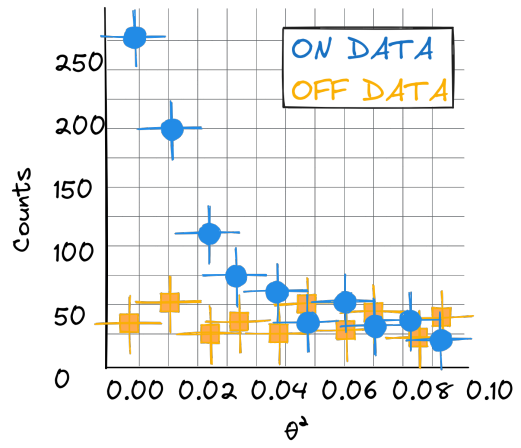
When used in inference mode, each event from the telescope observation dataset is reconstructed regardless its type, whether it is a gamma or not. At this stage, there is no strict manner to discriminate the particles. Furthermore, as illustrated in the introduction of this thesis, in Figure 1.2, gammas account for a small fraction of the incident particle flux. In



The **count map**, or **excess map**, is a graphical representation of the number of gamma-ray in excess per unit of surface area. When a source is detected, a characteristic excess must be spatially localized in the image, according to the corrected point spread function of the telescope. On the other hand, the background must be rather uniform. The surface area corresponds to a value of $S = 0.02^2 = 0.0004$ deg^2.

The **significance map** is another graphical representation displaying the significance calculated at each pixel. From the significance map is calculated the **position shift**, or **bias** of the reconstructed source compared to the official coordinates. The maximum value is simply the local peaks in the map above a specified threshold within a radius around each given pixel.





The θ^2 **plot** is a 1-D representation of the number of detected gamma particles in the ON and OFF regions as a function of the surface unit θ^2. The center of the source corresponds to $\theta^2 = 0$, and the distance to the source increases as θ^2 increases. Because of the point-wise localization of the source, both ON and OFF lines should merge for greater values of θ^2. The ON counts graph allows to visualize the angular resolution, which grows as both curve fusion late.

Figure 5.4: Figures of merit for the analysis of the model performance when applied to real data.

order to reject hadronic-initiated showers as most as possible, an event selection procedure based on cuts must be established.

Described in the last chapter, in Section 4.3, the gammaness corresponds to the probability that the current event is a gamma. The higher the value, the more likely a cosmic-ray is eliminated, but high values ultimately discard a significant amount of signal as well. There is therefore a trade-off regarding the quantity of particles that one desires to retain or reject. It is traditionally referred to as the specificity/sensitivity trade-off and is usually depicted as a ROC curve as described in 4.3 of the previous chapter. In our case, we select the gammaness that maximize a specific quantity, the significance, described in the next section.

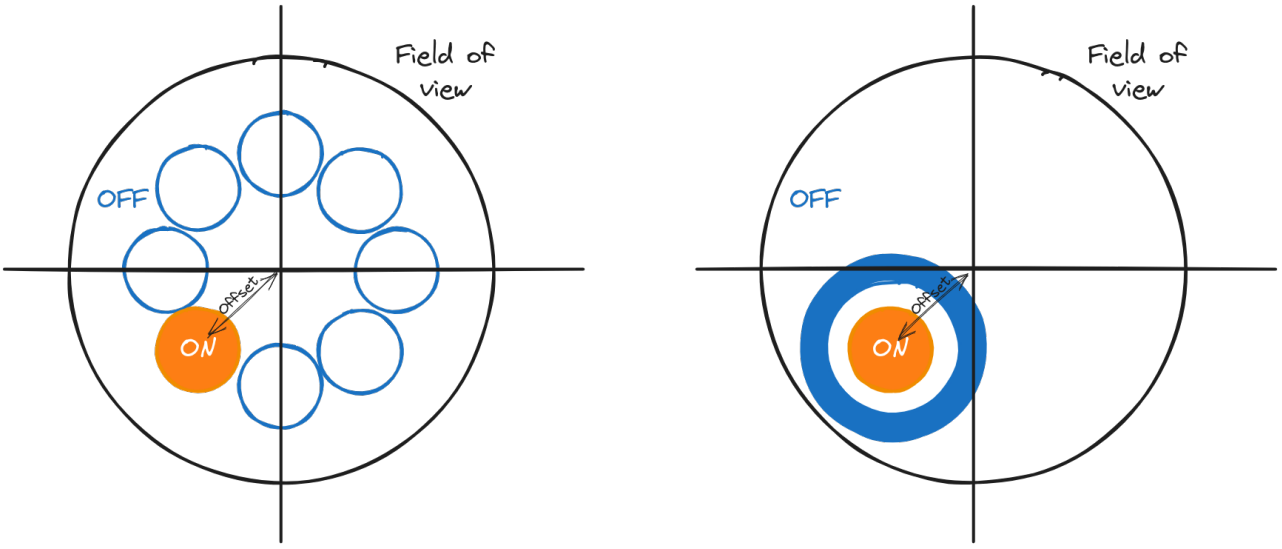


Figure 5.5: Background estimation using the multiple OFF (left) or the ring technique (right). Inspired from [Che17].

Significance

The statistical significance of an observation is crucial to determine whether an astronomical source has been detected by providing a probability that the observed excess N_{excess} is not merely due to background fluctuations. Introduced in [LM83], the *li&ma* significance estimation technique, derived from maximum likelihood principles, is the currently applied approximation method in LST acquisition analysis. It characterizes the probability that a gamma ray comes from a gamma-ray source rather than from the background noise. Based on hypothesis testing, the null hypothesis H_0 stipulates that all the observed events from the *ON* region are background signals and no sources are detected; thus $N_{excess} = 0$. The H_1 hypothesis, on the other hand, indicates the presence of a source in the *ON* region. The likelihood ratio λ is therefore defined as:

$$\lambda = \frac{L[X|H_1(N_{excess})]}{L[X|H_0(N_{excess})]} \tag{5.2}$$

For a great enough number of observations, if H_0 is true then $-2 \ln \lambda$ follows a χ^2 distribution with one degree of freedom (as H_0 only varies according to the N_{excess}). The significance is therefore calculated as:

$$S = \sqrt{-2 \ln \lambda} \tag{5.3}$$

Finally, a source is detected when the significance exceeds 5σ, which is equivalent to have a probability of $p < 5.97 \times 10^{-7}$ that the excess is caused by a fluctuation of the background and not from the gamma-ray source.

5.4 High-level analysis

In this section, we describe the high-level analysis that allows to create the necessary figures of merit and further conclude on the possible detection of a gamma-ray source.

We provided, in Sections 1.2 and 1.3 of Chapter 1, a description of the IACT based particle shower mechanism, and we dispense a concise summary in the following. When cosmic

particles penetrate the atmosphere, they generate extensive atmospheric showers that emit a Cherenkov light. If a IACT ground-based observatory is positioned within the cone of light, photons are captured by the telescope optical system. At a rate of 2−10kHz, the LST-1 trigger system records incoming events as sequences of 40 snapshots known as waveforms. These videos are calibrated and integrated into pixel charges and temporal maps, that are classically used as inputs of machine and deep learning models. This operation is performed with the *lstchain* software [Lop+22]. In the case of the γ-PhysNet, we perform from the calibrated and integrated images a full-event reconstruction of the energy, direction and type (whether it is a gamma or not) using the GammaLearn framework [Jac20]. The next step consists in evaluating the detection quality of the observed source by producing the corresponding figures of merit. To this end, we utilize GammaScan, an analysis software jointly developed by Cyann Plard and myself. The global framework is illustrated in Figure 5.6.

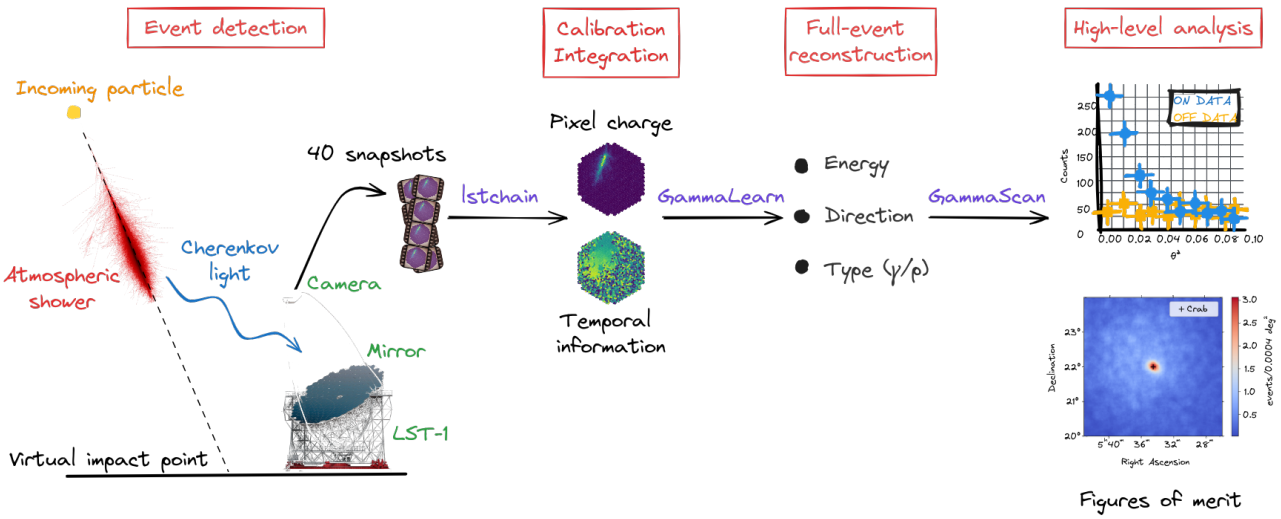


Figure 5.6: The detection and analysis workflow. The lstchain software computes the calibrated and integrated images that input the neural networks. Our framework, GammaLearn, generate the reconstructed parameters, which are used by GammaScan to create the corresponding IRFs and figures of merit.

GammaScan is the final link of the detection workflow, and inputs the reconstructed events generated either with GammaLearn or the standard analysis. In the specific case of the Crab Nebula detection, Figure 5.7 describes the optimization and analysis procedures applied on each run. GammaScan is used in this chapter for the calculation of the figures of merits for each mentioned method.

As explained previously, the significance can be computed as a function of the number of detected gamma N_{ON} and the number of particle classified as background N_{OFF} from which is derived the number of gammas in excess N_{excess}. Yet, the final proportion of each particle type depends on two aspects:

- The gammaness, that is to say the chosen cut value indicating the confidence of the model to classify the particle as a gamma. This is a classical optimization when considering a classifier, known as the sensitivity/specificity trade-off. Therefore, the gammaness must be optimized.

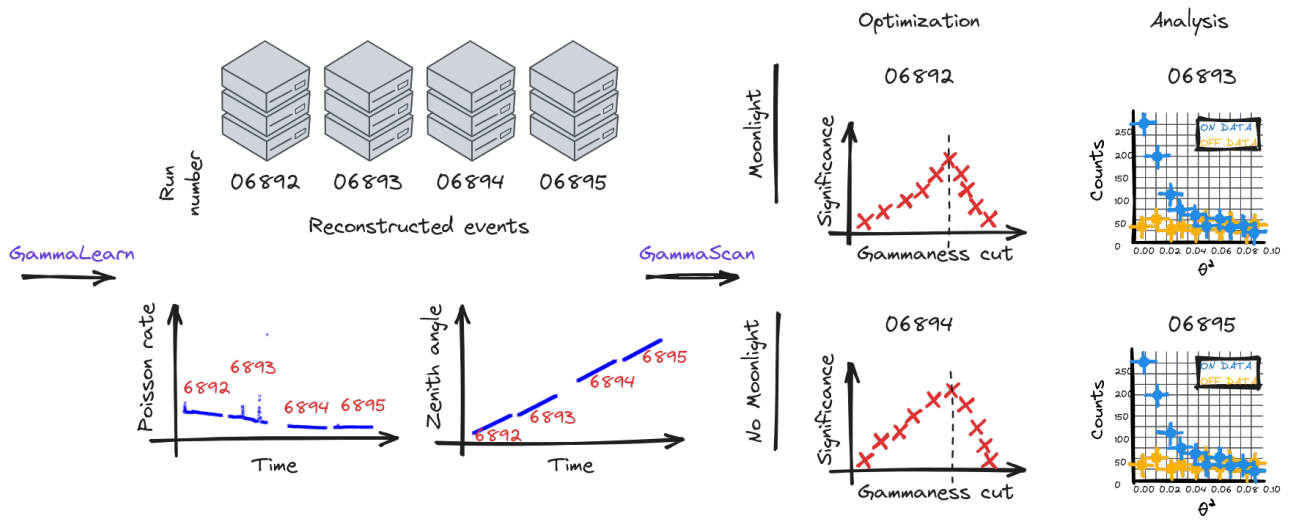


Figure 5.7: The analysis workflow. Each scenario, moonlight and no moonlight, contains two runs. In both cases, the optimization is processed on one run, and the analysis on the other one. However, it is important to take into consideration that, in our case, we have a limited amount of data, and each run might contain some variations, such as NSB and zenith angles amongst others.

- The θ containment, which corresponds to the size of the ON region. In that case, a region that is smaller than the source will miss a part of the signal, while a region that is bigger will contain more noise than necessary. Thus, there is a need to find the optimal θ^2. However, throughout the analysis we are conducting in this work, this optimization is not taken into account because of the lack of input data. Considering this part necessitates more statistics to have a significant influence and is let for future work.

As a result, the optimization step aims to find the set of cuts, gammaness and θ^2 that outputs the highest significance value. On the other hand, the analysis step computes the figures of merit based on the set of optimized cuts.

5.5 Application of Hillas+RF for the detection of the Crab Nebula

Training hyperparameters

The training hyperparameters of the Hillas+RF are identical to the one described in Section 4.9 of Chapter 4.

Results of Hillas+RF

The IRFs of the standard analysis are provided in Section 4.9 of the last chapter in Figure 4.13. As a reminder, although Hillas+RF gives the best results compared to the γ-PhysNet at the upper energy ranges, the neural network outperforms significantly the RFs at the lower ranges. Yet, low energies are where most of the Crab flux lies, as described in [Abe+23].

In accordance to the previous study in [Jac20], we can expect the same conclusions in the following.

The standard analysis constitutes the first baseline to compare our contributions. Its results are presented in Figures 5.12 and 5.13, in both moonlight (6892 and 6893) and no moonlight conditions (6894 and 6895). Notably, the standard analysis is resilient to perturbations of NSB. In fact, in both cases the contribution of data adaptation is not really discernible, with $+1.2\sigma$ in the moonlight condition and similar results otherwise. Regarding the reconstruction of the source position, it seems that the RFs trained on the standard dataset exhibit a better performance, although the selection of the displayed results is based on the best model initialization seed in terms of significance.

5.6 Application of the γ-PhysNet for the detection of the Crab Nebula

In this section, we shed light on the performance of the γ-PhysNet with and without data adaptation. As highlighted in previous work of [Vui+21], the background matching between the training simulations and the real Crab observations has shown an improvement in performance on the detection capabilities, increasing the number of detected events and the significance value. Although the analysis of [Vui+21] is performed on another Crab dataset, the same conclusions are expected in our case. Along with Hillas+RF, it constitutes the second method of the baseline to compare the other models.

Training hyperparameters

The training hyperparameters of the γ-PhysNet are identical to the one described in Section 4.4 of Chapter 4.

Results of the γ-PhysNet

Before assessing the performance in the detection of the Crab, we provide the IRFs evaluating the reconstruction capabilities of the γ-PhysNet with and without data adaptation in Figure 5.8. The former has been presented in the last chapter as the *best scenario*, while the latter corresponds to the best scenario in presence of a constant perturbation, with a rate $\lambda_{MC} = 0.46$, as computed in Section 5.2. It is interesting to notice that training and evaluating the model on degraded simulations yields the same IRF metrics as the *best scenario*. In fact, the introduction of noise didn't suppress the signal at the lowest energy levels and reconstructing the corresponding particle parameters is still possible.

Firstly, we present the results of the γ-PhysNet in moonlight conditions (runs 6892 and 6893) in Figure 5.14, and in no moonlight conditions (6894 and 6895) in Figure 5.15. In both cases, the left figures correspond to the application without data adaptation, and the right figures with data adaptation.

Secondly, we compare the count maps and θ^2 plots in Figure 5.14, reflecting the moonlight condition. In agreement with previous studies in [Jac20] and [Bon23], it is clear that data adaptation significantly enhance the performance of the neural network. In fact, much more excess gammas are detected (×9 more detection), and the significance value almost

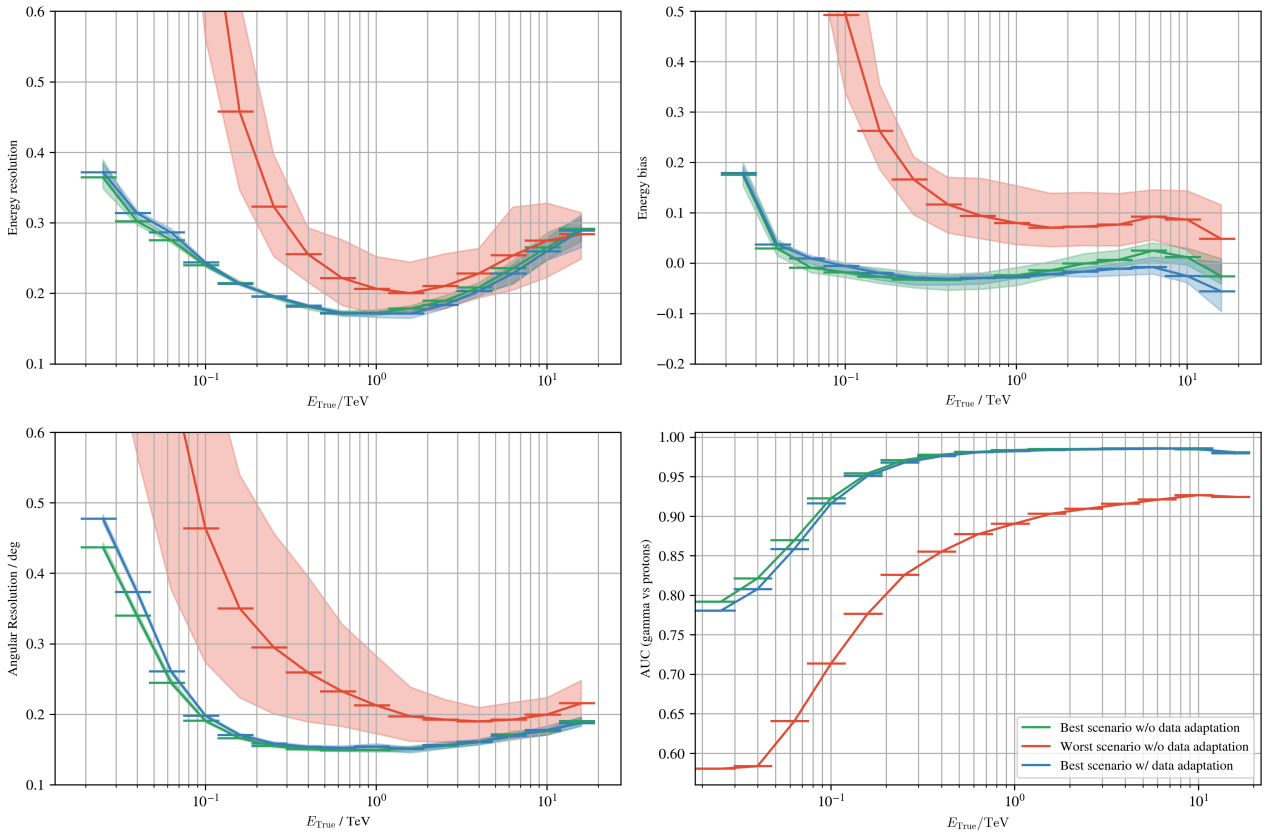


Figure 5.8: Performance of the γ-PhysNet with and without data adaptation. The green and red curves correspond respectively to the best and worst scenarios, as described in Section 4.3 of Chapter 4. The blue line on the other hand refers to the best scenario but trained and tested on degraded images.

doubles. Regarding the significance map, although the position of the source was not properly reconstructed in the first case, training on the tuned dataset allowed to divide the shift by a factor of 3. Moreover, the θ^2 plot shows an improvement in the angular resolution, with a convergence of the ON and OFF lines occurring at around 0.05deg^2 on the right panel, and around 0.10deg^2 on the left.

Regarding the results without the impact of the moonlight in Figure 5.15, although it is by nature a simpler case for the neural network, the change in zenith angle might have an impact on the performance. In that case, conclusions are more nuanced. Data adaptation remarkably increases the significance (+7.7 σ), but the reconstructed position is slightly at the advantage of the model trained on the standard dataset (+0.010 in distance).

5.7 Application of the γ-PhysNet-CBN for the detection of the Crab Nebula

Training hyperparameters

The training hyperparameters of the γ-PhysNet-CBN are identical to the one described in Section 4.5 of Chapter 4.

The γ-PhysNet-CBN inference workflow

During inference, the γ-PhysNet-CBN uses pedestal images to estimate the rate of Poisson noise in the acquisitions. As previously described in this chapter, data storage constraints impose that files, also known as sub-runs, contain roughly 53k images, which correspond to a few seconds of observation. In our inference framework, each sub-run is processed in parallel. Therefore, for each file, the images to be analysed are loaded into memory along with the corresponding pedestal images acquired simultaneously. The Poisson rate is then computed from these pedestals as the temporal and spatial mean, and it is used as the conditioning input for the entire sub-run. This approximation is sufficient because a few seconds of observation does not result in significant changes in NSB.

Results of the γ-PhysNet-CBN

The models used for inference have generated the IRFs depicted in Figure 4.7 presented in the last chapter. The model has been trained on the standard simulated dataset with a data augmentation procedure, consisting in applying a Poisson noise of rate $\delta\lambda$ sampled from a Uniform distribution of bounds 0 and 1. In that case, combined with the inherent perturbation of the simulations $\lambda_{MC} = 1.77$, the rate is contained in $\lambda_{MC} + \delta\lambda \in [1.77, 2.77]$, which covers all the possible noise range, as illustrated in Table 5.1.

The results of the application of the γ-PhysNet-CBN are depicted in Figure 5.16. On the contrary to the other presented methods, we didn't train this model on the tuned dataset with another $\delta\lambda$, as this current experimentation covers all cases. In this graph, the left panels show the moonlight condition and the right panels its absence. It turns out that the performance of our neural network performs better in the first scenario. Notably, the reconstruction of the position of the source is much more accurate, and despite the great difficulty to retrieve the physical properties of the particles in the degraded condition, the significance is higher by 1.6σ. Regarding both count maps and significance map, there is a clear shift towards lower declination in the no moonlight case. It is possible than our input conditioned model is more sensitive to the zenith angle. As illustrated in Figure 5.3, its value increase over time, but the value contained in the training simulations is of a fixed amount (20 degrees in particular). Moreover, the angular resolution is also more degraded, as shown in the θ^2 plot, with a converge of the background ensured after 0.10deg^2.

5.8 Applications of the γ-PhysNet-DANN and γ-PhysNet-CDANN for the detection of the Crab Nebula

In this section, we evaluate the contribution of the γ-PhysNet-DANN and its conditioning version in the case of real data. Preliminary results on unsupervised domain adaptation, reported in Section 4.6 of Chapter 4, has evinced a positive contribution of the domain task to reduce domain shifts. However, it importantly highlighted that, the more the noise in the target dataset increases, the more the domain performance decreases. Thus, in order to understand the impact of NSB and other factors on the γ-PhysNet-DANN results, the training will, in the following, be considered in two scenarios: Similarly to the standard analysis and

the γ-PhysNet, we evaluate unsupervised domain adaptation through γ-PhysNet-DANN and its conditioning version with and without data adaptation.

Training parameters

Including real data in the training set reveals a more challenging optimization process for the γ-PhysNet-DANN and its conditional version, because of the variety of conditions that it contained. Thus we pushed the number of epochs to 150 for both models, ensuring that each loss and weighting coefficients have properly converged. Similarly to the architecture proposed in Section 4.6 of the last chapter, the domain classifier is composed of two fully-connected layers of 100 features, combined with a ReLU activation function. The task balancing algorithm is Uncertainty Weighting. However, during our training phase, we noticed that the inclusion of the domain task into the task balancing strategy didn't allow for most of the seeds to converge. On average, 2 seeds out of 5 could be used in the end, regardless of the conditioning or the training dataset. Therefore, in the detection of the Crab Nebula, the domain loss function is weighted by a user-defined coefficient, that we set at 10^{-3}, and the results presented below take into account this modification.

Sampling of the real (target) dataset

Unsupervised domain adaptation incorporates real data within the unlabelled target dataset. In our case, the labelled source training dataset, described in both Section 4.2 of Chapter 4 and Appendix 8.3, is limited to 1.9M images. Therefore, considering a balanced amount of source and targets images, we need to carefully sample data from the available telescope observations, so that we extract enough variability for the training. As depicted previously in of this chapter, in Section5.2, there are a total of 36M acquisitions in our Crab dataset, split into 4 runs. Therefore, the sampling strategy consists in loading 20 sub-runs from each run. Regarding the training in moonlight condition, 20 files from the run numbered 6892 and 20 from 6893 are randomly selected, and all their events are considered. The same procedure is applied to the other scenario. In fact, as mentioned previously, a single sub-run contain roughly 53k events, which leads to the around 2M images with 40 files per training in total.

Results of the γ-PhysNet-DANN

Similarly to the γ-PhysNet, we aim to assess the benefits of domain adaptation on the IRFs, depicted in Figure 5.9. In total, four cases are represented and compared to the best scenario: training with the standard and the tuned datasets, both with moonlight and no moonlight target data. As a result, the energy resolution is at the advantage of the standard dataset, regarding the presence of the moon or not, especially at the first energy bin. The trend reverses slightly between 0.1 and 1 TeV in the moonlight scenario, and it is on the classification and energy bias metrics that the advantage on the tuned data in the most measurable, with a gain of almost 5% in accuracy at the lower energy level, and a gain of 0.1 on the bias on energies greater than 0.1 TeV. Finally, the γ-PhysNet-DANN is overall better in the context of moonlight, regardless of the training dataset that has been used.

The results on the detection of the Crab are presented in Figures 5.17 and 5.18. Regardless the run considered (6893 or 6895), it appears that the significance of domain adap-

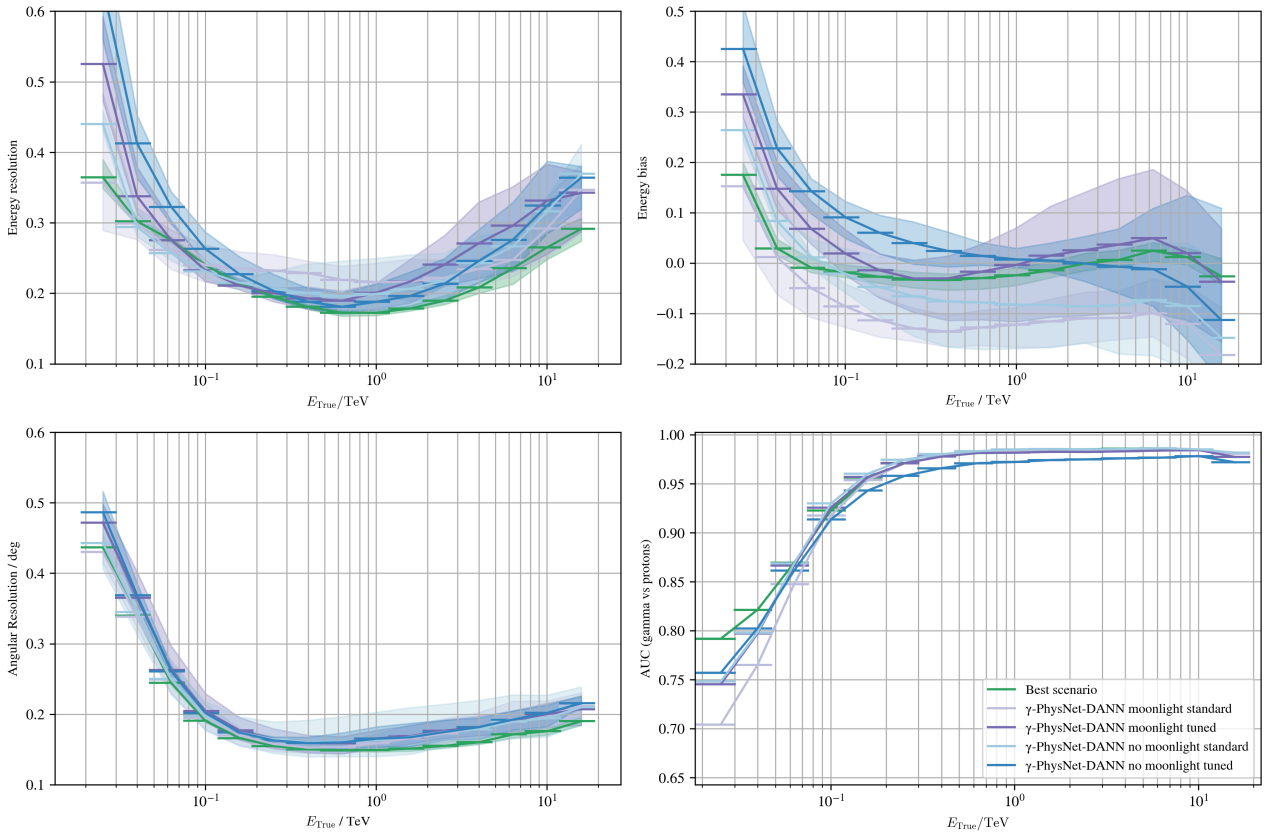


Figure 5.9: Performance of the γ-PhysNet-DANN on simulated data with and without data adaptation. In that case, targets are replaced by real data from the Crab dataset. The purple lines represent the moonlight condition, whereas the blue lines show the no moonlight condition. In total, 5 seeds have been produced to assess parameter initialization and data shuffling and target sampling.

tation using the tuned dataset is greater than compared to the standard training. This is expected from the previous analysis performed on simulated data in the previous chapter. In fact, Figure 5.2 illustrates the variety of NSB conditions affecting the considered runs. Yet, Figure 4.8 shows that, although the target set contains these kinds of variations, the performance of the model decreases as the noise increases. On the other hand, the spatial reconstruction of the source remains unchanged whether we perform data adaptation. Overall, the γ-PhysNet-DANN performance is better without moonlight when combined with data adaptation, although the angular resolution is degraded on the run 6893, with a convergence obtained for $\theta^2 > 0.10\mathrm{deg}^2$. Finally the results are rather similar when trained on the standard dataset.

Results of the γ-PhysNet-CDANN

Chapter 4 has demonstrated the benefits of conditioning when associated with DANN in the context of simulations. We presented the IRFs in Figure 4.10 with the target set corresponding to simulations. In Figure 5.10, targets are replaced with real data. The conclusions are identical to the DANN version. Although the classification power is at the advantage of the tuned dataset, other metrics exhibit slightly better performance in terms of mean and

min/max envelope with the standard dataset. In comparison to the IRFs depicted in Figure 5.9, results are rather consistent across both methods.

We now assess its performance in the real case. Results are shown in Figures 5.19 and 5.20. The conclusion is similar to the γ-PhysNet-DANN, but this time the reconstructed position of the source is clearly at the advantage of the standard training in the case of the moonlight condition (the distance increases almost threefold), with a significant shift present from the significance map. This shift is also visible from the count map. However, the gain in significance is consequential, with a gain of $+6.3\sigma$ in favour of the tuned dataset (right panel). Regarding the absence of moonlight, the gain is still present, although lower ($+3.2\sigma$). In this case, the position reconstruction is lower for the significance map of the model trained with data adaptation.

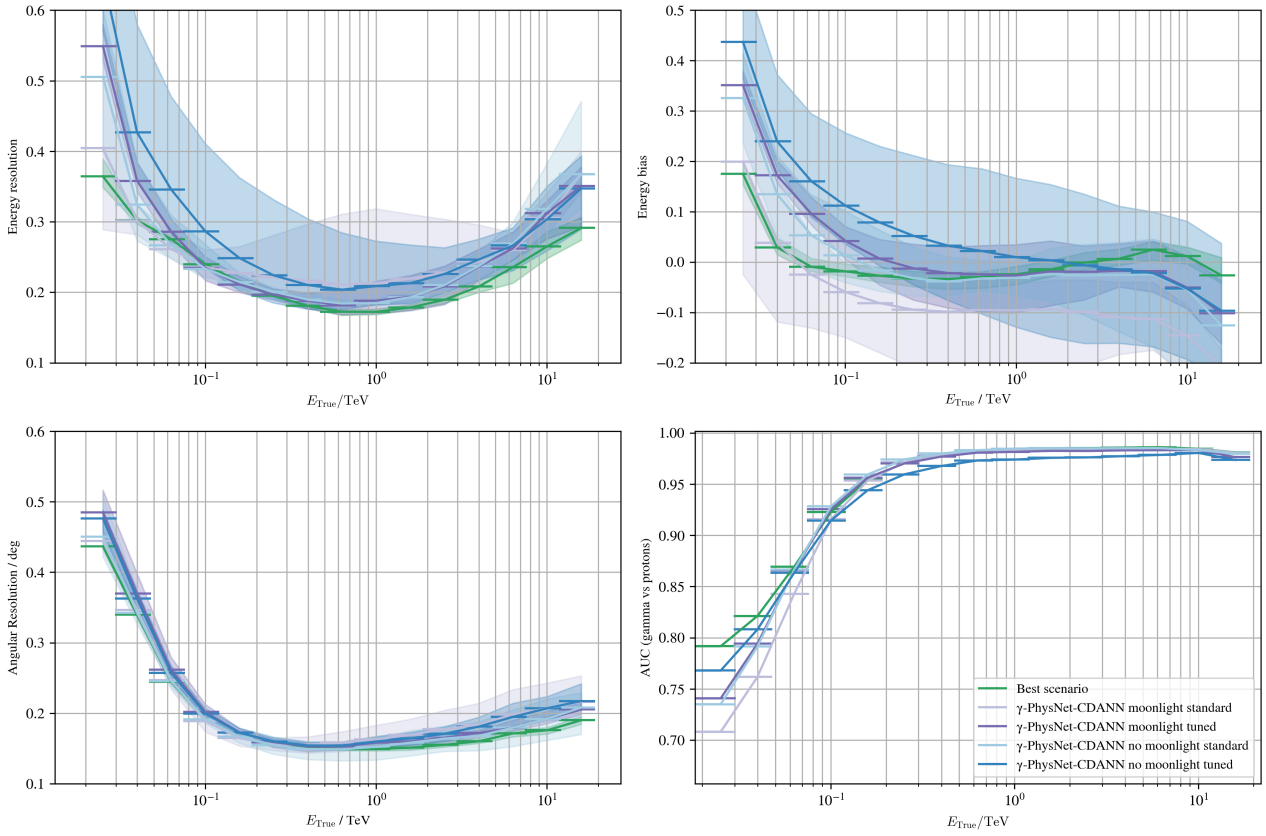


Figure 5.10: Performance of the γ-PhysNet-CDANN on simulated data with and without data adaptation. In that case, targets are replaced by real data from the Crab dataset. The purple lines represent the moonlight condition, whereas the blue lines show the no moonlight condition. In total, 5 seeds have been produced to assess parameter initialization and data shuffling and target sampling.

5.9 Application of the γ-PhysNet-Prime for the detection of the Crab Nebula

Training hyperparameters

The training hyperparameters of the γ-PhysNet-Prime are identical to the one described in Section 4.8 of Chapter 4.

Results of the γ-PhysNet-Prime

In agreement with the IRFs of the γ-PhysNet presented in Figure 5.8, our Transformer model, whether trained on the standard or tuned datasets, shows very good performance across all metrics. Regarding the angular resolution, our Transformer trained on the standard dataset exhibits slightly better results at the lower energy range (−0.05 in error). On the energy resolution chart, the same report can be drawn between 1 and 10 TeV. Yet, because only 1 seed has been produced, it is not possible to infer definitive conclusions, and these variations may be caused by statistical fluctuations. A more complete study is let to short term perspectives of this work.

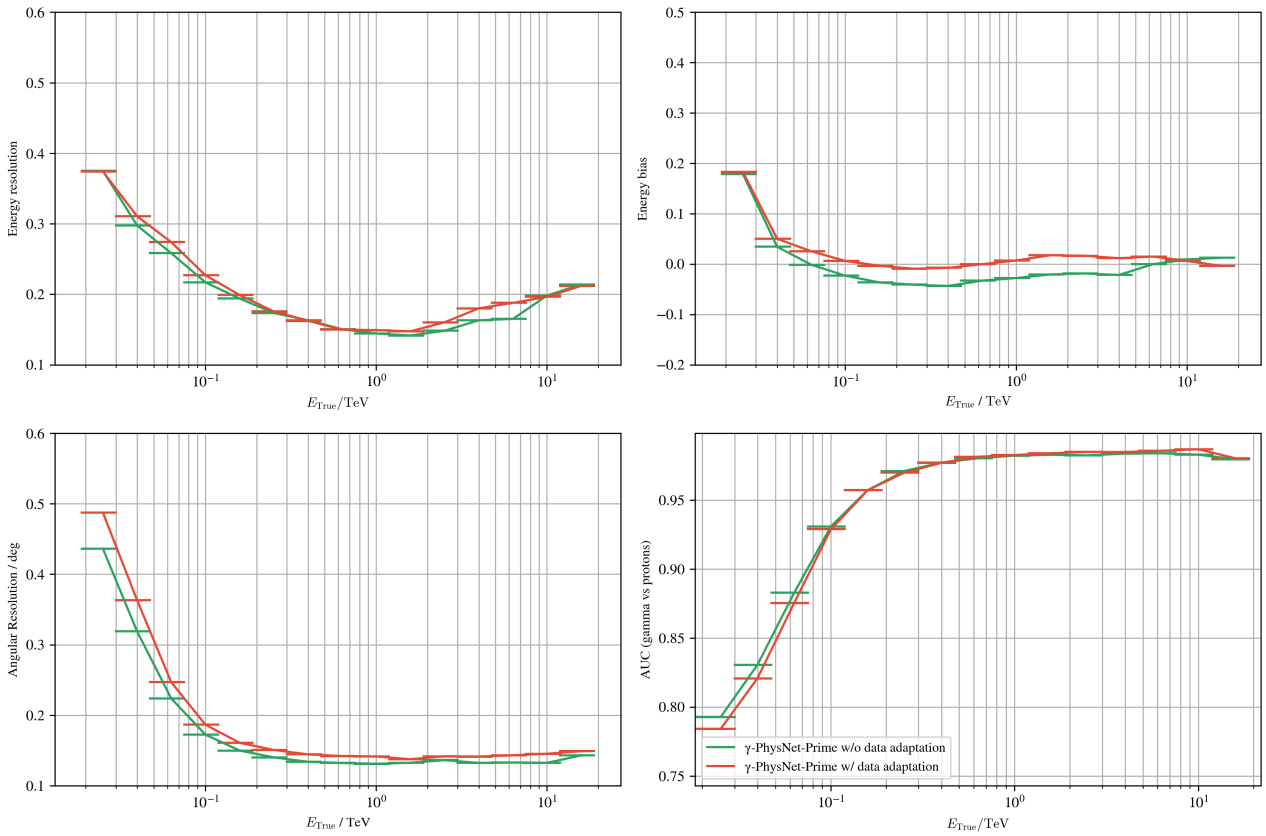


Figure 5.11: Performance of the γ-PhysNet-Prime with and without data adaptation. Only one seed has been produced.

We now compare the application of the γ-PhysNet-Prime for the detection of the Crab. In both Figures 5.21 and 5.22, it appears that data adaptation doesn't necessarily improve the performance of the model. In fact, the moonlight condition shows that the position is

better reconstructed with the use of the standard dataset. Yet, the last run (6895) exhibits the opposite conclusion. Overall, each case displays roughly the same significance value (between 15.8σ to 16.3σ), although it is not possible to conclude on any trend based on one initialization seed.

5.10 Method comparison for the detection of the Crab Nebula

In the section, results of each method are put in perspectives to ease their comparison. The summary in moonlight condition is displayed in Table 5.2. This is the most difficult case because the NSB difference between both the standard training and Crab inference distributions is at its maximum, as depicted in Table 5.1. Regarding the models trained on the standard dataset, Hillas+RF exhibits the best performance. On the other hand, the neural networks suffer more from the domain discrepancy. However, the gain in significance is notable between the γ-PhysNet baseline and our contributions, with an increase of 8.5σ compared to our Transformers, and almost $+10\sigma$ with our CBN-based approach. Yet, these results are less remarkable when the tuned dataset is considered. In fact, the γ-PhysNet recovers from its inherent loss, and our novel approaches increase the significance by $+2\sigma$ at maximum with our *gamma*-PhysNet-DANN and *gamma*-PhysNet-CDANN models, reaching the performance of the standard analysis. Nevertheless, our Transformer and Hillas+RF didn't really change with the data adaptation procedure. Regarding the position bias, there is no clear improvement brought by any method compared to the others, at the exception of the γ-PhysNet trained on the tuned dataset and its CBN-based version that, on average, show slightly better reconstructed positions.

Furthermore, Table 5.3 depicts the summary in absence of moonlight. In that case, the difference between the standard and tuned datasets is lesser compared to the first scenario. The standard analysis, the γ-PhysNet-CBN and the γ-PhysNet-Prime didn't improve with data adaptation. Yet, there is an increment of $2-3\sigma$ when considering our domain adaptation approaches. However, the greatest improvement comes from the γ-PhysNet baseline, with a gain of almost $+9\sigma$. Overall, this method shows the best results. Regarding the position bias, the conclusion is similar to previously, although in that case no model display an significant advantage.

5.11 Conclusion

In the chapter, we investigated the benefits of our novel approaches, the γ-PhysNet-CBN, γ-PhysNet-DANN, its conditional version and the Transformer γ-PhysNet-Prime in the concrete case of the Crab Nebula detection, a well-known reference gamma-ray source. Our contributions were compared to the baselines, constituted by the Hillas+RF standard analysis and the γ-PhysNet, in two different scenarios, with and without moonlight. In both cases, and as highlighted in previous work [Vui+21], the study is also conducted with and without data adaptation.

Comparing the performance on the standard procedure, each contribution increases the significance value by a notable margin ($+5$ to $+9\sigma$ in moonlight condition, and $+2$ to $+4\sigma$ in

the other scenario). This scenario, that is considered the most complicated for the baselines, illustrates the benefits of unsupervised domain adaptation and the CBN modules. However, for each of our contribution, their best results are obtained when they are trained on the tuned dataset. Then, because the application of data adaptation shows greater performance, this highlights that our models fail at completely reducing the gap between simulations and real data in the standard setting. Concerning the last observation run, the γ-PhysNet combined with data adaptation has overall the best result. Notably, the standard analysis appears resilient to the NSB level, and exhibit similar results regardless of the condition and training dataset.

Although enhancements have been observed on simulations in the presence of NSB perturbations and label shift, conditioning our DANN-based neural network seems to show slight improvement on real data in one case specifically, which is the absence of moonlight. In this context, we observe an increase of roughly $+2\sigma$. Yet, no relevant benefits are shown on the position bias.

Drawing conclusions on our Transformer is difficult, as only one seed has been produced, and this is subject to initialization uncertainty. However, its performance when trained on the standard dataset and applied to moonlight condition is very encouraging. In that specific case, it outperforms each other method remarkably.

Regarding the reconstruction of the source position, there is no significant difference that can be reported. With the exception of some specific seeds, each architecture manages on average to reconstruct a pertinent set of coordinates, although the zenith angle of the Crab observation varies over time.

Finally, two opposite effects are at stake during the acquisitions, a decreasing noise level and an increasing zenith angle. We can expect an improvement of the performance from the drop in NSB, which is not the case, most likely due to the other issue. Thus, this shed light on the necessity to integrate more information to the network, such as the pointing direction of the telescope, which can be effectuated using the CBN modules. This study is a short-term perspective that is left in a future work.

In conclusion, the analysis of the Crab Nebula leads to the following recommendations:

- Evaluating the performance of our contributions and the baselines with and without data adaptation, it appears that combining dataset tuning with the γ-PhysNet is the most promising approach in this specific case of the Crab Nebula. In fact, although Hillas+RF has slightly better results in the moonlight-affected runs, it has been demonstrated in Figure 4.13 of the last chapter that the standard analysis does not perform well at lower energy ranges, which is a limiting factor in the reconstruction of the flux of any gamma-ray sources.

- Regarding the applicability of neural networks, the current major drawback that can be attributed is the necessity to train the models for each specific observation. In fact, the application of data adaptation illustrates this constraint. Without the tuning of the dataset to match the NSB, the gain in performance is limited, but the step imposes to generate a new set of data for each acquisition. It appears that the standard analysis is the sensitive to the issue. Conversely, the γ-PhysNet-CBN, with its robustness capabilities, is able to take into account particular variations within the training procedure. Yet, the key lies in model generalization, which our Transformer model, through the MAE pre-training, is the most capable of providing.

- The optimization of domain adaptation techniques is challenging, for the reasons evoked in Section 3.5 of Chapter 3, notably the conflicting tasks between the domain and particle classifications. Similarly, precise training process control is necessary when considering Transformer approaches. The pre-training and fine-tuning necessitate tuning many hyperparameters, which requires expertise, computing time and resources. On the other hand, CBN-based methods are easier to converge, although they need to be optimized longer to reach their best performance. Yet, our baseline (the γ-PhysNet) remains the fastest and easiest model to train.

	Methods	**Significance**	**Position bias**	**# Seeds**
Standard	Hillas+RF	**18.3** (18.3)	0.034	1
	γ-PhysNet	7.54 (10.7)	0.060	5
	γ-PhysNet-DANN	12.9 (17.0)	0.051	5
	γ-PhysNet-CDANN	12.1 (16.1)	0.061	5
	γ-PhysNet-Prime	16.1 (16.1)	0.034	1
Tuned	Hillas+RF	19.5 (19.5)	0.044	1
	γ-PhysNet	17.8 (19.7)	0.036	5
	γ-PhysNet-DANN	19.2 (20.1)	0.053	5
	γ-PhysNet-CDANN	**19.9** (22.4)	0.041	5
	γ-PhysNet-Prime	15.8 (15.8)	0.060	1
	γ-PhysNet-CBN	17.0 (20.3)	0.036	4

Table 5.2: Summary table in moonlight condition. Results are average when multiple seeds are available. The highest significance value across the seeds is in parenthesis. Best average significance is displayed in bold.

	Methods	Significance	Position bias	# Seeds
Standard	Hillas+RF	**20.5** (20.5)	0.015	1
	γ-PhysNet	14.3 (16.1)	0.033	5
	γ-PhysNet-DANN	16.1 (16.8)	0.044	5
	γ-PhysNet-CDANN	18.0 (19.3)	0.037	5
	γ-PhysNet-Prime	16.3 (16.3)	0.074	1
Tuned	Hillas+RF	20.5 (20.5)	0.055	1
	γ-PhysNet	**23.1** (23.8)	0.044	5
	γ-PhysNet-DANN	19.8 (22.3)	0.037	5
	γ-PhysNet-CDANN	20.4 (22.5)	0.046	5
	γ-PhysNet-Prime	16.5 (16.5)	0.032	1
	γ-PhysNet-CBN	16.7 (18.7)	0.059	4

Table 5.3: Summary table in no moonlight condition. Results are average when multiple seeds are available. The highest significance value across the seeds is in parenthesis. Best average significance is displayed in bold.

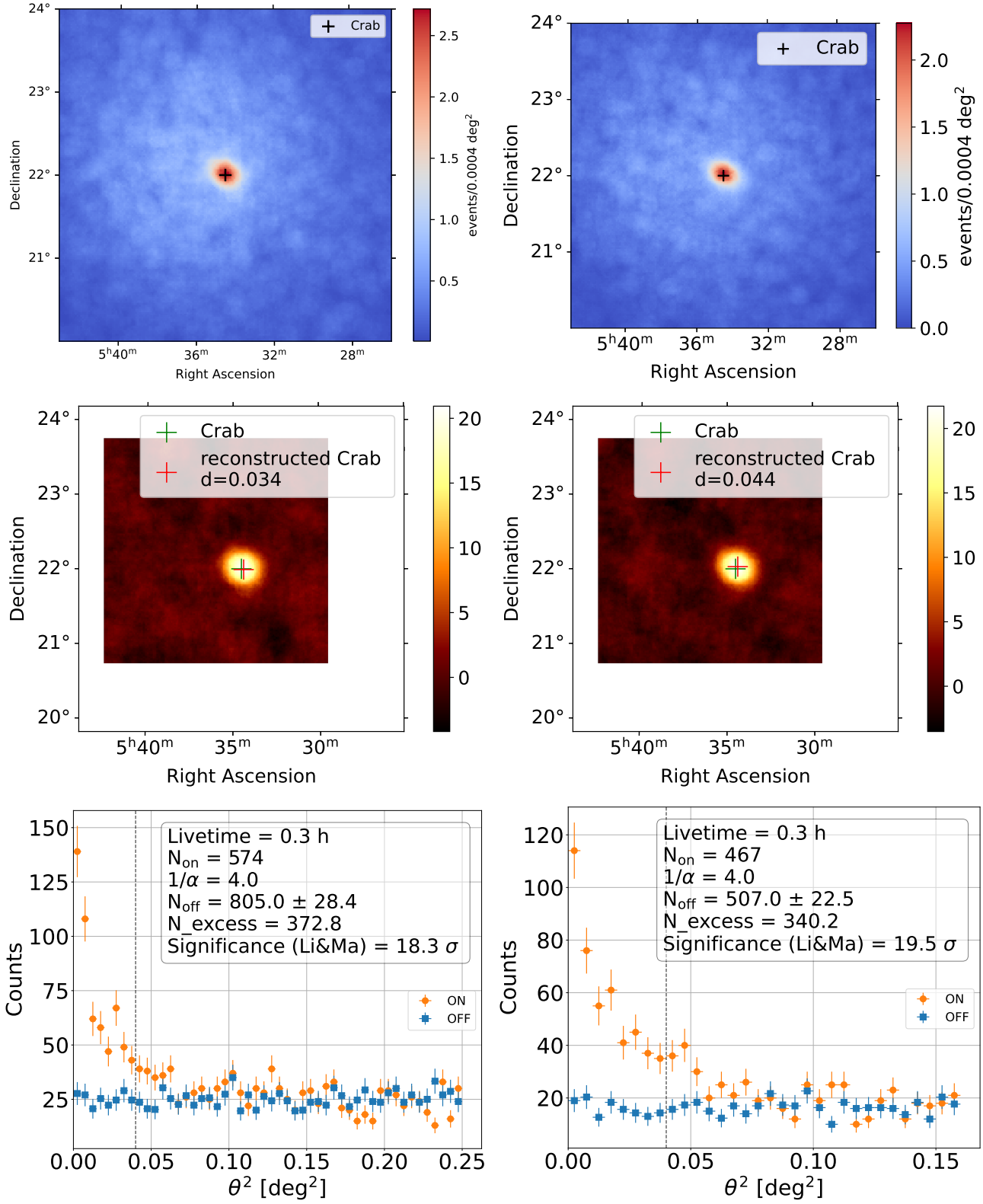


Figure 5.12: The Hillas+RF for the detection of the Crab in moonlight condition (run 6893) trained on the standard dataset (left) and the tuned dataset (right). Only the best seeds in terms of significance are reported.

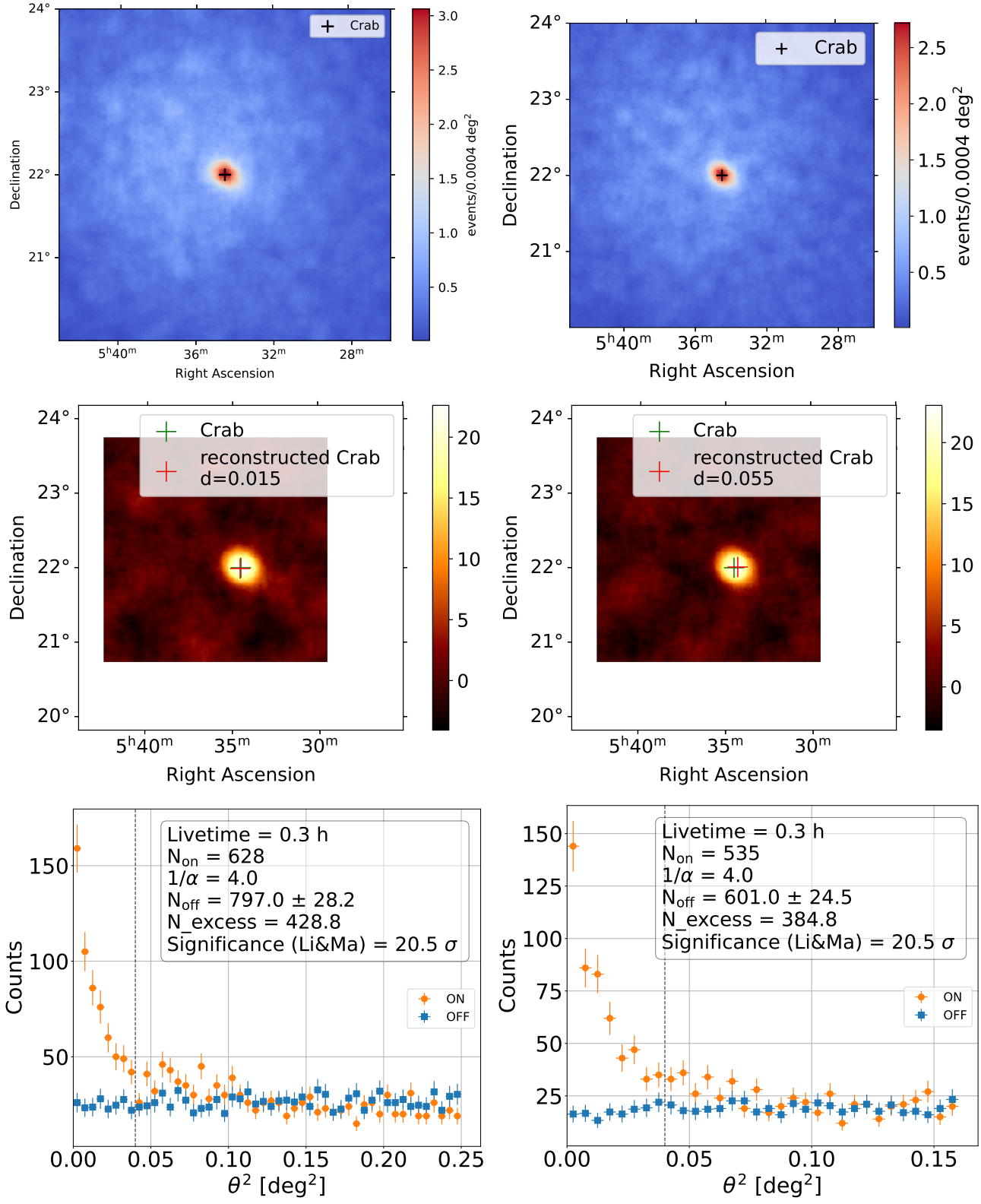


Figure 5.13: The Hillas+RF for the detection of the Crab in no moonlight condition (run 6895) trained on the standard dataset (left) and the tuned dataset (right). Only the best seeds in terms of significance are reported.

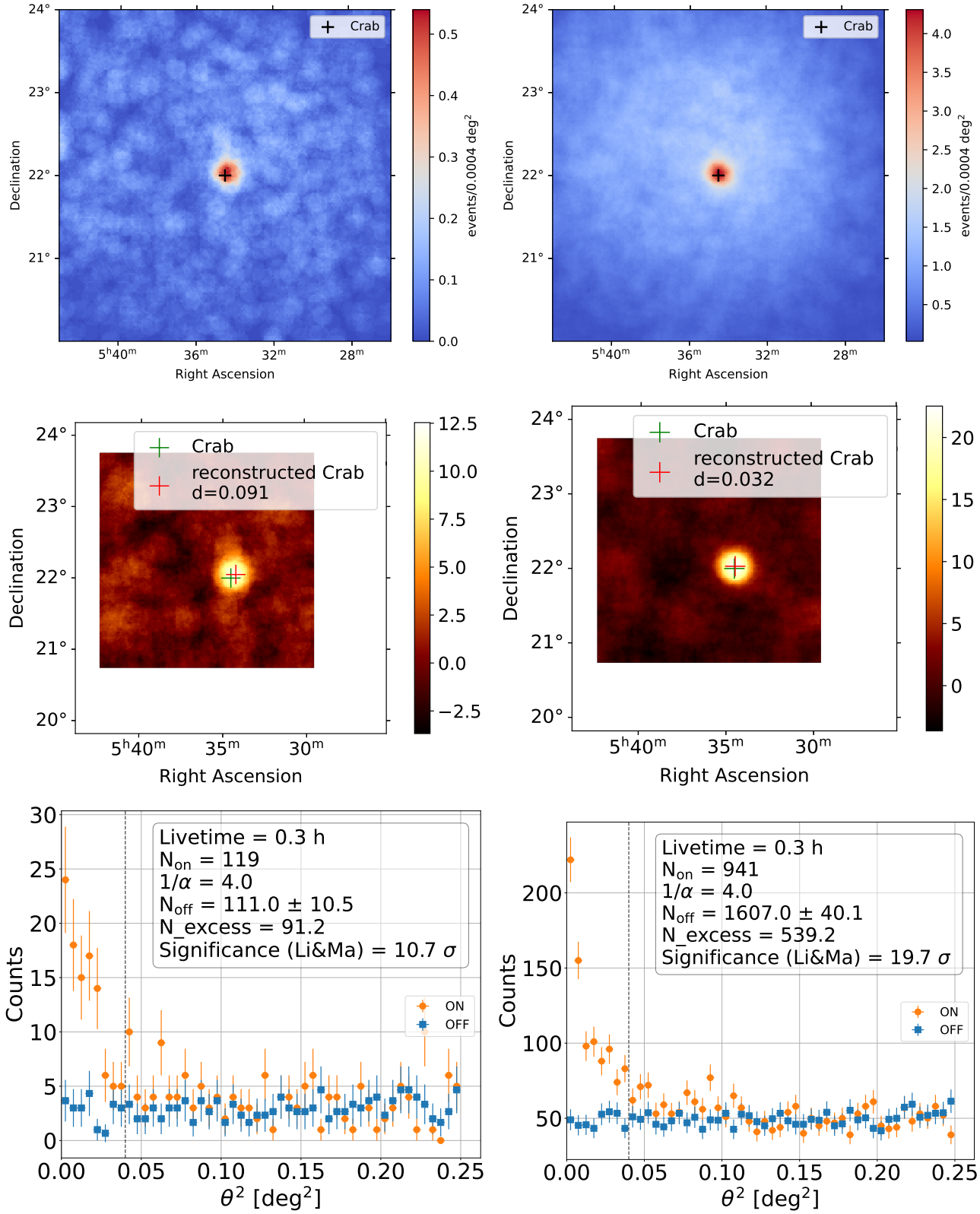


Figure 5.14: The γ-PhysNet for the detection of the Crab in moonlight condition (run 6893) trained on the standard dataset (left) and the tuned dataset (right). Only the best seeds in terms of significance are reported.

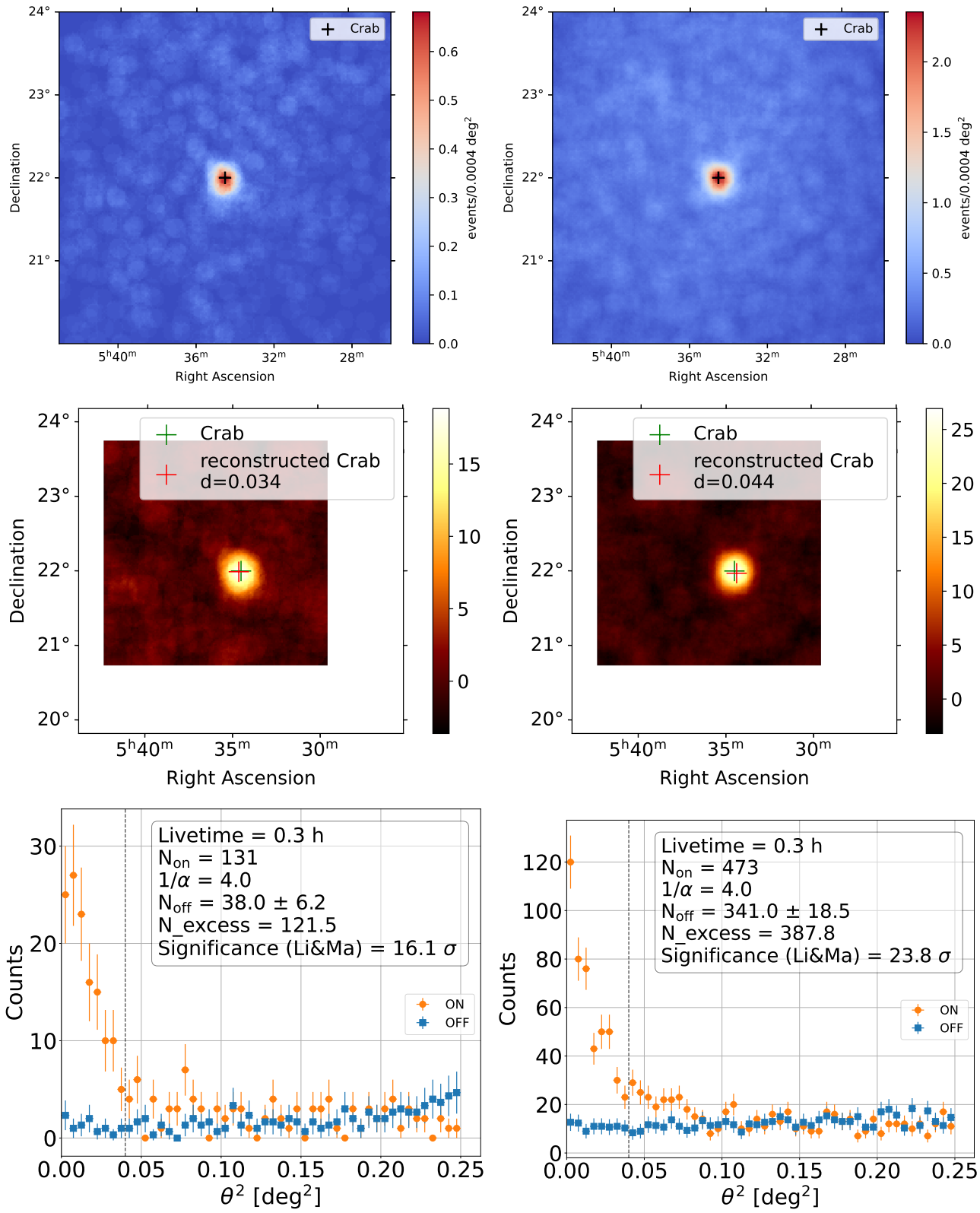


Figure 5.15: The γ-PhysNet for the detection of the Crab in no moonlight condition (run 6895) trained on the standard dataset (left) and the tuned dataset (right). Only the best seeds in terms of significance are reported.

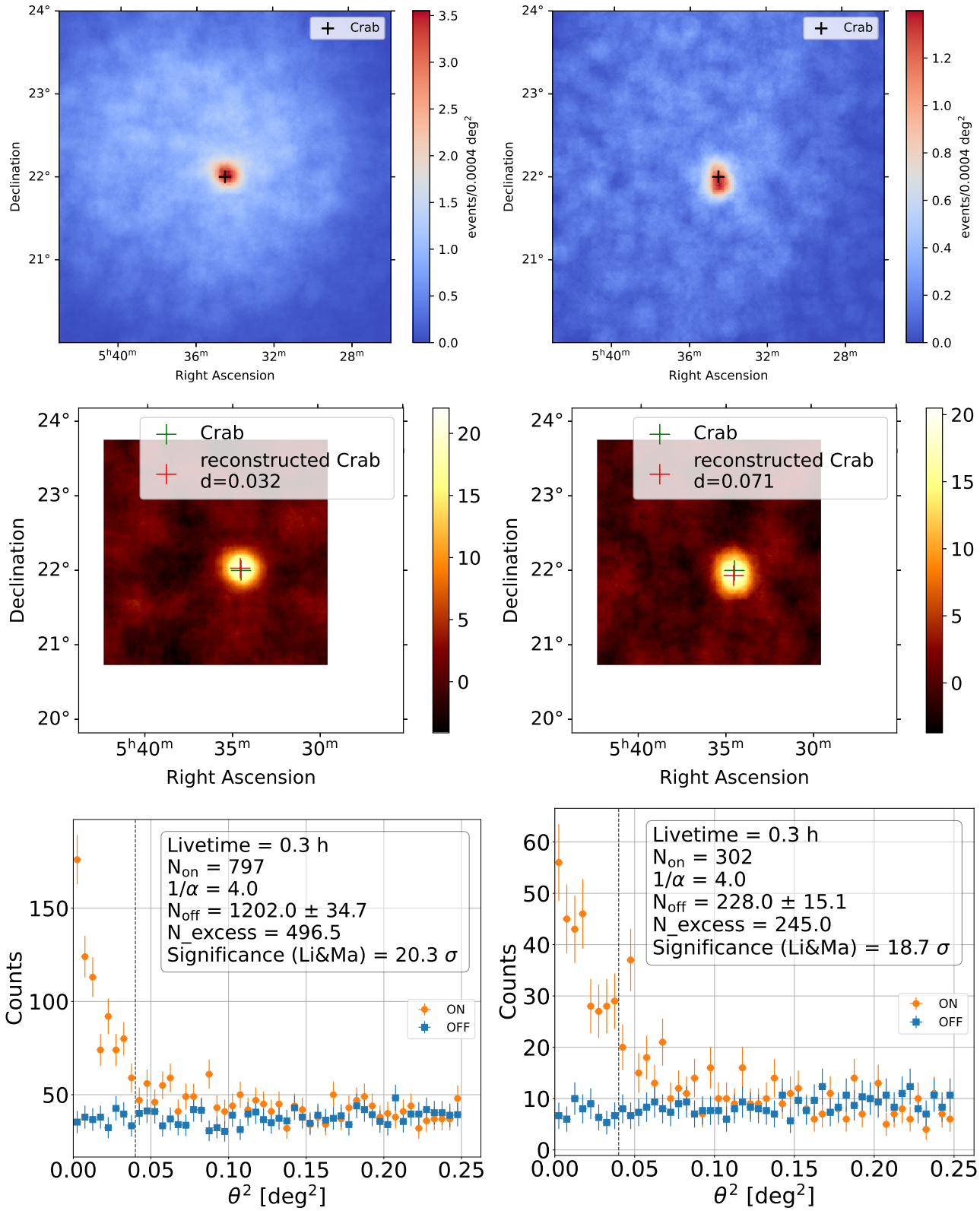


Figure 5.16: The γ-PhysNet-CBN for the detection of the Crab in moonlight condition (run 6893, left) and no moonlight (run 6895, right) conditions.

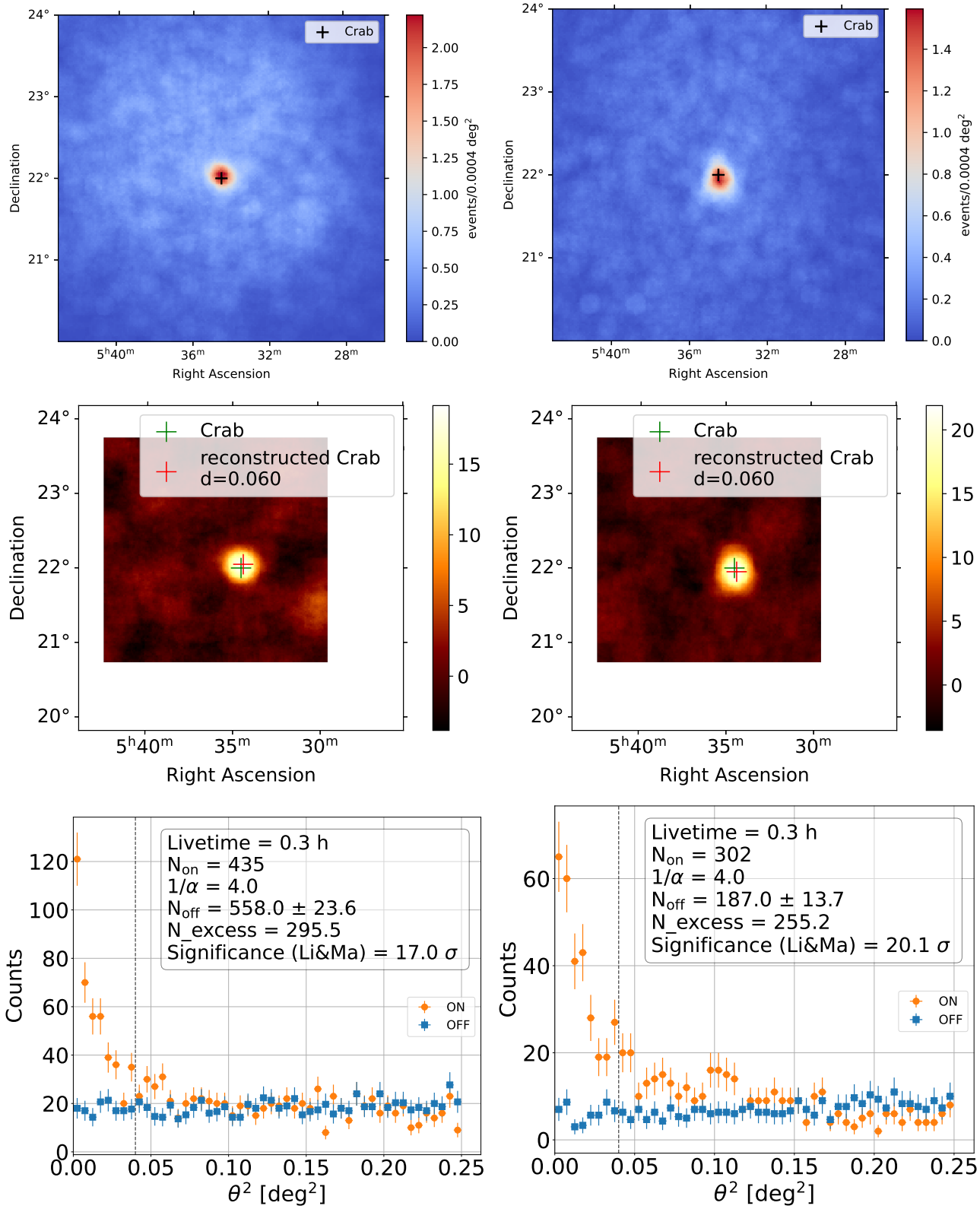


Figure 5.17: The γ-PhysNet-DANN for the detection of the Crab in moonlight condition (run 6893) trained on the standard dataset (left) and the tuned dataset (right).

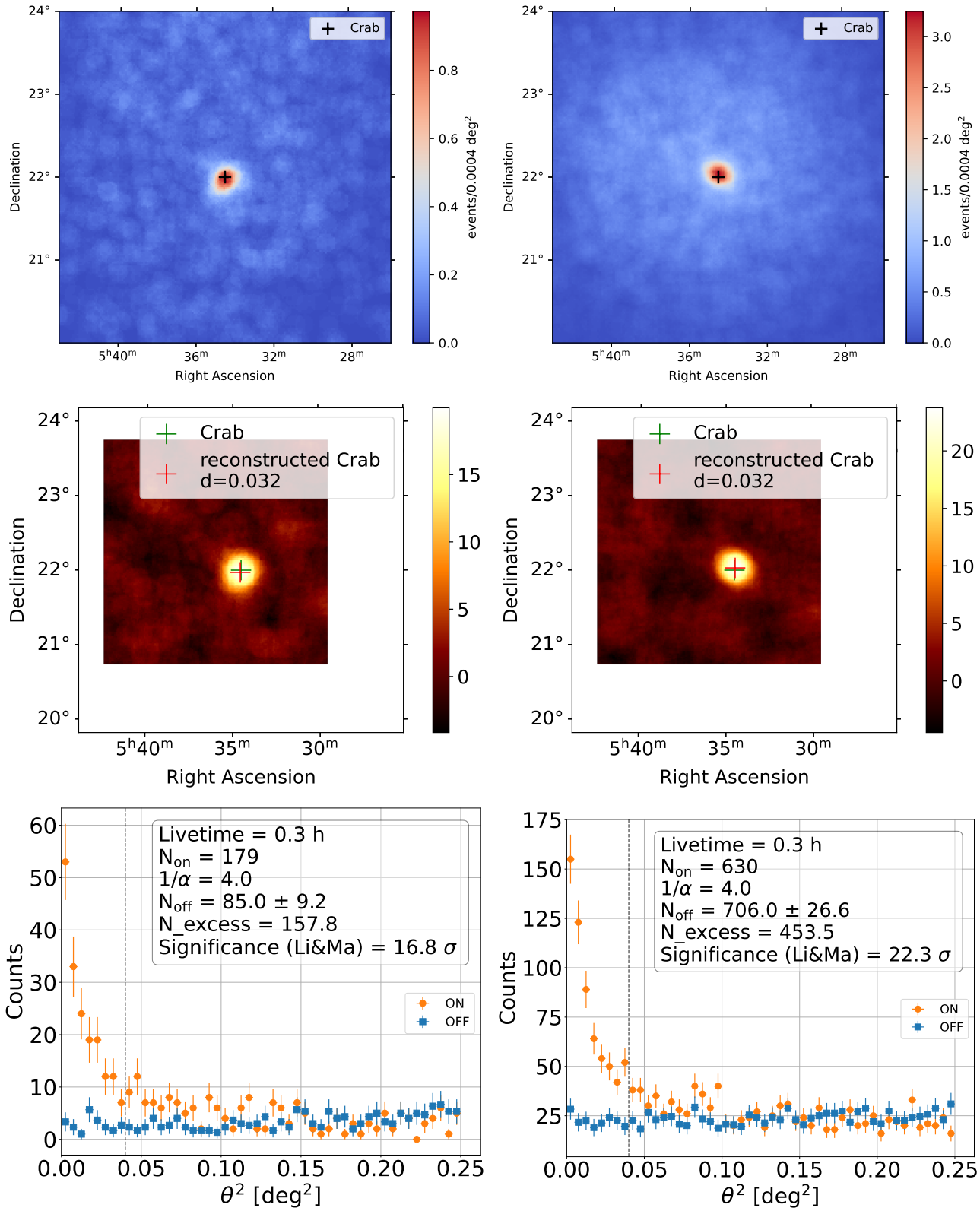


Figure 5.18: The γ-PhysNet-DANN for the detection of the Crab in no moonlight condition (run 6895) trained on the standard dataset (left) and the tuned dataset (right).

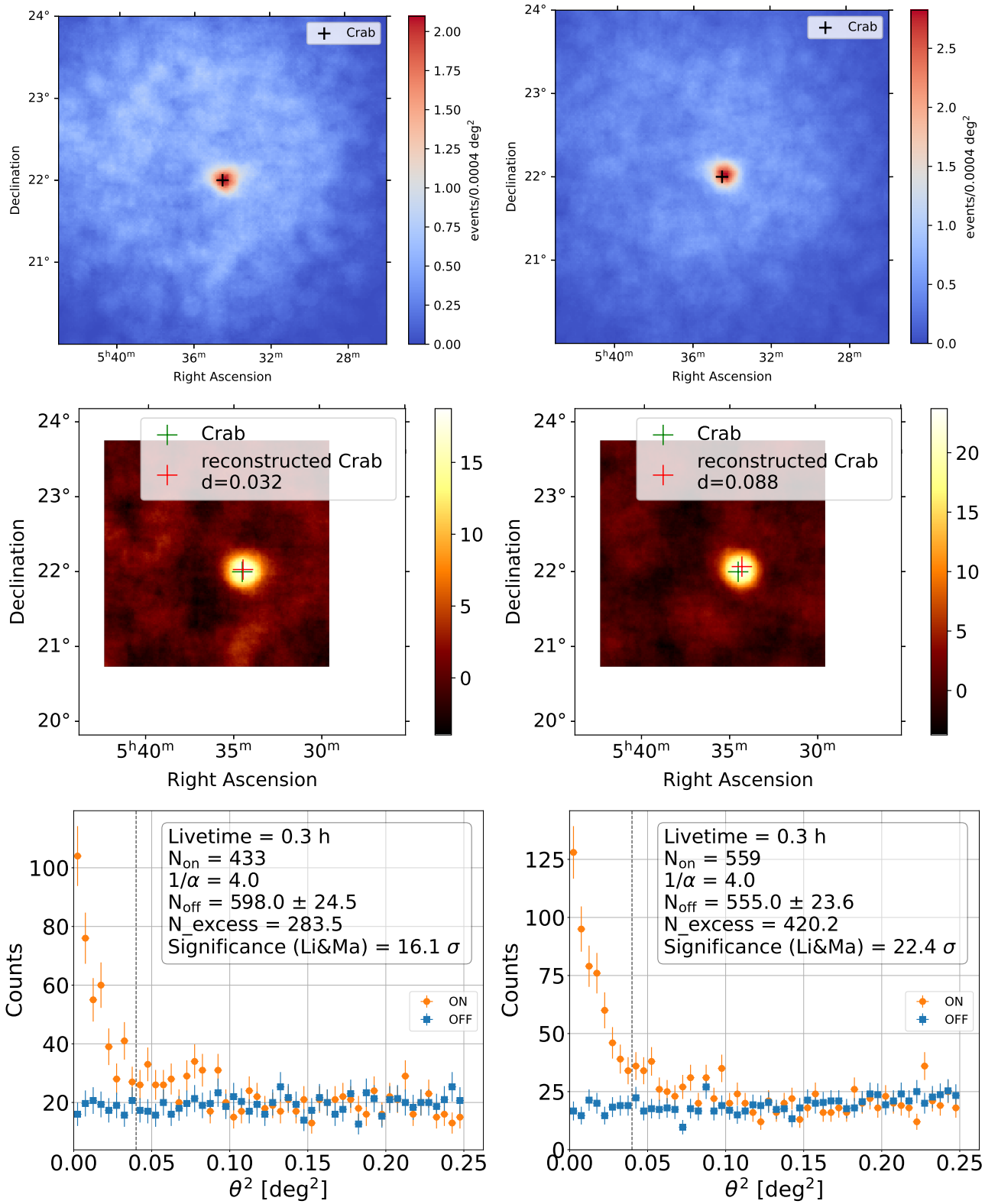


Figure 5.19: The γ-PhysNet-CDANN for the detection of the Crab in moonlight condition (run 6893) trained on the standard dataset (left) and the tuned dataset (right).

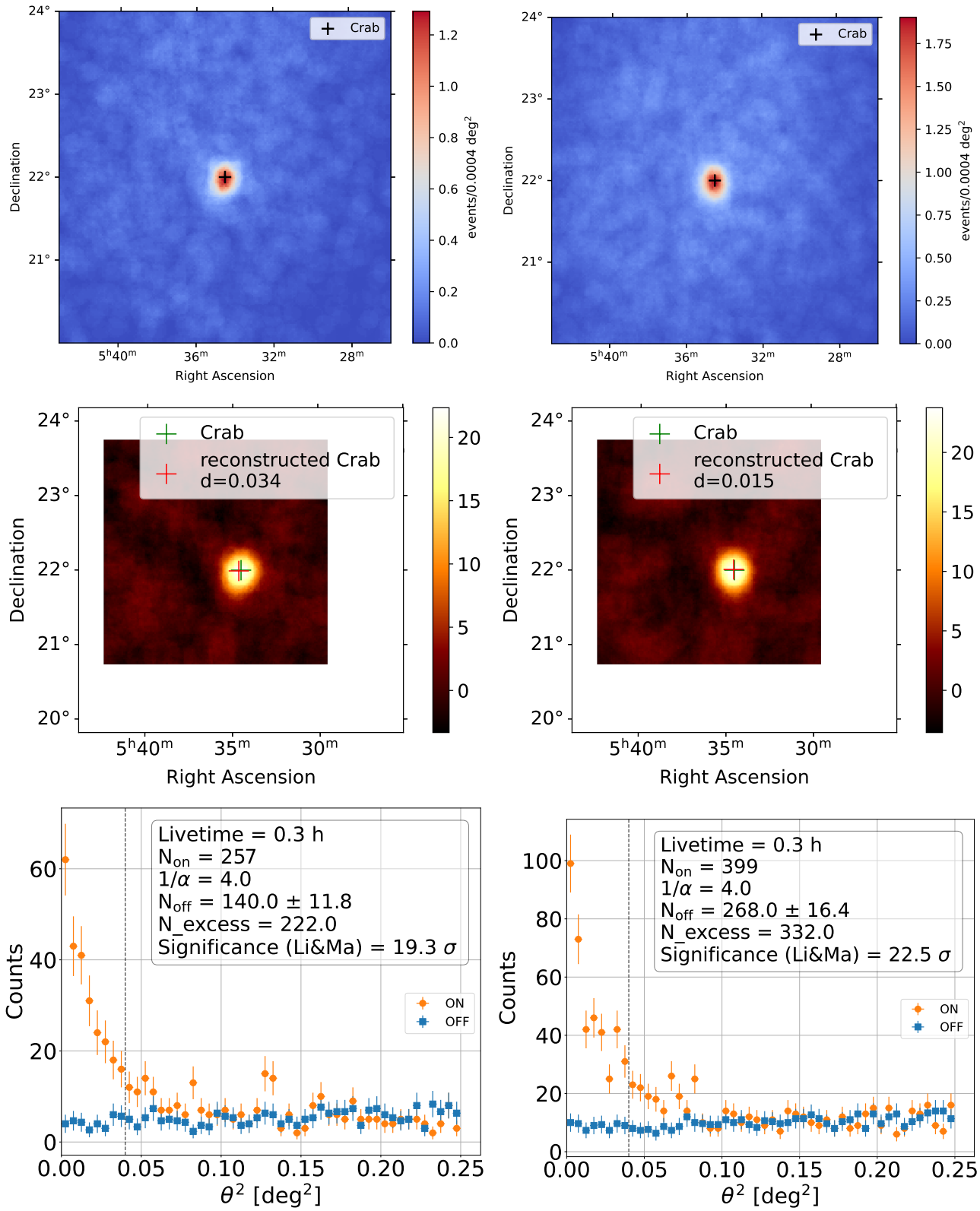


Figure 5.20: The γ-PhysNet-CDANN for the detection of the Crab in no moonlight condition (run 6895) trained on the standard dataset (left) and the tuned dataset (right).

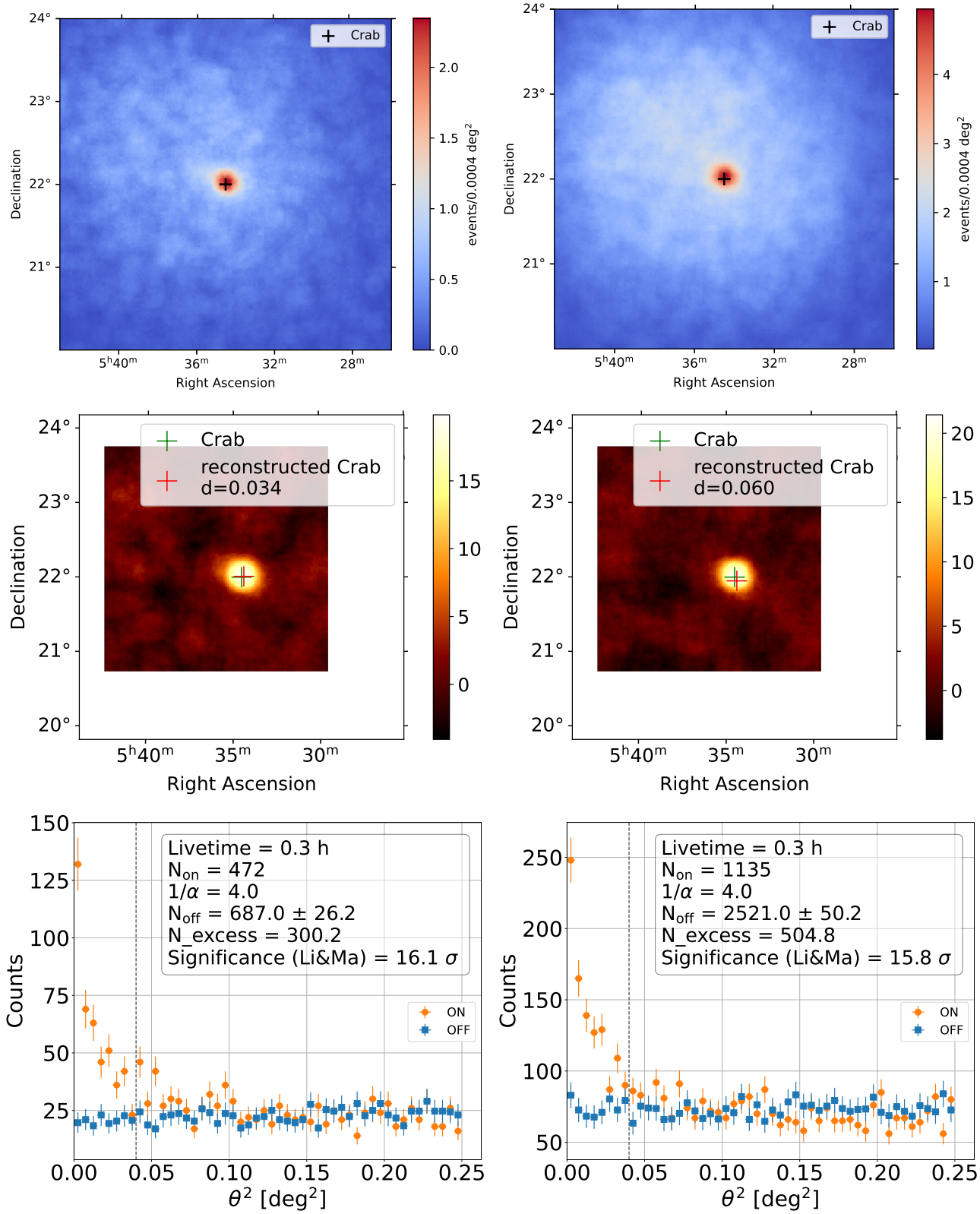


Figure 5.21: The γ-PhysNet-Prime for the detection of the Crab in moonlight condition (run 6893) trained on the standard dataset (left) and the tuned dataset (right). Only the best seeds in terms of significance are reported.

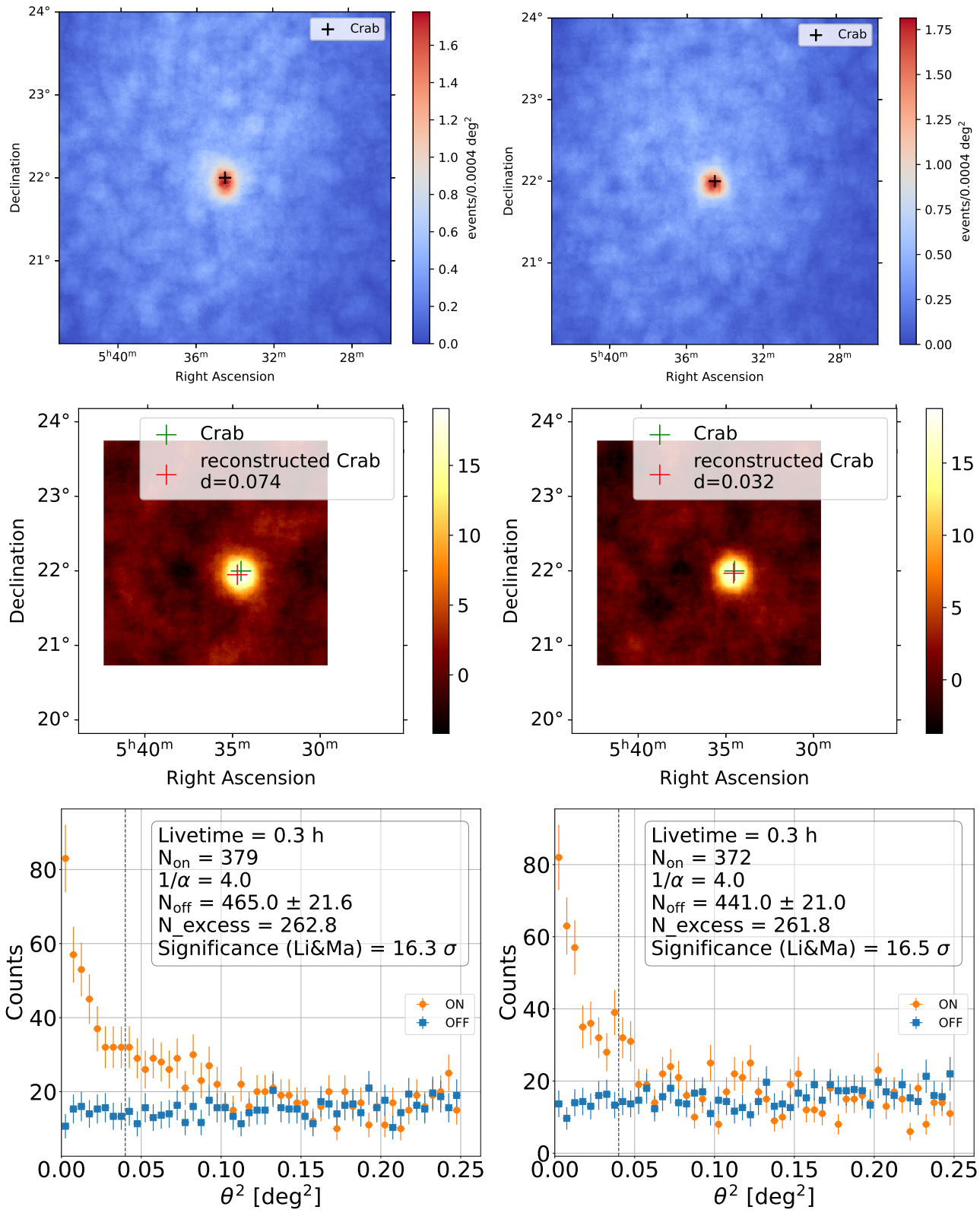


Figure 5.22: The γ-PhysNet-Prime for the detection of the Crab in no moonlight condition (run 6895) trained on the standard dataset (left) and the tuned dataset (right). Only the best seeds in terms of significance are reported.

6 Conclusion and perspectives

6.1 Conclusion

Astronomy is undoubtedly one of the oldest science that has captivated mankind's interest and has forged the culture of the most ancient and modern civilizations. Throughout history, countless beliefs have emerged, attributing sanctity to a wide range of unknown astronomical events. In fact, the lack of scientific knowledge, combined with the intrinsic curiosity of the human specie, pushed the population to find meaning in their observations and thus, giving birth to diverse cosmogonies.

In the relentless pursuit of truth and understanding, humanity has developed methodologies, tools, and instruments to extend the boundaries of what was once beyond reach. Nowadays, with the development of computing resources and simulations, deep learning has positively impacted a great variety of domains. Factually, with the vast amounts of data generated by modern telescopes and observatories, traditional methods of analysis have become insufficient. Therefore, this new paradigm is hoped to bring a brand new perspective in the field of astrophysics.

The focus of this thesis therefore lies within the realm of gamma-ray astronomy, specifically aiming to apply deep learning techniques to the analysis of images from the Cherenkov Telescope Array Observatory. Positioned at the intersection of astronomy and artificial intelligence, this work aspires to be a small piece of the puzzle in answering the difficult yet crucial astronomical questions currently at stake among the scientific community.

Contributions of this thesis

The current standard analysis, based on the Hillas algorithm combined with Random Forests, has been used to fully characterize the Crab Nebula with a compilation of LST acquisitions in the important work of [Abe+23]. Such approach has the appreciable advantage to be, to a certain extent, explainable. On the one hand, the extraction of a few parameters with the Hillas algorithm are based on human-designed morphological hypotheses. On the other hand, it is possible to compute the importance of each Random Forest input features. Nevertheless, it intrinsically suffer from limitations especially at the lower energy range, where lies most of the gamma-ray flux of a source.

Recently, the application of deep learning to reconstruct incident particle physical parameters has shown significant potential in enhancing detection performance in both simulations and real gamma-ray source detection. Notably, the γ-PhysNet is the first full-event reconstruction model for the LST-1. However, because RFs or neural networks are trained on simulations, they inevitably face domain shifts when applied to real observations. This shift results in decreased performance and generalization. It is particularly highlighted when

considering data adaptation. The modification of the training dataset with background matching, illustrated in the article [Vui+21], evinced much better results compared to a direct application.

In addition, it has been shown that the Night Sky Background is one of the most significant contributing factors affecting the detection performance, along with the pointing direction of the telescope, as investigated in the major work of [PMO22]. In both cases, it is possible to understand and theorize these concepts in depth. Firstly, the NSB can be evaluated using pedestals, allowing for the characterization and parameterization of its temporal variations across each pixel. Secondly, the impact of pointing can be mitigated by integrating a wide variety of pointing directions into the training dataset. Yet, despite these efforts, unknown or unparameterized differences may still exist and significantly contribute to the loss of performance.

Throughout this thesis, we have followed a structured guideline to address these issues. Built on the work of [Jac20], we have extended the research to different contexts:

- Known differences: These can be tackled using information fusion from external variables, such as pedestal images. We introduced γ-PhysNet-CBN, incorporating Conditional Batch Normalization modules to mitigate the negative impact of NSB variations.
- Unknown differences: To address these, we explored and evaluated unsupervised domain adaptation techniques. We proposed and compared several approaches, and we have selected the γ-PhysNet-DANN as the most promising. We also provide a comprehensive study to its conditioning version and its integration to multi-task balancing.
- Global model: As a plausible global solution, we introduced γ-PhysNet-Prime, a Transformer version of γ-PhysNet, designed to provide a more robust and adaptive model. It benefits from a increased learning capacity, optimized along with a first unsupervised learning step followed by a supervised specialization.

Input conditioning

Input conditioning, through the implementation of CBN modules in replacement to BN layers, has shown promising results in their application to simulations. Based on data augmentation and Poisson perturbations, it leverages the rate of the noise as an external variable that conditioned the neural network. The computation of the IRFs suggests that the model perfectly managed to learn the notion of noise and its performance are almost similar to the γ-PhysNet best scenario.

Regarding the detection of the Crab Nebula, two effects can be highlighted. As a reminder, the first two runs are degraded and affected by moonlight. The two last runs don't suffer from such perturbations but contain a higher zenith angle compared to the training set. The first conclusion concerns the success of their application with the standard dataset compared to the standard analysis and the γ-PhysNet in the same setup. It appears that the inclusion of CBN modules managed to provide the neural network resilience toward NSB variations. However, as a second point, it did not retrieve the performance of the baselines when trained on the tuned dataset, as it would have been expected. In fact, if the Poisson hypothesis was too weak, our model should at least have reached the results of data adaptation.

Unsupervised domain adaptation

The γ-PhysNet-CDANN has been selected among multiple approaches as the most promising for the detection of the Crab Nebula. To evaluate the contribution of conditioning, the DANN version has also been applied. As a result, the integration of domain adaptation, with and without the conditioning, clearly shows a improvement when applied to the moonlight condition. On the other hand, the application of data adaptation suggests that, after NSB matching, there is most likely no significant factor to domain shift, and the difficulty to optimize our model take the lead on a possible performance enhancement. It is necessary to test the γ-PhysNet-DANN is a more complicated setting to evaluate its capability to detect gamma-ray sources.

Regarding the application of conditioning, it has, on average, shown an improvement on the last observation run with the standard training dataset. Once trained with data adaptation, it has, however, no more impact.

Transformers

Currently, the γ-PhysNet-Prime has provided the best overall results when applied to simulations. The performance is yet more nuanced for the detection of the Crab Nebula, and two main conclusions can be drawn. Firstly, the difference in performance when trained on the standard or tuned datasets is negligible. This can probably be attributed to the power of the MAE to produce a noise robust latent representation. Secondly, the CNN-based γ-PhysNet, trained on the tuned dataset, currently outperforms our Transformer in the detection of the Crab Nebula. A possible explanation could lie in the difficulty to optimize such models. In fact, it has been noticed that training the γ-PhysNet-Prime over 500 epochs leads to overfitting. Thus, the fine-tuned model probably lacks regularization. Regarding the γ-PhysNet-Megatron, our domain adaptation Transformer, we unfortunately did not manage to train the model to converge, and more investigations are necessary to understand the causes.

6.2 Perspectives

The production of our work has shed light on two types of perspectives. The short term perspectives correspond to the prompt improvements and tests of our contributions and the high-level analysis. On the other hand, the long term perspectives correspond to novel approaches and some relevant research directions that need more efforts and investigations.

Short term perspectives: Enhancement of the proposed methods

The γ-PhysNet-CBN

There are many compelling possibilities to improve the current implementation of the γ-PhysNet-CBN.

Firstly, as a simple proof-of-concept, the conditioning variables used in this work consisted solely in the Poisson rate. However, the underlying assumptions of Poisson distribution and pixel independence might not be sufficient in tricky cases. The inclusion of a

λ-map as described in Section 3.3 of Chapter 3, or other kind of distributions, will undoubtedly improve the performance of the model, as more complex information can be taken into account.

Secondly, only the influence of the NSB has been investigated with such modules. Yet, another interesting application is addressing the pointing direction variations through their integration as conditioning inputs. Such experiments will be conducted in the future, following this evaluation plan:

- The baseline consists in training the γ-PhysNet on a specific pointing, for example a zenith of 20 degrees and an azimuth of 180 degrees. A first step consists in evaluating the loss in performance by testing it following a declination line, that is to say with rising zenith.

- The γ-PhysNet-CBN can in a second time be trained with a large training set with multiple pointings. The evaluation is therefore twofold: the performance on each training pointing will be assessed, followed by its analysis on a pointing that do not belong the training set. This way, the capability of the model to interpolate between angles will be investigated.

Lastly, it is possible to use our CBN-based neural network for fine tuning purposes, as it has originally been proposed in the founding article [Vri+17]. To this end, a general pre-trained model using Batch Normalization layers can be optimized, and then can be fine-tuned for any run-wise application to a specific observation condition or zenith angle replacing the corresponding layers with CBN modules. Fine-tuning will not only save time and computation cost, but also leverage the knowledge introduced by the pre-training, in the same way as Transformers and MAE.

The γ-PhysNet-DANN

Undoubtedly, the greatest opponent of the γ-PhysNet-DANN is the difficulty of its integration within the multi-task context. The conflicting gradients between domain adaptation and particle classification make it challenging to optimize properly. Thus, the implementation of other conflicting resolving strategies, using the gradient cosine similarities as inputs, may be beneficial to improve its results. Moreover, in the case of real scenario, the domain task coupled with Uncertainty Weighting does not allow the network to converge most of the time. Hence, it has been necessary in this work to set manually the weighting hyperparameter for this task, leading to time-consuming trials and errors.

It is also possible to improve the variety of observed conditions in the target training set by implementing another sampling strategy for the real acquisitions. In fact, the size of the observation run doesn't allow to store all the images in memory. This imposes to process the current dataset to create a new one. In that case, depending on the number of events to select from the acquisitions, a stride can be computed. The stride is an integer corresponding to the number of events to skip after selecting one. Because there are much more telescope acquisitions than simulations, it can be computed as the ratio of the number of real data divided by the number of simulated data. This new dataset will contain a better diversity of conditions compared to the method currently implemented, described in 5.8 of Chapter 5.

Finally, the combination of both unsupervised domain adaptation and CBN modules can be beneficial to successfully tackle both known and unknown differences.

The γ-PhysNet-Prime and Megatron

A lot of work is still needed to understand the benefits of the MAE pre-training. Furthermore, as proposed in Section 3.6 of Chapter 3, a modality-translation seems a promising approach to integrate physics knowledge into the network, by forcing the parameters of the model to learn the links between the pixel charge and temporal map.

In addition, only one initialization seed has been produced to compare our Transformer approaches with the standard analysis and the γ-PhysNet. Nevertheless, it is important to understand the impact of parameter initialization and data shuffling on both the pre-training and the fine-tuning steps, although this will require intense training computation.

Finally, as explained in Section 2.2 of Chapter 2, the rise in the size of the datasets, the time and computational resources have allowed the performance increase of deep learning models. Yet, a new paradigm, denominated frugal learning, has emerged as a recipe book that aims at reducing both energy consumption and the volume of data necessary for model training, without compromising the quality and robustness of the algorithms [Evc+21]. It mainly involves reduced training dataset size (input frugality), limited computational and memory resources (learning process frugality) and fewer capabilities (model frugality). The use of MAE and a smaller model size with the γ-PhysNet has been a first step towards the integration of this concept into our Transformer. Leveraging the system sensor properties or a more physics-based pre-training, as mentioned in Section 3.6 of Chapter 3, are interesting guidelines that could potentially lead to faster optimization. Yet, more research is necessary to characterize their benefits.

Short term perspectives: Improvement of the high-level analysis

The high-level analysis used in this thesis for the comparison of the methods has been performed with GammaScan. Several improvements can be considered for future studies.

As a first application of our proposed methods, four runs of the Crab have been selected, representing around 80 minutes of acquisitions. Although they provide a pertinent sandbox to test and compare models, they do not represent a large quantity of data. As a comparison, the full characterization of the Crab in [Abe+23] consisted in 50 hours of observation. In our case, the limited amount of data engenders statistical fluctuations, especially in the optimization and analysis phases, displaying possible differences in the optimized cuts from one run to another. Therefore, in the context of few data, another more relevant strategy can be applied: the optimization can be performed selecting even-numbered sub-runs, and the analysis selecting odd-numbered sub-runs. In that case, fluctuations introduced by NSB, zenith angles, and others external factors will be less significant.

Furthermore, in the proposed results, the optimization of the significance has only been performed on the gamma/hadron cut, also named gammaness. This imposes a strong assumption that the gamma-ray source is localized within the region of the sky of radius $\theta^2 = 0.2$ degree. However, because of the instrument, observation conditions, and source-related parameters, its size might vary. A constant value of θ^2 combined with a small-sized

source will ultimately contain more background than it should, degrading the value of the significance. Oppositely, if the cut value is too small, some amount of signal will be lost.

In addition, the optimization of the significance imposes to find which pairs of cuts, gammaness and θ^2, gives the best value. They are currently computed on the whole energy range, and each bin of energy receives the same optimal pair. Yet, the telescope Point Spread Function, the power-law characteristic of the gamma-ray flux, or a unfavorable Signal-Noise Ratio impose the cuts to be rather calculated as an energy-dependent function. In that case, each bin will receive a different optimized gammaness and θ^2. This improvement will especially benefit deep learning methods compared to the standard analysis, as they exhibit better performance as lower energy levels.

Finally, in this thesis, our methods have been tested on the Crab Nebula, a well-known and very bright reference gamma-ray source. Yet, it would be interesting to evaluate the behaviour of our contributions on fainter sources. It is possible to simulate this effect in the high-level analysis by considering different percentage of the flux, for example 10%. In that case, it will possible to compare each model and assess their performance in a more complicated scenario. We have selected two other interesting study cases: Markarian 421 and BLLac. Both sources are also well-known from the astrophysics community, and on the contrary to the Crab, their flux is not constant in time, making their analysis more complicated.

Long term perspectives: Stereoscopic reconstruction and analysis

Although both MAGIC observatory and LST-1 have already been associated for joint observations [Abe23], the end of the year 2024 will witness the crowning achievement of the second Large-Sized Telescope on the North CTAO of La Palma, and will give the starting signal to the debut of CTAO stereoscopic reconstruction. After completion, the project will possess more than 60 instruments working in symbiosis on the detection of EAS. Yet, the information combination of many instruments with a variety of sensors is a challenging step. In fact, the spatial coverage of the Cherenkov light will solely trigger a subset of telescopes, but rarely all of them. The fusion strategy has to take into account this aspect.

Previous works have tried to combine stereoscopy with IACT simulations. In [Shi+19] and [Mie+21], each telescope acquisition inputs a recurrent cell, and the problem of time ordering is solved by considering the trigger time of the detected event as the sorting element. Conversely, the strategy consisting in concatenating each image together into a multi-channel input comes with several drawbacks. Firstly, different sensors with different resolution will share the same backbone, which is not suitable as they don't necessarily share the same information. Secondly, the input data may be very sparse, depending on the number of telescopes that have triggered. As proposed in the thesis [Jac+19], indexed convolutions can be used in conjunction with Delaunay triangulation to merge each telescope features before the multi-task branches.

Graph Neural Network (GNN) are currently internally explored by Hana Ali Messaoud at LAPP as a promising information merging methodology. Introduced in 2009 in the pioneer work of [Sca+09], they became popular in 2017 when a variant, the Graph Convolutional Network (GCN), has been proposed in [KW17]. They are a class of neural networks designed to work with graph-structured data. Unlike CNN-based models that operate on grid-like data, GNNs can handle arbitrary structures, making them well-suited for applica-

tions where data is naturally represented as a graph, typically an array of telescopes. Thus, they are an encouraging method for the stereoscopy challenge. They have been applied for the first time for IACT background rejection in [Glo+23], evincing promising results compared to the baseline composed of Boosted Decision Trees.

Long term perspectives: Explicability

A complementary research direction relates to model explicability. Neural networks are often described as "black boxes". In fact, their internal functioning operates complex transformations as data passes through the network layers, making it difficult to understand how they produce specific predictions. Improving the explicability of neural networks is a current research trend commonly referred to as XAI [van+22]. Some approaches include visualization techniques which help detect which parts of the input data are influencing the model's decisions. The Grad-CAM algorithm, described by [Sel+19], has been implemented by [Jac20] as a preliminary step to make the γ-PhysNet model more interpretable. As expected, the neural network utilizes the signal as the main contributing factor to predict each task output. Yet, it is interesting to see that the background also has a small impact on the decision. However, such approaches don't allow understanding failure cases.

As another strategy, Bayesian neural networks [Jos+22] are a popular framework for computing uncertainty. Traditionally, computing the posterior distribution in these models is intractable due to the large number of parameters. To address this, approximation methods can be employed. In particular, sampling methods utilize Monte Carlo integration and approximate expected values of a finite set of random samples. Conversely, authors of [GG16] implements dropout at each stage of the network. During inference, dropout is applied, and each inference pass generates a distinct result from a different set of parameters. Such approaches provide an estimate of the uncertainty, but it remains unclear how much confidence can be placed in these uncertainty estimates. The more forward passes are performed, the more confident one can be in the results. However, this also increases the inference time, which can be problematic when working with huge data volumes. Nevertheless, XAI would enhance the credibility of neural network results and provide insights on both model understanding and the physics explored with CTAO.

Long term perspectives: Real-time analysis

Since 2024, GammaLearn has received new funding from the ANR through the project DIRECTA (ANR-23-CE31-0021) which aims at integrating deep learning into the Real-Time Analysis (RTA). Its implementation is crucial for an efficient observation of transient phenomena. On the one hand, the RTA can be used to emit alerts to other observatories in the world in the case of important fluctuations of the source flux, or for serendipitous discovery of a new one in the instrument field of view. On the other hand, when following alerts from other observatories, data must be analysed very quickly to ensure of the relevance to continue or stop their observations. Currently under testing phase, its application to the LST-1 aims to reconstruct events at the speed of 1500 events/s/CPU. As a reminder, a classic trigger rate is around $3-10$k images per second for a single telescope.

Relying on the Hillas+RF algorithm, the current pipeline would greatly benefit from deep learning methods, particularly for the lowest energetic part of the event distribution.

In particular, CBN-based or global models are adapted to multiple observation conditions, like varying zenith, azimuth, NSB and stars in the field of view. In fact, tackling such conditions in real time is critical because the results are used as feedback by the telescope operators during the observations.

The challenges that arise from the application of deep learning to real-time analysis are very different from the standard analysis. In our case, it scales mostly with the availability of computation resources and the memory consumption of the network. For example, more GPU memory allows a bigger batch size. Moreover, a trade-off can be expected between the size of the model and the time to compute the forward pass, as searching for frugality consider both model design and hardware considerations. In other words, refinements are mostly impacting the architecture of the network, while classical Computer Vision usually needs mathematical approximations, or shortcuts, to avoid computing difficult operations.

Several short term and long term directions are now unlocked and open to the use of deep learning approaches in the specific domain of astrophysics. Additionally, this work paves the way for broader research on domain adaptation and complementary strategies to enhance model generalization, making them applicable and robust in real-world scenarios.

7 Résumé long en français

7.1 Astronomie gamma

L'astronomie gamma et ses intérêts scientifiques

L'astronomie gamma est la branche de l'astronomie qui étudie les objets célestes et leurs phénomènes astrophysiques associés par la détection et l'analyse des rayons gamma, la forme de rayonnement électromagnétique la plus énergétique. Par définition, les rayons gamma possèdent des longueurs d'onde extrêmement courtes, et par conséquant des énergies très élevées, bien au-delà de celles de la lumière visible. Leurs implications dans la compréhension de nombreux événements physiques ont récemment suscité un intérêt accru, du fait qu'elles possèdent intrinsèquement des propriétés intéressantes. Par exemple, s'agissant de particles non chargées, elles ne sont pas déviées par les champs magnétiques présent sur leur trajectoire lors de leur parcours dans l'Univers. Cela rend possible le traçage de leur origine, contrairement aux hadrons.

L'origine et le rôle des particules cosmiques relativistes, accélérés à des énergies extrêmement élevées, intriguent les astrophysiciens depuis peu, bien qu'elles bombardent l'atmosphère terrestre depuis toujours. Les mécanismes d'accélération, ainsi que leur rôle dans la formation des étoiles et l'évolution des galaxies restent des sujets ouverts. De plus, ils pourraient fournir des précieux indices sur la physique des trous noirs, de la matière et de l'énergie noires. Par ailleurs, grâce aux récents innovations technologiques conduisant à la construction de nouveaux détecteurs, l'astronomie multi-messagers, combinant rayons cosmiques, neutrinos et ondes gravitationnelles, devient un champ majeur pour explorer les mystères de l'Univers. Les enjeux autour de la détection et de la compréhension des rayons gamma sont donc fondementaux.

Ils peuvent être détecter à travers les yeux d'instruments communément appelé Télescopes à Imagerie Tcherenkov Atmosphérique (Imaging Atmospheric Cherenkov Telescope, IACT). Ces instruments peuvent être à la fois spatiaux ou terrestes. Lorsque l'observatoire en question se situe sur Terre, l'observation des rayons gamma repose sur l'utilisation de l'atmosphère comme un calorimètre[1]. Lorsqu'un rayon pénètre dans l'atmosphère, il interagit avec les molécules environnantes et engendre la production d'une gerbe atmophérique. Comme ces gerbes se déplacent plus rapidement que la lumière dans ce milieu, elles émettent un faisceau de lumière Tcherenkov. Ces photons émis sont captés par le système optique du télescope, tel qu'illustré par la Figure 7.1. C'est en 1989 que l'observatoire Whipple a ou-

[1] Le mot *calorimètre* peut avoir des significations différentes selon les domaines de la physique dont on parle, et son usage en physique des particules est quelque peu figurative, puisqu'aucune mesure de chaleur n'est impliquée.

vert la voie à ce domaine en détectant pour la première fois une source de rayons gamma via ce procédé, la Nébuleuse du Crabe.

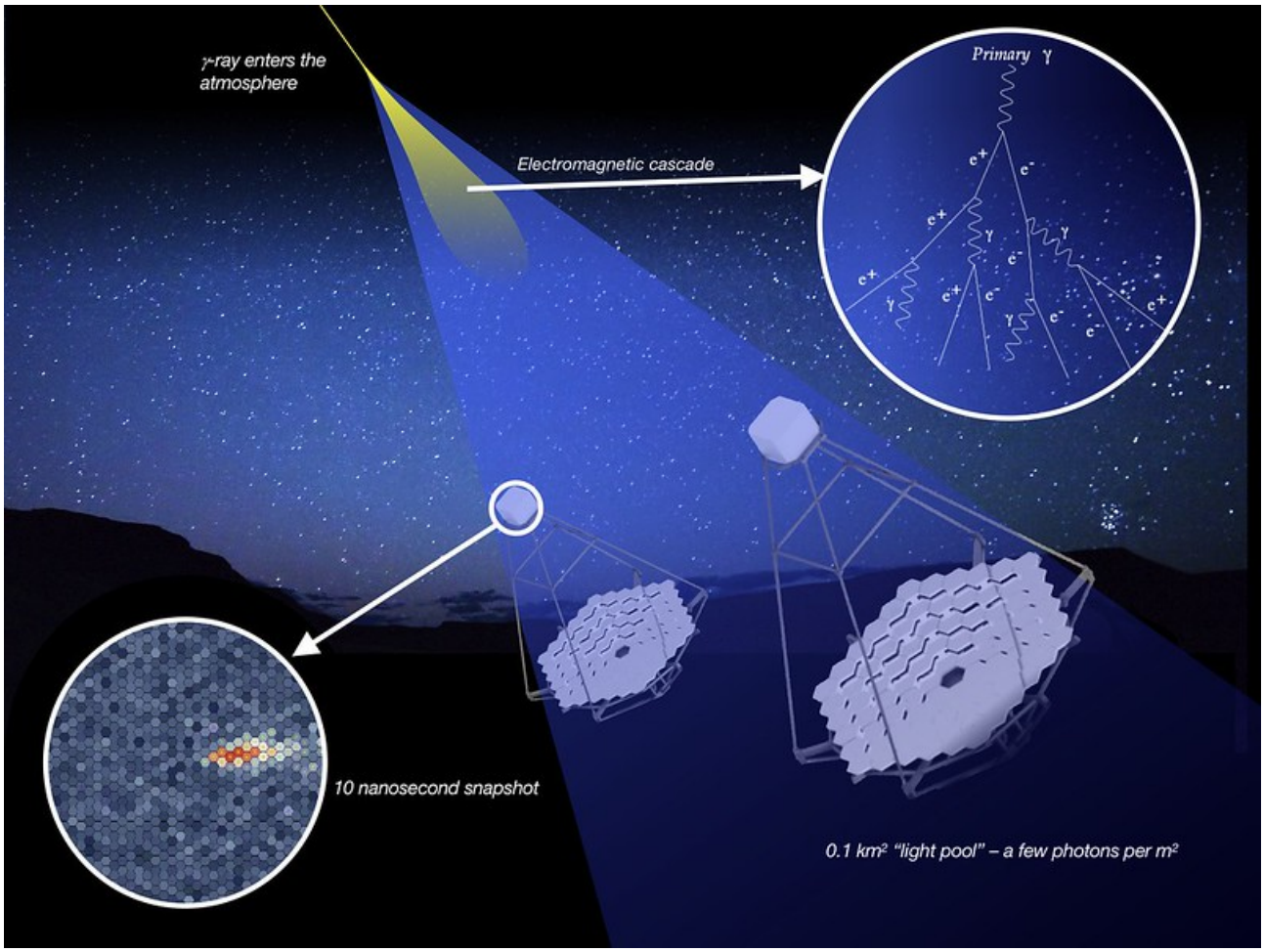


Figure 7.1: Principe de détection d'une gerbe atmosphérique. Source [].

Le projet Cherenkov Telescope Array Observatory

Motivé par le succès des observatoires précédents, le projet Cherenkov Telescope Array Observatory (CTAO) a été lancé en 2006 dans le but de développer la quatrième génération de IACT, améliorant considérablement la sensibilité des instruments par rapport à ses prédécesseurs [Mir22]. La couverture énergétique attendue s'étend entre autre de 20 GeV à 300 TeV.

Le Large-Size Telescope (LST) est l'instrument le plus grand du projet avec un diamètre de 24 m. Il est principalement conçu pour détecter les gerbes à basse énergie, c'est à dire peu lumineuse, jusqu'à environ 20 GeV. De part sa caméra, chaque image produite comporte 1855 pixels, disposés sur une grille hexagonale. Le premier LST (LST-1) a été inauguré en octobre 2018 et a terminé sa phase de mise en service fin 2023. Il est actuellement le seul instrument en opération, et a déjà conduit ses premières détections [Abe+23].

Analyse des données

De par la frugalité du flux de gamma incident, et de sa décroissance exponentielle en fonction de l'énergie, l'observation d'une source peut se faire grâce à des temps d'exposition relativement long, généralement compris au minimum entre quelque dizaine de minutes pour une source très brillante et quelque dizaine d'heures autrement. Au rythme d'environ 5000 images par seconde, cela peut représenter une immense quantité de données à analyser. Afin de limiter la prise en compte de fausses alarmes, la caméra déclenche son processus d'enregistrement que si l'intensité de la caméra dépasse un seuil prédéfini. L'événement considéré est alors capturé au rythme d'une image toutes les nanosecondes pendant 40 nanosecondes. Cette séquence, qui représente la lumière émise lors du déployement d'une gerbe dans l'atmosphère, est ensuite calibrée et intégrée en deux images correspondant à la charge des pixels et à l'information temporelle de déclenchement des capteurs. Cette dernière est importante pour lever l'ambiguité de la direction de la cascade. L'ensemble de ses images enregistrées pendant les observations forme la base de données de la source et constituent les acquisitions à analyser.

L'étude des sources de rayons gamma au sein du CTAO est actuellement réalisée avec des méthodes d'apprentissage automatique. Elle consiste à extraire quelques paramètres morphologiques du signal avec l'algorithme Hillas, présenté dans [Hil85], après un processus de nettoyage visant à éliminer le bruit de fond. Ce bruit, appelé Night Sky Background (NSB), correspond à la lumière résiduelle provenant de la lueur nocturne, des étoiles dans le champ de vision, etc. Les paramètres calculés sont ensuite injectés dans des Random Forest (RF) pour déterminer les grandeurs physiques de la particule. L'ensemble de la procedure est par la suite nommé Hillas+RF, et est décrite plus précisement dans [al08]. Cependant, cette procédure standard présente des limitations, en particulier aux faibles énergies, où il devient difficile d'extraire des paramètres morphologiques pertinents lorsque le signal est contenu dans quelques pixels. Par conséquent, l'apprentissage profond, qui utilise la connaissance contextuelle de toute l'image et qui fonctionne sans les contraintes imposées par les paramètres Hillas, pourrait potentiellement mener à des résultats améliorés.

7.2 Apprentissage profond et ses applications à l'astronomie gamma

Introduction

La science des données est une discipline récente qui vise à extraire de l'information à partir de données dont la nature et les questions qu'elles adressent peuvent varier considérablement. Bien qu'elle englobe de nombreuses techniques, l'Intelligence Artificielle (IA) se distingue comme l'une de ses composantes les plus importantes. L'IA fait référence à la notion large de création de machines capables de réaliser des tâches attribuées généralement à l'intelligence humaine, impliquant raisonnement, résolution de problèmes, prise de décision, créativité ou innovation [GBC16]. Cela peut être réalisé à l'aide de diverses techniques, y compris les systèmes basés sur des règles, la logique symbolique, les systèmes experts, l'apprentissage automatique, et plus récemment, l'apprentissage profond. Fondamentalement, il s'agit d'une discipline à la frontière entre les mathématiques et l'informatique, visant à résoudre une grande variété de tâches complexes, généralement liées à la vision

par ordinateur, au traitement du langage naturel ou à la reconnaissance audio. Récemment, l'apprentissage profond a été appliqué avec succès dans de nombreux cas de traitement de signal et d'image, et l'analyse d'images multimédias a encouragé l'extension de cette technique à d'autres domaines tels que l'astronomie [Jac20].

Application à l'astronomie gamma

L'application de l'apprentissage profond pour la reconstruction des attributs physiques des particules incidentes vise à résoudre les problèmes rencontrés par les approches classiques d'apprentissage automatique. Elle ne nécessite pas de prétraitement des données via des procédures de filtrages et exploite l'intégralité de l'information des contenues dans les images, avec le potentiel d'améliorer les performances, notamment à basse énergie. Une revue des méthodes existantes a été établie par [DK24]. Parmi ces approaches, le γ-Physnet, introduit dans la thèse [Jac20] est le premier réseau de neurones dédié à la reconstruction des événements pour le CTAO et à la détection des sources de rayons gamma. Contrairement à l'analyse standard, ce réseau reconstruit chaque paramètre simultanément au sein de la même architecture. Il a été appliqué avec succès à la fois sur des simulations et des données réelles dans [Vui+21]. Concernant son évaluation sur les données de synthèse, le γ-PhysNet surpasse clairement Hillas+RF sur l'ensemble des figures de mérite, notamment à basses énergies. De plus, l'analyse du Crab montre que plus de gamma sont détectés, et la capacité de détection est supérieure.

Limitations des approches d'apprentissage profond

L'application de γ-PhysNet aux données réelles révèle à la fois les forces et les limites des approches actuelles d'apprentissage profond. D'une part, l'exploitation du bruit de fond des images, combinée à l'optimisation des filtres de convolution par des méthodes basées sur les données, permet de détecter un plus grand nombre d'événements gamma. Cela permet ainsi de calculer des métriques de performance de haut niveau avec plus de statistiques par rapport à l'analyse standard. D'autre part, ces filtres personnalisés présentent un inconvénient majeur : l'entraînement du réseau sur des données simulées manque de généralisation et introduit des biais lorsqu'il est appliqué à des acquisitions réelles. Pour remédier à cela, l'adaptation des données, qui consiste à ajuster le NSB en fonction des données à analyser, s'est avérée essentielle pour améliorer les performances du réseau [Vui+21]. Cette nécessité met en évidence l'importance de rechercher des méthodes disposant de plus grande capacité de généralisation, menant techniques de réduction de divergence entre domaines, communément appelées adaptation de domaine et fusion d'informations.

Divergence de domaine

L'apprentissage automatique traditionnel repose sur l'hypothèse que les données de test et d'entraînement sont issues de la même distribution. Cependant, comme nous l'avons vu, en pratique, cette condition idéale est rarement respectée, entraînant des biais. Selon l'origine de la divergence, ce décalage peut être divisé en plusieurs catégories, comme expliqué par [KF14].

La notion de covariate shift survient lorsque la distribution des données d'entrée change, mais que la distribution conditionnelle des labels étant donnée l'entrée reste la même. Cela est courant dans les applications de l'IACT, principalement dû aux variations de NSB [PMO22]. D'autre part, le concept shift apparaît lorsque la distribution conditionnelle des labels étant donnée l'entrée change entre les distributions. Par exemple, dans les simulations LST, une particule simulée d'une énergie et origine données peut différer d'une particule réelle ayant les mêmes caractéristiques. Egalement, il est souvent impossible de garantir que les distributions des labels entre la source d'entraînement (simulations) et la cible (données réelles) soit identiques. Ce cas est appelé label shift. De nombreuses approches de la littérature reposent sur la pondération par importance, qui consiste à estimer le rapport des classes source et cible et à ajuster en conséquence les distributions [Liu+21a; Zha+13; Azi+19; LWS18]. Bien que cette approche améliore les performances, elle n'est applicable que lorsque ce ratio reste acceptable. Dans l'IACT, le flux de particules suit différentes lois de puissance et décroît de manière exponentielle avec l'augmentation des énergies. De plus, les données d'entraînement simulées doivent être équilibrées et représentatives pour chaque classe, faisant de l'astrophysique un cas d'application fortement affecté par le label shift.

Adaptation de domaine non-supervisé

Le problème de divergence nécessite de développer de nouvelles techniques permettant de généraliser sur des domaines connexes mais différents. D'un point de vue mathématique, les auteurs de [WD18] définissent un domaine $\mathcal{D}$ comme la combinaison d'un espace de caractéristiques $\mathcal{X}$ et d'une distribution de probabilité marginale $P(x)$ sur un ensemble de données D. Dans le cadre de l'apprentissage par transfert, l'adaptation de domaine consiste à transférer des connaissances d'un domaine source labelisé vers un autre domaine cible distinct. Selon la disponibilité des labels dans le domaine cible, cette adaptation peut être classée comme supervisée, semi-supervisée ou non supervisée, comme proposé dans [Zha+20]. Dans ce dernier cas, on en fait référence par l'abbréviation UDA (Unsupervised Domain Adaptation). En principe, elle repose sur l'hypothèse de covariate shift, stipulant que la distribution conditionnelle des labels selon les entrées ne change pas entre les deux domaines, comme expliqué dans [Ben+06]. Dans le cas de la détection de sources de rayons gamma, le modèle paramétrique f^θ permet de retrouver les propriétés physiques des événements incidents, où $f^\theta(x) = P_S(y|x) = P_T(y|x)$. Dans le contexte de l'UDA, les principales approches peuvent être classées en quatre groupes [Zha+20].

Méthodes basées sur la divergence

Les méthodes basées sur la divergence visent principalement à minimiser une mesure de distance (ou parfois de divergence) entre distributions. Bien que le calcul de ces distances nécessite théoriquement la connaissance des distributions sous-jacentes, les réseaux de neurones offrent la flexibilité de les estimer de manière stochastique. Il existe plusieurs métriques standard, impliquant par exemple le calcul des moments. Dans le cas des statistiques du premier ordre, on parle de Maximum Mean Discrepancy (MMD) [Tze+14], dont l'approximation empirique peut être calculée sur les couches de caractéristiques intermédiaires des réseaux de neurones. De plus, les auteurs de [Lon+15] introduisent des noyaux caractéristiques, qui contribue à améliorer la performance de classification. Deep Correlation Alignment (Deep-

CORAL) [SS16] vise à minimiser la distance via des statistiques d'ordre deux calculées à partir des couches latentes.

Réduire les moments d'ordre un et deux n'étant pas adapté aux distributions non gaussiennes, Higher-order Moment Matching (HoMM) [Che+19] unifie ces approches en les étendant à des statistiques d'ordre supérieur couplées à des noyaux caractéristiques. De manière similaire, Deep Joint Distribution Optimal Transport (DeepJDOT) [Dam+18] s'appuie sur le transport optimal pour estimer la distance de Wasserstein entre les deux domaines.

Méthodes adversaires discriminatives

Deuxièmement, les modèles discriminatifs adversariaux mettent en œuvre un discriminateur de domaine avec une fonction objective adversaire pour apprendre des caractéristiques qui soient invariantes en fonction des domaines considérés, comme le Domain Adversarial Neural Network (DANN) introduit dans [Gan+16]. Dans ce but, l'extracteur de caractéristiques est complété par un objectif supplémentaire qui consiste à tromper le classificateur de domaine, garantissant ainsi l'invariance au domaine tout en fournissant de bonnes performances sur la tâche de classification ou de régression. En tant que problème d'optimisation adversaire, il peut être formulé à l'aide d'une Gradient Reversal Layer (GRL), denotée $\mathcal{R}$, représentant une pseudo-fonction qui inverse le signe du gradient pendant la rétropropagation. En pratique, le gradient propagé retourné par le classificateur de domaine est pondéré par une fonction dépendante de l'époque, ce qui permet d'entraîner d'abord le réseau sur les tâches supervisées avant d'incorporer progressivement le processus d'adaptation.

Pour finir, le classificateur de domaine peut être remplacé par une métrique de domaine. Étant donné que la distance de Wasserstein présente de meilleures propriétés pour les gradients pendant la rétropropagation, elle est considérée comme un outil prometteur pour mettre à jour et améliorer les performances du modèle [She+18]. Cette approche nécessite d'imposer une contrainte de Lipschitz sur la partie du réseau qui calcule la métrique, ce qui peut être facilement implémenté avec une pénalisation appliquée à la dérivée des gradients [Gul+17].

Méthodes adversaires génératives

À l'inverse, les méthodes génératives reposent sur les Generative Adversarial Networks (GANs) [Goo+14] et reposent sur la translation d'images [Iso+16]. COGAN [LT16] est un exemple de ce type d'algorithme combinant des images non appariées avec l'utilisation de deux GANs synthétisant des images de chaque domaine. CycleGAN [Zhu+20] introduit la contrainte de cohérence cyclique pour effectuer cette translation garantir que les fonctions sont bijectifs. Alors que les modèles discriminatifs visent à obtenir des caractéristiques invariantes au domaine, ces approches cherchent à projeter une distribution sur une autre. Basé sur les Transformers, Image-to-image Translation with Transformers [Zhe+22] suit une architecture encodeur-décodeur et combine des convolutions avec des modules d'attention dans un bloc de perception hybride pour mettre en évidence les dépendances à court et long terme.

Méthodes auto-superisées

Enfin, les méthodes auto-supervisées intègrent des tâches d'apprentissage supervisées auxiliaires au sein du réseau de la tâche principale. Ces méthodes, comme le Multitask Auto-Encoder [Ghi+15], apprennent à transformer la distribution initiale en distributions analogues, de sorte que les caractéristiques extraites de l'encodeur soient robustes aux variations de domaine. Ensuite, ces caractéristiques invariantes sont utilisées comme entrée pour un classificateur ou un régresseur.

Fusion d'information

Les modèles d'apprentissage automatique rencontrent souvent des difficultés en matière de robustesse lorsque la distribution des données de test diffère de celle de l'entraînement. Même si la représentativité des données peut être suffisante, des limitations peuvent survenir car le réseau de neurones n'a pas accès à des paramètres externes pourtant cruciaux. Ainsi, il devient essentiel d'améliorer la visibilité de ces informations supplémentaires. Par exemple, dans le cas de la classification de terrains, les satellites sont équipés de plusieurs capteurs capturant des images multispectrales et les entrées deviennent naturellement multi-canales et offrent un contexte plus riche. Cette approche d'apprentissage vise à exploiter des informations complémentaires provenant de différents aspects afin d'améliorer les performances des modèles d'apprentissage automatique, les rendant plus aptes à comprendre et interpréter les données de manière similaire aux humains [Ion+14].

Les mécanismes d'attention [Vas+17] permettent aux réseaux de neurones de se concentrer sur les parties les plus pertinentes des données d'entrée en pondérant dynamiquement l'importance des différentes caractéristiques. Leur conception est bien adaptée à la fusion d'informations. Dans le cas des modèles Transformeurs, les variables externes peuvent être projetées à l'aide d'un MLP entraînable, qui peuvent ensuite être intégrés dans la représentation des tokens d'image. De plus, la combinaison d'entrée peut se faire à différent stades du réseau de neurones à travers des couches de concaténation et de projection. Pour finir, les variables supplémentaires peuvent être utilisées comme entrées de conditionnement via la conception de modules de Normalisation par Batch Conditionnelle (Conditional Batch Normalization, CBN), comme introduit dans [Vri+17].

7.3 Contributions

Fusion d'information

Initialement introduit comme une astuce pour moduler des modèles pré-entraînés, la CBN vise à remplacer les couches de normalisation pour permettre de fine-tuner les réseaux avec la possibilité d'ajouter des variables de conditionnement supplémentaires. Classiquement, elle offre la flexibilité de calculer des paramètres supplémentaires, notés $\Delta\gamma_c$ et $\Delta\beta_c$, et de les ajouter aux paramètres traditionnels de moyenne γ_c et d'écart-type β_c. Les décalages sont obtenus en projetant les variables de conditionnement c avec un encodeur supplémentaire denoté G_c dans un espace de caractéristiques latentes commun, noté $z = G_c(c)$. Chaque module CBN, indexé par l'entier k, possède ses propres paramètres qui calculent ces décalages. En résumé, $\Delta\gamma_c^k = f_\gamma^k(z)$ et $\Delta\beta_c^k = f_\beta^k(z)$, où les fonctions f_γ^k et f_β^k sont des MLP qui

calculent les décalages des $k^{\text{ème}}$ modules CBN à partir de l'espace latent conditionné. Dans ce qui suit, nous omettons l'indice k par soucis de simplification. La formulation générale peut facilement être étendue à partir de l'équation originale.

$$O_{i,c,x,y} = (\gamma_c + \Delta\gamma_c)\frac{F_{i,c,x,y} - \mu_c}{\sqrt{\sigma_c + \epsilon}} + (\beta_c + \Delta\beta_c), \forall i, c, x, y \tag{7.1}$$

Nous faisons référence à γ-PhysNet-CBN lorsque les modules CBN sont intégrés dans au modèle γ-PhysNet de [Jac20]. Les modules CBN permettent d'apporter au réseau de la robustesse face au NSB. Dans sa forme la plus simple, le bruit de fond simulé injecté dans le jeu de données synthétique correspond à un bruit de Poisson paramétré par λ_{MC} et est constant sur l'ensemble des pixels. Pour rappel, cette approche est basée sur l'idée que l'adaptation des données, comme décrite précédemment, vise à modifier le jeu de données d'entraînement pour correspondre aux variations réelles du bruit de fond. En tenant compte le fait que le NSB est soumis à des variations temporelles, nous proposons de considérer des perturbations plus réalistes en échantillonnant le taux λ_{MC} à partir d'une distribution uniforme, de sorte que $\lambda_{MC} \sim \mathcal{U}(a, b)$ où a et b sont les bornes inférieure et supérieure. Par conséquent, il est possible de considérer ce taux échantillonné comme entrée de conditionnement. Ce scénario correspond typiquement à une augmentation des données.

Adaptation de domaine non-supervisée

Selection des méthodes

Nous avons sélectionné trois méthodes pertinentes d'adaptation de domaine non-supervisée, à savoir DANN [Gan+16], DeepJDOT [Dam+18] et DeepCORAL [SS16]. Ces méthodes appartiennent aux catégories de l'adaptation adversaire et basées sur la divergence. Premièrement, DANN est l'une des approches les plus populaires et a été appliquée avec succès un grand nombre de contextes. Elle repose sur l'estimation de la $\mathcal{A}$-distance entre les domaines et bénéficie d'un fort soutien théorique [Ben+06]. Deuxièmement, la distance de Wasserstein présente d'excellentes propriétés de convergence par rapport à d'autres distances, comme démontré dans [ACB17], telles que la divergence de Jensen-Shannon ou de Kullback-Leibler. Enfin, étant donné que le NSB est reconnu comme l'un des principaux contributeurs de divergence entre les simulations et les données réelles, minimiser la différence entre les statistiques du premier et deuxième ordre semble être une stratégie adéquate, faisant de DeepCORAL un choix intéressant pour notre application. En particulier, le NSB est généralement approximé par un bruit de Poisson, dont le paramètre peut être estimer par le calcul de la moyenne ou de la variace de la distribution.

Prise en compte du décalage de labels extrême

Suivant les travaux de [Liu+21a] sur la pondération par importance, nous introduisons les versions conditionnelles de DANN (CDANN), DeepJDOT (CDeepJDOT) et DeepCORAL (CDeepCORAL) pour faire face au décalage de labels extrême rencontré avec le jeu de données du LST. Étant donné que le rapport entre les gammas et les protons est approximativement estimé à partir des observations, il est possible d'évaluer la valeur du paramètre de rééquilibrage par classe $\omega \in \mathbb{R}^2$. En effet, si $\omega = \begin{pmatrix} \omega(\gamma) & \omega(p) \end{pmatrix}$, le coefficient de pondération peut être déterminé comme suit :

$$\begin{aligned} \omega(\gamma) &= \frac{P^T(\gamma)}{P^S(\gamma)} = \frac{n_T^\gamma}{n_S^\gamma} = \frac{\epsilon}{n_S^\gamma} \approx 0 \\ \omega(p) &= \frac{P^T(p)}{P^S(p)} = \frac{n_T^p}{n_S^p} = \frac{n_S^p + n_S^\gamma - \epsilon}{n_S^p} > 1 \end{aligned} \tag{7.2}$$

où $\epsilon = n_T^\gamma$ désigne la quantité résiduelle de gammas dans l'ensemble cible, et la composition du mini-batch source ou cible respecte l'équation suivante :

$$n_S = n_S^\gamma + n_S^p = n_T^\gamma + n_T^p = n_T \tag{7.3}$$

Du point de vue de l'entraînement du modèle, cela équivaut à minimiser la divergence entre les protons simulés et l'ensemble de données cible. Nous notons alors $\mathcal{D}_S = \{p_S^{(i)}, \gamma_S^{(j)}\}_{i,j}$ les données sources contenant les protons sources $\{p_S^{(i)}\}_i$ et les gammas sources $\{\gamma_S^{(j)}\}_j$. De manière similaire, les données cibles sont définies comme $\mathcal{D}_T = \{p_T^{(i)}, \gamma_T^{(j)}\}_{i,j} = \{x_T^{(k)}\}_k$, puisqu'elles sont indifférenciées par nature. Nous introduisons ensuite des stratégies de conditionnement dans les méthodes d'optimisation suivantes.

Application à DANN Pour DANN, la mise en œuvre de notre repondération par importance adaptée peut être réalisée en réécrivant la fonction de perte de DANN établie dans [Gan+16]. La perte du classificateur de domaine est alors masquée en fonction des labels des particules sources, et la fonction de perte devient :

$$\mathcal{L}_{DANN} = \sum_{i | x_i \in \{p_S, x_T\}} CE\left(D(\mathcal{R}[G(x_i)]), d_i\right) \tag{7.4}$$

Application à DeepJDOT Dans le cas de DeepJDOT, le conditionnement est pris en compte lors du calcul de la matrice de coûts. Elle est calculée uniquement entre les protons sources et les mini-batch cibles échantillonnés uniformément afin de maintenir des matrices carrées. Le plan de transport optimal est ensuite déduit à partir de ces échantillons sélectionnés, suivant la procédure en deux étapes proposée par [Dam+18]. Si G définit l'extracteur de caractéristiques du réseau de neurones, alors la fonction de coût est décrite comme :

$$C_{i,j} = \|G(x_i) - G(x_j)\|^2, x_i \in \{p_S^{(i)}\}_i, x_j \in \{x_T^{(j)}\}_{j==i} \tag{7.5}$$

Enfin, la fonction de perte peut être calculée à l'aide de l'équation de transport optimal standard définie dans [Dam+18].

Application à DeepCORAL Enfin, DeepCORAL calcule les statistiques en utilisant les protons sources et l'ensemble complet du mini-lot cible. La matrice de covariance source est alors calculée comme suit :

$$C_S = \frac{1}{m-1}\left(D_S^T D_S - \frac{1}{m}(1^T D_S)^T (1^T D_S)\right) \tag{7.6}$$

où $D_S = \{p_S^{(i)}\}_{i=1}^m$ désigne le mini-batch contenant m protons sources. Le calcul de la matrice de covariance cible reste inchangé. Ensuite, la fonction de perte peut être calculée comme énoncé dans [SS16].

Modèle Transformers

Le succès des méthodes de réduction de données telles que la décomposition SVD ou les autoencodeurs (AE) provient du constat que les images contiennent beaucoup de redondances. Partant de cette idée, le Masked Auto-Encoder (MAE) [He+21] est une approche Transformers qui vise à supprimer aléatoirement 75 % des tokens, afin de les reconstruire à travers une tâche de complétion. Il s'agit d'une étape de pré-entraînement visant à améliorer la généralité du modèle sous jacent lors du fine-tuning.

Concrètement, l'architecture du MAE se compose de deux entités, qui sont un encodeur et un décodeur. Les données d'entrée sont d'abord transformées à l'aide d'une projection linéaire sur laquelle sont ajoutés des encodages positionnels. Ensuite, une quantité fixe de patches est supprimée via une procédure de masquage aléatoire. Cette stratégie permet d'entraîner des encodeurs très volumineux tout en n'utilisant qu'une fraction des ressources de calcul et de mémoire, ce qui améliore considérablement l'efficacité, la générabilité et l'évolutivité du modèle. D'autre part, le décodeur vise à récupérer les parties manquantes du puzzle. Cela se fait en réintroduisant tous les tokens comme entrées, y compris ceux masqués. Un token de masque indique la présence des tokens qui doivent être prédits.

Une fois que le modèle est pré-entraîné avec un MAE, il peut finalement être affiné pour la reconstruction des paramètres physiques. Pour cela, l'encodeur est conservé afin de projeter les images dans la nouvelle représentation, tandis que le décodeur n'est plus utile et est supprimé. Comme cela se fait classiquement dans le paradigme des Transformers, des tokens supplémentaires sont concaténés à la représentation encodés, correspondant aux différent tokens de tâche. Leur particularité réside dans le fait qu'ils sont définis comme des paramètres entraînables. Nous faisons référence à notre implémentation de ce Transformers par la dénomination γ-PhysNet-Prime.

7.4 Application aux données de simulations

Introduction

Dans de nombreuses applications de physique, obtenir des données labelisées est généralement impossible en raison de nombreux processus aléatoires indescriptibles qui surviennent lors des acquisitions, tels que les conditions météorologiques, le bruit électronique, et d'autres facteurs affectant la qualité des observations. De plus, les théories sous-jascentent fournissent une description des phénomènes en jeu grâce à des équations, et l'implémentation de ces équations dans les systèmes informatiques modernes permet d'accéder à un grand nombre de simulations.

Les données synthétiques présentent de nombreux avantages. À mesure que les images générées par ordinateur deviennent de plus en plus réalistes, il est possible de produire des images plus précises, et il peut être plus facile, tant en termes de temps que de ressources, de générer des simulations plutôt que de collecter et étiqueter des données réelles. De plus,

les simulations offrent un contrôle total sur les paramètres ayant servis à la génération des données, tels que le nombre d'interactions, la physique, etc.

Jeu d'entraînement du LST

Dans le cadre de cette thèse, le jeu de données du consortium LST est constitué de simulations Monte-Carlo, généré à l'aide du logiciel CORSIKA (COsmic Ray Simulations for KAscade) [Hec+98]. Ce package est utilisé pour simuler les gerbes électromagnétiques et hadroniques dans une atmosphère simulée, permettant ainsi le suivi des trajectoires des particules secondaires jusqu'à des énergies relativement faibles. Les processus utilisés dans CORSIKA sont soutenus par des études théoriques solides et des expériences menées dans des accélérateurs et des collisionneurs de particules. De plus, le NSB est injecté sous forme d'un bruit de Poisson de manière uniformément répartie sur les pixels.

La réponse du télescope aux gerbes atmosphériques est obtenue à l'aide du logiciel sim_telarray [Ber08] à un angle de zénith de pointage de 20 degré et un azimut de 180 degré, puis prétraitée avec cta-lstchain [Lop+22] en utilisant une intégration de charge par pixel, comme décrit dans [Abe+23]. Ce jeu de données est désigné ici sous le nom de Prod5 LST-1 monotrigger. Cela donne, pour chaque événement, des paires d'images correspondant à l'amplitude de charge intégrée et au temps moyen d'arrivée des photons pour chaque pixel. Ces échantillons bimodaux sont utilisés comme entrée pour notre réseau. Comme les images se trouvent au niveau du capteur sur une grille hexagonale de pixels, nous les interpolons sur une grille régulière en utilisant la bibliothèque dl1-data-handler [Kim+22]. Ainsi, nous profitons des implémentations performantes des convolutions de PyTorch, ce qui réduit le temps d'entraînement de nos modèles, tout en obtenant les mêmes performances comparé à une approche plus adaptée au capteur qui s'appuie sur des convolutions indexées [Jac+19] pour les images à pixels hexagonaux.

Les modèles γ-PhysNet-DANN et γ-PhysNet-CDANN

Nous évaluons la performance des modèles γ-PhysNet-DANN, γ-PhysNet-DeepCORAL, et γ-PhysNet-DeepJDOT dans différents scénarios. Premièrement, nous comparons l'inclusion de chacune de ces méthodes avec les techniques d'équilibrage multi-tâches Uncertainty Weighting (UW) [KGC18] et GradNorm (GN) [Che+18], avec différentes valeurs de α. Deuxièmement, nous examinons la contribution du conditionnement pour traiter le décalage de labels. Enfin, l'impact de la Gradient Layer (GL), qui est équivalent à la couche de gradient $\mathcal{R}$ introduit dans [Gan+16] avec la possiblité d'être inversé, sur les performances du modèle est analysé. A partir de notre étude d'ablation, nous avons sélectionné γ-PhysNet-CDANN comme l'approche la plus prometteuse pour la détection de la source de rayons gamma du Crabe.

La Figure 7.2 représente la comparaison entre la version DANN et CDANN du γ-PhysNet dans un cas proche de celui des données réelles, avec une dégradation engendrée par un taux de NSB constant ainsi qu'un label shift de ratio $r = 10^{-4}$. Les conséquences du label shift sont principalement visibles sur la plage intermédiaire des énergies. Cependant, le conditionnement permet de supprimer cette effect et de retrouver des performances proches du meilleur scénario.

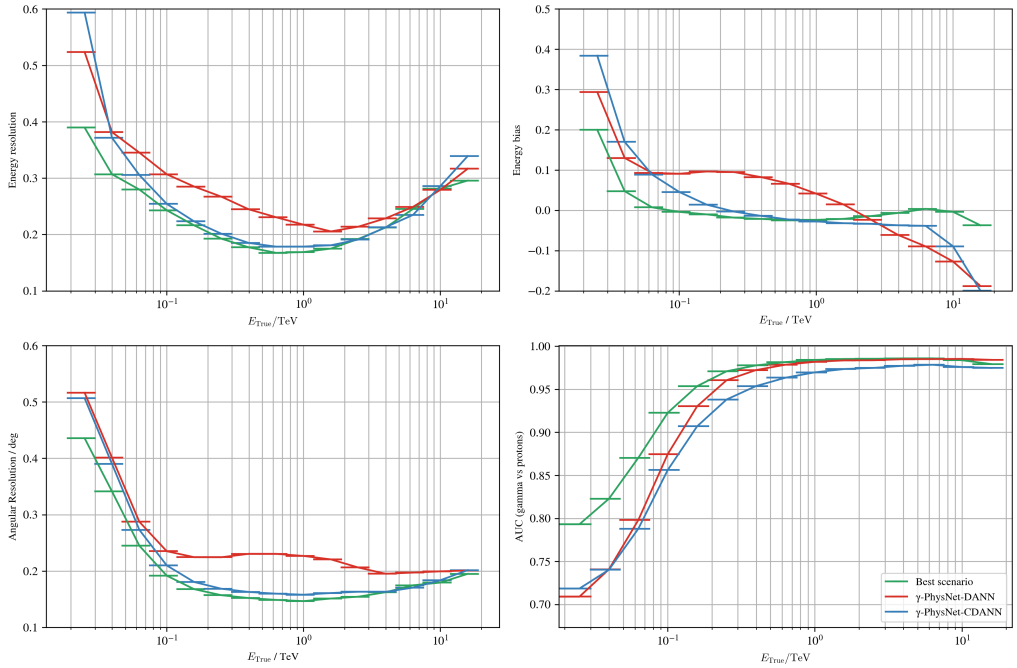


Figure 7.2: Comparaison entre γ-PhysNet-DANN et γ-PhysNet-CDANN, avec l'ajout d'un NSB de taux $\delta\lambda = 0.46$ et d'un label shift de $r = 10^{-4}$ dans les données cibles.

Le modèle γ-PhysNet-CBN

Les IRFs sont présentées sur la Figure 7.3. En prenant explicitement en compte la variation de NSB dans le cas de données synthétiques, le γ-PhysNet-CBN offre les meilleurs résultats sur toutes les métriques par rapport à l'UDA. Théoriquement, l'UDA présente de meilleures propriétés que la fusion d'information, car elle considère également les variations inconnues. Cependant, l'entraînement de ces modèles est complexe, et une étude d'optimisation approfondie doit être menée pour permettre à l'adaptation de domaine d'atteindre son plein potentiel. Néanmoins, comme le NSB a été identifié comme l'une des principales sources de divergences, s'attaquer précisément à ce problème à l'aide des modules CBN pourrait apporter des améliorations acceptables pour la communauté des astrophysiciens lorsqu'il est appliqué sur des données réelles.

Le module γ-PhysNet-Prime

Notre modèle γ-PhysNet-Prime est appliqué à la fois sur des simulations perturbées et non perturbées. Dans le premier cas, le bruit correspond au pire scénario, tel que décrit dans [Del+23], avec un paramètre de Poisson $\delta\lambda = 0.46$. Cela permet de nous comparer au γ-PhysNet classique et d'évaluer si l'intégration d'images réelles dans l'ensemble d'entraînement du MAE améliore les résultats dans un scénario dégradé. Les résultats sont illustrés dans la Figure 7.4. En comparant les deux meilleurs cas, il apparaît que le Transformers surpasse remarquablement le modèle de référence en termes de résolution en énergie pour des énergies supérieures à 510^{-2} TeV. Aux plus hautes énergies, le gain est de près de 0,1 en erreur relative, soit une amélioration considérable de 33%. De même, la résolution angulaire est grandement améliorée pour des énergies supérieures à 0,1 TeV. Dans ce cas, l'amélioration atteint environ 25%, avec une valeur de 0,2 pour le modèle de référence à la plus haute

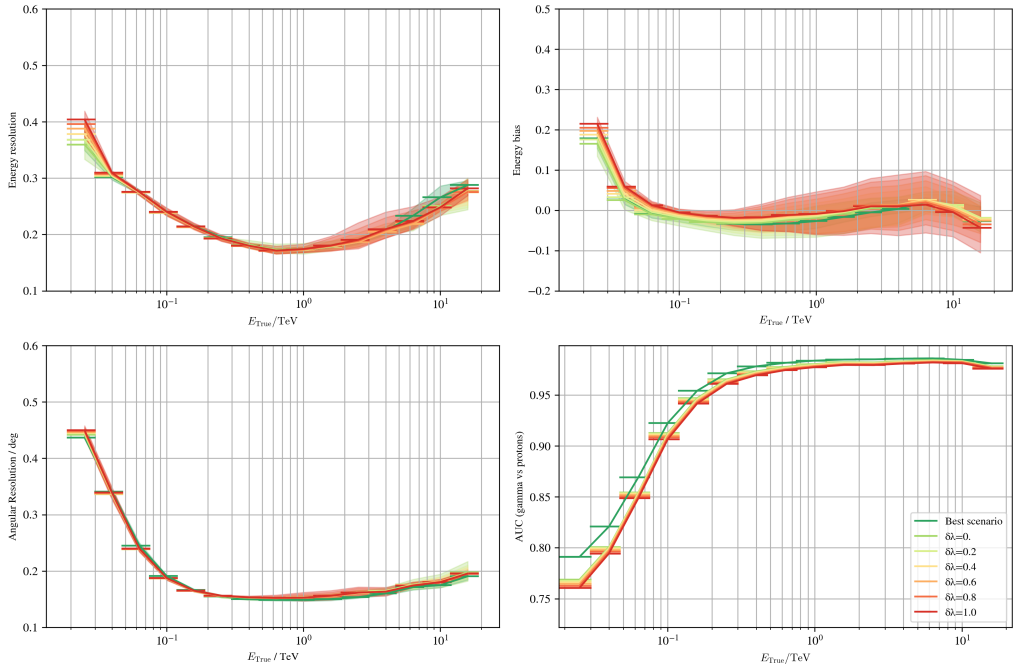


Figure 7.3: Impact du NSB sur le γ-PhysNet-CBN. Les lignes pleines montrent les valeurs moyennes sur les différentes graines d'entraînement, et les enveloppes représentent les valeurs minimales et maximales.

énergie, contre 0,15 pour le Transformer. Enfin, des performances similaires sont observées sur les autres métriques.

Concernant le pire scénario, nous observons finalement une baisse de performance par rapport au meilleur cas. Toutefois, le modèle reste exceptionnellement bon en termes de résolution angulaire et de pouvoir de classification, bien qu'une légère perte soit constatée aux niveaux des plus basses énergies. Cependant, il souffre d'un fort biais, et la résolution en énergie diminue fortement pour des niveaux d'énergie supérieurs à 0,1 TeV. Pourtant, même dans ce scénario, le Transformer reste soit dans les barres d'erreur, soit largement meilleur que le modèle de référence dans des conditions dégradées.

En outre, notre Transformeur γ-PhysNet-Prime que nous avons proposé montre les meilleurs résultats par rapport à chaque méthode étudiée dans le cas spécifique des données simulées dans le contexte du meilleur scénario. Notamment, l'intégration des données réelles dans la tâche de reconstruction du pré-entraînement basé sur MAE joue probablement un rôle clé dans les performances sur le jeu de données cibles. De manière remarquable, en comparant le pire scénario de notre transformeur à celui de γ-PhysNet, le modèle γ-PhysNet-Prime a montré une amélioration significative aux plus basses énergies. Sans affinage supplémentaire, cela pourrait déjà être une approche très prometteuse pour la détection de la Nébuleuse du Crabe, sans besoin d'adaptation de données.

7.5 Application à la détection de la nébuleuse du Crab

Nous étudions désormais l'application de γ-PhysNet-CBN, γ-PhysNet-DANN, sa version conditionnelle et de γ-PhysNet-Prime dans le cas concret de la détection de la Nébuleuse du Crabe, une source connue et très étudiée de rayons gamma. Nos contributions sont

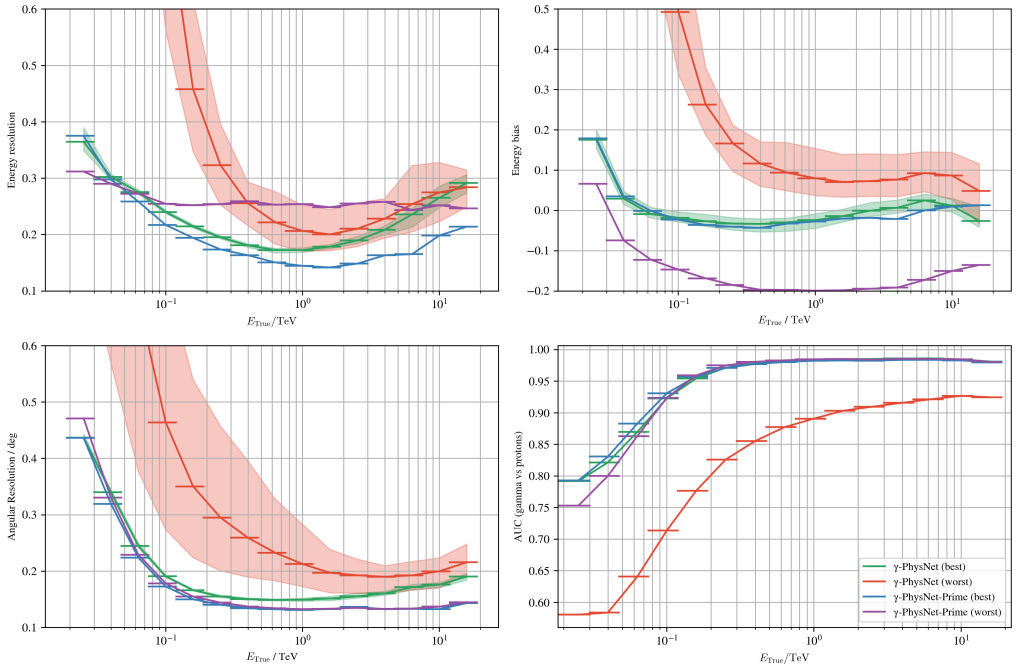


Figure 7.4: Application du γ-PhysNet-Prime sur les simulations.

comparées aux méthodes de référence, constituées de l'analyse standard Hillas+RF et du modèle γ-PhysNet, dans deux scénarios différents : avec et sans impact provenant de la lumière de la Lune. En accord avec les travaux réalisés précédents dans [Vui+21], cette étude comporte également l'analyse avec et sans l'adaptation de données. Les résultats sont données par les Tableaux 7.1 et 7.2.

En comparant les performances dans le cadre de la procédure standard avec le γ-PhysNet, chaque contribution augmente la valeur de signifiance d'une marge notable (+5 à $+9\sigma$ avec clair de Lune et +2 à $+4\sigma$ dans l'autre cas). Cependant, les meilleurs résultats pour chacune de nos contributions sont obtenus lorsqu'elles sont entraînées sur les ensembles de données ajustés. L'application de l'adaptation des données offrant de meilleures performances, cela souligne que nos modèles ne parviennent pas à réduire complètement l'écart entre les simulations et les données réelles sans cette ajustement. Globalement, le modèle γ-PhysNet combiné à l'adaptation des données obtient globalement le meilleur résultat. Notamment, l'analyse standard semble résiliente au niveau de NSB et montre des résultats similaires quel que soit le contexte et l'ensemble de données d'entraînement.

Bien que des améliorations aient été observées sur les simulations en présence de perturbations NSB et de décalage de labels extrèmes, le γ-PhysNet-CDANN semble montrer une légère amélioration sur les données réelles, dans le cas spécifique de l'absence de lumière lunaire. Dans ce contexte, on observe une augmentation d'environ $+2\sigma$. Cependant, aucun avantage pertinent n'est observé concernant le biais de position.

Il est difficile de tirer des conclusions sur notre modèle Transformer, car un seul entraînement a été produit, masquant l'incertitude liée à l'initialisation. Néanmoins, ses performances lorsqu'il est entraîné sur le jeu de données standard et appliqué en condition de clair de Lune sont très encourageantes. Dans ce cas spécifique, il surpasse remarquablement toutes les autres méthodes.

En ce qui concerne la reconstruction de la position de la source, il n'y a pas de différence

significative à signaler entre les différentes méthodes. À l'exception de quelques cas spécifiques, chaque architecture parvient en moyenne à reconstruire un ensemble pertinent de coordonnées de la source, bien que l'angle zénithal de l'observation du Crabe varie avec le temps.

Deux effets opposés se manifestent lors des acquisitions : une diminution du niveau de bruit et une augmentation de l'angle zénithal. Nous pourrions nous attendre à une amélioration des performances due à la baisse du NSB, mais l'augmentation de l'angle tend à contrebalancer ce phénomène. Cela met en lumière la nécessité d'intégrer plus d'informations dans le réseau, telles que la direction de pointage du télescope, ce qui peut être réalisé en utilisant les modules CBN. Cette étude est une perspective à court terme qui sera abordée dans un travail futur.

En conclusion, l'analyse de la Nébuleuse du Crabe mène aux recommandations suivantes :

- En évaluant les performances de nos contributions avec les méthodes de référence, avec et sans adaptation des données, il apparaît que la combinaison de l'adaptation des données avec le modèle γ-PhysNet est l'approche la plus prometteuse dans ce cas spécifique de la Nébuleuse du Crabe. Bien que Hillas+RF présente des résultats légèrement meilleurs lors des observations affectées par la lumière lunaire, l'analyse standard possède de forte limitations aux énergies plus faibles, ce qui constitue un facteur limitant pour la reconstruction du flux.

- Concernant l'applicabilité des réseaux de neurones, le principal inconvénient actuel est la nécessité d'entraîner les modèles pour chaque jeu d'observation, et la nécessité d'appliquer l'adaptation des données illustre cette contrainte. Sans ajustement de l'ensemble d'entrainement pour réduire l'impact du NSB, le gain de performance est limité. À l'inverse, le γ-PhysNet-CBN, grâce à sa robustesse, est capable de prendre en compte des variations particulières dans la procédure d'entraînement. Néanmoins, la clé réside dans la généralisation du modèle, et notre modèle Transformer, grâce à l'entraînement préliminaire MAE, semble le plus capable d'y parvenir.

- L'optimisation des techniques d'UDA est un défi difficile, notamment entre les tâches conflictuelles de classification de domaine et de particules. De même, un contrôle précis du processus d'entraînement est nécessaire lors de l'utilisation des approches Transformer. L'entraînement préliminaire et le fine-tuning nécessitent de nombreux ajustements d'hyperparamètres, ce qui demande de l'expertise et beaucoup de temps de calcul. En revanche, les méthodes basées sur les modules CBN sont plus faciles à faire converger, bien qu'elles doivent être optimisées plus longtemps pour atteindre leurs meilleures performances.

7.6 Conclusion

L'analyse standard actuelle, basée sur l'algorithme de Hillas et combiné aux RF, a été utilisée pour caractériser complètement la nébuleuse du Crabe avec une compilation d'acquisitions du LST dans l'importante étude de [Abe+23]. Cette approche présente l'avantage notable d'être, dans une certaine mesure, explicable. D'un côté, l'extraction de quelques paramètres

	Méthodes	Significance	Bias de position	# Graines
Standard	Hillas+RF	**18.3** (18.3)	0.034	1
	γ-PhysNet	7.54 (10.7)	0.060	5
	γ-PhysNet-DANN	12.9 (17.0)	0.051	5
	γ-PhysNet-CDANN	12.1 (16.1)	0.061	5
	γ-PhysNet-Prime	16.1 (16.1)	0.034	1
Adapté	Hillas+RF	19.5 (19.5)	0.044	1
	γ-PhysNet	17.8 (19.7)	0.036	5
	γ-PhysNet-DANN	19.2 (20.1)	0.053	5
	γ-PhysNet-CDANN	**19.9** (22.4)	0.041	5
	γ-PhysNet-Prime	15.8 (15.8)	0.060	1
	γ-PhysNet-CBN	17.0 (20.3)	0.036	4

Table 7.1: Tableau récapitulatif avec impact du clair de Lune. Les résultats sont obtenus en moyennant les valeurs lorsque plusieurs entraînements sont disponibles. La valeur de signifiance la plus élevée parmi les entraîenements est indiquée entre parenthèses. La meilleure signifiance moyennée est affichée en gras.

par l'algorithme de Hillas repose sur des hypothèses morphologiques conçues par l'humain. D'un autre côté, il est possible de calculer l'importance de chaque composante d'entrée des RF. Néanmoins, cette méthode souffre intrinsèquement de limitations, notamment dans les basses énergies, là où se trouve l'essentiel du flux de rayons gamma d'une source.

Récemment, l'application de l'apprentissage profond pour reconstruire les paramètres physiques des particules incidentes a montré un potentiel significatif dans l'amélioration des performances de détection, tant au niveau des simulations que dans la détection de sources réelles de rayons gamma. Notamment, le γ-PhysNet est le premier modèle de reconstruction des acquisiitons provenant du LST-1. Cependant, comme les RF et les réseaux de neurones sont entraînés sur des simulations, ils sont inévitablement confrontés à des divergence de domaine lorsqu'ils sont appliqués à des données réelles, entraînant inévitablement une baisse des performances. Cela est particulièrement visible lorsqu'on considère l'adaptation des données. La modification des données d'entraînement par ajustement du fond, telle qu'illustrée dans l'article [Vui+21], a montré des résultats bien meilleurs qu'une application directe de chacune des approches.

Il a été démontré que le NSB est l'un des facteurs les plus significatifs affectant les performances de détection, ainsi que la direction de pointée du télescope, comme étudié dans l'important travail de [PMO22]. Dans les deux cas, il est possible de comprendre et de théoriser ces concepts en profondeur. Premièrement, le NSB peut être évalué à l'aide des pédestaux, permettant la caractérisation et la paramétrisation de ses variations temporelles à travers chaque pixel. Deuxièmement, l'impact du pointée peut être atténué en intégrant

	Méthodes	Significance	Bias de position	# Graines
Standard	Hillas+RF	**20.5** (20.5)	0.015	1
	γ-PhysNet	14.3 (16.1)	0.033	5
	γ-PhysNet-DANN	16.1 (16.8)	0.044	5
	γ-PhysNet-CDANN	18.0 (19.3)	0.037	5
	γ-PhysNet-Prime	16.3 (16.3)	0.074	1
Adapté	Hillas+RF	20.5 (20.5)	0.055	1
	γ-PhysNet	**23.1** (23.8)	0.044	5
	γ-PhysNet-DANN	19.8 (22.3)	0.037	5
	γ-PhysNet-CDANN	20.4 (22.5)	0.046	5
	γ-PhysNet-Prime	16.5 (16.5)	0.032	1
	γ-PhysNet-CBN	16.7 (18.7)	0.059	4

Table 7.2: Tableau récapitulatif sans impact du clair de Lune. Les résultats sont obtenus en moyennant les valeurs lorsque plusieurs entraînements sont disponibles. La valeur de signifiance la plus élevée parmi les entraîenements est indiquée entre parenthèses. La meilleure signifiance moyennée est affichée en gras.

une grande variété de directions dans l'ensemble de données d'entraînement. Malgré ces efforts, des différences inconnues ou non paramétrées peuvent encore exister et contribuer de manière significative à la perte de performance.

Tout au long de cette thèse, nous avons suivi des lignes directrices structurées pour aborder ces problèmes. En nous appuyant sur le travail de [Jac20], nous avons étendu la recherche à différents contextes :

- Différences connues : Ces différences peuvent être traitées en intégrant des variables externes, telles que les images de pédestaux ou une information synthétique. Nous avons présenté le γ-PhysNet-CBN, incorporant des modules CBN en remplacement des traditionnels BN pour atténuer l'impact des variations du NSB.

- Différences inconnues : Pour traiter ces différences, nous avons exploré et évalué des techniques d'adaptation de domaine non supervisées. Nous avons proposé et comparé plusieurs approches, sélectionnant le γ-PhysNet-DANN comme la plus prometteuse. Nous avons également fourni une étude complète sur sa version conditionnée et son intégration à l'équilibrage des tâches.

- Modèle global : Comme solution globale plausible, nous avons introduit le γ-PhysNet-Prime, une version Transformer du γ-PhysNet, conçue pour fournir un modèle plus robuste et adaptatif. Il bénéficie d'une capacité d'apprentissage accrue, optimisée avec

une première étape d'apprentissage non supervisé via MAE suivie d'une spécialisation supervisée.

Fusion d'information

Le conditionnement des entrées, grâce à l'implémentation des modules CBN, a montré des résultats prometteurs lors de son application aux simulations. Basé sur l'augmentation des données par introduction de bruit poissonien, il exploite le paramètre de bruit comme variable externe conditionnant le réseau. Le calcul des IRFs suggère que le modèle a parfaitement appris la notion de bruit, avec des performances presque équivalentes au meilleur scénario du γ-PhysNet.

Concernant la détection de la nébuleuse du Crabe, deux effets peuvent être mis en évidence. Pour rappel, les deux premiers ensembles de données sont dégradés par la lumière de la Lune, tandis que les deux derniers ne souffrent pas de ces perturbations, mais contiennent un angle zénithal plus élevé par rapport à l'ensemble d'entraînement. Il semble que l'inclusion des modules CBN ait permis au réseau de mieux résister aux variations du NSB. Cependant, le modèle n'a pas retrouvé les performances des modèles bénéfiant de l'adaptation de donnée, contrairement aux attentes.

Adaptation de domain non-supervisée

Le γ-PhysNet-CDANN a été sélectionné parmi plusieurs approches comme la plus prometteuse pour la détection de la nébuleuse du Crabe. Afin d'évaluer la contribution du conditionnement, la version DANN a également été appliquée. Les résultats montrent que l'intégration de l'adaptation de domaine, avec ou sans conditionnement, améliore clairement les performances en présence de NSB. D'autre part, l'adaptation des données suggère qu'après un ajustement du NSB, il n'y a probablement plus de facteur significatif de divergence de domaine, et que la difficulté à optimiser notre modèle devient un obstacle majeur à une amélioration des performances. Il est nécessaire de tester le γ-PhysNet-DANN dans des environnements plus complexes afin d'évaluer sa capacité à détecter les sources de rayons gamma.

Concernant l'application du conditionnement, celui-ci a montré, en moyenne, une amélioration lors du dernier ensemble d'observation avec le jeu de données d'entraînement standard. Cependant, une fois entraîné avec l'adaptation des données, son impact n'est plus significatif.

Modèle Transformer

Actuellement, le γ-PhysNet-Prime a donné les meilleurs résultats sur les simulations. Cependant, les performances sont plus nuancées dans le cas des données réelles, et deux principales conclusions peuvent être tirées. Premièrement, la différence de performance entre l'entraînement sur des jeux de données standards ou ajustés est négligeable, probablement en raison de la capacité du MAE à produire une représentation latente robuste et générale. Deuxièmement, le γ-PhysNet entraîné sur des jeux de données ajustés surpasse actuellement notre Transformer dans la détection de la nébuleuse du Crabe. Cela pourrait s'expliquer par la difficulté à optimiser de tels modèles. En effet, il a été observé que

l'entraînement du γ-PhysNet-Prime sur 500 époques conduit à un surapprentissage, ce qui signifie que le modèle affiné manque probablement de régularisation. Quant au γ-PhysNet-Megatron, notre Transformer d'adaptation de domaine, nous n'avons malheureusement pas réussi à le faire converger, et des recherches supplémentaires sont nécessaires pour en comprendre les causes.

7.7 Perspectives

Notre travail a mis en évidence deux types de perspectives. Les perspectives à court terme correspondent aux améliorations immédiates de nos modèles et aux tests de nos contributions ainsi qu'à l'analyse de haut niveau. En revanche, les perspectives à long terme concernent des approches novatrices et certaines pistes de recherche pertinentes qui nécessitent davantage d'efforts et d'investigations.

Perspectives court terme

Améliorations de l'analyse haut-niveau

L'analyse de haut niveau utilisée dans cette thèse, réalisée avec GammaScan, ouvre plusieurs perspectives d'améliorations futures. Nos méthodes ont été testées sur quatre runs de la Nébuleuse du Crabe, représentant environ 80 minutes de données. Cela est bien moins que d'autres études plus complète telle que celle présentée dans [Abe+23], qui utilise 50 heures d'observation. Ce jeu de données limité entraîne des fluctuations statistiques, notamment lors des phases d'optimisation et d'analyse. Une approche plus pertinente dans le contexte de données limitées consisterait à utiliser les sous-runs pairs pour l'optimisation et les sous-runs impairs pour l'analyse, réduisant ainsi l'impact des facteurs externes.

Actuellement, l'optimisation de la signifiance repose sur une coupure gamma/hadron fixe et une coupure angulaire constante, qui ne reflète pas toujours la taille réelle de la source. Cela peut entraîner une perte de signal ou une augmentation du bruit de fond. Les futures études devraient explorer l'optimisation conjointe de ces coupures en tant que fonctions dépendantes de l'énergie, ce qui bénéficierait particulièrement aux méthodes d'apprentissage profond, étant donné leurs meilleures performances à faibles énergies.

De plus, les résultats actuels sont basés sur l'analyse de la Nébuleuse du Crabe, une source brillante et bien connue. Il serait intéressant de tester les méthodes sur des sources plus faibles, comme Markarian 421 et BLLac, afin de mieux évaluer leurs performances dans des scénarios plus complexes, ces sources ayant des flux variables dans le temps.

Perspectives long terme

Reconstruction stéréoscopique

Bien que l'observatoire MAGIC et le LST-1 aient déjà été associés pour des observations conjointes dans les traveux de [Abe23], la fin de l'année 2024 marquera l'inauguration du deuxième LST à La Palma, et donnera le coup d'envoi des premières reconstructions stéréoscopiques avec le CTAO. Une fois achevé, le projet comptera plus de 60 instruments travaillant en symbiose pour la détection des source de rayons gamma. Cependant, la combinaison

des informations provenant de nombreux instruments dotés de capteurs variés représente un défi majeur.

Des travaux précédents ont tenté la reconstruction d'événements en stéréoscopie sur des simulations. Dans [Shi+19] et [Mie+21], chaque acquisition de télescope alimente une cellule récurrente, et le problème de l'ordre temporel est résolu en considérant l'heure de déclenchement de l'événement détecté. À l'inverse, la stratégie consistant à concaténer chaque image ensemble dans une entrée multicanaux présente plusieurs inconvénients. Premièrement, différents capteurs avec différentes résolutions partagent la même structure de base, ce qui n'est pas optimal puisqu'ils ne partagent pas nécessairement les mêmes informations. Deuxièmement, les données d'entrée peuvent être très clairsemées, en fonction du nombre de télescopes en action. Comme proposé dans la thèse [Jac20], des convolutions indexées peuvent être utilisées en conjonction avec la triangulation de Delaunay pour fusionner les caractéristiques de chaque télescope avant les branches multi-tâches.

Les réseaux neuronaux de graphes (Graph Neural Network, GNN) sont actuellement explorés en interne au LAPP au sein de l'équipe GammaLearn comme une méthodologie prometteuse de fusion d'informations. Introduits en 2009 dans les travaux pionniers de [Sca+09], ils sont devenus populaires en 2017 lorsqu'une variante, le réseau de convolution de graphes, a été proposée dans [KW17]. Ils forment une classe de réseaux neuronaux conçus pour travailler avec des données structurées sous forme de graphes. Contrairement aux modèles basés sur les réseaux de neurones convolutionels qui opèrent sur des données strcturés en grille régulière, les GNN peuvent gérer des structures arbitraires, ce qui les rend bien adaptés aux applications où les données sont naturellement représentées sous forme de graphes, comme un ensemble de télescopes. Ils représentent donc une méthode prometteuse pour relever le défi de la stéréoscopie. Ils ont été appliqués pour la première fois à la suppression du bruit de fond dans les IACT dans [Glo+23], démontrant des résultats prometteurs par rapport à la méthode de référence basée sur des arbres de décision.

Explicabilité des modèles

Les réseaux neuronaux sont souvent qualifiés, à juste titre, de "boîtes noires". En effet, leur fonctionnement interne repose sur des transformations complexes des données à mesure qu'elles traversent les couches du réseau, ce qui rend difficile la compréhension des mécanismes par lesquels ils produisent leurs prédictions. L'amélioration de l'explicabilité des réseaux neuronaux est une tendance de recherche actuelle, souvent appelée XAI (Intelligence Artificielle Explicable) [van+22]. Certaines approches incluent des techniques de visualisation qui aident à identifier les parties des données d'entrée influençant les décisions du modèle. L'algorithme Grad-CAM, décrit par[Sel+19], a été mis en œuvre par [Jac20] comme une première étape pour rendre le γ-PhysNet plus interprétable. Comme attendu, le réseau utilise principalement le signal de la gerbe comme principal facteur pour prédire chaqu'une des tâches. Toutefois, il est intéressant de noter que le bruit de fond a également un léger impact sur la décision. Cependant, ces approches ne permettent pas de comprendre les cas d'échec.

Une autre stratégie repose sur l'utilisation de réseaux bayésiens [Jos+22], principalement mis en avant dans le calcul d'incertitude. Traditionnellement, le calcul de la distribution postérieure dans ces modèles est infaisable en raison du nombre trop important de paramètres. Pour y remédier, des méthodes d'approximation peuvent être employées. En

particulier, les méthodes d'échantillonnage utilisent l'intégration de Monte Carlo pour estimer les valeurs attendues d'un ensemble fini d'échantillons aléatoires. Inversement, les auteurs de [GG16] implémentent un dropout à chaque étape du réseau. Lors de l'inférence, le dropout est appliqué, et chaque passage génère un résultat distinct à partir d'un ensemble différent de paramètres. Ces approches fournissent une estimation de l'incertitude, mais il reste difficile de déterminer dans quelle mesure ces estmateurs peuvent être fiables. Plus il y a de passages d'inférence, plus on peut être confiant dans les résultats. Cependant, cela augmente également le temps d'inférence, ce qui peut poser problème lorsqu'on travaille avec de grands volumes de données. Néanmoins, l'XAI renforcerait la crédibilité des résultats des réseaux profonds et offrirait des perspectives sur la compréhension des modèles ainsi que sur la physique explorée avec le CTAO.

Analyse temps réel

Depuis 2024, GammaLearn a obtenu un nouveau financement de l'ANR via le projet DIRECTA (ANR-23-CE31-0021), qui vise à intégrer des modèles d'apprentissage profond dans l'analyse en temps réel (Real Time Analysis, RTA). Leur implémentation est cruciale pour observer efficacement les phénomènes transitoires. D'une part, le RTA peut être utilisé pour émettre des alertes à d'autres observatoires dans le monde en cas de fluctuations importantes du flux d'une source ou pour la découverte fortuite d'une nouvelle source dans le champ de vision de l'instrument. D'autre part, lorsqu'il s'agit de suivre des alertes provenant d'autres observatoires, les données doivent être analysées très rapidement afin de déterminer s'il est pertinent de poursuivre ou d'arrêter les acquisitions.

S'appuyant sur l'algorithme Hillas+RF, le pipeline actuel bénéficierait grandement des méthodes d'apprentissage profond, en particulier pour la partie de distribution d'événements à plus faible énergie. De plus, les modèles basés sur les modules CBN ou Transformers sont adaptés à diverses conditions d'observation, telles que les variations d'angle zénitaux, azimutaux, de NSB et de présence d'étoiles dans le champ de vision. Traiter ces conditions en temps réel est essentiel, car les résultats constituent un retour d'information important pour les opérateurs des télescopes.

8 Annexes

8.1 Introduction

The following annexes provide supplementary material that supports the main content of this thesis. They include detailed datasets, additional charts and graphs, the choice of the hyperparameters for the design of the γ-PhysNet, and other relevant documentation that is referenced throughout the research. Each annexe is intended to offer further clarity and depth to the findings and discussions presented, ensuring comprehensive understanding and validation of our procedures. The organization of these annexes follows the sequence mentioned hereafter.

Returning to the historical aspect of machine learning, Section 8.2 presents a graphical representation of the main milestones in the field of deep learning complementing the introduction in Chapter 2. Because these techniques aims to learn from data, it is important to analyse the composition of the training and test datasets. To this end, a presentation of the LST and digits datasets are provided respectively in Section 8.3 and 8.4. The LST dataset is wildly used in Chapter 4, and the digit datasets are used complementary as a validation and further understanding of the implementation of our models. Furthermore, the γ-PhysNet-CBN and the γ-PhysNet extended with domain adaptation are evaluated on the digits datasets sequentially in Section 8.5 and 8.6. Returning on the γ-PhysNet for IACT, the exploration of the benefits of initialization, normalization and activation are given in Section 8.7, 8.8 and 8.9. Moreover, we compare the difference in performance between indexed convolutions and image interpolation in Section 8.10. Because our neural networks are multi-task by design, a comparative study of multi-task balancing algorithms applied to the reconstruction of the particle parameters is given in 8.11. Finally, as Section 8.3 highlights biases in the LST dataset, especially at the energy range level, we explore the contribution of weighting the energy distribution in Section 8.12.

8.2 Annex: AI milestones

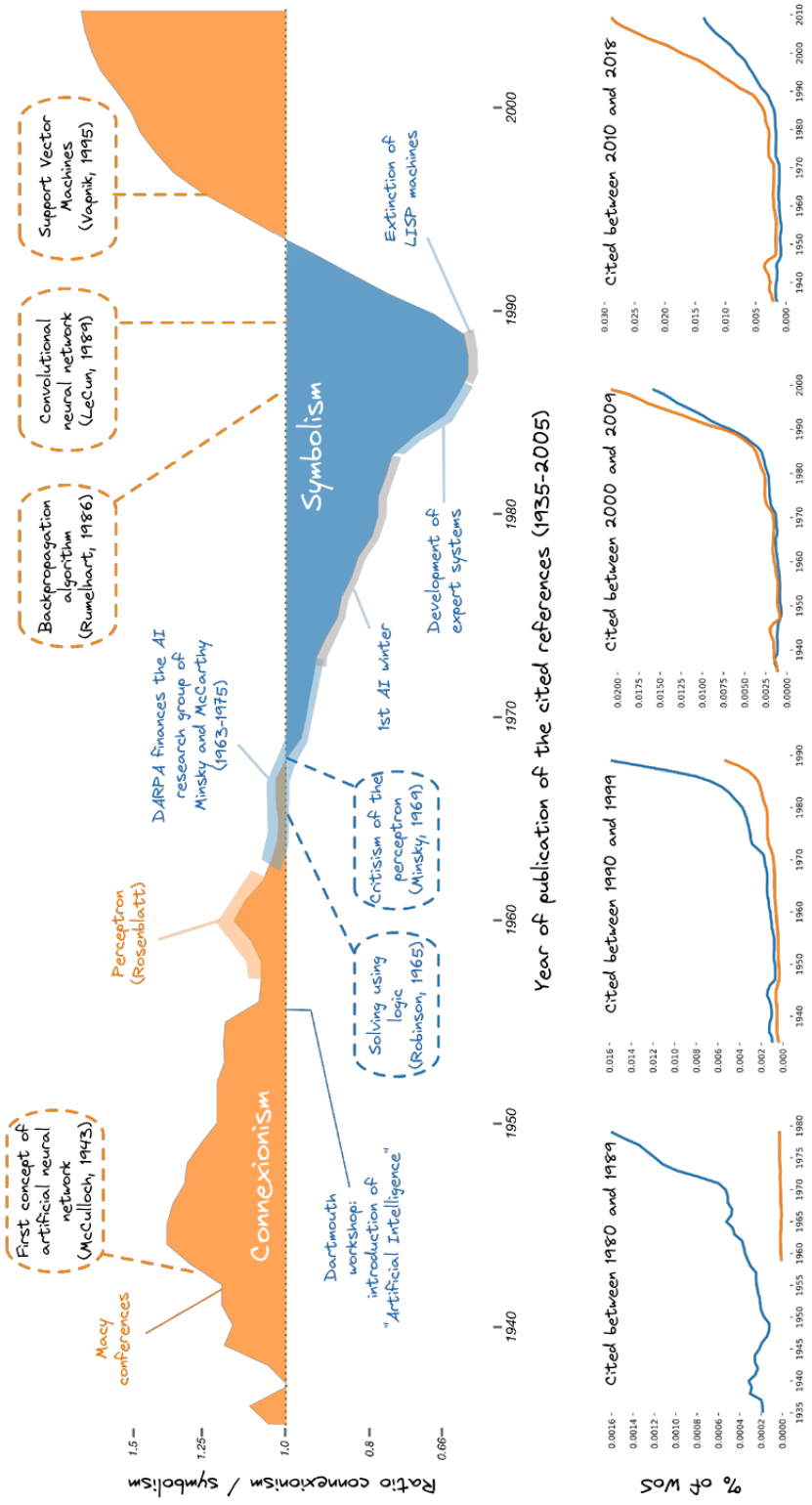


Figure 8.1: Graphical representation of the main milestones of AI. Source [CCM18], modified.

8.3 Annex: Presentation of the LST dataset

Introduction

This annex aims to provide an extensive description of the standard dataset used in Chapter 4, and to complete the data workflow presented in Section 4.2. As a reminder, data is split into a training and a test subsets, which contain images of particles, gammas and protons. Regarding the former, a distinction must be specified. In the training set, gammas are diffused, that is to say they come from different locations in space. On the other hand, in the test set, gammas are point source so that they represent a closer situation to reality. In the following, distributions are computed on pixel charge images uniquely. In that case, the direction (altitude and azimuth), the impact point, the energy and pixel intensity are depicted under a histogram representation.

Description of the training and test datasets

Figure 8.2 represents the distributions of the training diffused gammas. As explicated in the introduction, they simulate gammas emitted from a patch in the sky. Complementary, Figure 8.3 represents proton particles. In both cases, the azimuth is the most contributing directional factor impacting the incident particle energy; thus they are simulated on the broader range of position compared to the altitude, which has a limited influence. The test datasets are illustrated in Figure 8.4 and 8.5, respectively corresponding to the test gammas and test protons. Remarkably, the former pointing coordinates falls exclusively into a single bin because test gammas are coming from a single direction to simulate point-like sources. The latter must represent the isotropic flux on protons inherent to any real observations.

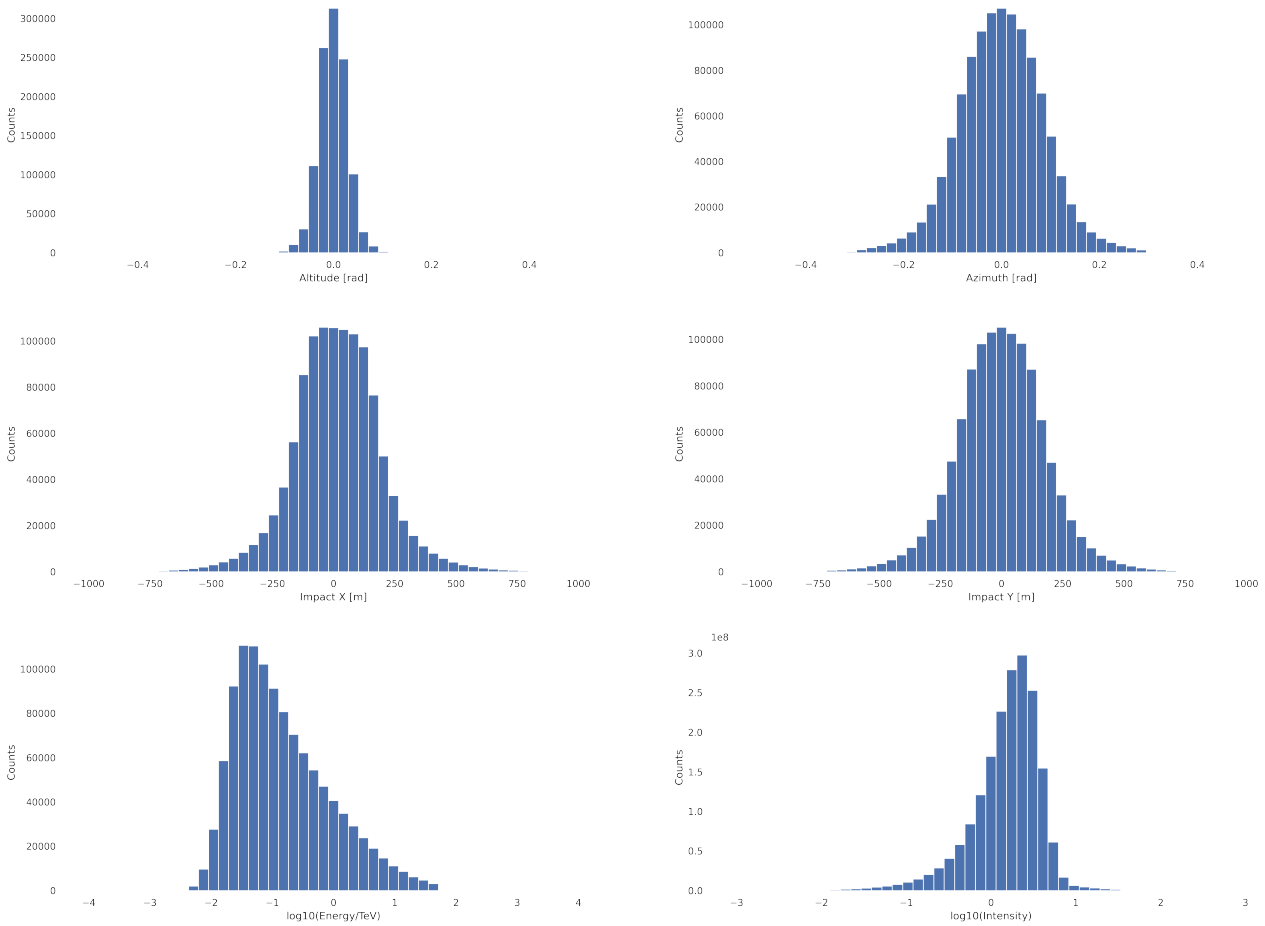


Figure 8.2: Distributions of the diffused gamma parameters in the training dataset.

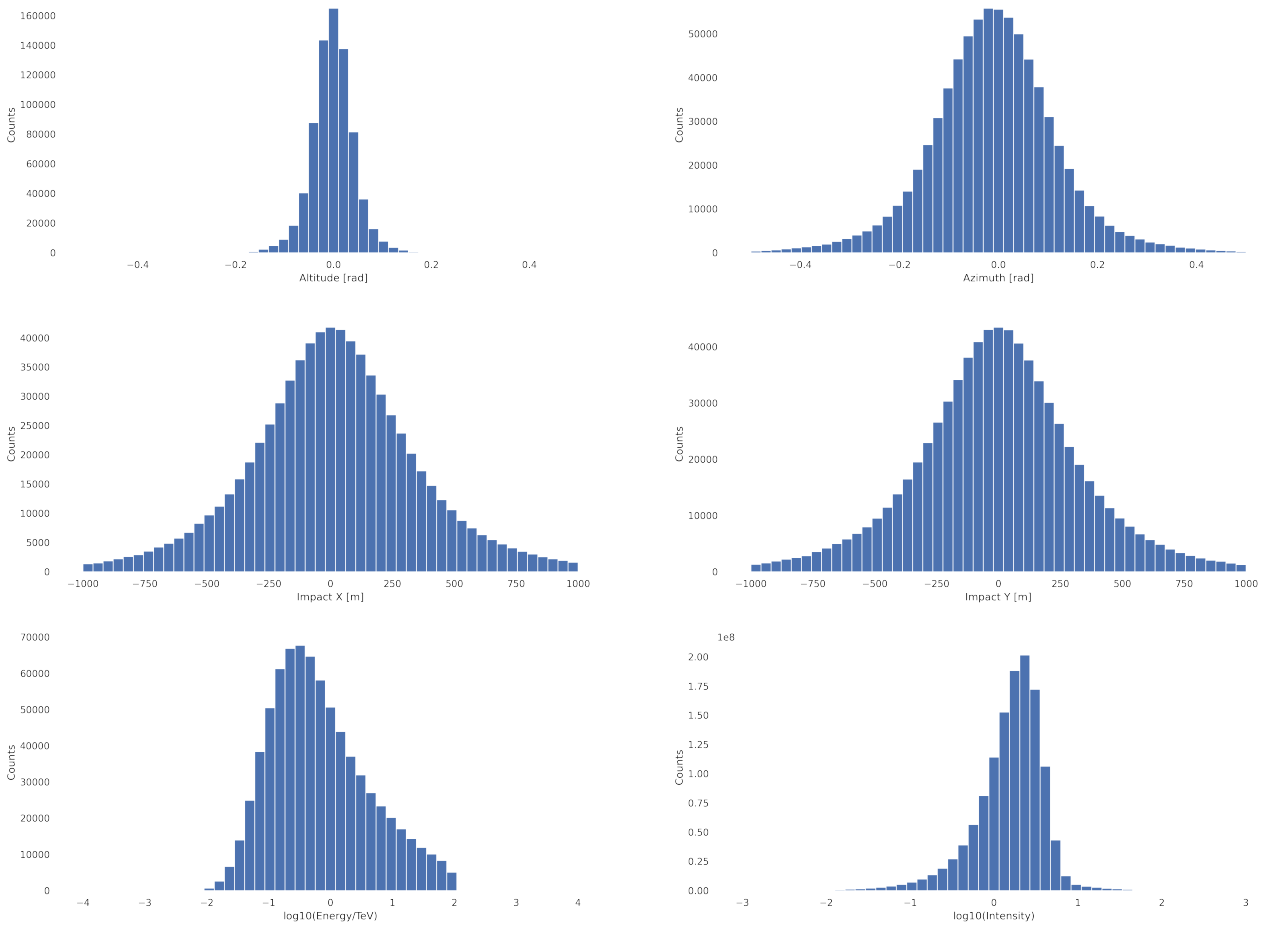


Figure 8.3: Distributions of the proton parameters in the training dataset.

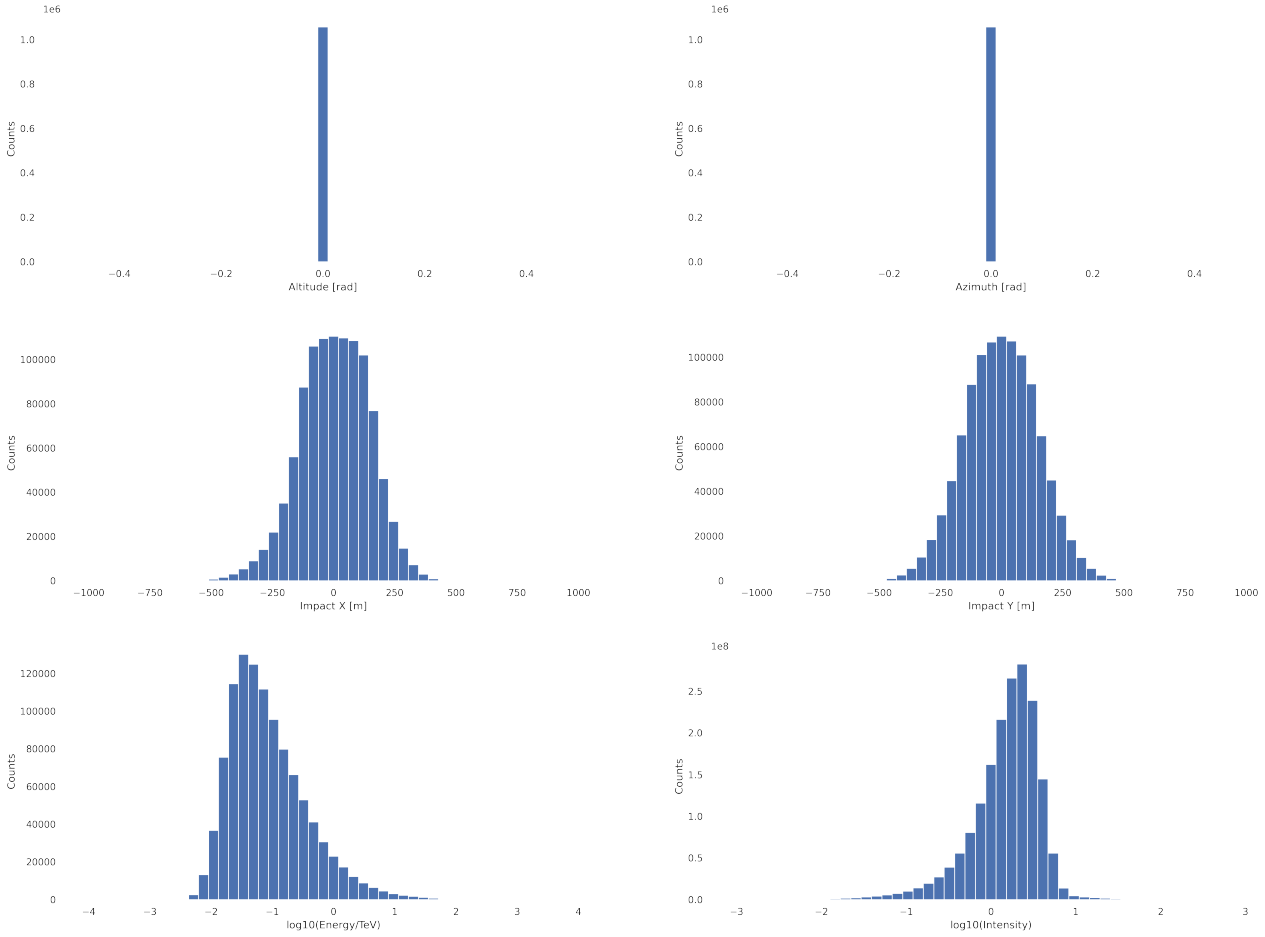


Figure 8.4: Distributions of the gamma point source parameters in the test dataset.

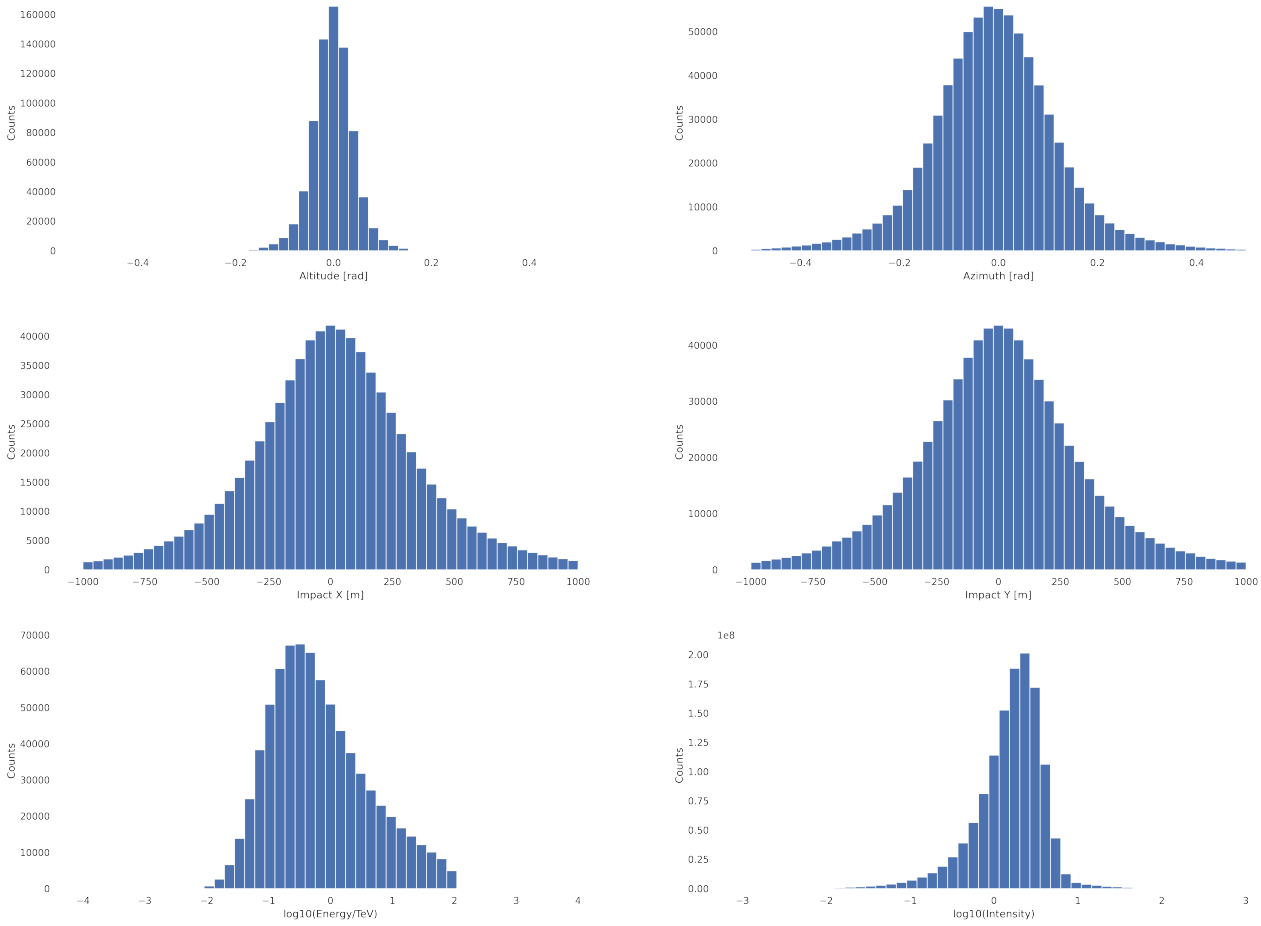


Figure 8.5: Distributions of the proton parameters in the test dataset.

8.4 Annex: Presentation of the digit datasets

The digit datasets offer a convenient way to test our approaches in a simple and computationally affordable setup. They are a traditional domain adaptation benchmark consisting of four different data collections: MNIST [Lec+98], USPS [Hul94], MNISTM [Gan+16], SVHN [Net+11], and some samples are plotted in Figure 8.6. They are an interesting use case as they represent different levels of difficulty: MNIST and USPS are closely related and the digits lay on a black background, whereas MNISTM and SVHN contain much more information with a complex RGB background. In addition, SVHN images can contain multiple digits, but only the one in the centre must be classified, strengthening the complexity of the analysis. A more detailed description of the datasets is given in Table 8.1. Each dataset sample is potentially pre-processed to provide 3 colour channels with the replication of the unique channel, and images are adjusted to 28x28 pixels using either zero-padding or subsampling with bilinear interpolation. Overall, this case study solely illustrates the covariate shift problem, as the digit datasets contain no shift in the labels.

Figure 8.6: The digit datasets.

	MNIST	MNISTM	SVHN	USPS
Resolution	28x28x1	28x28x3	32x32x3	16x16x1
Transforms	Duplicate channel	None	Resize to 28x28	Zero-padding, duplicate channel
#Training	48000	48000	14651	5833
#Validation	12000	12000	58606	1458
#Test	10000	10000	26032	2007

Table 8.1: Meta-data of the digits datasets and their related preprocessing.

8.5 Annex: Evaluation of γ-PhysNet-CBN on the digit datasets

Introduction

We validate and evaluate the classification performance of the γ-PhysNet extended with CBN modules, referred to as the γ-PhysNet-CBN. A formal description of the component is given in Section 3.3. For this purpose, we use the MNISTM dataset described in Appendix 8.4. In this context, digit images are degraded with a noise following a centred Normal distribution. An illustration of the application of the perturbation to the images is given in Figure 8.7. In more detail, the γ-PhysNet-CBN leverage data augmentation, with a noise sampled from a centred Gaussian distribution $n \sim \mathcal{N}(0, \sigma)$. Furthermore, the standard deviation σ is sampled from a Uniform distribution $\sigma \sim \mathcal{U}(0, 1)$. The conditioning variable of the CBNs consists in the single σ value applied on the inputs.

For comparison, we train a simplified γ-PhysNet model to classify the digits, and we evaluate its performance on a range of degraded inputs, with $\sigma \in \{i \times 0.1\}_{i=0}^{10}$. The results are presented in Figure 8.8.

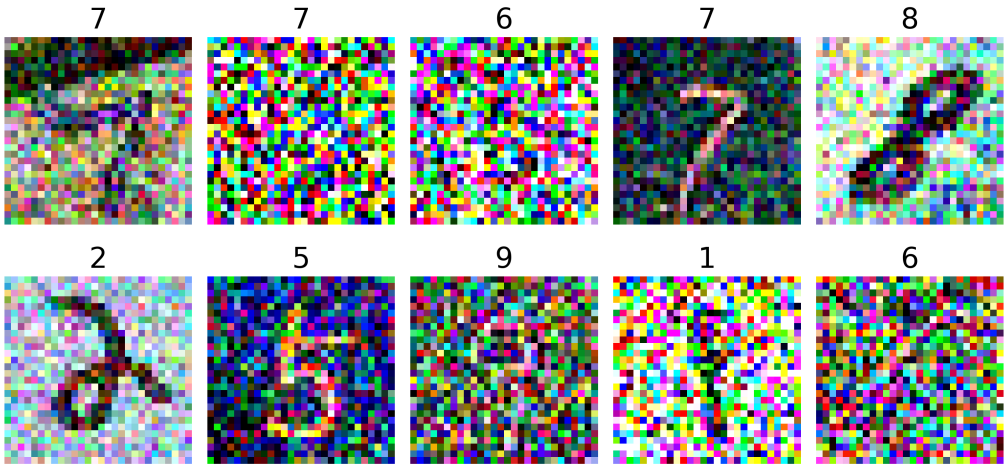


Figure 8.7: Application of a centered Gaussian noise with different standard deviation to the digit images. It is applied independently to each channel, then appears in color.

Training parameters

The γ-PhysNet converges within 20 epochs and the γ-PhysNet-CBN is trained on $20 - 100$ epochs to estimate the training time elongation introduced with data augmentation. In both cases, the batch size is set to 256 and the weights are updated using the Adam optimizer combined with a learning rate of 10^{-3} and a weight decay of 10^{-4}. Overall, the γ-PhysNet contains around $9k$ parameters, while its CBN extension contains $13k$ parameters.

Results

As expected, the accuracy of the γ-PhysNet drops rapidly as the noise increases, and reaches the random classification regimen from $\sigma = 0.5$ onwards. On the other hand, the integration

of the noise information within the network allows to compensate for the loss at the lower level of degradation, while $\sigma < 0.4$. Above this value, the performance decreases linearly with the augmentation of the perturbations. It is interesting to note that in order to retrieve the accuracy of the baseline on the initial images, it is necessary to train the CBN model much longer, from 20 to 100 epochs in our case.

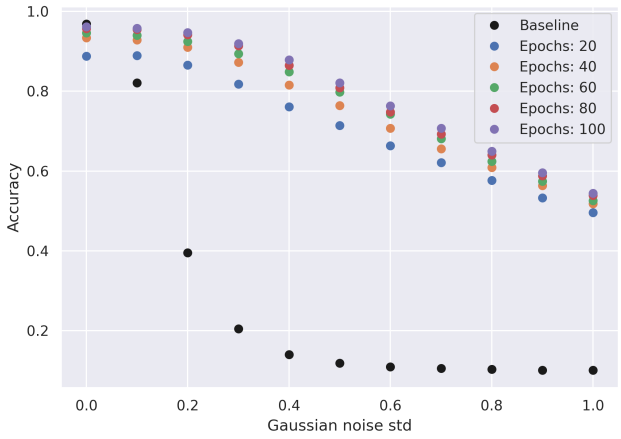


Figure 8.8: Comparison of the Accuracy of our baseline, the γ-PhysNet trained on 20 epochs, and its CBN version. The other labels represent the γ-PhysNet-CBN trained on the specified number of epochs.

In order to confirm the noise-invariance capabilities of the extracted features, we plot the t-distributed Stochastic Neighbor Embedding (tSNE) of the latent features. On Figure 8.9, it appears that the γ-PhysNet latent features are gathered among clusters for $\sigma = 0.1$, but as the standard deviation increases, the points are randomly spread across the latent space. Interestingly, the features extracted from the γ-PhysNet-CBN backbone exhibit a different pattern. They are gathered in clusters regardless of the noise parameter σ, and noise domains are overlapping. This illustrates the noise-invariance brought along with this approach.

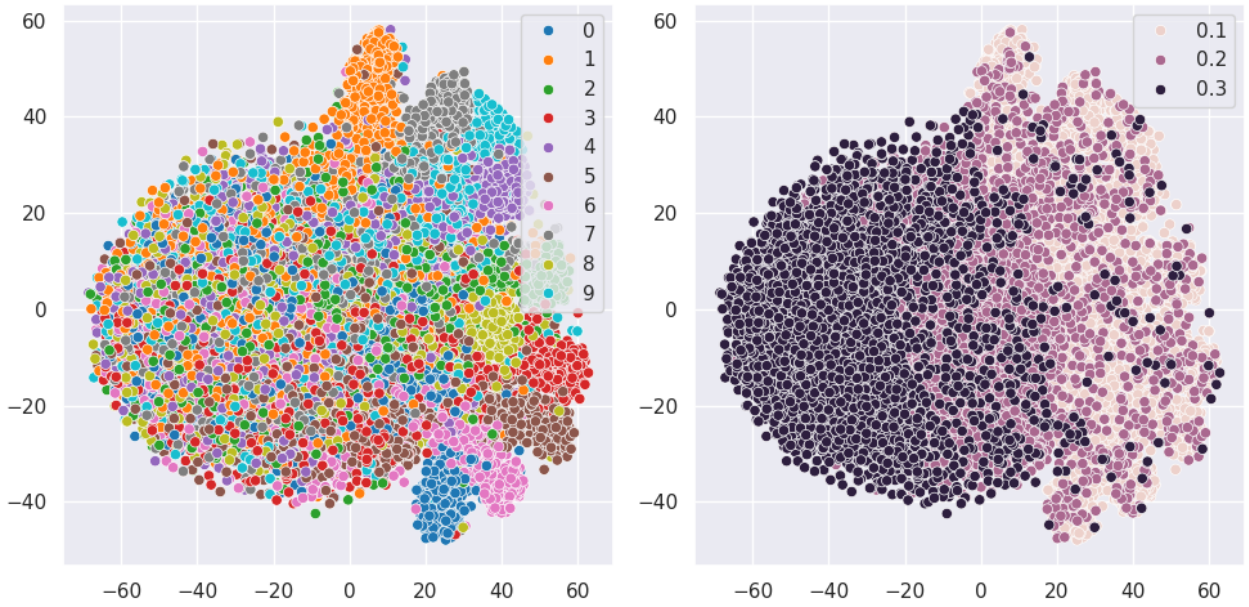


Figure 8.9: tSNE of the γ-PhysNet latent space. The left panel is labelled with the digit class, while the right panel is labelled with the noise standard deviation.

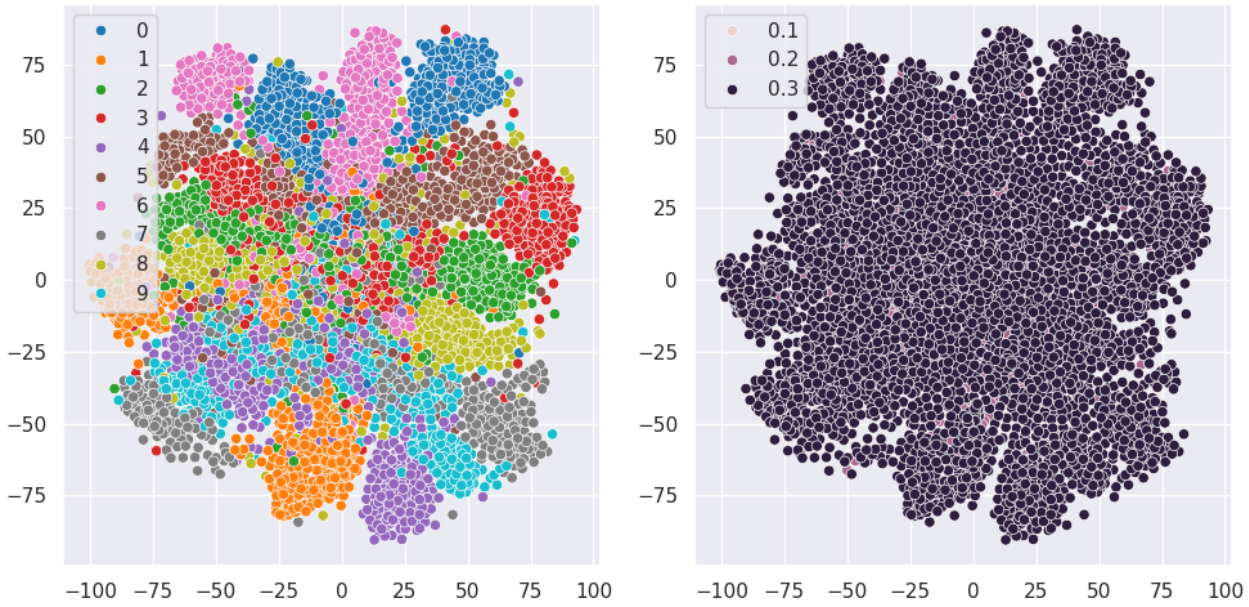


Figure 8.10: tSNE of the γ-PhysNet-CBN latent space. The left panel is labelled with the digit class, while the right panel is labelled with the noise standard deviation.

8.6 Annex: Evaluation of γ-PhysNet combined with domain adaptation and multi-task learning on the digit datasets

Introduction

We validate our domain adaptation approaches selected in Section 2.5 on the digit datasets described in Section 8.4 and compare the different selected multitask strategies combined

with our models. We then demonstrate the relevance of the gradient layer in the case of DeepJDOT.

Hyper-parametrization

We followed the ablation study proposed in [Dam+18] to validate the implementation of our methods. For all the considered experiments, the model is identical and consists in a feature extractor - a cascade of six 3×3 convolutional layers implemented with $(8, 8, 16, 16, 32, 32)$ filters, followed by an average pooling of size 5x5 - and a classifier - a fully connected layer with a SoftMax normalization. For DANN, the domain classifier is composed of two fully connected layers of size 100 and 2 (source vs target). The Adam optimizer with a learning rate of $1e^{-3}$ is used to update the models for 50 epochs. Batch size is set to 256 for both domains. In total, DeepJDOT, and DeepCORAL count 26306 parameters and DANN counts 106808 parameters. Batch Norm (BN) is used as normalization layers and the selected non-linearity is ReLU. The domain adaptation gradients are weighted using the GL with parameters $\gamma = 10$ when applied. The log-variances of UW and the parameters of GN are updated using the stochastic gradient descent (SGD) optimizer with no momentum. All optimizers have a learning rate of $1e^{-3}$ and a weight decay of $1e^{-4}$. A learning rate scheduler is employed and consists in reducing the learning rate by a factor of 10 every 10 epochs.

Finally, each experiment is repeated with ten different seeds to account for the variability in parameter initialization and data shuffling. We thus report, in the results section, the average measures over these ten repetitions for each configuration.

Results

The classification accuracy obtained from the considered approaches are presented in Table 8.2. In order to highlight the contribution of multi-task balancing and domain gradient weighting, a unified training and evaluation procedure is required. This induces a slightly different implementation of the models and the selection of hyperparameters (number of training epochs and the learning rates), compared to the state-of-the-art results described by [Dam+18]. Across all trials, MW is conducted using the weighting coefficients λ_{class} and λ_{domain} calculated with a grid search procedure based on MNIST→MNISTM only.

Firstly, we compare MW with UW and GN for each of the selected methods. On the one hand, DANN and DeepCORAL combined with UW either reach or outperform MW, sometimes by a significant margin (+6% for DANN on USPS→MNIST). On the other hand, DeepJDOT's best performance is obtained either with GN or UW depending on the scenario, but the gain in performance can be remarkable (+10% for DeepJDOT on SVHN→MNIST). This simple example shows that it is possible to use MTB not only to save on computing time but also to improve the performance of the model. Moreover, it is important to note that, even though DeepJDOT and DeepCORAL cannot be mathematically integrated into the UW theory, as they cannot be considered as likelihood following a Normal or Laplacian distribution, they still perform well in this context of digits datasets.

Furthermore, for fair comparison, all the methods rely on the same hyperparametrization, but it is noticeable that the performance of domain adaptation on the SVHN→MNIST scenario is relatively low for both DANN and DeepJDOT, especially compared to the literature results. However, doubling the number of training epochs and removing the learning

		Source → Target			
Method	MTB	MNIST → USPS	USPS → MNIST	SVHN → MNIST	MNIST → MNISTM
Vanilla	-	0.89	0.75	0.54	0.26
DANN	MW	0.90	0.86	**0.55**	0.95
DANN	UW	**0.95**	**0.92**	**0.55**	**0.97**
DANN	GN	0.92	0.89	0.49	0.96
DeepJDOT	MW	**0.95**	**0.94**	0.65	0.92
DeepJDOT	UW	0.92	0.92	**0.75**	**0.97**
DeepJDOT	GN	**0.95**	**0.94**	0.69	0.89
DeepCORAL	MW	0.94	**0.90**	0.61	0.79
DeepCORAL	UW	**0.95**	**0.90**	0.61	**0.82**
DeepCORAL	GN	0.94	0.89	**0.62**	0.72

Table 8.2: Ablation study of DANN, DeepJDOT and DeepCORAL corresponding to the mean accuracy over ten seeds on the target dataset. The source accuracy is not mentioned. Only are reported the best performing strategies, but the impact of the GN α hyperparameter and of the gradient weighting are described in the next sub-section.

rate scheduler increase the performance of DANN by 20 points for MW, UW and GN, finally reaching the state-of-the-art performance. This case illustrates a slower convergence of the considered methods. For DeepJDOT, we did not manage to retrieve the results of [Dam+18] in this particular scenario, but we noticed a high sensitivity to the training hyperparameters on the performances of this approach.

Impact of GradNorm hyper-parameter α

Despite GN reduces the need for costly grid search optimization of the weights, the influence of its hyperparameter α must be explored. We evaluate the performance sensitivity of our selected domain adaptation methods to α using a selection of $\alpha \in \{0.1, 0.5, 1.5, 3.0\}$. Experimentally, as presented in Table 8.3, we obtain that in the case of the digits datasets, our selected methods seem insensitive to the value of α on average. However, a finer analysis highlight differences depending on the scenarios. DANN on MNIST→USPS and USPS→MNIST works best for low values of α ($\alpha = 0.1$), but is not influenced on MNISTM→MNIST. DeepJDOT performance are at best on SVHN→MNIST with $\alpha = 0.5$ on the source, but with $\alpha = 1.5$ on the target. DeepJDOT is less sensitive to α when the gradient weighting is used, while DANN is more sensitive to this augmentation. DeepCORAL appears less sensitive with or without GL.

	0.1	0.5	1.5	3.0
DANN	**0.87**	0.86	0.86	0.86
DeepCORAL	**0.86**	**0.86**	**0.86**	**0.86**
DeepJDOT	0.90	**0.91**	**0.91**	**0.91**

Table 8.3: Mean accuracy across all experiments while varying the GradNorm hyperparameter α. Higher is better. In this table, DeepJDOT is combined with GL.

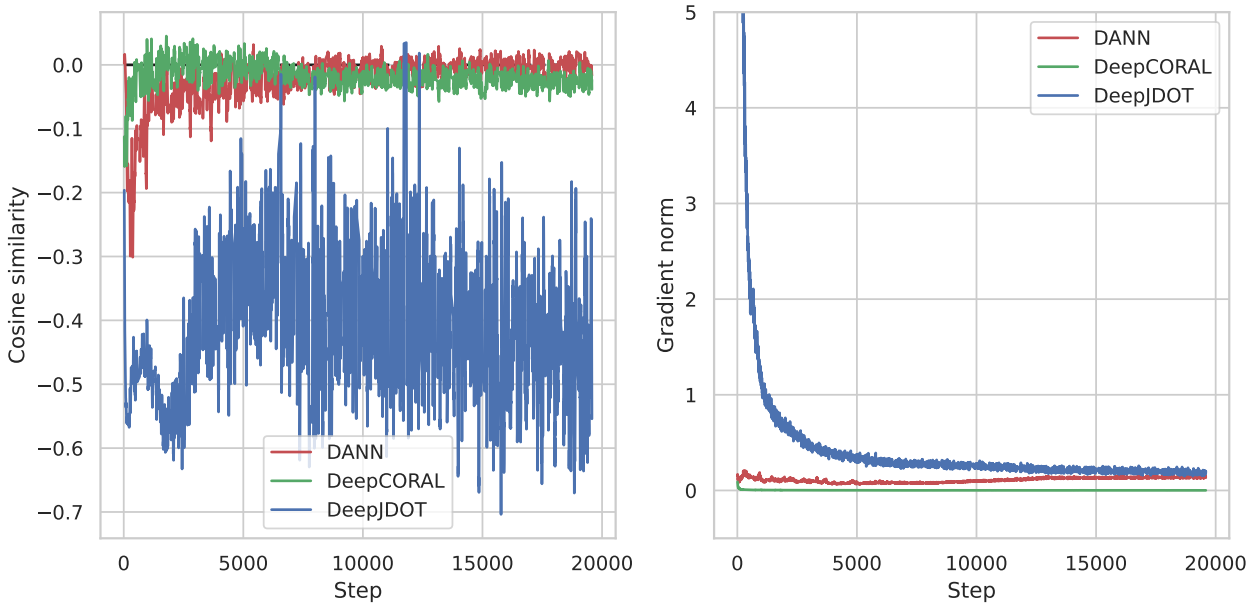


Figure 8.11: Cosine similarity between the domain task and the classification task (left), and gradient norms (right), using DANN, DeepCORAL and DeepJDOT on the SVHN→MNIST scenario. In the case of DeepJDOT, the domain and digit classification tasks are highly conflicting. Moreover, the norm of the DeepJDOT gradients is very high at the beginning of the training, suggesting that the domain task will not optimize properly.

Impact of the Gradient Layer

The gradient layer consists in integrating an additional pseudo-function to the domain adaptation branch to better control the resulting domain adaptation gradient during the backpropagation. As illustrated on Figure 8.11, the Wasserstein distance of DeepJDOT provides strong gradient values at the beginning of the training, which may lead the model to optimize very fast on the domain adaptation task to the detriment of the other tasks.

Weighting the domain adaptation gradients at the beginning of the training allows the model to optimize on the source data first, which appears to be the best strategy for DeepJDOT in this specific context. On the contrary, it reduces the performance of DANN and DeepCORAL, which work better without gradient weighting. A possible reason is that our model uses a learning rate scheduler, which reduces the contributions of the weight update over time, which can be detrimental when combined to GL in the case of DANN and DeepCORAL.

	UW	0.1	0.5	1.5	3.0
DANN	0.05	-1.56	-1.56	-0.83	-2.48
DeepCORAL	-0.47	-0.81	-0.73	-0.88	-0.91
DeepJDOT	21.8	18.9	17.1	15.1	15.0

Table 8.4: Performance gap between the base version and the GL-augmented version. The results correspond to the mean accuracy across all experiments. Values ~ 0 signify that both versions perform similarly on average. Values > 0 highlight a beneficial results of GL, while values < 0 indicate that no layer yields better performance.

8.7 Annex: Impact of parameter initialization on the γ-PhysNet

Introduction

Parameter initialization is a crucial aspect of training deep learning models. It refers to the process of setting the initial values of the weights and biases before training begins. Proper initialization can significantly impact the speed of convergence, the stability of the training process, and the final performance of the model. Conversely, poor initialization can lead to slow convergence, vanishing or exploding gradients, and suboptimal model performance [MM16]. In the early days of neural networks, it mostly consisted in setting values that don't fall into the saturating regime of the popular at the time sigmoid activation function. This trick allowed to avoid small gradients and thus a slow training process [LeC+12].

Modern approaches necessitate to compute the gain α of the selected activation function. It is a scaling factor used to adjust the either standard deviation of the weights or the bounds of the distribution according to the type of activation function used in the network. This adjustment helps maintain the variance of the activations and gradients as they propagate through the layers, preventing issues like vanishing or exploding gradients. More details on gains are available at [Kum17].

Zero-initialization

The parameters are all initialized with the value 0.

Uniform initialization

The parameters are sampled from a uniform distribution, $\theta \sim \mathcal{U}(-a, a)$, where a is a hyper-parameter.

Normal initialization

The parameters are sampled from a normal distribution, $\theta \sim \mathcal{N}(0, \sigma^2)$, where the standard deviation σ is a hyperparameter.

Orthogonal initialization

Widely used with CNNs, [SMG14] based an orthogonal matrices. This approach initializes weight matrices to be orthogonal, preserving the variance across layers.

He initialization

Introduced in [He+15b], the He Initialization, or Kaiming Initialization is a powerful and widely-used method for initializing parameters in deep neural networks with activations from the ReLU-like familly. By maintaining the variance of activations across layers, it helps ensure stable gradient propagation, enabling efficient and effective training of deep learning models. Two algorithms can be derived, depending on which distribution to sample from, namely He Uniform or He Normal. In the following, we refer to the number of inputs and outputs to a specific layer as fan in and fan out. In that case, fan mode can be either fan in or fan out.

- In the former, parameters θ are sampled from a Uniform law $\theta \sim \mathcal{U}(-a, a)$, with the bound a computed as:
$$a = \alpha \sqrt{\frac{3}{\text{fan mode}}} \tag{8.1}$$
- In the latter, parameters θ are sampled from a centred Normal law $\theta \sim \mathcal{N}(0, \sigma^2)$, with the standard deviation σ computed as:
$$\sigma = \frac{\alpha}{\text{fan mode}} \tag{8.2}$$

Xavier initialization

Proposed by [GB10], this method also merges two initialization procedures, namely Xavier Uniform and Xavier Normal. This approach helps maintain the variance of activations and gradients across layers, promoting more stable training. They both necessitates to compute the gain α as described previously.

- In the former, parameters θ are sampled from a Uniform law $\theta \sim \mathcal{U}(-a, a)$, with the bound a computed as:
$$a = \alpha \sqrt{\frac{6}{\text{fan}_{\text{in}} + \text{fan}_{\text{out}}}} \tag{8.3}$$
- In the latter, parameters θ are sampled from a centred Normal law $\theta \sim \mathcal{N}(0, \sigma^2)$, with the standard deviation σ computed as:
$$\sigma = \alpha \sqrt{\frac{2}{\text{fan}_{\text{in}} + \text{fan}_{\text{out}}}} \tag{8.4}$$

Results of initialization on the γ-PhysNet performance

The results are presented in Figure 8.12. Interestingly, initialization can have a huge impact of the end results when not conducted correctly. Uniform and Normal initializations, with default parameters (respectively bounds of 0 and 1 for the former and a mean and standard deviation of 0 and 1 for the latter) shows a deep degradation over all metrics, with a

great variation depending on the seed. On the other hand, Orthogonal, Xavier and Kaiming initializations are less sensitive to the seed. Furthermore, regarding the energy bias and resolution, although their results are on average quite similar, variations are less in favour to the Orthogonal method. Thus, Kaiming and Xavier, is their Uniform version, provide the best performance. In this thesis, we use Kaiming Uniform as the initialization method for the γ-PhysNet.

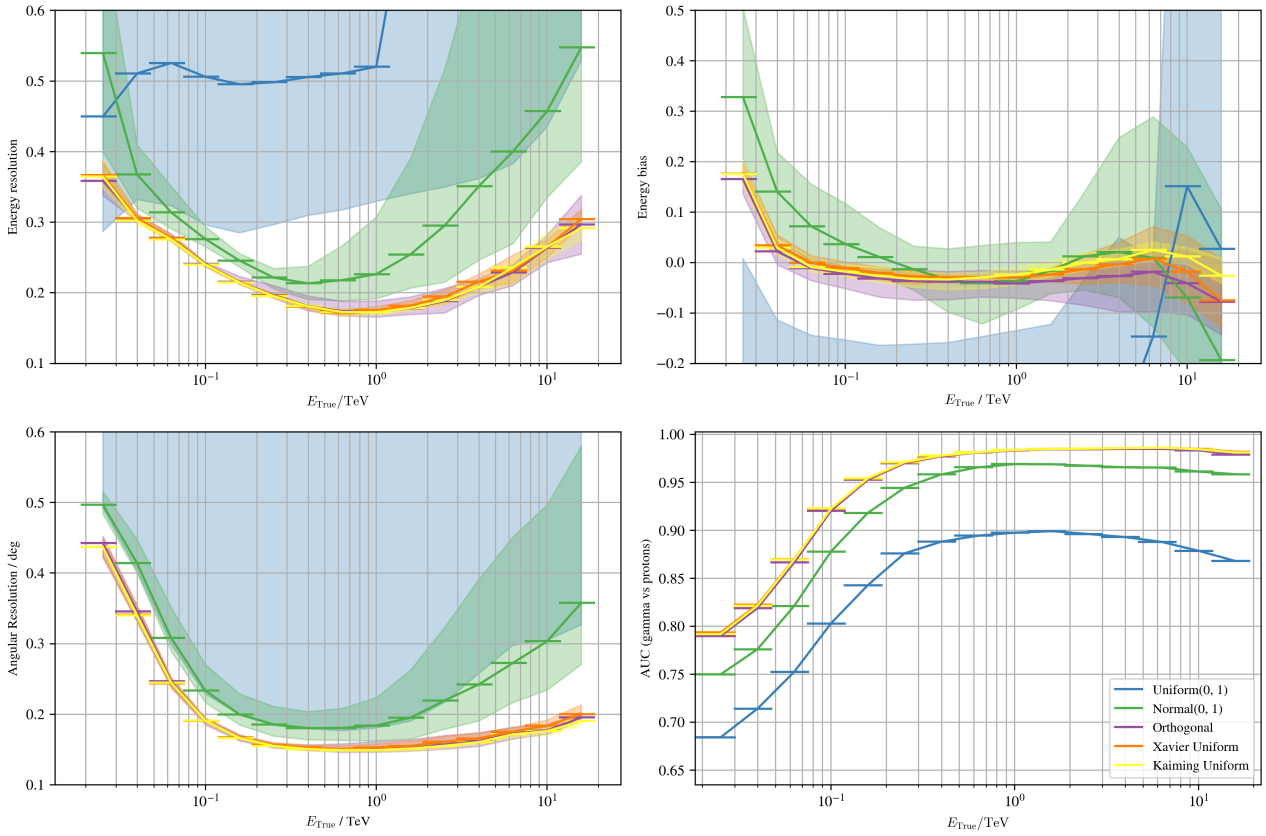


Figure 8.12: Comparison of initialization methods on the γ-PhysNet performance.

8.8 Annex: Impact of the normalization on the γ-PhysNet

Introduction

A decade ago, the training of deep learning models was an extensive process. The 2012 ImageNet competition is a good example. In order to win the challenge, the AlexNet model underwent a 6-day training on 2 once-very-performing GPUs. Consequently, it became crucial for researchers to develop tools to accelerate the learning procedure. Fortunately, established practices such as normalization and feature decorrelation, outlined in Lecun's deep learning recipe book [LeC+12], have long been recognized for their efficacy in accelerating training. In fact, during the optimization phase, back-propagation computes the model's internal gradients based on the inputs to each layer, and their directions guide the adjustment of the parameters. However, as layers are stacked successively, a slight update in the weights upstream can cascade to significant changes in the data distribution downstream, leading to suboptimal training signals. Normalization helps ensure more stable gradients for weight updates, preventing drastic shifts in data distribution, solving vanishing or exploding gradient issues and enhancing the overall training process. Consequently, it often facilitates a more stable and accelerated training regimen for neural networks, reducing the impact of parameter initialization and allowing the use of higher learning rates. An illustration of different normalization algorithms are depicted in Figure 8.13. They can be mathematically described using the same notations. In deep learning for computer vision, data is often organised as a mini-batch of size (N, C, H, W) where N refers to the batch size, C the number of channels, and H, W are the height and width of the feature map or image. An image can thus be indexed by a tuple $i = (i_N, i_C, i_H, i_W)$. Following the notations of [WH18], we can derive a formula for mean and the variance.

$$\begin{aligned} \mu_i &= \frac{1}{N} \sum_{k \in S_i} x_k \\ \sigma_i^2 &= \frac{1}{N} \sum_{k \in S_i} (x_k - \mu_i)^2 \end{aligned} \tag{8.5}$$

S_i represents the set of pixels on which the statistics are evaluated. The features maps are then updated following the standardization procedure. It becomes a part of the model architecture by implementing new trainable parameters (γ, β) so that the network auto-evaluates the need for the standardization.

$$y_i = f_n^{\gamma,\beta}(x_i) = \gamma \frac{x_i - \mu_i}{\sqrt{\sigma_i^2 + \epsilon}} + \beta \tag{8.6}$$

where f_n refers to the normalization algorithm, and ϵ is a constant added for numerical stability. The mini-batch statistics are usually computed at every training iteration.

Batch normalization Batch Normalization (BN) [IS15] standardizes the distribution of the input data to each layer along the (N, H, W) axis, thus $S_i = \{k \mid k_C = i_C\}$, with i_C and k_C denoting respectively the sub-index of i and k along the C axis. That is to say, each feature map is standardized with the other feature maps of the batch sharing the same location. However, BN is mini-batch-dependent in the way that it requires larger batch sizes to effectively

approximate the moments from the mini-batch, which can lead to poor results if memory constraints are an issue and the mini-batch size is small.

Layer normalization Layer normalization (LN) [BKH16] aims to overcome the limitations of batch normalization, that is to say its ineffectiveness in RNNs and its dependence to the batch size for the estimation of the statistics. The key idea resides in normalizing in the (C, H, W) axis, ie the feature direction instead of the mini-batch direction, so that $S_i = \{k \mid k_N = i_N\}$. Although LN yields results for RNNs, it doesn't recover BN's performance for pure CNN architecture.

Instance normalization Instance Normalization (IN) [UVL17] has been introduced in the context of image generation to remove instance-specific information during the training. Formally, it is defines using $S_i = \{k \mid k_N = i_N, k_C = i_C\}$.

Group normalization Group Normalization (GN) [WH18] is very similar to LN but adds an extra hyper-parameter G that defines the number of groups. The feature maps are divided into a few groups that are standardized separately. It can be defined using $S_i = \left\{k \mid k_N = i_N, \frac{k_C}{C/G} = \frac{i_C}{C/G}\right\}$, where C/G corresponds to the number of channels per group. If the hyper-parameter G is set to $G = C$, then GN becomes IN. Although GN performs better than LN, it doesn't achieve BN's performance in classification tasks.

Weight normalization Not really used in practise, weight normalization (WN) [SK16] is a reparametrization of the weights of the model. Each module's weight is normalized independently. Although WN speeds up the convergence of the model, it doesn't approach BN's accuracy in many vision applications.

Results of normalization layers on the γ-PhysNet performance

In this section, we compare BN, GN with hyperparameter $G = 4$, IN and LN with the context of no normalization. Results are presented on Figure 8.14. It appears that in general the addition of a normalization layer has a positive impact on all metrics, especially at the highest energy levels compared to case where no normalization is applied. Yet, BN remains the most performing layer compared to the selected ones, and its influence is particularly visible across all figures of merit, similarly at the higher energy range.

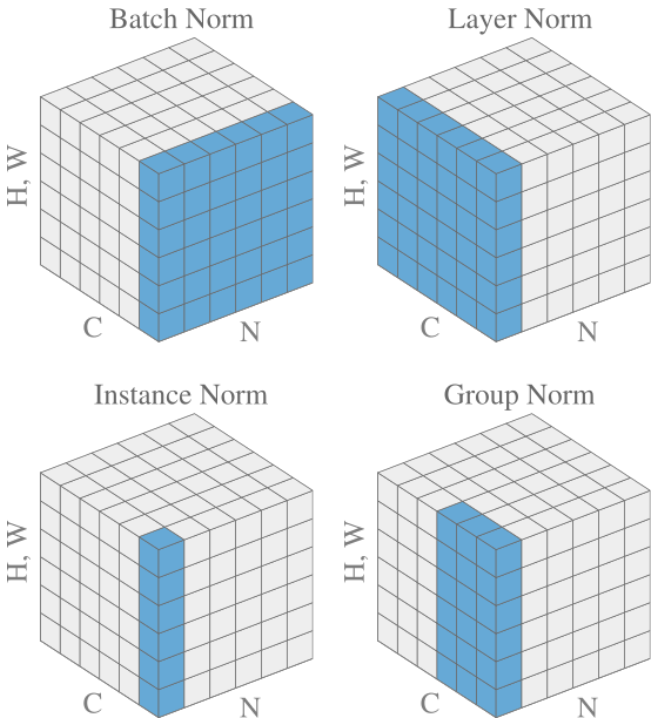


Figure 8.13: Illustration of the different type of Normalization. Source [Qia+20]

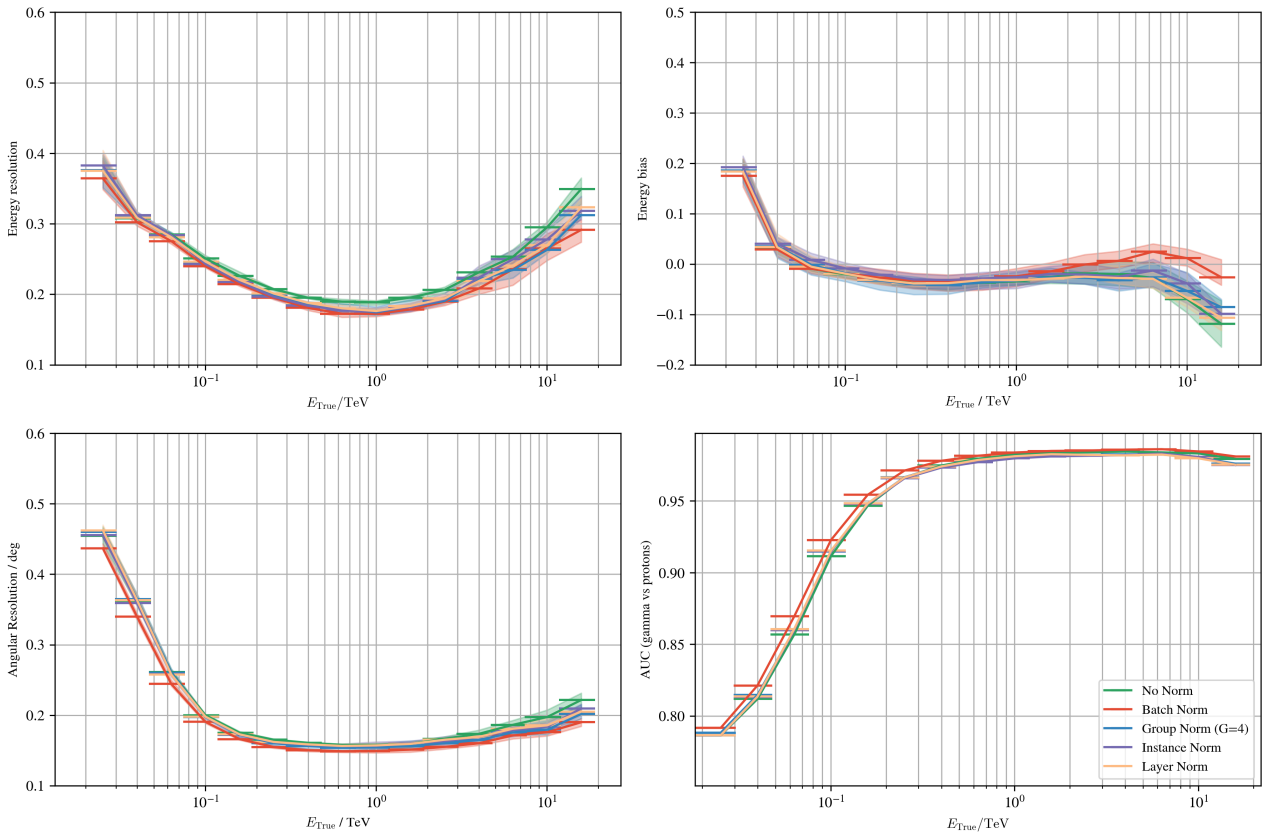


Figure 8.14: Comparison of normalization layers on the γ-PhysNet performance.

8.9 Annex: Impact of the activation functions on the γ-PhysNet

Introduction

Considering exclusively linear combinations of the inputs imposes significant limitations on a neural network's expressive power. The introduction of non-linearities give the model some non-linear capabilities, allowing it to learn complex patterns and relationships. However, they must preferably meet some criteria. Commonly, we seek for functions that avoid vanishing gradients, ensure the ease of computation for both the function and its derivative, and tend to zero-center the outputs. In the early days of neural network training, these specifications were not satisfied and artificial neurons utilized binary threshold units. Nowadays, with a greater understanding of such black box, it is possible to define strategies to implement more sophisticated activation functions and meet the requirements.

Sigmoid

Before the deep learning era, and the advent of deeper models, the sigmoid function was widely-used as a non-linear activation function. Its output is included within the $[0,1]$ segment, which can be interesting for any application that aims to compute values within this range, for example as the last layer of a binary classifier. It has two saturation regimes, and it doesn't center the output around the value 0 (if $X \sim \text{Sigmoid}$, $\mathbb{E}[X] \neq 0$). Thus, it is necessary to combine it with a normalization layer to prevent the inputs to fall one or the other saturation, and to remove the bias shift.

$$\forall x \in \mathbb{R},\ \text{sigmoid}(x) = \frac{1}{1+e^{-x}} \tag{8.7}$$

$$\forall x \in \mathbb{R},\ \text{sigmoid}'(x) = \frac{e^{-x}}{f(x)^2} \tag{8.8}$$

Notice that when $x \to +\infty$ or $x \to -\infty$, $f'(x) \to 0$, the derivative reaches its saturation regimes, leading to gradient vanishing.

Hyperbolic tangent

The hyperbolic tangent function, or tanh, slowly replaced the sigmoid function over time as it started to give better performance for multi-layer networks. Although tanh is also affected by the vanishing gradient problem associated with sigmoids, it centers the outputs around the value 0, which eliminates bias and facilitates faster convergence.

$$\forall x \in \mathbb{R},\ \tanh(x) = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}} \tag{8.9}$$

$$\forall x \in \mathbb{R},\ \tanh'(x) = 1 - \tanh^2(x) \tag{8.10}$$

The ReLU-like familly

ReLU The rectified linear unit (ReLU) [NH10] is a popular activation function for non-linearity. It preserves the positive activations and suppresses the negative ones. Oppositely to the sigmoid and tanh functions, it has a better gradient propagation. Moreover, it solely

consists in adding, multiplying and comparing values, which is very efficient to compute. It is scale-invariant because $\forall (a,x) in \mathbb{R}^2, f(ax) = af(x)$. Although it is not derivable in for $x = 0$, the derivative of the ReLU activation function is easily implementable as it corresponds to the Heaviside step function, with the extension $f'(0) = 0$. Furthermore, similarly to the sigmoid function, it doesn't center the output around the value 0.

$$\forall x \in \mathbb{R},\ \mathrm{ReLU}(x) = x\mathbf{1}_{x>0} \tag{8.11}$$

$$\forall x \in \mathbb{R},\ \mathrm{ReLU}'(x) = \mathbf{1}_{x>0} \tag{8.12}$$

Leaky-ReLU In order to tackle with the zero-centering problem, Leaky-ReLU (LReLU) [Maa13] has been designed as a replacement to ReLU that allows the activated units in the negative regime to have non-zero gradients, thus to always contribute to the training process. However, strong negative activations can have an undesirable impact on the cost function, and additional units may be needed to balance the constructive and destructive interference. LReLU is popular in specific tasks where sparse gradients are an issue, for example during generative adversarial networks training. The negative slope is often referred to as α, and a typical value of α is $\alpha = -10^{-2}$.

$$\forall x \in \mathbb{R},\ \mathrm{LReLU}(x) = \begin{cases} x & \text{if } x \geq 0 \\ \alpha x & \text{otherwise} \end{cases} \tag{8.13}$$

$$\forall x \in \mathbb{R},\ \mathrm{LReLU}'(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ \alpha & \text{otherwise} \end{cases} \tag{8.14}$$

There exist other extensions of LReLU, such as the parameterized LReLU (PLReLU) [He+15b] that has the particularity to implement the slope of the negative argument as a trainable parameter of the model, or also the randomized LReLU (RReLU) [Xu+15] which samples the slope in the negative arguments from a distribution, but these methods are not really used in practise anymore.

ELU Downstream layers are biased due to the non-zero mean activation of the upstream layers. Batch normalization aims to remove this effect, but it can also be tackled using a well-designed activation function. The exponential linear unit (ELU) [CUH16] aims to push to mean of the activations closer to zero using a converging function in the negative support to reach a saturation regime. In this case, for high negative values, the gradients will be very small.

$$\forall x \in \mathbb{R},\ \mathrm{ELU}(x) = \begin{cases} x & \text{if } x \geq 0 \\ \alpha(\exp(x) - 1) & \text{otherwise} \end{cases} \tag{8.15}$$

$$\forall x \in \mathbb{R},\ \mathrm{ELU}'(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ f(x) + \alpha & \text{otherwise} \end{cases} \tag{8.16}$$

GELU Gaussian error linear unit (GELU) [HG23] is an extension of ReLU to a great class of function, replacing $\mathbf{1}_{x>0}$ by Φ, where Φ is a function that approximate the cumulative distribution function of a normal distribution. It can be seen as the expectation of a stochastic regularizer. If the normal distribution is centered and reduced, then Φ can be computed using the Gauss error function erf. The mean and variance of the normal distribution can also be implemented as trainable parameters of the model. Φ can be the sigmoid function as it meets the approximation requirement, and in that case the activation function is referred to as SiLU. GELU is currently gaining popularity over time, notably because it is widely use in the growing NLP in the training of transformers.

$$\forall x \in \mathbb{R}, f(x) = x\Phi(x) \tag{8.17}$$

Results

In this section, we compare the most popular activation functions that are currently used in many deep learning applications that are ReLU, ELU and GELU. Results are presented on Figure 8.15. Regarding the energy resolution and bias metrics, its appears that ELU and GELU slightly outperform ReLU for energies greater than 1 TeV. On the other hand, concerning the angular resolution and the classification power, ELU has slightly less benefits compared to the other ones. On average, GELU proposes the best results compared to the selected non-linearities.

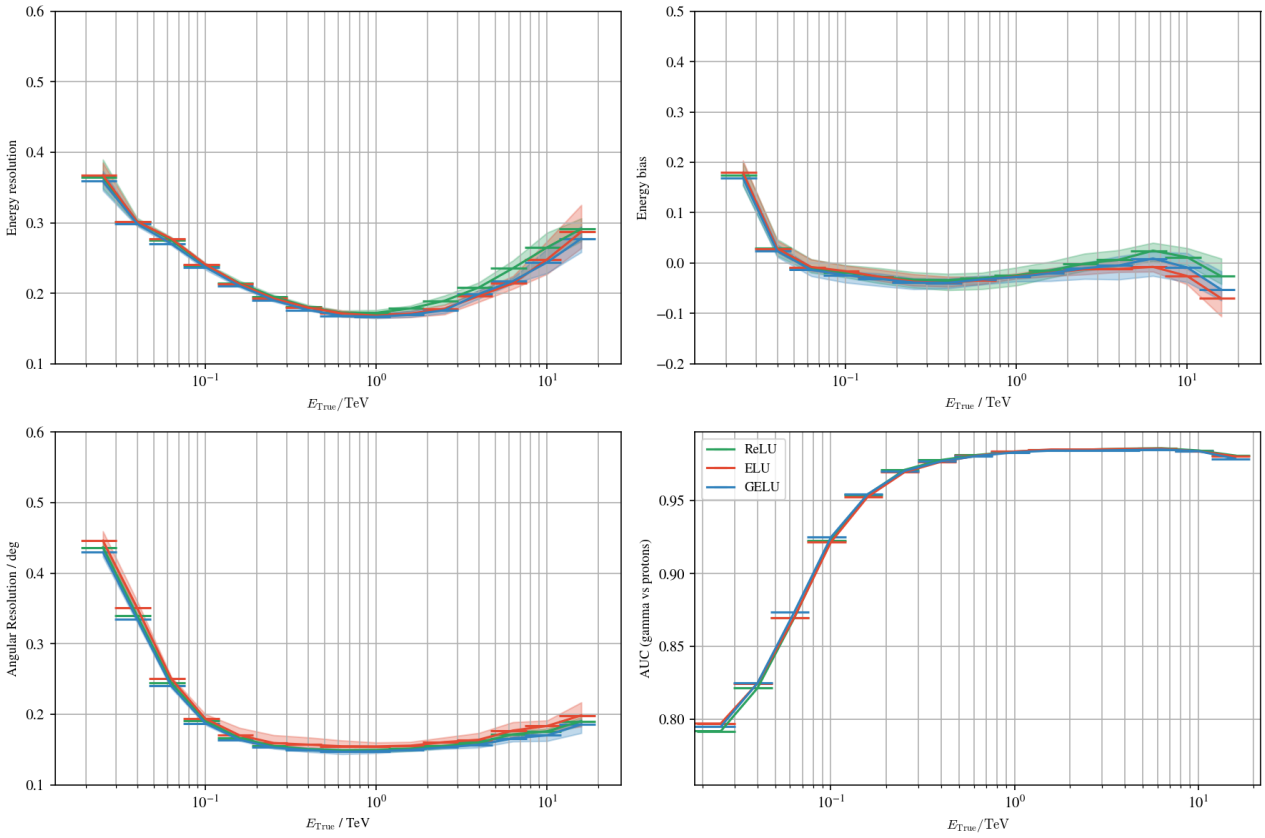


Figure 8.15: Comparison of activation functions on the γ-PhysNet performance.

8.10 Annex: Impact of indexed convolutions on the γ-PhysNet

Introduction

Unlike classical machine learning applications, IACT images often reside on a non-regular grid of pixels. The underlying configuration of the hardware, specifically the arrangement of photomultiplier tubes, determines the geometry of the sensors and, consequently, the grid shape of resulting images.

Indexed convolutions, first introduced in [Jac+19], are designed to account for the particular morphology of IACT data when computing convolutions. However, this approach does not take advantage of the highly optimized convolution operations available in deep learning frameworks like PyTorch, leading to a significant increase in computation time during training. In this case, the neural network is denoted γ-PhysNetIndexed

To mitigate this issue, it is necessary to consider interpolating the data onto a regular grid using an appropriate interpolation strategy. The study conducted in [Jac20] evaluated the performance of various interpolation strategies and identified performance variations depending on the chosen method. The study concluded that bi-linear interpolation offers the best results in terms of performance.

Results

The results are presented in Figure 8.16, and both experiments use the same hyperparameters. Five seeds are performed to account for the variability of the model initialization and data shuffling. The use of interpolation reduces the model uncertainty across all metrics, expected for the AUC. Moreover, it performs better on the energy bias. However, the γ-PhysNetIndexed performs slightly better in the classification and both energy and angular resolutions, at the lower energy levels.

Because interpolated images make use of the PyTorch convolution, the training time is much more in favour to the γ-PhysNet, and the gain of speed is around a factor of 4. In our experiments, the γ-PhysNet is trained within 1h30min using 3 A100 GPUs, while the Indexed version is trained within 6h using in the same scenario. Furthermore, classical grids allow to integrate more easily and quickly new kind of architectures. As a results, this thesis exclusively uses interpolated images for our CNN-based neural network, although the hexagonal grid is considered in the case of our Transformer models.

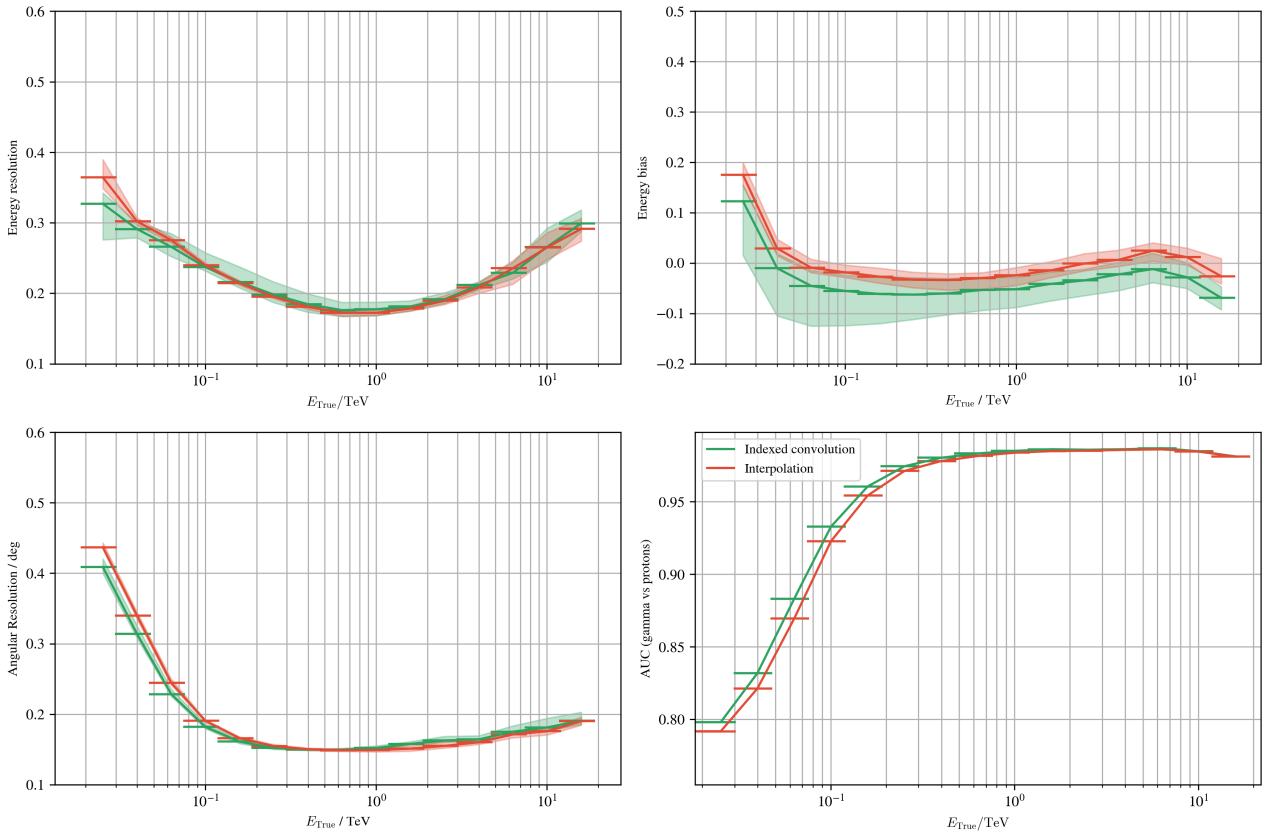


Figure 8.16: Caption

8.11 Annex: Application of multi-task balancing to the γ-PhysNet

Introduction

The full-event reconstruction of the incident particle involves retrieving its physics parameters that are the energy, direction and particle type. Previous works, notably [Jac20] in the case of IACT, have pointed out the interest of reconstructing the parameters simultaneously in a multi-task scenario for each task to benefit each other as they are closely correlated. Traditionally, the γ-PhysNet neural network utilizes the Uncertainty Weighting algorithm presented in [KGC18], but this methodology suffers from a limited loss function selection as it is mathematically constrained to the *L1*, *L2* and *Cross-Entropy* criteria. Furthermore, the introduction of domain adaptation involved the use of mathematical tools that are not considered as likelihood per se, prohibiting the extension of this balancing technique for most of the models. In this appendix, we are evaluating a selection of multi-task balancing approaches for the optimization of the γ-PhysNet and evaluate their performance on simulated data. The description of the selected methods are given in Section 2.8.

Results

Although multi-task balancing methods reduced the computational complexity of optimizing a neural network by cancelling the need for expensive grid search procedures, Dynamic

Weight Averaging and GradNorm are influenced by an additional hyperparameter. Figures 8.17 and 8.18 respectively illustrates the impact of the hyperparameter T for DWA and α for GN. In the former, we vary the value of $T \in \{0.1, 0.5, 2.0\}$, whereas in the latter, we cover the range $\alpha \in \{0.1, 0.5, 1.5, 3.0\}$. For all experiments, 5 seeds are produced to account for the variability introduced by the parameter initialization.

It appears that low values of T, $T = 0.1$ in our case, increase slightly the dependency to parameter initialization, as the area covered by the min and max increases. It is especially visible for the energy bias. On the contrary, higher values reduce this uncertainty. Overall, on average, DWA gives the same performance regarding the value of T in the case of IACT simulated image analysis. Regarding GN and the hyperparameter α, the results are more contrasted. Concerning the energy bias and resolution, it is clear that higher values of α degrades the performance at higher energy. This remark is still visible on the angular resolution metric, but its effect is not perceptible on the classification power. In conclusion, the value of $\alpha = 0.1$ is selected when the γ-PhysNet is combined with GN, and a value of $T = 2$ is chosen when the neural network is used with DWA.

Figure 8.19 depicts the performance of the γ-PhysNet with the selected hyperparameters of GN and DWA. The balancing methods are compared with the baselines EW and RLW. Three main conclusions can be drawn from these results. Firstly, improvements brought by the automatic algorithms are especially visible at higher energy, greater than 1 TeV. The low energy levels reflect a limited enhancement solely on the angular resolution metric. Secondly, although DWA slightly improves the performance compared to the baseline, the gain remains small. Lastly, two methods stand out from our selected, UW and GN and propose a remarkable upgrade of the angular resolution compared to the other ones. A demarcation is yet visible at the highest energy resolution, where UW performs best compared to any other approaches.

To conclude this annexe, UW remains the best performing balancing approach for the γ-PhysNet. However, as it is not possible to extend it to all possible loss functions, GN is a plausible solution to be considered. In our work, GN is combined with domain adaptation in the case of DeepJDOT and DeepCORAL as they do not fall in the mathematical framework of UW.

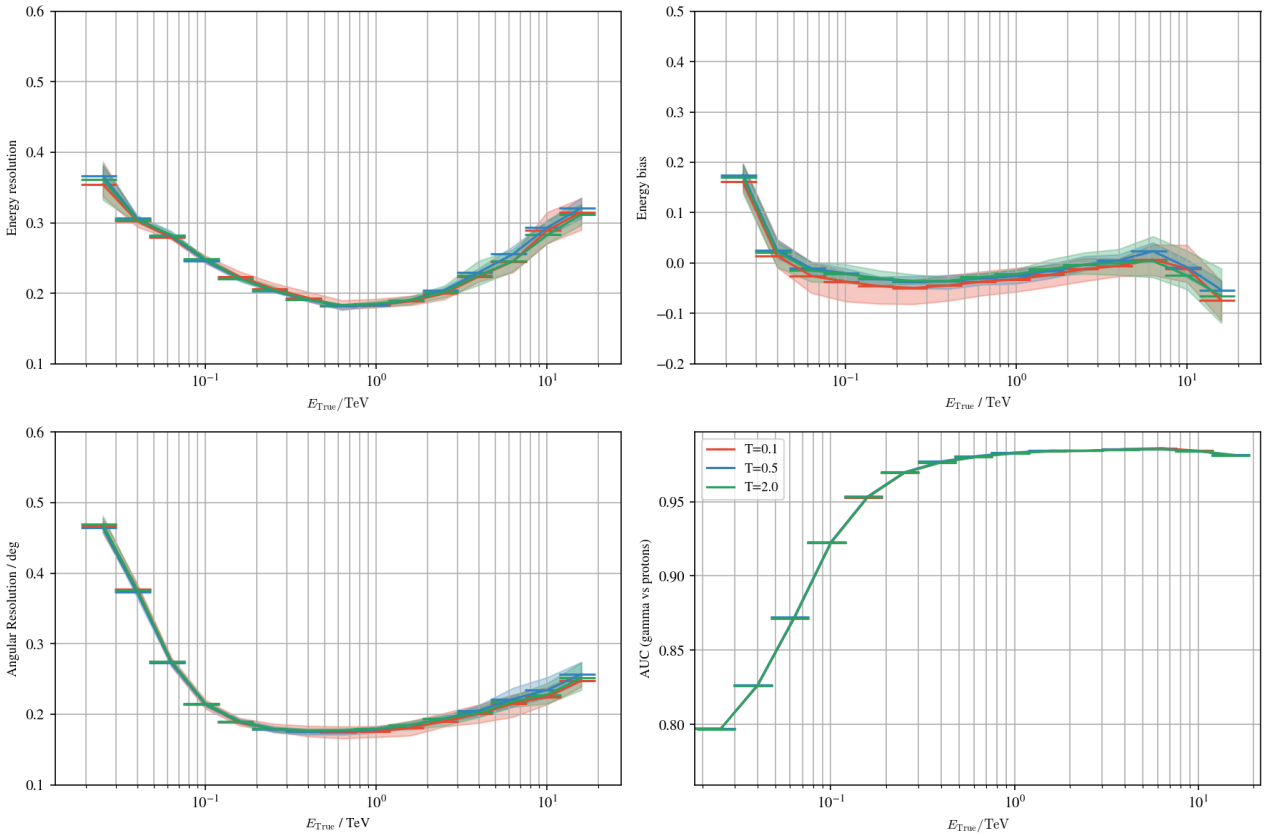


Figure 8.17: Impact of the hyperparameter T on the γ-PhysNet performance combined with DWA.

8.12 Annex: Weighting from the energy distribution

Introduction

From the description of the standard dataset, Annexe 8.3 has highlighted the presence of biases in the energy distribution. They arise from energy levels that are more frequently met during training. Consequently, the neural network tends to favor these over-represented quantities, potentially losing its generalization capabilities. A possible solution to mitigate this bias is to assign greater importance to the less frequently encountered energy levels through the introduction of a weighting function $w(t)$. By applying this weighting function, the loss is adjusted to become a weighted sum of the energy quantities, thereby encouraging the model to treat all energy levels more equally. The modified loss function can be expressed as:

$$\mathcal{L}_{\text{energy}}(\theta) = \sum_{i=1}^{N} w_i \|y_i - f^{\theta}(x_i)\|_{p=1} \tag{8.18}$$

This function can be computed from the energy distribution, and there is multiple possibility of calculating it. In our case, we chose to product the weighting function as follows:

- We calculate the histogram of the log-energy level with a total of 100 bins.
- We apply a Gaussian filter of standard deviation $\sigma = 1$ to smooth the histogram values.

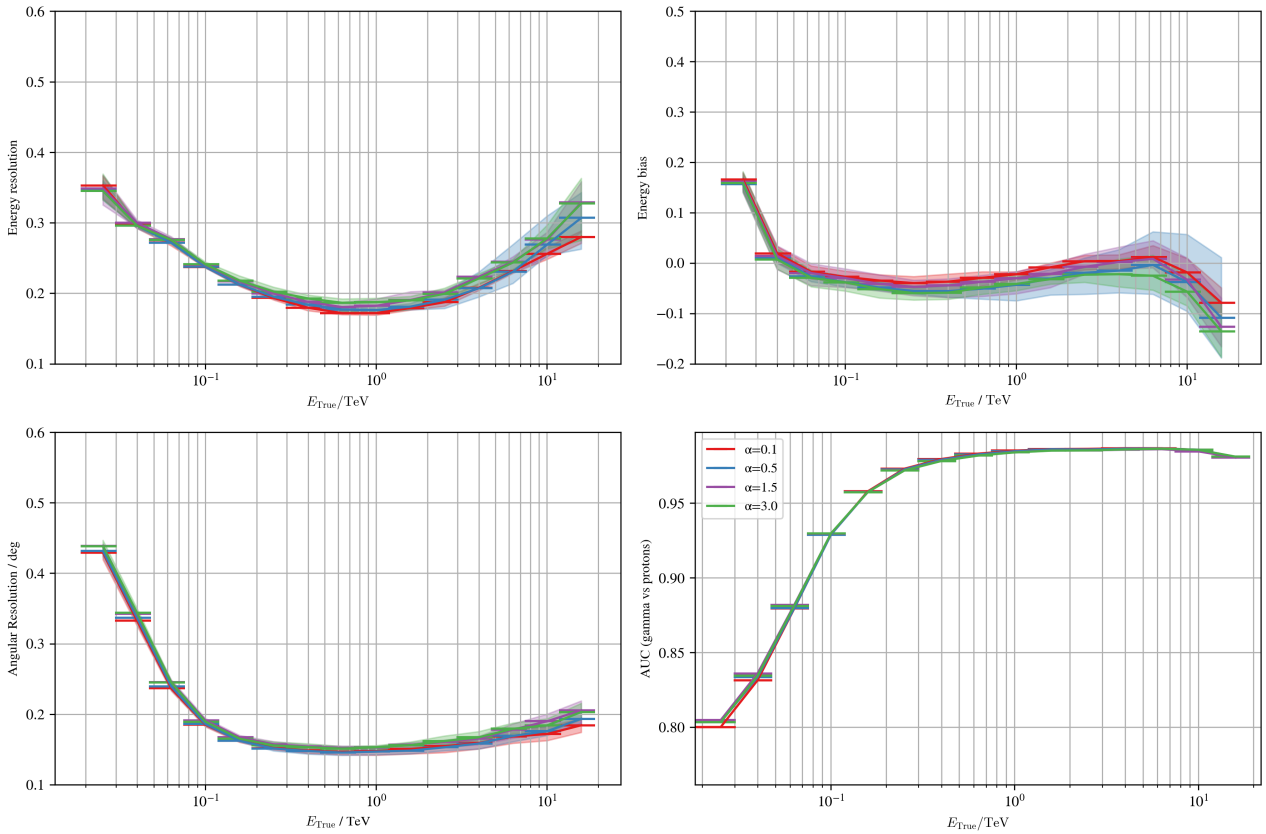


Figure 8.18: Impact of the hyperparameter α on the γ-PhysNet performance combined with GN.

- We compute the weights as the inverse of the smooth histogram.
- We clip the weights to a min and max values. In our case, we chose a minimum value of 1 and a maximum value of 50.

Because we don't have access to the continuous representation of the weighting function, weights are interpolated between the two closest positions. The calculated weighting function is illustrated in Figure 8.20.

Results

The results are depicted in Figure 8.21.Using the weighting function illustrated in Figure 8.20, the conclusion is twofold. Firstly, we notice that the energy resolution exhibits a degradation at the lower range. Nevertheless, it is quite remarkable that we manage to almost perfectly cancel the energy bias across the whole energy spectrum that is accessible. Finally, the other metrics, angular resolution and AUC, seems to be sightly deteriorated in favor of the traditional γ-PhysNet.

In conclusion, weighting the energies directly within the loss function offers several benefits. However, extensive work is required to determine the optimal weighting function that improves model performance without compromising the energy resolution. Additionally, the application of such a weighting function must be approached with caution, as it also affects the gradients of the considered task. This could potentially complicate the opti-

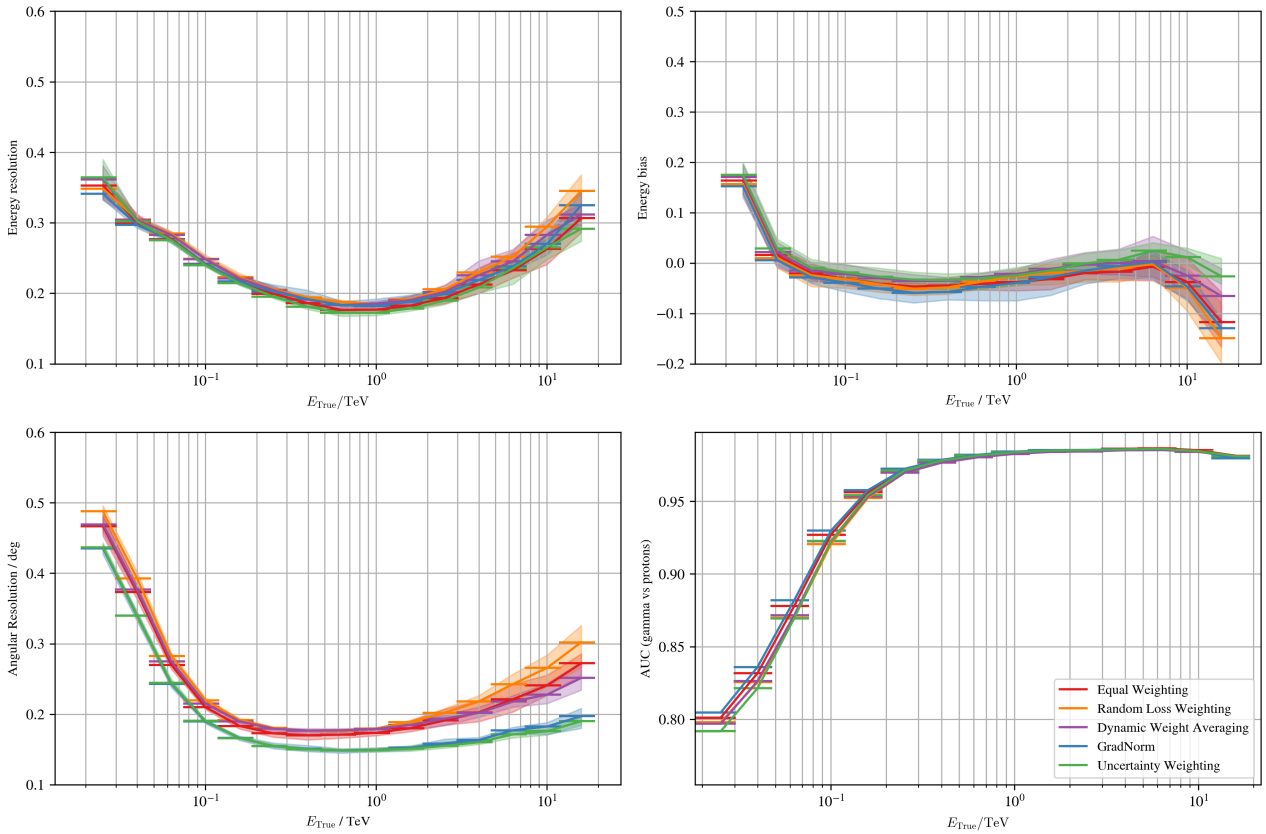


Figure 8.19: Comparison the EW, RLW, DWA, GN and UW on the γ-PhysNet.

mization process and make convergence more difficult. Therefore, the inclusion of energy weighting is left for a future work.

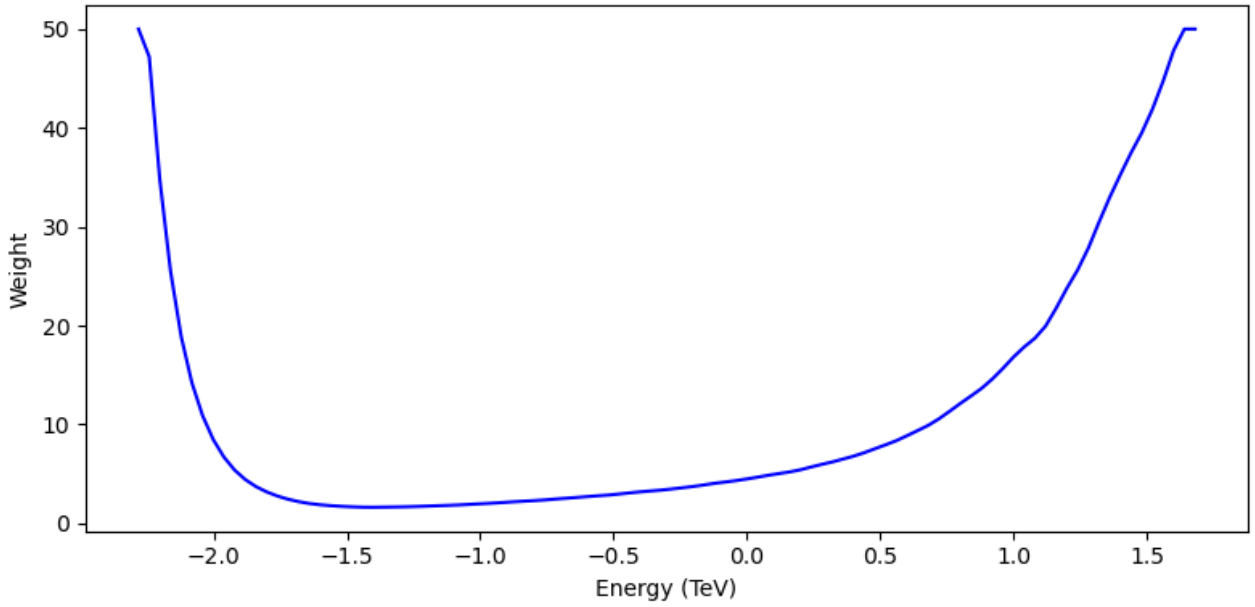


Figure 8.20: Computed weights as a function of the energy.

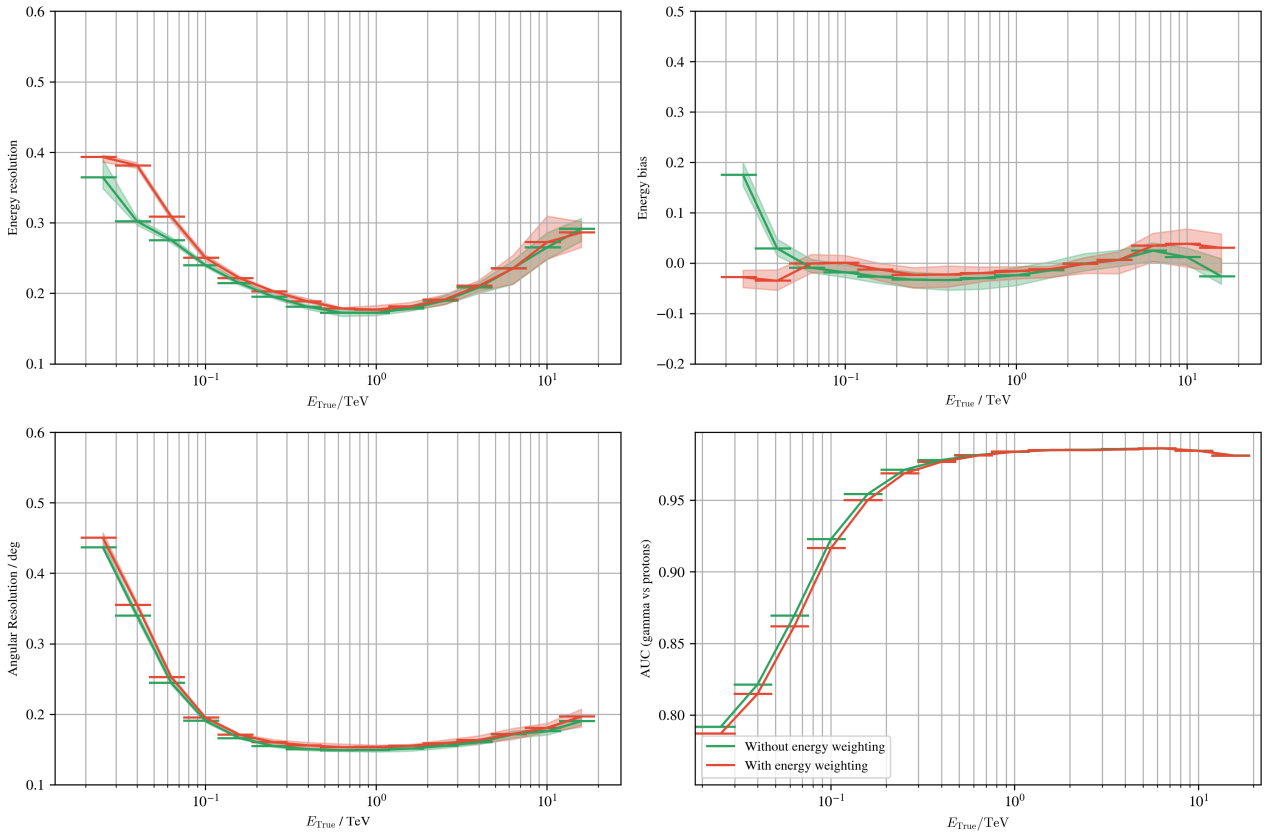


Figure 8.21: Comparison of the γ-PhysNet performance with and without energy weighting.

Bibliography

[79] *A Strategy for Space Astronomy and Astrophysics for the 1980s*. Washington, DC: The National Academies Press, 1979. DOI: 10.17226/12377. URL: https://nap.nationalacademies.org/catalog/12377/a-strategy-for-space-astronomy-and-astrophysics-for-the-1980s.

[Abd+18] H. Abdalla et al. "H.E.S.S. observations of RX J1713.7-3946 with improved angular and spectral resolution: Evidence for gamma-ray emission extending beyond the X-ray emitting shell". In: *Astronomy - Astrophysics* 612 (Apr. 2018), A6. ISSN: 1432-0746. DOI: 10.1051/0004-6361/201629790. URL: http://dx.doi.org/10.1051/0004-6361/201629790.

[Abd+21] H. Abdalla et al. "Revealing x-ray and gamma ray temporal and spectral similarities in the GRB 190829A afterglow". In: *Science* 372.6546 (June 2021), pp. 1081–1085. ISSN: 1095-9203. DOI: 10.1126/science.abe8560. URL: http://dx.doi.org/10.1126/science.abe8560.

[Abd+10] A. A. Abdo et al. "FERMI LARGE AREA TELESCOPE FIRST SOURCE CATALOG". In: *The Astrophysical Journal Supplement Series* 188.2 (May 2010), pp. 405–436. ISSN: 1538-4365. DOI: 10.1088/0067-0049/188/2/405. URL: http://dx.doi.org/10.1088/0067-0049/188/2/405.

[Abe+23] H. Abe et al. "Observations of the Crab Nebula and Pulsar with the Large-sized Telescope Prototype of the Cherenkov Telescope Array". In: *Astrophys. J.* 956.2 (2023), p. 80. DOI: 10.3847/1538-4357/ace89d. arXiv: 2306.12960 [astro-ph.HE].

[Abe23] H. Abe. "Performance of the joint LST-1 and MAGIC observations evaluated with Crab Nebula data". In: *Astronomy; Astrophysics* 680 (Dec. 2023), A66. ISSN: 1432-0746. DOI: 10.1051/0004-6361/202346927. URL: http://dx.doi.org/10.1051/0004-6361/202346927.

[Abe+24] S. Abe et al. "Dark matter line searches with the Cherenkov Telescope Array". In: *Journal of Cosmology and Astroparticle Physics* 2024.07 (July 2024), p. 047. ISSN: 1475-7516. DOI: 10.1088/1475-7516/2024/07/047. URL: http://dx.doi.org/10.1088/1475-7516/2024/07/047.

[al08] J. Albert et al. "Implementation of the Random Forest method for the Imaging Atmospheric Cherenkov Telescope MAGIC". In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 588.3 (2008), pp. 424–432. ISSN: 0168-9002. DOI: https://doi.org/10.1016/j.nima.2007.11.068. URL: https://www.sciencedirect.com/science/article/pii/S0168900207024059.

[al21] Yukiho Kobayashi et al. “Camera Calibration of the CTA-LST prototype”. In: *Proceedings of 37th International Cosmic Ray Conference — PoS(ICRC2021)*. Sissa Medialab, July 2021. DOI: 10.22323/1.395.0720. URL: https://doi.org/10.22323%2F1.395.0720.

[ACB17] Martin Arjovsky, Soumith Chintala, and Léon Bottou. *Wasserstein GAN*. 2017. arXiv: 1701.07875 [stat.ML].

[Azi+19] Kamyar Azizzadenesheli et al. *Regularized Learning for Domain Adaptation under Label Shifts*. 2019. arXiv: 1903.09734 [cs.LG].

[BKH16] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. *Layer Normalization*. 2016. arXiv: 1607.06450 [stat.ML].

[Bai03] C Baixeras. “The MAGIC telescope”. In: *Nuclear Physics B - Proceedings Supplements* 114 (2003). Proceedings of the XXXth International Meeting of Fundamentals Physics, pp. 247–252. ISSN: 0920-5632. DOI: https://doi.org/10.1016/S0920-5632(02)01910-2. URL: https://www.sciencedirect.com/science/article/pii/S0920563202019102.

[Ben+06] Shai Ben-David et al. “Analysis of Representations for Domain Adaptation”. In: *Advances in Neural Information Processing Systems*. Ed. by B. Schölkopf, J. Platt, and T. Hoffman. Vol. 19. MIT Press, 2006. URL: https://proceedings.neurips.cc/paper/2006/file/b1b0432ceafb0ce714426e9114852ac7-Paper.pdf.

[Ber08] Konrad Bernlöhr. “Simulation of imaging atmospheric Cherenkov telescopes with CORSIKA and sim_telarray”. In: *Astroparticle Physics* 30.3 (2008), pp. 149–158. ISSN: 0927-6505. DOI: https://doi.org/10.1016/j.astropartphys.2008.07.009. URL: https://www.sciencedirect.com/science/article/pii/S0927650508000972.

[BH34] H. Bethe and W. Heitler. “On the Stopping of Fast Particles and on the Creation of Positive Electrons”. In: *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* 146.856 (1934), pp. 83–112. ISSN: 09501207. URL: http://www.jstor.org/stable/2935479 (visited on 09/09/2024).

[Bjo+18] Johan Bjorck et al. *Understanding Batch Normalization*. 2018. arXiv: 1806.02375 [cs.LG].

[Bol+22] Julien Bolmont et al. “First Combined Study on Lorentz Invariance Violation from Observations of Energy-dependent Time Delays from Multiple-type Gamma-Ray Sources. I. Motivation, Method Description, and Validation through Simulations of H.E.S.S., MAGIC, and VERITAS Data Sets”. In: *The Astrophysical Journal* 930.1 (May 2022), p. 75. ISSN: 1538-4357. DOI: 10.3847/1538-4357/ac5048. URL: http://dx.doi.org/10.3847/1538-4357/ac5048.

[Bon23] Mathieu De Bony de Lavergne. “Gamma-Ray Bursts observations with Cherenkov telescopes : H.E.S.S. legacy observations and optimisation of follow-up and detection with the Large-Sized Telescop”. Theses. Université Savoie Mont Blanc, Feb. 2023. URL: https://theses.hal.science/tel-04496699.

[Bra+99] S. M. Bradbury et al. *The Very Energetic Radiation Imaging Telescope Array System (VERITAS)*. 1999. arXiv: astro-ph/9907248 [astro-ph]. URL: https://arxiv.org/abs/astro-ph/9907248.

[CCM18] Dominique Cardon, Jean-Philippe Cointet, and Antoine Mazières. "La revanche des neurones : L'invention des machines inductives et la controverse de l'intelligence artificielle". In: *Réseaux : communication, technologie, société* 5.211 (Dec. 2018), pp. 173–220. DOI: 10.3917/res.211.0173. URL: https://hal-sciencespo.archives-ouvertes.fr/hal-02005537.

[Car04] Rich Caruana. "Multitask Learning". In: *Machine Learning* 28 (2004), pp. 41–75.

[Che+19] Chao Chen et al. *HoMM: Higher-order Moment Matching for Unsupervised Domain Adaptation*. 2019. arXiv: 1912.11976 [cs.CV].

[Che+18] Zhao Chen et al. *GradNorm: Gradient Normalization for Adaptive Loss Balancing in Deep Multitask Networks*. 2018. arXiv: 1711.02257 [cs.CV].

[Che17] Jill Chevalier. "Active galactic nuclei population study at TeV with the H.E.S.S. telescopes and variability studies of the blazar PKS 2155-304 with SSC modelling". Theses. Université Grenoble Alpes, June 2017. URL: https://theses.hal.science/tel-03072832.

[Chu+22] Ren Chuan-Xian et al. *Towards Unsupervised Domain Adaptation via Domain-Transformer*. 2022. DOI: 10.48550/ARXIV.2202.13777. URL: https://arxiv.org/abs/2202.13777.

[CUH16] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. *Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs)*. 2016. arXiv: 1511.07289 [cs.LG].

[Col+24] MAGIC Collaboration et al. *Constraints on Lorentz invariance violation from the extraordinary Mrk 421 flare of 2014 using a novel analysis method*. 2024. arXiv: 2406.07140 [astro-ph.HE]. URL: https://arxiv.org/abs/2406.07140.

[CLJ20] Jean-Baptiste Cordonnier, Andreas Loukas, and Martin Jaggi. *Multi-Head Attention: Collaborate Instead of Concatenate*. 2020. DOI: 10.48550/ARXIV.2006.16362. URL: https://arxiv.org/abs/2006.16362.

[Cou+15] Nicolas Courty et al. *Optimal Transport for Domain Adaptation*. 2015. DOI: 10.48550/ARXIV.1507.00504. URL: https://arxiv.org/abs/1507.00504.

[Csu17] Gabriela Csurka. "Domain Adaptation for Visual Applications: A Comprehensive Survey". In: *CoRR* abs/1702.05374 (2017). arXiv: 1702.05374. URL: http://arxiv.org/abs/1702.05374.

[Cui+20] Shuhao Cui et al. *Gradually Vanishing Bridge for Adversarial Domain Adaptation*. 2020. arXiv: 2003.13183 [cs.CV].

[Cyb88] G. Cybenko. "Continuous Valued Neural Networks with Two Hidden Layers Are Sufficient". In: (1988).

[Dam+18] Bharath Bhushan Damodaran et al. "DeepJDOT: Deep Joint Distribution Optimal Transport for Unsupervised Domain Adaptation". In: (2018). DOI: 10.48550/ARXIV.1803.10081. URL: https://arxiv.org/abs/1803.10081.

[De+22] Songshaptak De et al. *Deep learning techniques for Imaging Air Cherenkov Telescopes*. 2022. DOI: 10.48550/ARXIV.2206.05296. URL: https://arxiv.org/abs/2206.05296.

[Del+23] Michaël Dell'aiera et al. "Deep unsupervised domain adaptation applied to the Cherenkov Telescope Array Large-Sized Telescope". In: *20th International Conference on Content-based Multimedia Indexing*. CBMI 2023. ACM, Sept. 2023. DOI: 10.1145/3617233.3617279. URL: http://dx.doi.org/10.1145/3617233.3617279.

[Del+24] Michael Dellaiera et al. *Deep Learning and IACT: Bridging the gap between Monte-Carlo simulations and LST-1 data using domain adaptation*. 2024. arXiv: 2403.13633 [astro-ph.IM]. URL: https://arxiv.org/abs/2403.13633.

[DK24] A. Demichev and A. Kryukov. "Using deep learning methods for IACT data analysis in gamma-ray astronomy: A review". In: *Astronomy and Computing* 46 (2024), p. 100793. ISSN: 2213-1337. DOI: https://doi.org/10.1016/j.ascom.2024.100793. URL: https://www.sciencedirect.com/science/article/pii/S2213133724000088.

[Dos+20] Alexey Dosovitskiy et al. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. 2020. DOI: 10.48550/ARXIV.2010.11929. URL: https://arxiv.org/abs/2010.11929.

[Evc+21] Mikhail Evchenko et al. *Frugal Machine Learning*. 2021. arXiv: 2111.03731 [cs.LG]. URL: https://arxiv.org/abs/2111.03731.

[Faw06] Tom Fawcett. "An introduction to ROC analysis". In: *Pattern Recognition Letters* 27.8 (2006). ROC Analysis in Pattern Recognition, pp. 861–874. ISSN: 0167-8655. DOI: https://doi.org/10.1016/j.patrec.2005.10.010. URL: https://www.sciencedirect.com/science/article/pii/S016786550500303X.

[FT37] I. M. Frank and I. E. Tamm. "Coherent visible radiation of fast electrons passing through matter". In: *Compt. Rend. Acad. Sci. URSS* 14.3 (1937). Ed. by V. L. Ginzburg, B. M. Bolotovsky, and I. M. Dremin, pp. 109–114. DOI: 10.3367/UFNr.0093.196710o.0388.

[Fu+19] Jun Fu et al. *Dual Attention Network for Scene Segmentation*. 2019. arXiv: 1809.02983 [cs.CV].

[GG16] Yarin Gal and Zoubin Ghahramani. *Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning*. 2016. arXiv: 1506.02142 [stat.ML]. URL: https://arxiv.org/abs/1506.02142.

[GJ53] W. Galbraith and J. V. Jelley. "Light Pulses from the Night Sky Associated with Cosmic Rays". In: *Nature* 171.4347 (Feb. 1, 1953), pp. 349–350. ISSN: 1476-4687. DOI: 10.1038/171349a0. URL: https://doi.org/10.1038/171349a0.

[Gan+16] Yaroslav Ganin et al. *Domain-Adversarial Training of Neural Networks*. 2016. arXiv: 1505.07818 [stat.ML].

[Ghi+15] Muhammad Ghifary et al. *Domain Generalization for Object Recognition with Multi-task Autoencoders*. 2015. arXiv: 1508.07680 [cs.CV].

[Glo+23] J. Glombitza et al. "Application of graph networks to background rejection in Imaging Air Cherenkov Telescopes". In: *Journal of Cosmology and Astroparticle Physics* 2023.11 (Nov. 2023), p. 008. ISSN: 1475-7516. DOI: 10.1088/1475-7516/2023/11/008. URL: http://dx.doi.org/10.1088/1475-7516/2023/11/008.

[GB10] Xavier Glorot and Yoshua Bengio. "Understanding the difficulty of training deep feedforward neural networks". In: *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. Ed. by Yee Whye Teh and Mike Titterington. Vol. 9. Proceedings of Machine Learning Research. Chia Laguna Resort, Sardinia, Italy: PMLR, May 2010, pp. 249–256. URL: https://proceedings.mlr.press/v9/glorot10a.html.

[GBC16] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. http://www.deeplearningbook.org. MIT Press, 2016.

[Goo+14] Ian J. Goodfellow et al. *Generative Adversarial Networks*. 2014. arXiv: 1406.2661 [stat.ML].

[Gua+22] SHI Guangyuan et al. "Recon: Reducing Conflicting Gradients From the Root For Multi-Task Learning". In: *The Eleventh International Conference on Learning Representations*. 2022.

[Gul+17] Ishaan Gulrajani et al. *Improved Training of Wasserstein GANs*. 2017. arXiv: 1704.00028 [cs.LG].

[17] "GW170817: Observation of Gravitational Waves from a Binary Neutron Star Inspiral". In: *Phys. Rev. Lett.* 119 (16 Oct. 2017), p. 161101. DOI: 10.1103/PhysRevLett.119.161101. URL: https://link.aps.org/doi/10.1103/PhysRevLett.119.161101.

[He+15a] Kaiming He et al. *Deep Residual Learning for Image Recognition*. 2015. arXiv: 1512.03385 [cs.CV].

[He+15b] Kaiming He et al. *Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification*. 2015. arXiv: 1502.01852 [cs.CV].

[He+16] Kaiming He et al. *Identity Mappings in Deep Residual Networks*. 2016. arXiv: 1603.05027 [cs.CV].

[He+17] Kaiming He et al. *Mask R-CNN*. 2017. DOI: 10.48550/ARXIV.1703.06870. URL: https://arxiv.org/abs/1703.06870.

[He+21] Kaiming He et al. *Masked Autoencoders Are Scalable Vision Learners*. 2021. DOI: 10.48550/ARXIV.2111.06377. URL: https://arxiv.org/abs/2111.06377.

[Hec+98] D. Heck et al. *CORSIKA: a Monte Carlo code to simulate extensive air showers*. 1998.

[HG23] Dan Hendrycks and Kevin Gimpel. *Gaussian Error Linear Units (GELUs)*. 2023. arXiv: 1606.08415 [cs.LG].

[Hey+09] Tony Hey et al. *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research, Oct. 2009. ISBN: 978-0-9825442-0-4. URL: https://www.microsoft.com/en-us/research/publication/fourth-paradigm-data-intensive-scientific-discovery/.

[Hil85] A. M. Hillas. "Cerenkov Light Images of EAS Produced by Primary Gamma Rays and by Nuclei". In: *19th International Cosmic Ray Conference (ICRC19), Volume 3*. Vol. 3. International Cosmic Ray Conference. Aug. 1985, p. 445.

[Hin04] J.A Hinton. "The status of the HESS project". In: *New Astronomy Reviews* 48.5–6 (Apr. 2004), pp. 331–337. ISSN: 1387-6473. DOI: 10.1016/j.newar.2003.12.004. URL: http://dx.doi.org/10.1016/j.newar.2003.12.004.

[Hof18] Werner Hofmann. "Perspectives from CTA in relativistic astrophysics". In: *Fourteenth Marcel Grossmann Meeting - MG14*. Ed. by Massimo Bianchi, Robert T. Jansen, and Remo Ruffini. Jan. 2018, pp. 223–242. DOI: 10.1142/9789813226609_0014.

[Hu+19] Jie Hu et al. *Squeeze-and-Excitation Networks*. 2019. arXiv: 1709.01507 [cs.CV].

[Hua+16] Gao Huang et al. *Deep Networks with Stochastic Depth*. 2016. arXiv: 1603.09382 [cs.LG].

[Hul94] J.J. Hull. "A database for handwritten text recognition research". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 16.5 (1994), pp. 550–554. DOI: 10.1109/34.291440.

[IS15] Sergey Ioffe and Christian Szegedy. *Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift*. 2015. arXiv: 1502.03167 [cs.LG].

[Ion+14] Bogdan Ionescu et al. *Fusion in Computer Vision: Understanding Complex Visual Content*. Springer Publishing Company, Incorporated, 2014. ISBN: 3319056956.

[Iso+16] Phillip Isola et al. *Image-to-Image Translation with Conditional Adversarial Networks*. 2016. DOI: 10.48550/ARXIV.1611.07004. URL: https://arxiv.org/abs/1611.07004.

[Jac20] Mikaël Jacquemont. "Cherenkov Image Analysis with Deep Multi-Task Learning from Single-Telescope Data". Theses. Université Savoie Mont Blanc, Nov. 2020. URL: https://hal.science/tel-03590369.

[Jac+21] Mikael Jacquemont et al. "First Full-Event Reconstruction from Imaging Atmospheric Cherenkov Telescope Real Data with Deep Learning". In: *2021 International Conference on Content-Based Multimedia Indexing (CBMI)*. IEEE, June 2021. DOI: 10.1109/cbmi50038.2021.9461918. URL: https://doi.org/10.1109%2Fcbmi50038.2021.9461918.

[Jac+19] Mikael Jacquemont. et al. "Indexed Operations for Non-rectangular Lattices Applied to Convolutional Neural Networks". In: *Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 5: VISAPP*, INSTICC. SciTePress, 2019, pp. 362–371. ISBN: 978-989-758-354-4. DOI: 10.5220/0007364303620371.

[Jos+22] Laurent Valentin Jospin et al. "Hands-On Bayesian Neural Networks—A Tutorial for Deep Learning Users". In: *IEEE Computational Intelligence Magazine* 17.2 (May 2022), pp. 29–48. ISSN: 1556-6048. DOI: 10.1109/mci.2022.3155327. URL: http://dx.doi.org/10.1109/MCI.2022.3155327.

[KGC18] Alex Kendall, Yarin Gal, and Roberto Cipolla. *Multi-Task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics*. 2018. arXiv: 1705.07115 [cs.CV].

[Kim+22] Bryan Kim et al. *DL1-Data-Handler: DL1 HDF5 writer, reader, and processor for IACT data*. Version v0.10.1. Jan. 2022. DOI: 10.5281/zenodo.5844134. URL: https://doi.org/10.5281/zenodo.5844134.

[KW17] Thomas N. Kipf and Max Welling. *Semi-Supervised Classification with Graph Convolutional Networks*. 2017. arXiv: 1609.02907 [cs.LG]. URL: https://arxiv.org/abs/1609.02907.

[KL80] V. Klema and A. Laub. "The singular value decomposition: Its computation and some applications". In: *IEEE Transactions on Automatic Control* 25.2 (1980), pp. 164–176.

[Koh+96] A. Kohnle et al. "Stereoscopic imaging of air showers with the first two HEGRA Cherenkov telescopes". In: *Astroparticle Physics* 5.2 (1996), pp. 119–131. ISSN: 0927-6505. DOI: https://doi.org/10.1016/0927-6505(96)00011-4. URL: https://www.sciencedirect.com/science/article/pii/0927650596000114.

[KMR18] Soheil Kolouri, Charles E. Martin, and Gustavo K. Rohde. "Sliced-Wasserstein Autoencoder: An Embarrassingly Simple Generative Model". In: *CoRR* abs/1804.01947 (2018). arXiv: 1804.01947. URL: http://arxiv.org/abs/1804.01947.

[KL21] Wouter M. Kouw and Marco Loog. "A Review of Domain Adaptation without Target Labels". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43.3 (Mar. 2021), pp. 766–785. ISSN: 1939-3539. DOI: 10.1109/tpami.2019.2945942. URL: http://dx.doi.org/10.1109/TPAMI.2019.2945942.

[KSH12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "ImageNet Classification with Deep Convolutional Neural Networks". In: *Advances in Neural Information Processing Systems*. Ed. by F. Pereira et al. Vol. 25. Curran Associates, Inc., 2012. URL: https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

[KF14] Meelis Kull and Peter A. Flach. "Patterns of dataset shift". In: 2014. URL: https://api.semanticscholar.org/CorpusID:211733031.

[Kum17] Siddharth Krishna Kumar. *On weight initialization in deep neural networks*. 2017. arXiv: 1704.08863 [cs.LG]. URL: https://arxiv.org/abs/1704.08863.

[LeC89] Y. LeCun. "Generalization and Network Design Strategies". In: *Connectionism in Perspective*. Ed. by R. Pfeifer et al. an extended version was published as a technical report of the University of Toronto. Zurich, Switzerland: Elsevier, 1989.

[Lec+98] Y. Lecun et al. "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324. DOI: 10.1109/5.726791.

[LeC+12] Yann A. LeCun et al. "Efficient BackProp". In: *Neural Networks: Tricks of the Trade: Second Edition*. Ed. by Grégoire Montavon, Geneviève B. Orr, and Klaus-Robert Müller. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 9–48. ISBN: 978-3-642-35289-8. DOI: 10.1007/978-3-642-35289-8_3. URL: https://doi.org/10.1007/978-3-642-35289-8_3.

[Lee+19] Chen-Yu Lee et al. *Sliced Wasserstein Discrepancy for Unsupervised Domain Adaptation*. 2019. arXiv: 1903.04064 [cs.CV].

[Lem04] Marianne Lemoine-Goumard. "Stéréoscopie de gerbes de gamma avec les télescopes H. E. S. S. : premières images de vestiges de supernovæ au TeV". 2006EPXX0014. PhD thesis. 2004, 1 vol. (297 p.) URL: http://www.theses.fr/2006EPXX0014.

[Les+01a] R.W. Lessard et al. "A new analysis method for reconstructing the arrival direction of TeV gamma rays using a single imaging atmospheric Cherenkov telescope". In: *Astroparticle Physics* 15.1 (2001), pp. 1–18. ISSN: 0927-6505. DOI: https://doi.org/10.1016/S0927-6505(00)00133-X. URL: https://www.sciencedirect.com/science/article/pii/S092765050000133X.

[Les+01b] R.W. Lessard et al. "A new analysis method for reconstructing the arrival direction of TeV gamma rays using a single imaging atmospheric Cherenkov telescope". In: *Astroparticle Physics* 15.1 (Mar. 2001), pp. 1–18. ISSN: 0927-6505. DOI: 10.1016/s0927-6505(00)00133-x. URL: http://dx.doi.org/10.1016/S0927-6505(00)00133-X.

[Li+17] Da Li et al. *Learning to Generalize: Meta-Learning for Domain Generalization*. 2017. DOI: 10.48550/ARXIV.1710.03463. URL: https://arxiv.org/abs/1710.03463.

[LM83] T. -P. Li and Y. -Q. Ma. "Analysis methods for results in gamma-ray astronomy." In: 272 (Sept. 1983), pp. 317–324. DOI: 10.1086/161295.

[Lin+22] Baijiong Lin et al. *Reasonable Effectiveness of Random Weighting: A Litmus Test for Multi-Task Learning*. 2022. arXiv: 2111.10603 [cs.LG].

[LWS18] Zachary C. Lipton, Yu-Xiang Wang, and Alex Smola. *Detecting and Correcting for Label Shift with Black Box Predictors*. 2018. arXiv: 1802.03916 [cs.LG].

[LT16] Ming-Yu Liu and Oncel Tuzel. *Coupled Generative Adversarial Networks*. 2016. arXiv: 1606.07536 [cs.CV].

[LJD19] Shikun Liu, Edward Johns, and Andrew J. Davison. *End-to-End Multi-Task Learning with Attention*. 2019. arXiv: 1803.10704 [cs.CV].

[Liu+21a] Xiaofeng Liu et al. *Adversarial Unsupervised Domain Adaptation with Conditional and Label Shift: Infer, Align and Iterate*. 2021. arXiv: 2107.13469 [cs.CV].

[Liu+21b] Ze Liu et al. *Swin Transformer: Hierarchical Vision Transformer using Shifted Windows*. 2021. arXiv: 2103.14030 [cs.CV].

[Liu+22] Zhuang Liu et al. *A ConvNet for the 2020s*. 2022. arXiv: 2201.03545 [cs.CV].

[Lon+15] Mingsheng Long et al. *Learning Transferable Features with Deep Adaptation Networks*. 2015. DOI: 10.48550/ARXIV.1502.02791. URL: https://arxiv.org/abs/1502.02791.

[Lop+22] Ruben Lopez-Coto et al. *cta-observatory/cta-lstchain: v0.9.6 - 2022-04-13*. Version v0.9.6. Apr. 2022. DOI: 10.5281/zenodo.6458862. URL: https://doi.org/10.5281/zenodo.6458862.

[LH19] Ilya Loshchilov and Frank Hutter. *Decoupled Weight Decay Regularization*. 2019. arXiv: 1711.05101 [cs.LG]. URL: https://arxiv.org/abs/1711.05101.

[Luo+17] Wenjie Luo et al. *Understanding the Effective Receptive Field in Deep Convolutional Neural Networks*. 2017. DOI: 10.48550/ARXIV.1701.04128. URL: https://arxiv.org/abs/1701.04128.

[Maa13] Andrew L. Maas. "Rectifier Nonlinearities Improve Neural Network Acoustic Models". In: 2013. URL: https://api.semanticscholar.org/CorpusID:16489696.

[Mal16] Stéphane Mallat. "Understanding deep convolutional networks". In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374.2065 (Apr. 2016), p. 20150203. ISSN: 1471-2962. DOI: 10.1098/rsta.2015.0203. URL: http://dx.doi.org/10.1098/rsta.2015.0203.

[Mar91] Stephen P. Maran. *The Astronomy and Astrophysics Encyclopedia*. 1991.

[Mas10] Julien Masbou. "Étude de la sensibilité de H.E.S.S. 2 en dessous de 300 GeV et recherche indirecte de matière noire dans les données de H.E.S.S." Theses. Université de Savoie, Sept. 2010. URL: https://tel.archives-ouvertes.fr/tel-00623972.

[McC+55] J. McCarthy et al. *A PROPOSAL FOR THE DARTMOUTH SUMMER RESEARCH PROJECT ON ARTIFICIAL INTELLIGENCE*. http://www-formal.stanford.edu/jmc/history 1955. URL: http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html.

[MP43] Warren Mcculloch and Walter Pitts. "A Logical Calculus of Ideas Immanent in Nervous Activity". In: *Bulletin of Mathematical Biophysics* 5 (1943), pp. 127–147.

[Mie+21] T. Miener et al. *IACT event analysis with the MAGIC telescopes using deep convolutional neural networks with CTLearn*. 2021. DOI: 10.48550/ARXIV.2112.01828. URL: https://arxiv.org/abs/2112.01828.

[MO14] Mehdi Mirza and Simon Osindero. *Conditional Generative Adversarial Nets*. 2014. arXiv: 1411.1784 [cs.LG].

[Mir22] Razmik Mirzoyan. "Technological Novelties of Ground-Based Very High Energy Gamma-Ray Astrophysics with the Imaging Atmospheric Cherenkov Telescopes". In: *Universe* 8.4 (2022). DOI: 10.3390/universe8040219. URL: https://www.mdpi.com/2218-1997/8/4/219.

[MM16] Dmytro Mishkin and Jiri Matas. *All you need is a good init*. 2016. arXiv: 1511.06422 [cs.LG].

[Mor+21] Katelyn Morrison et al. *Exploring Corruption Robustness: Inductive Biases in Vision Transformers and MLP-Mixers*. 2021. arXiv: 2106.13122 [cs.CV].

[18] "Multimessenger observations of a flaring blazar coincident with high-energy neutrino IceCube-170922A". In: *Science* 361.6398 (2018), eaat1378. DOI: 10.1126/science.aat1378. eprint: https://www.science.org/doi/pdf/10.1126/science.aat1378. URL: https://www.science.org/doi/abs/10.1126/science.aat1378.

[MGP15] Thomas Murach, Michael Gajdus, and Robert Daniel Parsons. *A Neural Network-Based Monoscopic Reconstruction Algorithm for H.E.S.S. II*. 2015. arXiv: 1509.00794 [astro-ph.IM].

[Mur+18] Zak Murez et al. “Image to Image Translation for Domain Adaptation”. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2018, pp. 4500–4509. DOI: 10.1109/CVPR.2018.00473.

[NH10] Vinod Nair and Geoffrey E. Hinton. “Rectified Linear Units Improve Restricted Boltzmann Machines”. In: ICML’10. Haifa, Israel: Omnipress, 2010, pp. 807–814. ISBN: 9781605589077.

[NR09] Mathieu de Naurois and Loïc Rolland. “A high performance likelihood reconstruction of gamma-rays for imaging atmospheric Cherenkov telescopes”. In: *Astroparticle Physics* 32.5 (Dec. 2009), pp. 231–252. DOI: 10.1016/j.astropartphys.2009.09.001. URL: https://doi.org/10.1016%2Fj.astropartphys.2009.09.001.

[Net+11] Yuval Netzer et al. “Reading Digits in Natural Images with Unsupervised Feature Learning”. In: *NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011*. 2011. URL: http://ufldl.stanford.edu/housenumbers/nips2011_housenumbers.pdf.

[Nie+17] D. Nieto et al. “Exploring deep learning as an event classification method for the Cherenkov Telescope Array”. In: (2017). DOI: 10.48550/ARXIV.1709.05889. URL: https://arxiv.org/abs/1709.05889.

[Nie+21] D. Nieto et al. *Reconstruction of IACT events using deep learning techniques with CTLearn*. 2021. DOI: 10.48550/ARXIV.2101.07626. URL: https://arxiv.org/abs/2101.07626.

[OC21] Cherenkov Telescope Array Observatory and Cherenkov Telescope Array Consortium. *CTAO Instrument Response Functions - prod5 version v0.1*. Version v0.1. Zenodo, Sept. 2021. DOI: 10.5281/zenodo.5499840. URL: https://doi.org/10.5281/zenodo.5499840.

[OEE09] S. Ohm, C. van Eldik, and K. Egberts. “hadron separation in very-high-energy astronomy using a multivariate analysis method”. In: *Astroparticle Physics* 31.5 (June 2009), pp. 383–391. DOI: 10.1016/j.astropartphys.2009.04.001. URL: https://doi.org/10.1016%2Fj.astropartphys.2009.04.001.

[Ope23] OpenAI. *GPT-4 Technical Report*. 2023. arXiv: 2303.08774 [cs.CL].

[PST18] Luca Pancioni, Friedhelm Schwenker, and Edmondo Trentin, eds. *Artificial Neural Networks in Pattern Recognition*. Springer International Publishing, 2018. DOI: 10.1007/978-3-319-99978-4. URL: https://doi.org/10.1007%2F978-3-319-99978-4.

[PMO22] R. D. Parsons, A. M. W. Mitchell, and S. Ohm. *Investigations of the Systematic Uncertainties in Convolutional Neural Network Based Analysis of Atmospheric Cherenkov Telescope Data*. 2022. arXiv: 2203.05315 [astro-ph.IM].

[PO20] R. D. Parsons and S. Ohm. “Background rejection in atmospheric Cherenkov telescopes using recurrent convolutional neural networks”. In: *The European Physical Journal C* 80.5 (May 2020). DOI: 10.1140/epjc/s10052-020-7953-3. URL: https://doi.org/10.1140%2Fepjc%2Fs10052-020-7953-3.

[PH14] R.D. Parsons and J.A. Hinton. "A Monte Carlo template based analysis for air-Cherenkov arrays". In: *Astroparticle Physics* 56 (Apr. 2014), pp. 26–34. DOI: 10.1016/j.astropartphys.2014.03.002. URL: https://doi.org/10.1016%2Fj.astropartphys.2014.03.002.

[Pir99] Tsvi Piran. "Gamma-ray bursts and the fireball model". In: *Physics Reports* 314.6 (1999), pp. 575–667. ISSN: 0370-1573. DOI: https://doi.org/10.1016/S0370-1573(98)00127-6. URL: https://www.sciencedirect.com/science/article/pii/S0370157398001276.

[Qia+20] Siyuan Qiao et al. *Micro-Batch Training with Batch-Channel Normalization and Weight Standardization*. 2020. arXiv: 1903.10520 [cs.CV]. URL: https://arxiv.org/abs/1903.10520.

[RMN09] Rajat Raina, Anand Madhavan, and Andrew Y. Ng. "Large-Scale Deep Unsupervised Learning Using Graphics Processors". In: *Proceedings of the 26th Annual International Conference on Machine Learning*. ICML '09. Montreal, Quebec, Canada: Association for Computing Machinery, 2009, pp. 873–880. ISBN: 9781605585161. DOI: 10.1145/1553374.1553486. URL: https://doi.org/10.1145/1553374.1553486.

[RL18] Zhongzheng Ren and Yong Jae Lee. "Cross-Domain Self-supervised Multi-task Feature Learning using Synthetic Imagery". In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018.

[Rey93] P. T. Reynolds. "Neural Networks to VHE Gamma-Ray Atmospheric Cherenkov Crab Nebula Imaging Data". In: *Irish Astronomical Journal* 21 (Sept. 1993), p. 118.

[Riq+23] Diego Riquelme et al. "Deep Learning Semi-supervised Strategy for Gamma/Hadron Classification of Imaging Atmospheric Cherenkov Telescope Events". In: (Mar. 2023).

[Ros58] F. Rosenblatt. "The perceptron: A probabilistic model for information storage and organization in the brain." In: *Psychological Review* 65.6 (1958), pp. 386–408. ISSN: 0033-295X. DOI: 10.1037/h0042519. URL: http://dx.doi.org/10.1037/h0042519.

[Rud17] Sebastian Ruder. *An overview of gradient descent optimization algorithms*. 2017. arXiv: 1609.04747 [cs.LG].

[RHW86] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. "Learning Representations by Back-propagating Errors". In: *Nature* 323.6088 (1986), pp. 533–536. DOI: 10.1038/323533a0. URL: http://www.nature.com/articles/323533a0.

[Sai+18] Kuniaki Saito et al. *Maximum Classifier Discrepancy for Unsupervised Domain Adaptation*. 2018. arXiv: 1712.02560 [cs.CV].

[SK16] Tim Salimans and Diederik P. Kingma. *Weight Normalization: A Simple Reparameterization to Accelerate Training of Deep Neural Networks*. 2016. arXiv: 1602.07868 [cs.LG].

[Sal+16] Tim Salimans et al. *Improved Techniques for Training GANs*. 2016. arXiv: 1606.03498 [cs.LG].

[SMG14] Andrew M. Saxe, James L. McClelland, and Surya Ganguli. *Exact solutions to the nonlinear dynamics of learning in deep linear neural networks*. 2014. arXiv: 1312.6120 [cs.NE].

[Sca+09] Franco Scarselli et al. "The Graph Neural Network Model". In: *IEEE Transactions on Neural Networks* 20.1 (2009), pp. 61–80. DOI: 10.1109/TNN.2008.2005605.

[19a] *Science with the Cherenkov Telescope Array*. WORLD SCIENTIFIC, 2019. DOI: 10.1142/10986. eprint: https://www.worldscientific.com/doi/pdf/10.1142/10986. URL: https://www.worldscientific.com/doi/abs/10.1142/10986.

[Sel+19] Ramprasaath R. Selvaraju et al. "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization". In: *International Journal of Computer Vision* 128.2 (Oct. 2019), pp. 336–359. ISSN: 1573-1405. DOI: 10.1007/s11263-019-01228-7. URL: http://dx.doi.org/10.1007/s11263-019-01228-7.

[Sev+22] Jaime Sevilla et al. *Compute Trends Across Three Eras of Machine Learning*. 2022. arXiv: 2202.05924 [cs.LG].

[She+18] Jian Shen et al. *Wasserstein Distance Guided Representation Learning for Domain Adaptation*. 2018. arXiv: 1707.01217 [stat.ML].

[Shi+19] I. Shilon et al. "Application of deep learning methods to analysis of imaging atmospheric Cherenkov telescopes data". In: *Astroparticle Physics* 105 (Feb. 2019), pp. 44–53. DOI: 10.1016/j.astropartphys.2018.10.003. URL: https://doi.org/10.1016%5C%2Fj.astropartphys.2018.10.003.

[Sim91] Herbert A. Simon. "The Architecture of Complexity". In: *Facets of Systems Science*. Boston, MA: Springer US, 1991, pp. 457–476. ISBN: 978-1-4899-0718-9. DOI: 10.1007/978-1-4899-0718-9_31. URL: https://doi.org/10.1007/978-1-4899-0718-9_31.

[Smi+18] Samuel L. Smith et al. *Don't Decay the Learning Rate, Increase the Batch Size*. 2018. arXiv: 1711.00489 [cs.LG].

[Spe+21] S. Spencer et al. "Deep learning with photosensor timing information as a background rejection method for the Cherenkov Telescope Array". In: *Astroparticle Physics* 129 (May 2021), p. 102579. DOI: 10.1016/j.astropartphys.2021.102579. URL: https://doi.org/10.1016%2Fj.astropartphys.2021.102579.

[SG03] F. Richard Stephenson and David A. Green. "Was the supernova of AD 1054 reported in European history?" In: *Journal of Astronomical History and Heritage* 6.1 (June 2003), pp. 46–52.

[SS16] Baochen Sun and Kate Saenko. *Deep CORAL: Correlation Alignment for Deep Domain Adaptation*. 2016. arXiv: 1607.01719 [cs.CV].

[Sun+17] Chen Sun et al. *Revisiting Unreasonable Effectiveness of Data in Deep Learning Era*. 2017. arXiv: 1707.02968 [cs.CV].

[19b] "Teraelectronvolt emission from the gamma-ray burst GRB 190114C". In: *Nature* 575.7783 (Nov. 2019), pp. 455–458. ISSN: 1476-4687. DOI: 10.1038/s41586-019-1750-x. URL: http://dx.doi.org/10.1038/s41586-019-1750-x.

[19c] "Teraelectronvolt emission from the gamma-ray burst GRB 190114C". In: *Nature* 575.7783 (Nov. 2019), pp. 455–458. ISSN: 1476-4687. DOI: 10.1038/s41586-019-1750-x. URL: http://dx.doi.org/10.1038/s41586-019-1750-x.

[TE11] Antonio Torralba and Alexei A. Efros. "Unbiased look at dataset bias". In: *CVPR 2011*. 2011, pp. 1521–1528. DOI: 10.1109/CVPR.2011.5995347.

[TCJ22] Hugo Touvron, Matthieu Cord, and Hervé Jégou. *DeiT III: Revenge of the ViT*. 2022. arXiv: 2204.07118 [cs.CV].

[Tze+14] Eric Tzeng et al. "Deep Domain Confusion: Maximizing for Domain Invariance". In: *CoRR* abs/1412.3474 (2014). arXiv: 1412.3474. URL: http://arxiv.org/abs/1412.3474.

[UVL17] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. *Instance Normalization: The Missing Ingredient for Fast Stylization*. 2017. arXiv: 1607.08022 [cs.CV].

[van+22] Bas H.M. van der Velden et al. "Explainable artificial intelligence (XAI) in deep learning-based medical image analysis". In: *Medical Image Analysis* 79 (2022), p. 102470. ISSN: 1361-8415. DOI: https://doi.org/10.1016/j.media.2022.102470. URL: https://www.sciencedirect.com/science/article/pii/S1361841522001177.

[Vas+17] Ashish Vaswani et al. *Attention Is All You Need*. 2017. arXiv: 1706.03762 [cs.CL].

[VH22] Pablo Villalobos and Anson Ho. *Trends in Training Dataset Sizes*. Accessed: 2024-01-06. 2022. URL: https://epochai.org/blog/trends-in-training-dataset-sizes.

[Vil08] C. Villani. *Optimal Transport: Old and New*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg, 2008. ISBN: 9783540710509. URL: https://books.google.fr/books?id=hV8o5R7%5C_5tkC.

[VB09a] Heinrich J. Völk and Konrad Bernlöhr. "Imaging very high energy gamma-ray telescopes". In: *Experimental Astronomy* 25.1–3 (Mar. 2009), pp. 173–191. ISSN: 1572-9508. DOI: 10.1007/s10686-009-9151-z. URL: http://dx.doi.org/10.1007/s10686-009-9151-z.

[VB09b] Heinrich J. Völk and Konrad Bernlöhr. "Imaging very high energy gamma-ray telescopes". In: *Experimental Astronomy* 25.1–3 (Mar. 2009), pp. 173–191. ISSN: 1572-9508. DOI: 10.1007/s10686-009-9151-z. URL: http://dx.doi.org/10.1007/s10686-009-9151-z.

[Vri+17] Harm de Vries et al. *Modulating early visual processing by language*. 2017. arXiv: 1707.00683 [cs.CV].

[Vui+21] Thomas Vuillaume et al. "Analysis of the Cherenkov Telescope Array first Large-Sized Telescope real data using convolutional neural networks". In: *arXiv preprint arXiv:2108.04130* (2021).

[WH08] S. P. Wakely and D. Horan. "TeVCat: An online catalog for Very High Energy Gamma-Ray Astronomy". In: *International Cosmic Ray Conference* 3 (2008), pp. 1341–1344.

[WD18] Mei Wang and Weihong Deng. "Deep visual domain adaptation: A survey". In: *Neurocomputing* 312 (Oct. 2018), pp. 135–153. ISSN: 0925-2312. DOI: 10.1016/j.neucom.2018.05.083. URL: http://dx.doi.org/10.1016/j.neucom.2018.05.083.

[Wee+89] T. C. Weekes et al. "Observation of TeV Gamma Rays from the Crab Nebula Using the Atmospheric Cerenkov Imaging Technique". In: *apj* 342 (July 1989), p. 379. DOI: 10.1086/167599.

[Wei+21] Guoqiang Wei et al. *MetaAlign Coordinating Domain Alignment and Classification for Unsupervised Domain Adaptation*. 2021. arXiv: 2103.13575 [cs.CV].

[WC18] Garrett Wilson and Diane J. Cook. "Adversarial Transfer Learning". In: *CoRR* abs/1812.02849 (2018). arXiv: 1812.02849. URL: http://arxiv.org/abs/1812.02849.

[WH18] Yuxin Wu and Kaiming He. *Group Normalization*. 2018. arXiv: 1803.08494 [cs.CV].

[Xu+15] Bing Xu et al. *Empirical Evaluation of Rectified Activations in Convolutional Network*. 2015. arXiv: 1505.00853 [cs.LG].

[Xu+21] Tongkun Xu et al. *CDTrans Cross-domain Transformer for Unsupervised Domain Adaptation*. 2021. DOI: 10.48550/ARXIV.2109.06165. URL: https://arxiv.org/abs/2109.06165.

[YY21] Shih-Min Yang and Mei-Chen Yeh. "Unsupervised Multi-Task Domain Adaptation". In: *2020 25th International Conference on Pattern Recognition (ICPR)*. 2021, pp. 1679–1685. DOI: 10.1109/ICPR48806.2021.9412458.

[Yos+14] Jason Yosinski et al. "How transferable are features in deep neural networks?" In: (2014). DOI: 10.48550/ARXIV.1411.1792. URL: https://arxiv.org/abs/1411.1792.

[Yu+20] Tianhe Yu et al. *Gradient Surgery for Multi-Task Learning*. 2020. arXiv: 2001.06782 [cs.LG].

[ZLO19] Jing Zhang, Wanqing Li, and Philip Ogunbona. "Unsupervised domain adaptation: A multi-task learning-based method". In: *Knowledge-Based Systems* 186 (2019), p. 104975. ISSN: 0950-7051. DOI: https://doi.org/10.1016/j.knosys.2019.104975. URL: https://www.sciencedirect.com/science/article/pii/S0950705119304010.

[Zha+13] Kun Zhang et al. "Domain Adaptation under Target and Conditional Shift". In: *International Conference on Machine Learning*. 2013. URL: https://api.semanticscholar.org/CorpusID:17069732.

[Zha19] Lei Zhang. "Transfer Adaptation Learning: A Decade Survey". In: *CoRR* abs/1903.04687 (2019). arXiv: 1903.04687. URL: http://arxiv.org/abs/1903.04687.

[ZY17] Yu Zhang and Qiang Yang. *A Survey on Multi-Task Learning*. 2017. DOI: 10.48550/ARXIV.1707.08114. URL: https://arxiv.org/abs/1707.08114.

[Zha+20] Sicheng Zhao et al. *A Review of Single-Source Deep Unsupervised Visual Domain Adaptation*. 2020. DOI: 10.48550/ARXIV.2009.00155. URL: https://arxiv.org/abs/2009.00155.

[Zhe+22] Wanfeng Zheng et al. *ITTR Unpaired Image-to-Image Translation with Transformers*. 2022. DOI: 10.48550/ARXIV.2203.16015. URL: https://arxiv.org/abs/2203.16015.

[Zhu+20] Jun-Yan Zhu et al. *Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks*. 2020. arXiv: 1703.10593 [cs.CV].

[Zhu+19] Fuzhen Zhuang et al. *A Comprehensive Survey on Transfer Learning*. 2019. DOI: 10.48550/ARXIV.1911.02685. URL: https://arxiv.org/abs/1911.02685.