Machine learning for fetal ultrasound image analysis using privileged information

Jules Bonnard

THÈSE DE DOCTORAT

Sorbonne Université – Spécialité doctorale "Informatique"

Présentée par

Jules BONNARD

Machine learning for fetal ultrasound image analysis using privileged information

---

Thèse soutenue le 7 mai 2024

devant le jury composé de

Pr. Isabelle Bloch	Sorbonne Université, LIP6	Présidente du jury
Pr. Céline Hudelot	CentraleSupélec, MICS	Rapportrice
Dr. Stefan Duffner	INSA Lyon, LIRIS	Rapporteur
Pr. David Picard	Ecole des Ponts ParisTech, LIGM	Examinateur
Dr. Bastien Rance	Université Paris Cité, INSERM	Examinateur
Pr. Bruno Gas	Sorbonne Université, ISIR	Examinateur
Dr. Ferdinand Dhombres	Sorbonne Université, INSERM	Encadrant
Dr. Arnaud Dapogny	Datakalab	Encadrant
Dr. Kévin Bailly	Sorbonne Université, ISIR	Directeur

Contents

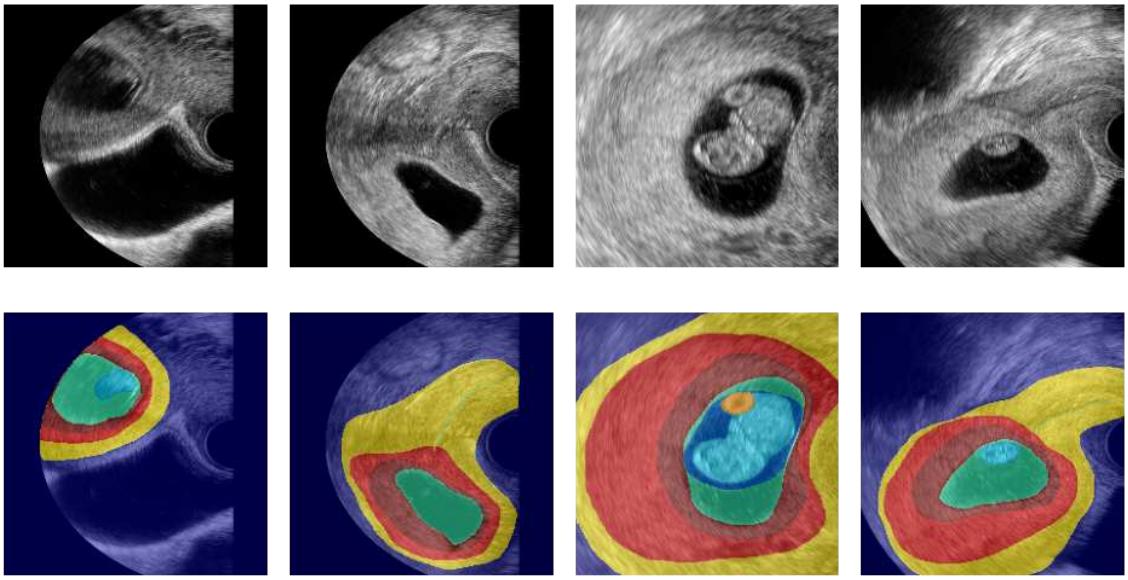
Chapter 1
Introduction

1.1 Context overview

Following my graduation in 2020 from the DAC Master's course in the Sorbonne University, which included an enriching experience as a data science intern at Reminiz, I decided to embark upon the thrilling yet tortuous journey of a PhD. My primary aim was to continue learning and doing so within an academic environment allowed me to work closely with highly experienced people who are experts in their fields and close to cutting-edge research. Another important advantage of working as a PhD student was the possibility to continue teaching, as I had been giving maths grinds to high school or university students since I was 17 years old. Finally, I also wanted to work in a challenging domain of machine learning where I would feel my work had an impact. I was therefore delighted to be offered to pursue a PhD thesis as part of the SUOG European project at the ISIR laboratory in Sorbonne Université, with Kévin Bailly, Ferdinand Dhombres, and Arnaud Dapogny whom I had as a teacher during my master's. This gave me the opportunity to work on groundbreaking developments in the domain of medical imaging, (with all the difficulties a truly innovative field entails...) to progress the practices of obstetrics and gynecology. It also allowed me to work closely with Gauthier Tallec and Edouard Yvinec, both PhD students who shared the same supervisors as me whose collaboration helped me for my personal work but also led to two publications out of the scope of this thesis. We will now present the research domain, the SUOG project and its challenges, and then deliver our chosen research directions.

1.2 Introduction to the SUOG project

Obstetrics and Gynecology (OB/GYN) denote the medical specialties that concentrate on the female reproductive system. Whether for pregnancies or breast cancer screening, medical imaging is essential in this area of medicine to correctly identify anomalies early enough to allow treatment. 3D and 4D ultrasound (US) scans, magnetic resonance imaging (MRI), which use strong magnetic fields and radio waves to create pictures of the anatomy and CT scans which use X-ray can be used in certain specific cases to produce the images for obstetrics and serve as a support for the expert to deliver a diagnosis. Nevertheless, 2D ultrasound scanning is still the primary imaging mechanism



Longitudinal view of the bladder | Longitudinal view of the cervix | Magnified view of the gestational sac | Longitudinal view of the uterus

Figure 1.1: Examples of images from the SUOG collection are shown in the first row and their segmentation maps are shown on the second row. The scan plane annotations corresponding to these examples are reported in this figure. Best viewed in colour.

when it comes to OB/GYN because of its safety and lower cost. Also, ultrasound scans are often preferred over MRI scans because they offer much faster acquisitions that however come at the cost of noisy images. These images are captured and stored in real time by an ultrasound operator.

However, ultrasound screening for OB/GYN is a complex task because of a large number of disorders (>1K) and an even larger number of signs or findings (>10K) that point to these disorders. This problem has become critical due to the number of pregnancy anomalies with 130K cases of congenital anomalies (e.g. structural or functional anomalies occurring during intrauterine life) and 50K cases of ectopic pregnancies (e.g. when the fertilized egg implants and grows outside of the main cavity of the uterus) per year in Europe. Moreover, it is exacerbated by the fact that there is only a limited number of ultrasound screening experts, especially in a field that bears significant medical responsibility. It is therefore evident that effective assistance would help improve pregnancy scans. The problematics for ultrasound screening in obstetrics and gynecology are therefore two-fold: (1) to collect and help filter all relevant images during the scan, and (2) to help analyze these images and achieve diagnosis.

My thesis is part of the SUOG (standing for Smart Ultrasound in Obstetrics and Gynecology) European project that aims to create an intelligent ultrasound assistant built to provide real-time support for the sonographer during the scanning process, and identify

the next relevant ultrasound acquisition or make the correct diagnosis. One of the starting points of this project was the creation of a knowledge base established by international experts from 9 fetal medicine centers in Europe which gathers hundreds of nodes, all clustered in three categories: disorders, findings and technical elements. Initially, the ultrasound assistant would use a rule-based method based on the semantic information as well as the implications provided by this ontology to determine which acquisition would be relevant. This would provide the non-expert sonographer with all the necessary images if ever an expert verification of the diagnosis was needed. To improve on this method, two tasks have been identified: scan plane recognition and image retrieval. Both these tasks are essential for an efficient ultrasound screening as it can guide the sonographer towards the next relevant ultrasound image acquisition. To deal with those tasks, as a complement to the SUOG ontology, thousands of images have been richly annotated by experts with labels extracted from this knowledge base. In particular, for scan plane recognition, there are 18 different views. For a classification task, they were merged into 8 classes, namely the longitudinal view of the uterus, oblique views of the uterus, the longitudinal view of the cervix, the longitudinal view of the bladder, the transverse view of the uterus, the interstitial portion view of the Fallopian tube, the longitudinal and transverse views of the adnexa and the ovary, and magnified views of the gestational sac. Furthermore, these 18 views all stem from 5 *meta classes*. Figure 1.2 illustrates this sub-graph extracted from the SUOG ontology. Teams from GE Healthcare also segmented 294 images into 10 zones that do not overlap: the amniotic sac, the embryo, the gestational sac, the midline echo, the ovary, the uterus borders, the yolk sac, the cervix external ostium, the endometrium and the trophoblast. Examples of these images along with their pixel-wise segmentation labels can be found in Figure 1.1. Although The SUOG project offers a limited amount of ultrasound images (a few thousand), they have been richly annotated in terms of pixel-wise segmentation and ontology entities. Making use of these annotations which include strong semantic information in order to yield interesting results with relatively little data is therefore an exciting challenge. For instance, as a part of the SUOG project and in order to fully use the structural information from the SUOG ontology, El Ghosh *et al.* worked on a graph-based similarity measure named *SimSUOG*. This work leverages symbolic AI to compute distances between pregnancy ultrasound images. It is illustrated in Figure 1.3.

Tasks such as scan plane recognition or image retrieval for medical imaging have already been dealt with using computer vision techniques [32] with relative success. We thus decide to tackle these problems using deep learning based approaches. From a machine learning standpoint, these tasks involve several challenges. First (**challenge 1**), most of the successful deep learning computer vision models (e.g. for tasks such as object recognition, object detection or face recognition) rely on large corpses (up to several millions) of annotated training examples, which is seldom the case when considering OB/GYN ultrasound imaging tasks (in particular, SUOG only offers a few thousand annotated images) and which also include noisy images. Second, (**challenge 2**), how to

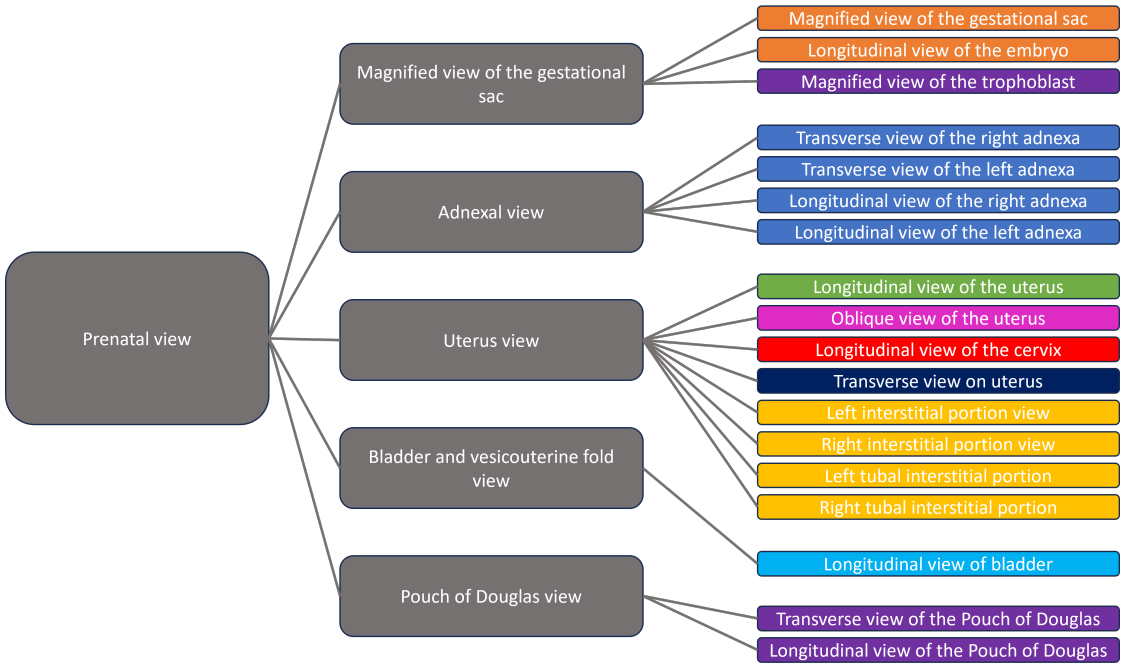


Figure 1.2: An overview of the scan plane annotations. In this figure we illustrate a sub-graph of the SUOG ontology. In particular, the 18 different scan plane annotations are depicted as the leaf entities of this sub-graph, and the *meta-views* at the second level ontology entities. Another way to group the different scan plane annotations was used in chapter 2, and is illustrated here with the 9 colours each representing a different group of views.

effectively leverage the additional data provided by the SUOG project (namely the rich structured annotations extracted from the SUOG ontology and the spatial information in the segmentation maps) remains to be determined. In a nutshell, a classic deep neural network would not be able to perform very well on medical ultrasound images because of the lack of a large-scale annotated training set, and would not be able to leverage the additional information available to enrich this dataset.

1.3 Research Directions

To mitigate the difficulties caused by the relatively small dataset size (**challenge 1**), we consider leveraging spatial prior information to guide the learning. This idea stems from the theory that certain areas of the input image may provide more pertinent information for making specific diagnostics. For instance, if we want to predict the scan plane in an early pregnancy ultrasound image, it is clear that the model will be more effective if it is able to precisely localize the gestational sac and the uterus in some way. To do so, we investigate forcing the model's attribution to resemble prior information heatmaps,

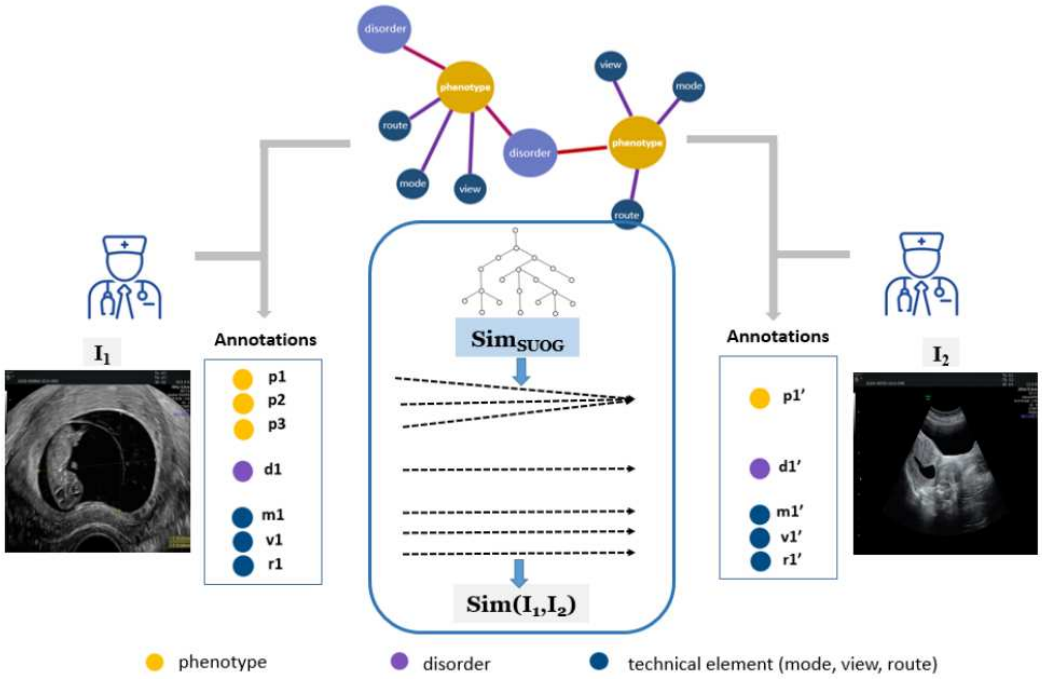


Figure 1.3: An illustration of the semantic-based approach for the similarity of ultrasound images proposed by El Ghosh *et al.*. Whereas classic image similarity measures would consider distances by comparing the annotations, SimSUOG first builds a similarity measure between all nodes of the SUOG ontology by combining the semantic distance between two entities and the Information Context of their Lowest Common Ancestor. This allows SimSUOG to build a sound measure with strong semantic context between two images annotated using entities extracted from the SUOG ontology.

thus teaching the model where the important information is. As the model would learn to concentrate and identify the important areas of the input image during training, the spatial prior would not be needed during inference.

The second track we explore to improve the predictive capacities of a deep neural network (thus adressing **challenge 1** and **challenge 2** at the same time) by exploiting the strong semantic information that can be found in the structured annotations. As mentioned previously, the SUOG dataset delivers rich annotations for each images and an ontology from which the annotations are extracted. This work builds upon the idea that all classes are not equally different. For instance, an African crocodile is closer semantically speaking to an American alligator than it is to a husky. In the same vein, the *left interstitial portion view* is closer to the *right interstitial portion view* than it is to the *Pouch of Douglas view*. The idea would be for the model to predict or separate different levels of hierarchical annotations using several embeddings. This would enable the model to build a sounder latent embedding space and therefore better optimize the inter-class

distances. Another way to do so is to directly integrate the semantic information through natural language guidance, through rich captioning or meta language guidance. The idea is to take advantage of the recent progress in the domain of Natural Language Processing (NLP) to improve the visual model's capacities. This may allow the model to have a more accurate representation of the input images through a sharper textual representation.

1.4 Related Work

1.4.1 Deep Learning for Obstetrics and Gynecology

In recent years, computer vision and image analysis have become central to medical imaging, for tasks such as anomaly or disorder classification, image segmentation or object detection. In particular, it is important for OB/GYN to better analyze pregnancy ultrasound scans or breast radiographies. In order to achieve interesting results, research has turned to artificial intelligence (AI) to analyze these images automatically.

At the beginning of my thesis, we worked on a systematic review of AI techniques for OB/GYN [32], in order to investigate and evaluate the methods, data and protocols. We collected all OB/GYN papers published between 2000 and 2020 that mentioned AI and easily divided them into two subcategories: those using symbolic AI, and those using statistical AI. Symbolic AI methods include formal logic, knowledge representation, and rule-based reasoning. These methods are usually explainable, do not need large amounts of data but rely on human supervision and design, which have made them popular in medical domains. Statistical AI, or more precisely machine learning methods, usually aim to optimize algorithms such as Artificial Neural Networks (ANN) and usually necessitate large amounts of data. Results and predictions given by machine learning methods are also more difficult to explain, as the mechanism inside the ANN is a form of "black box". We found that AI is still very seldom used to deal with image or video data (only 12% of the AI methods in OB/GYN) for OB/GYN tasks, and that all the machine learning methods that are employed in that field are in reality at the "proof of concept" or "proof of feasibility" stage. This showcases a lack of novelty in the area, and probably a number of constraints to the use of machine learning due to the sensitive nature of the OB/GYN examinations. In addition, as machine learning typically relies on large datasets to effectively learn a predictive model, this is not helped by the fact that such datasets are rarely available in OB/GYN. Finally, another reason may be due to the fact that most of the recent advances in machine learning are typically published in computer science journals and not medical reviews.

However, deep learning methods, which have become common for most computer vision tasks such as object detection, object recognition or image segmentation for example have slowly started to make their way into the medical imaging fields. Burgos-Artizzu *et al.* [18] evaluate the impact of deep learning methods such as Convolutional

Neural Networks (CNNs) on a relatively large medical imaging dataset (12K images) and show that these types of models can obtain results similar to humans for maternal-fetal ultrasound classification. In a similar manner, Qu *et al.* [70] propose to use transfer learning to deal with limited amounts of data, by using a network that is pretrained on a larger dataset and fine-tuning it on the available data. They compare these CNN-based methods to classical machine learning methods (e.g. SVMs and clustering) and show that using transfer learning limits overfitting and offers the most promising results.

In order to delve deeper into the ultrasound image analysis, we now present works that deal, on the one hand, with scan plane recognition, a central task in the SUOG project, and, on the other hand, with breast cancer detection, an ultrasound imaging classification task that has gained a lot of traction recently.

1.4.2 Scan Plane Recognition

One major task in medical image analysis is scan plane recognition, which aims at classifying an image into a certain category of standardised ultrasound planes. This task was originally carried out through manual screening by doctors. However, a high number of errors have been observed because of the large number of non-experts ultrasound operators. This task is crucial because it can guide a non-expert sonographer towards the needed scan plane ultrasound acquisitions, and thereby averting the need for additional scans due to errors. A significant volume of research concentrates on this task as a classification task and deal with it rather simply using a CNN and transfer learning [118, 57, 70, 18]. These works differ from each other through small changes in the evaluation protocol or database used to train and test. As it is pointed out by Fiorentino *et al.* [40], scan plane recognition has very few publicly available datasets, and there are various tasks even within the scan plane recognition. In particular, the planes that are typically evaluated are the fetal abdomen scan planes (FASP), the fetal brain scan plane (FBSP) and the femur standard planes (FFESP). Visual examples of these are highlighted in Figure 1.4. Therefore a lot of methods propose small adjustments to improve their prediction scores for a specific dataset, without necessarily comparing to other methods. For instance, Montero *et al.* [63] make use of Generative Adversarial Networks (GAN) to generate synthetic data and therefore learn a classification model with a larger training dataset. Chen *et al.* [23] aim to identify scan planes from ultrasound images and videos automatically on a self-collected dataset. They leverage a DenseNet architecture and mix high-level features from the shallow layers with low-level features from the deeper layers to improve their predictive capacity. Sundaresan *et al.* [93] use a Fully Convolutional Network (FCN) to simultaneously locate the center of the fetal heart and classify cardiac views. Similarly, Baumgartner *et al.* [10] introduce Sononet, a convolutional neural network (CNN) model designed for the recognition and localization of fetal standard scan planes. Their approach involves thorough preprocessing of a

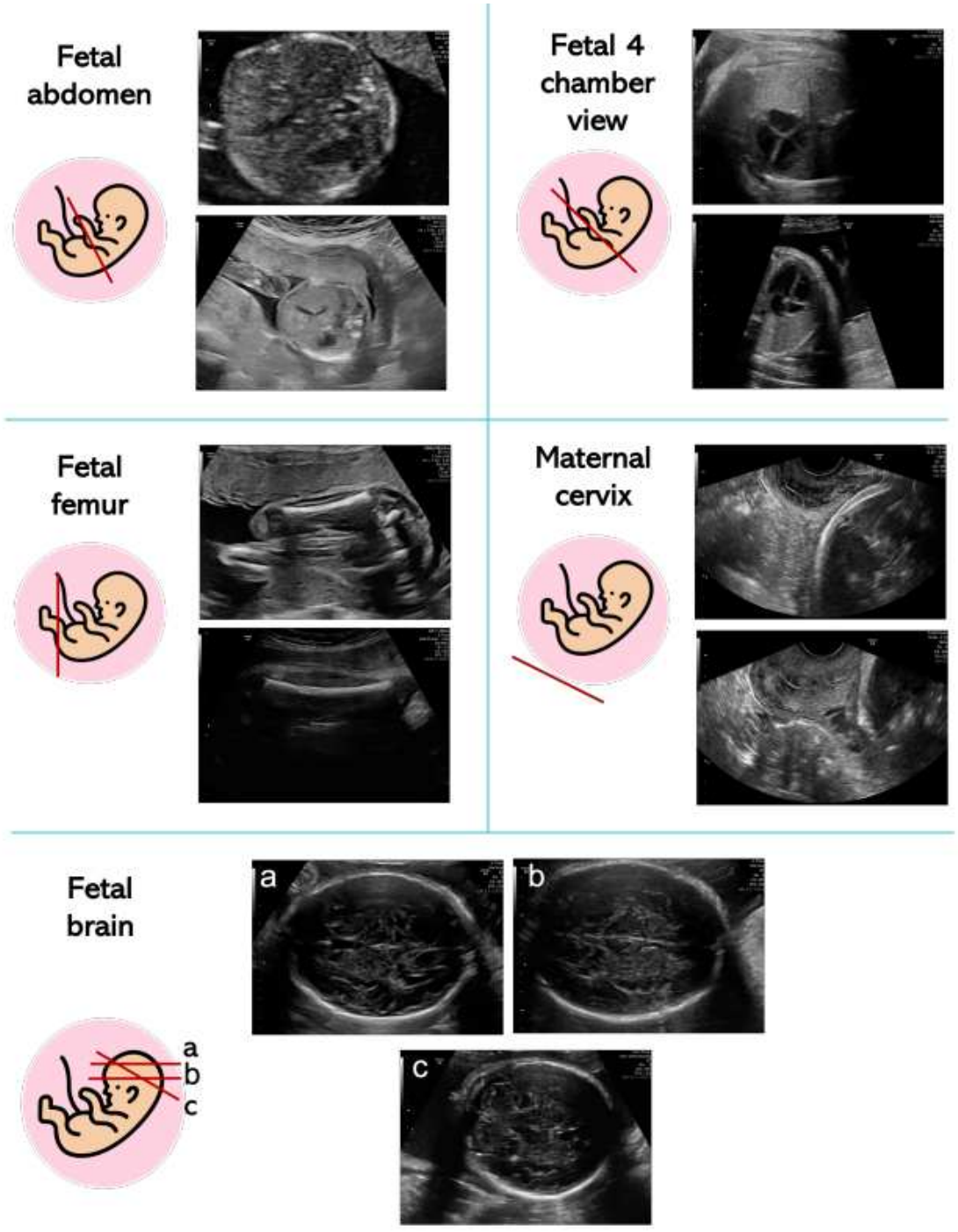


Figure 1.4: Visual examples of the most common fetal scan planes. Illustration taken from Fiorentino *et al.* [40].

substantial dataset containing over 30,000 2D ultrasound images, employing weakly-supervised learning to achieve real-time localization. While the previous work makes use of a localization mechanism to understand the the structure of the image and therefore improve the scan plane recognition predictions, SonoNet uses the scan plane recognition predictions to localize certain important structures using a backward pass.

Comparably, research has also investigated using attention mechanisms to better leverage the spatial information. Schlemper *et al.* [78] build on Sononet [10] to which they add attention gates for different ultrasound image analysis tasks such as segmentation or scan plane recognition. Cai *et al.* [20] use sonographer gaze tracking data to generate relevant attention maps. This offers valuable information to the model in order to better predict the scan plane.

Other methods leverage an auxiliary task to improve performance on the main task. Xu *et al.* [110] address scan plane recognition and landmark detection for abdominal ultrasound images in a multi-task framework. This way, the scan plane recognition task benefits from the latent information learnt from the landmark detection. Similarly, Zhao *et al.* [122] incorporates knowledge of fetal anatomy into scan plane recognition using a knowledge graph. The knowledge graph facilitates the creation of a co-occurrence relationship graph, connecting spatial features extracted by a detection module. This enhances the model's ability to reason about anatomical structures, thereby improving the predictive capabilities of the convolutional model. However, these methods necessitate additional information during inference.

1.4.3 Breast cancer detection

Another OB/GYN task that has gathered a lot of attention in automatic image analysis is breast cancer detection. Several works have shown an interest in AI methods for breast cancer detection. Le *et al.* [53] first showed that support vector machines (SVM) improved the predictive results when compared to rule-based methods, and Zheng *et al.* [125] apply an AdaBoost algorithm to the binary classification task and demonstrates its effectiveness. Nasser et Yusof [64] propose a systematic review of deep learning methods for breast cancer detection. They underline the fact that deep learning methods have obtained promising results using CNNs. For instance, Ha *et al.* [44] build a CNN with residual connections to predict breast cancer on MRI imaging data, and Wu *et al.* [106] train a custom ResNet-based model updated for high-resolution medical images on a large-scale database (1M images) and obtain very promising results. However, such large-scale datasets are seldom available publicly.

Strelcenia et Prakoonwit [90] generate synthetic data with a novel GAN-inspired method to alleviate this issue. Similarly, Tien *et al.* [97] combine variations of GAN (e.g. CycleGAN and DeblurGAN) to improve the quality of the training images. Albarqouni *et al.* [4] identify the problem of having limited access to expert annotations in the medical imaging field. They therefore leverage crowdsourcing to improve their model's capacity

to predict breast cancer malignancy.

Chen *et al.* [24] identify that shear-wave elastography (SWE) serves as important complementary data to the ultrasound images, but are scarce due to the lack of SWE devices in the majority of hospitals. To enhance the predictive capacities of the breast cancer classification network with this additional data, they propose cross-modal and semantic data augmentation simultaneously. More specifically, a modal translator synthesizes SWE images from ultrasound images, while two losses predict the presence of cancer: one on the ultrasound scans, and one on pairs of US-SWE images. Lee *et al.* [54] also use additional auxiliary images to improve the prediction task. They try to deal with cancer risk prediction using a prior image as additional information. To do so, both the present scan and the prior scan are passed through a shared CNN backbone for feature extraction. They then leverage a decoder based on a transformer architecture to fuse the relevant information from both inputs and improve the model's capacity to evaluate the breast cancer risk. Shareef *et al.* [82] also use a transformer architecture to capture local context in the input image. However, they deal with breast cancer detection in a multitask manner, both solving a classification task and a cancer segmentation task. They both use a CNN to extract hierarchical and local patterns and a SWIN transformer to leverage long-range dependencies. They show that learning to predict a segmentation map for the cancerous areas of the image greatly improves the classification predictions.

The take-home message of this section is that deep learning methods have recently been applied to OB/GYN imaging tasks such as scan plane recognition or breast cancer detection for instance. However, results have not been as convincing as in other computer vision domains because of the difficulty to access large-scale annotated datasets **(challenge 1)**. Since this data is not available in our research context, we aim to make use of available additional information to improve the deep learning model's predictive capacities and therefore reflect on different ways to guide the model through this data **(challenge 2)**. In particular, we leverage spatial prior information (i.e. pixel-wise segmentation maps, see Chapter 2 of this thesis) and structured annotations extracted from a hierarchical class ontology (i.e. meta-annotations, see Chapter 3) in order to add spatial or semantic context to our input samples.

1.5 Summary of contributions

In an attempt to leverage the power of deep learning methods in an OB/GYN context, we decide to integrate additional information (whether it be spatial information or structured annotations) to alleviate the problem of small datasets.

We first investigate the interest of adding a spatial prior to the scan plane recognition classification task. To do so, we guide the deep learning model's focus towards certain specific salient zones in order to improve its predictive capacities. We then aim to improve the model's capacity to separate scan plane recognition classes using structured

annotations. We train the model in order to be able to separate different levels of classes and therefore create better latent representations for the ultrasound images.

To sum it up, our contributions in this thesis are three-fold:

- In order to integrate spatial prior information, we introduce the Prior-Guided Attribution method (PGA). The aim of this method is to force the model to concentrate on certain specific regions of the input image that might be more discriminative for the model's predictions. We therefore implement a Privileged Attribution Loss (PAL) that maximizes the cross-correlation between the attribution map and a prior information heatmap where the salient areas are highlighted. We also propose a Prior Allocation Strategy, that enables the model to incorporate multiple spatial priors while still giving some liberty for the model to look elsewhere.

- To improve the capacities of a computer vision model for an image similarity task, we introduce a novel way to integrate structured annotations in the learning process. We introduce *meta-embeddings* that are pushed to encode hierarchical semantic information extracted from a class ontology by the Semantic Abstraction Loss (SAL). This loss is built as a weighted average of DML losses. We also introduce new ways to integrate this information through natural language. First we propose to input the hierarchical annotations as rich captioning. Second, we build on the work from Roth *et al.* [75] and guide the *meta-embeddings* with natural language, introducing Ontology Language Guidance (OLG).

- From an experimental standpoint, we validate both these methods on several use cases. For the classification task, we demonstrate that the PGA method is generic by conducting experiments on facial expression recognition, breast cancer detection and scan plane recognition. For the visual similarity learning task, we conduct experiments on an open-set birds classification dataset and validate its interest for OB/GYN tasks on scan plane recognition with the SUOG dataset.

1.6 List of Publications

The work presented in this thesis led to the following preprints and publications:

1.6.1 International Journals

◊ Dhombres, F., Bonnard, J., Bailly, K., Maurice, P., Papageorghiou, A. T., & Jouannic, J. M. (2022). Contributions of artificial intelligence reported in obstetrics and gynecology journals: systematic review. Journal of medical Internet research (JMIR), 24(4), e35465. [32]

◊ Bonnard, J., Dapogny, A., Zsamboki, R., De Braud, L., Jurkovic, D., Bailly, K., & Dhombres, F. (2023). Prior-Guided Attribution of Deep Neural Networks for Obstetrics and Gynecology. IEEE Journal of Biomedical and Health Informatics (JBHI). [14]

1.6.2 International Conferences

◊ Bonnard, J., Dapogny, A., Dhombres, F., & Bailly, K. (2022, August). Privileged attribution constrained deep networks for facial expression recognition. In 2022 26th International Conference on Pattern Recognition (ICPR) (pp. 1055-1061). IEEE. [13]

◊ Bonnard, J., Dapogny, A., Dhombres, F., & Bailly, K. Ontology-Guided Learning for Obstetrics and Gynecology. Under review at ICPR 2024

1.6.3 National Conferences

◊ Bonnard, J., Dapogny, A., Dhombres, F., & Bailly, K. Privileged attribution constrained deep networks for facial expression recognition. In 2022 Reconnaissance des Formes, Image, Apprentissage et Perception (RFIAP).

1.6.4 International Workshops

◊ Bonnard J. Numeric AI research : Neural network and Privileged Information. In 2022 SUOG Workshop

1.7 Outline

This thesis is divided into two main chapters where we investigate ways to better deal with OB/GYN imaging challenges. In chapter 2, we first present our work concerning the integration of prior spatial information to improve the model's predictions, and therefore answering **challenge 1**. More precisely, we present the novel Prior-Guided Attribution (PGA) method that guides the CNN-based network's attribution towards carefully-chosen salient areas. In this chapter, we first expose different methods that use prior information guidance or attribution constrained learning. We then delve deeper into the method and next demonstrate the effectiveness of the proposed method on several different tasks and datasets, while also discussing the importance of the prior information selection.

Second, in chapter 3, we introduce our work concerning the integration of structured annotations in metric learning, therefore addressing **challenge 2**. More specifically, we introduce our novel L_{SAL} and OLG methods that incorporate strong semantic information extracted form the annotations through meta-embeddings and language guidance. In this

chapter, we first present the most common deep metric learning baselines and methods that leverage hierarchical annotations in their learning framework. We then present both L_{SAL} and OLG in more detail and finally prove the interest of this method on a publicly-available birds classification DML dataset and the SUOG scan plane recognition task. Finally, we discuss the advantages and limitations of this method before suggesting interesting future research.

Chapter 2
Privileged Attribution Learning to deal with small datasets

2.1 Introduction

Ultrasound scanning is standard-of-care practice in obstetrics and gynecology, and the automatic analysis of these images has become ubiquitous for the development of efficient clinical decision support systems for non-expert operators, especially since the availability of ultrasound experts is insufficient. As it is often the case because of the cost of medical images, the SUOG project has access to a limited number of annotated images **(challenge 1)**, and dealing with tasks such as scan plane recognition with deep learning models can be difficult. However, structured spatial information is available through a few hundred segmentation maps **(challenge 2)**.

In order to alleviate the issues caused by the lack of annotated data, we therefore leverage this spatial prior information to guide the network. We build upon the idea that certain regions of the input image hold greater significance for accurate predictions (i.e. the eyes, eyebrows, mouth and nose are more important for facial expression recognition than the hair or the background). To improve the model's predictive power, we guide the model's *attribution maps*, which correspond to the importance or relevance of the input features with respect to the model's outputs. Specifically, we introduce the novel *Privileged Attribution Loss (PAL)* at train time that maximizes the cross-correlation between these attribution maps and carefully chosen priors, therefore encouraging the model to pay more attention to certain specific areas of the input image. This allows the model to capitalize on additional expert information that is available during training without needing it at test time, which means that the method comes at virtually no cost during inference. To enable the integration of multiple prior information maps at the same time, we propose the *Prior Allocation Strategy (PAS)* that allocates a fraction of channels of the network's attribution to each individual prior, while still leaving the network some freedom to focus on different areas. The proposed method, *Prior Guided Attribution (PGA)*, encompasses both PAL and PAS to guide the network towards any kind of spatial prior. The overall pipeline is illustrated in Figure 2.1. The method presented is generic and can be readily adapted for various computer vision tasks with distinct priors. To evaluate this, we initially implement PGA for Facial Expression Recognition (FER). The reasons for using FER as a proxy task are two-fold: First, the results of deep learning approaches for

FER are usually hindered by data scarcity, as the annotation is expensive compared to other computer vision subdomains. Second, face images can benefit from side annotations such as face bounding boxes and landmarks extracted using off-the-shelf methods. We then extend the application of PGA to two OB/GYN tasks: breast cancer detection (employing a single semantic segmentation map as a prior) and automatic scan plane recognition in early pregnancy ultrasound images, incorporating limited segmentation maps as spatial priors. The results demonstrate promising outcomes when compared to existing baselines. To summarize, the main contributions proposed in this chapter are:

- We introduce a Prior Attribution Loss (PAL) term that seeks to maximize the cross-correlation between the model's attribution and prior information maps, therefore ensuring that the model focuses on specifically chosen areas of the input image.
- We introduce a Prior Allocation Strategy (PAS) that allows the method to incorporate multiple maps of additional prior information while still compromising on the strength of the constraint. Additionally, we discuss and evaluate the choice of the number and granularity of prior heatmaps that are used to guide the network.
- We introduce PGA, a method that guides the model towards salient regions of the input image, built as the association of both PAL and PAS. PGA helps learning with few data for medical imaging, only necessitating additional prior information at train time and not at inference time.
- Experimentally, we show that PGA is generic and improves the predictive power of deep neural networks when applied to different tasks and images types. We then compare the impact of different priors on the method and therefore discuss the choice of said prior maps.

This chapter is divided as follows: in section 2.2 we present state-of-the-art methods for FER and methods that learn with spatial prior guidance and common attribution methods. In Section 2.3 we then present our method to integrate spatial prior information to improve the predictive power of the model. We then demonstrate the interest of this method with experimental results on various tasks and datasets in Section 2.4, and finally provide a discussion on this research and outline future directions for further work regarding the ideas introduced in this chapter in Section 2.5.

2.2 Related Works

In this section, we give an overview of methods that leverage spatially constrained attribution learning. In particular, we first present methods that deal with facial expression recognition then medical imaging tasks with spatial priors, then methods where the attribution is constrained to improve the predictive capacity of the model.

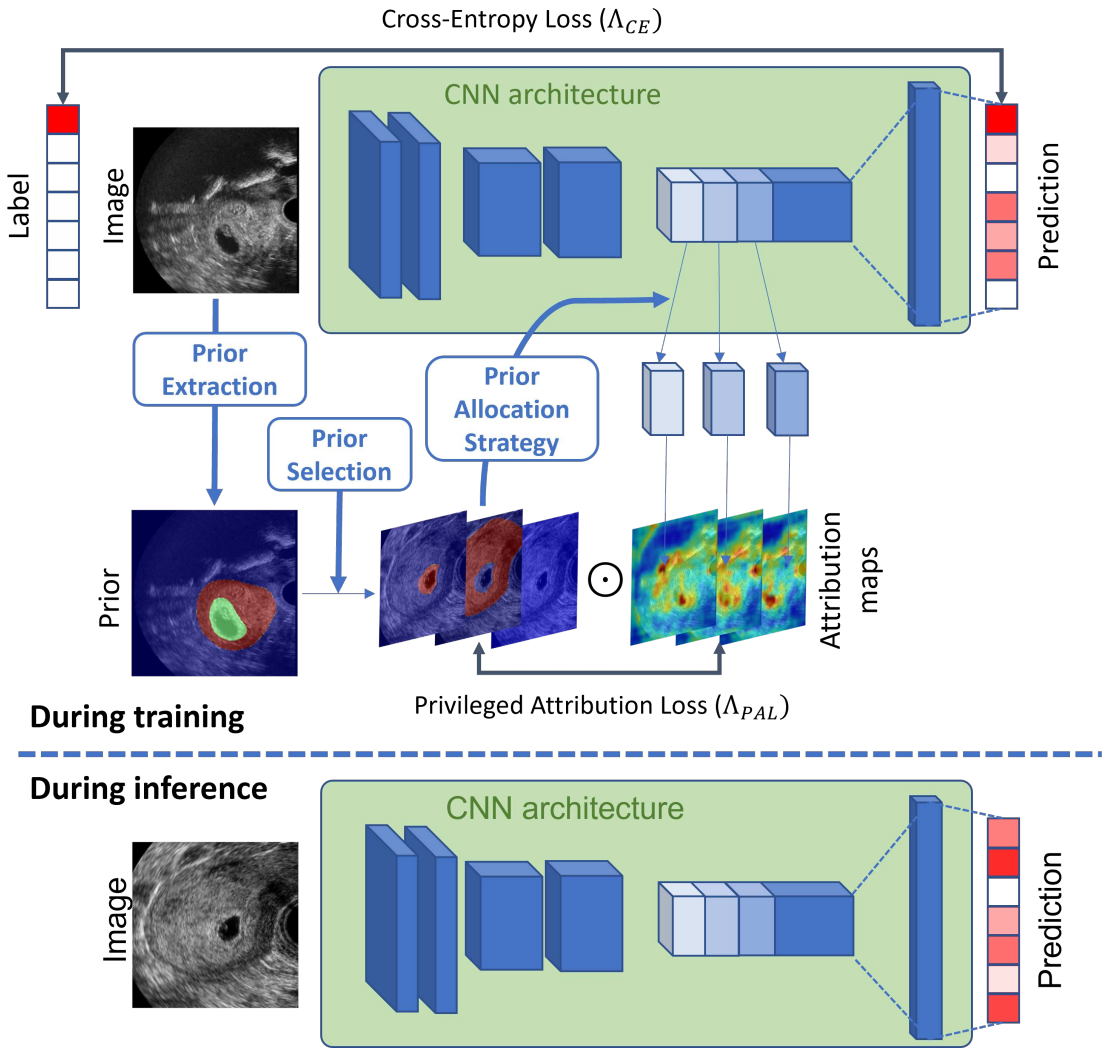


Figure 2.1: Overview of the Prior-Guided Attribution (PGA) pipeline. During training, the model uses the input image, the ground-truth labels and the prior heatmaps. The image is passed through a network, whose predictions are matched with the labels through a cross-entropy loss. The image's attribution is then calculated on a portion of channels of a specific layer, chosen using the Prior Allocation Strategy (PAS). For these channels, the attribution is then constrained to correspond to a certain prior heatmap with the Privileged Attribution Loss (PAL). During inference, only the image is needed for the network to produce its predictions.

2.2.1 Spatial prior guidance learning

Computer vision models learn to analyze structures inside images. This heavily relies on a sound understanding of the spatial information of the input image. Recent works have made use of prior spatial knowledge in order to improve the network's capacity to respond to specific computer vision tasks. We first broadly present face analysis and facial expression recognition methods that are spatially constrained, and then focus on medical imaging tasks that leverage prior spatial knowledge.

2.2.1.1 Constrained Learning for facial expression analysis

The face can hold a lot of semantic information about a person, its personal psychological state or can be decisive in human communication. Automatic face analysis has become a crucial topic in computer vision and is largely used in domains such as security, entertainment or healthcare. Multiple different ways to describe facial expressions or affect have been explored in the literature. One track based on the Facial Action Coding System (FACS) proposed by Ekman et Friesen [35] consists in characterizing facial expressions as a combination of 44 facial muscle activations, referred to as Action Units (AUs), therefore supposedly objective. A number of these AUs are highlighted in Figure 2.2. Another way to describe facial expressions is in a categorical manner. Ekman et Friesen [36] proposed a list of universally recognized basic emotions, namely anger, sadness, happiness, fear, surprise, disgust or neutral. Another important computer vision task for face analysis is facial landmark alignment, which consists in identifying all the facial landmark points (usually 68 or 80 landmarks) which describe the head pose and location. This task is often used as an additional input for feature extraction, used to then deal with facial expressions. In this chapter, we will work with the Facial Expression Recognition (FER) task that aims at predicting the basic emotion from a facial image, as a similar task to scan plane recognition (relatively small annotated datasets, approximately the same number of classes, available spatial prior information).

Dapogny *et al.* [28] have worked on predicting local facial expressions in order to better deal with occluded faces. They use random facial masks create local facial subspaces and then use randomized decisions trees to predict the local emotion. These local emotions predictions are then leveraged to help with occluded facial images. One of the conclusions of their work is that the spatial distribution of the facial features is paramount in the prediction of human emotions. Since then, a lot of recent research has focused on constraining their learning mechanism to leverage spatial prior information to improve the model's results on diverse face analysis tasks.

A first research track explored by several recent works is to guide their face analysis task using a spatial module. Authors in [123] improve their model's robustness and therefore perform better for FER in the wild. First they use a local feature extractor that divides the input features into multiple patches that therefore keep the spatial coherence

of the image and the local facial features, and then mix them to global features in a residual manner. This enables the model to learn salient global and local features, which improves the model's results on FER baselines and occluded or in-the-wild FER datasets.

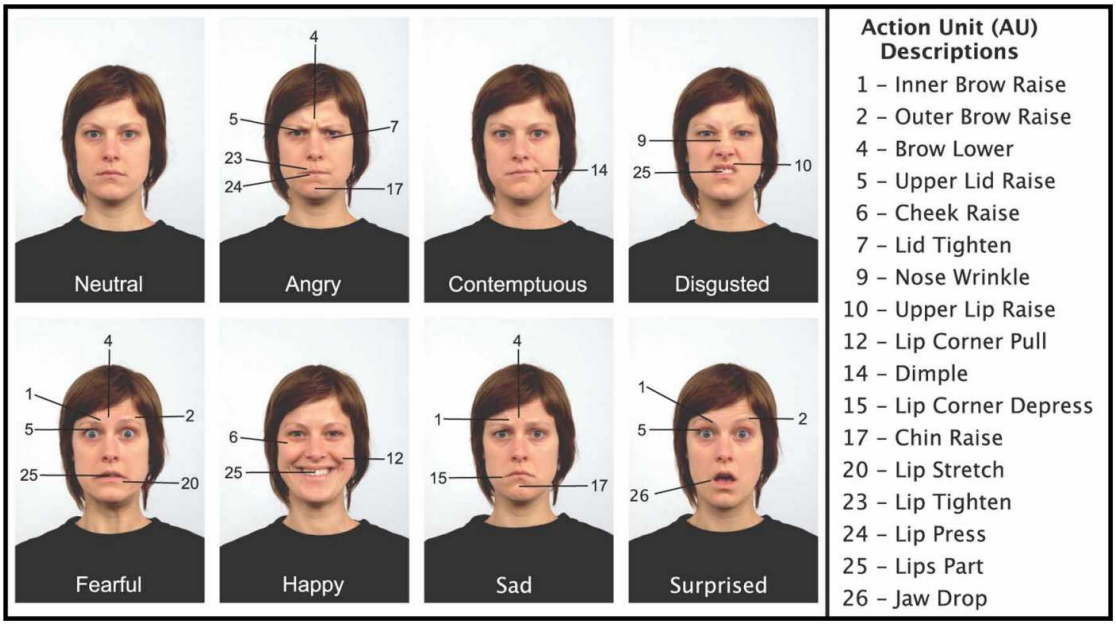


Figure 2.2: Visual examples of the most common action units activation for each of the 8 basic emotions. Illustration taken from Langner *et al.* [52].

Other methods leverage an adjacent task in order to improve the predictions on their main task. For instance, Pu *et al.* [69] use an Action Unit (AU) prediction module to improve their FER model. First a common shared CNN backbone extracts features for both tasks. They then use prior information in the form of an emotion-AU relationship graph that links the facial expressions with the most common AU activations (examples of such AU-Expression relations are depicted in Figure 2.2). In particular, a learned attention mechanism is constrained so that it would match the aforementioned relations. For instance, since the *"lip corner puller"* action unit (which usually represents a smile) is heavily correlated to the *"happiness"* emotion, the learned attentional coefficient is supposed to match the prior value. They therefore leverage prior knowledge to improve the FER predictions. Similarly, Shao *et al.* [81] aim at predicting facial action units by first predicting face alignment points. The prediction of the 68 face alignment points first give spatial context to the backbone that can be beneficial in order to predict action units. However, they go further to make use of both tasks. They use prior knowledge linking each action unit with their characteristic alignment points. They refine attention maps using heatmaps created from said facial landmarks, and use these attention maps to spatially guide the action unit detector towards the most discriminative areas of the input image. This work both leverages an adjacent task to integrate prior knowledge, but also constrains the network spatially to improve its predictive capacity.

Other methods directly take advantage of additional information to make sure their model focuses on the most salient areas of the input image instead of modelling it with an intermediate network. For instance, Jacob et Stenger [49] use heatmaps in which the areas of each action units are highlighted as prior spatial information for facial action unit detection. They propose to constrain the attention maps constructed by a Transformer decoder to resemble these heatmaps. These attention maps are created to accentuate the most relevant areas of an input image with respect to another (cross-attention) or the same (self-attention) input. This method therefore enables the network to mainly focus on the the areas that are most discriminative for the task at hand, in particular to predict the presence of a specific action unit. While this prior information can be quite costly because it necessitates a heatmap for each action unit for each training image, it has the advantage of needing no additional information during inference because the model has learnt to identify the important zones during training.

These works have demonstrated that leveraging spatial information to guide the learning has a positive impact on facial expression tasks. For medical imaging, we could argue that the presence and the position of certain structures can be crucial to analyze the image. We therefore present methods that spatially constrain the learning for medical images.

2.2.1.2 Spatially guided learning for medical imaging tasks

In order to remain closer to the main task in the framework of the SUOG project, we now focus on methods that constrain their learning for medical imaging tasks.

One way to integrate spatial information is to use it for preprocessing. For instance, Zeng *et al.* [121] use spatial priors for a multi-modal registration workflow. This task aims to align several images into a shared coordinate system. They propose an image segmentation network aimed at identifying the different liver structures and then predict the image alignment. However, they incorporate prior knowledge to improve the model's prediction capacity by using an initial rib cage segmentation to generate an initial alignment upon which they build their model.

Another way to do so is to guide the main task using an adjacent task. For example, Men *et al.* [61] propose a *Multimodal GuideNet* that aims to predict the probe motion from an ultrasound scan. The idea behind this work is to provide guidance for the less experienced sonographer to improve their scanning skills. To do so, they jointly learn to predict the gaze motion. The gaze trajectory prediction module therefore provides the model with crucial spatial information in order to predict the probe motion correctly. Similarly, Wang *et al.* [101] integrates gaze tracking information to improve a abdomen segmentation model. They argue that segmentation annotations is very costly in terms of time and effort and necessitates expert knowledge. They propose to weaken the demand for these high-cost labels by using the more convenient, but less precise gaze-tracking heatmaps. More specifically, they highlight the most relevant areas of the input image

by computing the cross-attention between both the input image and the gaze-tracking heatmap. This provides the model with human cognitive expert information since it follows the attention of the highly professional medical experts.

Other methods make use of the nature of attention methods, aimed at discovering strong semantic relations between local spatial features. For instance, Cai *et al.* [19] propose to deal with vulvovaginal candidiasis detection (e.g. a specific disorder classification task) using attention guidance. They observe that the disorder is usually hard for the model to pick up automatically because of the small size of the symptoms and the lack of labeled data. Therefore, they use strong spatial information to guide their model's attention. First, they use an image encoder backbone pre-trained for the detection of the main symptom (candida detection). They then introduce an attention-based module meant to recognize strong relevant information for the disorder classification task. To do so, they apply a cross-attention between low-level fine-grained feature maps and high-level coarse feature maps. The idea behind that is that different structures that identify the candidiasis disorder are visible at different scales, therefore the model performs better when it is able to use both coarse and fine-grained spatial information. In the context of chest X-ray segmentation, Miao *et al.* [62] also use a spatial prior to guide a self-supervised vision transformer. The authors use prior knowledge segmentation masks that highlight certain relevant structures of the input image and introduce a loss designed to make different attention maps correspond to the aforementioned spatial priors. They demonstrate that this guidance not only improves the segmentation results but also yields attention maps that are more interpretable.

For lung and heart 3D segmentation, Xie *et al.* [108] also introduce a prior-guided model that learns region-wise local consistency in the latent feature space. Built upon the philosophy of BYOL [43], a local consistency loss forces the voxel areas to correspond for different transformations of the same image. This provides the model with semantic information and context about the different structures. They prove that this spatial information significantly improves the segmentation results on multiple datasets.

These methods, whether they guide the model's attention or use an adjacent task to incorporate spatial expert information, do not ensure that the model makes use of this specific information. Some research has aimed at directly guiding the model's relevance or *attribution*.

2.2.2 Attribution methods

Attribution methods aim at evaluating the relevance of specific input features on the output. In a computer vision context, a model's attribution directly computes how important each pixel of the input image is for the final prediction.

Two large families of attribution methods have emerged, namely the occlusion-based methods and the gradient-based methods. Occlusion-based methods [119] consist

in computing the attribution of a certain patch of pixels by comparing the outputs given by the input feature with the outputs given by the image where this patch of pixels is occluded. However, the occlusion-based methods usually involve prohibitive computational costs on images.

In order to reduce the runtime, gradient-based methods were introduced. First, Simonyan *et al.* [86] introduced image-specific class saliency maps. These are computed as the derivative of the wanted class with respect to the input image. This straightforward method allows for a good evaluation of the input feature's impact at the cost a single backward pass. Shrikumar *et al.* [84] developed on this idea a proposed a technique to sharpen these attribution maps called *Gradient*Input*. They obtain these maps by taking the signed partial derivative of the output with respect to the input and multiplying it by the input itself. The idea is that the gradient represents how important a certain feature is, and the input represents how strongly it is expressed in the final prediction. Chen *et al.* [26] even proved that this was the exact relevance of the input features for any ReLU-based convolutional models. In a similar manner, Selvaraju *et al.* [80] presented the Grad-CAM method. The method computes class-specific attribution heatmaps by multiplying each channel of the feature map given by the last convolutional layer by its "importance" (e.g. the mean -over that specific channel- of the gradient of the chosen output class with respect to the last convolutional layer's output). Examples of such attribution maps are illustrated in Figure 2.3. Finally, another popular gradient-based method is Integrated Gradients, introduced by Sundararajan *et al.* [92]. They build this attribution as the path integral of the gradients along a straight line between a baseline input (usually a black image) and the actual image. In practice, they compute a certain number of interpolations (between 20 and 1000) between the baseline image and the actual input considered. Then, the values of their gradients are computed after a forward pass and are summed together to give the attribution values. This method is theoretically better but is not derivable, as it is a built as a Riemann approximation.

These attribution methods have become very popular for explainability reasons as they offer visual cues to understand "black-box" models' predictions. However, recent research has made use of these attribution methods to guide the network's focus towards certain areas of the input image. This allows for a certain regularization of the model and alleviates the problem of data scarcity.

2.2.2.1 Attribution guided learning

Several works have investigated constraining their network's attribution in order to directly guide the learning towards important input features. For instance, Du *et al.* [34] constrained the occlusion-based attribution of their Natural Language Processing (NLP) network by reducing the values of words deemed non-important by expert annotated clauses in the input sentence. However, due to the fact that these methods are not derivable and their computational inefficiency, most works have focused on gradient-

Original Image | Grad-CAM "Dog" | Grad-CAM "Cat"

Figure 2.3: Examples of visualizations for the Grad-CAM method. The first image shows the picture of a cat and a dog. The other two images show the attribution maps for both predictions dog and cat, which both highlight the right areas of the input image. The figure is taken from Selvaraju *et al.* [80]. Best viewed in colour.

based methods. Erion *et al.* [37] constrain their attribution so that it would have satisfying properties such as sparsity or smoothness, but did not explicitly constrain it using prior information to guide it towards crucial input features. Conversely, in an NLP context, Liu et Avci [58] added an L2 norm between their attribution (computed with Integrated Gradients) and an external prior that evaluated if words were toxic or not. For a computer vision task, Ross *et al.* [73] investigated the idea to be "right for the right reasons". They tried to improve the model's explanations while still keeping the same level of predictive performance. To do so, they penalize gradients of non-relevant features using a prior information binary mask as additional annotations. Similarly, Fel *et al.* [39] integrate prior heatmaps of human-chosen features as a kind of "correct explanation" for object recognition predictions. Both these works guide the network towards input features that are considered important beforehand using prior information. Ismail *et al.* [48] introduce "saliency-guided training" in order to get rid of small and noisy gradients. More specifically, they use a binary mask to get rid of pixels with small gradients, and then minimize the KL-divergence between the output given by the input image and the masked input image. This prevents the model from concentrating on noisy gradients that could pollute the final attribution while maintaining the predictive performance. However, these four methods constrain the learning in order to improve the model's explanations, but not to increase the predictive results. However, Bertoin *et al.* [12] argued that, in order to enhance deep reinforcement learning models, they should be able to identify the most important input features to focus on them and ignore the others. To do so, the model is trained to predict its saliency maps so that it learns to "look where it looks". This circumvents the risk that the model would make wrong decisions because of distracting visual features and therefore improves the models' policies.

In this section, we presented works that highlight the advantages gained by constraining the model spatially on facial expression and medical imaging tasks. To ensure that the model concentrates on the selected prior areas of the input image, we decide to guide the network's attribution, instead of constraining cross-attention or self-attention modules, or leveraging an adjacent task to improve results on the main task. Unlike most methods that constrain the attribution in order to improve the network's explainability, we use these spatial constraints to improve the prediction results. In particular, our work guides the model's attribution towards carefully-chosen prior information heatmaps **(challenge 2)** in order to increase the model's predictive capacities with *PGA*, while still leaving some freedom for the model to concentrate on other input features. The integration of spatial priors enables the model to better deal with the lack of annotated data **(challenge 1)**.

2.3 Prior-Guided Attribution of deep neural networks

In this section, we present a novel Prior-Guided Attribution (PGA) method, built to guide the network towards the most informative areas of the input image. First, we introduce the Privileged Attribution Loss (PAL), that allows the model to concentrate on certain areas of the input image by forcing its attribution maps to match a privileged information heatmap. Second, we present the Prior Allocation Strategy (PAS), which describes the way multiple prior information maps are integrated at the same time, while still leaving some freedom for the network to focus on other areas. Third, we discuss the selection of the prior heatmaps, and finally, we discuss PGA as a regularization tool.

Let f denote the CNN, L the number of layers in f, f_o the output vector, I the input image, $a_{i,j,c}(I)$ the attribution of the pixel (i,j) of the c-th channel of I and $a^l_{i,j,c}$ the attribution of the intermediary feature map $f^l(I)$ taken at layer l.

2.3.1 Privileged Attribution Loss

In this work we force the model to focus on certain areas of the input image. To do so, we leverage the model's attribution. Attribution methods compute the *relevance* or *importance* of certain input features with regards to the prediction. The most common gradient-based method is named *saliency maps*, first introduced by Simonyan *et al.* [85]. The aim is to compute the gradient of the output with respect to certain input pixels. In our work, we decide to compute the absolute value of the sum of derivatives of the outputs, since we want to constrain both negative and positive values of relevance on all output features. In the rest of this work, we refer to this method as the *Grad* attribution

method. For a pixel (i, j) at channel c, it can be written as:

$$a_{i,j,c}(I) = |\frac{\partial \Sigma f_o}{\partial I_{i,j,c}}(I)| \tag{2.1}$$

In particular, the structure of convolutional neural networks (CNN) can be taken advantage of. They can be written as a composition of functions or layers:

$$f(I) = f^L \circ f^{L-1} \circ ... \circ f^1(I) \tag{2.2}$$

where f is either an activation function, a convolutional layer, a batch-normalization layer of a pooling function.

This enables us to compute the relevance of feature maps from any intermediary layer instead of pixels from the input image. This might be of interest because each layer contains a different level of semantic information, therefore constraining deeper or shallower layers might require different kinds of spatial priors. The attribution for a channel c at layer l can therefore be written:

$$a^l_{i,j,c}(I) = |\frac{\partial \Sigma f_o}{\partial f^l_{i,j,c}}(I)| \tag{2.3}$$

In order to sharpen the attribution maps, Shrikumar *et al.* [84] introduced a new attribution method named *Gradient*Input* that consists in multiplying the derivative with the input value. One intuition of the difference between the two is that *Grad* corresponds to how a small change in the input will impact the output of the network, whereas *Gradient*Input* corresponds to the total contribution of a feature on the output of the network. Furthermore, Chen *et al.* [26] proved that every ReLU-based CNN can be decomposed as a piece-wise affine function of each pixel of f^l and that the contribution of each pixel i, j from channel c of a feature map f^l can be written as follows:

$$a^l_{i,j,c}(I) = |\frac{\partial \Sigma f_o}{\partial f^l_{i,j,c}}(I)| \cdot f^l_{i,j,c}(I) \tag{2.4}$$

We now aim at pushing the intermediary attribution maps towards resembling a normalized prior information heatmap denoted a^*. This encourages the model to "pay more attention" to the areas highlighted in a^* by having a larger impact on the predictions. We therefore introduce the *Privileged Attribution Loss (PAL)*, which optimizes the cross-correlation between the attribution map a^l and the spatial prior heatmap a^*. We choose to compute the cross-correlation instead of a negative Euclidian distance for instance because we do not really mind the scale of the attribution map. Formally, if we have $\mu(a^l) = \Sigma_{i,j} a^l_{i,j,c}$ and $\sigma^2(a^l) = \sum_{i,j}(a^l_{i,j,c} - \mu(a^l))(a^l_{i,j,c} - \mu(a^l))$, then PAL constraint can be written as follows:

$$\Lambda^l_{PAL}(\Theta) = -\sum_{i,j,c} \frac{a^l_{i,j,c} - \mu(a^l)}{\sigma(a^l)} * a^*_{i,j} \tag{2.5}$$

This *PAL* loss term is then added to a classical cross-entropy loss term, denoted $\Lambda_{CE}(\Theta)$:

$$\Lambda^l_{total}(\Theta) = \Lambda_{CE}(\Theta) + \alpha * \Lambda^l_{PAL}(\Theta) \tag{2.6}$$

where α denotes a weighting scalar hyperparameter denoting the importance of the constraint. Similarly to L1 or L2 regularization for example, the PAL loss term can be viewed as a regularization term for the classification objective.

2.3.2 Prior Allocation Strategy

The loss presented above offers a certain limitation as it only takes into account one single prior heatmap. However, we intend on incorporating several prior maps to guide the model towards several salient areas, since the SUOG project offers 11 pixel-wise segmentation maps for a small subset of the ultrasound images. To solve this problem, we introduce the *Prior Allocation Strategy* which consists in allocating a certain portion of the attribution maps to each prior. The first simple strategy involves forcing the first r channels of a considered layer to resemble the first prior, the next r channels to resemble the second prior, and so on. We name this the *All Channels* strategy. However, not allowing the network any freedom might be too strong a constraint, and a weaker formulation would be to only force the *mean* of a certain portion of these channels to resemble a certain prior. This allows a certain lenience in the learning mechanism while still guaranteeing that the model globally focuses on the right areas. In particular, with P denoting the number of different priors we want to use, C_1 the number of channels in f^l that we choose to constrain and $r = \lfloor \frac{C_1}{P} \rfloor$, our new attribution maps are:

$$a^l_{i,j,p} = \sum_{c=pr}^{(p+1)r} \frac{1}{r} a^l_{i,j,c} \tag{2.7}$$

the loss term therefore becomes

$$\Lambda^l_{PAL}(\Theta) = \sum_{i,j,p} \frac{a^l_{i,j,p} - \mu(a^l)}{\sigma(a^l)} * a^*_{i,j,p} \tag{2.8}$$

This formulation implies that all images from the training set have a spatial prior. However, the SUOG dataset only offers segmentation maps for a small subset of the training images. We therefore introduce a masking term m_p indicating if this spatial prior is present in the image or not. The loss therefore becomes:

$$\Lambda^l_{PAL}(\Theta) = -m_p \sum_{i,j,p} \frac{a^l_{i,j,p} - \mu(a^l)}{\sigma(a^l)} * a^*_{i,j,p} \tag{2.9}$$

The above-mentioned approach forces a constraint on all channels when $C_1 = C$. Here again, this constraint might be too strong in the sense that it would not allow the

model to explore other areas of the input image to find discriminative information (in other words, the prior knowledge maps might not be exhaustive). This is specially the case for the SUOG challenges, where the segmented objects might not be the only decisive structures in order to predict the scan plane. Thus, we decide to only apply this constraint to a certain portion of channels. This allows certain channels to stay free and investigate different areas of the input image that might contain crucial information absent from the prior knowledge maps.

In what follows, we explore four channel strategies, namely *All Channels* (discussed earlier), *Mean* (where $C_1 = C$ and C the number of channels in the layer), *Mean of Half* (where $C_1 = \frac{C}{2}$) and *First* P (where $C_1 = P$, and P is the number of prior maps).

The impact of the Prior Allocation Strategy is therefore two-fold, as it allows not only to choose which channels will be constrained and which ones will be free, but also the correspondence between priors and the allocated channels.

2.3.3 Prior generation and selection

The proposed PGA method is generic and can be used for different computer vision classification tasks. During training, it utilizes prior spatial information to guide the network's attention towards important areas of the input image. Importantly, this additional information is not needed during inference. Hence, the selection of the prior plays a crucial role in determining the final predictions. This study examines various types of priors and assesses how their nature might affect the results on different data. Specifically, we conduct experiments on three datasets: RAF-DB, a facial expression recognition dataset (and 3 variations), BUSI, a breast cancer detection dataset and finally SUOG for scan plane recognition. We now extensively present the datasets and their prior information heatmaps.

2.3.3.1 RAF-DB

The Real-world Affective Faces Database (RAF-DB) [56] is a dataset for facial expression where all examples were manually annotated by several annotators. It contains a train set of 12,271 images and a test set of 3068 images, all annotated with 7 basic emotions, namely surprise, fear, sadness, happiness, anger, disgust and neutral. In order to better validate the hyperparameter selection, we created a validation set (15% of the training set) sampled with the same label distribution as the test set. Facial expression recognition shares many similarities with scan plane recognition. It is a task limited by the relatively small number of annotated images in the datasets, it has a similar number of classes (<10), and has easily available prior spatial information. We therefore use RAF-DB as a toy dataset to test PGA while the medical images were not yet available. In particular, we created three different versions of this dataset to highlight certain characteristics of this method. First, we created **RAF-Aligned**, where the images are aligned using similarity

RAF-Cropped-Aligned RAF-Aligned RAF-In-The-Wild

Figure 2.4: Three versions of an image taken from the RAF-DB dataset. First row depicts the RAF-Cropped-Aligned version, where the image is cropped around the face, the second shows the RAF-Aligned version, where the image is aligned according to the face, and finally the RAF-In-The-Wild version where the image is not aligned according to the face.

transformation according to two eye locations and the center of the mouth. Then we used **RAF-Cropped-Aligned**, where the images are also aligned using the eyes and the mouth but are closely cropped around the face (this is the version most commonly used in the FER literature). We name the third version **RAF-In-The-Wild**, where a random crop around the face is applied on all the RAF-Aligned images, creating images that are not aligned according to the facial landmarks. This means that, contrary to RAF-Aligned and RAF-Cropped-Aligned, the crucial facial information is not mainly located in the same areas. On both RAF-Aligned and RAF-In-The-Wild, we guide the network using a face bounding box (*bbox* prior), available in the dataset, or a heatmap representing all 51 facial landmarks given by an off-the-shelf method [7] (*all landmarks* prior) to evaluate the impact of spatial priors that have different scales and compare a dense, coarse prior to a more fine-grained and sparse heatmaps. More specifically, from a 2D face image *I*, the face alignment model *g* locates 68 facial landmarks that form the face shape $\mathbf{y} \in \mathbf{R}^{68^*2}$ (the *n*-th row of this matrix corresponding to the 2D coordinates of the *n*-th facial landmark). Let $\mathbf{1}^L_{i,j}$ denote the indicator function which equals 1 when on a landmark, and 0 otherwise. From these facial landmark coordinates, we create an image a^* with:

$$a^*_{i,j} = \mathbf{1}^L_{i,j} \tag{2.10}$$

Which ultimately gives an image with pixel values of 1 at the landmarks and 0 elsewhere. We then apply a gaussian filter with a standard deviation σ of 3 to this image a^*. We then have:

$$a^{*filtered}_{i,j} = \sum_{k=1}^{68} \frac{1}{\sqrt{2\pi\sigma^2}} exp(-\frac{(i-y_{k,1})^2 + (j-y_{k,2})^2}{2\sigma^2}) \tag{2.11}$$

For the RAF-Cropped-Aligned version, which encompasses less variability and where

the crucial information is easier to locate, we force the model's focus using the previous *all landmarks* prior, but also using a prior containing 51 heatmaps, each one representing a single face alignment point (*independent landmarks* prior), and a prior that contains four heatmaps, each representing a group of face alignment points: eyes, eyebrows, nose and mouth (*grouped points* prior). This allows a sound evaluation of the impact of having several specific spatial prior maps instead of having all the crucial structures in one single map and how semantic groupings of prior information structures can benefit the learning. Examples of these images can be found in Figure 2.4 and examples of priors in the three first columns of Figure 2.5.

2.3.3.2 BUSI

The BUSI dataset [3] is composed of 780 breast ultrasound scans (see Table 2.1), collected in order to predict breast cancer. These images are annotated into three different classes: benign, malignant and normal. Images annotated as benign or malignant also come with a binary segmentation map that locates the areas of the tumors. These segmentation maps are used as our prior information in this work. We observe images with a certain variability in the location of the highlighted prior information (similarly to RAF-In-The-Wild), as well as in its shape and size. We separate this dataset into 5 train/val/test folds with similar label distributions in order to compare the results of different methods. In particular, we perform a 5-fold cross-evaluation, meaning that we evaluate the model as the mean accuracy scores on all 5 test sets. Examples of these images can be found in the two middle columns of Figure 2.5.

2.3.3.3 SUOG

The ten expert centers of the SUOG project collected over 200K ultrasound images at all stages of pregnancy. The SUOG dataset used in this experiment is composed of 1297 images annotated by the early pregnancy expert group of the consortium. The scan plane labels are grouped into 8 different labels, namely the longitudinal view of the uterus, oblique view of the uterus, the longitudinal view of the cervix, the longitudinal view of the bladder, the transverse view of the uterus, the interstitial portion view of the Fallopian tube, the longitudinal and transverse views of the adnexa and the ovary, and magnified views of the gestational sac. An illustration of these groups of views can be found in Figure 1.2. Among those images, 294 are segmented into 10 zones. We form several different priors from these segmentation maps:

- *All Zones*: This prior is composed of the 10 segmentation maps: the amniotic sac, the embryo, the gestational sac, the midline echo, the ovary, the uterus borders, the yolk sac, the cervix external ostium, the endometrium and the trophoblast. The pixel-wise segmentation allows us to have 10 maps that don't overlap.

- *Only Interior*: This prior is built as the union of all segmented zones. It therefore only contains one map.
- *Three Zones*: For this prior, we group semantically similar segmentation maps together. We merge all the gestational structures (the amniotic sac, the embryo, the gestational sac, the yolk sac and the trophoblast) in one map, the uterus structures (the midline echo, the cervix external ostium, the endometrium and the uterus borders) in another map and finally the ovary as the final map.

The dataset is split into train and test sets (with a train/test ratio of 5/1), the train set is then separated into 5 folds of the same size with similar label distribution, which allows us to choose our hyperparameters with a 5-fold cross-validation (see Table 2.1). Examples of these images can be found in Figure 2.5.

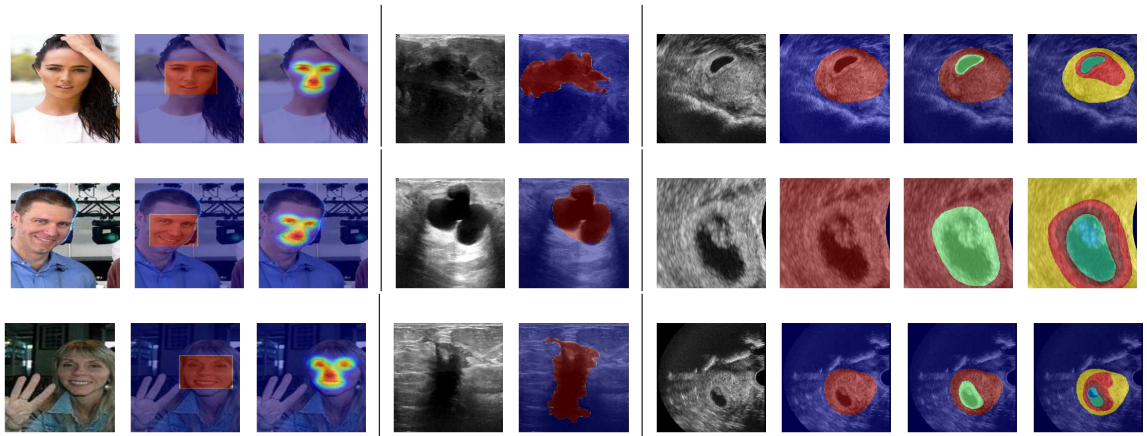


Figure 2.5: Examples of images and priors for the three datasets explored (RAF-DB, BUSI, SUOG). For RAF-DB, the first column depicts an image from the RAF-In-The-Wild dataset, the second column shows the *bbox* prior and the third column depicts the *all landmarks* prior. For BUSI, the the second column depicts the spatial prior highlighted. For SUOG, the first column shows the *Only Interior* prior, the second shows the *Three Zones* one and the last column depicts the *All Zones* prior. Best viewed in colour.

The nature of the prior information heatmaps is an important factor in the effectiveness of the method in several ways. First, the method's results can be significantly influenced by the different semantic granularities of the spatial priors. For instance, a more fine-grained prior could identify key areas and structures more accurately, but might encounter challenges generalizing to test examples (as it intuitively may require allocating more resources within the network to integrate this information), whereas a coarser prior might recognize broader or less precise areas but could be more easily learned during training. To explore this phenomenon, we examine the impact of guiding the networks with more

or less fine-grained priors, e.g. in the context of face analysis, a coarse bounding box and a more precise landmark heatmap prior. A similar argument could be made about the fact that some priors have dense highlighted areas whereas other highlight sparse zones. Identifying dense areas of the input image such as bounding boxes might be a simpler task than identifying sparse areas such as face alignment points.

Second, the number of different priors can highly affect the method. If the model is learnt using multiple maps that separate the spatial information into smaller distinct areas, it might be able to identify these zones more easily. However, it might not benefit from the relationships between these regions that would be taken into account separately in the Prior Allocation Strategy, as opposed to using a single prior built as the union of these areas. To assess this, we compare results obtained by guiding the model with both a heatmap representing all 51 face landmarks and another prior with 51 heatmaps representing one landmark individually to investigate the impact of multiple spatial prior maps for FER. For automatic scan plane recognition, we benchmark with a single, merged segmentation map as well as multiple segmentation priors.

2.3.4 PGA as a regularization method

Deep learning models often struggle to obtain good predictions for test examples because it has overfit the training data to some degree. This is particularly relevant when dealing with small training datasets. Methods such as L2 regularization [33], LASSO [96] or dropout [89] aim at avoiding this phenomenon, possibly at the cost of increased training errors. These methods usually consist in enforcing a prior structure on the network weights or activations. From this point of view, PGA can be viewed as regularization method, as its goal is to encourage the network to preferably look at certain regions to decipher the correct classification. As with the aforementionned regularization techniques, successful integration shall involve finding a successful trade-of between the strength of this regularization, and the classification loss.

This trade-of depends on a number of settings. First of all, an important such setting is at what level in the network (i.e. layer) PGA is applied. We know that feature maps from deeper layers are more precise but hold lower-level semantic information such as contours or textures, whereas feature maps from shallow layers hold higher-level semantic information. Constraining layers closer to the output might be better to integrate strong semantic information, but might penalize every layer coming before in the model if the localization task is too hard. We could also argue that the optimal layer to constrain might depend on the nature of the prior. In practice a coarse prior representing the contours of a specific structure might be more useful in the deeper layers of the network, whereas a more fine-grained prior with strong semantic information might be better adapted to layers closer to the output. Second, an important regularization parameter of the proposed method concerns how many channels are constrained. Forcing all of the channels individually to resemble a spatial might be too constraining and actually prevent the

network to learn any interesting patterns. Only constraining the mean of the attribution channels might offer the network more leeway to shape its attribution maps. Finally, only constraining a certain fraction of these channels allows the method to guide the model while giving some freedom for the network to identify and focus on other areas of the input image that might be interesting for the task at hand. Third, PGA might be sensitive to noisy spatial priors. Priors that highlight certain zones incorrectly or imperfectly might propagate errors to the model and therefore force it to focus on unimportant areas of the input image. However, the constraint on the learning mechanism might still work as a regularization method and therefore increase predictive performances. The *Mean of Half* channel strategy might boost the performances because of the regularization, while still being able to identify interesting zones that are not highlighted in the spatial priors. The proposed method also aims at alleviating the problem caused by the limited amount of annotated images, which often prevents deep learning models to generalize well to test examples. Finally, the weighting term α in the Privileged Attribution Loss (PAL) is essential. It enables the model to benefit from the semantic information and the regularization, while having some freedom to concentrate on the main classification task. A good compromise needs to be chosen for this weighting term between fine-grained priors where the highlighted areas are harder to locate, and coarser priors where the crucial areas are easy to identify. An α term too strong might penalize the model for the main classification task, but an α term too small might not enable the network to identify the correct zones. In what follows, we answer those questions and empirically validate the proposed approach.

2.4 Experiments

In this section, we evaluate the interest of the PGA method for multiple tasks: facial expression recognition, breast cancer detection and scan plane recognition. First, we briefly present the implementation details of the method. Second, we discuss results on the facial expression recognition task in Section 2.4.2, and more specifically the impact of certain settings and the choice of the prior information used to guide the learning. We then demonstrate the genericity of the method in Section 2.4.3 by discussing results on two OB/GYN tasks, breast cancer detection and scan plane recognition. In particular, we discuss the impact of the nature of the prior to optimize PGA as a regularization tool, and attach qualitative results to support this discussion.

2.4.1 Implementation Details

We train our model using ADAM optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. For facial expression recognition, we use a batch size of 16 and a base learning rate of $5e^{-5}$ with polynomial decay, whereas we use a batch size of 32 and a base learning rate of $1e^{-6}$ for the

Table 2.1: Description of the RAF-DB, BUSI and SUOG datasets.

Dataset	Train	Val	Test	Eval Protocol
RAF-DB	11044	1227	3068	Single Train/Test/Val split
BUSI	561	63	156	Cross-Evaluation
SUOG	944	237	150	5-fold Cross-Validation

OB/GYN tasks. On the one hand, for RAF-DB, all the results comparing the baseline and the model trained using PGA are given by a VGG16 architecture, pre-trained on VGGFace [67] for face recognition. The face images are resized to 244x244 and augmented with random rotation [-10°, 10°] followed by a random horizontal flip. On the other hand, for the BUSI and SUOG datasets, we use a VGG16 pre-trained on ImageNet and use data augmentation methods better adapted to medical imaging on-the-fly, so that each image has a different transformation at each epoch. The images are only augmented with a random vertical flip as it allows to keep the ultrasound imaging structure.

As it is depicted in Table 2.1, for RAF-DB, the best model is chosen by keeping the weights that maximize the accuracy on a validation set sampled with the same label distribution as the test set, while the best model is chosen by keeping the weights that maximize the accuracy on a 5-fold validation set for BUSI and SUOG. All results reported are accuracy results, computed as the number of correct predictions divided by the total number of predictions.

2.4.2 PGA for Face Image Analysis

In this section we validate the impact of the proposed method for facial expression recognition. This allows us to investigate which settings of the method fit the best, such as the layer to which PGA is applied, the attribution method, the choice of the channel strategy, the sensitivity to incorrect priors and the impact of PGA on very small datasets.

2.4.2.1 Which layer

One of the first settings of this method to look into, is which layer we are aiming to constrain to optimize the impact of PGA. Figure 2.6 shows the accuracy results of the PGA method applied to different layers of a VGG model. It demonstrates an important performance increase when the method is applied to the last layers of the model. This could be explained by the fact that the first layers encode low-level information and might struggle to identify the important zones. In particular, Selvaraju *et al.* [80] indicate that the last convolutional layers of a CNN are a good "compromise between high-level semantics and detailed spatial information". Another reason for poor performances when PGA is applied in the first layers of the CNN is that they contain very few channels, and

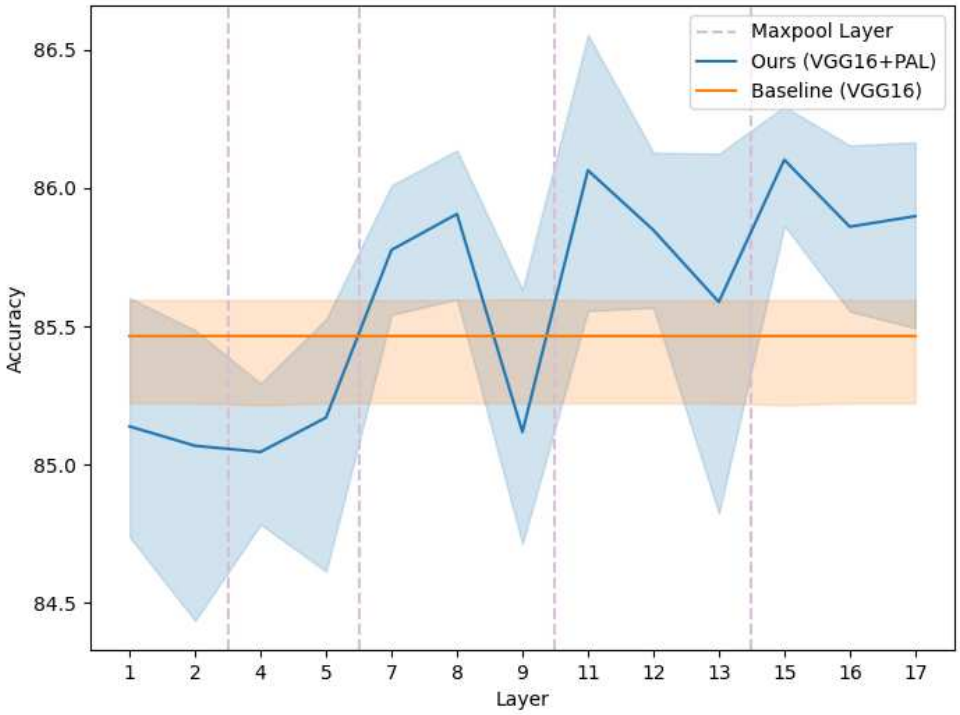


Figure 2.6: Performance of the model learnt with PAL applied to different convolutional layers of the VGG with Grad attribution method and Mean of Half channel strategy. The plot shows the mean accuracy score and 95% confidence interval. The method yields better results when applied near the end of the model.

therefore even applying the *Mean of Half* channel strategy might be too constraining. Finally, we can observe a drop in performance when we apply the method to the layers 9 and 13, because they precede max-pooling layers, represented in Figure 2.6 by a vertical line. Indeed, it means that these attribution maps are naturally sparse (e.g. three pixels out of four have a zero value for a (2, 2) pooling layer) and therefore are unable to match a the prior information heatmap. We observe that the variance is higher for layers 2 and 5 that also precede a max-pooling layer.

2.4.2.2 Which attribution method

Another sensitive parameter of the method is which attribution method is used. In this work, we evaluate the PGA method using the *Grad* and *Grad*Input* attribution methods. We see in Table 2.2 that the latter offers better performance results than the former, whichever channel strategy is used. Some research [84, 6] argues that *Grad*Input* provides sharper attribution maps, as the gradient accounts for the importance of a certain

Table 2.2: Ablation study on RAF-Cropped-Aligned dataset comparing different attribution methods and channel strategies.

Method	*Attribution*	*Channels*	*Acc*
$VGG16$	$---$	$---$	85.17
$VGG16 + PGA$	$Grad$	All Channels	85.59
$VGG16 + PGA$	$Grad$	Mean	86.34
$VGG16 + PGA$	$Grad$	Mean of half	86.47
$VGG16 + PGA$	$Grad * Input$	Mean	86.64
$VGG16 + PGA$	$Grad * Input$	Mean of half	**86.86**

input feature, and the input accounts for how strongly it is expressed in the output prediction. However, we can observe that both attribution methods lead to a steady boost in performance.

Table 2.3: Accuracy results on RAF-Cropped-Aligned dataset comparing of different values for C_1 on the channel strategy.

Method	C_1	*Acc*
$VGG16$	$---$	85.4 ± 0.2
$VGG16 + PGA$	$C/4$	86.55 ± 0.56
$VGG16 + PGA$	$C/2$	**86.86 ± 0.1**
$VGG16 + PGA$	$3C/4$	86.32 ± 0.25
$VGG16 + PGA$	C	86.38 ± 0.17

2.4.2.3 Which channel strategy

We also evaluate the impact of the PAS and channel strategies with a single prior in the first place. First of all, we can observe in Table 2.2 that constraining all the channels of the attribution map improves the accuracy results on RAF-Cropped-Aligned but is a very strong constraint. For instance, using the *Mean* channel strategy (e.g. $C_1 = C$) already greatly improves the baseline results by 0.98 points, as is shown in Table 2.3. This demonstrates that a weaker constraint allows the model to integrate the spatial information and give sufficient freedom for the network to better predict the facial expression. We therefore assess the importance of the value of C_1. We can observe on Table 2.3 that taking $C_1 = \frac{3C}{4}$ gives similar results to the *Mean* channel strategy, which would mean that the constraint is still too strong. However, taking $C_1 = \frac{C}{4}$ allows the model to reach an accuracy score of 86.55%, accounting for a 1.15 point increase compared to the results obtained without PGA. This shows that a weaker constraint allows the model some freedom to look wherever it wants while still benefiting from the spatial prior information. Finally, the best results are obtained with the *Mean of Half* channel strategy (e.g. $C_1 = \frac{C}{2}$), with an accuracy of 86.86%, which is a 1.46 point increase compared to the baseline.

2.4.2.4 Imprecision in landmark annotation

We now assess the importance of informative prior information to fully leverage the power of *PGA*. To do so, we train a facial expression model with *PGA* using incorrect prior information.

Figure 2.7: Examples of noisy heatmaps. The landmark points were sampled from a gaussian distribution with values of sigma going from 1 to 5 (from left to right).

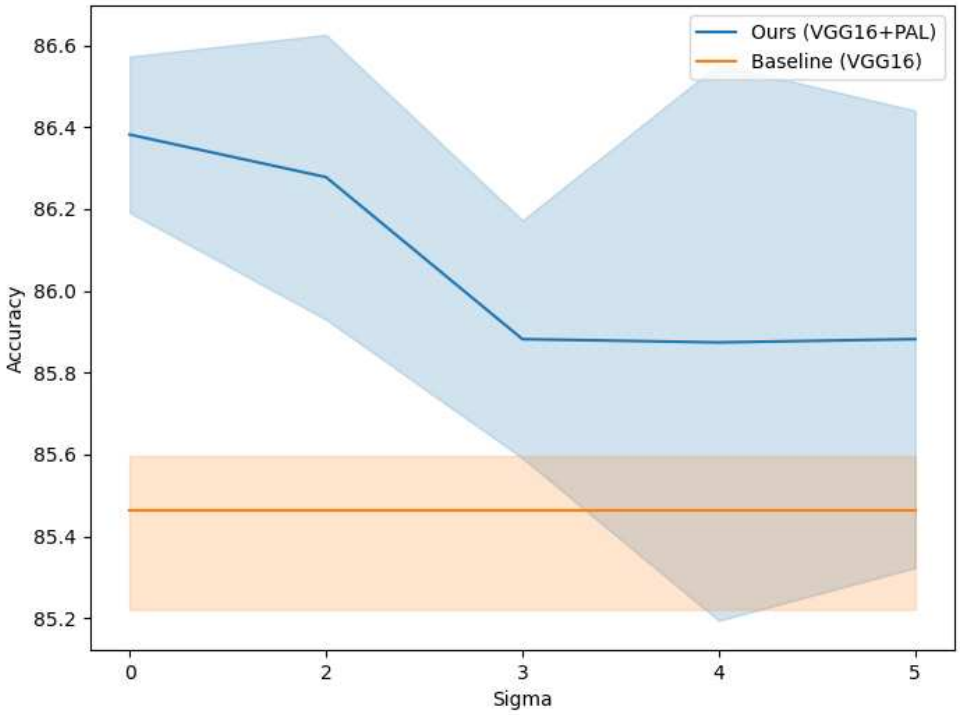


Figure 2.8: Performance of the model learnt with heatmaps created with noisy landmarks (evaluation for different values of sigma).

To understand the importance of the prior information, we train a face analysis model with *PGA* using heatmaps of noisy facial landmarks as a prior. We sampled landmarks

from a gaussian distribution with sigma equal to 2, 3, 4 and 5, examples of these heatmaps can be found in Figure 2.7. First of all, we can observe in Figure 2.8, that reports the classification results of the models trained using the different priors, that the model learnt with the correct prior ($\sigma = 0$) outperforms all the other models, therefore demonstrating the importance of having a correct prior. However, even though the mean accuracy results decrease while the value of sigma increases, we can observe that the models still all outperform the baseline. This can be explained by the fact that the heatmaps still highlight roughly interesting areas of the input image, giving a slight indication to the network as to where to focus. It also reinforces the idea that PGA works as a regularization method, allowing the model to generalize better on the test set. This result is important because it means that weaker priors might be sufficient to improve the network's predictive capacities, one could therefore rely on less costly annotations to guide the network.

2.4.2.5 Preciseness of the prior information

Table 2.4: Accuracy results on RAF-Aligned. The mean accuracy is computed over 5 runs.

Method	*Prior*	*MeanAcc*	*std*	*Maxacc*
Baseline	$---$	77.18	0.46	77.87
PGA	*bbox*	78.2	0.23	78.65
PGA	*all landmarks*	**78.5**	0.31	**78.94**

In order to evaluate the importance of the nature of the prior information, we apply PGA to the RAF-Aligned dataset (see the second image of Figure 2.4 for an example) with two different priors, the first being a simple bounding box around the face, and the second being a heatmap highlighting the 51 face alignment points. Table 2.4 shows us that guiding the attribution using a bounding box improves the results from 77.18% to 78.2%, whereas using all the landmarks as prior information improves the results up to 78.5%. This shows that for images where the crucial information is always located in the same areas, using a strong and precise prior is more adapted than using a weaker one. This can be explained by the fact that the model is easily able to locate the important structures and can therefore leverage the precise information, in this case the face alignment points.

More qualitative results can be analyzed in Figure 2.9, which depicts attribution maps from the RAF-Aligned test set. We can see in the first row that while the baseline model does focus on the faces, it also largely highlights irrelevant information such as background or other faces. The models trained with PGA however always manage to concentrate on the principal face only. In particular, the model trained with the *all landmarks* prior has no trouble identifying the faces, since the faces are always located in the same areas in the RAF-Aligned dataset. This backs our previous claim that, as the salient areas are

easy to locate, the more precise prior yields the best results, as we can see in the first and last columns.

Figure 2.9: troisieme colonne problèmeAttribution maps of images from the RAF-Aligned test set. The attribution maps are taken from models learnt with PGA using both a bounding box prior and face alignment heatmap prior, and compared to the baseline model. This model is trained using the *Mean of Half* channel strategy, which means that the second half of channels is not constrained.

We therefore evaluate the impact of PGA on data where the crucial information is not always located in the same areas of the input image. We therefore report the results of PGA on the RAF-In-The-Wild dataset (an example can be found in the third column of Figure 2.4) in Table 2.5. Results show that a model trained using PGA with a precise *all landmarks* prior improved the baseline results, taking the accuracy from 79.73% up to 80.97%, accounting for a 1.24 point increase, while using a broader *bbox* prior yields an

Table 2.5: Accuracy results on RAF-In-The-Wild. The mean accuracy is computed over 5 runs.

Method	*Prior*	*MeanAcc*	*std*	*Maxacc*
Baseline	$---$	79.73	0.34	80.15
PGA	*bbox*	**81.5**	0.36	**81.98**
PGA	*all landmarks*	80.97	0.5	81.65

accuracy score of 81.5%.

Conversely to the results on the RAF-Aligned dataset, where the faces are always in the same areas of the image, using a bounding box prior, which does not discern the precise face alignment points, gives the best results. We therefore argue that using a broader or less precise prior might be more helpful when the discriminative information is more difficult to locate.

2.4.2.6 Number of prior spatial information maps

After having evaluated the impact of a more or less precise spatial prior for the PGA method, we now evaluate our *Prior Allocation Strategy* and therefore assess how multiple prior information maps might affect the model's performance on more stable images from the RAF-Cropped-Aligned dataset (an example can be found in the third row of Figure 2.4).

Table 2.6: Accuracy results on RAF-Cropped-Aligned. Mean results computed over 5 runs.

Method	Prior	Channels	Accuracy
Baseline	—	—	85.4 ± 0.2
PGA	Independent	Mean	86.07 ± 0.24
PGA	Independent	Mean of half	86.26 ± 0.35
PGA	Independent	First 51	$\mathbf{86.42 \pm 0.15}$
PGA	Grouped	Mean	85.78 ± 0.33
PGA	Grouped	Mean of half	86.31 ± 0.37
PGA	Grouped	First 51	86.08 ± 0.29
PGA	Grouped	First 4	$\mathbf{85.59 \pm 0.25}$

First of all, we can observe on Table 2.6 that using the *independent points* (each map highlighting a single alignment point) prior outperforms the baseline regardless of the channel strategy used. As discussed earlier, we can see that the *Mean of Half* channel strategy yields better results than the *Mean* channel strategy as it allows the model more freedom to focus elsewhere and identify interesting areas or structures in the input image.

We can also see that guiding only the *First 51* channels also improves the results, with an accuracy of 86.42%, which accounts for a 1.02 point increase compared to the baseline. This shows that guiding the attribution of only one channel at a time of with a very precise prior (e.g. one face alignment point) is sufficient to improve the predictive results, while also giving more freedom to the network than the *Mean of Half* channel strategy. We can therefore argue that guiding the network with very precise priors such as single face alignment points might require constraining less channels of the attribution maps to provide a predictive improvement.

We also evaluate the interest of grouping different priors that are semantically similar. Table 2.6 shows that constraining the *First 4* channels with the *grouped points* prior only slightly improves the baseline accuracy scores by 0.19 points. This can be easily explained by the fact that the model is not constrained enough and does not benefit fully from the PGA method. To fairly compare with the *independent points* prior, we therefore constrain the *First 51* channels and observe that we obtain a 0.68 point increase compared to the baseline accuracy (constraining the first 51 channels corresponds to 10% of all the channels approximately). The use of PGA with semantically grouped priors offers an interesting increase in performance, taking the accuracy from 85.4% for the baseline to 85.78% with the *Mean* channel strategy. However, the *Mean of Half* channel strategy yields a large improvement with an accuracy score of 86.31%, probably because the prior is rather coarse and requires more channels to be constrained.

In a nutshell, these results might indicate that training a model with more fine-grained priors might require giving more freedom to the network, while fewer and coarser priors might requirea stronger constraint.

2.4.2.7 PGA helps with small datasets

In general, training deep learning networks on small amounts of annotated data is a cause of overfitting. We could argue that a relatively small training set could lead to the network struggling to identify the most discriminative structures of the input image for its predictions. In this sense, we evaluate the impact of PGA as a regularization method to help the model learn with few data. We therefore compare the results obtained by the baseline and the model trained with *PGA* on RAF-DB when trained with a fraction of the training dataset. Specifically, the model is trained with 10%, 40% and 70% of the RAF-DB training set and we report results on the test set. In Figure 2.10, we can observe a 13 percentage point increase for the model trained with PAL on only 10% of the training dataset. When trained with 40% and 70% of the train set, the proposed method obtains 10 and 8 point increase respectively in terms of accuracy. These results show that PGA works as a regularization method and helps the network fight overfitting when trained with less data by guiding its focus towards the salient areas.

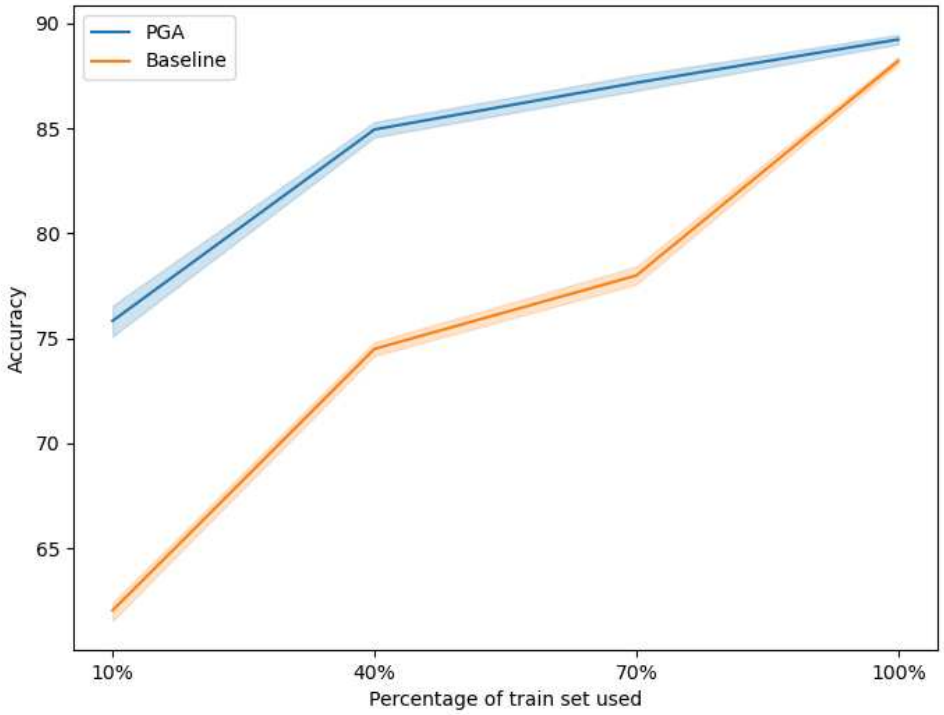


Figure 2.10: Comparison of the performance of the baseline and the model guided by PGA when learnt on a subset of the training data. Results are from the RAF-DB dataset. Results show that guiding the model with prior information constantly improves the predictions even when the priors are not perfectly correct.

2.4.2.8 Comparison with state-of-the-art results

We now compare PGA to other state-of-the-art methods. Table 2.7 shows a comparison of state-of-the-art methods for two very commonly used FER public datasets, RAF-DB and AffectNet.

First, we can see that our method significantly improves the accuracy results of the already very competitive baselines on both datasets (+0.96 points for RAF-DB and +3.69 points for AffectNet) with the exact same number of parameters, inference time and memory footprint. Also, the privileged prior information is only needed during training but not during inference, thus this performance upgrade comes at virtually no cost in inference.

Second, we extend the current state-of-the-art classification results up to 89.54%, while on AffectNet, we upgrade the results up to 65.83%. This is most likely due to the fact that, unlike its closest contenders (FDLR [76] and DMUE [83] on RAF-DB and PAENet [47] on AffectNet), the proposed method guides the network's focus towards

Method	RAF-DB	AffectNet
IPA2LT [120]	86.77	57.31
THIN [8]	87.81	63.97
DACL [38]	87.78	65.20
EfficientFace [123]	88.36	63.70
PSR [99]	88.98	63.77
PAENet [47]	–	65.29
DMUE [83]	89.42	–
FDLR [76]	89.47	–
Baseline (Resnet50)	88.6	62.14
Ours (Resnet50 + PGA)	**89.54**	**65.83**

Table 2.7: Accuracy results comparison for RAF-DB and AffectNet datasets. Results in %. PGA proves to outperform the state-of-the-art methods on both datasets.

areas of the input image that are discriminative for the prediction of the facial expression., while still letting the model have some freedom in its attribution. Last but not least, our method could in theory be used on top of these methods.

These results are taken from our publication [13] in 2022. Since then, other methods such as [112, 113, 60] have outperformed these results, most of them relying on transformer architectures.

In conclusion, we used a facial expression recognition task as a case study of a relatively small dataset, on which strong spatial priors can be extracted using off-the-shelf algorithms. In this context, we evaluated the impact of the proposed PGA method to improve the network's predictive capacities. We first showed that the loss is best applied to the attribution maps given by layers towards the end of the network and not behind a pooling layer. We also proved that PGA yields better results when applied using *Grad*Input* attribution method instead of the *Grad* attribution method. We showed that applying PGA to all the channels of an intermediate layer is too strong a constraint, and that guiding only the mean of half of the channels is a good compromise, since it gives valuable information as to where the crucial information is located, while still leaving tome freedom to the model to look elsewhere. Last but not least, we evaluated the impact of the nature, preciseness and number of prior information heatmaps in the results and showed that using PGA was all the more relevant that the training dataset is small.

2.4.3 PGA for obstetrics and gynecology

After having evaluated the proposed PGA method on a facial expression recognition task, and pinpointed its interest for using extra annotation to help the network learn on small

datasets, we now concentrate on the impact of the method for OB/GYN tasks such as breast cancer detection and scan plane recognition.

2.4.3.1 Breast cancer detection

Table 2.8: Accuracy results on the BUSI dataset, using the *Mean of Half* channel strategy. Results computed over 5 runs.

Method	α	$fold1$	$fold2$	$fold3$	$fold4$	$fold5$	$Mean$
$Baseline$	$---$	74.74 ± 1.59	84.87 ± 0.76	90.86 ± 0.62	76.41 ± 1.78	82.56 ± 1.25	81.88
PGA	1	77.69 ± 1.24	86.92 ± 0.65	90.51 ± 1.15	76.02 ± 1.65	82.17 ± 1.10	82.66
PGA	5	75.38 ± 2.27	87.98 ± 0.97	91.66 ± 0.81	76.66 ± 3.45	80.51 ± 0.51	82.42
PGA	10	79.10 ± 0.95	88.07 ± 1.49	92.05 ± 0.77	74.74 ± 1.49	80.89 ± 0.62	**82.97**
PGA	20	79.64 ± 1.89	90.06 ± 1.89	91.6 ± 0.53	71.31 ± 1.72	81.41 ± 0.79	82.80

In the previous section, we demonstrated the interest of *PGA* for facial expression recognition task. We now extend its interest on an OB/GYN image classification task, namely breast cancer detection on the BUSI dataset. Table 2.8 shows that applying *PGA* consistently improves the network's ability to detect the breast cancer, with an increase in the mean accuracy over 5 folds. In particular, we can observe a 0.54 point increase for $\alpha = 1$ and a 0.78 point increase for $\alpha = 5$, whereas the largest increment can be seen for $\alpha = 10$ (1.09 increase). An explanation for these results could be that the prior segmented zones are smaller and might be hard for the network to detect, therefore PGA might not provide sufficient insight into the spatial distribution of the structures present in the image. The network might be penalized by the fact that many images do not have a tumor to locate (the images annotated as *normal*), which might make the network struggle to know where to focus. An interesting aspect to notice is that we use a prior which directly indicates the object that the network is trying to detect, whereas the face alignment points for the RAF-DB and the segmentation maps for SUOG are only structures meant to help in the final prediction, and not exactly the task that we focus on. These results therefore prove the genericity of *PGA* as well as its interest for OB/GYN tasks.

2.4.3.2 Scan plane recognition

In this section, we evaluate the proposed PGA method on the SUOG dataset for scan plane recognition. In particular, we discuss the impact of different natures of priors. We first compare priors that are more or less precise and identify when each can be more suitable. We then assess when having multiple priors might be useful.

We now validate the interest of *PGA* and the different priors on an OB/GYN task: scan plane recognition on the SUOG dataset. This dataset mixes both the specificities

Table 2.9: Results on the SUOG dataset, using the *Mean of Half* channel strategy.

Method	*zones*	α	*Accuracy*
Baseline	− − −	− − −	79.76 ± 1.99
PGA	Only interior	1	81.46 ± 0.97
PGA	Only interior	5	83.73 ± 1.91
PGA	Only interior	10	83.86 ± 0.77
PGA	Three zones	1	82.83 ± 2.27
PGA	Three zones	5	$\mathbf{84.53 \pm 1.36}$
PGA	Three zones	10	83.19 ± 1.42
PGA	All zones	1	78.21 ± 2.13
PGA	All zones	5	83.06 ± 1.08
PGA	All zones	10	82.79 ± 1.14

of the facial expression task, as it contains multiple prior information heatmaps, and the specific crucial areas are not always located in the same zones of the input image.

First of all, we can see in Table 2.9 that the baseline accuracy reaches 79.76%, and applying *PGA* generally improves the predictive capacity of the classification model. Second, applying both *PAL* and *PAS* coupled with the *All Zones* prior and $\alpha = 1$ decreases the accuracy to 78.21%. This can be explained by the fact that these strong and precise priors can be a difficult to locate, therefore the network might focus on incorrect zones because the cross-entropy loss term is too important compared to the *PAL* loss term. However, increasing the value of α improves the accuracy results since it allows the model to focus more on identifying the right zones. *PGA* even achieves an accuracy score of 83.06% which accounts for a 3.40 points increase.

Working with a broader and less precise prior such as *Only Interior*, already provides an interesting increment working with $\alpha = 1$, with an accuracy score of 81.46%. This can be explained by the fact that the prior might be easier to identify than the *All Zones* one and therefore perform better with a smaller weight on the *PAL* loss term. However, increasing the value of α drastically improves the baseline scores by 4.10 points (from 79.76% to 83.86%), most likely because the network is able to identify the crucial zones with more ease.

Finally, we evaluate the impact of the *Three Zones* prior, working as a good compromise between the two previous priors. Again, this remarkably improves the predictive capacities of the model, climbing up to 84.53% in accuracy (a 4.77 point increase compared to the baseline). The *Three Zones* prior therefore constitutes a good trade-off between a very precise, fine-grained prior that is harder to locate, and an easier, broader prior that is however less informative.

Following experiments conducted on a facial expression recognition task and presented in section 2.4.2, we have extended these experiments and have demonstrated the interest

of PGA for OB/GYN and the importance of a well-chosen prior adapted to the task. In a nutshell, we showed that a broader prior might be more useful when the important areas are harder to locate, while a more precise and informative prior will yield the best results when the model will identify these areas more easily.

Furthermore, we prove the method's genericity by showing its interest for two different OB/GYN ultrasound imaging classification tasks, validating a significant increment in multiple different settings and versions of the RAF-DB dataset and extending the state-of-the-art for both public FER datasets RAF-DB and AffectNet.

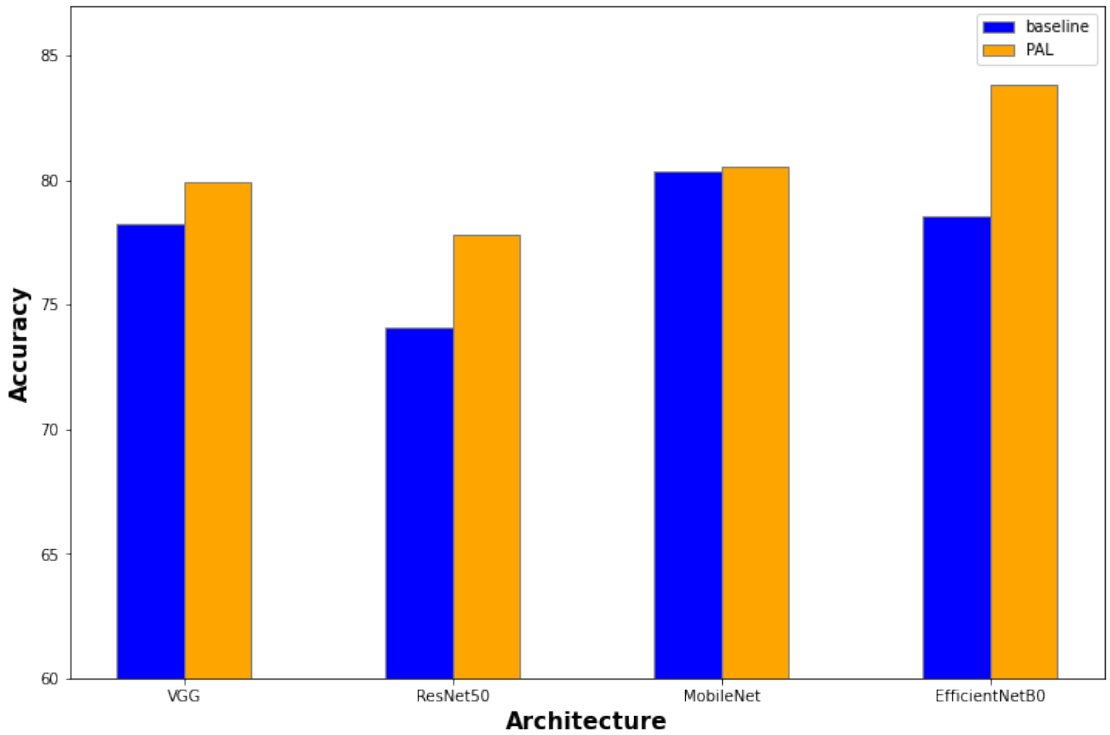


Figure 2.11: Study of the impact of *PGA* on different architectures. Results from the SUOG dataset.

2.4.3.3 PGA works on different architectures

After having proven that the proposed method was of interest in several different classification settings, we assess its interest for different CNN-based architectures.

To further explore the adaptability of the *PGA* method to different settings, we measured the impact of its increment on 4 convolutional architectures, namely VGG16, ResNet50, MobileNetV1 with a width multiplier set to 1 and EfficientNetB0. We can observe on Figure 2.11 that applying *PGA* to different architectures always improves the mean classification results for scan plane recognition on the SUOG dataset. In particular, *PGA* offers an increment of 0.17 points for the MobileNet architecture, which could be

explained by the size of the model. Indeed, the smaller size of the model and the fact that it is naturally heavily regularized enables it to fight overfitting but therefore might not benefit fully from the spatial guidance given by *PGA*, that also works as a regularization method. However, as it was shown earlier, both VGG16 and Resnet50 networks gain predictive power when coupled with *PGA* (e.g. +1.76 points and +3.73 points respectively). Finally, EfficientNetB0 benefits the most from the attribution guidance, with its accuracy results climbing from 78.58% without *PGA* to 83.82% with *PGA*, accounting for a 5.24 points increase. This therefore demonstrates that all CNN-based models can benefit from the proposed method with a carefully-chosen prior.

2.4.3.4 Qualitative Results

We have proved the interest of *PGA* for multiple tasks and settings with different kinds of spatial priors. We now analyze the results qualitatively.

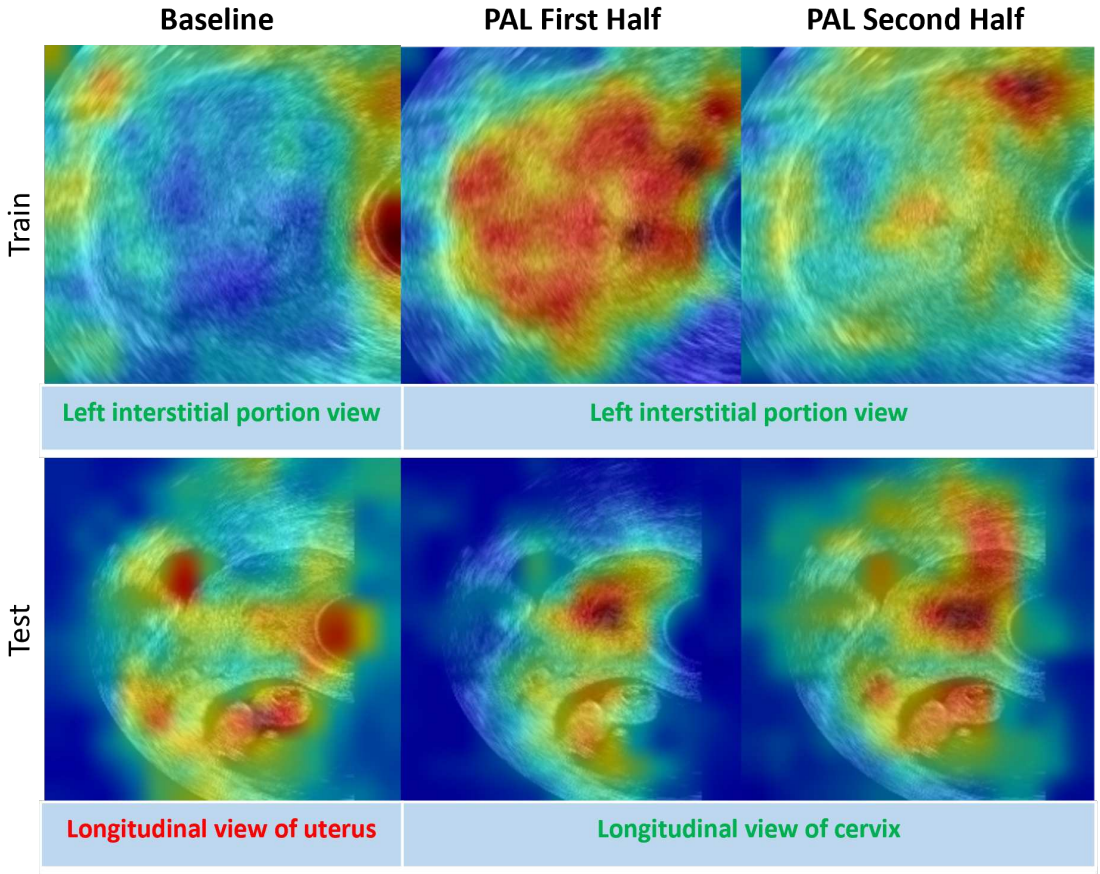


Figure 2.12: Attribution maps of two images from the SUOG dataset. The maps presented correspond to the model learnt PGA using the Only Interior prior. Best viewed in colour.

2.4.3.4.1 Impact of prior We use the attribution maps to better understand the model's reasoning. In Figure 2.12, we can see that the baseline model manages to focus well on the gestational structures of the the test image, but however focuses on the probe for both the test and training image, visible on the right side of both image. In general, a general CNN model might wrongly focus on certain artefacts of the ultrasound images due to noisy acquisition methods, which is a classic side-effect of working on medical images, and this might be avoided by working with *PGA* for classification tasks. Also note that, unlike the model guided by *PGA*, the baseline model concentrates on a dark spot next to the top-left uterus border of the test image that is only a cavity and has not been identified as an important area of this image for the task at hand. In Figure 2.13, the model guided by the *Three Zones* prior also focuses on this cavity and identifies it as a gestational structure, while the model guided by the *All Zones* prior identifies it as amniotic and gestational sac in Figure 2.14. This shows that a method learnt with a more fine-grained, precise prior might be more prone to such errors, whereas a larger and coarser prior might avoid this kind of mistakes because they look for less precise areas or structures.

More globally, we can assume that more fine-grained priors are more informative since they are more precise. Nonetheless, they are more likely to make mistakes and highlight unimportant zones or look for a structure that is absent from the image. For example, in Figure 2.14, the channels allocated to the ovary or the yolk sac are useless because used for missing structures. Conversely, a coarser prior such as *Only Interior* might guarantee to roughly focus on the correct areas of the image, but will provide less in-depth information. Therefore, we can conclude that the *Three Zones* obtains the best accuracy results because it a good compromise between strong and insightful priors, but large and easy enough for the model to correctly locate them. These results are essential because they allow the proposed method to yield very interesting classification results using a single segmentation map and even significantly improve the model's predictive capacity with only three segmentation maps. This means that the model does not need eleven extremely precise prior segmentation maps to perform well, as labelling images with such precise ground truth is very costly.

2.4.3.4.2 Impact of PAL weighting coefficient We can observe a certain impact of weighting term α for the *PAL* loss. Figure 2.14 depicts attribution maps of a model trained with the proposed method using the *All Zones* prior, the first couple of rows showing results for $\alpha = 1$, the third and fourth rows for $\alpha = 5$ and the fifth and sixth rows for $\alpha = 10$. We can observe that some maps are nearly blank for the $\alpha = 5$ and $\alpha = 10$, and we can argue that it is because these objects are seldom present in the training set, and are therefore difficult to identify for the model. For instance, the ovary is only present in a few training images and certainly seldom present in the test set. Therefore another explanation might be that these structures or objects are actually absent from the current image, and the network might struggle finding similar patterns.

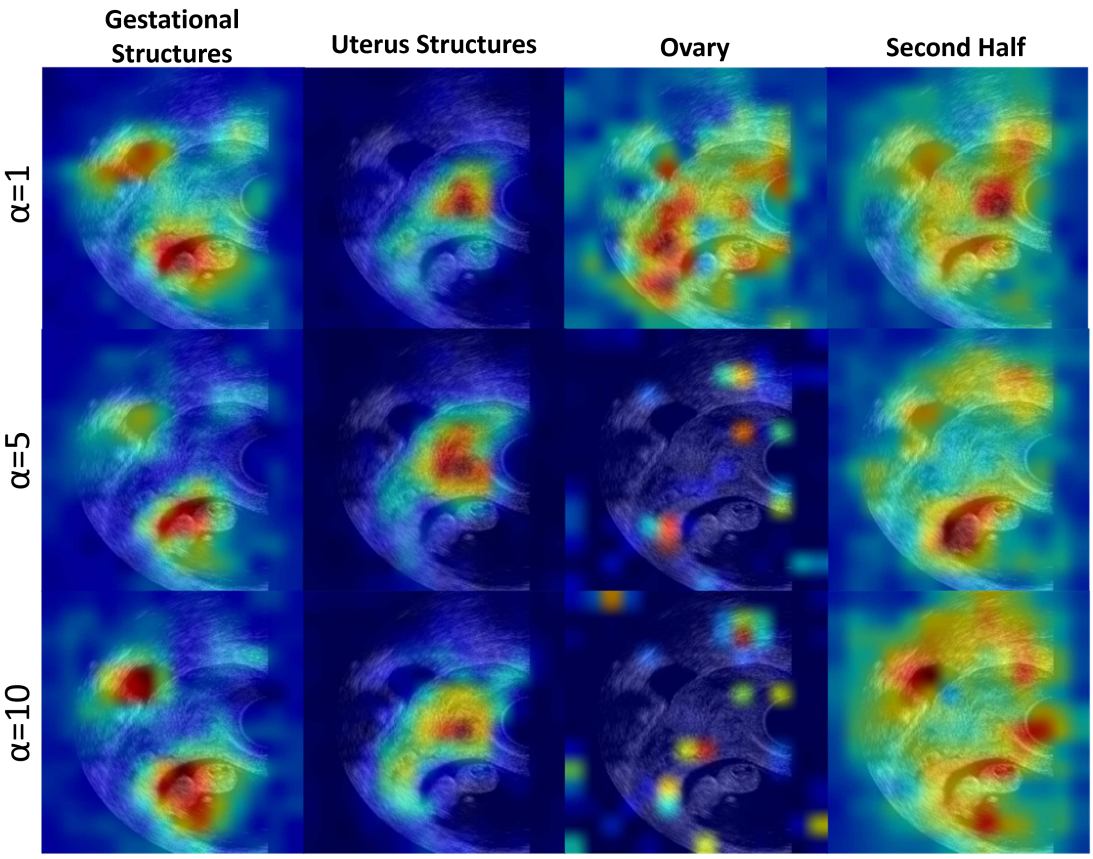


Figure 2.13: Attribution maps of a test image (midsagittal view through the uterus cervix) from the SUOG dataset. The maps presented on the first row correspond to the model learnt with $\alpha = 1$, the second $\alpha = 5$ and the last $\alpha = 10$, using the Three Zones prior. Best viewed in colour.

Finally, structures such as the midline echo, which is essentially just a line, are extremely hard to locate. This means that in the *Prior Allocation Strategy*, structures that are less useful in the scan plane prediction actually limit the model's capacities because a part of its channels is actually not focusing on anything important. This might be an explanation as to why the *Three Zones* and the *Only Interior* prior yield better results than the *All Zones* one, because they cut down useless priors and build a good compromise that enables the model to focus on more important structures.

However, another important detail to note is that the zones identified by the model trained with $\alpha = 5$ and $\alpha = 10$ are much neater than those from the first couple of rows. Indeed, the model might have some trouble identifying the very precise zones for $\alpha = 1$ because it focuses too much on the classification loss term, and this could therefore explain the large gap in performance for a model trained using *PGA* and the *All Zones* prior between $\alpha = 1$ and $\alpha = 5$ or 10 (see Table 2.9). We can also note that the difference in performance between the values of α is much less important for the *Only Interior* and

Three Zones priors, as the structures might be easier to identify.

2.5 Conclusion

2.5.1 Discussion

In this chapter, we investigated the idea to leverage prior spatial information to guide a network's focus towards the most discriminative areas of the input image. In particular, we noted that OB/GYN ultrasound imaging analysis was crucial, but that deep learning methods were still seldom used because of the lack of large annotated datasets **(challenge 1)**. Indeed, tasks such as scan plane recognition or breast cancer detection are still dealt with using rule-based methods or simple machine learning methods such as SVMs. We build on the idea that some objects or areas of the input image are more informative in order to make a good prediction. For instance, we argue that the eyes, eyebrows, mouth or nose hold more useful information to predict the emotion than the hair or the background. Therefore, in order to leverage the predictive strength of deep learning networks with limited amounts of annotated data, we take advantage of rich annotations **(challenge 2)** and specifically prior spatial information (e.g. segmentation maps) to help the network pay more attention to certain specific salient areas of the input image during training. In particular, we introduce the *Privileged Attribution Loss* (PAL) which forces the model's *attribution maps*, corresponding to the relevance of input features with respect to the outputs, to resemble a prior information heatmap by maximizing their cross-correlation. This enables the model to capitalize on additional expert information during training without needing it during inference, therefore coming a virtually no cost at test time. We also introduce the *Prior Allocation Strategy* (PAS) that allows the model to integrate several privileged information heatmaps at the same time while still leaving some freedom so that the network can investigate different areas of the image. The proposed *Prior-Guided Attribution* (PGA) method combines both PAL and PAS to guide the network's attribution towards any kind of spatial prior, and works as a regularization method to fight overfitting.

We first conducted experiments on the face analysis RAF-DB dataset to better understand and therefore optimize the way to apply PAL to a classification task. We observed that the *Grad*Input* attribution method worked best and that the method obtained better results when applied towards the end of the CNN network while leaving the network some freedom to focus on other areas of the input image. We then observed the impact of different priors on several versions of the RAF-DB dataset. Results showed that a coarser and larger prior might be more useful when the important areas are difficult to identify, whereas several smaller and more precise priors might be more useful when these zones are easier to locate. Heatmaps highlighting incorrect or imperfect spatial information proved to still consistently improve the baseline accuracy results,

demonstrating the interest of using PGA as a regularization method. This was further confirmed by results showing that PGA significantly increased the accuracy results when applied to a model trained on a small fraction of the training set, therefore proving to attenuate the burden of lack of data. We then proved that the proposed method was generic as it showed to consistently improve the predictive capacities of the model for a face analysis task and important increase in accuracy for two OB/GYN tasks, namely breast cancer detection for the BUSI dataset and scan plane recognition for the SUOG dataset. The method also proved to work on multiple CNN-based architecture, while coming at virtually no additional cost during inference.

As a conclusion, it is interesting to note that guiding a CNN-based model for a classification task using spatial prior knowledge drastically improves the accuracy results for many different tasks and in many different setups. However, choosing the right spatial prior for the specific task is essential. A more fine-grained prior is more informative but might have trouble locating the precise important zones, whereas a coarser prior might hold less information but will be easier to locate, and as such to learn for the network. In particular, for the SUOG dataset, the *Three Zones* prior, built as the union of semantically similar zones, works best as it is a good compromise between the broad *Only Interior* prior and the very precise *All Zones* prior. This result is important because it means that three large segmentation maps (which are rare and costly) for only a small fraction of the training set are sufficient to drastically improve the performances of the classification model.

2.5.2 Future Works

In the short term, we would like to be able to apply the method to different OB/GYN tasks. In particular, dealing with disorder prediction can be an interesting path, because it could help the sonographer with stronger information. This might be especially interesting because we could guide the model to focus on structures that are directly linked to the prediction. For instance, we could highlight a certain structure that implies a certain disorder in order to better predict it (i.e. a gestational sac separate from the uterus is key to an abdominal ectopic pregnancy, therefore if the network is able to identify both these structures it will be able to predict the disorder more easily). Another way to do so would be to use auxiliary tasks to improve the main task, similarly to works such as [117, 122, 110]. This would be interesting because we could adapt the prior information needed for the tasks at hand, and therefore guide the attribution maps of the different networks or branches with specifically chosen prior information.

Another track that we could explore would be to introduce spatial guidance to transformers and attention mechanisms. Transformers have shown to perform poorly when trained on small datasets, and using spatial constraints might work as an interesting regularization tool. Works such as Tallec *et al.* [95] and Tallec [94] have investigated the application of privileged spatial guidance for transformers but through the constraint of

attention maps. We could explore guiding the transformer's attribution, similarly to what was presented in this chapter. In particular, we could explore constraining transformer-specific attribution methods [46, 22].

Finally, since attribution methods were originally proposed for visual explanation [119], we could investigate evaluating the use of PGA as a way to improve the explainability of certain classification models. This work could possibly be of great interest because of the important responsibility in medical domains. In this sense, we could also explore the idea of updating PGA so that during training it would optimize both the model's predictions and its explainability.

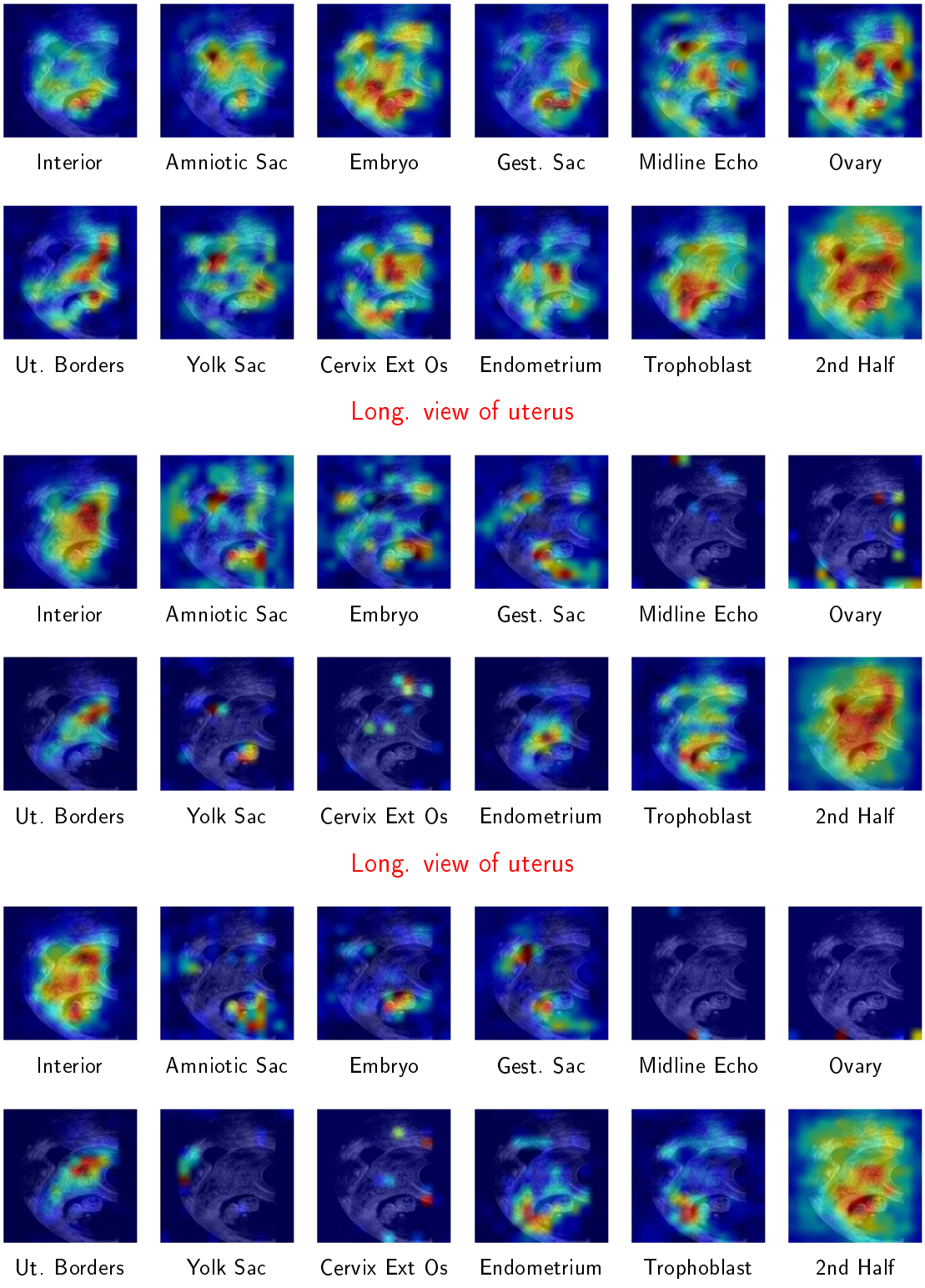


Figure 2.14: Attribution maps of a test image from the SUOG dataset. The maps presented on the first and second rows correspond to the model learnt with $\alpha = 1$, the third and fourth to $\alpha = 5$ and the last two rows to $\alpha = 10$, using the *All Zones* prior. The predictions of the three models were added on the right. The models trained with $\alpha = 1$ and 5 wrongly predicted *Longitudinal view of uterus*, whereas the model trained with $\alpha = 10$ correctly predicted *Longitudinal view of cervix*. Best viewed in colour.

Chapter 3
Ontology-Guided Learning

3.1 Introduction

In most of the usual machine learning tasks, the semantic space learnt by the model is relatively simple and can be formulated as multi-class classification tasks. This means that all ground-truth classes are considered as distinct and exclusive. For large datasets such as ImageNet, defining the problem as a classification is largely sufficient because of the sheer number of training examples, as well as the disparity between classes and the lack of additional semantic information, and treating it with softmax-based methods [15, 31, 59, 102] renders satisfactory results. Conversely, in obstetrics and gynecology, images are scarce (**challenge 1**) and furthermore pregnancy ultrasound images are characterized by a large number of rare illnesses (more than 1000) and an even larger number of signs. In particular, the SUOG dataset images are enriched by various annotations extracted from the SUOG ontology, created by experts from 10 hospitals across Europe. In the previous chapter, we treated the scan plane recognition problem as a classification task, which did not allow us to make use of the rich annotations available under the form of the ontology (**challenge 2**). Furthermore, the classification framework is suboptimal from an image retrieval standpoint (e.g. finding similar images to a query image from large database) which would allow the SUOG assistant to offer similarly annotated images to help guide the ultrasound operator in real time. We thus decide to explore methods that concentrate on optimizing the similarity between images, and mostly Deep Metric Learning (DML) that aims at modeling a sound embedding space for input images.

In a naïve DML setup, one would optimize models built to learn an euclidean embedding space that satisfies a semantic distance among training examples based on available labels. Yet, using this formulation, one would struggle to capture strong semantic links between classes as each individual classes are considered equally different (i.e. when considering ImageNet classification, one could easily assert that an *African crocodile* is closer to an *American alligator* than it is to a *saxophone*, yet they are all treated similarly different as they are distinct ImageNet classes). Another problem that stems from the DML framework is that it is usually used on large-scale datasets, which is seldom the case in medical imaging tasks, especially for pregnancy ultrasound images. In this chapter, we therefore aim at integrating strong semantic information to DML methods through higher-level annotations and language guidance to improve the inter-class similarity distances and alleviate the size-related problems.

For this purpose, we introduce a novel way to integrate specific domain knowledge through meta-annotations extracted from the SUOG ontology. We first introduce meta-embeddings that are meant to encode the information linked to annotations from different semantic levels (*animal* and *reptile* as meta-annotations for an *African crocodile*, for instance) and introduce a novel *Semantic Abstraction Loss (SAL)* that consists of a combination of multiple DML losses applied at different semantic levels. In order to make use of the strong semantic textual information that is carried in the annotations extracted from the SUOG ontology, we then use a language guidance module as presented by Roth *et al.* [75]. We demonstrate that the integration of strong semantic information as meta-labels or as textual information improves the organization of the latent representation space and helps for image similarity, but also helps make mistakes that are semantically closer to the ground-truth, which is crucial to assist in the practice of fetal medicine. These experiments are first validated on a classic DML dataset (CUB200 [100]) as well as the SUOG OB/GYN dataset. In summary, the main contributions introduced in this chapter are:

- We introduce novel meta-embeddings built to encode hierarchical semantic information extracted from a class ontology. To ensure a sound latent representation space, we learn a meta-loss *Semantic Abstraction Loss (SAL)*, constructed as a weighted average of DML losses applied to training pairs of meta-embeddings and meta-classes.

- We propose two different ways to integrate the rich annotations as textual inputs to better guide the visual model during training. We first integrate the hierarchical nature of the annotations as *rich captions* for language guidance. We also build on the language guidance proposed by Roth *et al.* [75], and introduce an *Ontology Language Guidance (OLG)* that guides the meta-embeddings using natural language.

- Experimentally, we validate the interest of introducing higher-level semantic information for visual similarity on the CUB-200 [100] dataset and prove the efficiency of this method for assistance in fetal medicine with the SUOG dataset. The interest of adding strong semantic information in the shape of textual privileged information is also showed in this chapter. We also demonstrate quantitatively and qualitatively that the prediction mistakes made by the model are more sound when learnt with *SAL*. This method is shown to be generic and to work on many different DML losses.

This chapter is divided as follows: in section 3.2, we present the historic methods and state-of-the-art approaches to DML. We then present a DML learning framework and several commonly used baselines in section 3.3, before introducing the methodology used to integrate ontology-extracted strong semantic information in 3.4. We then provide

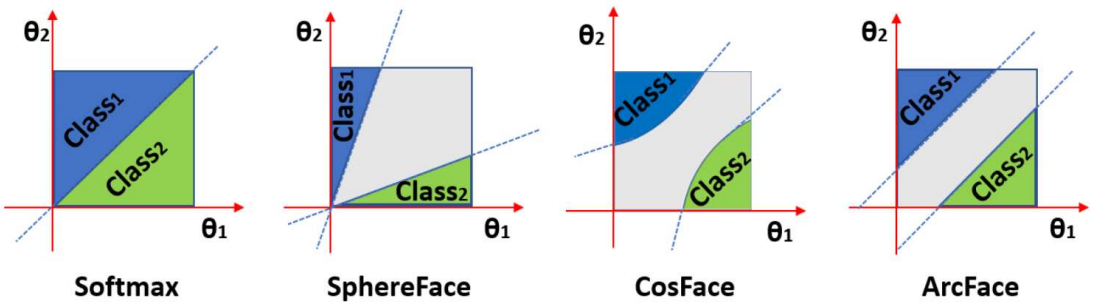


Figure 3.1: Illustration of the decision boundary margin for a binary classification case for 4 different losses. Θ_1 and Θ_2 are the angles between the weight of output and the input feature. ArcFace has a constant linear angular margin throughout the whole interval. This figure is taken from Deng *et al.* [31].

experimental results that demonstrate the interest of our method in 3.5, and finally offer a discussion on this work and quickly explore possible future works in 3.6.

3.2 Related Works

In this section, we will first give an overview of the basic methods to deal with a DML framework using deep learning architectures. Then, we will present methods that integrate language modalities to improve their visual representations, and finally the methods that leverage structured annotations during learning.

3.2.1 Deep Metric Learning

Deep Metric Learning aims to learn informative embedding spaces that encompass strong and meaningful semantic context, where the representations of similar images are close and those of dissimilar images are further away. These types of learning frameworks have mostly been used in the case of open-set classification where train and test labels aren't the same (e.g. face verification), and therefore the problem can't be resolved by a classification framework. This has led to a great interest in DML for tasks such as zero-shot learning ([71, 74, 77, 79]), clustering ([42, 88, 105, 114]) or person re-identification([31, 79, 59]).

We present here the most popular methods based on the classification framework, different losses designed to learn on tuples of examples and said tuple selection heuristics introduced to improve the visual similarity learning. Finally, we put forward methods that split the embedding metric subspace to improve their predictive model.

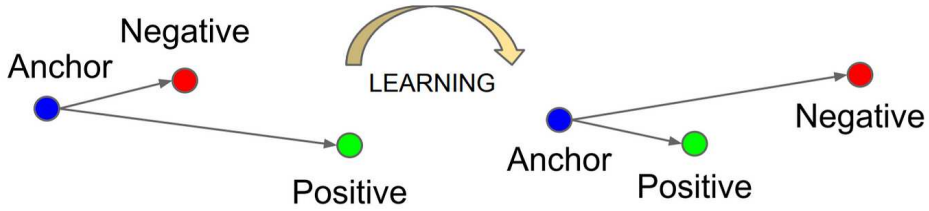


Figure 3.2: Illustration of the concept of triplet loss from Schroff *et al.* [79]. It aims at pulling similar pairs of examples closer and push dissimilar pairs of examples further apart.

3.2.1.1 Classification-based methods for DML

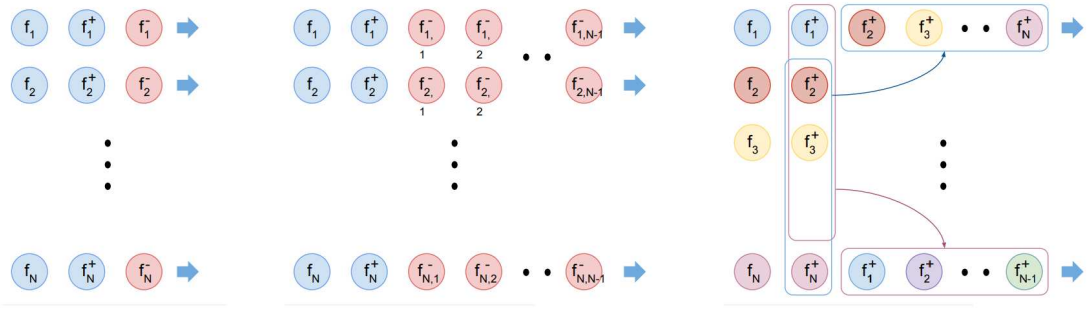
Deep Metric Learning aims to learn and optimize an embedding space that has discriminative properties. This cannot be treated the same as a classification task because this framework aims to learn a finer semantic information and notably has been used to solve multiple open-set challenges (i.e. where the labels from the train and test set are different). However, a naive way that has rendered satisfactory results of dealing with DML problematics is to use a softmax classifier [15] to separate the training classes, and then use the latent representation during inference. While this method works well on classification tasks and manages to well separate the training classes, research showed that it struggled on large open-set problems. In particular, tasks such as person re-identification or face recognition needed an expressive and semantically well organized embedding space to be able to separate thousands of new identities in the test set.

Therefore, a lot of research concentrated on softmax-based losses that explicitly worked towards creating a sound embedding space with discriminative properties. All these methods leverage a *modified softmax definition,* where an angular definition is presented. One way to better separate classes during training is to implement a margin penalty to that angular definition. For instance, Liu *et al.* [59] present SphereFace (Figure 3.1.b) which updates the softmax loss with the *A-Softmax loss*, which has a multiplicative angular margin. Wang *et al.* [103] build on this work and present CosFace (Figure 3.1.c) and the associated loss *LMCL*, where the margin term isn't multiplicative but added to the cosine of the angular term. Finally, Deng *et al.* [31] introduce ArcFace (Figure 3.1.d), where the margin is added directly to the angle (e.g. inside the cosine function instead of outside for LMCL). Deng *et al.* [31] argue that they obtain a constant angular margin whatever the angle is, as opposed to LMCL. These results are depicted in Figure 3.1. These works, focusing mainly on face recognition, mostly leverage very large amounts of data (millions of training images), which is far from the medical imaging problematic.

3.2.1.2 Metric Learning on tuples of samples

To improve the capacity of a vision model to encode sound semantic information, many methods have decided to explore the comparison of tuples of examples.

Hadsell *et al.* [45] introduced the notion of *Siamese networks* and contrastive loss. Instead of optimizing a loss over a sum of individual examples like most conventional machine learning methods do, the idea is to work with pairs of examples. This loss explicitly pulls the embeddings of similar pairs of examples closer and pushes embeddings of dissimilar pairs of examples further apart, with the embeddings outputted by the said *Siamese networks*, which are feature extractors with shared parameters. However, Schroff *et al.* [79] argues that a contrastive loss encourages all images from one class towards the same embedding without worrying about the distance between different classes. They therefore introduce a triplet loss, which works on (anchor, positive, negative) triplets instead of pairs. The anchor is compared to the positive example (e.g. they share the same class) and to the negative example (e.g. they have different classes). The idea is to reduce the intra-class distances and increase the inter-class distances, by forcing a certain margin between the positive and negative distance for each triplet. This is illustrated in Figure 3.2. Alternatively, Chen *et al.* [25] build on the triplet loss for person re-identification. They introduce a quadruplet loss that takes an anchor, a positive example and two negative samples. They prove that adding a loss term that pushes away negative pairs and positive pairs with a smaller margin helps reduce intra-class variance and increase the inter-class variance. This leads to an overall better generalization. Sohn [87] creates a (N+1)-tuple loss, where a positive example is pulled towards the anchor while N-1 diverse negative examples all from different classes are pushed away. The novelty here lies in the construction of the N-pair mini-batches for highly scalable training, using only $2N$ examples instead of $(N+1)N$ to build N tuplets of length $N+1$. They only use two samples from each class and are therefore able to build a positive pair for each anchor and $N-1$ negative samples all from different classes. This strategy is illustrated in Figure 3.3.



(a) Triplet loss (b) $(N$+1)-tuplet loss (c) N-pair-mc loss

Figure 3.3: An illustration of the N-pair loss introduced by Sohn [87]. While, for a batch consisting of N distinct queries, the triplet loss requires 3N different examples, and the (N+1)-tuplet loss needs N(N+1) examples, the N-pair loss only uses 2N examples. Furthermore, it integrates more variety and leads to a faster convergence, as each query example is compared with at least one example of each negative class.

However, these methods need to optimize the way they select their training tuples, because most tuples are uninformative as they are too easy to classify. In the next section, we present the numerous tuple selection heuristics mentioned earlier.

3.2.1.3 Tuple Selection heuristics

In this paragraph, let x^a, x^p and x^n denote the anchor, the positive sample and the negative sample respectively, and f denote a certain feature extractor. Schroff *et al.* [79] introduce the idea of carefully selecting triplets to improve the learning dynamics. The naive way to do it, for a mini-batch, is to average the triplet loss on all valid triplets, called *batch-all* mining strategy. However, quickly most triplets become uninformative, because the distance between the anchor and negative examples is larger than that of the positive example. Therefore, the informative triplets are diluted in the middle of the trivial triplets, and the model learns too slowly or stagnates. To alleviate this problem, they argue that selecting the hardest positive x^p (s.t. $argmax_{x^p}\|f(x^a) - f(x^p)\|$) and the hardest negative x^n (s.t. $argmin_{x^n}\|f(x^a) - f(x^n)\|$) for each anchor example x^a would lead to the best performances, and call it *hard negative mining*. The problem with *hard negative mining* is that it can lead to collapsed models where all embeddings are pushed towards 0. Schroff *et al.* therefore decide to introduce *semi-hard triplet mining*, where negative examples are further from the anchor than the positive example, but still have a positive loss. They have to meet the specific criterion : $\|f(x^a) - f(x^p)\| < \|f(x^a) - f(x^n)\| < \|f(x^a) - f(x^p)\| + \alpha$ with α being the margin parameter. Additionally, to ensure that there is a certain representation of each class in each mini-batch of n examples, they build batches with k examples from $\frac{n}{k}$ different classes. Carvalho *et al.* [21] propose to divide the sum of losses induced by all triplets by the number of informative triplets. In practice, this means that the model uses a *batch-all* mining strategy at the beginning and ends with a *batch-hard* strategy at the end of the training.

Such tuple selection methods that use all the batch examples can often lead to a trade-off between efficiency and accuracy. Suh *et al.* [91] propose a new stochastic hard negative mining method. They assume that distances between embeddings of the same class are relatively small. They therefore decide to replace every pair-to-pair distance by a pair-to-class distance using class signatures. This allows them to reduce computing time and also improve predictive performances. Wu *et al.* [105] argue that a smart and efficient sampling strategy is the most important part of a DML pipeline. They advance that hard negative sampling isn't optimal because the gradient has a high variance and is dominated by noise, while random sampling mostly yields examples that are too easy and that induce no loss. Semi-hard negative mining, presented earlier in this subsection, seems a good compromise between the two. However, after learning very well at the beginning of the training, semi-hard mining isn't as efficient because there are no examples left in the semi-hard distance space. Distance weighted sampling is therefore introduced as a solution to this problem, where negative examples are sampled uniformly according to the

distance. This allows the model to see a wide range of examples, from easy to hard, and therefore control the variance.

For each negative pair, Wang *et al.* [104] make use of three different kinds of similarities in their DML framework : self-similarity (computed from the pair itself), negative relative similarity (computed in relation to neighbouring negative pairs) and positive relative similarity (computed in relation to neighbouring positive pairs). They use positive relative similarity to mine informative pairs and use both the self-similarity and negative relative similarity to weight the chosen pairs. This method is the one giving that gives the best results, and we will therefore discuss it in more detail in 3.3.2.4 and use it in the experiments.

In this section, we presented an overview of the most commonly used losses and tuple selection heuristics. In this work, we will concentrate on those that perform the best. We now concentrate on the nature of annotations, and more specifically, methods that leverage textual modalities to add semantic information during the learning.

3.2.2 Leveraging textual modalities for visual Deep Metric Learning

To better represent the semantic relations between training examples, many researchers have explored integrating textual modalities during training. In particular, we will present works on cross-modal retrieval with a language modality and then more specifically visual DML methods that guide their learning using privileged language representations to improve their similarity predictions.

3.2.2.1 Cross-Modal Retrieval

The concept of integrating textual modalities for visual similarity shares several aspects of methodology with the research done on cross-modal retrieval. This task consists in correlating multiple modalities of inputs, for instance images and texts. One way to do so is to project the multiple modality inputs into a shared latent representation space. Zhen *et al.* [124] optimize a discrimination loss in the common embedding space and in the label space to learn rich semantic features. Carvalho *et al.* [21] leverages the classic DML triplet loss framework for cross-modal retrieval in the cooking context, using a list of ingredients or cooking instructions as the textual input and an image of the dish as the visual input. They introduce a double triplet loss that relies on two levels of annotations (fine and coarse-grained) learned jointly to update their representation space. Similarly, Xu *et al.* [109] also take inspiration from DML methods by optimizing the projections of pairs of samples in order to increase the inter-class variability and reduce the intra-class variability. To do so, they project the pairs of samples from different modalities in a common subspace, which are then compared by a *feature correlation loss* which pulls them closer if they are of the same class and pushes them apart if they are not. The novelty here

is that they also use an *adversarial loss* that should predict which modality the sample is from originally. Peng *et al.* [68] integrate Generative Adversarial Networks (GANs) to improve the model's representation power. First, both text and image modalities are projected into a shared representation space and compared using an inter-modality discrimination loss. Second, a generative model is used to reconstruct both the original image and text, and are compared by an intra-modality discrimination loss. This proves to increase the model's performance in terms of retrieval metrics.

These methods create and optimize a shared latent representation space for different modalities in order to retrieve samples from one modality with a query from another modality. In our work however, we leverage natural language during training to improve the visual model's capacity but do not need it during inference.

3.2.2.2 Guiding DML with privileged language representations

Language models leverage the rich information found on the Internet and are able to encode very powerful semantic information, most notably thanks to training these models on few-shot learning tasks [71, 17, 72] using huge corpora of text (e.g. CLIP [71] is trained on 400 million training samples, corresponding pairs of image and text). They have recently become ubiquitous in most large-scale deep learning tasks and have also been used in order to improve other deep learning tasks (e.g. using language to improve visual encoders). Audio retrieval using natural language queries has been explored by Oncescu *et al.* [65], that compare different benchmarks and prove the interest of guiding the audio retrieval task using language guidance. Radford *et al.* [71] emphasize on the idea of learning visual features through natural language supervision. They introduce a method that replaces class annotations by the rich language representations. They exploit a very large dataset of 400M (text, image) pairs and very large mini-batches and evaluate their model using zero-shot transfer on large image classification datasets. To do so, they use a categorical cross-entropy loss on the similarity matrix between text and image embeddings, and therefore push for a similarity matrix that looks like the identity matrix. This method earns excellent results on most common datasets but uses very large amounts of data and computing power. In a similar vein, although applied to a DML learning framework more adapted to reasonably small datasets, Roth *et al.* [75] capitalizes on a (frozen) language encoder, guiding the visual similarity matrix towards the textual similarity matrix by means of a *matching loss*.

We have presented a way to add semantic context to a DML framework through language guidance. However, this necessitates a text encoder that is knowledgeable in the training domain, which is not always the case, especially in the medical context. Another way to improve the semantic information in a vision model is to leverage rich and structured annotations to guide the learning. This is presented in the next section.

3.2.3 Using hierarchical annotations as a prior to guide the learning

Another way to improve visual similarity learning is to integrate hierarchical annotations into the learning framework. In particular, the idea is to break the common mistakes done in classification that treat all classes equally different, and therefore better optimize inter-class distances. We first present methods that leverage the class hierarchies to better represent the ground-truth labels in a latent embedding space, then methods that update the model's architecture to mimic the class hierarchy, and finally methods that create custom losses that take into account the hierarchical structure of the data.

3.2.3.1 Using class hierarchies for label-embedding methods

One way to introduce hierarchical class annotations is to map labels to latent representations that have the potential to better encode the semantic similarity between pairs of classes. In particular, Frome *et al.* [41] create a label embedding from a skip-gram language model and then use a ranking loss between the output of a vision model and the label embedding. Similarly, Barz et Denzler [9] map images onto a hypersphere such that the distances represent similarities derived from the Lowest Common Ancestor (LCA) in the label annotation hierarchy tree. LCA is a similarity metric that indicates a certain similarity between two nodes of an ontology or a graph, by computing the distance with their lowest common ancestor.

Akata *et al.* [2] apply this method for zero-shot classification. They compare and therefore evaluate multiple output embedding methods: textual similarities (Glove, Word2Vec, weakly-supervised Word2Vec, BOW), human-annotated attributes and finally hierarchical embeddings. In another work, Xian *et al.* [107] create multiple latent visual embeddings and select one using a ranking loss.

These methods try to create sounder visual similarities by embedding label annotations using certain heuristics. More specifically, they learn to project samples towards fixed positions in the latent space and defined by the labels. However, since the fixed target space is very dependent on the training classes, the methods still present generalization issues.

3.2.3.2 Hierarchical architectures

Other methods try to integrate the hierarchical nature of the data annotations into the architectures. Ahmed *et al.* [1] implement a network of experts, where a generalist CNN learns to classify a subset of K higher-order classes where $K < C$, with C being the number of classes. The backbone of the generalist network then feeds K expert branches that learn to discriminate the classes within a specialty. Similarly, Alsallakh *et al.* [5] discuss the nature of CNNs and how different convolutional blocks learn different levels of features. They also implement an AlexNet model where classification branches are added

after each convolutional block and learn to classify different levels of the class hierarchy. Yan *et al.* [115] simplify the class hierarchy by dividing it into coarse and fine-grained categories. Again, a shared feature extractor feeds both a coarse component classifier and K fine component classifiers. They however offer a novelty with a probabilistic averaging layer which allows the model to weight the fine-grained predictions by the coarse class predictions.

These methods smartly embed the class hierarchy into the model architecture, but mostly do it for classification tasks and are very specific to certain architectures. This is particularly problematic because these specific architectures have to be retrained from scratch and do not benefit from a consistent pretraining on a large-scale database. In order to bypass that problem, other methods integrate the hierarchical nature of the data by creating a custom loss.

3.2.3.3 Hierarchical Losses

Different works have highlighted the importance of having a specific loss function that takes into account higher-level semantic context. Deng *et al.* [29] argue that most classification tasks have been trained using unrealistic datasets with a number of class that isn't large enough and where classes can be easily separated. To build a stronger classifier that would work in other scenarios, especially with numerous classes (they test it on ImageNet10K for instance), they propose to minimize a hierarchical cost. This hierarchical cost is computed as the Lowest Common Ancestor (LCA) metric extracted from WordNet on SVM and kNN-based classifiers. Verma *et al.* [98] propose a "context-sensitive loss" where the LCA extracted from the class hierarchy provides valuable insights to learn similarity metrics between pairs of classes. Bertinetto *et al.* [11] argue that recent methods that have obtained impressive results on most computer vision tasks, have not really improved the mistakes that were made by these models. This means that even though the accuracy has increased, when the model makes a mistake (e.g. predicts a class different from the ground-truth class), the predictions are still semantically very different to the ground-truth class. This can be an important problem in tasks such as medical imaging, where we would want for the predictions to be as close as possible to the ground-truth label semantically even if they are wrong. They therefore introduce a "hierarchical cross-entropy" as the reweighted sum of the cross-entropies of the conditional probabilities that are derived from the label taxonomy tree. For image similarity, Ge *et al.* [42] update the Triplet Loss by introducing a novel dynamic margin instead of a fixed one in the original Triplet Loss. This new margin takes the class hierarchy tree into account to produce the loss term. In particular, negative pairs that are further apart in the class hierarchy tree will be pushed further apart than negative pairs whose classes are semantically closer.

Unlike the majority of the works presented in this section, our work builds on DML frameworks (and not classification or regression tasks) and allows for good generalization

on classes absent from the training set. Also, it doesn't require a custom architecture or heavy ensemble methods. In particular, we leverage structured annotations to better input semantic context into the vision encoder model and help improve the inter-class distances with a *Semantic Abstraction Loss* L_{SAL}. We also make use of the semantic information given by a text encoder by guiding the vision encoder with rich captions of these images extracted from the structured annotations, and by guiding the meta-embeddings using natural language.

3.3 Deep Metric Learning framework and baselines

In this section, we will first introduce the reader to the common Deep Metric Learning framework and then present the most used baselines that will be compared in section 3.5.

3.3.1 DML Framework

Metric Learning makes use of distances to mirror similarity measures between data points, such that similar data will be closer with respect to that distance, and dissimilar data will be further apart. One of the earliest examples of that is the famous nearest neighbours classifier (Cover et Hart [27]).

In particular, Deep Metric Learning consists in learning a distance metric $d(x_i, x_j)$ over data points $x_i \in X$, with X being the training dataset of images. This distance metric d is parametrized by a deep feature extraction model $\phi : X \to \Phi$ followed by a linear projection to the target metric space $f : \Phi \to \Psi \subset \mathbb{R}^d$. In general, Ψ is normalized to the unit hypersphere, for better regularization. As opposed to Metric Learning, which usually builds on fixed feature extraction methods, DML learns an end-to-end embedding function $\psi = f \circ \phi$ that will bring latent representations of similar images closer, and push the latent representations of dissimilar images further apart. This ψ function is updated such that the distances $d(\psi(x_i), \psi(x_j))$ matches the semantic similarity between both data samples.

The two main ways to characterize the ground-truth similarity are usually either categorical classes $y_i \in [\![1, C]\!]$ with C being the number of classes in the training set, or using binary pairwise relations. However, in most works, the binary pairwise relations are extracted from the categorical classes in the first place, with $S(x_i, x_j) = \mathbb{1}_{y_i = y_j}$. Thus, we start our DML framework using labels as the ground-truth and the basis for our image similarity. Different DML baselines composed of training losses and tuple selection heuristics are presented in 3.3.2.

3.3.2 DML Baselines

We present here a brief overview of some of the most commonly used DML methods to better introduce the domain to the readers. We first present the softmax, then the contrastive and triplet losses, then the margin and multisimilarity losses, and finally the CLIP method.

3.3.2.1 Softmax

The softmax function, first introduced by Bridle [15], maps a vector of real values into a vector of probabilities:

$$\text{Softmax}(z)_i = \frac{e^{z_i}}{\sum_{j=1}^{C} e^{z_j}} \tag{3.1}$$

This function is mostly used as the last activation function in a neural network model for classification. It is built to output a vector of C values, each of which roughly representing a probability for each possible class. Even though it is mostly used for classification tasks, it can easily be adapted to a DML framework. To do so, one can learn the model with the softmax loss on all training data, and use the final embedding (the input to the softmax function) as the embedding representation during inference. This can be efficient because the layers prior to the softmax activation have learned discriminative features. This method is particularly effective when there is a large sample-to-class setup, and while it manages to separate training classes very well, it has trouble generalizing to new classes, or struggles when the number of classes increases greatly.

As it was discussed earlier in section 3.1, novel research has built upon the softmax function and focused on providing a better separation between classes. These works have concentrated on problems with large numbers of classes, such as face verification or person re-identification for example [31, 59, 102, 103].

3.3.2.2 Contrastive and Triplet Losses

Early research works such as Bromley *et al.* [16] or Hadsell *et al.* [45] have introduced a new learning paradigm to improve the latent representations given by deep architectures with siamese networks and a contrastive loss. The idea behind this method is to pull positive pairs of examples closer together, and push dissimilar pairs of examples further than a certain margin. To do so, both examples are passed through *Siamese Networks* that have shared weights, and their representations are then compared with a contrastive loss. We can write the contrastive loss for examples (x_i, y_i) and (x_j, y_j):

$$L_{contrastive}(x_i, y_i, x_j, y_j) = \mathbb{1}_{i=j} D_{i,j}^2 + \mathbb{1}_{i \neq j} max(margin - D_{i,j}, 0)^2 \tag{3.2}$$

with $D_{i,j} = \|f(x_i) - f(x_j)\|$ for a euclidian distance, with f being an end-to-end embedding function, and with $f(x_i)$ usually a normalized embedding vector.

This loss uses a margin to separate class clusters. However, Schroff *et al.* [79] argue that a contrastive loss comparing pairs of samples only encourages all examples of one class to be projected onto a single point in the latent representation space, without forcing a distance between the clusters of different classes. They therefore build on the contrastive loss for (anchor, positive, negative) triplets instead of (positive, negative) pairs, as it is illustrated in 3.2. The advantage of this method is that it enforces a margin between class clusters, but allows a certain lenience towards embeddings of the same class. For a triplet (x_a, x_p, x_n), we can write the triplet loss as:

$$L_{triplet} = \sum_{i}^{N} max(\|f(x_a) - f(x_p)\|_2^2 - \|f(x_a) - f(x_n)\|_2^2 + \alpha, 0) \tag{3.3}$$

In order to satisfy the constraint:

$$\|f(x_a) - f(x_p)\|_2^2 + \alpha < \|f(x_a) - f(x_n)\|_2^2 \tag{3.4}$$

for all triplets in the training set.

3.3.2.3 Distance weighted tuple sampling

Many DML frameworks and losses such as the contrastive and triplet loss work comparing tuples, and a lot of research has been done to optimize the tuple selection methods (see 3.2.1.3). In this work, we decide to follow a tuple selection heuristic presented by Wu *et al.* [105] called *distance weighted sampling*. The rationale behind this method is that, given a certain anchor, sampling negative examples randomly would not work because it would induce no loss for most examples, while sampling negative examples that are too hard would yield a high variance for the gradient. Therefore, the idea is to sample negative examples uniformly according to the distance, which leads to negative examples that are scattered instead of amassed in a small specific region.

Figure 3.4 shows how distance weighted sampling offers negative samples that have different distances whereas the other sampling methods propose examples that are biased towards certain clusters of distances.

3.3.2.4 Margin Loss and Multisimilarity

Wu *et al.* [105] introduce a *margin loss* that better fits the distance weighted tuple sampling method. This loss can be written as:

$$L_{margin}(x_i, x_j, y_{i,j}) = max(\alpha + y_{i,j}(\|f(x_i) - f(x_j)\| - \beta), 0) \tag{3.5}$$

with $\alpha, \beta \in \mathbb{R}$ and $y_{i,j} = 1$ if examples x_i and x_j share the same class, and $y_{i,j} = -1$ otherwise.

In this loss, α serves as a margin of separation in the same way as in the triplet loss, while β serves as a boundary between the positive and negative pairs.

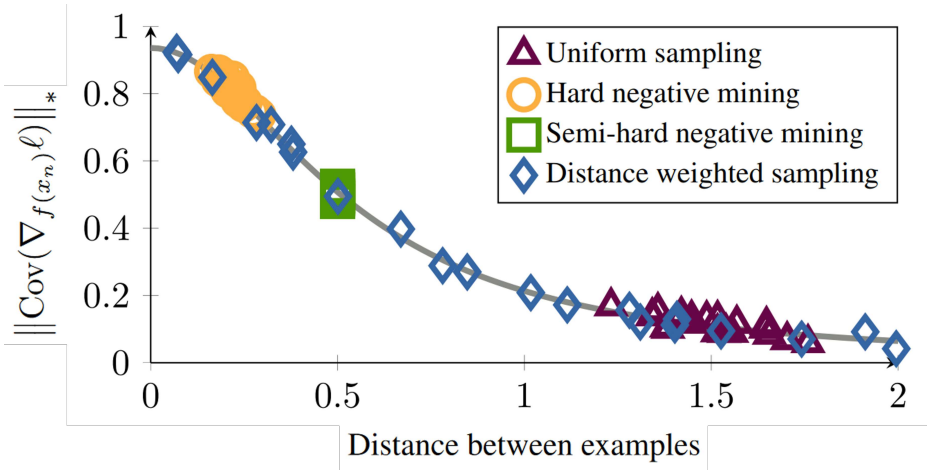


Figure 3.4: Figure that depicts the empirical distribution of samples drawn from different strategies. In particular, it shows the distance between examples with respect to the the norm of the covariance of the gradients. This figure is drawn from Wu *et al.* [105]

In the very popular *Multi-Similarity* work, Wang *et al.* [104] argue that in all DML frameworks working with tuples, negative pairs are usually weighted by three kinds of similarities: the *self-similarity (Sim-S)*, precisely the similarity between the positive and negative example, the *positive relative similarity (Sim-P)*, meaning the relative similarity of the negative pair with other positive pairs (which is done in the triplet loss), and *negative relative similarity (Sim-N)* which weights the negative pair relatively to other negative pairs. This is illustrated in Figure 3.5. They use the *Sim-P* to mine informative pairs (e.g. a negative pair is selected if its similarity is higher than the *hardest* positive pair for the same anchor) and integrate *Sim-S* and *Sim-N* to weight these examples in the multisimilarity loss. The weight of a negative pair is computed as such:

$$w_{i,j}^{-} = \frac{e^{\beta(S_{i,j}-\lambda)}}{1+\sum_{k\in N_i} e^{\beta(S_{i,k}-\lambda)}} \tag{3.6}$$

with β, λ and tuneable hyper-parameters. This negative pair weighting compares the Sim-S $e^{\beta(S_{i,j}-\lambda)}$ with the other negative similarities, therefore Sim-N, with $e^{\beta(S_{i,k}-\lambda)}$. In other words, the importance of a certain negative sample in the loss is larger if its distance with the anchor is smaller than the distance of positiv examples to the same anchor.

Finally, the multisimilarity loss can be written as:

$$\begin{aligned} L_{MS} = \frac{1}{B}\sum_{i=1}^{B}[\frac{1}{\alpha}\log(1+\sum_{k\in P_i} e^{-\alpha(S_{i,k}-\lambda)})] \\ +[\frac{1}{\beta}\log(1+\sum_{k\in N_i} e^{\beta(S_{i,k}-\lambda)})] \end{aligned} \tag{3.7}$$

Both these methods have proven their efficiency with great results on most common DML datasets.

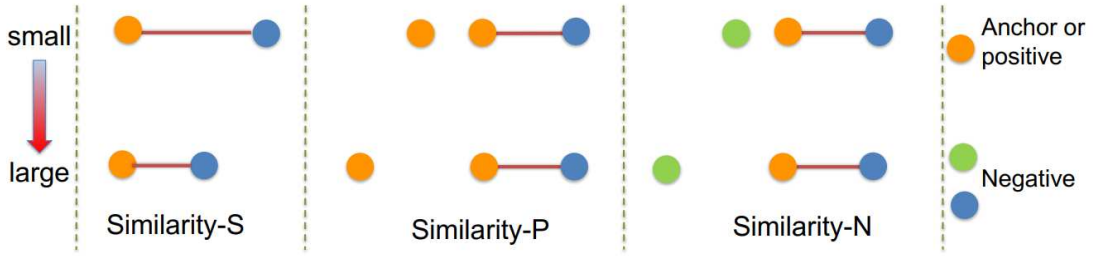


Figure 3.5: This figure illustrates the concepts of Sim-S, Sim-P and Sim-N. It is taken from Wang *et al.* [104]. Sim-S denotes the similarity between the anchor and the negative sample, Sim-P denotes a relative similarity with positive examples. In other words, for each anchor, a negative sample is more important if its distance with the anchor is smaller than the positive examples. Alternatively, Sim-N denotes the importance of a negative pair with respect to other negative pairs. If the negative sample is closer to the anchor than the other negative samples, it is considered more relevant in the final loss.

3.3.2.5 CLIP

Radford *et al.* [71] decide to train computer vision models with natural language supervision and introduce a method named CLIP. At the heart of this work is the idea that the semantic information comprised in the textual data can be leveraged easily (because textual models do not require large annotation resources and are learnt from large corpuses of text from the internet) and can be a very powerful tool in order to guide the training of visual models. The idea of this method is to match both corresponding textual and image latent representations. In theory, the visual encoder should benefit from the semantic context inherent to the textual data and the expressivity of the language model pretrained on millions of data from large-scale web-scrapped datasets. Although this work was originally intended zero-shot predictions, it can easily be used to adapt to a DML pipeline, comparing visual representations during inference rather than image-pair embeddings.

In practice, to implement CLIP, models take (image, text) sample pairs as inputs rather than the more common (image, label) pair. The method then compares the text embedding created by a text encoder with an image embedding created by a image encoder. Using large mini-batches, they simultaneously update the weights for both encoders using a double cross-entropy loss on both axes of the similarity matrix between the text and image embeddings. The categorical cross-entropy loss (CCE) can be written :

$$L_{CCE} = -\sum_{i=1}^{N} y_i \cdot \log(f(x_i)) \tag{3.8}$$

Although this method has shown very high performance results on many computer vision tasks, it requires a large training dataset (i.e. CLIP was trained on 400 million pairs of images and text). This can be difficult to reproduce on smaller DML datasets.

3.3.2.6 Language Guidance

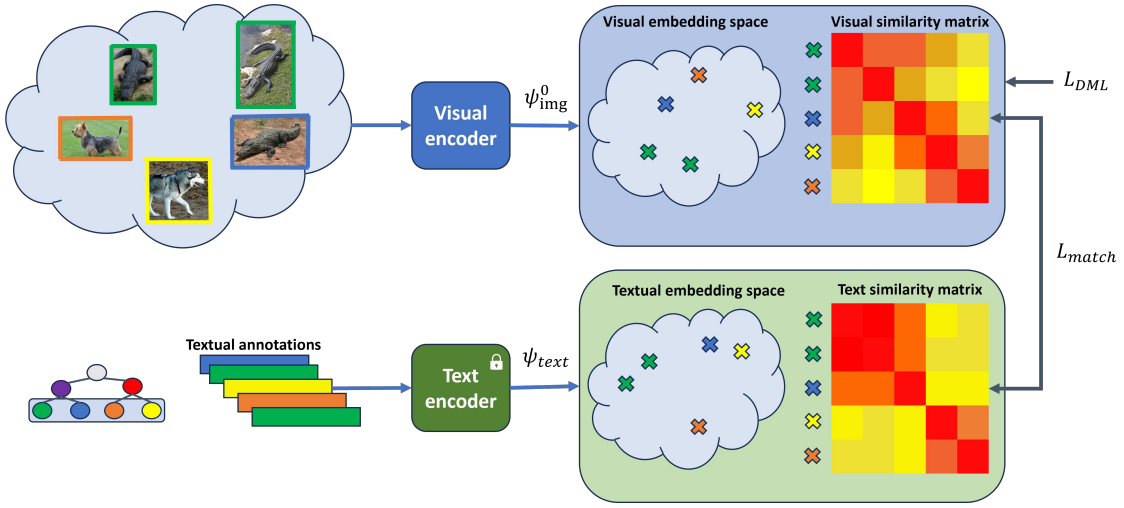


Figure 3.6: Overview of the Expert Language Guidance (ELG) method presented by Roth *et al.* [75]. A dual encoder embeds both the mini-batch of images and the mini-batch of captions associated to these images. While the visual encoder is updated by a classic DML loss, it is also guided by L_{match}, a loss designed to make the visual similarity matrix match the textual similarity matrix for each mini-batch. The pre-trained text encoder is frozen and aims to add context to the visual encoder.

Another way to integrate language guidance in the frame of visual similarity learning without needing extremely large datasets and enormous computing resources was introduced by Roth *et al.* [75]. In this paper, the authors still use a dual encoder, however, the text encoder is frozen to only update the image representations. They leverage the strong semantic information learnt by the text encoder and guide the image similarity matrix S_{img} to resemble the text similarity matrix S_{text}. To do that, they introduce a new matching loss based on the KL-divergence :

$$L_{match}(S_{img}, S_{text}) = \frac{1}{B}\sum_{i}^{B} \sigma(S_{img}) \log(\frac{\sigma(S_{img})}{\sigma(S_{text})}) \tag{3.9}$$

with B the batch size and σ a row-wise softmax.

The final loss term is a weighted average of L_{match} and a classic DML loss that ensures to learn a sound representation space. It is named *Expert Language Guidance (ELG)* and is written as :

$$L_{ELG} = L_{DML} + w \cdot L_{match} \tag{3.10}$$

The methodology is presented in Figure 3.6. This method exploits textual information by matching the visual similarity matrix with the textual similarity matrix. It therefore heavily relies on the language models, that can be very powerful nowadays in most use cases.

However, these methods can have limits when applied to some specific niche contexts such as medical imaging where the language models do not perform as well because it isn't as present in the web-scrapped resources used to train these models. In the next section we overcome these limitations by introducing a *Semantic Abstraction Loss (SAL)* that uses hierarchical annotations as ground-truth labels. We also introduce a novel way to guide these meta-embeddings with rich textual annotations, both through rich captioning and language guidance.

3.4 Exploiting rich semantic annotations for Deep Metric Learning

In naive DML setups, where all classes are considered exclusive, inter-class distances are not encoded optimally. To improve the model's performance in that regard, we propose in 3.4.1 a novel loss L_{SAL} that integrates annotations from multiple abstraction levels as ground-truth labels. In 3.4.2, we introduce several ways of integrating the rich structured annotations as textual information, both through rich captioning and through higher-order ontology language guidance.

3.4.1 Leveraging structured annotations for image similarity

Let x_i denote an image, y_i^l the class label associated to the image at the l-th depth of the class hierarchy (with y_i^0 being the leaf, and natural class annotation). For instance, the class hierarchy represented in Figure 3.7 would yield y_i^0 as *African crocodile*, y_i^1 as *reptile* and y_i^2 as *animal*.

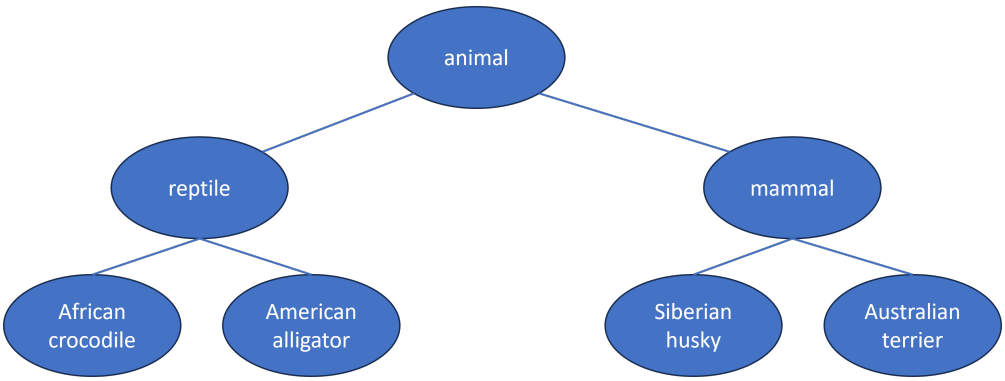


Figure 3.7: An example of a possible ontology for ImageNet classes "African crocodile", "American alligator", "Siberian husky" and "Australian terrier".

To improve the latent representations of the input images and obtain a better organized embedding space, we would like for semantic information extracted from higher-order annotations to be encrypted in the image encoder. In order to obtain these semantically

rich embeddings, we want to supervise the training of a vision encoder using classes and meta-classes.

The main idea introduced here is to create meta-embeddings that encode the information issued from the meta-classes. In the general case, the end-to-end image encoder ψ_{img} is created as a composition of derivable *layers*, and particularly as the composition of a common feature extractor f and a linear projection ϕ. It can be written as:

$$\psi^0_{img}(x_i) = f \circ \phi^0(x_i) \tag{3.11}$$

We present auxiliary *meta-embeddings* $\psi^l_{img}(x_i)$ output by *meta-projections* ϕ^l built on a single common feature extractor. They can be written as:

$$\psi^l_{img}(x_i) = f \circ \phi^l(x_i) \tag{3.12}$$

These auxiliary embeddings encode the semantic information carried by the meta-annotations. To push the feature extraction embedding space to reflect these semantic relations, we introduce a novel loss L_{SAL} written as:

$$L_{SAL} = \sum_{l=0}^{L} \alpha_l \cdot L_{DML}(\psi^l(x_i), y^l_i) \tag{3.13}$$

with α_l the weight associated to each abstraction level l in the loss and L_{DML} the DML loss (i.e. triplet loss, margin loss or multi-similarity loss for example).

This loss enables the feature extractor f to learn features that are able to discriminate all meta-classes, and therefore semantic information stemming from the rich class ontology. This way, the inter-class distances can be represented better in the embedding space. The method overview is illustrated in Figure 3.8.

To evaluate this method, one only needs the leaf-level *meta embeddings* $(\psi^0_{img}(x_i))$, and therefore does not necessitate any additional information during inference. This method also presents the advantage of being generic and working with different encoder architectures and different DML losses.

3.4.2 Integrating language information

The rich information that arises from the SUOG ontology comes in multiple forms. We presented in the previous section how to integrate the structured nature of the annotations to improve the predictive capacities of the visual similarity model. Another way to make the most of this ontology is to use the strong and rich textual annotations to guide the visual encoder. For instance, the textual caption *magnified view of the gestational sac* holds a strong semantic meaning in itself that could be extracted by rich language models.

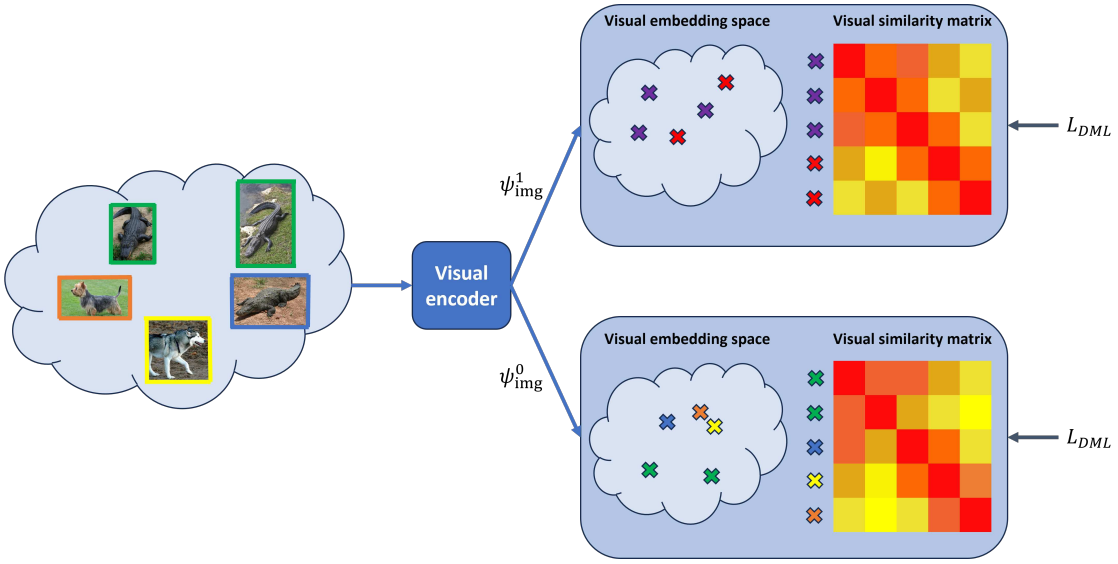


Figure 3.8: Overview of the proposed approach. A mini-batch of images with two levels of annotations y^0 and y^1 is passed through a vision encoder, and through two projections ψ^0 and ψ^1. For each image, these two embeddings are compared to their ground-truth annotations y^0 and y^1 respectively using a DML loss L_{DML}.

3.4.2.1 Rich Captioning

The first idea that came to mind was to integrate the information that comes from the class hierarchy in a textual manner. For both works that use language guidance and that are tested in this work, such as CLIP [71] presented in 3.3.2.5 and Language Guidance [75] presented in 3.3.2.6, we enrich the textual input with hierarchical information.

For instance, for the CUB-200 dataset [100], a bird classification dataset used in this work (that present more extensively in section 3.5, we have three levels of annotation hierarchy, with species, genus and family. The textual primer for example x_i becomes *"a photo of a y_i^0 from the genus y_i^1 and the y_i^2 family."*. For instance, for an image of a baltimore oriole, the textual primer used by the model for language guidance would be changed from *"a photo of a Baltimore Oriole"* to *"a photo of a Baltimore Oriole from the genus Icteridae and the Icterus family"*. For the SUOG view dataset, where there is only one level of label hierarchy, the additional text for the sample x_i becomes *"y_i^0 from the y_i^1"*. This method has the advantage of offering a certain liberty towards the primer, allowing anyone to change the primer to better fit the domain task. However, it also puts all the semantic abstraction information at the same level. The information that stems from the species, genus and family are put together in a single caption, and one of these might overshadow the representation of the two others.

3.4.2.2 Language guidance over meta embeddings

To optimize the increment in visual similarity learning brought by the *meta embedding learning* presented in 3.4.1, we introduce *Ontology Language Guidance* (OLG), where we apply the language guidance loss introduced by Roth *et al.* [75] and presented in 3.3.2.6 to the aforementioned *meta embeddings*. We therefore obtain :

$$L_{OLG} = \sum_{l=0}^{L} \alpha_l \cdot L_{DML}(\psi^l(x_i), y_i^l) + w_l \cdot L_{match}(S_{img}^l, S_{text}^l) \tag{3.14}$$

This loss allows to integrate the rich annotations, both using the multiple levels of class hierarchy, while also guiding the representations to be sound using a strong textual guidance. Compared to the rich captioning, OLG also enables the model to better benefit from the hierarchical nature of the annotations by separating the representations of the different levels of abstraction of the data. The overview of this method is presented in Figure 3.9.

3.4.2.3 Making use of a domain specific language encoder

The above-mentioned extensive loss term heavily relies on the good semantic representations of the textual encoder. These tend to be good enough for large-scale tasks such as object classification, where a lot of classes are relatively common in the web-scrapped text datasets that have served for the text encoder pre-training. However, and it was pointed out as a limitation for this work in [75], the text encoders might have more trouble working with very specific domains, and might typically struggle to separate OB/GYN medical terms for instance, where the class separation can be too fine-grained.

We therefore decide to switch the text encoder to a domain specific one, where the class separations will be broader. Lee *et al.* [55] build BioBERT on a pre-trained BERT model that is updated on a biomedical corpus from PubMed and then fine-tuned for biomedical Named Entity Recognition, Relation Extraction and Question Answering.

Adapting the textual encoder to a domain specific one such as BioBERT improves the predictive capacity of the model, as is shown in 3.5, and demonstrates that semantic relations extracted from the textual corpora are highly helpful for visual similarity. However, using domain specific text encoders for language guidance can sometimes be detrimental to the visual task at hand. In practice, and as it is highlighted in the experiments of this chapter, a text encoder trained on the same domain is not sufficient to improve the semantic context, one needs to take into account the task that the encoder has been trained on.

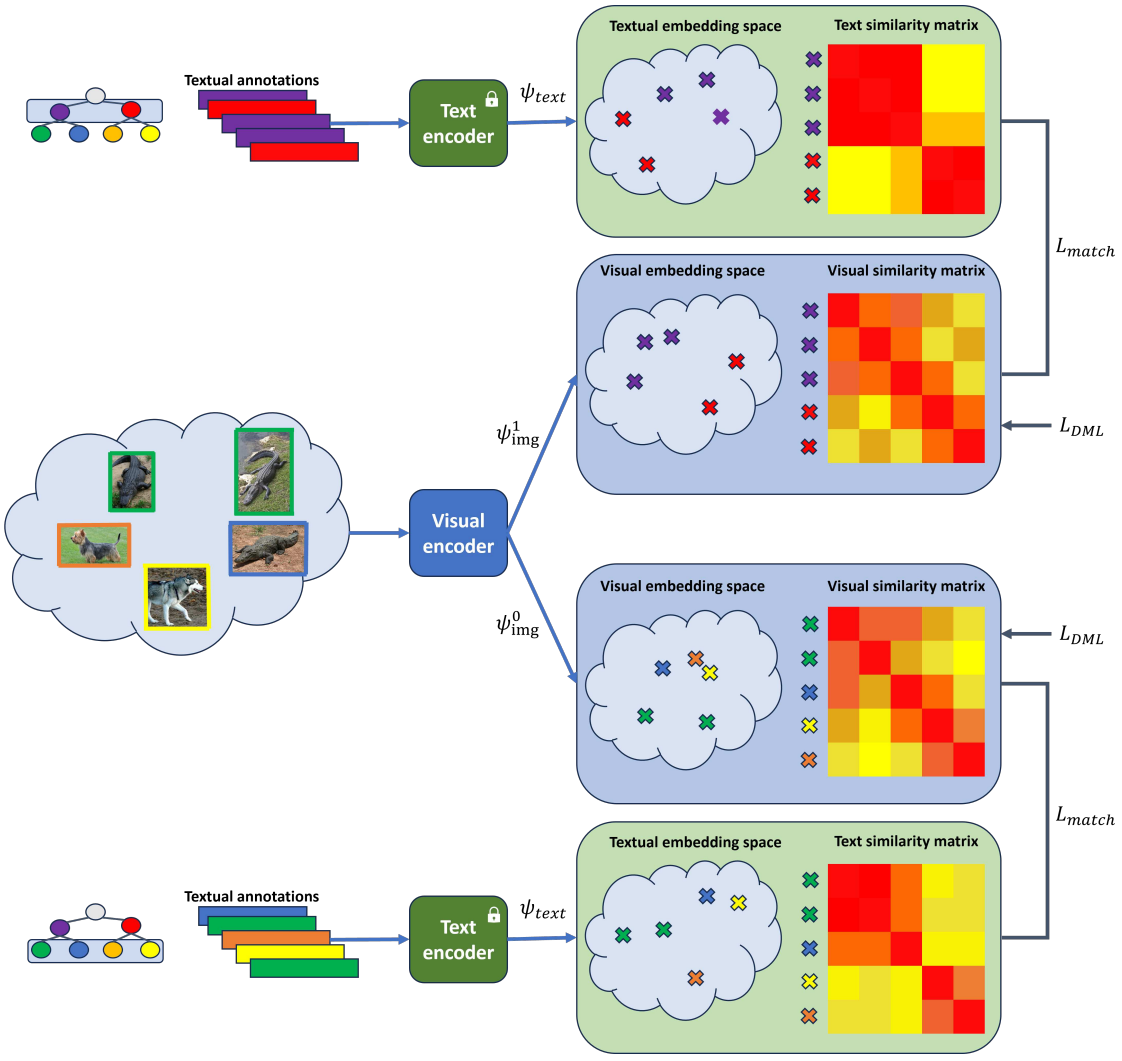


Figure 3.9: Overview of the Ontology Language Guidance (OLG) method. While a vision encoder takes a mini-batch of images as input and outputs two visual embeddings, a frozen text encoder takes two mini-batches of captions stemming from y^0 and y^1 (e.g. the hierarchical annotations) as input and outputs two textual embeddings. A DML loss is applied to both visual embeddings and a matching loss L_{match} is applied to both pairs of similarity matrices in order to help guide the vision encoder with the context encoded by the text model. This example only depicts two levels of hierarchical annotations but it can be generalized to more.

3.5 Experiments

In this section, we evaluate the interest of guiding a visual similarity framework with higher-level semantic information. First, we broadly present the tasks, datasets and implementation details in Section 3.5.1, then we present how higher-order annotations can improve the predictive capacities of an image similarity model in section 3.5.2 and finally the impact of language guidance in Section 3.5.3.

3.5.1 Tasks and Datasets

3.5.1.1 Datasets

We present results to experiments led on two datasets: the birds classification dataset CUB-200 [100] and the scan plane recognition SUOG dataset. CUB-200 is a public dataset, less noisy than SUOG and where classes are more easily separable for a visual model pretrained on ImageNet. Let us present both of them:

3.5.1.1.1 CUB-200 CUB-200 is a dataset very widely used for Metric Learning. It contains 11788 images of birds belonging to 200 different species as a ground-truth class. The first row of Figure 3.10 highlights examples from CUB-200. We manually extract the *genus* and *family* to which these species belong from the Avibase world bird database (https://avibase.bsc-eoc.org/avibase.jsp) to create higher-level annotations. The training set contains all images from 100 species (with 69 different *genuses* and 28 *families*), while the test set contains all images from the other 100 species. This allows us to evaluate the visual similarity performances of the model on data samples and classes that haven't been seen during training.

3.5.1.1.2 SUOG The SUOG project collected 200K ultrasound images at all stages of pregnancy from ten different expert centres across Europe. The subset of images used in these experiments contain 4323 pregnancy ultrasound images, with 649 used in the test set, randomly sampled to follow the same label distribution as the train set. The second row from Figure 3.10 highlights examples from SUOG. We use the *view* annotations as the ground-truth label to perform DML on. There are 18 classes that all belong to a set of 5 *metaclasses*, all extracted from the SUOG ontology created by OB/GYN experts. These classes and metaclasses are shown in Table 3.1.

3.5.1.2 Implementation Details

For most experiments, we use a ResNet50 pretrained on the ImageNet dataset (Deng *et al.* [30]) as the image encoder, while different text encoders have been tested. We set the final embedding size of the image encoder to 128 as it is the case in most state-of-the-art works to ensure a fair comparison. For all trainings, we used ADAM optimizer with

Meta Label	Label
adnexal view	transverse view of right adnexa
	transverse view of left adnexa
	longitudinal view of right adnexa
	longitudinal view of left adnexa
magnified view of the gestational sac	magnified view of the gestational sac
	longitudinal view of the embryo
	magnified view of the trophoblast
uterus view	longitudinal view of the uterus
	oblique view of the uterus
	longitudinal view of cervix
	transverse view on uterus
	left interstitial portion view
	right interstitial portion view
	right tubal interstitial portion
	left tubal interstitial portion
bladder and vesicouterine fold view	longitudinal view of bladder
pouch of Douglas view	transverse view of Pouch of Douglas
	longitudinal view of Pouch of Douglas

Table 3.1: Groups of labels extracted from the SUOG ontology. Another larger illustration of these annotations can be found in Figure 1.2.

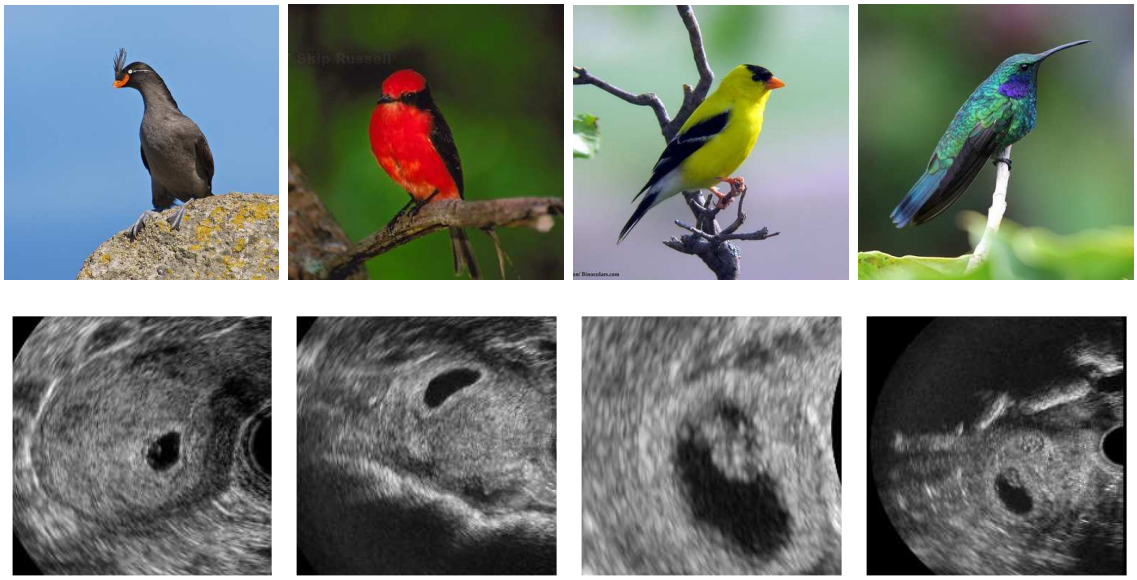


Figure 3.10: Examples from the CUB-200 dataset are shown in the first row. Examples from the SUOG dataset are shown in the second row.

β_1=0.9 and β_2=0.999, and a batch size of 64 unless stated otherwise. For the CUB-200 dataset, we used the common data augmentations used in the state-of-the-art methods: the training images are randomly cropped while keeping the same aspect ratio and then are randomly flipped. The test images are cropped and centred. For the SUOG images, only a simple random vertical flip is applied as it allows keeping the ultrasound imaging structure.

For all the CLIP experiments, we update multiple details to effectively train on relatively small datasets. We add a triplet loss term to regularize the training, and work with pretrained vision (ResNet50 pretrained on ImageNet) and text encoders (BERT-small model) instead of learning everything from scratch as in the original paper. We also replace the original CCE by a binary cross-entropy loss (BCE). Using a CCE with large mini-batches is acceptable in the original paper because of the extremely large number of classes, whereas we only have 100 classes for CUB-200 and 18 for SUOG. This allows each image to match several textual inputs (e.g. several examples of the same class in the same mini-batch).

3.5.1.3 Evaluation Metrics

In order to evaluate DML models, classic metrics such as *accuracy* or *F1-score* aren't optimal because these models output latent representations and not class predictions so to speak.
In most state-of-the-art DML works, the main evaluation metric used is *recall@k*. Unlike what the name suggests, *recall@k* does not share the same concept as the *recall* (also

known as sensitivity) evaluation metric, which counts the number of true positives over all positive samples. *Recall@k* equals to 1 when at least one of the K nearest neighbours of one specific query sample is of the same class as that sample, and 0 otherwise. We also introduce *meta-recall@k*, following the same mechanism as *recall@k* but using the meta labels as the ground truth. This enables a coarser assessment, and more globally evaluates the capacity of the model to make "sounder" mistakes.

Table 3.2: Results for the CLIP model on the CUB-200 dataset

Method	Recall@1	std
CLIP	44.98	0.34
CLIP rich text	45.48	0.37
CLIP + Genus-SAL	46.63	0.2
CLIP + Family-SAL	46.57	0.23
CLIP + Genus-SAL + Family-SAL	**47.00**	0.4

3.5.2 Guiding the Metric Learning with prior meta annotations

In this section we validate the impact of the proposed *Semantic Abstraction Loss (SAL)* that leverages hierarchical annotations to better characterize the embedding space.

We can observe in Table 3.2 that performing visual similarity at metaclass level allows the CLIP model to improve its predictive performances with regard to recall@1 on the CUB-200 dataset. The results go from 44.98% when trained with a CLIP model, to 46.63% when *SAL* takes into account the genus (y^1) of the training sample, 46.57% when taking the family into account (y^2) and finally 47.00% when taking both into account, which demonstrates that the integration of semantic context leads to better predictions. This accounts for a 2.02 points global increase, while we can also note that the use of two different levels of hierarchical annotations works better than only using one. This could demonstrate that even stronger semantic information leads to better representations. Alternatively, Table 3.3 shows that applying the novel *SAL* loss to the multisimilarity [104] DML loss offers a very short increase in performance (+0.09 points when guiding with the family classes). The small improvement on the multisimilarity loss could be explained by the fact that the CUB-200 dataset contains a relatively large number of images and classes. We will see that *SAL* offers a considerable increment on multiple DML losses for the SUOG dataset, that has less images and less classes.

Indeed, Table 3.4 shows the impact of the meta-loss *SAL* for the SUOG dataset. It shows that, for five different DML losses, the addition of the *SAL* loss term consistently improves the *recall@k* scores (+2.34 points for CLIP, +1.95 points for the triplet loss, +1.58 points for the softmax, +2.69 points for the margin loss and +0.50 points for the

Table 3.3: Results for the Multism. model on the CUB-200 dataset

Method	Recall@1	Genus-Recall@1	Family-Recall@1
Multism.	63.41 ± 0.45	71.64 ± 0.43	86.16 ± 0.22
Multism. + Genus-SAL	63.36 ± 0.19	**71.93 ± 0.26**	86.38 ± 0.28
Multism. + Family-SAL	**63.50 ± 0.34**	71.73 ± 0.50	86.21 ± 0.27
Multism. + Both-SAL	63.22 ± 0.16	71.61 ± 0.30	**86.43 ± 0.35**

multisimilarity loss). This shows that the integration of a single level hierarchical semantic information greatly benefits the model to separate echographic views that can be highly similar visually. It also demonstrates that the method is generic to all DML losses in this use case, with the *recall@1* scores always higher than the baseline when the weighting of *SAL* is well-chosen.

Table 3.4: Results for SAL on the SUOG dataset.

DML method	α	Recall@1	Meta-Recall@1
CLIP	0	36.83 ± 1.58	81.97 ± 1.56
CLIP + SAL	0.1	37.39 ± 1.05	**82.37 ± 0.34**
CLIP + SAL	0.5	**39.17 ± 1.05**	82.16 ± 1.44
CLIP + SAL	1	38.84 ± 1.13	81.97 ± 0.77
Triplet	0	52.32 ± 1.31	90.91 ± 0.70
Triplet + SAL	0.1	**54.26 ± 1.14**	91.28 ± 0.58
Triplet + SAL	0.5	**54.27 ± 0.38**	91.65 ± 0.40
Triplet + SAL	1	**54.27 ± 0.69**	**91.83 ± 0.50**
Softmax	0	55.71 ± 0.85	91.74 ± 0.73
Softmax + SAL	0.1	**57.29 ± 0.93**	92.17 ± 0.38
Softmax + SAL	0.5	56.73 ± 0.65	**92.35 ± 0.45**
Softmax + SAL	1	54.73 ± 0.74	92.29 ± 0.83
Margin loss	0	53.19 ± 0.95	89.30 ± 0.90
Margin loss + SAL	0.1	**55.9 ± 0.88**	91.19 ± 0.65
Margin loss + SAL	0.5	54.3 ± 0.65	**92.02 ± 0.87**
Margin loss + SAL	1	52.36 ± 0.49	91.96 ± 0.34
Multisimilarity	0	56.16 ± 0.81	90.45 ± 0.83
Multisimilarity + SAL	0.1	**56.66 ± 1.00**	91.44 ± 0.55
Multisimilarity + SAL	0.5	54.85 ± 0.50	**91.96 ± 0.34**
Multisimilarity + SAL	1	54.52 ± 1.19	91.89 ± 0.66

Another important aspect of the method to point out is that it allows the model to make "better" mistakes. Table 3.3 shows results of experiments done on the CUB-200

dataset, and indicates that the model trained with *SAL* obtains better results in terms of genus-recall@1 and family-recall@1 than the baseline. This means that when the nearest neighbours of the query don't share the same class as the query sample, most of them still share the same genus or family.

We can also observe in Table 3.4 that for the SUOG dataset, all five DML methods, when coupled with *SAL*, perform better than the baseline in terms of meta-recall@1 even when they have a lower recall@1 (at best, the input of the *SAL* loss term increases the meta-recall@1 by 0.99% for CLIP, 0.92% for the triplet loss, 0.61% for the softmax and 2.72% for the margin loss). For instance, while applying *SAL* with $\alpha = 1$ on the multisimilarity loss makes the recall@1 drop from 56.16 to 54.52, the meta-recall@1 still increases by 1.44%. Also, when the compromise is done between both losses by choosing a smaller and better adapted α at 0.1, both metrics are higher when the model is guided by rich annotations. This is important because it ensures that even when the model is not able to represent a test sample correctly, the most similar images in the latent embedding space are still relatively close semantically. In practice, for a query image labelled as a *"longitudinal view of right adnexa"*, it is preferable to present as similar an image labelled *"longitudinal view of left adnexa"* than an image labelled as *"magnified view of the trophoblast"*. This property is also interesting in cases such as scan plane recognition where the datasets are relatively small, because the model generalizes better as the embedding space is better organized.

One last important detail to highlight is the difference in results between the multisimilarity-trained model and the CLIP-trained one. Although multisimilarity obtains very positive results on the CUB-200 dataset, the *SAL* loss term only provides a very slight improvement in performance. The CLIP model, however, achieve a 2.02 point increase form the *SAL* loss term. This could be explained by the actual nature of the CLIP method, that aims to match the visual embeddings and the text embeddings through a cross-entropy loss (see section 3.3.2.5 for further details). This means that the multisimilarity model benefits from the semantic information through structured annotations and *meta*-embeddings, while the CLIP model also takes advantage of this information through natural language and textual representations too, and therefore probably profits from large text encoders trained on large-scale databases. Therefore, we explore the idea of integrating textual data so that the visual model can learn semantic relations from the textual data. In the next section, we present results from these experiments.

3.5.3 Integrating structured prior information through natural language

Language models have recently taken the deep learning world by storm because they manage to encode strong semantic information due to the large-scale training datasets available. We try to input this information in our deep metric learning framework through rich textual data, guiding auxiliary embeddings with natural language, and finally by using

a domain-specific text encoder.

3.5.3.1 Impact of rich textual data during language-guided learning

As leveraging natural language to improve visual deep learning models capacities has become common, we leverage the strong semantic information obtained through the rich annotations using textual representations.

We evaluate the impact of the *rich captioning* method on different datasets. To do so, we apply ELG to a classic DML method such as multisimilarity but update the textual input by introducing the higher-level annotations in the captions (i.e. the simple caption *"a photo of a sooty albatross"* would be changed to *"a photo of a sooty albatross from the genus diomedeidae from the family phoebastria"*). On the CUB-200 dataset, tables 3.2 and 3.5 show that using the *rich caption* method slightly improves the predictive performance. For the CLIP method, using a rich textual input improves the predictive performance by 0.5 points, whereas it improves the recall@1 by 0.14 points for multisimilarity. To understand the short improvement given by the rich captioning compared to the use of L_{SAL}, we decide to compare the embeddings given by the text encoder for the simple caption and the rich caption. Results found in Table 3.6 show that the mean cosine similarity (over all classes) between the embeddings of the rich caption and the simple caption for the species is very high (0.937), whereas it is much smaller for the genus and family (0.617 and 0.655 respectively). These results demonstrate that the text encoding for the rich captions is extremely similar to that of the simple captions and therefore doesn't fully capture the strong semantic information given by the higher-order annotations, but rather still focuses on the most precise terms. Also, the difference in performance between multisimilarity and CLIP can be explained by the fact that for CLIP, the textual representations are updated, whereas for the multisimilarity loss, the textual encoder is frozen.

However, on the SUOG dataset, we can observe a slight decrease in performance when applying the rich textual inputs instead of the simple ones in Table 3.7. This can be explained by the poor performances of the textual encoder on specific OB/GYN terms. In particular, we can observe in Table 3.6 that the rich representations are very similar to both the textual representation of the classes and to the meta-classes (the mean cosine similarity is respectively equal to 0.937 and 0.916). The main theory behind these results is that the frozen text encoder has very limited knowledge concerning early pregnancy and OB/GYN in general, and therefore embeds all the SUOG classes and meta-classes in a very tight and small space.

3.5.3.2 Guiding meta embeddings using natural language

As we demonstrated the interest of using the *SAL* loss term in a DML context, we hereby prove the impact of guiding these meta-embeddings using natural language.

Table 3.5: Ablation study of Multisim. and Language Guided method on CUB-200 dataset.

Method	ELG	Rich capt.	OLG	Recall@1
Multisim.	✗	✗	✗	63.41 ± 0.45
Multisim.	✓	✗	✗	67.19 ± 0.12
Multisim.	✓	✓	✗	67.33 ± 0.22
Multisim.	✓	✓	Genus+Family ($\alpha = 0.25$)	67.5 ± 0.33
Multisim.	✓	✓	Genus+Family ($\alpha = 0.5$)	**67.74 ± 0.30**
Multisim.	✓	✓	Genus+Family ($\alpha = 0.75$)	67.61 ± 0.26
Multisim.	✓	✓	Genus+Family ($\alpha = 1$)	66.92 ± 0.32
Multisim.	✓	✓	Genus ($\alpha = 1$)	66.92 ± 0.25
Multisim.	✓	✓	Family ($\alpha = 1$)	67.62 ± 0.35

Table 3.6: Mean cosine similarity between the embeddings of the rich caption and the simple caption with the CLIP encoder.

Dataset	Level-0 Similarity	Level-1 Similarity	Level-2 Similarity
CUB-200	0.937	0.617	0.655
SUOG	0.937	0.916	—

For the CUB-200 dataset, results in Table 3.5 show the impact of guiding the DML training with natural language, as presented by Roth *et al.* [75]. We can also observe that adding the OLG loss term helps improve the model's predictive performance, as it helps structure the embedding space and therefore improve the inter-class distances. When guiding the model using both genus and family annotations and language guidance, the model goes from 63.41% to 67.74% at best, the OLG loss term accounting for a 0.55 points increase compared to the model that used simple ELG language guidance.

Table 3.7 shows results for ELG language guidance and OLG on the SUOG dataset. We can observe that the simple ELG only very slightly improves the recall@1 results by 0.17 points when guided using a simple caption. As expected, adding a L_{SAL} loss term improves the recall@1 up to 56.49, with the meta-recall@1 increasing by 0.43 points, which confirms the results discussed previously, while guiding the model using OLG slightly improves the results, with a recall@1 climbing up to 56.61%. However, we can observe that the results with OLG are still slightly below those for the model only guided by L_{SAL}. These results are interesting because they show that language guidance is only realistically useful when the language model can offer good context to better separate the different classes. In the case of the SUOG dataset, it seems that the text encoder isn't powerful or knowledgeable enough to significantly improve the results on its own, and even slightly degrades the model's performance, as it is not able to encode the semantic context between the annotations.

Table 3.7: Ablation study on SUOG dataset for language guided meta-learning.

DML method	$SAL\ \alpha$	Rich capt.	ELG	OLG	r@1	meta-r@1
Multisim.	✗	—	✗	✗	56.16 ± 0.81	90.45 ± 0.83
Multisim.	✗	✗	✓	✗	56.33 ± 0.98	90.17 ± 0.54
Multisim.	✗	✓	✓	✗	56.21 ± 1.31	90.35 ± 0.89
Multisim.	✓	✗	✓	✗	56.49 ± 0.14	**90.88 ± 0.86**
Multisim.	✓	✗	✓	✓	**56.61 ± 0.77**	90.57 ± 0.79
Multisim.	✓	—	✗	✗	**56.66 ± 1.00**	**91.96 ± 0.34**

Table 3.8: Impact of the text encoder on language-guided DML.

DML method	Text encoder	recall@1
Multisimilarity	CLIP	56.33 ± 0.98
Multisimilarity	BioBERT	55.29 ± 0.62
Multisimilarity	LargeBioBERT	55.96 ± 0.88

We can therefore conclude that using textual information to guide a DML model is effective when the text encoder is knowledgeable about the information in the training data. We therefore investigate the interest of using a language representation model trained on specific data.

3.5.3.3 Using a domain specific text encoder

In order to better understand the results given by language guidance on the SUOG dataset, we decide to try different text encoders to potentially better encode the captions extracted from the rich annotations. Results shown in Table 3.7 show us that using a Transformer language model trained with CLIP to guide the training of a DML model only slightly improves the recall@1 results on the SUOG dataset. The main hypothesis behind this result is that the text encoder, trained on a very large object recognition dataset containing 400 million images and text captions, cannot provide sufficient insight on the SUOG classes to better separate them because of the domain gap. A language encoder trained on CLIP will therefore have trouble building coherent distances between annotations such as *magnified view of the gestational sac*, *transverse view of the right adnexa* and *transverse view of the Pouch of Douglas*.

We therefore try changing the text encoder to introduce a model trained on medical data, that could possibly help build better embeddings for the SUOG dataset captions. We therefore investigate both BioBERT [55] and BioBERT-Large, two language models trained for biomedical text mining on a large biomedical corpus extracted from PubMed, with the latter using a larger architecture than the former. However, we can see on Table 3.8 that the DML model guided by BioBERT and BioBERT-Large does not perform as

well as the one guided by the CLIP text encoder. Even though these results are surprising at first, Figure 3.11 demonstrates an explanation.

This figure shows the pairwise distance between the embeddings of all SUOG classes, with the groups belonging to the same metaclass are delimited by the blue lines. It shows that the CLIP text encoder (as opposed to what one could have thought initially) and the BioBERT-Large text encoder build embeddings that are closer relatively to the semantic hierarchy between classes than BioBERT. Even though both BioBERT and BioBERT-Large build relatively sound embeddings, the CLIP text encoder seems to concentrate on simpler things to separate the classes, bringing closer all embeddings of *adnexal views* for example, as the caption is very similar semantically. Another example of that is that the *transverse view of the Pouch of Douglas* is considered very close to *longitudinal view of the Pouch of Douglas* by the CLIP model and the BioBERT-Large model whereas the BioBERT model considers them further apart. This can be explained by the fact that BioBERT is trained on a large corpus of biomedical data for named entity recognition or relation extraction, and therefore would give more importance to the terms *longitudinal* and *transverse* rather than Pouch of Douglas, which is very specific to OB/GYN problematics.

Alternatively, both BioBERT and BioBERT-Large represent the adnexal views very close to several uterus views as well as interstitial portion views, as opposed to the CLIP model. These representations don't match the hierarchy given by the SUOG ontology, and this therefore confirms that the semantic relations given by the ontology (built by experts from 10 European centres) and leveraged by L_{SAL} is more accurate than that given by text encoders such as CLIP or BioBERT and BioBERT-Large.

In a nutshell, the integration of hierarchical annotations through a dedicated loss like L_{SAL} improves the vision encoder's predictive capacity, and using natural language to guide the visual similarity through rich captioning or OLG also significantly increases the results in terms of *recall@1* and *meta-recall@1* when the text encoders perform well on these input domains.

3.6 Conclusion

3.6.1 Discussion

In this chapter, we explored the integration of rich annotations to improve a deep metric learning framework. In particular, we started from the problem that classic DML and classification methods deem different class annotations as exclusive, and therefore treat any prediction other than the ground-truth annotation as equally wrong. We argue that, in most cases, that is a suboptimal way to formulate the deep learning problem, (e.g. it would consider the ImageNet class *American alligator* to be equally distant to a *saxophone* than it is to an *African crocodile*). Additionally, the semantic representation space for

SUOG images is structured, as all annotations are extracted from a large ontology created by experts from 10 hospitals across Europe. To alleviate these issues, we first leveraged the hierarchical annotations extracted from the class ontology by creating auxiliary *meta-embeddings* that are pushed to encode different levels of meta-annotations. To do so, we propose L_{SAL}, which is a weighted average of DML losses applied to the auxiliary embeddings. L_{SAL} enables the model to better encode inter-class relations, bringing closer samples from classes that share the same metaclass. This method improves the representation capacity of the image encoder by ensuring a better integration of the hierarchical semantic information. Second, we made use of the strong textual information linked to the annotations. We built our work on the language guidance method presented by Roth *et al.* [75], and propose a rich captioning method that integrates the hierarchical nature of the annotations in the caption used for language guidance. Finally, to be able to separate the impact of the multiple levels of annotations, we introduced *Ontology Language Guidance (OLG)*, a method that specifically guides the *meta-embeddings* using natural language. One positive takeaway from these methods is that while they require additional input information during training with the meta-annotations, they do not necessitate it during inference.

We validated the interest of L_{SAL} method on the CUB-200 dataset, achieving better results in terms of recall@1 and meta-recall@1, and on the SUOG dataset, where the addition of L_{SAL} consistently improves both metrics for multiple DML baselines. Then, the increment of rich captioning provided a limited improvement on both datasets, as the text encoders mainly focused on the leaf-level classes in the rich captions. The incorporation of *OLG* circumvented this issue by optimizing the impact of all levels of annotations. It proved to be efficient on the CUB-200 dataset where the text encoder managed to provide sufficient semantic context, but however did not improve the representation capacity of the model on the SUOG dataset, because the text encoders tested were not able to separate the classes and metaclasses well, due to a domain gap.

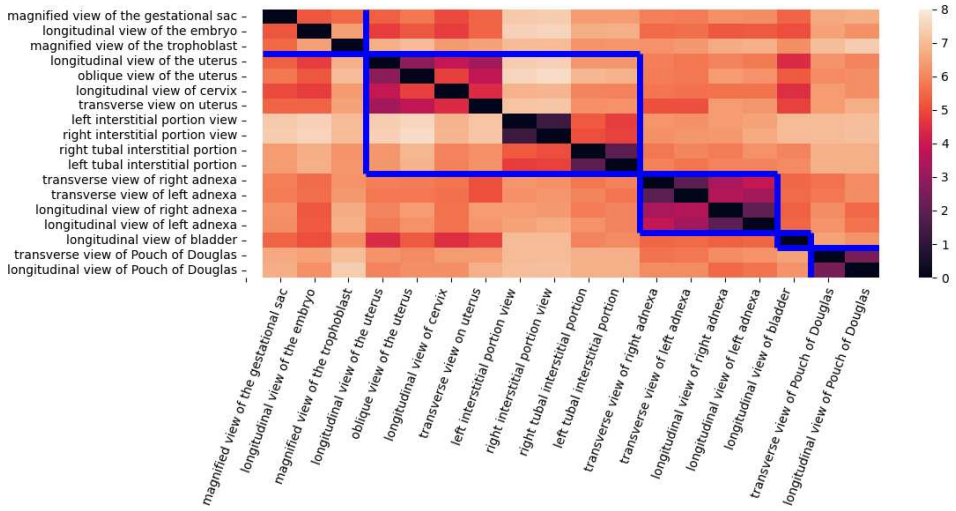
As a conclusion, it is interesting to note that guiding a DML model using rich annotations, whether it be through auxiliary embeddings with L_{SAL} or language guidance with rich captioning or OLG, attest to generally improve the representations given by the model. However, it is usually more interesting to use textual representations when the text encoder can provide semantic context in the input domain. This can be verified qualitatively by comparing distances between the embeddings of all classes or different visualization techniques. In parallel, using hierarchical annotations to learn auxiliary embeddings usually improves the performance of the visual encoder.

3.6.2 Future Work

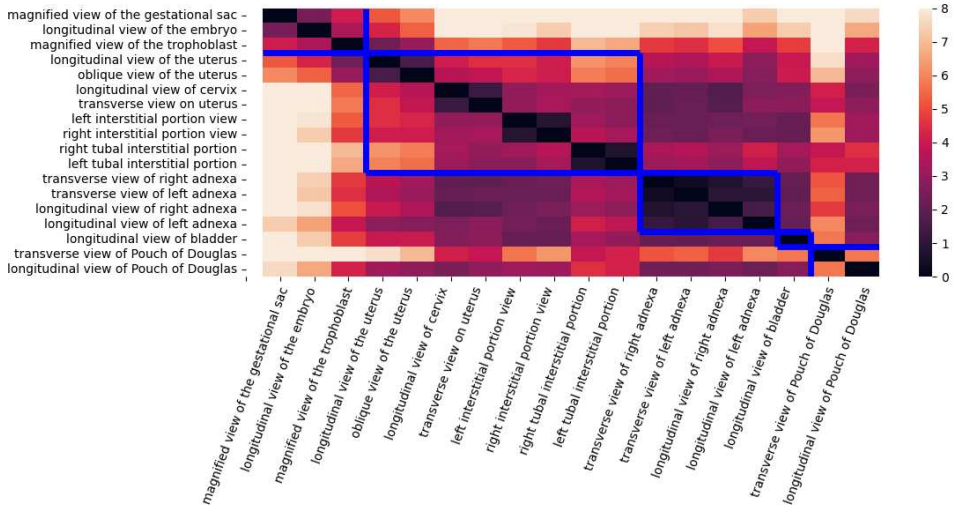
In the short term future, we would like to analyse the interest of the method on another OB/GYN task on the SUOG dataset. It would be interesting to see if the method brings forward positive results on DML using disorder annotations as the ground-truth labels.

Indeed, the current results show encouraging results on DML for scan plane recognition, which is an important task to help the ultrasound operator in real-time, but for the sonographer to be able to present similar images in terms of disorder might be a very important advance for the SUOG project. Similarly, we could investigate the idea of using the edges of the ontologies as more semantic information. In particular, we could use the causality links given by the ontology instead of only using the hierarchical links (i.e. *"unilocular cystic liver mass"* suggests *"fetal abdomen disorder"*) to improve the semantic context in the representations created by the deep learning model. Another way of integrating semantic knowledge into the model's representations is to use a similarity metric computed on the SUOG ontology. We could therefore use the metric created by Mirna El Ghosh for the SUOG project to guide the visual similarities and introduce graph-based semantic similarity.

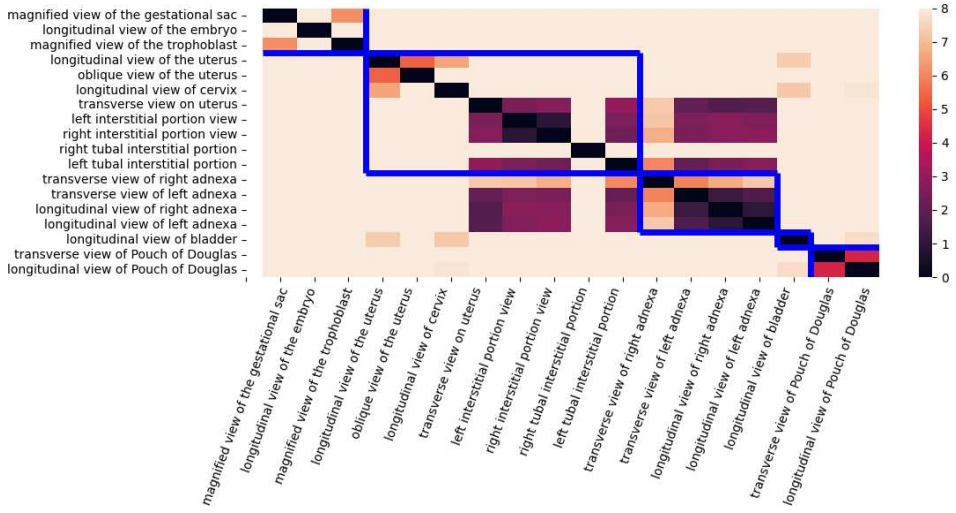
Another track we could investigate would be to update the architecture to better fit the hierarchical nature of the annotations. Works such as [111, 50] leverage ensemble methods to learn different levels of annotations, while others like [77, 66] divide the embedding layer into multiple subspace learners following online clustering. We could investigate mixing both methods, and divide the embedding layer using higher-level annotations. Then the multiple subspace learners would only need to concentrate on separating different classes from the same metaclass, without the extra parameters needed for ensemble methods.



CLIP



BioBERT



BioBERT-Large

Figure 3.11: Similarity matrices for the SUOG classes for three different text encoders: CLIP, BioBER and BioBERT-Large. Groups of classes from the same meta-class are surrounded by blue lines.

Chapter 4
Conclusion

4.1 Discussion

The aim of this thesis was to improve ultrasound image analysis within the framework of OB/GYN pregnancy scans. In this regard, we identified two main challenges for this problem, namely the lack of large-scale annotated datasets and the difficulty to integrate additional rich information such as spatial priors or structured annotations in the learning framework.

In fact, as part of the SUOG project, my thesis focused on helping the ultrasound operator in real time. In particular, it aimed at creating deep learning models, integrated in the ultrasound machine, that would guide the sonographer towards the next acquisition or towards a sound diagnosis automatically. This is essential because of the complexity of ultrasound screening caused by the large number of disorders, the even larger number of signs or findings, and the insufficient number of experts. However, these deep learning models rely heavily on large amounts of annotated images, which is seldom available for medical imaging tasks because the acquisition of these images is costly and sensitive **(challenge 1)**. In practice, the SUOG project only contained a few thousand annotated ultrasound scans. Nevertheless, different types of additional data were made available for the project, such as a limited number of pixel-wise segmentation maps, a knowledge base created by OB/GYN experts containing entities such as disorders, findings and technical elements, and finally rich annotations extracted from the said knowledge graph. Thus, finding a way to integrate this strong spatial and semantic information into the learning framework constitutes an appealing task **(challenge 2)**.

In order to answer these challenges, in chapter 2, we presented our work that leveraged spatial priors during training in order to improve the model's predictive capacity for small and noisy datasets **(challenge 1)**. In particular, we introduced *Prior-Guided Attribution*, a novel method that guides the CNN-based network to focus towards the most salient areas of the input image. More precisely, the method forces the network's attribution maps (which highlight the most relevant pixels of the input image with respect to the output) to resemble prior information heatmaps by means of a *Privileged Attribution Loss* that maximizes the cross-correlation between the two aforementioned maps. The *Prior Allocation Strategy* enables the model to integrate multiple spatial priors while still leaving some freedom for the model to look into other areas that might be interesting for the final prediction. Experimentally, we demonstrated that the proposed method was generic and

consistently increased baseline predictions scores for several tasks and datasets such as facial expression recognition, breast cancer detection and scan plane recognition, without needing any additional information during inference. Moreover, we proved that a good compromise between fine-grained, precise priors and larger, less informative priors offered the best results. This result is important as this spatial information can be costly and difficult to obtain. We argue that the ideas presented in this chapter are therefore of interest for computer vision in general and for OB/GYN in particular to better process small and noisy datasets.

Second, in chapter 3, we exposed our work concerning the integration of higher-order annotations extracted from the SUOG ontology to add semantic context to a Distance Metric Learning learning framework created to deal with scan plane recognition **(challenge 2)**. In a naive DML setup, examples from separate classes would be considered equally different, and therefore similarly pushed apart in the latent space regardless of the semantic distances between classes. For instance, considering ImageNet classification, it could easily be argued that an *American alligator* is more similar to an *African crocodile* than it is to a *saxophone*. Thus, we introduced a method aimed at improving the inter-class similarity distances. In particular, we introduced the *Semantic Abstraction Loss* L_{SAL} built as a weighted average of multiple DML losses applied at different semantic levels. More precisely, we introduced meta-embeddings that shall encode the higher-level semantic information from the meta-annotations. We also proposed to integrate the semantic information extracted from the higher-level annotations as natural language. To do so, we use language guidance (as introduced by Roth *et al.* [75]) with rich captions, and also introduced *Ontology Language Guidance*, aimed at guiding the aforementioned meta-embeddings to better separate the increment given by multiple levels of annotations. Through thorough experimentation, we demonstrated the interest of L_{SAL} and *OLG* on a public birds classification dataset and the SUOG scan plane recognition dataset. We specifically showed that adding semantic context through auxiliary embeddings consistently helped improve the visual similarity capacity of the model for different DML losses, and enabled the model to "make better mistakes" as the results in terms of *recall@1* and *meta-recall@1* increased using L_{SAL}. While the rich captioning method provided only limited improvement because it focused mostly on leaf-level annotations, *OLG* obtained better results as it allowed the model to better separate the impact of each annotations. Also, it is paramount to note that language guidance and *OLG* only offer consequent improvement when the language model is able to provide semantic context in the specific input domain. We argue that the proposed methods therefore hold interest for OB/GYN imaging tasks to integrate strong semantic information through hierarchical annotations.

4.2 Future Works

As it was presented in the conclusions of chapters 2 and 3, short term future works include applying both methods to disorder recognition, integrating in different and evaluating different natures of spatial priors or customizing the training architecture to the ground-truth ontology to better integrate the hierarchical semantic information. We now propose other possible perspectives for further future works.

4.2.1 Using different additional information to guide the learning

During this thesis, we leveraged both spatial information and structured annotations as priors to improve our model's predictions. One track we could explore for further research would be to find different priors that we could use as additional information to guide the learning.

In chapter 2, we use segmentation maps to guide our model spatially. However, annotating images in terms of pixel-wise segmentation maps can be very costly. We could therefore investigate other types of spatial priors, such as points that indicate the center of a structure for instance, or an outline of these important structures to reduce the cost of additional annotations.

In chapter 3, we use hierarchical annotations in a DML framework to improve the inter-class distances for a scan plane recognition task. We could therefore investigate a multi-task framework using both hierarchical labels for scan planes and findings for instance. This could be of interest because all these annotations are linked in the SUOG ontology.

An interesting track to investigate would be to leverage off-the-shelf models such as SAM [51] for instance, as it is able to produce segmentation maps. PGA working as a regularization technique, we demonstrated in chapter 2 that leveraging incorrect or imperfect spatial priors was not detrimental to the network's predictions. This could therefore improve the model's results without needing human annotations in the process.

We could also work with other kinds of additional information, such as different imagery. In the SUOG project for instance, most studies have 2D ultrasound scans but also use Doppler or 3D ultrasound scans. An interesting idea would be to add semantic context to the model by learning correspondences between the different modalities, in a similar fashion to Xie *et al.* [108]. This would help the model locate important structures in order to improve its predictions.

4.2.2 Combining statistical and symbolic AI

Another one of the initial goals of the SUOG project was to be able to get the best of the impressive results obtained by deep learning methods with the transparency and

explainability of the symbolic methods. This was important because of the responsibility of these models in the medical domain. One way to investigate this track would be to mix semantic reasoning or rule-based methods on the SUOG ontology with deep learning visual reasoning methods.

We could explore a framework similar to VQA, where a symbolic AI model from the ontology would be able to reason using semantic features extracted by a deep learning model. For instance, Yi *et al.* [116] first create an abstract scene representation from the input image using an object detection module. After having parsed the question using a neural network, they use a symbolic approach to answer the question. This approach is very interesting because its prediction or answer can easily be traced back. In particular, we could implement a similar method in the medical domain that leverages the large and informative SUOG ontology to answer questions from the ultrasound operator in real time. In particular, the ontology contains a lot of implications concerning findings and disorders (e.g. "lack of decidual layer" is key to "caesarean scar ectopic pregnancy"). This could also mean that, the sonographer could also input certain findings manually and therefore not only rely on deep learning methods to identify structures or objects in the first place. In a nutshell, such a predictive framework could be highly beneficial to help non-expert operators during ultrasound screenings.

4.2.3 SUOG project

Finally, as it was discussed in the introduction (chapter 1), the aim of the SUOG project, and by extension that of my thesis work was the conception of an intelligent ultrasound assistant able to guide the sonographer in real time using artificial intelligence. Both methods presented in this thesis work have been validated on several datasets, but have not been integrated into the SUOG assistant prototype as I am writing these lines. One area of improvement that could be considered would be to re-train both PGA and OLG on the latest SUOG data (new acquisitions and annotations have since been collected), and integrate the best model into the SUOG assistant for scan plane recognition. To further extend this work and maybe increase the model's predictive capacity, both PGA and OLG could be combined as they are not exclusive, to get the best of additional spatial information and strong semantic context. In practice, we could investigate guiding the attribution of a visual encoder in a DML framework.

Another track to be explored would be to apply both these methods to different OB/GYN tasks on the SUOG dataset. As all images from the SUOG dataset are annotated in terms technical elements, findings and disorders, it could be very interesting to apply both PAL and OLG to different classification tasks such as disorder recognition, that could help the non-expert ultrasound operator perhaps more directly than view classification, or a model able to recognize certain signs or findings. These models could enhance the power of the ultrasound assistant, as they would be able to provide several predictions or highlight several similar images from the annotated database.

Bibliography

[1] Karim Ahmed, Mohammad Haris Baig et Lorenzo Torresani : Network of experts for large-scale image categorization. *In* Bastian Leibe, Jiri Matas, Nicu Sebe et Max Welling, éditeurs : *Computer Vision – ECCV 2016*, pages 516–532, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46478-7. 65

[2] Zeynep Akata, Scott Reed, Daniel Walter, Honglak Lee et Bernt Schiele : Evaluation of output embeddings for fine-grained image classification. *In 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2927–2936, 2015. 65

[3] Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled et Aly Fahmy : Dataset of breast ultrasound images. *Data in Brief*, 28:104863, 2020. ISSN 2352-3409. 33

[4] Shadi Albarqouni, Christoph Baur, Felix Achilles, Vasileios Belagiannis, Stefanie Demirci et Nassir Navab : Aggnet: Deep learning from crowds for mitosis detection in breast cancer histology images. *IEEE Transactions on Medical Imaging*, 35(5):1313–1321, 2016. 13

[5] Bilal Alsallakh, Amin Jourabloo, Mao Ye, Xiaoming Liu et Liu Ren : Do convolutional neural networks learn class hierarchy? *IEEE Transactions on Visualization and Computer Graphics*, 24:152–162, 2017. URL https://api.semanticscholar.org/CorpusID:192425. 65

[6] Marco Ancona, Enea Ceolini, Cengiz Öztireli et Markus Gross : Towards better understanding of gradient-based attribution methods for deep neural networks. *In 6th International Conference on Learning Representations (ICLR)*, 2018. 38

[7] Estèphe Arnaud, Arnaud Dapogny et Kévin Bailly : Tree-gated deep mixture-of-experts for pose-robust face alignment. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2:122–132, 2020. 32

[8] Estèphe Arnaud, Arnaud Dapogny et Kévin Bailly : Thin: Throwable information networks and application for facial expression recognition in the wild. *Transactions on Affective Computing*, 2022. 46

[9] Björn Barz et Joachim Denzler : Hierarchy-based image embeddings for semantic image retrieval. *In 2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 638–647, 2019. 65

[10] Christian F. Baumgartner, Konstantinos Kamnitsas, Jacqueline Matthew, Tara P. Fletcher, Sandra Smith, Lisa M. Koch, Bernhard Kainz et Daniel Rueckert : Sononet: Real-time detection and localisation of fetal standard scan planes in freehand ultrasound. *IEEE Transactions on Medical Imaging*, 36 (11):2204–2215, 2017. 11, 13

[11] L. Bertinetto, R. Mueller, K. Tertikas, S. Samangooei et N. A. Lord : Making better mistakes: Leveraging class hierarchies with deep networks. *In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12503–12512, Los Alamitos, CA, USA, jun 2020. IEEE Computer Society. URL https://doi.ieeecomputersociety.org/10.1109/CVPR42600.2020.01252. 66

[12] David Bertoin, Adil Zouitine, Mehdi Zouitine et Emmanuel Rachelson : Look where you look! Saliency-guided Q-networks for visual RL tasks. *In NeurIPS 2022*, novembre 2022. 27

[13] Jules Bonnard, Arnaud Dapogny, Ferdinand Dhombres et Kevin Bailly : Privileged attribution constrained deep networks for facial expression recognition. *In 2022 26th International Conference on Pattern Recognition (ICPR)*, pages 1055–1061. IEEE, 2022. 16, 46

[14] Jules Bonnard, Arnaud Dapogny, Richard Zsamboki, Lucrezia De Braud, Davor Jurkovic, Kévin Bailly et Ferdinand Dhombres : Prior-guided attribution of deep neural networks for obstetrics and gynecology. *IEEE Journal of Biomedical and Health Informatics*, 2023. 16

[15] John Bridle : Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimation of parameters. *In* D. Touretzky, éditeur : *Advances in Neural Information Processing Systems*, volume 2. Morgan-Kaufmann, 1989. URL https://proceedings.neurips.cc/paper_files/paper/1989/file/0336dcbab05b9d5ad24f4333c7658a0e-Paper.pdf. 57, 60, 68

[16] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger et Roopak Shah : Signature verification using a "siamese" time delay neural network. *In* J. Cowan, G. Tesauro et J. Alspector, éditeurs : *Advances in Neural Information Processing Systems*, volume 6. Morgan-Kaufmann, 1993. URL https://proceedings.neurips.cc/paper_files/paper/1993/file/288cc0ff022877bd3df94bc9360b9c5d-Paper.pdf. 68

[17] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish

Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever et Dario Amodei : Language models are few-shot learners. *In* H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan et H. Lin, éditeurs : *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf. 64

[18] Xavier Burgos-Artizzu, David Coronado-Gutiérrez, Brenda Valenzuela, Elisenda Bonet-Carne, Elisenda Eixarch, Fatima Crispi et Eduard Gratacós : Evaluation of deep convolutional neural networks for automatic classification of common maternal fetal ultrasound planes. *Scientific Reports*, 10, 06 2020. 10, 11

[19] Jiangdong Cai, Honglin Xiong, Maosong Cao, Luyan Liu, Lichi Zhang et Qian Wang : Progressive attention guidance for whole slide vulvovaginal candidiasis screening. *In* Hayit Greenspan, Anant Madabhushi, Parvin Mousavi, Septimiu Salcudean, James Duncan, Tanveer Syeda-Mahmood et Russell Taylor, éditeurs : *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pages 233–242, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-43987-2. 25

[20] Yifan Cai, Harshita Sharma, Pierre Chatelain et Julia Noble : *Multi-task SonoEyeNet: Detection of Fetal Standardized Planes Assisted by Generated Sonographer Attention Maps*, volume 11070, pages 871–879. 09 2018. ISBN 978-3-030-00927-4. 13

[21] Micael Carvalho, Rémi Cadène, David Picard, Laure Soulier, Nicolas Thome et Matthieu Cord : Cross-modal retrieval in the cooking context: Learning semantic text-image embeddings. *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 2018. URL https://api.semanticscholar.org/CorpusID:13755946. 62, 63

[22] Hila Chefer, Shir Gur et Lior Wolf : Transformer interpretability beyond attention visualization. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 782–791, June 2021. 55

[23] Hao Chen, Qi Dou, Dong Ni, Jie-Zhi Cheng, Jing Qin, Shengli Li et Pheng-Ann Heng : Automatic fetal ultrasound standard plane detection using knowledge transferred recurrent neural networks. *In International Conference*

on Medical Image Computing and Computer-Assisted Intervention, 2015. URL https://api.semanticscholar.org/CorpusID:6096203. 11

[24] Kun CHEN, Yuanfan GUO, Canqian YANG, Yi XU, Rui ZHANG, Chunxiao LI et Rong WU : Enhanced breast lesion classification via knowledge guided cross-modal and semantic data augmentation. *In* Marleen de BRUIJNE, Philippe C. CATTIN, Stéphane COTIN, Nicolas PADOY, Stefanie SPEIDEL, Yefeng ZHENG et Caroline ESSERT, éditeurs : *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pages 53–63, Cham, 2021. Springer International Publishing. ISBN 978-3-030-87240-3. 14

[25] Weihua CHEN, Xiaotang CHEN, Jianguo ZHANG et Kaiqi HUANG : Beyond triplet loss: A deep quadruplet network for person re-identification. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1320–1329, 2017. URL https://api.semanticscholar.org/CorpusID:14795862. 61

[26] Yifu CHEN, Antoine SAPORTA, Arnaud DAPOGNY et Matthieu CORD : Delving Deep into Interpreting Neural Nets with Piece-Wise Affine Representation. *In IEEE ICIP*, pages 609–613, 2019. 26, 29

[27] T. COVER et P. HART : Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, 1967. 67

[28] Arnaud DAPOGNY, Kevin BAILLY et Séverine DUBUISSON : Confidence-weighted local expression predictions for occlusion handling in expression recognition and action unit detection. *International Journal of Computer Vision*, 126(2):255–271, 2018. 22

[29] Jia DENG, Alexander C. BERG, Kai LI et Li FEI-FEI : What does classifying more than 10,000 image categories tell us? *In* Kostas DANIILIDIS, Petros MARAGOS et Nikos PARAGIOS, éditeurs : *Computer Vision – ECCV 2010*, pages 71–84, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg. ISBN 978-3-642-15555-0. 66

[30] Jia DENG, Wei DONG, Richard SOCHER, Li-Jia LI, Kai LI et Li FEI-FEI : Imagenet: A large-scale hierarchical image database. *In 2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 78

[31] Jiankang DENG, Jia GUO, Niannan XUE et Stefanos ZAFEIRIOU : Arcface: Additive angular margin loss for deep face recognition. *In 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4685–4694, 2019. 57, 59, 60, 68

[32] Ferdinand DHOMBRES, Jules BONNARD, Kévin BAILLY, Paul MAURICE, Aris T PAPAGEORGHIOU et Jean-Marie JOUANNIC : Contributions of artificial

intelligence reported in obstetrics and gynecology journals: Systematic review. *J Med Internet Res*, 24(4):e35465, Apr 2022. ISSN 1438-8871. URL https://www.jmir.org/2022/4/e35465. 7, 10, 15

[33] H. Drucker et Y. Le Cun : Improving generalization performance using double backpropagation. *IEEE Transactions on Neural Networks*, 3(6):991–997, 1992. 35

[34] Mengnan Du, Ninghao Liu, Fan Yang et Xia Hu : Learning credible deep neural networks with rationale regularization. *In 2019 IEEE International Conference on Data Mining (ICDM)*, pages 150–159, 2019. 26

[35] Paul Ekman et Wallace V Friesen : Facial action coding system. *Environmental Psychology & Nonverbal Behavior*. 22

[36] Paul Ekman et Wallace V Friesen : Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17(2):124, 1971. 22

[37] Gabriel Erion, Joseph D Janizek, Pascal Sturmfels, Scott M Lundberg et Su-In Lee : Improving performance of deep learning models with axiomatic attribution priors and expected gradients. *Nature Machine Intelligence*, pages 1–12, 2021. 27

[38] Amir Hossein Farzaneh et Xiaojun Qi : Facial expression recognition in the wild via deep attentive center loss. *In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2402–2411, January 2021. 46

[39] Thomas Fel, Ivan Felipe, Drew Linsley et Thomas Serre : Harmonizing the object recognition strategies of deep neural networks with humans. *In NeurIPS*, 11 2022. 27

[40] Maria Chiara Fiorentino, Francesca Pia Villani, Mariachiara Di Cosmo, Emanuele Frontoni et Sara Moccia : A review on deep-learning algorithms for fetal ultrasound-image analysis. *Medical Image Analysis*, 83:102629, 2023. ISSN 1361-8415. URL https://www.sciencedirect.com/science/article/pii/S1361841522002572. 11, 12

[41] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato et Tomas Mikolov : Devise: A deep visual-semantic embedding model. *In* C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani et K.Q. Weinberger, éditeurs : *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper_files/paper/2013/file/7cce53cf90577442771720a370c3c723-Paper.pdf. 65

[42] Weifeng GE, Weilin HUANG, Dengke DONG et Matthew R. SCOTT : Deep metric learning with hierarchical triplet loss. *In* Vittorio FERRARI, Martial HEBERT, Cristian SMINCHISESCU et Yair WEISS, éditeurs : *Computer Vision – ECCV 2018*, pages 272–288, Cham, 2018. Springer International Publishing. ISBN 978-3-030-01231-1. 59, 66

[43] Jean-Bastien GRILL, Florian STRUB, Florent ALTCHÉ, Corentin TALLEC, Pierre RICHEMOND, Elena BUCHATSKAYA, Carl DOERSCH, Bernardo AVILA PIRES, Zhaohan GUO, Mohammad GHESHLAGHI AZAR, Bilal PIOT, koray KAVUKCUOGLU, Remi MUNOS et Michal VALKO : Bootstrap your own latent - a new approach to self-supervised learning. *In* H. LAROCHELLE, M. RANZATO, R. HADSELL, M.F. BALCAN et H. LIN, éditeurs : *Advances in Neural Information Processing Systems*, volume 33, pages 21271–21284. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/f3ada80d5c4ee70142b17b8192b2958e-Paper.pdf. 25

[44] Richard S. HA, Simukayi MUTASA, Jenika KARCICH, Nishant GUPTA, Eduardo Pascual Van SANT, John S. NEMER, Mary SUN, Peter D. CHANG, Michael Z. LIU et Sachin R. JAMBAWALIKAR : Predicting breast cancer molecular subtype with mri dataset utilizing convolutional neural network algorithm. *Journal of Digital Imaging*, 32:276–282, 2019. URL https://api.semanticscholar.org/CorpusID:59524493. 13

[45] Raia HADSELL, Sumit CHOPRA et Yann LECUN : Dimensionality reduction by learning an invariant mapping. *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, 2:1735–1742, 2006. URL https://api.semanticscholar.org/CorpusID:8281592. 61, 68

[46] Yaru HAO, Li DONG, Furu WEI et Ke XU : Self-attention attribution: Interpreting information interactions inside transformer. *In AAAI Conference on Artificial Intelligence*, 2020. URL https://api.semanticscholar.org/CorpusID:264653968. 55

[47] Steven C. Y. HUNG, Jia-Hong LEE, Timmy S. T. WAN, Chein-Hung CHEN, Yi-Ming CHAN et Chu-Song CHEN : Increasingly packing multiple facial-informatics modules in a unified deep-learning model via lifelong learning. *In Proceedings of the 2019 on International Conference on Multimedia Retrieval*, page 339–343, 2019. 45, 46

[48] Aya Abdelsalam ISMAIL, Héctor Corrada BRAVO et Soheil FEIZI : Improving deep learning interpretability by saliency guided training. *In NeurIPS*, 2021. 27

[49] Geethu Miriam JACOB et Bjorn STENGER : Facial action unit detection with transformers. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7680–7689, June 2021. 24

[50] Wonsik KIM, Bhavya GOYAL, Kunal CHAWLA, Jungmin LEE et Keunjoo KWON : Attention-based ensemble for deep metric learning. *In Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 89

[51] Alexander KIRILLOV, Eric MINTUN, Nikhila RAVI, Hanzi MAO, Chloe ROLLAND, Laura GUSTAFSON, Tete XIAO, Spencer WHITEHEAD, Alexander C. BERG, Wan-Yen LO, Piotr DOLLÁR et Ross GIRSHICK : Segment anything. *arXiv:2304.02643*, 2023. 93

[52] Oliver LANGNER, Ron DOTSCH, Gijsbert BIJLSTRA, Daniel WIGBOLDUS, Skyler HAWK et Ad KNIPPENBERG : Presentation and validation of the radboud face database. *Cognition & Emotion - COGNITION EMOTION*, 24:1377–1388, 12 2010. 23

[53] E.P.V. LE, Y. WANG, Y. HUANG, S. HICKMAN et F.J. GILBERT : Artificial intelligence in breast imaging. *Clinical Radiology*, 74(5):357–366, 2019. ISSN 0009-9260. URL https://www.sciencedirect.com/science/article/pii/S0009926019301163. 13

[54] Hyeonsoo LEE, Junha KIM, Eunkyung PARK, Minjeong KIM, Taesoo KIM et Thijs KOOI : Enhancing breast cancer risk prediction by incorporating prior images. *In* Hayit GREENSPAN, Anant MADABHUSHI, Parvin MOUSAVI, Septimiu SALCUDEAN, James DUNCAN, Tanveer SYEDA-MAHMOOD et Russell TAYLOR, éditeurs : *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pages 389–398, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-43904-9. 14

[55] Jinhyuk LEE, Wonjin YOON, Sungdong KIM, Donghyeon KIM, Sunkyu KIM, Chan Ho SO et Jaewoo KANG : Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, septembre 2019. ISSN 1367-4811. URL http://dx.doi.org/10.1093/bioinformatics/btz682. 76, 86

[56] Shan LI, Weihong DENG et JunPing DU : Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. *In IEEE CVPR*, pages 2584–2593. IEEE, 2017. 31

[57] Jiajun LIANG, Rian HUANG, Peiyao KONG, Shengli LI, Tianfu WANG et Baiying LEI : Sprnet: Automatic fetal standard plane recognition network for ultrasound

images. page 38–46, Berlin, Heidelberg, 2019. Springer-Verlag. ISBN 978-3-030-32874-0. URL https://doi.org/10.1007/978-3-030-32875-7_5. 11

[58] Frederick Liu et Besim Avci : Incorporating priors with feature attribution on text classification. *In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019. 27

[59] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj et Le Song : Sphereface: Deep hypersphere embedding for face recognition. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6738–6746, 2017. URL https://api.semanticscholar.org/CorpusID:206596594. 57, 59, 60, 68

[60] Jiawei Mao, Rui Xu, Xuesong Yin, Yuanqi Chang, Binling Nie et Aibin Huang : Poster v2: A simpler and stronger facial expression recognition network. *arXiv preprint arXiv:2301.12149*, 2023. 46

[61] Qianhui Men, Clare Teng, Lior Drukker, Aris T. Papageorghiou et J. Alison Noble : Multimodal-guidenet: Gaze-probe bidirectional guidance in obstetric ultrasound scanning. *In* Linwei Wang, Qi Dou, P. Thomas Fletcher, Stefanie Speidel et Shuo Li, éditeurs : *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, pages 94–103, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-16449-1. 24

[62] Kevin Miao, Akash Gokul, Raghav Singh, Suzanne Petryk, Joseph Gonzalez, Kurt Keutzer et Trevor Darrell : Prior knowledge-guided attention in self-supervised vision transformers. *arXiv:2209.03745*, 2022. 25

[63] Alberto Montero, Elisenda Bonet-Carne et Xavier Paolo Burgos-Artizzu : Generative adversarial networks to improve fetal brain fine-grained plane classification. *Sensors*, 21(23), 2021. ISSN 1424-8220. URL https://www.mdpi.com/1424-8220/21/23/7975. 11

[64] Maged Nasser et Umi Kalsom Yusof : Deep learning based methods for breast cancer diagnosis: A systematic review and future direction. *Diagnostics*, 13(1), 2023. ISSN 2075-4418. URL https://www.mdpi.com/2075-4418/13/1/161. 13

[65] Andreea-Maria Oncescu, A. Sophia Koepke, João F. Henriques, Zeynep Akata et Samuel Albanie : Audio retrieval with natural language queries. *In Interspeech*, 2021. URL https://api.semanticscholar.org/CorpusID:233740062. 64

[66] Michael Opitz, Georg Waltner, Horst Possegger et Horst Bischof : Deep metric learning with bier: Boosting independent embeddings robustly. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):276–290, 2020. 89

[67] Omkar M. Parkhi, Andrea Vedaldi et Andrew Zisserman : Deep face recognition. *In British Machine Vision Conference*, 2015. 37

[68] Yuxin Peng, Jinwei Qi et Yuxin Yuan : Cm-gans: Cross-modal generative adversarial networks for common representation learning. *ArXiv*, abs/1710.05106, 2017. URL https://api.semanticscholar.org/CorpusID:8355505. 64

[69] Tao Pu, Tianshui Chen, Yuan Xie, Hefeng Wu et Liang Lin : Au-expression knowledge constrained representation learning for facial expression recognition. *In 2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11154–11161, 2021. 23

[70] Ruowei Qu, Guizhi Xu, Chunxia Ding, Wenyan Jia et Mingui Sun : Deep learning-based methodology for recognition of fetal brain standard scan planes in 2d ultrasound images. *IEEE Access*, 8:44443–44451, 2020. 11

[71] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger et Ilya Sutskever : Learning transferable visual models from natural language supervision. *In International Conference on Machine Learning*, 2021. URL https://api.semanticscholar.org/CorpusID:231591445. 59, 64, 71, 75

[72] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei et Ilya Sutskever : Language models are unsupervised multitask learners. 2019. 64

[73] Andrew Slavin Ross, Michael C. Hughes et Finale Doshi-Velez : Right for the right reasons: Training differentiable models by constraining their explanations. *In Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 2662–2670, 2017. 27

[74] Karsten Roth, Biagio Brattoli et Björn Ommer : Mic: Mining interclass characteristics for improved metric learning. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7999–8008, 2019. URL https://api.semanticscholar.org/CorpusID:202749912. 59

[75] Karsten Roth, Oriol Vinyals et Zeynep Akata : Integrating language guidance into vision-based deep metric learning. *In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16156–16168, 2022. 15, 58, 64, 72, 75, 76, 85, 88, 92

[76] Delian Ruan, Yan Yan, Shenqi Lai, Zhenhua Chai, Chunhua Shen et Hanzi Wang : Feature decomposition and reconstruction learning for effective facial expression recognition. *In 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7656–7665, 2021. 45, 46

[77] A. Sanakoyeu, V. Tschernezki, U. Buchler et B. Ommer : Divide and conquer the embedding space for metric learning. *In 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 471–480, Los Alamitos, CA, USA, jun 2019. IEEE Computer Society. URL https://doi.ieeecomputersociety.org/10.1109/CVPR.2019.00056. 59, 89

[78] Jo Schlemper, Ozan Oktay, Michiel Schaap, Mattias P. Heinrich, Bernhard Kainz, Ben Glocker et Daniel Rueckert : Attention gated networks: Learning to leverage salient regions in medical images. *Medical image analysis*, 53:197 – 207, 2018. URL https://api.semanticscholar.org/CorpusID:52091450. 13

[79] Florian Schroff, Dmitry Kalenichenko et James Philbin : Facenet: A unified embedding for face recognition and clustering. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 59, 60, 61, 62, 69

[80] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh et Dhruv Batra : Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359, 2019. 26, 27, 37

[81] Zhiwen Shao, Zhilei Liu, Jianfei Cai et Lizhuang Ma : JÂa-net: Joint facial action unit detection and face alignment via adaptive attention. *International Journal of Computer Vision*, 129(2):321–340, Sep 2020. 23

[82] Bryar Shareef, Min Xian, Aleksandar Vakanski et Haotian Wang : Breast ultrasound tumor classification using a hybrid multitask cnn-transformer network. *In* Hayit Greenspan, Anant Madabhushi, Parvin Mousavi, Septimiu Salcudean, James Duncan, Tanveer Syeda-Mahmood et Russell Taylor, éditeurs : *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pages 344–353, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-43901-8. 14

[83] Jiahui She, Yibo Hu, Hailin Shi, Jun Wang, Qiu Shen et Tao Mei : Dive into ambiguity: Latent distribution mining and pairwise uncertainty estimation for facial expression recognition. *In 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6244–6253, 2021. 45, 46

[84] Avanti Shrikumar, Peyton Greenside et Anshul Kundaje : Learning important features through propagating activation differences. *In ICML*, pages 3145–3153, 2017. 26, 29, 38

[85] Karen Simonyan, Andrea Vedaldi et Andrew Zisserman : Deep inside convolutional networks: Visualising image classification models and saliency maps. *CoRR*, abs/1312.6034, 2013. 28

[86] Karen Simonyan, Andrea Vedaldi et Andrew Zisserman : Deep inside convolutional networks: Visualising image classification models and saliency maps. *CoRR*, abs/1312.6034, 2014. 26

[87] Kihyuk Sohn : Improved deep metric learning with multi-class n-pair loss objective. *In* D. Lee, M. Sugiyama, U. Luxburg, I. Guyon et R. Garnett, éditeurs : *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/6b180037abbebea991d8b1232f8a8ca9-Paper.pdf. 61

[88] Kihyuk Sohn, Wenling Shang, Xiang Yu et Manmohan Chandraker : Unsupervised domain adaptation for distance metric learning. *In International Conference on Learning Representations*, 2018. URL https://api.semanticscholar.org/CorpusID:108299626. 59

[89] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever et Ruslan Salakhutdinov : Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014. URL http://jmlr.org/papers/v15/srivastava14a.html. 35

[90] Emilija Strelcenia et Simant Prakoonwit : Improving cancer detection classification performance using gans in breast cancer data. *IEEE Access*, pages 1–1, 2023. 13

[91] Yumin Suh, Bohyung Han, Wonsik Kim et Kyoung Mu Lee : Stochastic class-based hard example mining for deep metric learning. *In 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7244–7252, 2019. 62

[92] Mukund Sundararajan, Ankur Taly et Qiqi Yan : Axiomatic attribution for deep networks. *In ICML*. PMLR, 2017. 26

[93] Vaanathi Sundaresan, Christopher P. Bridge, Christos Ioannou et Julia Alison Noble : Automated characterization of the fetal heart in ultrasound images using fully convolutional neural networks. *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, pages 671–674, 2017. URL https://api.semanticscholar.org/CorpusID:24996139. 11

[94] Gauthier TALLEC : *Deep learning methods for Action Unit detection*. Theses, Sorbonne Université, juillet 2023. URL https://theses.hal.science/tel-04205382. 54

[95] Gauthier TALLEC, Jules BONNARD, Arnaud DAPOGNY et Kévin BAILLY : Multi-task transformer with uncertainty modelling for face based affective computing, 2022. 54

[96] Robert TIBSHIRANI : Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996. ISSN 00359246. URL http://www.jstor.org/stable/2346178. 35

[97] Hui-Ju TIEN, Hsin-Chih YANG, Pei-Wei SHUENG et Jyh-Cheng CHEN : Cone-beam ct image quality improvement using cycle-deblur consistent adversarial networks (cycle-deblur gan) for chest ct imaging in breast cancer patients. *Scientific Reports*, 11, 01 2021. 13

[98] Nakul VERMA, Dhruv MAHAJAN, Sundararajan SELLAMANICKAM et Vinod NAIR : Learning hierarchical similarity metrics. *In 2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2280–2287, 2012. 66

[99] Thanh-Hung VO, Guee-Sang LEE, Hyung-Jeong YANG et Soo-Hyung KIM : Pyramid with super resolution for in-the-wild facial expression recognition. *IEEE Access*, 8:131988–132001, 2020. 46

[100] C. WAH, S. BRANSON, P. WELINDER, P. PERONA et S. BELONGIE : The caltech-ucsd birds-200-2011 dataset. Rapport technique CNS-TR-2011-001, California Institute of Technology, 2011. 58, 75, 78

[101] Chong WANG, Daoqiang ZHANG et Rongjun GE : Eye-guided dual-path network for multi-organ segmentation of abdomen. *In* Hayit GREENSPAN, Anant MADABHUSHI, Parvin MOUSAVI, Septimiu SALCUDEAN, James DUNCAN, Tanveer SYEDA-MAHMOOD et Russell TAYLOR, éditeurs : *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pages 23–32, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-43990-2. 24

[102] Feng WANG, Jian CHENG, Weiyang LIU et Haijun LIU : Additive margin softmax for face verification. *IEEE Signal Processing Letters*, 25(7):926–930, 2018. 57, 68

[103] H. WANG, Yitong WANG, Zheng ZHOU, Xing JI, Zhifeng LI, Dihong GONG, Jin ZHOU et Wei LIU : Cosface: Large margin cosine loss for deep face recognition. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5265–5274, 2018. URL https://api.semanticscholar.org/CorpusID:68589. 60, 68

[104] Xun Wang, Xintong Han, Weilin Huang, Dengke Dong et Matthew R. Scott : Multi-similarity loss with general pair weighting for deep metric learning. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5017–5025, 2019. URL https://api.semanticscholar.org/CorpusID:118646482. 63, 70, 71, 81

[105] Chao-Yuan Wu, R. Manmatha, Alexander J. Smola et Philipp Krähenbühl : Sampling matters in deep embedding learning. *In 2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2859–2867, 2017. 59, 62, 69, 70

[106] Nan Wu, Jason Phang, Jungkyu Park, Yiqiu Shen, Zhe Huang, Masha Zorin, Stanislaw Jastrzebski, Thibault Févry, Joe Katsnelson, Eric Kim, Stacey Wolfson, Ujas N Parikh, Sushma Gaddam, Leng Leng Young Lin, Kara Ho, Joshua D. Weinstein, Beatriu Reig, Yiming Gao, Hildegard Toth, Kristine Pysarenko, Alana A. Lewin, Jiyon Lee, Krystal Airola, Eralda Mema, Stephanie H Chung, Esther Hwang, Naziya Samreen, S. Gene Kim, Laura Heacock, Linda Moy, Kyunghyun Cho et Krzysztof J. Geras : Deep neural networks improve radiologists' performance in breast cancer screening. *IEEE transactions on medical imaging*, 39:1184 – 1194, 2019. URL https://api.semanticscholar.org/CorpusID:84186546. 13

[107] Yongqin Xian, Zeynep Akata, Gaurav Sharma, Quynh Nguyen, Matthias Hein et Bernt Schiele : Latent embeddings for zero-shot classification. *In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 69–77, 2016. 65

[108] Yutong Xie, Jianpeng Zhang, Zehui Liao, Yong Xia et Chunhua Shen : Pgl: Prior-guided local self-supervised learning for 3d medical image segmentation. *ArXiv*, abs/2011.12640, 2020. 25, 93

[109] Xing Xu, Li He, Huimin Lu, Lianli Gao et Yanli Ji : Deep adversarial metric learning for cross-modal retrieval. *World Wide Web*, 22:657–672, 2019. URL https://api.semanticscholar.org/CorpusID:4560834. 63

[110] Zhoubing Xu, Yuankai Huo, JinHyeong Park, Bennett Landman, Andy Milkowski, Sasa Grbic et Shaohua Zhou : Less is more: Simultaneous view classification and landmark detection for abdominal ultrasound images. *In MICCAI 2018*, 2018. 13, 54

[111] Hong Xuan, Richard Souvenir et Robert Pless : Deep randomized ensembles for metric learning. *In* Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu et Yair Weiss, éditeurs : *Computer Vision – ECCV 2018*, pages 751–762, Cham, 2018. Springer International Publishing. ISBN 978-3-030-01270-0. 89

[112] Fanglei Xue, Qiangchang Wang et Guodong Guo : Transfer: Learning relation-aware facial expression representations with transformers. *In Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3601–3610, 2021. 46

[113] Fanglei Xue, Qiangchang Wang, Zichang Tan, Zhongsong Ma et Guodong Guo : Vision transformer with attentive pooling for robust facial expression recognition. *IEEE Transactions on Affective Computing*, 2022. 46

[114] Jiexi Yan, Lei Luo, Cheng Deng et Heng Huang : Unsupervised hyperbolic metric learning. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12465–12474, June 2021. 59

[115] Zhicheng Yan, Hao Zhang, Robinson Piramuthu, Vignesh Jagadeesh, Dennis DeCoste, Wei Di et Yizhou Yu : Hd-cnn: Hierarchical deep convolutional neural networks for large scale visual recognition. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2740–2748, 2014. URL https://api.semanticscholar.org/CorpusID:206770495. 66

[116] Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli et Joshua B. Tenenbaum : Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. *In Advances in Neural Information Processing Systems*, pages 1039–1050, 2018. 94

[117] Xiang Yu, Shui-Hua Wang et Yu-Dong Zhang : Cgnet: A graph-knowledge embedded convolutional neural network for detection of pneumonia. *Information Processing & Management*, 2021. 54

[118] Zhen Yu, Ee-Leng Tan, Dong Ni, Jing Qin, Siping Chen, Shenli Li, Baiying Lei et Tianfu Wang : A deep convolutional neural network based framework for automatic fetal facial standard plane recognition. *IEEE Journal of Biomedical and Health Informatics*, PP:1–1, 05 2017. 11

[119] Matthew D Zeiler et Rob Fergus : Visualizing and understanding convolutional networks. *In European conference on computer vision*, pages 818–833, 2014. 25, 55

[120] Jiabei Zeng, Shiguang Shan et Xilin Chen : Facial expression recognition with inconsistently annotated datasets. *In Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 46

[121] Qi Zeng, Shahed Mohammed, Emily H. T. Pang, Caitlin Schneider, Mohammad Honarvar, Julio Lobo, Changhong Hu, James Jago, Gary Ng, Robert Rohling et Septimiu E. Salcudean : Learning-based us-mr liver image

registration with spatial priors. *In* Linwei WANG, Qi DOU, P. Thomas FLETCHER, Stefanie SPEIDEL et Shuo LI, éditeurs : *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, pages 174–184, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-16446-0. 24

[122] Lei ZHAO, Kenli LI, Bin PU, Jianguo CHEN, Shengli LI et Xiangke LIAO : An ultrasound standard plane detection model of fetal head based on multi-task learning and hybrid knowledge graph. *Future Generation Computer Systems*, 2022. 13, 54

[123] Zengqun ZHAO, Qingshan LIU et Feng ZHOU : Robust lightweight facial expression recognition network with label distribution training. *In Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 3510–3519, 2021. 22, 46

[124] Liangli ZHEN, Peng HU, Xu WANG et Dezhong PENG : Deep supervised cross-modal retrieval. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 63

[125] Jing ZHENG, De nan LIN, Zhongjun GAO, Shuang WANG, Mingjie HE et Jipeng FAN : Deep learning assisted efficient adaboost algorithm for breast cancer detection and early diagnosis. *IEEE Access*, 8:96946–96954, 2020. URL https://api.semanticscholar.org/CorpusID:218950464. 13