Development of artificial vision based intelligent tools for supervising the wellbeing of animals

Ruiwen He

Université de Lille

Ecole doctorale Science de l'ingénierie et des systèmes (ENGSYS-632)
IEMN - Institut d'électronique de microélectronique et de nanotechnologie
UMR CNRS-8520

Thèse financée par

Thèse présentée par

RUIWEN HE

En vue de l'obtention du grade de

Docteur de l'Université de Lille

Spécialité: Electronique, microélectronique, nanoélectronique et micro-ondes

Développement d'outils intelligents par vision artificielle pour la supervision du bien-être animal

Présentée et soutenue publiquement le 24 Novembre 2022

Membres du jury:

Directeur de thèse	M. Dominique COLLARD	Directeur de recherche CNRS, Université de Lille
Co-directeur de thèse	M. Abdelmalik TALEB-AHMED	Professeur, Université Polytechnique Hauts-de-France
Rapporteurs	M. Amir NAKIB	Professeur, Université Paris Est Créteil
	M. Frédéric CHAUSSE	Professeur, Université Clermont Auvergne
Examinateurs	Mme Ouiddad LABBANI-IGBIDA	Professeur, Université de Limoges
	M. Sebastien JACQUES	MCF, Université de Tours
Encadrants	M. Halim BENHABILES	PhD, enseignant-chercheur, JUNIA ISEN Lille
	Mme Feryal WINDAL	PhD, enseignante-chercheuse, JUNIA ISEN Lille
Invités	M. Christophe AUDEBERT	Gènes Diffusion / GD Biotech
	M. Gaël EVEN	Gènes Diffusion / GD Biotech

University of Lille

Université de Lille

Doctoral school engineering and system sciences (ENGSYS-632)
IEMN - Institute of Electronics, Microelectronics and Nanotechnology
UMR CNRS-8520

Thesis funded by JUNIA ISEN, GENES DIFFUSION Nous innovons, vous progressez., Hauts-de-France, UNION EUROPÉENNE, L'Europe s'engage en Nord Pas de Calais - Picardie, Ce projet fait l'objet d'une demande de cofinancement par l'Union Européenne. Programme Opérationnel Nord Pas de Calais – CCI : 2014FR16M0OP012

Thesis submitted by

RUIWEN HE

In order to obtain the grade of

Doctor from the University of Lille

Specialty: Electronics, microelectronics, nanoelectronics and microwaves

Development of intelligent tools by artificial vision for the supervision of animal well-being

Presented and publicly supported on November 24, 2022

Jury members:

Theis director	Mr. Dominique COLLARD	Directeur de recherche CNRS, Université de Lille
Thesis co-director	Mr. Abdelmalik TALEB-AHMED	Professeur, Université Polytechnique Hauts-de-France
Rapporteurs	Mr. Amir NAKIB	Professeur, Université Paris Est Créteil
	Mr. Frédéric CHAUSSE	Professeur, Université Clermont Auvergne
Examiners	Mrs. Ouiddad LABBANI-IGBIDA	Professeur, Université de Limoges
	Mr. Sebastien JACQUES	MCF, Université de Tours
Supervisors	Mr. Halim BENHABILES	PhD, enseignant-chercheur, JUNIA ISEN Lille
	Mrs. Feryal WINDAL	PhD, enseignante-chercheuse, JUNIA ISEN Lille
Guests	M. Christophe AUDEBERT	Gènes Diffusion / GD Biotech
	Mr. Gaël EVEN	Gènes Diffusion / GD Biotech

I dedicate my dissertation work to my family and many friends. A special feeling of gratitude to my loving parents, Xiang GAO and Shibin HE whose words of encouragement and push for tenacity ring in my ears.

Remerciements

Je tiens à remercier en premier lieu pleinement mes encadrants Monsieur Halim BENHABILES, Madame Feryal WINDAL et mes directeurs de thèse Monsieur Dominique COLLARD, Monsieur Abdelmalik TALEB-AHMED qui ont su me guider et diriger le long de la thèse. Ils m'ont appris à être plus autonome tout au long de ce travail de recherche.

Je voudrais remercier tout particulièrement Monsieur Halim BENHABILES, Madame Feryal WINDAL avec qui j'ai eu la chance de pouvoir travailler. Ils ont toujours été disponibles, à l'écoute de mes nombreuses questions et à m'encourager, ils se sont toujours intéressés à l'avancée de mes travaux. Surtout durant la pandémie de COVID19, ils m'ont aidé à avancer dans ma thèse.

Un grand merci à Frédéric CHAUSSE et Amir NAKIB pour leur relecture de ce manuscrit de thèse et l'intérêt qu'ils ont porté à mes travaux de recherche. Je remercie également Sebastien JACQUES, Ouiddad LABBANI-IGBIDA d'avoir accepté de participer à mon jury de thèse.

Mes remerciements vont à notre partenaire industrielle Gènes Diffusion pour tous leurs supports de ma thèse. Les échanges que nous avons eus entre nous ont été enrichissants et m'ont permis d'évoluer et d'approfondir mon savoir dans le domaine de l'elevage, en particulier Gaël Even et Christophe Audebert.

Je ne saurai oublier les membres de l'équipe DSLS de L'ISEN junia avec qui j'ai passé trois ans de travail. Je tiens à remercier les doctorants avec qui j'ai passé des moments agréables au laboratoire, en particulier Ziheng YANG, Tarek MAYLAA, Bilel GUETARNI.

Je tiens à remercier mes amis qui m'ont encouragé à aller de l'avant, en particulier Walid AMAÏEUR.

Enfin, un plus grand « merci » à mes parents Shibin HE et Xiang GAO aussi ma femme, Xuefeng ZHOU, qui m'ont toujours soutenu et supporté. Leurs aides et encouragements étaient très précieux dans mes choix, parfois difficiles, tout au long de mon parcours universitaire.

"*When you have a dream, you've got to grab it and never let go. "*

Carol Burnett

Author's publications

International journals

- **Ruiwen HE**, Halim BENHABILES, Feryal WINDAL, Gaël EVEN, Christophe AUDEBERT, Agathe DECHERF, Dominique COLLARD, Abdelmalik TALEB-AHMED,
  "A CNN-based methodology for cow heat analysis from endoscopic images",
  Applied Intelligence, DOI: 10.1007/s10489-021-02910-5, vol. 52, no 8, p. 8372-8385, 2022. IF: 5.019, Scopus highest percentile: Q1

- Ziheng YANG, Halim BENHABILES, Karim HAMMOUDI, Feryal WINDAL, **Ruiwen HE**, Dominique COLLARD,
  "A generalized deep learning-based framework for assistance to the human malaria diagnosis from microscopic images",
  Neural Computing and Applications, DOI: 10.1007/s00521-021-06604-4, published on line, 2021. IF: 5.102, Scopus highest percentile: Q1

International conferences

- **Ruiwen HE**, Halim BENHABILES, Feryal WINDAL, Gaël EVEN, Christophe AUDEBERT, Dominique COLLARD, Abdelmalik TALEB-AHMED,
  "A cervix detection driven deep learning approach for cow heat analysis from endoscopic images",
  In : 2022 IEEE International Conference on Image Processing (ICIP). IEEE DOI: 10.1109/ICIP46576.2022.9897442, p. 3672-3676, 2022. Acceptance rate: 24%.

Contents

List of Figures

List of Tables

Résumé

De nos jours, le problème de la population mondiale et de l'alimentation n'a pas été complètement résolu. La production agricole intensive est inévitable, en particulier dans le domaine de l'élevage. De plus en plus d'animaux doivent être confinés dans un petit espace, et les fermiers doivent s'occuper d'un plus grand nombre d'animaux. Dans ce context, le bien-être des animaux ne peut pas être bien assuré, ce qui peut causer divers problèmes au niveau de l'environnement, de la qualité et de la sécurité des aliments, et de la morale sociale. Le développement de l'élevage de précision rend possible le bien-être des animaux. Dans l'élevage de vaches, l'insémination artificielle peut offrir de nombreux avantages potentiels en termes de bien-être pour les vaches, notamment la prévention de la propagation des maladies, la prévention des blessures des taureaux pendant l'accouplement et la sélection du sexe pour éviter de réformer les veaux mâles indésirables. Dans cette thèse, nous avons développé des outils intelligents par vision artificielle pour répondre à deux défis principaux de l'insémination artificielle : la détection incertaine des chaleurs et l'indisponibilité des vétérinaires. Nous avons proposé les contributions suivantes :

1. Nous avons développé un système basé sur l'apprentissage profond pour la classification de l'état des chaleurs des vaches. Dans ce but, une approche originale s'appuyant sur l'analyse de l'appareil génital de la vache à partir d'images endoscopiques a été adoptée. Le système développé permet aux vétérinaires et aux fermiers de détecter les phases de chaleur chez les vaches pour une insémination optimale tout en respectant le bien être de l'animal.

2. Pour remédier à l'indisponibilité des vétérinaires, nous avons développé également un système d'aide à l'insémination artificielle, qui permet de prédire les coordonnées d'une fenêtre pour la localisation du col de l'utérus. À cette fin, un modèle de détection basé sur un transformateur a été spécifiquement conçu pour localiser le col de l'utérus. De plus, nous avons exploité ce modèle pour augmenter les performances de notre modèle de classification de l'état des chaleurs.

Mots clés : bien-être animal, insémination artificielle, apprentissage profond, traitement endoscopique d'images/vidéos, détection des chaleurs, détection du col de l'utérus.

Abstract

Nowadays, the world population and food problem has not been properly resolved. Intensive agricultural production is inevitable, especially in livestock farming. More and more animals have to be confined in a small spaces, and farmers have to take care of more animals. In this context, the welfare of animals can not be ensured, which may cause various problems in the environment, food quality and safety, and social morals. The development of precision Livestock Farming makes animal welfare possible. In cow farming, artificial insemination can provide many potential welfare benefits for cows, including preventing the spread of disease, preventing bull injuries during mating, and sex selection to avoid culling unwanted bull calves. In this thesis, we have developed intelligent tools by artificial vision to address two main challenges of artificial insemination: uncertain heat detection and the unavailability of veterinarians. We have proposed the following contributions:

1. We developped a deep learning-based cow heat state classification system. To this end, an original artificial vision-based approach relying on the analysis of the genital tract of the cow from endoscopic images has been apdoted. The developed system permits to veterinarians and farmers to detect heat phases in cows for an optimal insemination while respecting the well-being of the animal.

2. To remedy the unavailability of veterinarians, we also proposed an artificial insemination assistance system, which permits to predict a window coordinates for the localization of the cervix. To this end, a transformer-based detection model has been specifically designed to localize the cervix. Furthermore, we have exploited this model to increase the performance of our heat state classification model.

Key words: animal welfare, artificial insemination, deep learning, endoscopic image/video processing, heat detection, cervix detection.

Chapter 1

Introduction

There is a growing awareness that the animal is entitled to be treated humanely, even if the animal is for animal production [34]. Thus, more and more consumers demand transparency about the origin of animal products [84]. Products with poor animal welfare have a great potential to be boycotted [34]. However, the problem of feeding the world has not been adequately addressed. In particular, the COVID-19 pandemic has been raging around the world since 2019. To contain the epidemic, many countries have imposed various stringent sanitary measures, including quarantines and teleworking, lockdowns, transportation closures, travel bans and border controls, restrictions on import and export activities, and the closure of many industrial/agricultural activities. Livestock and related industries have felt significant impacts [94].

To address this problem, intensification of livestock production is inevitable. However, intensive livestock farming means that a large number of animals are crammed into a small space, as shown in Figure 1.1 [1], which is detrimental to animal welfare and has led to an environmental crisis and a food security crisis [144]. Therefore, the development of a sustainable livestock system is imperative. A sustainable livestock system is regarded to be environmentally friendly, economically viable for farmers, socially acceptable, and especially beneficial for animal welfare. This requires farmers to gather and analyze information about the farming environment and animal welfare in time to intervene when necessary. Precision livestock farming (PLF) is the use of advanced technologies based on artificial intelligence, IoT, Big Data, and robotics in livestock that allows farmers to monitor and control the health and welfare of their animals at any given time. Animal welfare is a complex issue. In the context of intensive livestock, artificial insemination can provide potential welfare benefits in cow farming,

[1] By Magdalena Pistorius | EURACTIV France | translated by Daniel Eck
https://www.euractiv.com/section/agriculture-food/news/support-for-banning-intensive-farming-grows-despite-cost-of-animal-welfare/

FIGURE 1.1: An exemple of intensive chicken farm

such as reducing the risk of disease transmission and injury or enabling the selection of specific beneficial traits [25], controlling the birth of male calves which are to be euthanized due to being undesirable. Nevertheless, for more than 70 years, the insemination procedure has been performed by blind rectal palpation, which is an invasive method, that can have a negative effect on the animal organism, leading to a deterioration of the cow's well-being [120, 119]. In addition, as an old biotechnology, artificial insemination still faces several challenges, such as uncertain detection of oestrus (heat) and lack of artificial insemination technicians [171, 18, 158]. In this context, this thesis aims to develop a Deep Learning (DL) based computer vision analysis tools to address these artificial insemination issues.

This thesis is part of a collaborative project between Junia-Lille and Gènes Diffusion company (2019-2022) and has been funded by the FEDER European program, JUNIA French Engineering school, and Gènes Diffusion French company.

1.1 Contributions

In this thesis, we have made two distinct contributions: the first one is related to the problem of cow heat state classification from vaginal endoscopic video, which allows farmers to inseminate a cow at the right time, and the second one is related to the

problem of detecting and tracking the cervix of a cow using vaginal endoscopic video, which allows farmers to deposit stored semen in the right place. In addition, detection and tracking of the cervix can be used in the diagnosis of uterine inflammation in cows.
Cow heat state classification – The main condition for the success of artificial insemination within cattle is the heat (or estrus) detection [52]. Traditional heat detection methods are mainly based on behavior monitoring by using various devices, such as sensor devices or cameras. They are considered as incertain, especially, in the case of a ”silent heat” where cows do not show any sign [138, 267]. Furthermore, some behavioral indicators exploited for detecting heat may be affected by a general change in animal physiology [241]. To address the difficulty of heat detection in cows, we proposed an original artificial vision-based approach for cow heat state classification relying on the analysis of the genital tract of the cow from endoscopic images. It is worth noting that our approach is complementary to the existing activity analysis methods. Indeed, these methods offer a global analysis of the herd to detect a set of cows potentially in heat. Hence, our method can be applied to each identified cow to provide a more precise analysis. In this work, we built a cow estrus analysis dataset consisting of 31,360 tagged endoscopic images. These images were extracted from simulated insemination videos of 46 Holstein cows using the Eye Breed device. The estrus state of each cow has been predetermined by experts. We also designed a CNN model named "InceptionVGG8" for endoscopic image analysis, which has been deployed in Android applications for practical use by farmers.

Cow cervix detection – To face the issue of the lack of artificial insemination technicians or experts, we proposed an artificial vision-based guidance system to assist the farmers in the insemination operation. To this end, a transformer-based detection model has been designed for cow cervix detection from endoscopic images and combined with our previous system (heat state classification) to accomplish the whole process of image analysis namely, cervix localization and heat state classification as shown in Figure 1.2.

1.2 Organization of thesis

The remaining of this manuscript is organized into 4 chapters summarized below:

- **Chapter 2** presents the definition of animal welfare and points out the challenge to be overcome for ensuring it. The chapter shows the importance of protecting

FIGURE 1.2: Illustration of the final result displayed on the screen.

animal welfare in terms of increasing production, ensuring food safety, enhancing economic benefits, and improving employee well-being and psychological health. Moreover, the chapter reviews the state of the art of sensoring technology and Deep Learning (DL) based PLF (Precision Livestock Farming) system for improving the welfare of the animal. The chapter concludes by breifly explaining how artificial insemination may provide potential welfare advantages in the context of intensive livestock.

- **Chapter 3** begins with a detailed description of the benefits that artificial insemination can provide in cow farming and points out that heat detection is one of the main conditions and challenges for successful artificial insemination. It then reviews the state of the art of heat detection methods. It presents then our innovative CNN-based approach for heat detection. The chapter describes all the parameters used for model design, as well as the different steps and protocols to create the dataset for model training. Finally, this chapter presents a series of experiments to determine the best components of our model, show the advantages of the proposed model compared to 19 methods from the state of the art, and explain the different steps to deploy our model in an Android application via a

smartphone for practical use.

- **Chapter 4** proposes a new problem, cow cervix detection, to address the shortage of artificial insemination technicians and to assist farmers in the insemination operation. It points out that cervix detection is a problem of object detection from images. The chapter reviews the state of the art in object detection algorithms based on Deep Learning (DL) and points out that these existing architectures are designed and trained to detect at least one target object in the image, which is not suitable for our problem. To this end, we present a novel training data generation protocol and a novel transformer-based model for cervix detection. Finally, the chapter presents a set of experiments to detect the best components of our model, showing the advantages of the proposed model compared to 6 methods from the state of the art.

- **Chapter 5** gives the conclusions and draws some recommendations for future work.

Chapter 2

Animal well-being

2.1 Introduction

In general, an animal is in well-being if it is healthy, comfortable, well-nourished, free from pain, fear, distress, and able to express its innate behavior, as described by the World Organization for Animal Health (OIE) [1]. Ensuring the welfare of animals is our moral obligation and one of the hallmarks of a civilized society. The progress of society is reflected not only in the development of science and technology but also in the attitude toward animals in society.

Nevertheless, the world population and food problem thus far has not been properly solved. The Food and Agricultural Organization (FAO) of the United Nations stated that the population can reach around 9.1 billion by 2050 [79], with 650 million people will still be malnourished[53]. To meet the escalating global demand for food, animal production is being intensified. In this context, ensuring animal welfare is worrisome in livestock farming. Especially in commercial livestock farming that is profit-oriented, animal welfare is often neglected. As in the undercover investigation by Spanish animal protection charity Equilia [2], the raw footage shows multiple decaying birds being eaten by other birds. Unable to stand up straight or walk, there are many birds with deformities and broken bones. Some birds are so broken, that they can not even reach the drinking trough. It is frightening and disturbing. As a matter of fact, this is not an accident nor an exception. Over the past few years, environmental crises such as the exploitation and inhumane use of animals, biodiversity loss, climate change, and pollution have become increasingly prominent. To this end, the united nations resolution

[1] https://www.oie.int/en/what-we-do/animal-health-and-welfare/animal-welfare/

[2] https://www.onegreenplanet.org/animalsandnature/decomposing-deformed-and-trampled-to-death-horrific-footage-shows-conditions-on-italian-chicken-farms/

places animal welfare at the heart of sustainable development [3] at the 5th Session of the United Nations Environment Assembly (UNEA-5.2). Indeed, animal welfare links to animal health and food safety, and animal welfare is of increasing concern in the whole world. Moreover, while ensuring animal welfare, we could get many important benefits. In what follows we briefly discuss each of these benefits.

Improving animal productivity –In livestock, good animal productivity is the main condition for farm profitability and sustainable farm development. It is also the key to meeting the demand for food. Animal productivity refers to animal reproductivity, such as egg and milk products. Good animal productivity means good animal reproductivity. Good animal reproducibility depends on the good condition of the animal, including physical and mental conditions. For this purpose, good feeding, good housing, and good health are necessary. In addition, the attitude of the stockpeople is very important. It has also been proven that the decrease in milk production is also relevant to the negative behavior and attitude of stockpeople [247]. In brief, animals need to be in well-being. Indeed, the reproductive performance of animals is related to their state of animal welfare[85]. Consequently, to ensure a high-quality food supply, animal welfare could not be neglected in any way.

Improving meat and product quality – The quality and security of meat and products have also been the subject of widespread concern in recent years. In order to guarantee the quality and security of meat and products, animal health is the precondition. In livestock, animals are completely dependent on humans for basic needs such as shelter, food, and water. As a result, animal health is dependent on animal welfare. The study [83] reported that improved animal welfare has the potential to improve meat and product quality. In addition to physiological welfare, it has also been demonstrated that animal stress is also a key factor affecting meat quality and that controlling animal stress can also improve meat quality [67, 257, 82]. The 'Welfare-friendly' animal products are growing in popularity, such as free-range eggs, which have increased in recent years [33, 205]. Several studies showed [142, 161, 174]that consumers are willing to pay a higher price for animal products obtained using production processes that enhance animal welfare. From this point of view, most consumers seem to support that comprehensive protection of animal welfare is the best way to improve the quality of meat products.

Avoiding economic loss – Reducing mortality and preventing disease are the need of

[3] https://www.worldanimalprotection.org.nz/news/united-nations-resolution-places-animal-welfare-heart-sustainable-development

ensuring animal welfare, which is also one of the main challenges for livestock, and the way to avoid loss for the farmer. Throughout all stages of livestock farming, various diseases are the main threats affecting the normal physiological activities of animals, including feeding, lying, reproducibility, or even death. On a more serious note, certain animals have to be culled due to the disease. Especially for some zoonotic diseases, 4 million domestic animals were culled for the purpose of disease control during the 2001 FMD epidemic [96], this is catastrophic for the livestock industry. Therefore, providing sufficient feed, arranging a comfortable and clean environment, and timely disease detection can ensure animal welfare while avoiding huge losses.

Human-Based Benefits: physical health and psychological wellbeing – As described above, the prevention and treatment of some zoonotic diseases and the quality of livestock products depend on the state of animal welfare. Ensuring animal welfare not only brings economic benefits but also ensures human health to a certain extent. In addition, the ensured animal welfare helps to establish a good relationship between humans and animals, which could reduce excessive reactions in animals to avoid employee accidents during farm daily management. Moreover, high farmer occupational well-being and a low level of stress are proved to have a direct positive association with the animal welfare indicator [92]. Indeed, the link between the poor treatment of animals and poor treatment of other people was also recognized.

In order to effectively ensure animal welfare, the World Organization for Animal Health has proposed five freedoms principles[4] (as shown in table 2.1), which are widely recognized. These five freedom principles describe society's expectations of the conditions that animals should experience under human control.

Nevertheless, to meet the increasing demand for animal products, livestock management is intensive. In this context, the confined and crowded nature of livestock housing makes it difficult for farmers to closely monitor animal health and welfare [101]. Moreover, it costs money and labor to improve the welfare of farm animals [126]. For the large farm, due to a large number of animals, it is inevitable to neglect the animal welfare. Furthermore, hiring more employees means more expenses, which is unacceptable for commercial farms. For small private farms, which are often located in remote areas, due to the lack of animal welfare decision support systems and the unavailability of veterinarians, animal welfare is hindered. To this end, more than more livestock farms have used various electronic tools for managing livestock. It involves automated monitoring of animals to improve their production/reproduction, health and

[4] https://www.legislation.gov.uk/ukpga/2006/45/contents

Freedom from hunger and thirst	by ready access to fresh water and a diet to maintain full health and vigour
Freedom from discomfort	by providing an appropriate environment including shelter, a comfortable resting area
Freedom from pain, injury or disease	by prevention or by rapid diagnosis and treatment
Freedom to express normal behaviour	by providing sufficient space, proper facilities and company of the animals' own kind.
Freedom from fear and distress	by ensuring conditions and treatment which avoid mental suffering

TABLE 2.1: The Farm Animal Welfare Council's Five Freedoms (FAWC).

welfare, and impact on the environment.

2.2 Animal welfare-based precision livestock farming

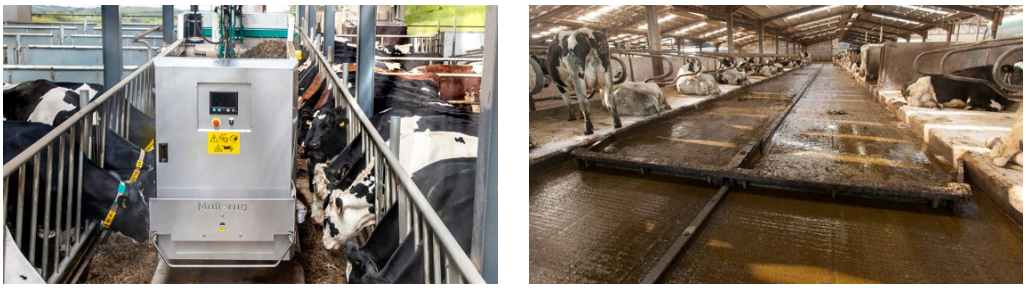
According to statistics from the International Labour Organization (ILO) [5], the number of people working in agriculture in the world has been declining every year, with the regular agriculture labor force in the world declined by 180 million persons between 2003 and 2018. However, statistics from FAO stated that the world's annual meat production in 2018 was 93.49 million tons higher than in 2013. This means that each farmer has to care for a growing number of animals, while there is an increase in demand from a society that the right of animals to individual attention is respected. In this context, the use of various electronic tools for managing livestock namely **Precision Livestock Farming (PLF)** is imperative, one of the aims of PLF is to improve animal welfare. More precisely, the purpose of PLF is to detect deviations at an early stage and improve animal health, welfare, and efficiency, expecting an improvement in production sustainability [23].

As a reliable way of improving animal welfare, a range of precision livestock monitoring and control technologies have been developed, such as automatic feeding systems (AFS) (see in Figure 2.1 - (a)[6]), automatic manure scraper (AMS) (see in Figure 2.1 -

[5] https://www.ilo.org/global/lang--en/index.htm

[6] https://www.gea.com/en/articles/33-years-feeding-green/index.jsp

(b)[7]), which have become basic equipment in modern livestock farm for animal feeding and barn cleaning. Moreover, the application of sensing technology, expert system



(a) Automatic Feeding Systems (AFS) **(b)** Automatic Manure Scraper (AMS)

FIGURE 2.1: Illustration of two basic equipments in modern livestock farm

and computer vision technology have been woven into modern livestock farming, especially the deep learning based computer vision PLF system, that has been very popular in recent years due to higher precision and faster speed. In this section, we will explore and discuss the state of these technologies in modern livestock farming.

2.2.1 Sensing technologies in Precision Livestock Farming (PLF)

The main idea of such a system is to use various sensors to collect data and then use specialized algorithms to process the raw sensor data to provide biologically relevant information. At the same time, thanks to ZigBee, RFID, or WiFi networks, farmers can obtain a variety of detailed and accurate target data information at any time, anywhere, in any environment. Livestock farming has adopted the use of sensing technologies as a way to monitor the farm environment and animal behavior. Some systems can also analyze and diagnose diseases according to the received data. Table 2.2 summarizes some commonly available sensors and their applications in animal welfare, current available sensors fall into two categories: Non-attached and Attached.

[7] https://www.dairymaster.com/products/manure-scrapers/automatic-scraper-system/

Sensors	Application	Categories
Temperature and humidity sensors	Monitoring of temperature and humidity in the ban	Non-attached
Gas sensor	Monitoring of harmful gases in the barn especially ammonia	Non-attached
Opto sensor	Monitoring of lightness in the barn	Non-attached
Microphones/ acoustic sensor	Monitoring of sound in the barn	Non-attached
GPS sensors	Animal tracking and positioning	Attached
Accelerometers	Measuring the movement acceleration of animal	Non-attached/ Attached
Pedometer	Animal activity monitoring, such as counting number of steps	Attached
ECG, Pulsoxymetry	Monitoring of heart rate or respiratory rate	Attached
Body temperature sensors	Measuring body temperature of animal	Attached
Electrical conductivity sensor	Measuring the electrical conductivity of cow rumen	Attached
Rumen PH-sensor	Measuring pH-value of cow rumen	Attached
Radio-frequency identification (RFID)	Individual identification	Attached

TABLE 2.2: Current available sensors for animal monitoring.

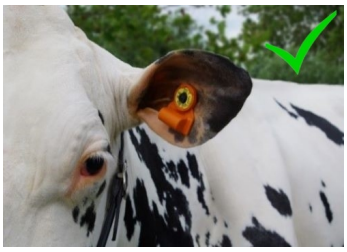
2.2.1.1 Non-attached sensors

Most Non-attached sensors are used for environmental monitoring in livestock farming. The productivity and welfare of an animal in livestock farming are directly related to the environmental conditions of livestock [270]. Based on the welfare of animals, there are some special industry standards on the livestock and poultry field environment to provide an appropriate environment for animals. In this sense, constant monitoring of environmental parameters such as temperature, humidity, lightness, CO2, and all kinds of harmful gases is crucial, if necessary, farmers can intervene more quickly with management measures to ensure a suitable environment. To this end, various environmental sensors are developed for keeping track of the environment in livestock. In poultry farming, sensor systems have been designed for temperature, lighting, carbon dioxide and ammonia monitoring [229, 148, 222, 190]. In addition, acoustic analysis

is an important way in which sensors can provide important information about animal welfare. The analysis of sound information collected by microphones/ placed in barns has been used to detect squeals and to monitor coughs in animals to identify respiratory diseases [22].

2.2.1.2 Attached sensors

Attaching sensors to individual cows is considered the most reliable way to monitor cows throughout a day [101]. Several attached sensors have been developed to monitor behavioral and physiological parameters of animal in livestock farming, which make the state of animal be known overtime. Current attached sensors are either external or internal, most external sensors are easily attached to the animal by using ear tag, collar, ankle straps, halter and belt as shown in Figure 2.2.



(a) Ear tag [8]

(b) Collar [9]

(c) Ankle straps [10]

(d) Halte [11]

(e) Belt [125]

FIGURE 2.2: Illustration of five common external attached sensors

[8] https://support.cowmanager.com/s/article/Attach-sensor-to-the-ear?language=en_US

[9] https://www.dairymaster.com/products/moomonitor/

[10] https://www.icerobotics.com/news/health-sensors-for-livestock/

[11] https://dilaag.boku.ac.at/innoplattform/en/2021/02/10/putting-sensor-technology-to-the-acid-test/

GPS sensors has been widely used for animal monitoring in livestock farming [229, 252, 157], which can be attached in the collar of animal. The main aim of the GPS-based collar is to identify various behaviors of animals, such as feeding, walking, lying, and standing. However, the GPS aims at providing the positional information, and the effectiveness of behaviors classification is insufficient [101].

Compared with GPS sensors, **accelerometers** is more precise on activity tracking, allowing to measure the movement acceleration of the X, Y, and Z-axis, namely triaxial accelerometers. Accelerometers have been deployed in multi-position, including animal bodies and around the environment. In the work of Wang et al [249], the data collected by triaxial accelerometer-equipped ankle straps attached to the cow's leg is used to classify common cow behaviors including standing, lying down, normal walking, active walking, standing, and lying down. Accelerometers can be also deployed in the collar to keep track of the action of the animal head, which refers to the feeding and drinking behaviors of the animal [168]. For cattle, accelerometers can be also mounted in the halter of cattle to identify the rumination behavior of cattle [195]. Moreover, accelerometers can be used for measuring the vibration event. In the work of Bonde.A et al[30], accelerometers based vibration sensors are installed under the barn floor to evaluate overall animal activity. Similarly, accelerometers based vibration sensors mounted on vehicles and belts around the chest of animals are used to evaluate or predict possible discomfort and stress reaction to animals during animal transport [74].

Pedometer is considered as a kind of cheap and simple sensor, which is usually deployed in ankle straps to count the number of steps or to evaluate the walking intensity [88, 253, 252]. The pedometer has been also used in the study of heat behavior [209, 160].

ECG, Pulsoxymetry are generally located in a belt attached around the chest of the animal and are used for heart/respiratory rate monitoring [179]. With the wireless technology, it allows farmers to constantly monitor the heart/respiratory rate of animals [226, 137].

In addition, measuring some physiological parameters requires internal sensors that are deployed within the animal. Since internal sensors are unaffected by external conditions, it is considered more reliable than external sensors [101]. Therefore, internal sensors have various application for animal health monitoring in livestock farming, such as **electrical conductivity sensor** [70], **ruminal pH-value sensor**[176, 38, 268] for measuring the electrical conductivity and pH-value of cow rumen, respectively. Moreover, **body temperature sensor** is also a surgically implantable sensor, which

has been surgically implanted in pigs or cows [177] for body temperature measurement. Radio-frequency identification (RFID) can be also implanted in the animal, which enables the identification of animals [231].
For animal welfare testing, the use of a single sensor is clearly insuffisant. To this end, based on the integration of available sensors, many efficient monitoring systems have been developed for monitoring of animal behavior [88, 136, 195, 249], animal health [254, 141, 137, 89] and detection of some common diseases in livestock, such as lameness [133, 132, 90], mastitis [121, 107], fever [186, 185, 228].
Indeed, the deployment of sensing technology in animal husbandry brings hope and vision to animal welfare, especially with wireless communication technology, farmers can monitor the state of animals in real-time, which can reduce intensive labor in the context of intensified production. However, currently available sensors are in most instances battery-powered, and the lifetime of the sensor needs to be verified [63]. The effectiveness external sensor may be affected by external conditions in combination with continuous observation, [101]. For internal sensors, the reusability is a huge challenge [101], as an invasive sensor, it may make inconfort of animals which is contrary to protecting animal welfare. Moreover, for the wireless system, the quality of wireless communication is also a challenge on a remote farm.

2.2.2 Expert system in PLF

An expert system is a computer system emulating the decision-making ability of a human expert [12], which aims to address the problem of availability of experts. As shown in Figure 2.3, an expert system has two subsystems: the inference engine and the knowledge base. The main work of the knowledge base system is to collect human knowledge, express it systematically or modularize it so that the computer can make inferences and solve problems. The inference engine uses algorithms or decision-making strategies to infer various specialized knowledge in the knowledge base and deduce the correct answer according to the user's question. The common application of expert system is disease diagnosis in human, such as breast cancer detection [130], knee problems diagnosis [213], arthritis diseases diagnosis [62], as well as current pandemic COVID19 diagnosis [211].
Moreover, expert systems have been widely used in precision agriculture, such as crop

[12] https://en.wikipedia.org/wiki/Expert_system

management, which mainly aims at the diagnosis of crop diseases [1, 10, 61]. Similarly, the main application of expert systems in precision livestock farming is to adresse various diseases of animals [206, 57, 182, 5, 232]. Suharjito et al [232] proposed a mimic expert system using fuzzy Tsukamoto for determining the level of risk of endometritis in cows based on six clinical symptoms: 1) Body temperatures (most commonly rectal temperatures); 2) Frequency of breath; 3) Retensio Sekundinae (the speed to expel fetal membranes after parturition; 4) Purulent; 5) Urinate; 6) Lochia (the duration of vaginal discharge). These symptoms were measured by veterinarians and processed by Linguistic Terms as an input variable for their system. The resulting output is obtained by making the membership function of the range between 0 and 100. It contains three level of endometritis: mild symptoms [0-50]; severe symptoms [30-70]; Acute [50-100]. By comparing the system diagnosis results with the expert diagnosis results of 12 cows, the accuracy of the system's prediction of the level of cow metritis reached 100%.

Although the expert system does provide many significant advantages, including providing consistent solutions and reasonable explanations and overcoming human limitations, it does have its drawbacks as well. 1) It cannot respond as creatively and innovatively as human experts in unusual situations; 2) It has high development costs and the recurring cost of subsequently upgrading the system to adapt to the new environment; 3) Domain experts will not always be able to explain the logic and reasoning required for their knowledge engineering process; 4) It may provide wrong solutions.

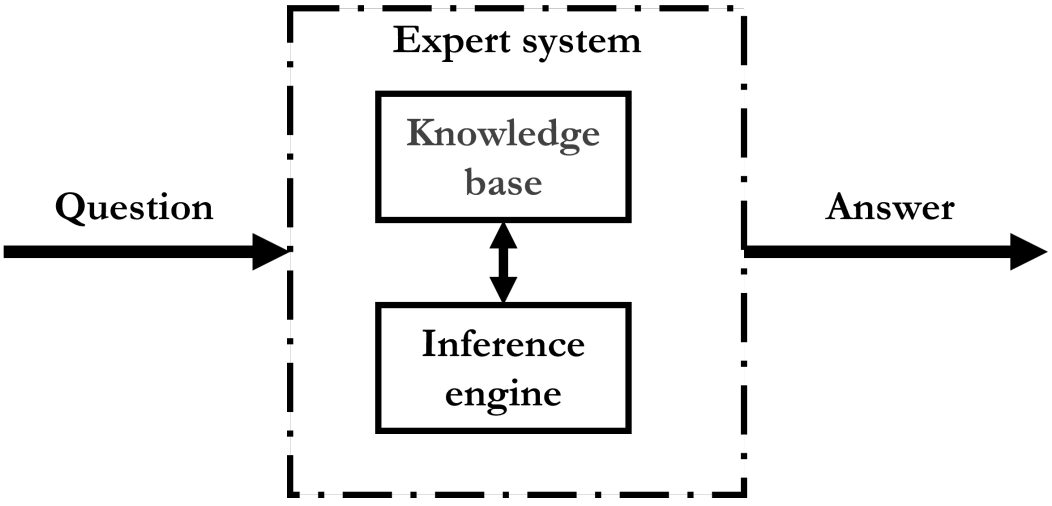


FIGURE 2.3: Schematic diagram of expert system structure

2.2.3 Deep learning-based computer vision in PLF

Computer vision has achieved great success in recent years, mainly due to the advent of a specific machine learning applications as deep learning. Compared with traditional computer vision algorithms, deep learning-based computer vision has higher precision and faster speed, which can perfectly meet the need for real-time monitoring of animals. The studies of animal monitoring have developed from computer vision to deep learning. Image classification, object detection, and image segmentation are the three most common tasks of the deep learning algorithms used for image analysis. In this context, considering the various welfare issues of animals, many deep learning-based computer vision systems have been developed for livestock for animal welfare monitoring. Aiming at different welfare issues, deep learning-based computer vision has been used for different livestock applications including animal identification, behavior recognition, face recognition, etc. We will discuss these applications in this section.

2.2.3.1 Animal identification

Nowadays, animal identification is recognized as an important component related to human health and food safety. Some studies have proposed the use of deep learning-based models of classification or segmentation for animal identification [167, 93, 127, 269]. Since they did not consider individual animal identification as a multiclass problem, but rather as a binary classification problem (presence of an animal in the image). This is insufficient to meet the need for individual animal tracking.

To accurately identify an individual animal, several identification methods have been developed, which can be categorized into semi-permanent recognition methods and permanent recognition methods. Collar ID is one semi-permanent recognition method. In the work of [27], an object detection model namely Faster R-CNN [223] is used for cow individual identification by recognizing the numbers on the collars on the cows' necks. Semi-permanent recognition methods such as collar ID-based methods are considered unsuitable for long-term usage due to susceptibility to damage, duplication, loss, and unreadability [140]. Contrariwise, permanent identification methods, including ear tattoo-based identification technology, embedded with microchip-based markings have been proved to have higher accuracy [139]. Nevertheless, based on animal welfare, this identification method has been banned in many countries due to its invasiveness to animals [131].

In addition to these artificial marking methods, animals also have biometrics such as

nose patterns, and coat patterns, that are used for individual identification, which is similar to human biometrics like faces and fingerprints. The deep learning-based classification of cattle nose pattern image [140, 20] and coat patterns image [280, 149] has been used for individual cattle identification. To simplify the image collection, a two-stage YOLOv3-ResNet50 object detection algorithm has been proposed by Shojaeipour et al.[224], the proposed algorithm can precisely localize the muzzle of cattle. Moreover, human facial recognition has been an active area of research for the past few decades. Facial recognition applications have been used to improve surveillance systems, identify threats and create high-security access systems. In recent years, several deep learning-based systems have been developed for animal facial recognition [93, 258, 106], with promising results. They proposed the use of animal faces as biometric information for individual animal identification. In the work of Hhitelman et al [106], they collected 5625 facial images from 81 Assaf breed sheep. The Faster R-CNN [223] deep learning object detection algorithm was applied to localize the sheep's face in an image. The detected face was provided as input to a CNN model to predict the ID of sheep with average identification accuracy of 97%.

Indeed, deep learning-based biometrics image analysis shows great performance in animal identification. However, it is a challenge to identify individuals animal of the same species with the same body size and single-color breeds, such as Holstein cows for the methods based on nose pattern, and coat pattern image analysis. For facial recognition methods, the biggest challenge comes from the complex farm environment, such as light intensity. Animals are living creatures and do not always face the camera at an angle. Moreover, in the work of Hhitelman et al [106], the case of multiple sheep appearing in the same image which is the common case of real-time animal tracking in practice is not considered. For large commercial farms, the classification of the faces of hundreds of animals would be a challenge. Therefore, more data and more experiments are also required.

2.2.3.2 Behaivor recognition

Monitoring and assessing the behaviors of livestock is important as it can be used to indicate their welfare state and well-being state [194, 23, 59]. For this purpose, a variety of deep learning-based computer vision systems have been developed for animal behavior monitoring, such as feeding and drinking behavior, aggressive behavior, and some behaviors (lameness, mounting) related to disease and heat events.

Feeding and drinking behavior recognition – In the work of Betzen et al [27], based

on individual cow identification as described above, they also proposed a computer vision system based on a deep convolutional neural network (CNN) model for individual feed intake measurement in dairy cows by combining information from RGB and depth images. Zhang et al. [271] proposed a deep learning-based algorithm for sow behavior detection using a MobileNet classification network with depth separable convolutional operations, which allow for detect drinking, urinating, and mounting behaviors in the sow. Feeding and drinking behavior are the two most basic behavior of all animals. Achour et al. [2] used the top of the cow head image as a region of interest (ROI) and different classifiers based on convolutional neural network (CNN) models to identify the feeding behavior of Holstein cows [2]. Tsai et al. [241] used imaging modules installed above the drinking troughs to collect video streams, which are analyzed by a CNN model to estimate the drinking time and frequency of dairy cows.

As the two most basic behaviors of all animals, some systems have been reported to detect both feeding and drinking behaviors. Yang et al. [260] used a fully convolutional network (FCN) to extract the roundness of the pig's head and the overlap area between the head and the feeding bunk as spatial features and used optical flow vectors to define the intensity of the head motion as temporal features, combining spatial and temporal features for feeding behavior detection. The recognition of drinking behavior is the same as the recognition of feeding behavior, except that the distance from the snout to the drinking nipple and head circularity for distinguishing head-up/head-down is used as the extracted spatial features. In the work of Jiang et al. [122], they obtained exciting results in goat eating, drinking, activity, and inactivity behavior detection by using an object detection method namely YOLOv4 [6]. Moreover, the CNN-LSTM (Long Short-term Memory) network is a classic combination for video processing, the spatial features extracted from the CNN are used as inputs in the LSTM, which then extracts spatial-temporal features. have been used for the detection of feeding and drinking behavior in pigs [45, 42, 256].

Aggressive behavior recognition – In livestock farming, existing computer vision-based studies on aggressive behavior recognition are based on pigs. Several computer vision-based systems for aggressive behavior recognition in pigs have been proposed [245, 43, 41], which were focused on spatial feature extraction of pigs in each frame. To directly use spatial-temporal features for detecting the aggression behavior between pigs, the CNN-LSTM network has been used in the work of Chen et al [44]. In feeding and drinking behavior recognition, the spatial-temporal features and extracted by CNN and LSTM. Finally, through the fully connected layer, the prediction function Softmax

is used to determine whether the current episode is aggressive.

Lameness recognition – Lameness occurs when an animal has leg or foot pain that affects how they move. Lameness is an animal health and welfare concern, as well as a production issue. Pain due to lameness often limits growth because animals may be reluctant to eat or drink [13]. Several studies about computer vision-based systems for lameness recognition in cattle have reported mainly cattle [117, 273, 9]. With the development of deep learning technology, Wu et al.[255] used YOLOv3 [198] to detect the leg targets of cows in each frame of video. Finally, they used a LSTM model to classify lame and non-lame cows. Similarly, RFBNetSSD [155] was used in the work of Kang et al. [129] to detect the locations of cow legs. Subsequently, they proposed a classification algorithm based on the analysis of data for the supporting phase for lameness recognition. Moreover, a study has reported the possibility of horses' lameness identifying by using convolutional network models for motion trajectory analysis [251].

Mounting behavior recognition – Commonly, mounting is thought of as being sexually driven. Therefore, the mounting behavior recognition is frequently used for animal heat detection in livestock farming [208, 138, 81, 21]. A single shot multibox detector (SSD) [156] and MobileNet [112] base optimised deep learning network namely SBDA-DL [272] is proposed for real-time detection of three typical sow behaviors: drinking, urination, and mounting. The detection speed of 7 frames per second and the detection Average Precision (AP) of more than 90% meet the requirements of most pig farm support staff for daily monitoring. In addition, Li et al. [147] proposed a pig segmentation network using Mask R-CNN to extract multidimensional feature vectors from the bounding box parameters and the mask file generated by instance segmentation. The feature vectors are extracted from the segmentation results and then classified by a kernel extreme learning machine (KELM) [114, 115]. Based on the classification results, it is determined whether the mounting behavior has occurred.

Behaviour posture recognition – Animal pose estimation is a key step in analyzing animal behaviors and evaluating animal health, hence, study on deep learning for animal pose estimation has also made great progress. The analysis of depth image is used for identifying five postures (standing, sitting, sternal recumbency, ventral recumbency, and lateral recumbency) in the work of Zheng et al.[275]. The depth image is obtained by Kinect v2 sensor, the object detection model Faster R-CNN [223] is

[13] https://www.beefresearch.ca/research-topic.cfm/lameness-64

used for depth image analysis to predict sow postures. As an excellent object detection, with Neural Architecture Search (NAS) base network [281], Faster R-CNN [223] is also used by Riekert et al. [202] to analyze the images collected by a 2D-cameras placed on top of the barn for identifying lying and standing postures of the sow. Based on the idea of Faster R-CNN [223], Zhu et al. [278] proposed a multi-stage object detection framework for recognizing five postures of lactating sows (standing, sitting, sternal recumbency, ventral recumbency, and lateral recumbency) in scenes at a pig farm. They used two CNNs to extract the RGB image features and depth image features. Then, they used a Region Proposal Network (RPN) to generate the regions of interest (ROIs) for the two types of image feature maps in RGB-D, and a feature fusion layer is used to extract and merge the features of the RGB-D ROIs, which are the input of a Fast R-CNN to obtain the recognition results. Nasirahmadi et al. [180] tested three deep learning-based object detector including Faster R-CNN [223], SSD [156] and region-based fully convolutional network (R-FCN) [49], combined with InceptionV2 [235], Residual Network(ResNet)[98] and InceptionResNetV2 [234] to detect the standing and lying (belly and side) postures of pigs under commercial farm conditions. Finally, ResNet101-based R-FCN gets the best performance with mean average precision (mAP) of more than 93%. Similarly, Faster R-CNN-based object detector for sow pose estimation has been reported in the work of Zhang et al .[274]. In addition to object detection technology, the used of deep cascaded convolutional models [150] and panoptic segmentation [35] have been reported for animal pose estimation.

In fact, in addition to the above-mentioned behaviors, other animal behaviors are also of interest and related deep learning-based monitoring systems have been developed, such as tail-biting behavior [153], nursing behavior [261], playing behavior [40], hen egg breeders behavior [248] and cow rumination behavior [14], etc. From the article review, we can find that CNN-LSTM and various object detectors are frequently used for animal behavior monitoring. Indeed, CNN-LSTM allows the identification of animal behavior considering both temporal and spatial features, and advanced object detectors can quickly predict animal behavior and localize animals, which is suitable for real-time monitoring systems. Moreover, these proposed systems all report high performance. However, most of them lack practical experience. Commercial farms are known to have complex environments and a large number of animals. In such a situation, the effectiveness of animal behavior monitoring and the system's resistance to interference face challenges. Especially for possible silent behaviors.

2.2.3.3 Facial expression recognition

As described in section 2.2.3.1, animal facial recognition has been used for animal individual identification. In addition, these advances in facial recognition technology can be extended to several other useful applications, such as helping us learn more about the emotional and attentional states of animals. Pain is one of the most basic animal emotions, correct recognition of pain in animals is essential to ensure animal welfare and to provide successful and rapid treatment when needed [69, 51]. In the work of Guesgen [86], the painful facial expressions of the lambs have been quantified and coded, which demonstrates changes in lamb facial expressions associated with pain. Moreover, Goldberg et al.[80] describe in detail how to identify signs of pain specific to livestock (pig, cattle, and goat) based on their facial expressions.
As deep learning and computer vision-based methods show a great powerful performance in the recognition and classification of images. Noor et al[181] proposed the use of the CNN model for pain facial expression image classification in sheep. They trained a CNN-based binary classifier for sheep pain recognition. As shown in Figure 2.4, the facial expression images are divided into two classes, i.e Normal (pain) or Abnormal (no pain). VGG16 model [227] has been proved as the best classification model for pain facial expression images classification in sheep with a testing accuracy of 100%. Several studies have shown that pain causes changes in the position of the

FIGURE 2.4: Exemple of pain or no pain facial expression in lambs[181]

animal's eyes, ears, lips, and jaw profile, as well as the position of the nostrils and midperson [78, 169]. To more accurately recognize the horse's painful expressions, Dalla

Costa et al. [146] designed a three levels of pain (Not Present, Moderately Present, and Obviously Present)assessment system based on three CNN models to recognize the pain levels of ears, eyes, mouth, and nostrils, respectively. From the three trained CNN models, a fine-tuning process was performed to train the ANN(Artificial neural network)-based classifier that combines the individual recommendations and confidence value of each model. The system was able to achieve an overall accuracy of 75.8%. If not consider the level of pain (only pain or no pain), the overall accuracy of reached 88.3%.

2.3 Purpose and significance of this study

Animal welfare is a complex issue, which includes direct and indirect indicators. In cow farming, artificial insemination is usually considered as a measure for the improvement of fertility, which is the key to farm profitability. In fact, in the context of intensive livestock, artificial insemination may provide potential welfare advantages as these practices reduce the risk of disease transmission and injury or enable the selection of specific beneficial traits [25].

Nevertheless, as an ancient biotechnology, artificial insemination still faces many challenges that hinder its development and promotion. Heat (or estrus) detection is the first challenge and one of the main conditions for a successful insemination [52]. Current heat detection methods are mainly based on the observation of the behavior of the cow or bull, some of which are uncertain and inaccurate, and some of which are harmful to cow welfare. In addition, for more than 70 years, the insemination procedure has been performed by rectal palpation, which allows the operator to move his catheter through the cervix and then deposit the semen in the uterine body. However, rectal palpation is also an invasive method, that can have a negative effect on the animal organism, resulting in deterioration in the well-being of cow [120, 119]. Meanwhile, an effective examination by this method is not easy and requires theoretical and practical preparation. It requires the operator to put his arm into the cow's vagina. If the cow is stressed caused to improper operation, it can lead to dislocation and fracture of the operator's arm. Coupled with the decline in farm labor on dairy units, it is getting more difficult to successfully artificially inseminate (AI) dairy cows to get them pregnant when required.

In this context, the main objective of this study was to develop a deep learning computer vision-based artificial insemination system for dairy cows to improve the insemination success rate by improving the accuracy of estrus detection and locating the cow's cervix, while ensuring the welfare of the cow.

Chapter 3

Cow heat classification

3.1 Introduction

In cattle farming, the artificial insemination technique [1] is a biotechnology that clearly contributes to the development of sustainable agriculture. Indeed, several studies have shown the wide range of benefits brought by this technique [25, 110, 16, 215], especially with regard to genetic gain, advantageous economic cost, health security, and potential welfare advantages for animals. In what follows, we briefly discuss each of these advantages.

Welfare advantages for male – As mentioned in the section 2.3, people have rarely noticed that artificial insemination can provide potential welfare advantages for cattle on commercial farms. Natural mating can cause a range of injuries for bulls. Adult bulls can weigh between 500 and 1000 kg, and hooves hitting the ground may cause lameness at the end of mating. Furthermore, the struggle of the cow cause damage to the bull's penis. The study has reported that lameness and penile damage are estimated to increase by a factor of 2.5 in cows after mating compared to before mating in Australia [91]. In contrast, artificial insemination can well avoid these cases. Furthermore, dairy cow farming aims at milk production. Therefore, male calves are generally unwanted [54]. Unfortunately, the welfare of bull calves is frequently overlooked given they often leave the farm within a week or two after birth and are often of low-value [108], their fate is to wait to be euthanized. Considering the welfare of bull calves, artificial insemination using pre-determined sexed semen allows dairy farmers to avoid the surplus production of male calves and maximize female [108].

[1] An assisted animal reproduction method consisting of artificially introducing, by a trained breeder, the semen of the bull into the reproductive tract of the cow [25].

Health security for both animal and humain – The development of artificial insemination in cattle farming is mainly due to health reasons [25, 110, 16, 215]. Indeed, the spread of disease is better controlled when ensuring that selected bulls for the insemination process are safe from any venereal affection. Otherwise, many hereditary diseases can shorten the life of newborn calves, such as Bovine Leukocyte Adhesion Deficiency can cause immune disorders in the calf, and the calf will not survive past eight months of age, which seriously damages their welfare. More seriously, some zoonotic diseases are heritable, which can cause a huge number of newborn calves to be culled. Obviously, the control of the semen donors must be highly rigorous since the risk of a resisting pathogen to the artificial insemination process will lead to a larger spread of the disease compared to a natural insemination. For instance, in [173] it has been shown that twelve heifers have been contaminated by bovine virus diarrhoea virus (BVDV) after an artificial insemination with semen from an infected bull. However, considering the current standards followed by the semen production centers over the world, the risk of spreading disease remains weak [237].

Genetic gain – Over the past fifty years, genetic improvement programs have highly increased the productivity in most animal species [103]. In the case of cattle farming, these selection programs notably genomic selection aim to improve several points [24]: i) the robustness of the cattle by targeting different skills such as fertility, disease resistance, and longevity, ii) the food efficiency while limiting discharges and iii) the milk through its sanitary, nutritional and technological quality. The application of artificial insemination using semen collected from these selection programs will help to spread the genetic gain and progress. Therefore, the continuation of good offspring also invariably improves the welfare of the cow.

Economic cost – Artificial insemination may bring to farmers a real financial benefit [17, 215] notably with respect to the transport, animal inventory, and associated labor costs. Ball and Peters [16] pointed out some brakes related to the possession of a bull such as the required purchase investment and the long lead times before getting back profits as well as maintenance costs at the end of career. Moreover, artificial insemination allows to avoid financial losses caused by possible infertility problems of a bull. In this context, several studies have been carried out to evaluate the economic impact of artificial insemination over natural mating [17, 159, 11, 215]. For instance, Barrientos-Blanco et al. [17] developed a NPV model (Net Present Value) that includes different parameters such as dystocia and stillbirth costs, as well as improved fertility of crossbred cattle. Indeed, the obtained results have shown that the NPV was

clearly advantageous following a specific artificial insemination strategy. On the other hand, long-distance transportation is also painful for cattle, and reducing long-distance transportation is also a boon for cattle.

The main condition for the success of artificial insemination within cattle is the heat (or estrus) detection [52]. Indeed, detecting cow heat permits the farmer to determine the right time of insemination and lead to: i) increase the conception rate, ii) avoid economic loss due to expenses related to extended calving and additional semen. A cow in heat manifests a primary behavioral sign corresponding to firm footing and allowing herd mate to mount it [73]. Obviously, this sign may be observed only if the interactions within the herd are allowed. Furthermore, as previously mentioned, mounting behaviors may cause damage to cows, and threat the welfare of cow. In addition to primary sign, several secondary signs may be manifested by the cow such as mucus discharge, swelling and reddening of the vulva, bellowing, restlessness and trailing [184, 73]. Consequently, the required time for observing all these signs by the farmer turns this task into a fastidious process specially within herds of a large size. For this reason, several cow heat detection systems, have been recently proposed in the literature to assist the farmer in this task [214, 138, 8, 87, 102, 195, 13, 39], They are mainly based on the analysis of the cow behavior in order to detect eventual signs of changes with respect to its physical activity. Nevertheless, analyzing only the cow activity is not enough for efficiently detecting whether it is in heat. Indeed, the activity of the cow may be affected by multiple factors such as feeding, type of housing, cow density, feet and leg problems [184, 267]. Moreover, in case of a "silent heat", cows do not show any sign [138, 267]. Nevertheless, the finding of the scientific community remains unanimous about the fact that the human visual observation is an efficient method to detect heat [138, 8]. As formulated by Kumar et al. [138], the visual observation is efficient notably: "*if it is done three times a day for at least 30 minutes every time*". Moreover, it is recognized that combining heat detection aids tools with the visual observation gives better detection results [138, 8].

In this sense, we propose a deep learning-based system for cattle heat automatic detection from vaginal video endoscope imagery. The system allows to analyze a sequence of endoscope images acquired in real-time using an innovative insemination technology named *Eye Breed* [55] [2]. As illustrated in figure 3.1, the Eye Breed device, which is equipped with an embedded camera and is connected to a smartphone, is introduced

[2]Eye Breed is the 1st TECHNOLOGICAL INNOVATION to INSEMINATE WITHOUT RECTAL PALPATION. https://www.axce-repro.com/en/eye-breed-2/

into the genital tract of the cow allowing the operator (farmer) to monitor in real-time a simulated insemination process. The recorded video is then automatically analyzed by a CNN classifier at the frame level in order to detect whether the cow is in heat. The contributions of this work are the following:

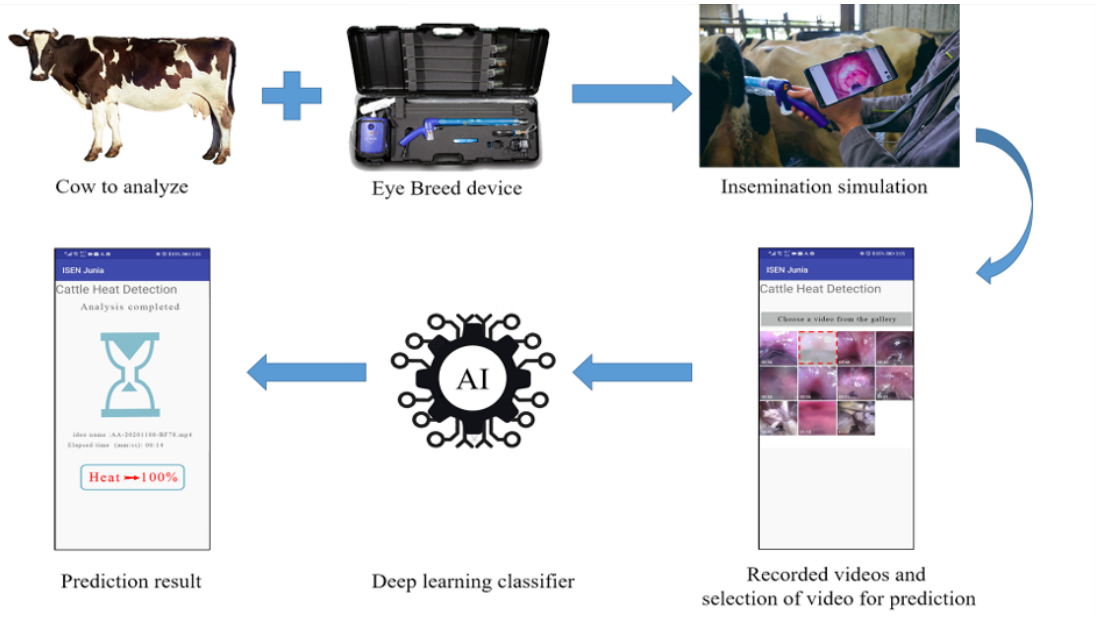


FIGURE 3.1: An overview of our cow heat detection system using Eye Breed technology [55].

- An original artificial vision-based approach for cow heat state classification relying on the analysis of the genital tract of the cow from endoscopic images. It is worth noting that our approach is complementary to the existing activity analysis methods. Indeed, these methods offer a global analysis of the herd to detect a set of cows potentially in heat. Hence, our method can be applied on each identified cow to provide a more precise analysis.

- A dataset for cow heat analysis composed of 31360 labeled endoscopic images. The images have been extracted from videos of simulated insemination on 46 Holstein dairy cows using the Eye Breed device. The heat state of each cow has been pre-identified by experts.

- A CNN model named “InceptionVGG8” tailored for endoscopic image analysis with an efficiency demonstrated over two datasets namely our proposed cow heat

dataset and Kvasir dataset for human digestive system analysis [193]. Indeed, the conducted experimental studies on the two datasets show the high accuracy of our model outperforming 19 methods from the state of the art.

- An optimized version of our CNN model for an Android deployment by exploiting several techniques namely quantization, GPU acceleration and video downsampling. The conducted experiments on a smart-phone show that our Android application renders a decision in a few seconds.

3.2 State of the art

3.2.1 Previous research on heat detection

As mentioned into the introduction, the human visual observation is an efficient method to detect heat [138, 8]. However, the required time for observation and the relatively high number of secondary signs to consider for the achievement of this task, turn it out into a fastidious process specially within herds of a large size. Furthermore, in this latter case the detection efficiency may be affected due to the repetition, by the farmer, of the same actions at a large scale. For these reasons, several heat detection aid techniques ranged from simple annotation techniques to more elaborate ones, have been developed to assist the farmer in the heat detection process. In what follow, we present the different techniques that have been proposed in the literature with a categorization according to the type of the used technology. Apart from the first category of methods, the performance of the presented systems in what follow has been evaluated by calculating standard metrics namely accuracy, sensitivity and positive predictive value. Indeed, the accuracy reported in some recent systems exceeds 80%. However, it is not possible to make an objective comparison between the systems based on the reported performance since they have been evaluated by adopting different experimental protocols notably in term of size of herd, the number of cows in silent heat and the farm characteristics. For these reasons, we choose to highlight their advantages/drawbacks in term of usage ease and adaptation to particular contexts.

Annotation and recording – This type of methods is mainly based on the maintain of a calendar by the farmer to register cows observed in heat and then to anticipate their next heats by considering a cycle of 21 days [8]. Hence, the annotation result endorsed by the visual observation allow the farmer to monitor the cycle of cows and

detect possible cyclicity resumption defaults as well as heat expression issues. This means that the maintain of the calendar is extremely important and must be regularly updated in order to preserve the usefulness of the method. Such calendars are generally provided by artificial insemination centers and are available under the form of special charts. Information systems for farming management propose also this function [163, 60, 26, 183]. In this latter case, several data with respect to the cow characteristics (e.g. sanitary state) and its environment (e.g. type of housing, footing surface and ambient temperature) are required by the system in order to suggest a date at which the cow will need a particular observation. Nevertheless, this type of methods does not provide a detection tool but rather an advice tool to the farmer.

Mount detectors – These methods use a mechanical or an electronic device to identify which cow has been mounted. More precisely, the device is glued on the tail-head of the cow and is triggered if she undergoes a pressure with a sufficient intensity and duration. Indeed, the contact between two cows due to a mount allows to trigger the detector which is glued on the mounted cow. The mechanical mount detectors correspond generally to a flat patch with one adhesive side and on the other side a large transparent storage capsule occupying almost the whole surface of the patch. The large capsule wraps a small one that contains a red fluid. Hence, the pressure due to the mount on the cow leads to the spread of the red fluid, from the small capsule toward the larger one, making it much more visible. The Kamar heat mount detector is one of the oldest and most popular mechanical detectors [207]. However, several mechanical detectors have been developed thereafter [203, 95, 31, 170] offering more options such as the possibility of being used several times. The electronic detectors are based on the same concept as in the mechanical detectors; however they use an electronic pressure sensor for detecting the mount. In addition, these detectors are generally connected via a wireless network to a central unit or a software that allow to keep the farmer informed about the mount data in the herd and the associated cows identities [220, 259, 262]. The drawbacks of these detectors are the risk of dropout and/or the increase number of false detections [109]. For instance, a mount may be detected due to a pressure from a contact with a wall or a barrier. Moreover, these detectors are not efficient with cows in silent heat.

Locomotor activity detectors – Activity increase within cows during heat period [243] have promoted the development of locomotor activity monitoring systems. In this context, several developed systems use a pedometer for monitoring the activity of cows [160, 210, 204]. More precisely, the pedometer is generally placed on one of the

cow legs and is exploited to count the number of steps made by the cow during multiple time intervals. The data collected by the pedometer are transmitted via a wireless network to a central processing unit where they are analyzed to eventually identify a behavior changing in the cow activity. The data analysis stage is performed by algorithms that generally calculate the ratio between the number of steps taken by the cow for a short slot time (for instance the last 2 hours) and the standard deviation of the average number of steps taken in a defined preceding period (for instance the last 10 days) [160, 210, 204]. If the ratio exceeds a given threshold which is defined empirically, the monitoring system identifies then an activity increase and consequently the heat. In recent proposed monitoring systems [239, 195, 219], more sophisticated detectors namely accelerometers have replaced the pedometers. More specifically, these motion sensors are usually placed on the head or the neck of the cow and allow to acquire data into the three dimensional space offering accurate information with regard to the activity type of the cow (e.g. grazing, resting, walking, and standing). The use of these sensors allowed to improve the heat accuracy detection of the monitoring systems, nevertheless the proposed systems still suffer from weakness to false detections due to several factors related to the cow environment [239].

Odors analysis – It has been shown that during the heat period, the cow body emits odors which are different from those emitted outside of this period [28]. The scientific community have investigated this axis and proposed various methods to analyze the volatile compounds emitted by the cow [165, 214, 124, 68, 175, 143]. More precisely, in [175, 214, 143], the developed heat detection systems exploit electronic noses technology to analyze odors changes from perineal and perigenital samples collected using special cotton bud swabs. In the system proposed recently by Manzoli et al. [165], authors designed and made their own electronic nose with a particular attention to its sensitivity and reversibility properties against humidity. They adopted a principal component analysis (PCA) technique for the heat state detection. In [124, 68], authors proposed to train sniffer dogs to distinguish between vaginal mucus samples of cows in heat and in no-heat. In the case of Dorothea et al. [124], they developed a specific training protocol to train six dogs, each one during 50 hours. Nevertheless, the authors highlighted in the perspectives the need of optimizing the training protocol as well as the training hourly volume to improve the detection results. Overall, one of the drawbacks of this category of methods is the need of a human operator intervention to prepare samples which make the process laborious especially within herds of large size.

Video monitoring– Several video analysis based automatized monitoring systems for heat detection have been developed [189, 105, 279, 240]. The proposed systems exploit video cameras installed in the farms to analyze the behavior of cows and identify those expressing heat signs notably the riding. To reach this goal, the developed systems consist of a standard processing pipeline including the scene segmentation (extraction of objects of interest namely cows), cows recognition and tracking, behavior classification for heat detection. In [189], the behavior classification step is achieved by classifying a set of feature points extracted automatically from the cow body parts. The authors experimented multiple learning algorithms for the classification task and reported that the Support Vector Machine algorithm with Radial Basis Function gave the best results in term of heat detection. It is worth mentioning that this type of methods aims to mimic the human method for heat detection namely the visual observation. Hence, the short videos of interest (i.e. the ones identified by the system as containing heat signs) may be rapidly double checked and validated by the farmer which represent a real asset for him in order to improve the accuracy of the heat detection. However, the main drawback with these methods is that they are unable to detect cows in silent heat. Additionally, they can hardly be exploited on grazing cows due to cameras deployment issues.
The wealth of the presented state of the art on the heat detection within cattle clearly shows the great significance of this task. Moreover, the continuous development of new automatic systems show that this task is challenging and not solved yet due to many factors related to the physiological state of the cows and the farms environment. In our work, we present a new analysis way for detecting heat. More precisely, the proposed systems in the literature are based either on the behavior changes analysis or the odor changes analysis, while the system we propose is based on vaginal endoscope imagery deep analysis. One of the main advantages of our system is its all in one device offering to the farmer both of heat detection and the artificial insemination functions.

3.2.2 Image classification

In our work, we define the heat detection in cow as an image classification problem, which is one of the most popular task of computer vision. The main goal of image classification is identifying if a given object appears in an image. In our case, we aim

at identifying heat features from cow vaginal endoscopic image to classify two categorie of heat state of cow, i.e heat or no heat.

3.2.2.1 Public datasets and benchmarks

In the context of deep learning based computer vision, the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) evaluates algorithms for object detection and image classification at large-scale. The goal of image classification in this challenge is to train a model that use a dataset namely ImageNet [56], which has roughly 1.2 million images for training, 50,000 for validation, and 100,000 for testing, capable of classifying images into 1,000 different classes. Since 2012, almost all the proposed methods are validated on ImageNet [56] to demonstrate their performance. Moreover, fine-tuning a model pre-trained on ImageNet has been considered as a common CNN model training strategy.
In addition to these benchmark datasets, there are also a lot of special dataset. Dogs vs. Cats dataset is a competition dataset of Kaggle [3], which is comprised of photos of dogs and cats provided as a subset of photos from a much larger dataset of 3 million manually annotated photos. For livestock, several datasets of cattle or cow [12, 72, 4] and pig [4] have been built for animal recognition and animal behavior classification. Furthermore, endoscopic image classification is also one of the main application of image classification, which has been used to several diseases diagnosis of digestive system in humain. Pogorelov et al [193] created a endoscopic image dataset for computer aided gastrointestinal disease detection namely Kvasir, which has two versions. They have 4000 and 8000 images of 8 classes, respectively, which can be used for anatomical landmarks recognition, pathological findings and polyps recognition.

3.2.2.2 Deep learning based endoscopic image classification

Endoscopy-aided techniques have been used in the health checks and surgical treatments of the pet (dog, cat) [71] and livstock animal (cows, mares and sheep) [216, 162]. This technique is a noninvasive method for a pathological diagnosis of living tissue. However, it can produce thousands images during one examination at least. The reading of a large amount of endoscopic image data has exceeded the limit of human attention, which is easy to cause misdiagnosis and reduce the diagnostic accuracy.

[3] Kaggle DogVCat Competation: http://www.kaggle.com/c/dogs-vs-cats

Moreover, it requires a diagnosis from experienced expert or veterinarian.
In the context of the great success of deep learning in the field of computer vision, to face this issue, some computer aided diagnosis (CAD) systems have been developed in humain gastrointestinal (GI) disease disgnosis. By using the Kvasir dataset presented in section 3.2.2.1, several CNN models have been trained for anatomical landmarks recognition, pathological findings and polyps recognition [192, 3, 164] . Furhermore, corresponding models have also been developed for the disgnosis of gastric cancer [104], recognition of celiac disease [276], hookworm [97], and small intestine motility characterization [217]. Indeed, deep learning based endoscopic image analysis has become widespread in humain disease diagnosis. Nevertheless, there is very little deep learning based CAD systems for animal endoscopic image analysis. Therefore, this work will be the frist deep learning based methodology for cow heat analysis from endoscopic images.

3.3 Methodology

The use of deep learning techniques for image analysis has become widespread over multiple domains including agriculture one [128]. Indeed, these techniques have shown their efficiency in several tasks including image classification and segmentation while providing higher performance in comparison with traditional techniques of image analysis. In this context and to tackle the problem of heat detection from an endoscopic video, we propose a method based on a deep learning approach that permits to analyze the video and to assign the appropriate class, i.e. “heat” or “no-heat”. Algorithm 1 summarizes the different steps of our method. First, the video is subdivided into a sequence of individual frames (or images) according to a given image rate, each frame is then standardized in term of resolution and normalization in order to feed a CNN classifier, which permits to predict a class. The decision with respect to the final class of the video is attributed based on a major voting which allows to identify the dominant class over the sequence. A heat state prediction is also calculated as shown in steps 14 and 17 of our algorithm.

Algorithm 1 InceptionVGG8-based cow heat state prediction

Input: Cow vagnial endoscopic video (V), FPS rate (R)
Output: Decision (D), Heat state Prediction (HSP)

$L_F \leftarrow (empty\ list\ of\ frames)$
$L_F = \text{SPLIT}(V, R)$
$L_F = \text{STANDARDIZATION}(L_F)$
for each f *in* L_F **do**
    $P_{heat}, P_{no-heat} = \text{INCEPTIONVGG8}(f)$ ▷ Model in Figure3.2
    if $P_{heat} > P_{no-heat}$ **then**
        $Total_{heat} + 1$
    else
        $Total_{no-heat} + 1$
    end if
end for
if $Total_{heat} > Total_{no-heat}$ **then**
    D = Heat
    HSP $= \frac{Total_{heat}}{\text{SIZE}(L_F)}$
else
    D = No-Heat
    HSP $= \frac{Total_{no-heat}}{\text{SIZE}(L_F)}$
end if

3.3.1 CNN backbone

To build the main core of our system namely the image classifier, we designed and developed a variant of the VGG16 CNN model[227]. As illustrated in Figure 3.2 our variant named "InceptionVGG8" differs from the original architecture on 2 major points:

- The convolution blocks 2×128, 3×256, 3×512 and 3×512 are replaced by a customized inception module which is a variant of the block A of InceptionV4 architecture [234].
- The two dense layers of 4096 dimensions are replaced by a global average pooling and a dense layer of 2 dimensions.

The choice of a VGG16-like architecture is explained by its shallow design which offers the advantage of reducing over-fitting on image datasets of limited quantity. The inception module offers through its multiple channels the advantage of being able to extract image features at different scales, which should reinforce the generalization

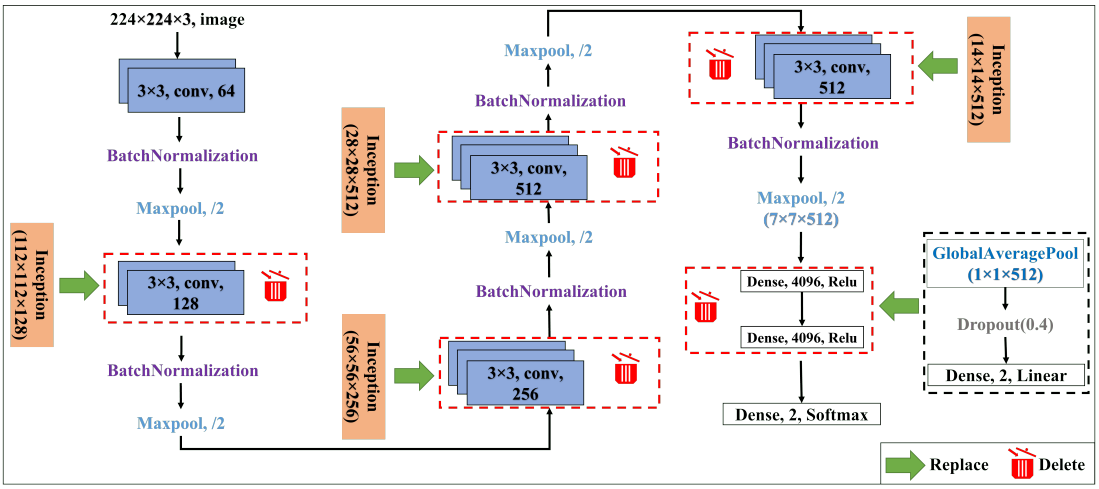


FIGURE 3.2: Design of our CNN model: from VGG16[227] to InceptionVGG8.

potential of the CNN model. The global average pooling and the dense layer of 2 dimensions allow to reduce the model complexity. The design of our inception module is shown in figure 3.3. In comparison with the block A of InceptionV4 architecture, we have: i) removed data downsampling step (average pooling) from the first channel to avoid loss of possible relevant information, ii) extended the depth of channels 2, 3 and 4 by adding one convolution block with a kernel size of 3×3 in order to increase the size of their receptive fields. Hence, these changes have permitted to design a 4-scale feature extractor module namely 1×1, 3×3, 5×5 and 7×7 via channels 1, 2, 3 and 4, respectively.

Each convolution layer of our architecture is represented by a stack of feature maps. One feature map contains $(N \times N)$ units (or neurones), each one associated with a *Relu* (Rectified linear unit) activation function calculated as follow:

$$x = max(x, 0) \tag{3.1}$$

where x is resulting from a convolution operation between the feature maps $f_{k=1,m}$ of the previous layer and associated filters of weights (kernels):

$$x = (\sum_{f_{k=1,m}} (\sum_{i,j}^{w,h} w_{ij} \times x_{ij})) + b \tag{3.2}$$

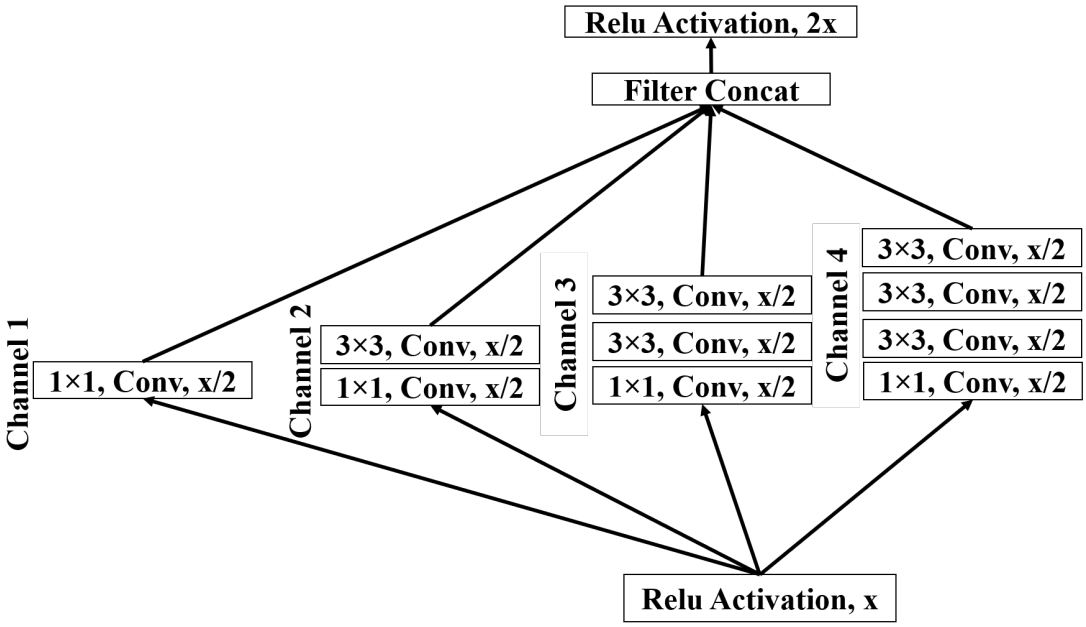


FIGURE 3.3: Proposed inception module for the InceptionVGG8 architecture. The x is the number of layers.

Here, the (w,h) correspond to kernel width and height, w_{ij} the weight at position i,j of the kernel, x_{ij} the unit at position i,j of the considered feature map and b is the bias. Each convolution layer of the architecture is followed by a batch normalization [116] which corresponds to a normalization of the feature maps over a batch of images. The normalization algorithm is illustrated in figure 3.4 for one unit x and may be generalized to a feature map by replacing x by f.

A max pooling function follows batch normalization step which permits to half the size of each feature map while keeping its most significant features. To this end, a square window of 2x2 units is shifted with a step of 2 over the feature map and output each time the unit having the maximum value. The last max pooling is followed by a Global Average Pooling (GAP) function which is calculated for each feature map $f_{k=1,512}$ as follow:

$$GAP(f) = \frac{1}{7\times 7}\sum_{i,j}^{7} x_{ij} \tag{3.3}$$

After GAP function comes a dropout function (0.4 to cancel in the training stage the contribution of 40% of the units picked randomly in order to reduce over-fitting [230]),

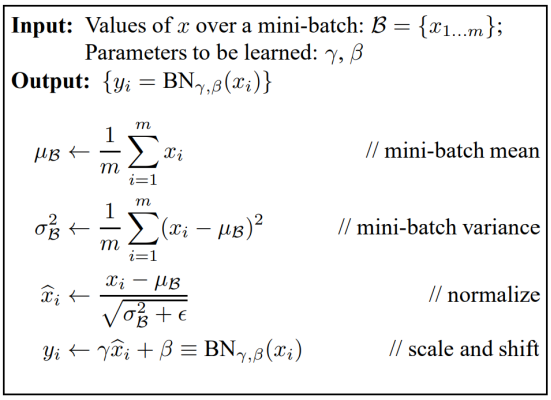


FIGURE 3.4: Batch normalization algorithm over a batch of images. Image extracted from [116].

a fully connected layer (with identity function) and an output layer having both a size of 2 neurones. Each neurone $x_{i=1,2}$ of the output layer permits to calculate the corresponding prediction probability $p_{i=1,2}$ of the input image being in class $c_{i=1,2}$ using a $softmax$ function defined as follow:

$$p_i = softmax(x_{i,c_i}) = \frac{e^{x_i}}{\sum_{j=1,2} e^{x_j}} \tag{3.4}$$

3.3.2 Loss function

In the CNN training, the loss function is used to quantify the difference between the expected outcome and the one predicted by the CNN model. Furthermore, gradients derived from the loss function is used to update the weights of model. In our case, the overall architecture has been trained to reduce the categorical cross entropy loss (*Loss*) which is calculated for the output layer over a batch of images as follow:

$$Loss = -\frac{1}{n}\sum_{n}\Big(\sum_{c=1,2} y_c log(p_c)\Big) \tag{3.5}$$

where n corresponds to the batch size, y_c is a binary indicator whether the class c is the correct one for the given image from the batch and p_c is the probability predicted (see equation 3.4) by the model on the same image.

3.3.3 Optimization function

The CNN training aims to minimize the difference between the expected outcome and the one predicted by the CNN model, which require an optimization function. In our case, the weights are optimized using a Root Mean Square propagation technique (RMSprop) [47]. We denote $w_{ij_{t+1}}$ a weight to update during the forward pass (from step t to $t+1$ of a training epoch) between two neurones x_i and x_j of our architecture. Based on RMSprop technique, the weight is updated as follow:

$$\begin{cases} w_{ij_{t+1}} = w_{ij_t} - (\frac{\eta}{\sqrt{v_t+\varepsilon}} * \delta_{j_t}) & \\ v_t = \beta v_{t-1} + (1-\beta)\delta_{j_t}^2, & \beta = 0.9, v_0 = 0 \\ \eta = 10^{-3}, & \text{learning rate} \\ \varepsilon = 10^{-8}, & \text{calculation stability} \end{cases} \tag{3.6}$$

where δ_{j_t} is the gradient error of the neurone x_j back-propagated from the *Loss* of the predicted class of the input image at step t.
To find the best optimization function which is suitable for our model, we consider also other 5 common optimization functions (SGD, Adadelta, Adam, Adamax, Nada)[47] (see in section 3.4.2)

3.4 Experimental study

3.4.1 The generated endoscopic image dataset for cattle heat analysis

This dataset is provided and established by Gènes Diffusion [4], CECNA [5], and Elexinn companies [6]. Gènes Diffusion and CECNA are pioneering companies in the field of genetics and animal reproduction, Elexinn is a start-up developing the gynecological gun (Eye Breed), these companies have been a key partner in animal husbandry in France.

[4] https://www.genesdiffusion.com/
[5] http://www.cecna.fr/
[6] http://www.cecna.fr/category/elexinn/

3.4.1.1 Eye Breed technology for data acquisition

Eye Breed [55] is the first innovative technology that permits insemination without rectal search, which is considered as a painful procedure . Indeed, it has been designed and developed with special attention to the animal well-being, ease of use by the operator and data visualization/connection capabilities. Specifically, the device is equipped with 1) an embedded camera to visualize the entrance of the uterine cervix and 2) an insemination tube with atraumatic nozzle allowing to optimize the operator's work and the animal's conditions. The device has been used to record videos of simulated insemination using the default parameters of the camera notably an image rate of 20 FPS (frames per second) and an image resolution of 640×480 as shown in Figure 3.5.

FIGURE 3.5: Artificial insemination illustration using Eye Breed technology[55].

3.4.1.2 Data preparation and preprocessing

To build the dataset, 46 videos, representing 46 Holstein dairy cows, have been collected from several farms located in the north region of France between 2017 and 2018. For each cow, the device has been introduced by a human operator to carry out the insemination operation without triggering it (i.e. without semen injection). It is worth

mentioning that the time of this operation depends on both the human operator experience in the use of the Eye Breed device as well as the anatomy of the cow genital tract. For these reasons, the length of the collected videos varies between 40 and 80 seconds. The heat state of the considered cows has been identified by CECNA experts which allowed to label each video into the appropriate class. The labeling step gave 36 videos with "heat" label and 10 videos with "no-heat" label. Two videos from each class have been selected randomly and excluded from the set for a final test of the global system. The 4 videos represent 5600 images. The remaining 42 videos (34+8) have been used to create the training and validation sets with balanced classes. In total, 21760 images, artificially augmented to 239360 images, have been generated for the training set and 4000 images for the validation set. The pre-processing chain that allowed to generate the image sets is described in what follow.

Video split – Each video is split into a set of images (or frames) with a rate of 20 images per second in accordance with the image rate of the embedded camera in the Eye Breed device. Each set of images is then manually checked in order to remove noisy images which are generally positioned at the beginning or at the end of the videos. As illustrated in figure 3.6, these images correspond to the ones captured from outside of the cow genital tract. This cleaning step, produced a total of 32000 and 12880 images for heat and no-heat classes respectively. The produced set of images for both classes has been then subdivided into two subsets: 10880×2 images picked randomly for training and 2000×2 for validation. At this stage, we have a training set of 21760 images and a validation set of 4000 images with balanced classes for both sets.

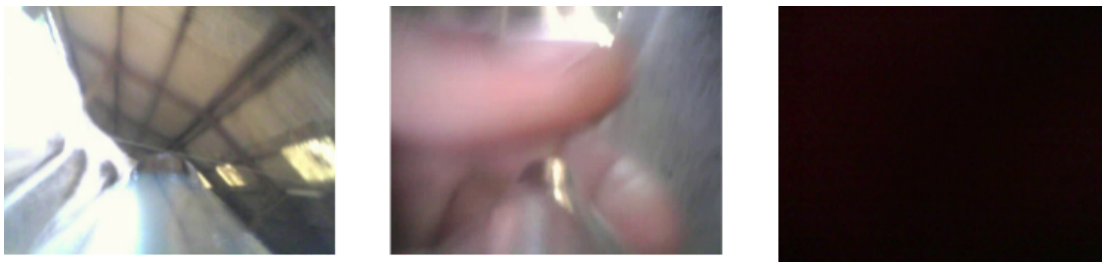


FIGURE 3.6: Examples of noisy images.

Data standardization – In order to be in accordance with the standard image resolution required by the designed architecture, we have down-scaled the dataset images to a square resolution of 224 x 224. We first applied a partial mirroring copy technique

to move from 640×480 which is the original resolution of the images to a square resolution of 640×640 (see Figure 3.7), then down-scaled the images based on a bi-linear interpolation to reach the targeted resolutions namely 224×224. Indeed, this processing methodology allows to avoid image distortion. All the images have been normalized in the range $[0, 1]$.

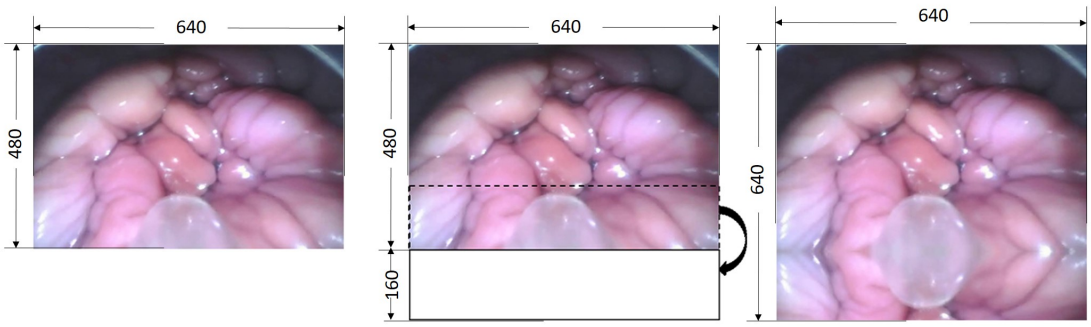


FIGURE 3.7: Partial mirroring copy technique illustration.

Data augmentation– Previous works have shown that data augmentation is an effective technique to increase the learning capacity of deep architectures and limit the over-fitting [225]. For this reason, we have artificially augmented our training set from 21760 to 239360 images by applying 10 simple label-preserving transformations. Table 3.1 summarizes the different transformations together with their associated parameters . An example of each transformation is illustrated in Figure 3.8.

Type	Range
Rotation	[-30, +30]
Width shift	0.1
Height shift	0.1
Shear	0.2
Zoom	[0.8, 1.1]
Horizontal_flip	NA
Vertical_flip	NA
Contrast enhancement	2.0
Brightness	[0, 2.0]
Channel shift	10.0

TABLE 3.1: Summary of label-preserving transformations used for data augmentation.

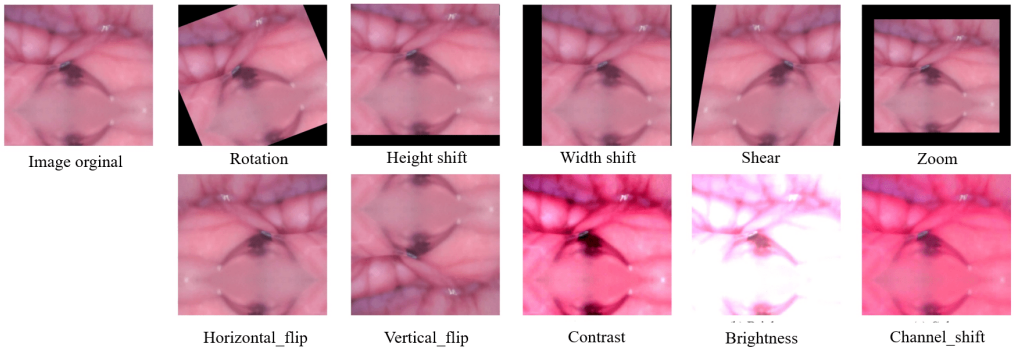


FIGURE 3.8: Illustration of the ten label-preserving transformations from an original image.

3.4.2 Ablation study

In order to evaluate the effectiveness of our CNN model components and the data augmentation, we have conducted an ablation study. First, we studied the influence of 6 common optimizers (Adadelta, Adam, Adamax, Nadam, RMSprop and SGD)[47] on the performance of the model. To this end, we trained our architecture over 500 epochs on the 239360 training image set. Table 3.2 presents the obtained classification accuracies on the validation set (4000 images). One can observe that the RMSprop outperforms the other optimizers with an accuracy of 99.47%.

Adadelta	Adam	Adamax	Nadam	RMSprop	SGD
96.72	96.21	98.82	95.51	**99.47**	95.03

TABLE 3.2: Classification accuracies in (%) of our CNN model on the validation set (4000 images) using different optimizers.

We have then removed from our model the batch normalization (BN), the GAP and the dropout functions and trained it using the best optimizer (RMSprop). Table 3.3 summarizes the obtained accuracies. The results show that removing the 3 functions from our model caused a drop in its performance with notably an accuracy loss of 2.77%.

We have also analyzed the impact of changing the inception module. To this end, we have replaced this module by the ones proposed in the inceptionV series (V1 [233], V2[235], V3[235] and V4[234]) and trained our model accordingly. We emphasize

No BN-GAP	No BN-GAP-Dropout	Original
97.9	96.7	**99.47**

TABLE 3.3: Classification accuracies in (%) of our CNN model on the validation set (4000 images) by leaving out some components.

that the V3 and V4 architectures include several inception modules. For this reason, we tested them separately in our architecture and we kept the ones giving the best performances. Table 3.4 summarizes the obtained accuracies on the validation set. The table shows that our proposed module (original) gives the best classification accuracy. Nevertheless, we may observe that our model still gives a high accuracy when using the inception module of V4 architecture named block A in [234]. This result was expected due to the design similarity between this block and our inception module.

InceptionV1	InceptionV2	InceptionV3	InceptionV4	Our Inception
89.74	94.5	89.88	98.89	**99.47**

TABLE 3.4: Classification accuracies in (%) of our CNN model on the validation set (4000 images) by replacing its inception module by the ones of InceptionV series architectures.

Finally, we have studied the impact of data augmentation on our model. To this end, we have trained it following 4 scenarios of data augmentation:

- NDA for no data augmentation and using only the original 21760 images,
- TDA for texture data augmentation (contrast, brightness and channel shift) with a total of 87040 images,
- SDA for spatial data augmentation (rotation, height shift, width shift, shear, zoom, horizontal flip and vertical flip) with a total of 174080 images,
- TSDA for both texture and spatial data augmentations with a total of 239360 images.

The obtained accuracies on the validation set according to each scenario are summarized in Table 3.5. The table shows that texture data augmentation scenario has permitted to the trained model to correctly classify the validation set at 100%. Indeed, this latter scenario improved the model accuracy by 8.5% in comparison with no data

augmentation scenario.

NDA	TDA	SDA	TSDA
91.5	**100**	97.4	99.47

TABLE 3.5: Classification accuracies in (%) of our CNN model on the validation set (4000 images) following 4 scenarios of data augmentation.

3.4.3 Performance comparative study

We have conducted on our dataset an extensive comparative study between the performances of our CNN model and those of 19 methods from the state of the art. More precisely, we considered 14 existing CNN models: VGG16 [227], VGG19 [227], InceptionV1 [233], InceptionV2 [235], InceptionV3 [235], InceptionV4 [234], IV1VGG14 [250] (a variant of VGG16 exploiting an InceptionV1 module), ResNet50 [98], InceptionResNetV2 [234], Xception [48], MobileNetV3 [111], DenseNet201 [113], NASNetMobile [191] and EfficientNetB7 [236]. The architectures have been adapted to our problem by: i) setting the output layer to 2 classes, ii) applying a transfer learning strategy from the ImageNet dataset [56], except for IV1VGG14 [250]. Indeed, we have implemented from scratch this latter architecture since it is not publicly available. We emphasize that each CNN model has been trained and validated by testing 6 common optimizers (Adam, SGD, RMSprop, Adadelta, Adamax and Nadam)[47]. In addition to CNN models, we have compared our model with 5 image descriptor methods: i) two moment-based descriptors namely Fractional-order Jacobi-Fourier Moments (FJFM) [264] and Fast Quaternion Generic Polar Complex Exponential Transform (FQGPCET) [263], ii) Riemannian covariance descriptor (RieCovDs) [46], iii) two texture based descriptors namely Local Binary Pattern with Histogram Refinement (LBPHR) and Local Derivative Pattern with Histogram Refinement (LDPHR) [238]. Each descriptor has permitted to train 2 classifiers based on KNN (K Nearest Neighbor) and SVM (Support Vector Machine) machine learning techniques. In total, 94 classifiers have been trained namely 84 CNN-based ones and 10 descriptor-based ones for being compared with our model.

The performances of these classifiers have been evaluated based on standard metrics namely recall, precision, specificity and accuracy. We define TP (True Positive) the

number of heat image samples identified by the classifier as heat, TN (True Negative) the number of no-heat image samples identified by the classifier as no-heat, FP (False Positive) the number of no-heat image samples identified by the classifier as heat, FN (False Negative) the number of heat image samples identified by the classifier as no-heat. The metrics are calculated as follow:

- Recall (Sensitivity): the capacity of the classifier to identify heat images.

$$Recall = \frac{TP}{TP+FN} \tag{3.7}$$

- Precision: the probability of an image to represent a heat state when it is identified by the classifier in heat.

$$Precision = \frac{TP}{TP+FP} \tag{3.8}$$

- Specificity: the capacity of the classifier to identify no-heat images.

$$Specificity = \frac{TN}{FP+TN} \tag{3.9}$$

- Accuracy: the ratio of correct predictions.

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \tag{3.10}$$

Table 3.6 shows the performances of our model and those of the 19 methods obtained on the validation set. We emphasize that we present for each of these methods the results of the best classifier. One can observe that our model outperforms all the methods with an accuracy of 100% offering a gain between 3% and 47% in comparison with the other methods. An unexpected result is the ranking of a moment descriptor-based method namely FQGPCET-KNN [263] at the second position giving an accuracy of 96.72% and outperforming 14 CNN-methods.

Table 3.7 reports the results of the top 7 methods from Table 3.6 (methods that gave an accuracy greater than 90%) in term of heat state prediction on the test set composed of 4 videos as well as the classification accuracy over the corresponding 5600 images. We remind that the heat state prediction of a video corresponds to the ratio between the frames of the dominant class predicted by a method and the total number of frames of the video. The table shows that except the FQGPCET-KNN[263] method which

Category	Method	Recall	Precision	Specificity	Accuracy
Descriptor	FJFM-SVM [264]	55.35	63.76	68.55	61.95
	FQGPCET-KNN [263]	94.9	98.49	98.55	96.72
	Riemaniancov-SVM [46]	80.15	75.47	73.95	77.05
	LBPHR-KNN [238]	88.15	64.1	50.56	69.4
	LDPHR-KNN [238]	83.75	58.73	41.15	62.45
CNN	InceptionV1 [233]	93.2	97.5	97.6	95.4
	InceptionV2 [235]	95.75	96.47	96.5	96.13
	InceptionV3 [235]	50.35	63.77	71.4	60.7
	InceptionV4 [234]	62.25	53.62	43.95	53.1
	InceptionResNetV2 [234]	93.95	94.27	94.3	94.12
	Xception [48]	59.8	78.68	83.8	71.8
	MobileNetV3 [111]	67.1	68.19	68.7	67.9
	VGG16 [227]	83.3	94.43	94	91
	VGG19 [227]	88.4	94.55	94.9	91.67
	ResNet50 [98]	77.4	94.97	95.9	86.65
	DenseNet201 [113]	58.9	61.32	62.85	60.88
	NASNetMobile [191]	65	69.15	71	68
	EfficientNetB7 [236]	49.55	52.52	55.2	52.38
	IV1VGG14 [250]	**100**	50.59	29.12	58.92
	Our model	**100**	**100**	**100**	**100**

TABLE 3.6: Comparison between the performances in (%) of our CNN model and those of 19 methods from the state of the art obtained on the validation set (4000 images).

miss-predicted video_4, the other methods have correctly predicted the heat state of the 4 videos. Nevertheless, our method gives the highest performances on the 4 videos with notably a classification accuracy of 99.5% over the 5600 corresponding images. Indeed, it offers a gain between 5% and 23%.

It is worth mentioning that to the best of our knowledge there exist no public endoscopic image dataset for cow heat analysis. Nevertheless, to examine the efficiency of our model on other datasets, we have trained and tested it on a public dataset of endoscopic images for human digestive system analysis [193]. Indeed, this dataset named kvasir version 2 has been the subject of several works from the state of the art [164, 192, 3]. It is composed of 8000 images representing 8 classes and covering anatomical landmarks, pathological findings or endoscopic procedures inside the gastrointestinal tract. Table 3.8 summarizes the global performances obtained by our model on the test set of this dataset and those reported in [164, 192, 3]. The obtained results show that the methods proposed by Majid et al. [164] and by Pogorelov et al. [192] have

Decision system	Heat class		No-Heat class		Accuracy
	video_1	video_2	video_3	video_4	
FQGPCET-KNN [263]	86.75	78.45	95.1	47.62	76.35
VGG16 [227]	78.4	89.5	89.9	94.2	89.31
VGG19 [227]	81.72	83.1	91.51	93.33	88.27
InceptionV1 [233]	91.8	98.38	99.36	86.15	94.41
InceptionV2 [235]	91.55	100	94.39	90.26	94.42
InceptionResNetV2 [234]	80.62	98.45	96.16	92.51	93.58
Our model	**97.8**	**100**	**100**	**99.3**	**99.5**

TABLE 3.7: Results in (%) of heat state prediction on the test set composed of 4 videos* as well as the classification accuracy over the corresponding 5600 images.

* duration: video_1 (40s), video_2 (77.75s), video_3 (86s), video_4 (76.16s).

competitive performances. Nevertheless, our model outperforms them notably in term of accuracy (97.8%).

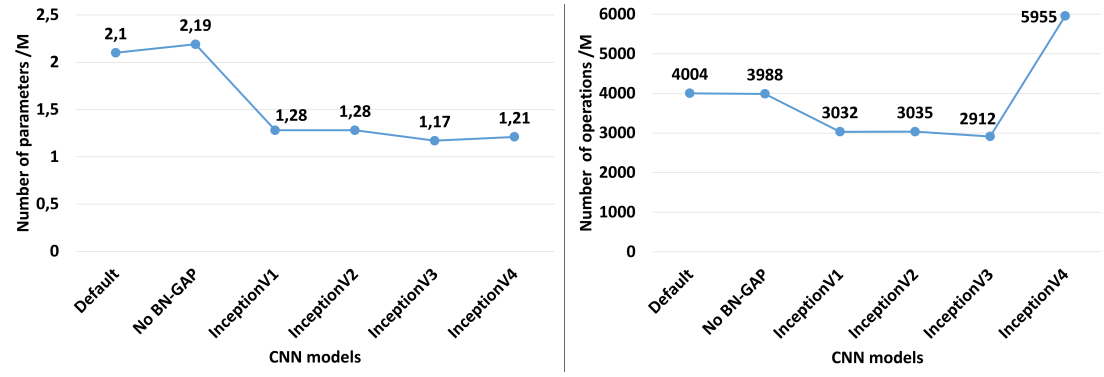
Method	Recall	Precision	Specificity	Accuracy
[164]	-	**96.5**	-	96.5
[192]	87.2	87.2	97.4	95.7
[3]	75.5	75.4	-	-
Our model	**92.1**	92	**98.9**	**97.8**

TABLE 3.8: Comparison between the performances in (%) of our CNN model and those of 3 methods from the state of the art obtained on the Kvasir dataset [193].

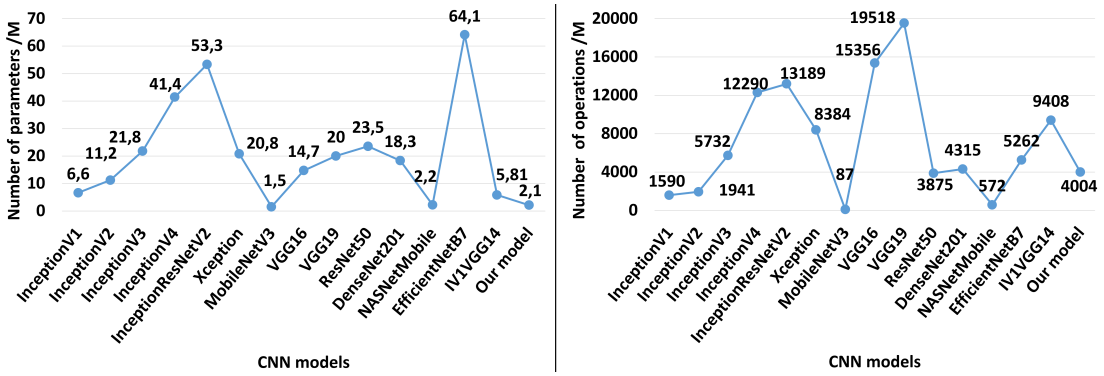
3.4.4 Complexity comparative study

We have calculated the complexity of our model and those of the 14 CNN models indicated in Table 3.6. More specifically, following the study conducted by Howard et al. [111] on the complexity of their MobileNetV3 CNN model, we retained the same complexity measures they have proposed namely the number of parameters to optimize, the number of MAdds (MultiplyAdds) operations and the latency of the model. Based on these measures, we have calculated the complexity of our model and compared it with: i) our model variants reported in Tables 3.3 and 3.4 from the ablation study section, ii) the 14 CNN models from the state of the art indicated in Table 3.6. We remind that the architectures of the associated 11 CNN models have been adapted to our problem, hence the calculated complexities are different from the ones reported in the original articles. The comparison results in term of parameters and operations

are shown in Figure 3.9(see Figure 3.9a for the model variants and Figure 3.9b for the 14 models) From Figure 3.9a, we may observe that the complexity level of the variants



(a) Our model vs. its variants reported in the ablation study



(b) Our model vs. 14 CNN models from the state of the art

FIGURE 3.9: Parameter complexities (left side) and operation complexities (right side) in million of our CNN model and those of 14 CNN models from the state of the art.

of our model notably Inceptions V1-V3 is slightly better than that of the default version for both parameters and operations. Nevertheless, it is worth mentioning that this difference shows a negligible impact on the improvement of the model latency which is equal to 10.71 milliseconds. Figure 3.9b shows that in term of number of parameters to optimize, our model is ranked at the second position after the MobileNetV3 model. At this level, we observe that our model has a better complexity than the VGG, the Inception and InceptionVGG architectures. In term of number of operations, our model is ranked at the fifth position after MobileNetV3, InceptionV1, InceptionV2, and ResNet50 models. Nevertheless, Table 3.9 shows that the latency of our model is close to the one from InceptionV1 and InceptionV2 which are the only models among the 4 mentioned previously that exceeded an accuracy of 90% on the test images (see

Table 3.7). We precise that the latencies have been calculated on NVIDIA Tesla V100 GPU with 32 GB of memory and a batch size equal to 1.

InceptionV1 [233]	InceptionV2 [235]	Our model
7.69	8.68	10.71

TABLE 3.9: Latency in milliseconds of our model compared with the top 2 CNN models in term of accuracy from Table 3.7.

3.4.5 Visual analysis of the decision of our model

In order to highlight on an image the regions of interest that drive our model to predict a specific class, we have exploited the Grad-CAM (Gradient-weighted Class Activation Mapping) method [218]. To conduct our experiment, we applied the Grad-Cam on the images of the 4 videos from the test set. Figure 3.10 shows an example of the generated heat maps for three images extracted from each video. The images are located at the beginning, middle and end of the videos. We precise that the classes of the selected images have been correctly predicted by our model. The figure shows

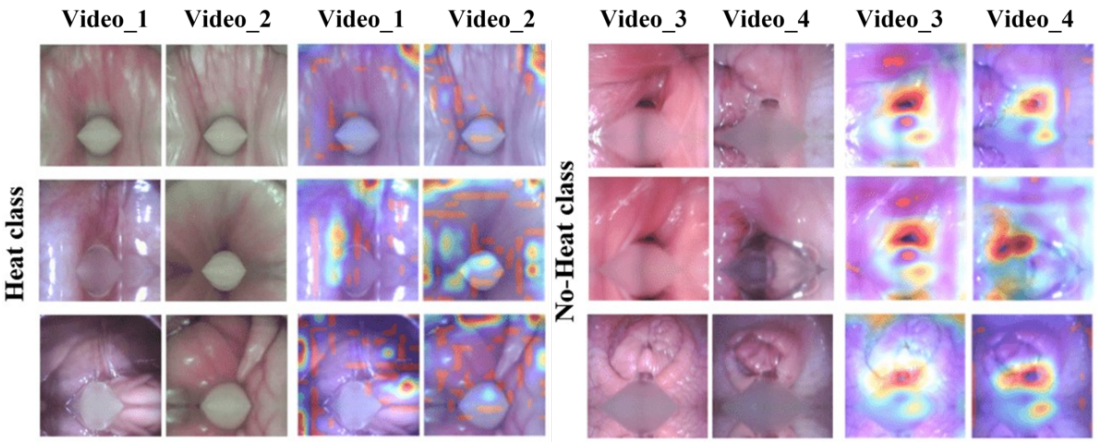


FIGURE 3.10: Examples of original images correctly classified by our model and their associated class activation maps. From top to bottom, the images are located on the begin, middle and end of the videos.

that our model focuses on the borders of the image for detecting the heat class and approximately in the center for the no-heat class. A possible explanation of this way of decision is the fact that for "Heat" class, the heat tends to be overspread over the image surface including borders while for "No-heat" class, the heat is absent in the borders

and in the center of the image.

Interestingly, Figure 3.11 shows an example of maps generated on misclassified images by the InceptionVGGNet from video_1 and video_4. One may observe that the positioning of ROIs on these images by the classifier has been altered which explain its wrong prediction. This is probably due to a noise, invisible to the human eye, that affected the characterization of these images by the classifier.

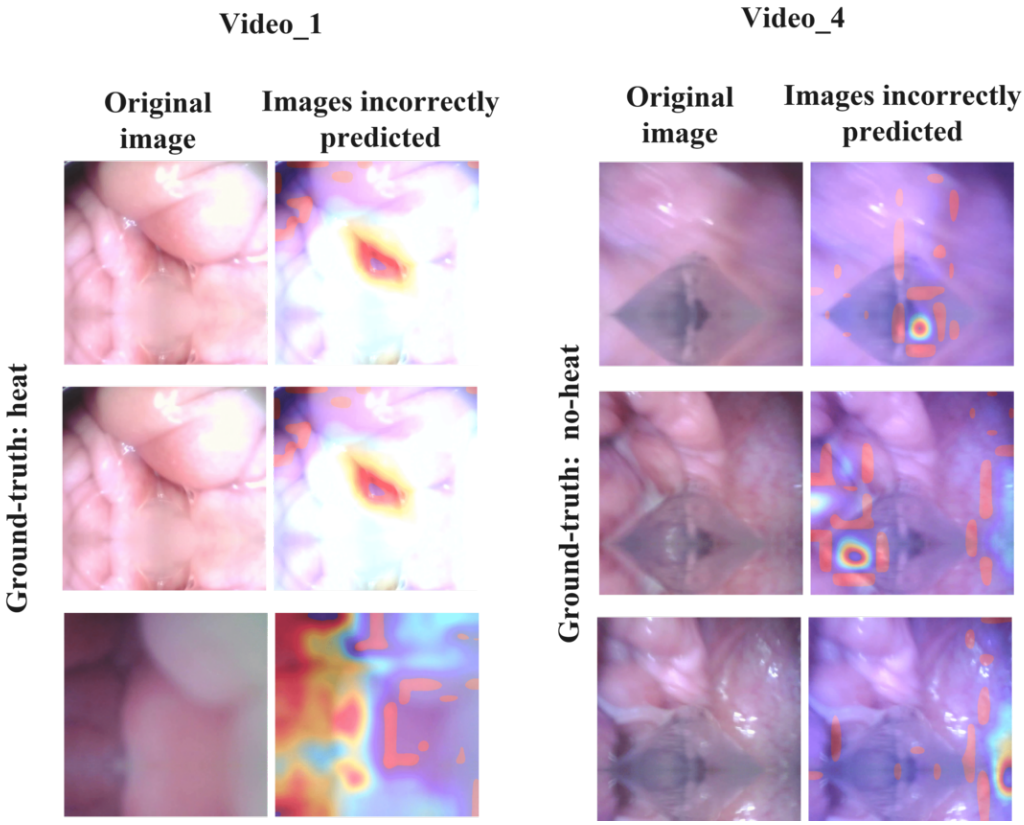


FIGURE 3.11: Examples of original images misclassified by the InceptionVGGNet classifier and their associated class activation maps. From top to bottom, the images are located on the begin, middle and end of each video.

3.4.6 Android deployment and optimization

The overall InceptionVGG8 based decision system has been deployed on an Android HUAWEI MATE 30 PRO smart-phone with the following characteristics: Kirin 990 Hisilicon 8-core with 8 GB of RAM and a Mali-G76 GPU. A video demonstration of our smart-phone application is available in the supplementary material. To shorten

the decision time of our model, we have optimized it using a post-training quantization technique [178] and we have exploited the TensorFlow Lite framework[7] to support smart-phone GPU acceleration. Moreover, as our model has shown a high efficiency in term of image classification, we made the choice to downsample the videos to analyze by lowering the image rate from 20 FPS to 1 FPS. Table3.10 summarizes the performances and the running times obtained by the Android raw version and the optimized one named QGAVD (Quantization + GPU Acceleration + Video Downsampling) on the test videos. As shown in the table, the QGAVD version has permitted to drastically improve the decision time of our system, decreasing from several minutes to just a few seconds, while keeping a high level of performances.

Android CNN version	Heat class				No-Heat class			
	video_1		video_2		video_3		video_4	
	HSP	RT	HSP	RT	HSP	RT	HSP	RT
Raw	97.8	06:11	**100**	11:40	**100**	12:54	99.3	11:25
QGAVD	95.12	**00:08**	**100**	**00:14**	**100**	**00:16**	**100**	**00:14**

* duration: video_1 (40s), video_2 (77.75s), video_3 (86s), video_4 (76.16s).

TABLE 3.10: Results in (%) of heat state prediction (HSP) and associated running time (RT) in mm:ss format obtained by the two versions of our Android CNN model on the test set composed of 4 videos*.

3.4.7 Generalization of model

To test the generalization ability of our model, 55 videos (27 videos labeled as “heat” and 28 videos labeled as “no-heat”) have been recently recorded on 2020 for cow heat state prediction. The results show that our model correctly predicted 21videos of 27 videos labeled as “heat”, and our model correctly predicted 13 videos of 28 videos labeled as “no-heat”, which is insuficant for pratical use. We noticed that the quality of the 55 recently recorded videos is better than previous ones. Furthermore, the image rate of cameras used to record the new videos is 100 FPS (frames per second) instead of 20 FPS.

Therefore, It is necessary to retrain our model with these new data to enchance the robustness and generalization of our model. To this end, we fristly selected eight videos from each class of the new data, half of the selected videos are incorrectly predicted with previous our model. With the original four test videos, we have a total of 10

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

videos per class for a final test of the retained model. We then proposed four different training strategies with different protocols of dataset creation

Training from scratch – Our model is trained from scratch with two different protocols of dataset creation as following:

- **SCS** – Based on the previous dataset, we supplement 13432 (6716 × 2) new images into training set and 1496 (748 × 2) images into into validation set from splitting of new videos with balanced classes. In total, 35192 ((10880+6716) × 2) images have been generated for training set, 5496 ((2000+748) × 2) images have been generated for validation set.
- **SCR** – Eliminate 20 pre-selected videos for testing, we reintegrate the remaining 81 videos (42 original videos, 39 new videos; 53 videos labeled as "heat", 28 videos labeled as "no-heat") to split into images for dataset generation. In total, 35192 (17596 × 2) images have been generated for training set, 5496 (2748× 2) images have been generated for validation set. It should be pointed out that the number of images labeled as heat is far larger than the number of images labeled as no-heat. Therefore, compared with the first strategy, the training and validation set of "no-heat" have been remained. To keep balanced classes, we randomly chose the same number of image from the images which are generated from the 53 videos labeled as "heat" for the training and validation set of "heat" class.

Transfer learning from previous data – We only use the new images splitted from the new videos for dataset generation with two transfer learning strategies. In total, 13432 (6716 × 2) images have been generated for training set, 1496 (748 × 2) images have been generated for validation set. We use two transfer learning strategies as following:

- **TLA** – We used all the weights from the previous model as initial weights to form our model, and the weights from all layers are trainable.
- **TLC** – The fine tuning strategy served to train a new pattern.Specifically, the weights of the convolutional layers of the previous model are transferred as initial weights into the new model, and the convolutional layers of the new model are frozen.We trained two last dense layers from scratch.

Table 3.11 summarized the training strategies. In each strategy, we trained our InceptionVGG8 CNN model over 500 epochs with the best optimizer RMSprop of ablation study. Table reports the result of our InceptionVGG8 CNN model with these 4 pro-

Strategy	Training set	Validation set	Training configuration
SCS	35192	5496	Scratch
SCR	35192	5496	Scratch
TLA	13432	1496	Transfer learning No frozen layer
TLC	13432	1496	Transfer learning Convolutional layer frozen

TABLE 3.11: Summarization of the four training strategies, we emphasize that the training and validation set of "heat" class of strategy SCR are different from ones of strategy SCS.

posed training strategies in term of heat state prediction on the test set composed of 20 videos. One can observe that your model reaches the best accuracy of 95 % (19/20) in term of the number of correct predicted video with the first strategy SCS. Especially, all the videos incorrectly predicted (*) by previous model have been correctly predicted by retrained model. Moreover, with TLA strategy, even though our model has been trained on the previous dataset, but after 500 epochs of training on the new dataset only, it mispredicted the two previously correctly predicted videos (video_11† and video_12†). For the TLS strategy, thanks to the fact that we froze the convolutional layer, he did not make similar mistakes. However, compared to the SCS strategy, it is still slightly inferior. Nevertheless, with all the four strategies, our model has a significant improvement in view of the prediction of videos miscorrectly predicted by previous model (*). This means that we need to continuously add new data to upgrade the model, and to maintain the robustness and generalization ability of the model.

3.5 Conclusion

A deep learning based system has been proposed to address the heat detection problem within cattle. The system allows to analyze a vaginal endoscope video of a cow at the level of frames and predicts its heat state in a few seconds. To train the system, a dataset of endoscopic images has been generated and labeled by experts. To limit the overfitting of the trained system, the dataset has been artificially augmented by exploiting texture label preserving transformations. The conducted experiments on

Videos		SCS	SCR	TLA	TLC
Heat class	1†	95.4	97.5	100	97.2
	2†	100	99.7	93.8	100
	3⋆	99.5	99.5	100	99.5
	4⋆	87.4	95.3	99.9	97
	5⋆	98.9	100	100	100
	6	97.5	99	99.7	98.9
	7	87	99.7	100	98.3
	8	71.2	59.4	73.1	71.6
	9⋆	61	25.7	65.4	39.9
	10	0.06	1.24	2.7	3.4
No-heat class	11†	100	81.5	0	86.1
	12†	100	99.5	0	99.3
	13	93.2	96	93.8	97
	14	100	94	79.3	96.9
	15 ⋆	84.3	69	68.7	96.2
	16	100	100	99.7	100
	17	100	99.4	99.4	99.2
	18 ⋆	91.4	98.1	99.7	98.1
	19⋆	86.5	89.7	83.1	83.3
	20 ⋆	96.7	97	97.7	99.6

† The videos from previous dataset, corresponding to the video_1 to video_4 of Table 3.7.

⋆ The videos incorrectly predicted by previous model.

TABLE 3.12: Results in (%) of heat state prediction on the test set composed of 20 videos

our dataset and a public dataset show the high performance of the system. Indeed, the trained image classifier exceeded the 97% of accuracy for both datasets. In addition, the feasibility of the system for real usage on Android smart-phone has been demonstrated. A video clip is provided in the supplementary material.

In the short term, the system will be deployed in several farms in the north region of France in coordination with Gènes Diffusion, CECNA and Elexinn companies. In the next future, new research opportunities will be investigated related to the development of systems for assistance to cow pathology diagnosis.

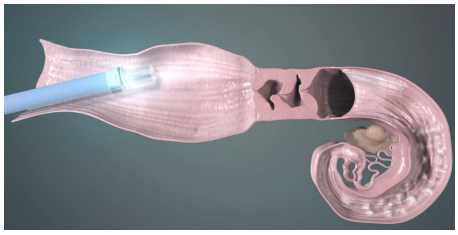
Chapter 4

Cow cervix detection

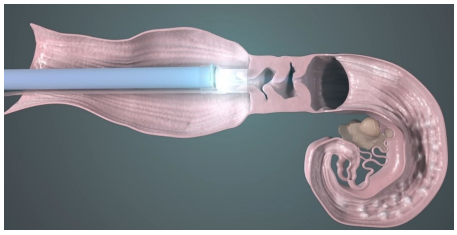
4.1 Introduction

In cow farming, artificial insemination is a reproduction biotechnology which is widely spread [25]. Nevertheless, to successfully accomplish cow insemination, its heat period needs to be correctly detected, a stage that the farmer can possibly observe through certain behaviors of the cow [73], without complete certainty [267]. The expert or the veterinarian then intervenes to confirm the state of heat and locate the cervix to introduce the spermatozoa. On that point, farmers face two challenges, the availability of experts at the right time (cow heat) and the cost associated with their interventions, particularly if the state of heat is already over [171].
In this context, we presented a new deep learning vision-based approach for cow heat detection by using an innovative insemination technology namely *Eye Breed* [55] in the previous chapter. The principle of this insemination technique is shown in Figure 4.1, the device enters the vagina to find the cervix and immobilizes it through aspiration, then a catheter is passed through the cervix to deposit the semen in the uterus. In this way, insemination can get rid of the need for rectal palpation to improve the welfare status of the cow. However, cervix localization is not a simple task, especially when some secretions are present. It requires mastering the insemination procedure. Therefore, in addition to heat detection, precise cervix localization is the other main condition for successful insemination. To this end, we propose a new approach for cow heat detection, which allows to localize the cervix. As illustrated in Figure 4.2, our approach permits identifying on the fly the cow's heat state through two main stages namely cervix detection and then heat classification. For this purpose, each frame of the input video stream, collected by the endoscopic camera of the Eye breed device, is analyzed by a Transformer based detection model to localize the cervix, in

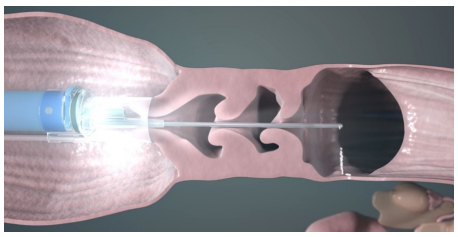
which case the frame is analyzed by the CNN based heat classification model. The final result of the frame analysis (cervix box + heat state) is displayed on the screen. Our proposed approach permits to significantly improve the two functionalities of our



(a) Device into the vagina



(b) Immobilization of cervix by aspiration



(c) Deposit the semen in the uterine body

FIGURE 4.1: Illustration of insemination process by *Eye Breed* [55]

previous system:

- It further facilitates the insemination operation performed by the farmer thanks to the integration of the cervix detection model. Indeed, the model is able to detect and localize on the fly and in an accurate way the cow cervix offering thus to the farmer an assistance in the device guidance inside the cow genital tract.
- It increases the performance of our original heat state classification model by excluding from its analysis noisy video frames. To this end, the model only proceeds to the analysis of frames that are identified by the cervix detection model as positive.

Additionally, we show through the experimental study conducted on our dataset that the proposed Transformer architecture for object detection outperforms the state-of-the-art ones [123, 37, 29, 198, 197, 277].

4.2 State of the art

In this chapter, cervix detection is an object detection problem, which is another essential research topic in computer vision covered by extensive literature. The task of object detection is to find objects of interest in an image or video and simultaneously detect their location and size. Different from the image classification task, object detection not only needs to solve the classification problem but also solve the positioning problem, which is a multi-task problem.

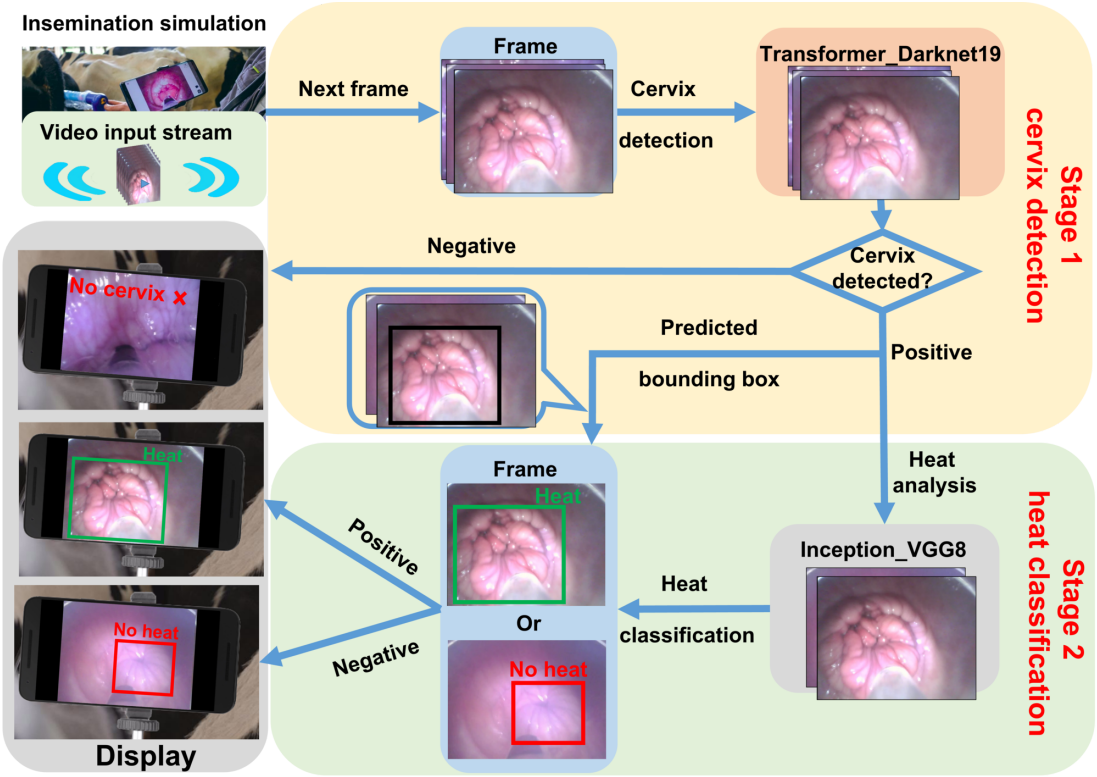


FIGURE 4.2: Flowchart of our cow heat analysis method

4.2.1 Public dataset and benchmarks

In object detection of computer vision, PASCAL VOC (Visual Object Classes) [64, 65], Microsoft COCO (Common Objects in Context) [152] are the two most popular public datasets for object detection.

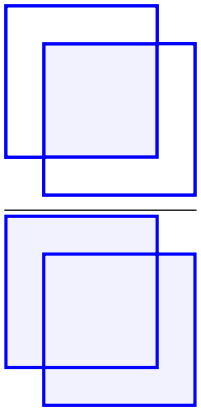
PASCAL VOC (Visual Object Classes) [64, 65] is a an challenges dataset, which is one of the benchmarks of object detection technology. the competitions are held from 2005 to 2012. Between 2005 and 2012, the Computer Vision Challenge was held

every year. VOC dataset has two version VOC2007 [64] and VOC2012 [65], the latest version VOC2012 [64] contains 11,530 images with 27,450 ROI-annotated objects in 20 classes,including:

- **Person**: person
- **Animal**: bird, cat, cow, dog, horse,sheep
- **Vehicle**: aeroplane, bicycle, boat,bus, car, motorbike, train
- **Indoor**: bottle, chair, dining table,potted plant, sofa, tv/monitor

For evaluation of object Detection results on the VOC dataset, the IoU (Intersection over Union) between the prediction and ground-truth is calculated as:

$$IoU = IoU = \frac{\text{area of overlap}}{\text{area of union}} = \frac{\text{}}{\text{}} \tag{4.1}$$



A positive prediction is defined as a box detected with an IoU$\geq$ 0.5 compared to the ground-truth box. In the case of multiple detections of the same object, it counts the one with the first ranking confidence score as a positive while the rest as negatives. For the VOC2007 challenge, the 11-points interpolated average precision (AP)[212] was calculated to evaluate both classification and detection. The interpolated AP summarises the shape of the precision/recall curve, which is defined as the mean precision at a set of eleven equally spaced recall levels [0, 0.1,..., 1]:

$$AP = -\frac{1}{11} \sum_{r=[0,0.1,.....,1]} p_{interp}(r) \tag{4.2}$$

The precision at each recall level r is interpolated by taking the maximum measured

precision for a method with the corresponding recall in excess of r:

$$p_{interp}(r) = \max_{\tilde{r}:\tilde{r}\geq r} p_{interp}(\tilde{r}) \tag{4.3}$$

Where $p_{interp}(\tilde{r})$ is the measured precision at recall $\tilde{r}$. For later Pascal VOC competitions (VOC2010–2012), it calculated the area enclosed by the smoothed precision/recall curve and the recall axis. Moreover, for the 20 different classes of PASCAL VOC, it computed an AP for each class and provided the average of these 20 AP results called mAP (mean average precision).

Microsoft COCO (Common Objects in Context) [152] is a large-scale object detection, segmentation, and captioning dataset created by Microsoft Corporation. Microsoft COCO is another benchmark of object detection technology for object detection tasks, having 121,408 images with 883,331 ROI-annotated objects in 80 classes. The latest research papers tend to give results on the COCO dataset rather than PASCAL VOC due to the larger amount of data and the higher challenge.

For the evaluation, a 101-point interpolated AP definition is used in the calculation. Moreover, instead of using a fixed IoU threshold, MS COCO introduces multi-metrics (summarized in Table 4.1) to characterize more completely the performance of an object detector

In addition to the large public image datasets for object detection, there are also several

mAP@[.5,0.95]	mAP averaged over ten IOUs: {0.5:0.05:0.95};
mAP@.5	mAP at IoU=.50 (PASCAL VOC metric)
mAP@.75	mAP at IoU=.75 (strict metric)
mAP_S	mAP for small objects: area <322
mAP_M	mAP for medium objects: 322 <area <962
mAP_L	mAP for large objects: area >962

TABLE 4.1: Summary of commonly used object detectors evaluation metrics

datasets for specific issues, such as pedestrian detection (EuroCity [32]), face detection (WildestFaces [134], WiderFace [266], Fddb [118]), vehicle detection (CompCars[265], Stanford Cars Dataset [135]), and endoscopy artefact detection (EAD [7]). Amongst them, EAD is one of the few publicly available endoscopic image data sets for object detection, aiming to identify hindrances like saturations, motion blur, specular reflections, bubbles, imaging artefacts, contrast and instrument using revolutionary

techniques in artificial intelligence as shown in Figure 4.3. The training dataset for detection consists in total 2147 annotated frames of endoscopic video overall 7 artifact classes.

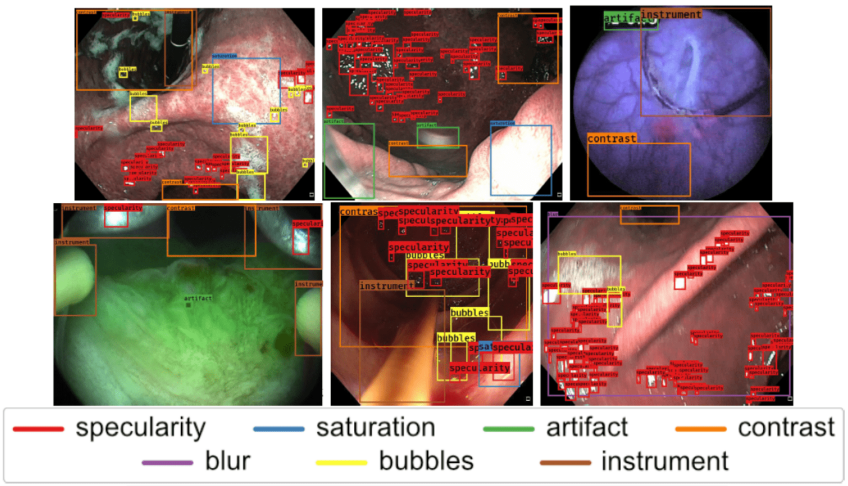


FIGURE 4.3: Example annotated training detection boxes illustrating the 7 different artifact classes in the EAD2019 challenge dataset. (Image from Sharib.A et al [7])

4.2.2 Deep learning-based object detection

Prior to the development of deep learning, the use of image feature descriptors to extract image features is the core strategy of most early object detection algorithms called handcrafted features-based algorithms. Viola-Jones Detector [246], histogram of oriented gradients (HOG) [50] and deformable part model (DPM)[66] are the three most representative handcrafted features based object detection algorithms. In particular, DPM Detector has won the championship of the VOC 2007-2009 Object Detection Challenge. However, the handcrafted features-based algorithms are slow to progress and have low performance. Compared with traditional methods, thanks to the advent of convolutional neural networks, the accuracy of deep learning-based object detection is getting higher and higher, and the detection speed is getting faster and faster. As raised in the survey made by Liu et al. [154], object detection is mainly divided into two-stage region-based frameworks and unified region-based frameworks.

4.2.2.1 Two stage region-based frameworks

The two-stage detection algorithms are mainly divided into the following two stages: 1) Generate region proposals from images; 2) Generate final object bounding box from region proposals.
R-CNN (Region-CNN) [76] is the first model to successfully apply deep learning to object detection. As shown in Figure 4.4, it first extracts 2000 region proposals from

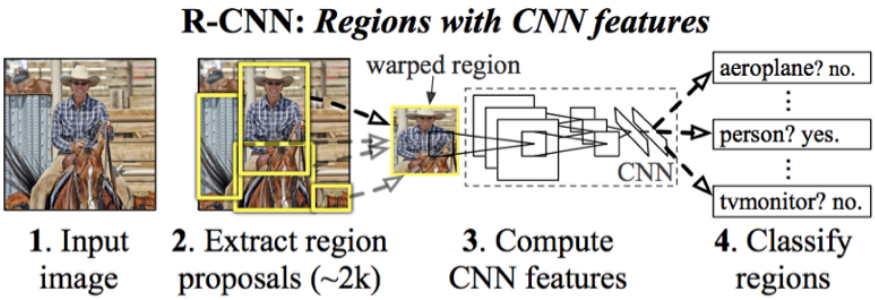


FIGURE 4.4: Region-CNN object detection system overview (Image from Ross Girshick et al [76])

the image by using the selective search algorithm. The generated region proposals are then resized to a fixed size image, and fed into a CNN model trained on the ImageNet to extract features. Finally, the extracted features are sent to the SVM classifier to confirm whether an object exists in the region proposals, and further predict which class the detected object belongs. Compared with traditional detection algorithms, the RCNN algorithm achieved significant results on the VOC2007 dataset, showing a qualitative leap in accuracy (rising average precision from 33.7% (DPM-V5) to 58.5%).
However, it needs to generate 2000 feature maps corresponding to 2000 region proposals for an image, which requires a huge count of calculations, leading to a lower detection speed (40-50s per frame). To reduce the redundant computation caused by a large number of overlapping boxes, K. He et al. [99] proposed a network using SPP (Spatial Pyramid Pooling) layer namely SPPNet. As shown in Figure 4.5, the main idea of the SPP layer is to generate fixed-length outputs, which is the fusion of multi-scale division of the input feature maps (divided into 1, 4, and 16 parts). The fixed-length outputs are then fed into the fully-connected layers (or other classifiers). In this way, an image only needs to be calculated once to generate the corresponding feature map, which prevents the convolution feature map from being counted repeatedly. Furthermore, since the outputs of the SPP layer are fixed length, SPPNet removes

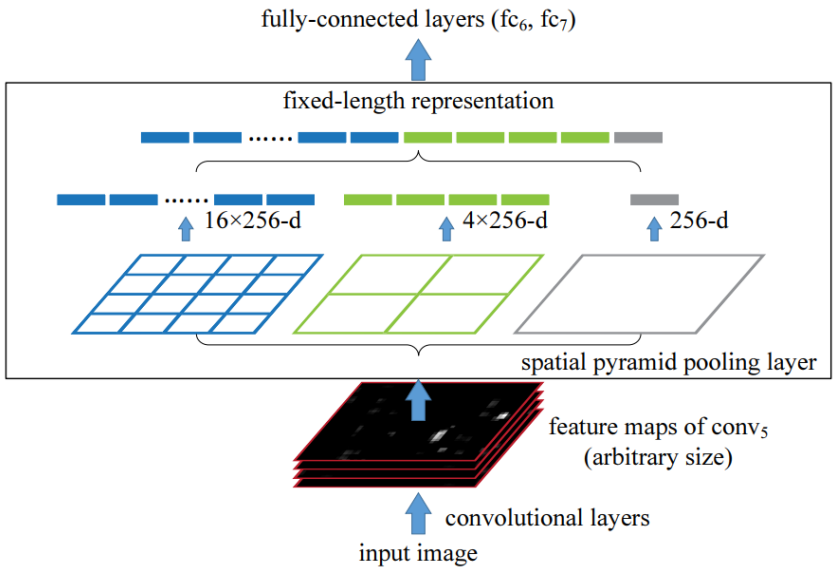


FIGURE 4.5: Illustration of SPP (Spatial Pyramid Pooling) layer. (Image from K. He et al. [99])

the fixed network size constraint. Compared with the RCNN algorithm, SPPNet accelerates the detection speed by more than 20 times without sacrificing detection accuracy (VOC2007, mAP@.5=59.2%).

Fast R-CNN [75] solved some of the drawbacks of R-CNN. As can be seen from Figure 4.6, instead of feeding the region proposals to the CNN, the input image is fed into the CNN to generate a convolutional feature map. The feature maps of each region proposal can be identified from the convolutional feature map of the input image. The feature maps of each region proposal are then warped into squares by using a RoI pooling layer to be fed into a fully connected layer. Finally, two softmax layers are used to predict the class of the proposed region and the offset values for the bounding box from the RoI feature vector obtained by a RoI pooling layer, respectively. Fast R-CNN improved the detection accuracy (mAP@.5) to 70.0% on the VOC2007 dataset, and the speed of detection is reduced to 2 seconds per image. However, the generation of region proposals also depends on the selective search algorithm.

To this end, Faster R-CNN [200] is the first end-to-end object detection model. The basic difference between Faster R-CNN and the previous models is the use of a CNN network namely RPN (Region Proposal Network) to generate the region proposals in Faster R-CNN. As illustrated in Figure 4.7, the input image is fed into a convolutional network to generate a feature map of the image. The generated feature map is fed into

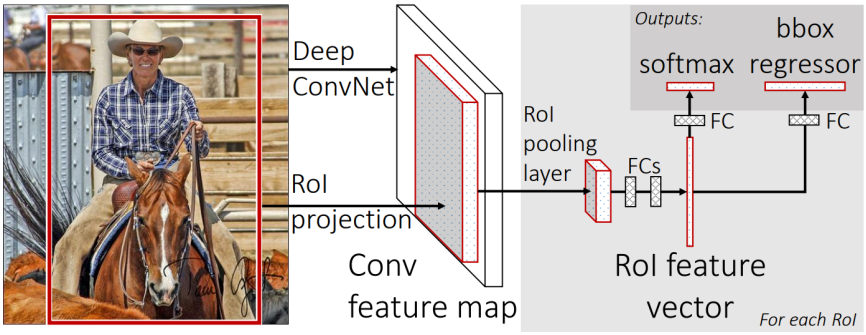


FIGURE 4.6: Fast R-CNN architecture. (Image from Ross Girshick et al [75])

RPN to generate the region proposals. These region proposals and feature maps are then fed into the ROI pooling layer for further procedure, which is similar to Fast R-CNN. Faster R-CNN improved furtherly the detection accuracy (mAP) to 73.2% on the VOC2007 dataset and 42.7%(mAP@.5) on the COCO dataset, and the speed of detection is reduced to 0.2 seconds per image. Faster R-CNN has higher accuracy, and faster speed, and is very close to meet the need for real-time detection. Therefore, several works are proposed for the improvement of Faster R-CNN. Lin et al [151] proposed a top-down network architecture namely FPN (Feature Pyramid Networks) with lateral connections for constructing high-level semantic information at all high-level layers with different scales. The use of FPN technology in the Faster R-CNN network permit to greatly improve the detection accuracy of the network (mAP@.5=59.1% on the COCO dataset). Moreover, Cai et al [36] proposed a cascade network namely Cascade R-CNN, the main idea of this proposition is the use of cascading several detection networks based on different IOU thresholds to address the difficulty of the training IOU setting. Cascade R-CNN improved the detection accuracy (mAP@.5) to 62.1% on the COCO dataset.

4.2.2.2 Unified region-based frameworks

As presented in 4.2.2.1, several two-stage-based algorithms have shown impressive performance. Nevertheless, the detection speed is also insufficient for real-time detection. In contrast, the unified region (one-stage) based detection algorithm can directly generate the class probability and position coordinate value of the object without the region proposal generation stage. It makes this type of algorithm have a faster detection speed.

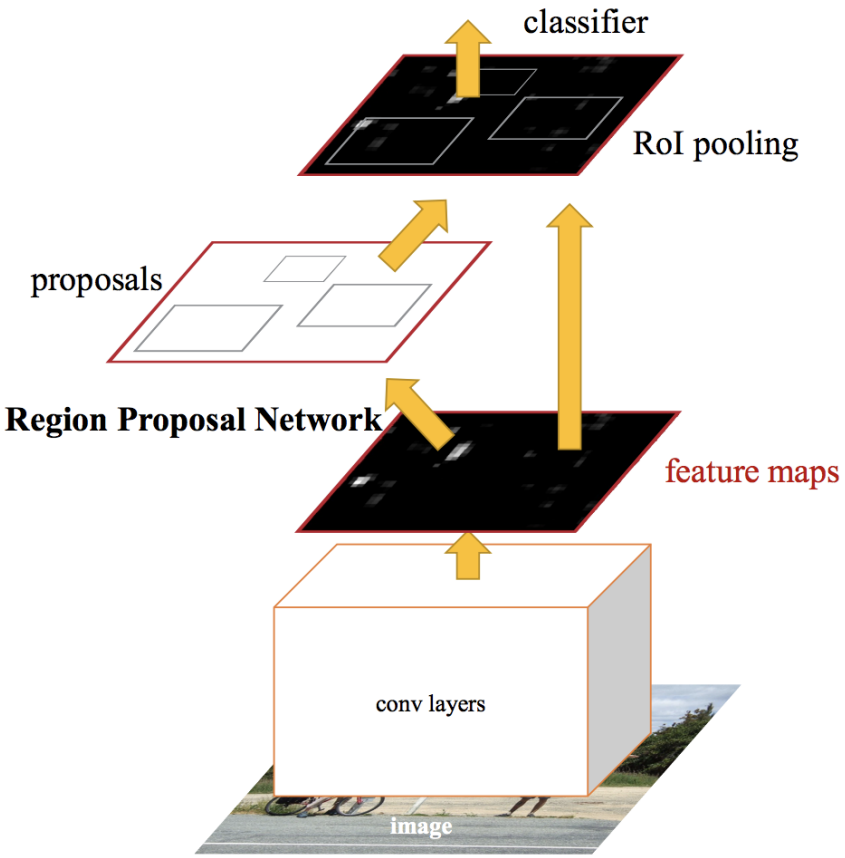


FIGURE 4.7: Faster R-CNN architecture. (Image from Ren et al [200])

YOLO [199] is the first CNN-based one-stage detector and its detection speed is very fast. The idea of this algorithm is to divide the image into multiple grids, and then predict the bounding box for each grid at the same time and give the corresponding probability. Finally, NMS (Non-Max Suppression) is used to remove the overlapping bounding box of the same object. It should be noted that only a single CNN network (VGG16) has been used in all detection processes, enabling YOLO to reach a detection speed of 21 FPS. Compared with the two-stage target detection algorithm, although the detection speed of YOLO has been greatly improved, the accuracy is relatively low (mAP@.5 =66.4% on the VOC2007 dataset), especially for some small objects detection problems.

YOLOv2 [197] made many improvements in accuracy, speed, and the number of classifications. DarkNet19 is proposed as the CNN backbone for feature extraction, which

is faster than the VGG16 used by YOLO. FC (Fully Connected) layers can cause escape information loss, leading to imprecise localization. Therefore, the FC layers are removed in YOLOv2. Moreover, the use of anchor boxes generated by k-means also accelerates the model training. YOLOv2 uses the joint training skills of target classification and detection, combined with methods such as Word Tree, to expand the detection types of YOLOv2 to thousands. YOLO v2 has achieved mAP@.5 of 76.8% on the VOC2007 and mAP@.5 of 44% on the COCO dataset. The detection speed of YOLOv2 has reached 67 FPS.

Since the FPN technology in Faster R-CNN made a great improvement in detection precision, FPN has become one of the most important techniques for improving the precision of major networks. To address the difficulty of multiscale object detection, YOLOv3 [198] proposed a novel CNN backbone namely DarkNet53 (see in Figure 4.8), which borrows the idea of FPN and adopts three branches (feature maps of three different scales/different receptive fields) to detect objects of different sizes. YOLO v2 has achieved mAP@.5 of 57.9% on the COCO dataset. However, the detection speed of YOLOv3 has droped to 20 FPS. Recently, YOLOv4 [6] and YOLOv5 [123]

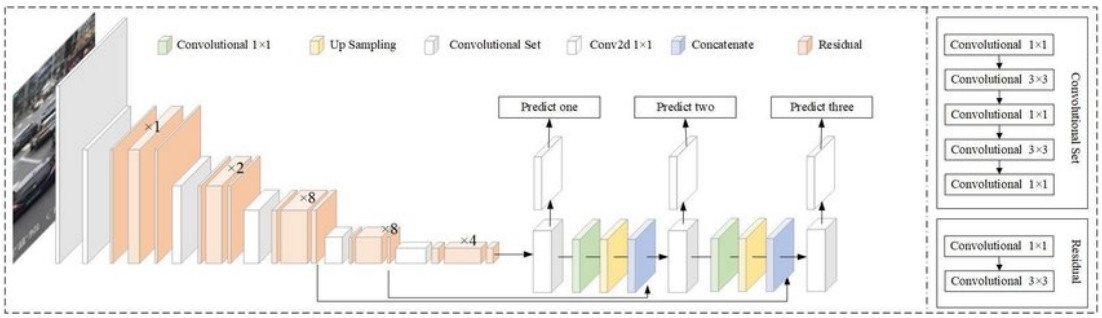


FIGURE 4.8: YOLOv3 architecture. (Image from Mao et al [166])

have used a variety of SOTA (State-of-the-art) tricks in the area of deep learning in recent years, which significantly improved the detection performance of YOLO on objects. They achieved mAP@.5 of 65.7% and 72.7% on the COCO dataset, respectively. Moreover, the detection speed of YOLOv5 can astonishingly reach 140 FPS.

In addition to the YOLO series, DETR (Detection Transformer) [37] proposed a new idea for object detection, it applies transformers which have taken impressive progress in natural language processing (NLP) tasks to the field of object detection, replacing the manual design work of current models (such as non-maximum suppression and anchor generation). DETR regards object detection as a set prediction problem and

proposes a very concise object detection pipeline. As shown in Figure 4.9, a CNN backbone is used to extract basic features, the extracted features are fed into the transformer module for relationship modeling, and the obtained output is matched with ground truth by the bipartite graph matching algorithm. DETR achieved mAP@.5 of 64.7% on the COCO validation dataset and a detection speed of 10 FPS.

However, the training speed of DETR [37] is not comparative. It has a long training

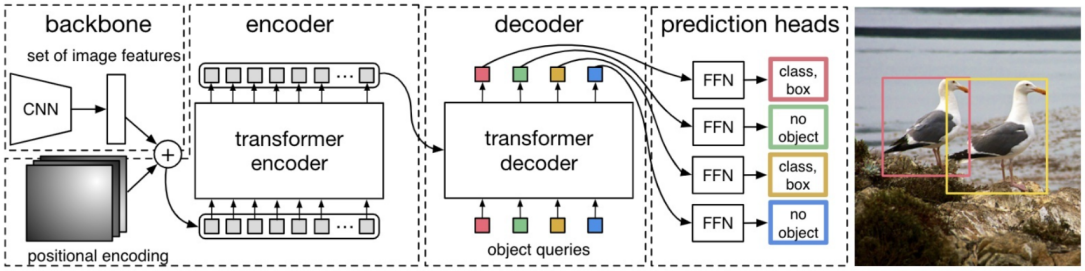


FIGURE 4.9: DETR architecture. (Image from Carion.N et al [37])

period that is 10-20 times slower than faster-RCNN. On the other hand, DETR [37] has difficulty with small object detection. To this end, Zhu et al [277] improved the DETR model by using scale-level embedding, multi-scales deformable Attention Module. The improvement version of DETR is called Deformable-DETR, with a mAP@.5 of 65.2% on the COCO validation dataset and a detection speed of 19 FPS.

Nevertheless, in reason of the requirements of the existing datasets (VOC and COCO), these architectures are designed and trained in such a way to detect at least one target object in the image. Hence, their direct exploitation for cervix detection from endoscopic videos will lead to a high rate of false positives detection which is a major issue in our case. Indeed, detecting a false cervix will increase the number of noisy frames to be analyzed by the classifier and will mislead the farmer in the insemination operation. To address this issue, our detection module has been designed to perform exclusively the detection task and has been trained on a balanced image dataset that includes positive and negative cervix examples.

4.3 Methodology

Our proposed methodology for cow heat analysis exploits two CNN architectures namely Transformer-Darknet19 and InceptionVGG8 for cervix detection (stage 1) and

heat classification (stage 2) respectively. In this section, we present the first architecture (Transformer-Darknet) which is a variant of the original DETR architecture [37]. Figure 4.10 illustrates the design of our Transformer-Darknet19. More specifically,

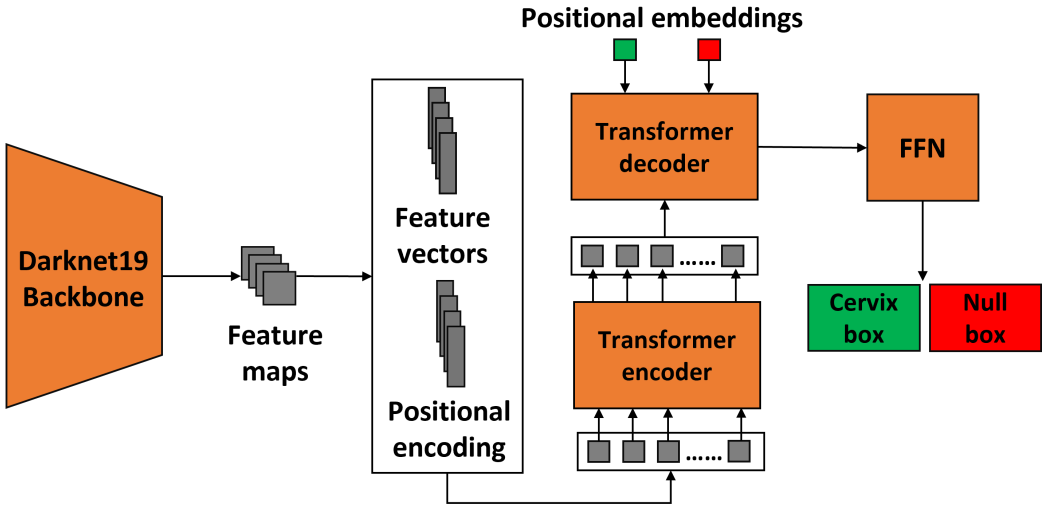


FIGURE 4.10: Overview of our Transformer-Darknet19 architecture for cervix detection.

image features are extracted using a CNN backbone corresponding to the Darknet19 [197] architecture. The features are then flattened and given to the encoder of the transformer in the form of a data sequence which is the expected form by transformers. To avoid losing spatial positioning information of the image pixels, positional encoding of the flattened features are joined to the input data sequence of the encoder. The encoder proceeds then to learn, through a self-attention mechanism [267], how to focus on relevant object patterns from the input sequence. The encoder output is then exploited by the decoder to learn, through a self-attention mechanism, how it can affect the prediction of the two positional embeddings that represent the existence and absence of a cervix object respectively. Finally, the decoded output is passed to an FFN (Feed Forward network corresponding to a 3-layer perceptron of size 2048 each) for box prediction.

It is worth mentioning that to adapt the original DETR architecture [37] to our problem. In addition to using Darknet19 as the CNN backbone, we have: 1) used the Darknet CNN backbone for image features extraction instead of ResNet, 2) limited the number of encoders and decoders to one instance instead of six, 3) limited the number of positional embeddings, for encoding the cervix existence and absence to 2 instead of 100. 4) modified the loss function by replacing the $L1$ loss component by the $smooth_{L1}$ loss

in order to faster the convergence of the learning [75].

4.3.1 CNN backbone

We use Darknet19 [197] (see in Figure 4.11) as our network backbone, which is the features extractor of YOLOv2 [197], including 19 convolutional layers and 5 max-pooling layers. The design principles of the Darknet-19 and VGG16 [227] models are consistent, mainly using 3×3 convolutional layers, and a 1×1 convolutional layer between 3×3 convolutions to compress feature map channels to reduce model computation and parameters. The 2×2 max-pooling layer is then used to reduce the dimension of the feature map by a factor of 2. The FC (fully connected) layer is removed from

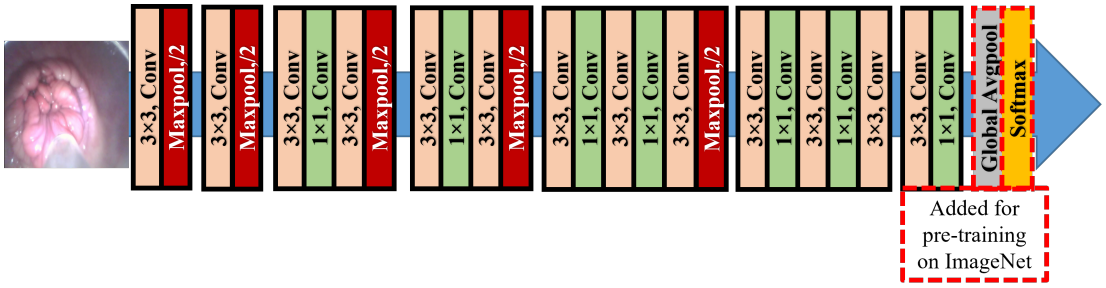


FIGURE 4.11: Darknet19 architecture

DarkNet19, which greatly reduces the parameters of the network and accelerates the speed of detection. Furthermore, a BN (Batch Normalization) layer is used after each convolutional layer instead of dropout, which aims to the convergence speed of the model is improved, and it can play a regularization effect and reduce the overfitting of the model. The choice of Darknet19 backbone is explained by the fact that, contrary to ResNet which is used in the original DETR [37] model, it has a reasonable depth that helps to limit the overfitting during the training process of the whole architecture. The backbone has been pre-trained on the ImageNet dataset by adding a global average pooling layer and a classifier Softmax with the input image of 448 × 448. To extract image features, we considered the output of the 18th convolutional layer (1024 feature maps). Therefore, beginning with the input color image $I \in \mathbb{R}^{H_0 \times W_0 \times 3}$, a lower-resolution activation map (feature map) $f \in \mathbb{R}^{H \times W \times C}$ (H $=\frac{H_0}{32}$, W $= \frac{W_0}{32}$, C = 1024) is generated by our CNN backbone (Darknet19).

4.3.2 Transformer module

Since our model is a variant model of the DETR [37] model, we remain the general form of the Transformer module, which consists of the encoder module and decoder module. The difference with the transformer of DETR is that we proposed the use of a lighter transformer module, considering the feasibility of the deployment in mobile. We remind the general form of the Transformer module in this section.

Multi-head attention layers – In a Transformer based model, the multi-head attention layer is the most important component that is used in both the encoder and the decoder of the Transformer module. Multi-head Attention is a module for attention mechanisms that runs through an attention mechanism several times in parallel. Algorithm 2 summarizes the different steps of the multi-head attention layer. Let us denote by {Q

Algorithm 2 Multi-head Attention Layer

Input: $Q \in \mathbb{R}^{L\times E}$, $K \in \mathbb{R}^{S\times E}$, $V \in \mathbb{R}^{S\times E}$, N (Number of heads)
Output: $attn\_output \in \mathbb{R}^{L\times E}$, $attn\_output\_weights \in \mathbb{R}^{L\times S}$

Initialization : $W_Q \in \mathbb{R}^{E\times E}, W_K \in \mathbb{R}^{E\times E}, W_V \in \mathbb{R}^{E\times E} \leftarrow (Weight\ Matrix)$
$q = \text{LINEAR}(Q, W_Q)$, $k = \text{LINEAR}(K, W_K)$, $v = \text{LINEAR}(V, W_V)$
$q_i = \text{SPLIT}(q, N)$
$k_i = \text{SPLIT}(k, N)$
$v_i = \text{SPLIT}(v, N)$
$head\_dim = \frac{E}{N}$
for each q_i, k_i, v_i *in* q, k, v **do**
    $attn\_weights_i == \text{SOFTMAX}(\frac{q_i \otimes k_i}{\sqrt{head\_dim}})$
    $attn_i = attn\_weights_i \otimes v_i$
end for
$attn\_output = \text{CONCAT}(attn_i)_{i=1}^{N}$
$attn\_output\_weights = \frac{\text{SUM}(attn\_weights_i)_{i=1}^{N}}{N}$

$\in \mathbb{R}^{L\times E}$, K $\in \mathbb{R}^{S\times E}$, V $\in \mathbb{R}^{S\times E}$} the input set of a multi-head attention layer, and the multi-head attention has N heads. Where L is the target sequence length, S is the source sequence length, and E is the embedding dimension and also the total dimension of the model. In the first step, we initial three weight matrices W_Q, W_K, W_V of shape(E, E), which allows to get the three parameters of attention mechanism (q (query), k (key), v

(value)) in the second step :

$$q = Q \times W_Q \tag{4.4}$$

$$k = K \times W_K \tag{4.5}$$

$$v = V \times W_V \tag{4.6}$$

Where $q \in \mathbb{R}^{L\times E}$, $k \in \mathbb{R}^{S\times E}$, $v \in \mathbb{R}^{S\times E}$. In the third step, according to the number of heads of multi-attention (N), we reshape the q, k, v into N parts: $\{q_i \in \mathbb{R}^{L\times N\times head\_dim}\}_{i=0}^{N}$, $\{k_i \in \mathbb{R}^{S\times N\times head\_dim}\}_{i=0}^{N}$, $\{v_i \in \mathbb{R}^{S\times N\times head\_dim}\}_{i=0}^{N}$, where $head\_dim$ is the dimention of each head ($head\_dim = \frac{E}{N}$).
The attention weights of each $head_i$ ($attn\_weights_i$) is defined as:

$$attn\_weights_i\ (q,k) = softmax(\frac{q_i \otimes k_i^T}{\sqrt{head\_dim}}) \tag{4.7}$$

Where $\otimes$ is the multiply operation between the multi-dimensional tensors. Afterward, the output of each $head_i$ ($attn_i$) can be calculated as:

$$attn_i = attn\_weights_i \otimes v_i \tag{4.8}$$

Finally, the output of the multi-attention layer ($attn\_output$) is the concatenation of $attn_i$ of each $head_i$ and the weights of multi-attention layer output($attn\_output\_weights$) is the mean of the $attn\_weights_i$ of each $head_i$:

$$attn\_output = Concat(attn_1, attn_2, ...attn_i)_{i=1}^{N} \tag{4.9}$$

$$attn\_output\_weights = \frac{1}{N}\sum_{i=1}^{N} attn\_weights_i \tag{4.10}$$

Positional encoding – Two kinds of positional encoding are used in the Transformer module:

- Spatial positional encodings (SPE): For the object detection problem, the position of each pixel is very important. However, the structure of the transformer is permutation-invariant. In other words, it is not sensitive to the position. Moreover, collapsing the input image feature map cause the 2D spatial structure disappears. Therefore, spatial position encoding is also required to represent the original complete 2D information.

- Object queries (OQ): The anchor box is used to constrain the predicted object range and add size priors to achieve the purpose of multi-scale learning. However, the anchor box is manually designed in the previous object detection, which may lead to redundancy of bounding boxes, requiring NMS to remove duplicate bounding boxes. To this end, the object queries are a learned anchor in the form of embedding, the purpose is to make the network learn the anchor by itself according to the dataset.

LN (Layer Normalization) – Normalization can ensure that the input of each layer of the network maintains the same distribution. The LN (Layer Normalization) [15] is used in the transfomer module. For a given input sequence $X = \{x_1, x_2, ..., x_n\}$, each sample x_i has m elements. We calculate the mean $E|x_i|$ and variance $Var|x_i|$ of each sample, and the LN (Layer Normalization) can be calculated as:

$$LN_{\gamma,\beta}(x_i) = \frac{x - E|x_i|}{\sqrt{Var|x_i| + \varepsilon}} \times \gamma + \beta \tag{4.11}$$

Where: γ and β are learnable parameters.

$$E|x_i| = \frac{1}{m}\sum_{m}^{j=1} x_{i,j} \tag{4.12}$$

$$Var|x_i| = \frac{1}{m}\sum_{m}^{j=1} (x_{i,j} - E|x_i|) \tag{4.13}$$

$$\tag{4.14}$$

Transformer Encoder – To be fed into the encoder, the feature map f generated by CNN backbone is reduced by a convolutions with dimension of 1×1 f from $C = 1024$ to a smaller dimension of $E = 256$ to create a new map $f' \in \mathbb{R}^{H \times W \times E}$. The generated low-level activation map is then collapsed into an one-dimensional sequence $X = flatten(f')$ ($X \in \mathbb{R}^{HW \times E}$, E=256). As shown in the left of the Figure 4.12, the encoder consists mainly of an eight-heads attention module and a feed-forward network (FFN). The residual structure is used in both of them and a Layer Normalization is used after each residual structure. We first duplicate the input sequence X to three sequences $\{Q, K, V\}$ ($Q = K = V = X$), SPE is added in Q and K ($Q = Q + SPE$, $K = K + SPE$) to supplement the spatial information. The sequences Q, K, V are then fed into the eight-head attention module to generate $attn\_output$ of shape (HW, 256),

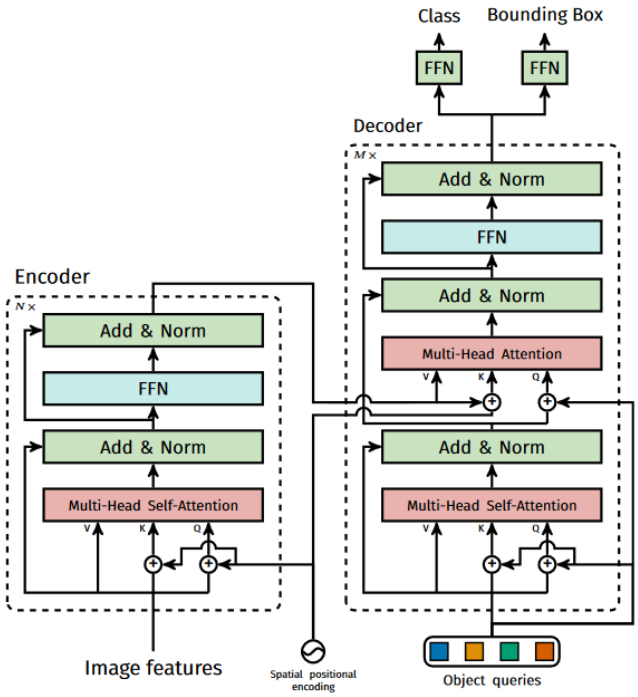


FIGURE 4.12: Architecture of transformer (Image from Carion.N et al [37]

and an associated matrix of weight $attn\_output\_weights^{enc}$ of shape (HW, HW), which can be used for self-attention visualization. Afterward, the original mapping is recast into $X + dropout(attn\_output^{enc})$ by a skip connection as the input to be sent to the feed-forward network (FFN) to generate the output of encoder $output^{enc}$ of shape (HW, 256).

Transformer Decoder – The decoder consists mainly of two eight-head attention modules and a feed-forward network (FFN) (see in the right of the Figure 4.12). Similar to the encoder, each multi-head attention module and feed-forward network use the residual structure and the Layer Normalization. We first create n ($n = 2$) input embeddings of dimension $E = 256$ which refer to the object queries (OQ) as a matrix of shape (2,256). As in the encoder, we then duplicate the input embeddings OQ to three sequences $\{Q, K, V\}$ ($Q = K = V = OQ$). OQ is added in Q and K ($Q = Q + OQ$, $K = K + OQ$) as the learnt positional encodings to feed into the first eight-head attention module. We obtain the outputs of the first eight-head attention module ($attn\_output_1^{dec} \in \mathbb{R}^{2\times256}$). For the second eight-head attention module, V is the output of the encoder, K is the addition of the output of the encoder and SPE, Q is the addition of the outputs of the first eight-head attention ($attn\_output_1^{dec}$), and the

learnt positional encodings (OQ). Afterward, the output of the second eight-head attention module ($attn\_output_2^{dec} \in \mathbb{R}^{2\times HW}$) is sent to the feed-forward network (FFN) to generate the output of encoder $output^{dec}$ of shape (2, 256). The associated weight matrix ($attn\_output\_weights_2^{dec} \in \mathbb{R}^{2\times HW}$) can be used for self-attention visualization of each object query.

Prediction FFN (Feed-forward network) – FFNs are used multiple times in the transformer module, the FFN used in the encoder and the decoder consists of two linear transformations with a ReLU activation in between.

$$FFN(x) = \max(0, W_1 + b_1)W_2 \times b_2 \tag{4.15}$$

In addition, the FFN for final prediction is a 3-layer perceptron with ReLU activation function and the hidden dimension of 2048, which is different from the FFNs in encoder and decoder. This FFN predicts the bounding box of the object (normalized center coordinates, height, and width of the box), and object class is predicted by the linear layer with a softmax function. For each input image, a fixed-size set of n bounding boxes are predicted, which refers to the number of object queries. In our case, one image contains at most the cervix. Therefore, for the purpose of avoiding false-positive boxes, we reduce the number of object queries from 100 (set in [37]) to 2.

4.3.3 Loss function

In our case, an image has at most one cervix as the target object, and our model generates a fixed-size set of n predictions ($n \equiv 2$, the number of object queries). For the prediction of an image, we denote by $x = \{x_1, \varnothing\}$ (padded with$\varnothing$ for class balance) the ground-truth set of object, and $\hat{x} = \{\hat{x_1}, \hat{x_2}\}$ the set of 2 predictions. We find the bipartite matching between the two sets by searching for a permutation of 2 elements with the lowest cost:

$$\hat{\sigma} = argmin \sum_{i}^{2} \mathcal{L}_{match}(x_i, \hat{x}_{\sigma(i)}) \tag{4.16}$$

Where $\mathcal{L}_{match}(x_i, \hat{x}_{\sigma(i)})$ is the matching cost between ground truth x_i and optimal matching prediction $\hat{x}_{\sigma(i)}$. The matching cost considers two parts of cost: class prediction and the similarity of the predicted box and the box of ground truth. Each element of both prediction set and ground truth set consists of the target class label (cervix or $\varnothing$) c_i

and a vector $b_i \in [0,1]^4$ represents the center coordinates and the height and width relative to the image size of the object bounding box . We can define the $\mathcal{L}_{match}(x_i, \hat{x}_{\sigma(i)})$ as:

$$\mathcal{L}_{match}(x_i, \hat{x}_{\sigma(i)}) = -\mathbb{1}_{\{c_i \neq \varnothing\}} \hat{p}_{\sigma(i)}(c_i) + \mathbb{1}_{\{c_i \neq \varnothing\}} \mathcal{L}_{bbox}(b_i, \hat{b}_{\sigma(i)}) \tag{4.17}$$

Where $\hat{p}_{\sigma(i)}(c_i)$ is the class propablity of the prediction at index $\sigma(i)$. For the bounding box loss, we take into account both IoU loss [201] $\mathcal{L}_{IoU}$ and $smooth_{L1}$ loss [75] $\mathcal{L}_{smooth_{L1}}$ insteading of L1 loss:

$$\mathcal{L}_{bbox}(b_i, \hat{b}_{\sigma(i)}) = \lambda_{IoU} \mathcal{L}_{IoU}(b_i, \hat{b}_{\sigma(i)}) + \lambda_{smooth_{L1}} \mathcal{L}_{smooth_{L1}}(b_i, \hat{b}_{\sigma(i)}) \tag{4.18}$$

Where $\lambda_{IoU}, \lambda_{smooth_{L1}} \in \mathbb{R}$ are hyperparameters. The choice of $\mathcal{L}_{smooth_{L1}}$ is explained by the experiment of 4.4.2, and it is defined as following:

$$\mathcal{L}_{smooth_{L1}}(b_i, \hat{b}_{\sigma(i)}) = \begin{cases} 0.5(||b_i, \hat{b}_{\sigma(i)}||_1)^2, & \text{if } |||b_i, \hat{b}_{\sigma(i)}||_1| < 1 \\ |||b_i, \hat{b}_{\sigma(i)}||_1| - 0.5, & \text{otherwise} \end{cases} \tag{4.19}$$

4.4 Experimental study

4.4.1 Data preparation

Our dataset consists of 12732 labeled endoscopic images which have been extracted from 79 recorded videos of simulated insemination operations on several cows using the Eye breed device. The device has been used to record videos of simulated insemination using the default parameters of the camera notably an image rate from 20 FPS to 102 FPS and an image resolution of 640×480. The labeling of each image corresponds to i) its heat class (heat or no-heat) which is the video class assigned by the expert and ii) the associated cervix bounding box which is set up to null if there is no cervix. The set of images has been split into a training set of 10734 images and a validation set of 1998 images. Both two sets are balanced in terms of positive and negative cervix boxes. The pre-processing chain that allowed to generate the image sets is described in what follow.

Video split and data partitioning –To avoid the images for training and validation coming from the same video, the dataset split has been done in such a way that the

frames of a given video are completely included either in the training set or in the validation set. We have randomly split a set of videos into 69 videos for training and 10 for validation. All the video frames have been extracted using an FPS rate set to 5. To evaluate the generalization level of our models, we exploited the 10 validation videos to extract a larger set of labeled images namely 32552 and considered it as a test set. For this purpose, the video frames have been extracted using the default FPS rate of the concerned videos which is ranged between 20 and 102. Indeed, this parameter can be finetuned by the operator during the recording.

Image labellization – According to the heat status of the video represented cow, all

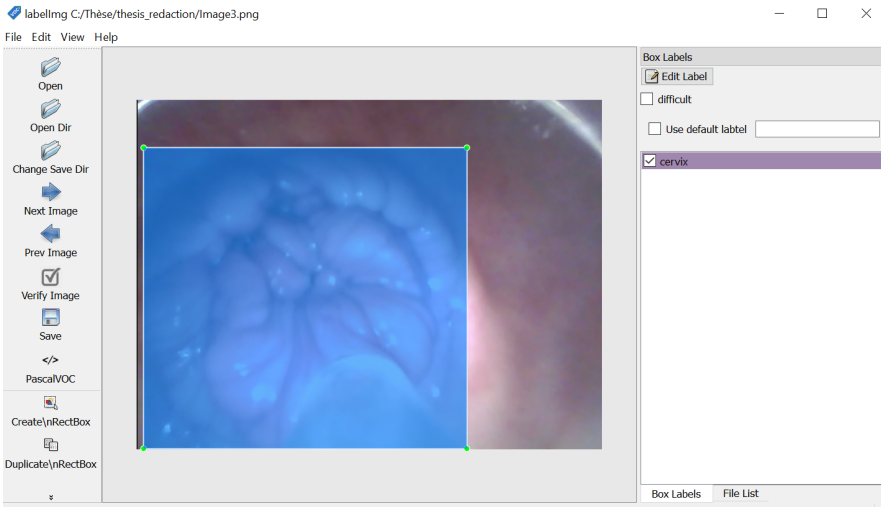


FIGURE 4.13: Cervix localization mark

the videos have been labeled in two classes (i.e heat or no-heat) by CECNA experts. Moreover, for each cervix positive image, its cervix bounding box is labeled by the labeling tool called labellmg [242] which allows to manually label each image (see in Figure 4.13) and get a corresponding XML tag file is formed, which contains the four coordinates (x_{min}, y_{min}, x_{max}, y_{max}) of the target object in the image and the given category (PASCAL VOC format).

Coordinate transformation – To meet the various requirement of input image resolution and the format of bounding box coordinates, we transformed the bounding box coordinates. For a given bounding box $bbox(x_{min}, y_{min}, x_{max}, y_{max})$ of an image with resolution of $h_0 \times w_0$, the model requires the image resolution of $h \times w$. In addition to resize of image, the bounding box coordinates should be transformed as $bbox^t(x^t_{min}, y^t_{min}, x^t_{max}, y^t_{max})$:

- $x^t_{min} = x_{min} \times \frac{w}{w_0}$

- $y^t_{min} = y_{min} \times \frac{h}{h_0}$
- $x^t_{max} = x_{max} \times \frac{w}{w_0}$
- $y^t_{max} = y_{max} \times \frac{h}{h_0}$

Moreover, some models require the bounding box coordinates with format (x, y, w_1, h_1), which represents the center coordinates and the height and width relative to the image size of the object bounding box. Therefore, we need to transfer bounding box $bbox(x^t_{min}, y^t_{min}, x^t_{max}, y^t_{max})$ to format (x, y, w_1, h_1):

- $x = \frac{x^t_{min}+x^t_{max}}{2} \times \frac{1}{w}$
- $y = \frac{y^t_{min}+y^t_{max}}{2} \times \frac{1}{h}$
- $w_1 = \frac{x^t_{max}-x^t_{min}}{w}$
- $h_1 = \frac{y^t_{max}-y^t_{min}}{h}$

4.4.2 Ablation study

In order to validate the effectiveness of our Transformer-Darknet19 model design, we have conducted an ablation study. In our ablation analysis, we explore how the other protocol of dataset generation, configurations of the CNN backbone and the transformer module, and loss influence the final performance. To this end, we trained our architecture over 500 epochs on the 10734 training image set for each case of the ablation study. Moreover, we use a multiscale training strategy instead of using a fixed-size input image. For every 10 batches, our network randomly chooses a new image dimension size from a set of scales: [480, 512, 544, 576, 608, 640, 672, 704, 736, 768, 800]. Thus the smallest option is 480 × 480 and the largest is 800 × 800. To measure this influence, we consider the ability of the models to distinguish between positive and negative cervix frames we calculated their accuracy (ACC) on the 32552 images of the test set as:

$$ACC = \frac{TP+TN}{TP+TN+FN+FP} \tag{4.20}$$

Where the TPs, FPs, TN, and FNs are identified with respect to the ground truth boxes based on the presence of cervix. They correspond for a given frame to what follows:

- TP: The number of positive cervix image samples with at least one cervix box detected by the detector.
- FP: The number of negative cervix image samples with at least one cervix box detected by the detector.
- TN: The number of negative cervix image samples without cervix box detected by the detector.
- FN: The number of positive cervix image samples without cervix box detected by the detector. We precise that the experiments have been calculated on NVIDIA Tesla V100 GPU with 32 GB of memory, and the resolution of the image used for performance evaluation is 800×800.

Protocol of dataset generation – The detection task of the public large dataset as COCO, VOC is to detect at least one target object in the image. In our case, the ability of the models to distinguish between positive and negative cervix frames is an important indicator for model performance evaluation. To this end, the half samples of our dataset are negative cervix frames. To evaluate the effectiveness of this dataset generation protocol, we have trained our model following 2 training set generation protocols:

- CP for the dataset without negative cervix images and using only the positive 5367 images.
- CNP for the dataset with both positive and negative cervix images with a total of 10734 (5367×2) images.

Protocols	CP	CNP
Accuracy(%)	63.7	**87.1**

TABLE 4.2: Cervix detection accuracies in (%) of our model on the test set (32552 images) following training set generation protocols

Table 4.2 summarized the obtained accuracies on the test set according to the 2 dataset generation protocols. The results show that removing the negative samples from our dataset caused a drop in its performance with notably an accuracy loss of 23.4%.

Importance of CNN backbone pre-training – To evaluate the importance of CNN

backbone pre-training, we consider we have trained it following 3 scenarios of CNN backbone pre-training:

- SCR for the training without CNN backbone pre-training.
- PR224 for the training with CNN backbone pre-training on ImageNet dataset with the input image size of 224×224.
- OR for the training with CNN backbone pre-training on ImageNet dataset with the input image size of 448×448, as described in 4.3.1.

Scenarios	SCR	PR224	OR
Accuracy(%)	80.6	85.2	**87.1**

TABLE 4.3: Cervix detection accuracies in (%) of our model on the test set (32552 images) following different CNN backbone pre-training scenarios

The obtained accuracies on the test set according to each scenario are summarized in Table 4.3. One can observe that our original scenario allows to correctly distinguish between positive and negative cervix frames on the test set at 87.1 %. Moreover, the latter scenario (PR224) improved the model accuracy by 4.6 % in comparison with the scratch scenario.

Different CNN backbones – In addition to Darknet19, we have also considered several CNN models: VGG16 [227], ResNet50 [98], ResNet101 [98], and Darknet53 [198] as the CNN backbone. The CNN models have been adapted to our problem by i) removing the fully connected layers and the global or average pooling (if exist); ii) pre-training on the ImageNet dataset. For the table 4.4, we can observe that our

CNN backbones	VGG16 [227]	ResNet50 [98]	ResNet101 [98]	Darknet53 [198]	Darknet19 (Ours) [197]
Accuracy(%)	82.1	79.1	62.5	77.3	**87.1**

TABLE 4.4: Cervix detection accuracies in (%) of our model on the test set (32552 images) using different CNN backbones

proposed module (original) gives the best cervix detection accuracy. Moreover, our model can also give an accuracy superior 80% when using the VGG16 [227] as the CNN backbone. This can be expected to the design similarity between this VGG16 [227] and Darknet19 [197], especially having a similar depth of the network. From

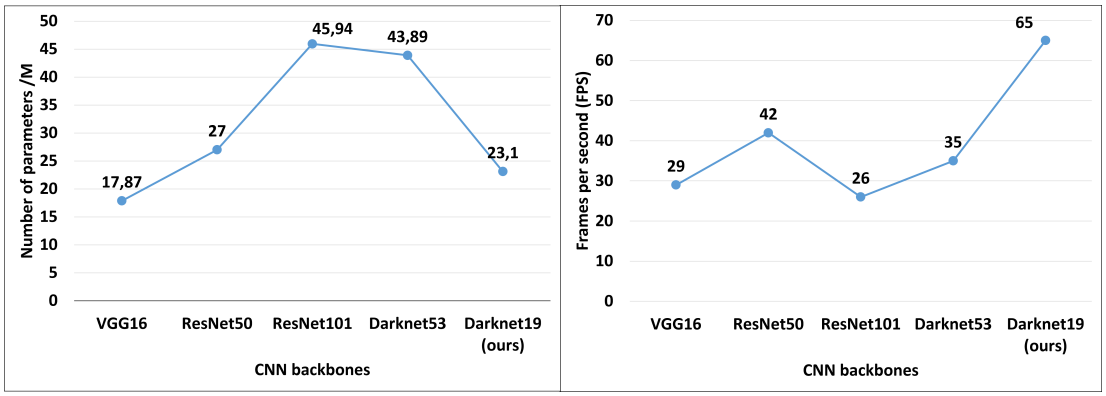


FIGURE 4.14: Complexity comparative results of our model using different CNN backbones: parameter complexities (left side) and FPS (right side)

Figure 4.14, we may also observe that our model using the recommended CNN backbone (Darknet19 [197]) has the fastest detection speed of 65 FPS, and is second only to VGG16 [227] in terms of the number of parameters.

Number of encoder layers and decoder layers – To train a lighter model for mobile deployment, we consider limiting the number of encoder layers and decoder layers (NED) from 6 to 1. The table 4.5 summarizes the results of our model on the test set

NED	NED_6	NED_3	NED_1
Accuracy(%)	79.8	86.4	**87.1**

TABLE 4.5: Cervix detection accuracies in (%) of our model on the test set (32552 images) using different numbers of encoder-decoder modules

(32552 images) using different numbers of encoder-decoder modules. It can be observed that our model obtains the best accuracy while using one encoder module and one decoder module.

From Figure 4.15, we may also observe that our model using one encoder module and one decoder is significantly better than that of the other versions for both parameters and FPS. Furthermore, it is worth mentioning that our model can also give a near-best accuracy of 86.4% when using three encoder modules and three decoder modules. However, it requires an additional 5.8M parameters.

Loss ablations – We also explore the fact that replacing the L1 loss with $smooth_{L1}$ loss faster the convergence of the learning. For this purpose, we optimize our model by using the L1 loss and $smooth_{L1}$, respectively. We get the best accuracy of 87.1%

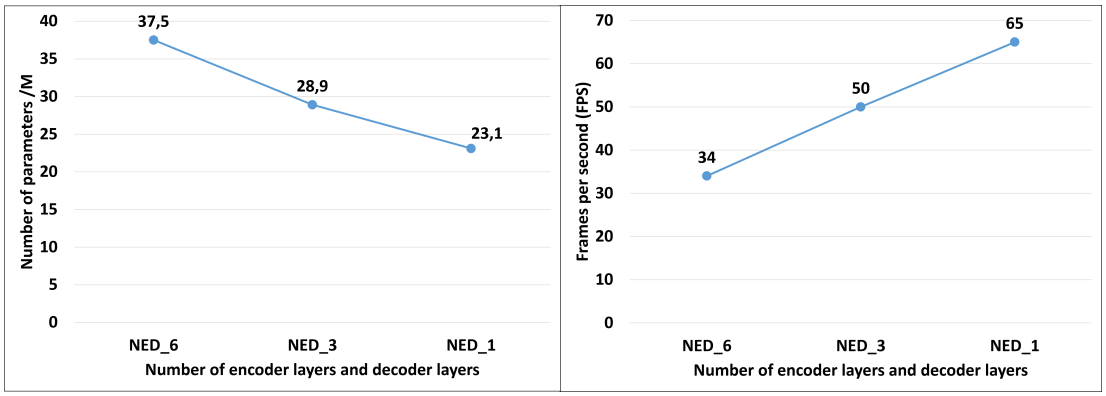


FIGURE 4.15: Complexity comparative results of our model using numbers of encoder layers and decoder layers: parameter complexities (left side) and FPS (right side)

in the 327^{th} epoch when using the $smooth_{L1}$ loss, et the best accuracy of 82.7% in the 492^{th} epoch when using the L1 loss. Indeed, in addition to a faster convergence, we improved further the accuracy by using the $smooth_{L1}$ loss to optimize our model.

4.4.3 Performance comparative study

To evaluate the performance of our method we considered two scenarios: 1) we have evaluated the performance of the cervix detection model (stage 1) and 2) we have evaluated the performance of the global pipeline namely cervix detection and heat classification (stage 1 + stage 2). The obtained performances in both scenarios have been compared with those of the state of the art namely HOG detector [50], YOLO series [197, 198, 29, 123], DETR-ResNet50/100 [37], Deformable-DETR-R50 [277] and Inception-VGG8 [100]. To this end, each method has been trained, validated, and tested on the sets presented in the section 4.4.1. To train the state-of-the-art methods we used their respective source codes which are publicly available and set up their parameters following the recommendations from the referenced articles.

4.4.3.1 Cervix detection performance

Positive vs. Negative frames – We firstly consider the ability of the models to distinguish between positive and negative cervix frames. In addition to their accuracy (ACC)

used in the section 4.4.2, we also calculated their precision (PRE) and recall (REC) as:

$$PRE = \frac{TP}{TP+FP} \tag{4.21}$$

$$REC = \frac{TP}{TP+FN} \tag{4.22}$$

Table 4.6 summarizes the obtained results on the 32552 images of the test set. We

Methods	Accuracy (%)	Precision (%)	Recall (%)
HOG [50]	10	16.7	20
YOLOv2 [197]	51.4	51.4	**100**
YOLOv3 [198]	62.5	67.7	95.8
YOLOv4 [29]	78.1	78.1	85.1
YOLOv5s [123]	83.8	78.9	93.5
DETR-R50 [37]	81	80.4	83.4
DETR-R101 [37]	51.8	58.2	23.3
Deformable-DETR-R50[277]	67.3	63.7	83.2
Our method	**87.1**	**98.9**	76.3

TABLE 4.6: Cervix detection models performance comparison obtained on the test set (32552 images).

can observe that our detection model reached an accuracy of 87.1% which is the best one compared to the other models. The table also shows that, contrary to the other models, our model tends to favor the precision over the recall (98.9% vs. 76.3%). This means that our model has a weak rate of FP which is more suited to the context of our insemination application. Furthermore, it can be observed that the accuracy of the "sliding window" based image descriptor method HOG [50] is far from the deep learning-based method. Indeed, the size of the "sliding window" influences the detection effect due to the multiscale objects. However, in real-time detection, the camera is moving, which causes the size of the cervix in the image to change in real-time. Therefore, it is difficult to determine an universal size for the "sliding window".

Cervix box localization – To measure the ability of the models to detect the cervix and localize it, we calculated their average precision (AP) as:

$$AP = \frac{TP}{TP+FP+FN} \tag{4.23}$$

Where the TPs, FPs and FNs are identified with respect to the ground-truth boxes based on an IoU (Intersection over Union) metric. They correspond for a given frame to what

follow:

- TP a box detected with an IoU$\geq$ threshold compared to the ground-truth box.
- FP all additional boxes detected within the same frame.
- FN missed box (empty frame) or box detected with an IoU$<$threshold compared to the ground-truth box.

Figure 4.16 shows the obtained results of the top 3 models on the 16773 positive frames of the test set using several IoU thresholds [0.5, 0.95]. The curves in the figure permitted to observe that our model reached a precision that is slightly less than the one obtained by the YOLOv5s (75.6 vs. 79.5) for a small IOU threshold (0.5). Nevertheless, our model and contrary to the others succeeded to keep the same level of performance with relatively a high threshold (0.75) which clearly indicates that is has a better cervix localization ability.

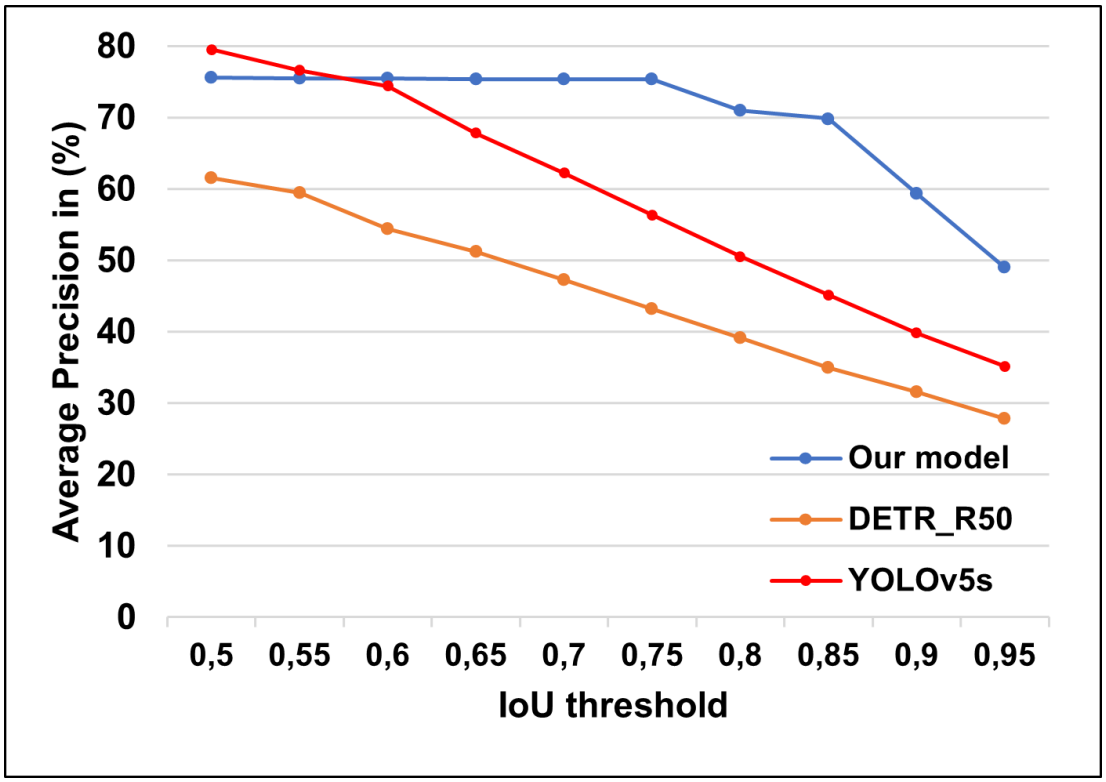


FIGURE 4.16: Performance comparison in term of cow cervix localization obtained on the test set of positive frames (16773 images).

4.4.3.2 Cervix detection and heat classification performance

We measured for each model the mean of average precision (mAP) which takes into consideration the detection quality and classification as well. Indeed, it is not relevant to evaluate separately the classification of the state of the art methods in reason of their unified detection and classification strategy.

$$mAP = \frac{AP_{heat} + AP_{no-heat}}{2} \tag{4.24}$$

$$AP_{class} = \frac{TP_{class}}{TP_{class} + FP_{class} + FN_{class}} \tag{4.25}$$

Where $class = heat, no\text{-}heat$. To this end, we used the 16773 positive frames of the test set and split them according to their heat category. Each AP_{class} has been calculated using several IoU thresholds [0.5,0.95] and has been averaged to obtain a global performance for each class. Table 4.7 shows the obtained results by the best 3 models together with their complexities. We can observe that our method has reached the highest percentage for both classes outperforming widely the state of the art ones. In addition, it has a reasonable complexity making possible its deployment on a smartphone.

Figure 4.17 shows two examples of ground truth and results of the YOLOv5s [123],

Method	AP_{heat}	$AP_{no-heat}$	mAP	Params	FPS
YOLOv5s [123]	59.6	56.2	57.9	**7M**	**151**
DETR-R50 [37]	44.5	40.6	42.5	36.7M	20
Our method	**68.5**	**70.7**	**69.6**	26.5M	46

TABLE 4.7: Performance comparison of the final decision on cervix detection and heat classification obtained on the test set of positive frames (16773 images).

DETR-R50 [37] and our model [123]. From the first line of Figure 4.17, we can observe that both YOLOv5s [123] and DETR-R50 [37] have detected a "false cervix". In contrast, our model can well distinguish the difference between the absence and presence of cervix from an image. Furthermore, we can also find that the prediction of our model is more precise than the others compared to the ground truth from the second line of Figure 4.17 in term of both cervix detection and heat state classification. To compare the performance of our new method with our previous one [100], we evaluated both of them on the validation set of 10 videos and calculated their rates of correct

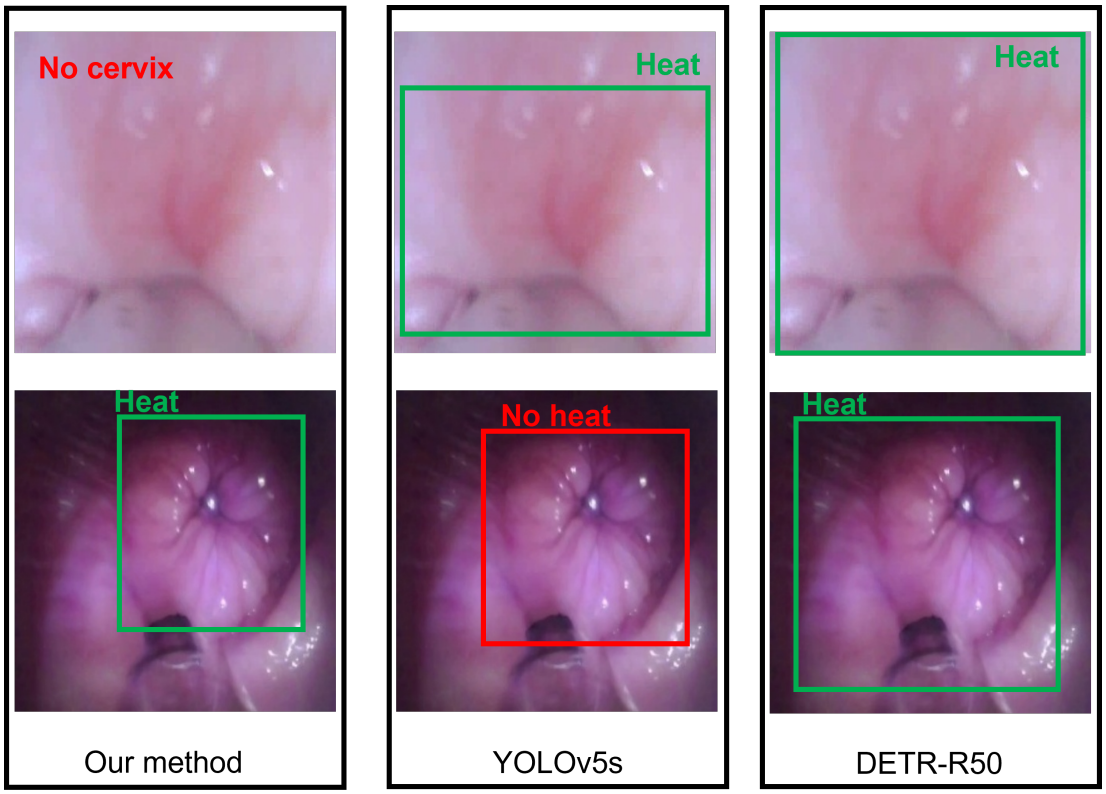


FIGURE 4.17: Some ground truth examples and result examples of the YOLOv5s [123], DETR-R50 [37] and our model.

prediction (heat or no-heat). We also exprimented two other conbinations : DETR-R50 [37] + IVGG8 [100] and YOLOv5s [123] + IVGG8 [100]. Figure 4.18 shows the obtained results on 5 videos from each class. We can observe that all methods are able to correctly predict the heat state. However, our new method shows more confidence in its decision since it gives the highest rates for all the videos.

4.4.4 Attention Visualization analysis

To better explain the choice of our architecture components, we visualize the Attention Heat Maps (AHM) of the last decoder layer of our trained Transformer-Darknet19 model with two object queries. For this purpose, we first consider extracting the attention output weights of the encoder ($atten\_output\_weight^{enc} \in \mathbb{R}^{HW \times HW}$) to visualize the attention map using the reference points as described in [37]. Where H and W are the dimension of feature map $f \in \mathbb{R}^{H \times W \times 1024}$ generated by the CNN backbone (Darknet19). Afterward, we reshape $atten\_output\_weight^{enc}$ so that it has a more interpretable representation of (H, W, H, W). Figure 4.19 shows the attention maps of

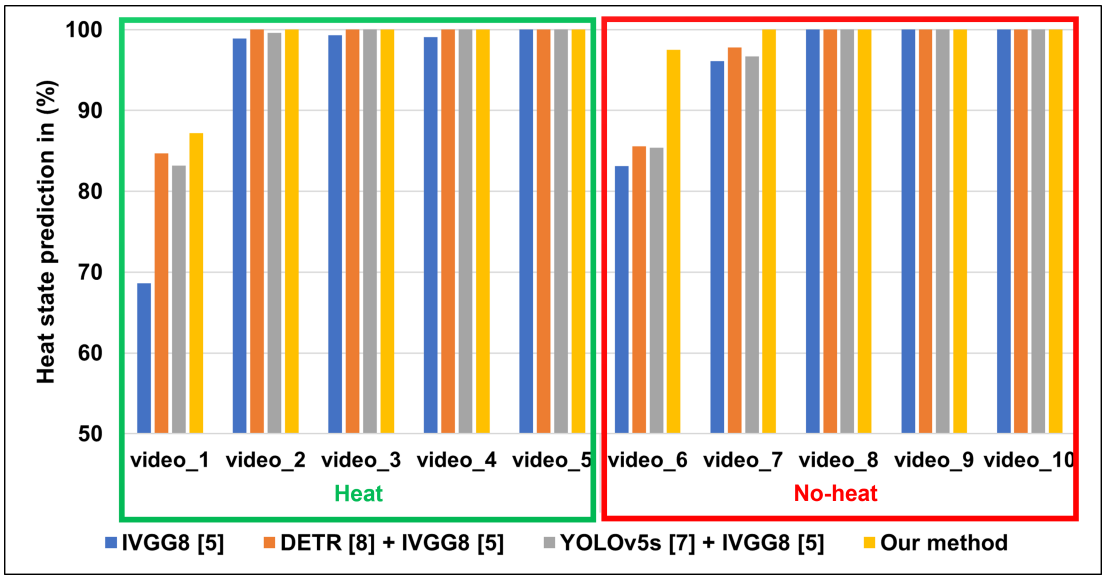


FIGURE 4.18: Results in (%) of heat state prediction on the validation set composed of 10 videos.

the last encoder layer of our trained model, focusing on one point sited at the cervix. Indeed, the encoder seems to be able to focus on the target region.

Furthermore, we also consider extracting the attention output weights of decoder

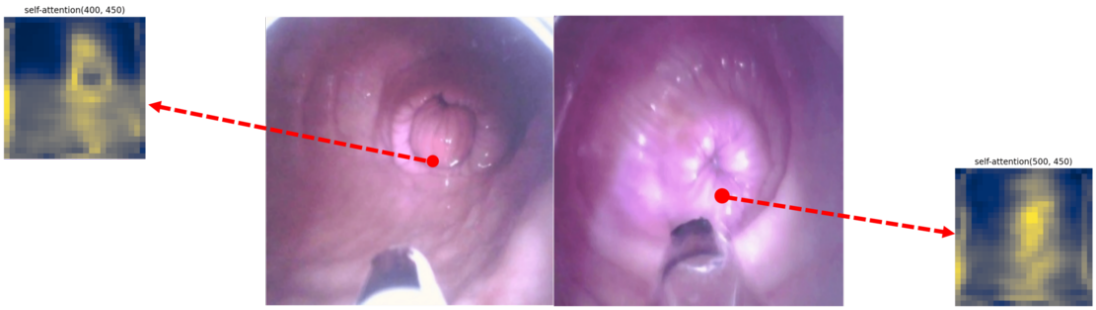


FIGURE 4.19: Encoder self-attention for a reference point

($atten\_output\_weight^{dec} \in \mathbb{R}^{2\times HW}$) to visualize the final attention map. Algorithm 3 summarizes the different steps of Attention Heat Maps (AHME) visualization. First, an input image $I \in \mathbb{R}^{H_0\times W_0\times 3}$ is fed into generate the associated feature map ($H = \frac{H_0}{32}$, $W = \frac{W_0}{32}$). The flattened feature maps $\vec{f}$ are then fed into the transformer module to get the attention output weights of decoder $atten\_output\_weight^{dec}$. Similarly, we reshape $atten\_output\_weight^{dec}$ so that it has a more interpretable representation of $(2, H, W)$, and then split it into two matrices (m_1, m_2) of dimension (H, W). The values of each matrix are normalized between 0 to 255 for drawing the heat maps. Finally, the heat

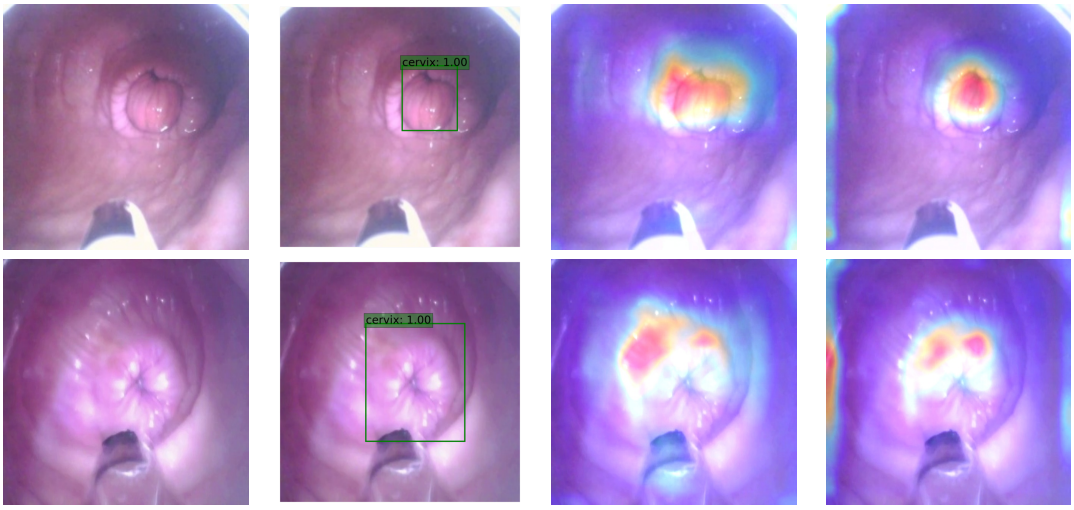
Algorithm 3 Attention Heat Maps of decoder (AHMD) visualization

Input: Cow vagnial endoscopic Image (I) $\in \mathbb{R}^{H_0 \times W_0 \times 3}$
Output: Attention Heat Maps of decoder (AHMD)

$f \in \mathbb{R}^{H \times W \times 1024}$ (H $=\frac{H_0}{32}$, W $= \frac{W_0}{32}$)$=$ DARKNET19(I) ▷ Model in Figure 4.11
$\vec{f} =$ FLATTEN(f)
$attn\_output\_weights_{dec} =$ TRANSFORMER$(\vec{f})$ ▷ Model in Figure 4.12
$attn\_output\_weights_{dec} =$ RESHAPE$(attn_output_weights_{dec})$
$M = \{m_1, m_2\} =$ SPLIT$(attn_output_weights_{dec})$
for each m_i *in* M **do**
    $map =$ NORMALIZATION$(m)_{\,[0,255]}$
    $Heat\_map =$ COLORMAP(map)
    $Heat\_map =$ RESIZE$(Heat_map)$
    $AHM_i = 0.5 \times Heat\_map + I$
end for
$AHMD = \{AHM_i\}_{i=1}^{2}$

maps are resized to f $H_0 \times W_0$ and then overlaid with the original image.

Figure 4.20 shows the obtained decoder self-attention maps of two examples. We



(a) Image orgnial **(b)** Prediction **(c)** Object query1 **(d)** Object query2

FIGURE 4.20: Decoder self-attention for two object queries

can observe that the self-attention of each object query seems to be able to focus on the cervix already, the predictions seem to be the result of taking into consideration the object query1 and object query 2. Therefore, it can explain that our model no longer needs additional object queries, and using two object queries is suitable for our problem.

4.5 Conclusion

A new deep learning-based approach for cow heat analysis from endoscopic images has been proposed. The approach goes through two main stages namely cervix detection and heat classification. The effectiveness of our approach has been demonstrated on our dataset outperforming the state of the art methods. More specifically, our transformer-based detection model reached an accuracy of 87.1% which permitted to increase the confidence level of the final decision of our method for heat prediction in comparison with our previous method [100]. We believe that this new method will further assist the farmer in the insemination operation offering him a precision and fast detection and analysis tool.

Chapter 5

Conclusion

5.1 Summary

Ensuring animal welfare can provide many benefits at the level of environmental protection, the quality and security of food, and also the moral society. This is a social need, but also a challenge. Indeed, to face the increasing population, the intensification of livestock farming is inevitable, which means that more and more animals are confined in a small space, and a farmer has to take care of more and more animals. In this context, animal welfare is being threatened.

Artificial insemination is considered as a biotechnology that can ensure the quality and security of food, improve livestock production, and provide some potential animal welfare, such as avoiding bull injuries and culling unwanted bull calves. However, the two main challenges for artificial insemination are precise heat detection which is the main condition for its sucess and the availability of veterinarians or experts to accomplish this operation at the right time. To this end, we have proposed to develop a computer vision-based intelligent system for artificial insemination assistance. The proposed system permits to analyze endoscopic videos of the genital tract of the cow collected by the Eye Breed device for cervix localization and heat state classification. This can simplify and accelerate the insemination process which is a gospel for farmers. Simultaneously, this technology frees artificial insemination from reliance on blind rectal palpation and on observations of behaviors such as mounting. Moreover, cervix localization has a huge potential to be used for pathology diagnosis such as cow uterine inflammation.

For heat state classification, we proposed a CNN model namely InceptionVGG8. The conducted experiments on two datasets namely our own dataset and a public dataset (kvasir [193]) show the high accuracy of our CNN model (more than 97% for both

datasets) outperforming 19 methods from the state of the art. Moreover, we propose an optimized version of our model for an Android deployment by exploiting several techniques namely quantization, GPU acceleration and video downsampling. The conducted tests on a smartphone show that our heat detection system has a response time of a few seconds.
To address the issue of unavailability of veterinarians, we also proposed an artificial insemination assistance system, which allows to predict the bounding box for cervix localization. For this purpose, a Transformer based detection model namely Transformer-Darknet19 has been specially designed to localize the cervix while considerably reducing the rate of false positive detections. More precisely, the conducted experiments show a high cervix detection accuracy (87.1%) of our Transformer-Darknet19 model outperforming 6 methods from the state of the art and a significant improvement in heat state classification compared with our basic version.

5.2 Recommendations for further research

In the context of cow heat detection and its relevance for insemination [58], it has been reported that the highest conception rates for artificial insemination occurred between 4 and 12 hours after the onset of heat as shown in table 5.1. Indeed, we can observe

Interval from onset of heat to artificial insemination (hours)	Number of inseminations	Conception rate (%)
0 to 4	32	43.1
4 to 8	735	**50.9**
8 to 12	677	**51.1**
12 to 16	459	46.2
16 to 20	317	28.1
20 to 24	139	31.7
24 to 26	7	14.3

TABLE 5.1: Conception rates of dairy cows inseminated at different times after the onset of heat [58]

from the table 5.1 insemination at different times can lead to a huge difference in the conception rate. Therefore, it would be interesting to construct a cow estrus stage classification dataset and train a related classifier that not only predicts whether a cow is in heat, but also the heat stage. The classification of heat stages is a more refined classification, which requires a stronger ability of classification. This would be a challenge

for the current model. In recent years, attention mechanisms [244] have achieved great success in the field of computer vision, such as the transformer-based detection of the cervix that we developed. Therefore, I recommend developing a CNN-Transformer-based model for this task in the future.

Metritis and endometritis are common uterine diseases that impact up to 40% of lactating cows early postpartum [221]. Metritis (inflammation of the uterus) is a bacterial infection and refers to the early post-partum inflammation of the whole uterus, while endometritis is limited to the lining of the uterus (endometrium). Both diseases are a consequence of sustained infection of the uterus caused by pathogenic bacteria, such as Arcanobacterium pyogenes [145]. This type of disease reduces milk production[196], reproductive performance [187, 172], increases risk of culling during lactation [77] and increases veterinary costs. This will bring a huge loss to farmers. The total costs per case of metritis have been calculated to approximate to US$ 329–386 [188]. Therefore, early identification of sick cows is crucial to ensure a healthy herd.

Cytological examination of the reproductive tract is considered as the standard diagnosis method. However, this method is laborious since it is performed in specialized laboratories and requires experts for results interpretation. It has been shown in [19] that metritis and endometritis involve changes in the color and shape of the genital tract. Form this obsevation, it is reasonable to consider that computer vision based detection techniques can be exploited to identify such pathologies. In this sense, I recommend as a direction of future research, the development of this type of techniques.

Bibliography

[1] Mohammed N Abu Al-Qumboz and Samy S Abu-Naser. “Spinach expert system: diseases and symptoms”. In: *International Journal of Academic Information Systems Research (IJAISR)* 3.3 (2019), pp. 16–22.

[2] Brahim Achour et al. “Image analysis for individual identification and feeding behaviour monitoring of dairy cows based on Convolutional Neural Networks (CNN)”. In: *Biosystems Engineering* 198 (2020), pp. 31–49.

[3] Jamil Ahmad et al. “Endoscopic image classification and retrieval using clustered convolutional features”. In: *Journal of medical systems* 41.12 (2017), p. 196.

[4] Ali Alameer, Ilias Kyriazakis, and Jaume Bacardit. “Automated recognition of postures and drinking behaviour for the detection of compromised health in pigs”. In: *Scientific reports* 10.1 (2020), pp. 1–15.

[5] Abel Alarcón-Salvatierra et al. “A Rule-Based Expert System for Cow Disease Diagnosis”. In: *2nd International Conference on ICTs in Agronomy and Environment*. Springer. 2019, pp. 29–37.

[6] Hong-Yuan Mark Liao Alexey Bochkovskiy Chien-Yao Wang. “YOLOv4: YOLOv4: Optimal Speed and Accuracy of Object Detection”. In: *arXiv* (2020).

[7] Sharib Ali et al. “Endoscopy artifact detection (EAD 2019) challenge dataset”. In: *arXiv preprint arXiv:1905.03209* (2019).

[8] Rodney D Allrich. “Estrous behavior and detection in cattle.” In: *The Veterinary clinics of North America. Food animal practice* 9.2 (1993), pp. 249–262.

[9] Maher Alsaaod, Mahmoud Fadul, and Adrian Steiner. “Automatic lameness detection in cattle”. In: *The veterinary journal* 246 (2019), pp. 35–44.

[10] Izzeddin A Alshawwa, Abeer A Elsharif, and Samy S Abu-Naser. “An Expert System for Coconut Diseases Diagnosis”. In: *International Journal of Academic Engineering Research (IJAER)* 3.4 (2019).

[11] John D Anderson et al. “Economic impact of artificial insemination vs. natural mating for beef cattle herds”. In: (2008).

[12] William Andrew, Colin Greatwood, and Tilo Burghardt. “Visual localisation and individual identification of holstein friesian cattle via deep learning”. In: *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 2017, pp. 2850–2859.

[13] Claudia Arcidiacono, M Mancino, and SMC Porto. “Moving mean-based algorithm for dairy cow’s oestrus detection from uniaxial-accelerometer data acquired in a free-stall barn”. In: *Computers and Electronics in Agriculture* 175 (2020), p. 105498.

[14] Safa Ayadi et al. “Dairy cow rumination detection: A deep learning approach”. In: *International Workshop on Distributed Computing for Emerging Smart Networks*. Springer. 2020, pp. 123–139.

[15] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. “Layer normalization”. In: *arXiv preprint arXiv:1607.06450* (2016).

[16] Peter JH Ball and Andy R Peters. *Reproduction in cattle*. John Wiley & Sons, 2008.

[17] Jorge A Barrientos-Blanco et al. “Expected value of crossbred dairy cattle artificial insemination breeding strategies in virgin heifers and lactating cows”. In: *Livestock science* 211 (2018), pp. 66–74.

[18] Pietro Sampaio Baruselli et al. “Timed artificial insemination: current challenges and recent advances in reproductive efficiency in beef and dairy herds in Brazil”. In: *Animal Reproduction (AR)* 14.3 (2018), pp. 558–571.

[19] TP Basarab and VY Stefanyk. “Hysteroscopic investigation of dairy cows uterus with subclinical endometritis”. In: *Scientific Messenger of LNU of Veterinary Medicine and Biotechnologies. Series: Veterinary Sciences* 18.3 (71) (2016), pp. 218–220.

[20] Rotimi-williams BELLO, Abdullah Zawawi Hj TALIB, and Ahmad Sufril Azlan Bin MOHAMED. “Deep learning-based Architectures for recognition of cow using cow nose image pattern”. In: *Gazi University Journal of Science* 33.3 (2020), pp. 831–844.

[21] Said Benaissa et al. "Calving and estrus detection in dairy cattle using a combination of indoor localization and accelerometer sensors". In: *Computers and Electronics in Agriculture* 168 (2020), p. 105153.

[22] Madonna Benjamin and Steven Yik. "Precision livestock farming in swine welfare: a review for swine practitioners". In: *Animals* 9.4 (2019), p. 133.

[23] Daniel Berckmans. "Precision livestock farming technologies for welfare management in intensive livestock systems". In: *Rev. Sci. Tech* 33.1 (2014), pp. 189–196.

[24] Donagh P Berry. "Symposium review: Breeding a better cow—Will she be adaptable?" In: *Journal of dairy science* 101.4 (2018), pp. 3665–3685.

[25] DP Berry et al. "Choice of artificial insemination beef bulls used to mate with female dairy cattle". In: *Journal of Dairy Science* 103.2 (2020), pp. 1701–1710.

[26] Jeffrey Bewley. "Precision dairy farming: advanced analysis solutions for future profitability". In: *Proceedings of the first North American conference on precision dairy management, Toronto, Canada*. 2010, pp. 2–5.

[27] Ran Bezen, Yael Edan, and Ilan Halachmi. "Computer vision system for measuring individual cow feed intake using RGB-D camera and deep learning algorithms". In: *Computers and Electronics in Agriculture* 172 (2020), p. 105345.

[28] NB Blazquez et al. "A pheromonal function for the perineal skin glands in the cow." In: *The Veterinary Record* 123.2 (1988), pp. 49–50.

[29] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. "Yolov4: Optimal speed and accuracy of object detection". In: *arXiv preprint arXiv:2004.10934* (2020).

[30] Amelie Bonde et al. "Structural vibration sensing to evaluate animal activity on a pig farm". In: *Proceedings of the First Workshop on Data Acquisition To Analysis*. 2018, pp. 25–26.

[31] Joe T Braden. *Estrus detector*. US Patent 7,137,359. 2006.

[32] Markus Braun et al. "EuroCity Persons: A Novel Benchmark for Person Detection in Traffic Scenes". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2019), pp. 1–1. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2019.2897684.

[33] Heather J Bray and Rachel A Ankeny. "Happy chickens lay tastier eggs: motivations for buying free-range eggs in Australia". In: *Anthrozoös* 30.2 (2017), pp. 213–226.

[34] Donald M Broom. "Animal welfare: an aspect of care, sustainability, and food quality required by the public". In: *Journal of veterinary medical education* 37.1 (2010), pp. 83–88.

[35] Johannes Brünger et al. "Panoptic Segmentation of Individual Pigs for Posture Recognition". In: *Sensors* 20.13 (2020). ISSN: 1424-8220. DOI: 10.3390/s20133710. URL: https://www.mdpi.com/1424-8220/20/13/3710.

[36] Zhaowei Cai and Nuno Vasconcelos. "Cascade r-cnn: Delving into high quality object detection". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 6154–6162.

[37] Nicolas Carion et al. "End-to-end object detection with transformers". In: *European conference on computer vision*. Springer. 2020, pp. 213–229.

[38] E Cha et al. "The cost of different types of lameness in dairy cows calculated by dynamic programming". In: *Preventive veterinary medicine* 97.1 (2010), pp. 1–8.

[39] Jung-woo Chae and Hyun-chong Cho. "Identifying the mating posture of cattle using deep learning-based object detection with networks of various settings". In: *Journal of Electrical Engineering & Technology* 16.3 (2021), pp. 1685–1692.

[40] Chen Chen et al. "A computer vision approach for recognition of the engagement of pigs with different enrichment objects". In: *Computers and Electronics in Agriculture* 175 (2020), p. 105580.

[41] Chen Chen et al. "A kinetic energy model based on machine vision for recognition of aggressive behaviours among group-housed pigs". In: *Livestock Science* 218 (2018), pp. 70–78.

[42] Chen Chen et al. "Classification of drinking and drinker-playing in pigs by a video-based deep learning method". In: *Biosystems Engineering* 196 (2020), pp. 1–14.

[43] Chen Chen et al. "Image motion feature extraction for recognition of aggressive behaviors among group-housed pigs". In: *Computers and Electronics in Agriculture* 142 (2017), pp. 380–387.

[44] Chen Chen et al. “Recognition of aggressive episodes of pigs based on convolutional neural network and long short-term memory”. In: *Computers and Electronics in Agriculture* 169 (2020), p. 105166.

[45] Chen Chen et al. “Recognition of feeding behaviour of pigs and determination of feeding time of each pig by a video-based deep learning method”. In: *Computers and Electronics in Agriculture* 176 (2020), p. 105642.

[46] Kai-Xuan Chen et al. “Covariance descriptors on a Gaussian manifold and their application to image set classification”. In: *Pattern Recognition* 107 (2020), p. 107463.

[47] Dami Choi et al. “On empirical comparisons of optimizers for deep learning”. In: *arXiv preprint arXiv:1910.05446* (2019).

[48] François Chollet. “Xception: Deep learning with depthwise separable convolutions”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 1251–1258.

[49] Jifeng Dai et al. “R-fcn: Object detection via region-based fully convolutional networks”. In: *Advances in neural information processing systems* 29 (2016).

[50] Navneet Dalal and Bill Triggs. “Histograms of oriented gradients for human detection”. In: *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR’05)*. Vol. 1. Ieee. 2005, pp. 886–893.

[51] Emanuela Dalla Costa et al. “Development of the Horse Grimace Scale (HGS) as a pain assessment tool in horses undergoing routine castration”. In: *PLoS one* 9.3 (2014), e92281.

[52] JC Dalton et al. “Artificial insemination of cattle: Description and assessment of a training program for veterinary students”. In: *Journal of Dairy Science* 104.5 (2021), pp. 6295–6303.

[53] Matthieu De Clercq, Anshu Vats, and Alvaro Biel. “Agriculture 4.0: The future of farming technology”. In: *Proceedings of the World Government Summit, Dubai, UAE* (2018), pp. 11–13.

[54] A De Vries et al. “Exploring the impact of sexed semen on the structure of the dairy industry”. In: *Journal of dairy science* 91.2 (2008), pp. 847–856.

[55] Agathe Decherf and Pierrick Drevillon. *Device for the atraumatic transfer of a material or substance with a reproductive, therapeutic or diagnostic purpose into female mammals*. US Patent 10,675,133. 2020. URL: https://patentimages.storage.googleapis.com/7c/ff/f4/48b9ea854760e9/US10675133.pdf.

[56] Jia Deng et al. "Imagenet: A large-scale hierarchical image database". In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255.

[57] Dewa Gede Hendra Divayana. "Development of duck diseases expert system with applying alliance method at bali provincial livestock office". In: *USA: Corllins University* (2014).

[58] MBG Dransfield et al. "Timing of insemination for dairy cows identified in estrus by a radiotelemetric estrus detection system". In: *Journal of dairy science* 81.7 (1998), pp. 1874–1882.

[59] Ritaban Dutta et al. "Dynamic cattle behavioural classification using supervised ensemble classifiers". In: *Computers and electronics in agriculture* 111 (2015), pp. 18–28.

[60] CR Eastwood, DF Chapman, and MS Paine. "Networks of practice for co-construction of agricultural decision support systems: case studies of precision dairy farms in Australia". In: *Agricultural Systems* 108 (2012), pp. 10–18.

[61] Mohammed Ibraheem El Kahlout and Samy S Abu-Naser. "An Expert System for Citrus Diseases Diagnosis". In: (2019).

[62] Hosni Q El-Mashharawi et al. "An Expert System for Arthritis Diseases Diagnosis Using SL5 Object". In: (2019).

[63] Felicia Engmann et al. "Prolonging the lifetime of wireless sensor networks: a review of current techniques". In: *Wireless Communications and Mobile Computing* 2018 (2018).

[64] M. Everingham et al. *The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results*. URL: http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html.

[65] M. Everingham et al. *The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results*. URL: http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

[66] Pedro Felzenszwalb, David McAllester, and Deva Ramanan. "A discriminatively trained, multiscale, deformable part model". In: *2008 IEEE conference on computer vision and pattern recognition*. Ieee. 2008, pp. 1–8.

[67] DM Ferguson and Robin Dorothy Warner. "Have we underestimated the impact of pre-slaughter stress on meat quality in ruminants?" In: *Meat science* 80.1 (2008), pp. 12–19.

[68] Carola Fischer-Tenhagen et al. "Training dogs on a scent platform for oestrus detection in cows". In: *Applied Animal Behaviour Science* 131.1-2 (2011), pp. 63–70.

[69] P Flecknell. "Analgesia from a veterinary perspective". In: *British Journal of Anaesthesia* 101.1 (2008), pp. 121–124.

[70] Masato Futagawa et al. "Study of a wireless multimodal sensing system integrated with an electrical conductivity sensor and a temperature sensor for the health control of cows". In: *IEEJ Transactions on Electrical and Electronic Engineering* 6.2 (2011), pp. 93–96.

[71] Alexander Gallagher. "Interventional radiology and interventional endoscopy in treatment of nephroureteral disease in the dog and cat". In: *Veterinary Clinics: Small Animal Practice* 48.5 (2018), pp. 843–862.

[72] Jing Gao et al. "Towards Self-Supervision for Video Identification of Individual Holstein-Friesian Cattle: The Cows2021 Dataset". In: *arXiv preprint arXiv:2105.01938* (2021).

[73] Ina Gaude et al. "Estrus signs in Holstein Friesian dairy cows and their reliability for ovulation detection in the context of visual estrus detection". In: *Livestock Science* 245 (2021), p. 104449.

[74] Girma Gebresenbet et al. "Vibration levels and frequencies on vehicle and animals during transport". In: *Biosystems Engineering* 110.1 (2011), pp. 10–19.

[75] Ross Girshick. "Fast r-cnn". In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 1440–1448.

[76] Ross Girshick et al. "Rich feature hierarchies for accurate object detection and semantic segmentation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014, pp. 580–587.

[77] Mauricio Javier Giuliodori et al. “Metritis in dairy cows: Risk factors and reproductive performance”. In: *Journal of dairy science* 96.6 (2013), pp. 3621–3631.

[78] Karina B Gleerup et al. “An equine pain face”. In: *Veterinary anaesthesia and analgesia* 42.1 (2015), pp. 103–114.

[79] H Charles J Godfray et al. “Food security: the challenge of feeding 9 billion people”. In: *science* 327.5967 (2010), pp. 812–818.

[80] ME Goldberg. “Pain recognition and scales for livestock patients”. In: *J. Dairy Vet. Anim. Res* 7 (2018), pp. 236–239.

[81] Serap Göncü and Nazan Koluman. “The sensor technologies for more efficient cow reproduction systems”. In: *MOJ Eco Environ Sci* 4.3 (2019), pp. 128–131.

[82] Paula A Gonzalez-Rivas et al. “Effects of heat stress on animal physiology, metabolism, and meat quality: A review”. In: *Meat Science* 162 (2020), p. 108025.

[83] T. Grandin. “12 - Animal welfare and food safety at the slaughter plant”. In: *Improving the Safety of Fresh Meat*. Ed. by John N. Sofos. Woodhead Publishing Series in Food Science, Technology and Nutrition. Woodhead Publishing, 2005, pp. 244–258. ISBN: 978-1-85573-955-0. DOI: https://doi.org/10.1533/9781845691028.2.244. URL: https://www.sciencedirect.com/science/article/pii/B9781855739550500121.

[84] Temple Grandin. “Animal welfare and society concerns finding the missing link”. In: *Meat science* 98.3 (2014), pp. 461–469.

[85] Bénédicte Grimard et al. “Relationships between welfare and reproductive performance in French dairy herds”. In: *The Veterinary Journal* 248 (2019), pp. 1–7.

[86] MJ Guesgen et al. “Coding and quantification of a facial expression for pain in lambs”. In: *Behavioural processes* 132 (2016), pp. 49–56.

[87] Yangyang Guo et al. “Detection of cow mounting behavior using region geometry and optical flow characteristics”. In: *Computers and Electronics in Agriculture* 163 (2019), p. 104828.

[88] Ying Guo et al. “Animal behaviour understanding using wireless sensor networks”. In: *Proceedings. 2006 31st IEEE Conference on Local Computer Networks*. IEEE. 2006, pp. 607–614.

[89] Erika Gusterer et al. “Sensor technology to support herd health monitoring: Using rumination duration and activity measures as unspecific variables for the early detection of dairy cows with health deviations”. In: *Theriogenology* 157 (2020), pp. 61–69.

[90] Juan Haladjian et al. “A wearable sensor system for lameness detection in dairy cattle”. In: *Multimodal Technologies and Interaction* 2.2 (2018), p. 27.

[91] AS Hancock et al. “An assessment of dairy herd bulls in southern Australia: 1. Management practices and bull breeding soundness evaluations”. In: *Journal of dairy science* 99.12 (2016), pp. 9983–9997.

[92] Bjørn Gunnar Hansen and Olav Østerås. “Farmer welfare and animal welfare-Exploring the relationship between farmer’s occupational well-being and stress, farm expansion and animal welfare”. In: *Preventive veterinary medicine* 170 (2019), p. 104741.

[93] Mark F Hansen et al. “Towards on-farm pig face recognition using convolutional neural networks”. In: *Computers in Industry* 98 (2018), pp. 145–152.

[94] Nesrein M Hashem, Antonio González-Bulnes, and Alfonso J Rodriguez-Morales. “Animal welfare and livestock supply chain sustainability under the COVID-19 outbreak: An overview”. In: *Frontiers in veterinary science* 7 (2020), p. 679.

[95] Ira Steven Hatch. *Estrus detection device*. US Patent 8,133,188. 2012.

[96] Daniel T Haydon, Rowland R Kao, and R Paul Kitching. “The UK foot-and-mouth disease outbreak—the aftermath”. In: *Nature Reviews Microbiology* 2.8 (2004), pp. 675–681.

[97] Jun-Yan He et al. “Hookworm detection in wireless capsule endoscopy images with deep learning”. In: *IEEE Transactions on Image Processing* 27.5 (2018), pp. 2379–2392.

[98] Kaiming He et al. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[99] Kaiming He et al. “Spatial pyramid pooling in deep convolutional networks for visual recognition”. In: *IEEE transactions on pattern analysis and machine intelligence* 37.9 (2015), pp. 1904–1916.

[100] Ruiwen He et al. “A CNN-based methodology for cow heat analysis from endoscopic images”. In: *Applied Intelligence* (2021), pp. 1–14.

[101] Amruta Helwatkar, Daniel Riordan, and Joseph Walsh. “Sensor technology for animal health monitoring”. In: *International Journal on Smart Sensing and Intelligent Systems* 7.5 (2020).

[102] Shogo Higaki et al. “Estrus detection using background image subtraction technique in tie-stalled cows”. In: *Animals* 11.6 (2021), p. 1795.

[103] William G Hill. “Is continued genetic improvement of livestock sustainable?” In: *Genetics* 202.3 (2016), pp. 877–881.

[104] Toshiaki Hirasawa et al. “Application of artificial intelligence using a convolutional neural network for detecting gastric cancer in endoscopic images”. In: *Gastric Cancer* 21.4 (2018), pp. 653–660.

[105] Tetsuya Hirata et al. “A study on estrus detection of cattle combining video image and sensor information”. In: *International Conference on Big Data Analysis and Deep Learning Applications*. Springer. 2018, pp. 267–273.

[106] Almog Hitelman et al. “Biometric identification of sheep via a machine-vision system”. In: *Computers and Electronics in Agriculture* 194 (2022), p. 106713.

[107] Henk Hogeveen et al. “Novel ways to use sensor data to improve mastitis management”. In: *Journal of Dairy Science* 104.10 (2021), pp. 11317–11332.

[108] SA Holden and ST Butler. “Review: Applications and benefits of sexed semen in dairy and beef herds. Animal 12, s97–s103”. In: *Proceedings-New Zealand Society of Anima l Production* (2018).

[109] A Holman et al. “Comparison of oestrus detection methods in dairy cattle”. In: *Veterinary Record* 169.2 (2011), pp. 47–47.

[110] Richard M. Hopper. *Bovine reproduction*. John Wiley & Sons, 2014.

[111] Andrew Howard et al. “Searching for mobilenetv3”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 1314–1324.

[112] Andrew G Howard et al. “Mobilenets: Efficient convolutional neural networks for mobile vision applications”. In: *arXiv preprint arXiv:1704.04861* (2017).

[113] Gao Huang et al. “Densely connected convolutional networks”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 4700–4708.

[114] Guang-Bin Huang, Qin-Yu Zhu, and Chee-Kheong Siew. "Extreme learning machine: a new learning scheme of feedforward neural networks". In: *2004 IEEE international joint conference on neural networks (IEEE Cat. No. 04CH37541)*. Vol. 2. Ieee. 2004, pp. 985–990.

[115] Guang-Bin Huang, Qin-Yu Zhu, and Chee-Kheong Siew. "Extreme learning machine: theory and applications". In: *Neurocomputing* 70.1-3 (2006), pp. 489–501.

[116] Sergey Ioffe and Christian Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift". In: *International conference on machine learning*. PMLR. 2015, pp. 448–456.

[117] K Abdul Jabbar et al. "Early and non-intrusive lameness detection in dairy cows using 3-dimensional video". In: *Biosystems engineering* 153 (2017), pp. 63–69.

[118] Vidit Jain and Erik Learned-Miller. *Fddb: A benchmark for face detection in unconstrained settings*. Tech. rep. UMass Amherst technical report, 2010.

[119] Jedrzej M Jaskowski et al. "Modern techniques of teaching bovine rectal palpation: Opportunities, benefits and disadvantages of new educational devices". In: *Med Weter* 76 (2020), pp. 5–10.

[120] Jędrzej M Jaśkowski et al. "Rectal palpation for pregnancy in cows: A relic or an alternative to modern diagnostic methods". In: *Med. Weter* 75.5 (2019), pp. 259–264.

[121] Dan B Jensen, Henk Hogeveen, and Albert De Vries. "Bayesian integration of sensor information and a multivariate dynamic linear model for prediction of dairy cow mastitis". In: *Journal of Dairy Science* 99.9 (2016), pp. 7344–7361.

[122] Min Jiang et al. "Automatic behavior recognition of group-housed goats using deep learning". In: *Computers and Electronics in Agriculture* 177 (2020), p. 105706.

[123] Glenn Jocher. *ultralytics/yolov5: v3.1 - Bug Fixes and Performance Improvements*. Version v3.1. accessed on 18 May 2020. Oct. 2020. DOI: 10.5281/zenodo.4154370. URL: \url{https://github.com/ultralytics/yolov5}.

[124] Dorothea Johnen, Wolfgang Heuwieser, and Carola Fischer-Tenhagen. "How to train a dog to detect cows in heat—training and success". In: *Applied Animal Behaviour Science* 171 (2015), pp. 39–46.

[125] Julia Johns, Antonia Patt, and Edna Hillmann. "Do bells affect behaviour and heart rate variability in grazing dairy cows?" In: *PloS one* 10.6 (2015), e0131632.

[126] Maria Jorquera-Chavez et al. "Computer vision and remote sensing to assess physiological responses of cattle to pre-slaughter stress, and its impact on beef quality: A review". In: *Meat science* 156 (2019), pp. 11–22.

[127] Zhao Kaixuan and He Dongjian. "Recognition of individual dairy cattle based on convolutional neural networks." In: *Transactions of the Chinese Society of Agricultural Engineering* 31.5 (2015).

[128] Andreas Kamilaris and Francesc X Prenafeta-Boldú. "Deep learning in agriculture: A survey". In: *Computers and electronics in agriculture* 147 (2018), pp. 70–90.

[129] X Kang, XD Zhang, and G Liu. "Accurate detection of lameness in dairy cattle with computer vision: A new and individualized detection strategy based on the analysis of the supporting phase". In: *Journal of Dairy Science* 103.11 (2020), pp. 10628–10638.

[130] Murat Karabatak and M Cevdet Ince. "An expert system for detection of breast cancer based on association rules and neural network". In: *Expert systems with Applications* 36.2 (2009), pp. 3465–3469.

[131] Amanpreet Kaur, Munish Kumar, and MK Jindal. "Cattle identification with muzzle pattern using computer vision technology: a critical review and prospective". In: *Soft Computing* (2022), pp. 1–25.

[132] Kevin G Keegan et al. "Assessment of repeatability of a wireless, inertial sensor–based lameness evaluation system for horses". In: *American journal of veterinary research* 72.9 (2011), pp. 1156–1163.

[133] Kevin G Keegan et al. "Evaluation of a sensor-based system of motion analysis for detection and quantification of forelimb and hind limb lameness in horses". In: *American journal of veterinary research* 65.5 (2004), pp. 665–670.

[134] Mehmet Kerim Yucel et al. "Wildest Faces: Face Detection and Recognition in Violent Settings". In: *arXiv e-prints* (2018), arXiv–1805.

[135] Jonathan Krause et al. "3D Object Representations for Fine-Grained Categorization". In: *4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13)*. Sydney, Australia, 2013.

[136] Sanya Kuankid, Tanadon Rattanawong, and Apinan Aurasopon. "Classification of the cattle's behaviors by using accelerometer data with simple behavioral technique". In: *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific*. IEEE. 2014, pp. 1–4.

[137] Anuj Kumar and Gerhard P Hancke. "A zigbee-based animal health monitoring system". In: *IEEE sensors Journal* 15.1 (2014), pp. 610–617.

[138] Pankaj Kumar et al. "Heat detection techniques in cattle and buffalo." In: *Veterinary World* 6.6 (2013).

[139] Santosh Kumar et al. "Analytical study of animal biometrics: A technical survey". In: *Animal Biometrics*. Springer, 2017, pp. 21–78.

[140] Santosh Kumar et al. "Deep learning framework for recognition of cattle using muzzle point image pattern". In: *Measurement* 116 (2018), pp. 1–17.

[141] Kae Hsiang Kwong et al. "Adaptation of wireless sensor network for farming industries". In: *2009 Sixth International Conference on Networked Sensing Systems (INSS)*. IEEE. 2009, pp. 1–4.

[142] Carl Johan Lagerkvist and Sebastian Hess. "A meta-analysis of consumer willingness to pay for farm animal welfare". In: *European Review of Agricultural Economics* 38.1 (2011), pp. 55–78.

[143] AJP Lane and DC Wathes. "An electronic nose to detect changes in perineal odors associated with estrus in the cow". In: *Journal of dairy science* 81.8 (1998), pp. 2145–2150.

[144] Thérésa Lebacq, Philippe V Baret, and Didier Stilmant. "Sustainability indicators for livestock farming. A review". In: *Agronomy for sustainable development* 33.2 (2013), pp. 311–327.

[145] SJ LeBlanc et al. "Defining and diagnosing postpartum clinical endometritis and its impact on reproductive performance in dairy cows". In: *Journal of dairy science* 85.9 (2002), pp. 2223–2236.

[146] Gabriel Carreira Lencioni et al. "Pain assessment in horses using automatic facial expression recognition through deep learning-based modeling". In: *PloS one* 16.10 (2021), e0258672.

[147] Dan Li et al. "Mounting behaviour recognition for pigs based on deep learning". In: *Sensors* 19.22 (2019), p. 4924.

[148] Hua Li et al. “Development of a remote monitoring system for henhouse environment based on IoT technology”. In: *Future Internet* 7.3 (2015), pp. 329–341.

[149] Shijun Li et al. “Individual dairy cow identification based on lightweight convolutional neural network”. In: *Plos one* 16.11 (2021), e0260510.

[150] Xiangyuan Li et al. “Deep cascaded convolutional models for cattle pose estimation”. In: *Computers and Electronics in Agriculture* 164 (2019), p. 104885. ISSN: 0168-1699. DOI: https://doi.org/10.1016/j.compag.2019.104885. URL: https://www.sciencedirect.com/science/article/pii/S016816991831874X.

[151] Tsung-Yi Lin et al. “Feature pyramid networks for object detection”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 2117–2125.

[152] Tsung-Yi Lin et al. “Microsoft coco: Common objects in context”. In: *European conference on computer vision*. Springer. 2014, pp. 740–755.

[153] Dong Liu et al. “A computer vision-based method for spatial-temporal action recognition of tail-biting behaviour in group-housed pigs”. In: *Biosystems Engineering* 195 (2020), pp. 27–41.

[154] Li Liu et al. “Deep learning for generic object detection: A survey”. In: *International journal of computer vision* 128.2 (2020), pp. 261–318.

[155] Songtao Liu, Di Huang, et al. “Receptive field block net for accurate and fast object detection”. In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 385–400.

[156] Wei Liu et al. “Ssd: Single shot multibox detector”. In: *European conference on computer vision*. Springer. 2016, pp. 21–37.

[157] Alvaro Llaria et al. “Geolocation and monitoring platform for extensive farming in mountain pastures”. In: *2015 IEEE International Conference on Industrial Technology (ICIT)*. IEEE. 2015, pp. 2420–2425.

[158] P Lonergan. “Historical and futuristic developments in bovine semen technology”. In: *animal* 12.s1 (2018), s4–s18.

[159] Emily D Lord et al. “Evaluating the expected value of beef reproduction strategies in an era of volatile feed and cattle prices”. In: *Livestock Science* 174 (2015), pp. 113–125.

[160] P Løvendahl and MGG Chagunda. "On the use of physical activity monitoring for estrus detection in dairy cows". In: *Journal of Dairy Science* 93.1 (2010), pp. 249–259.

[161] Jayson L Lusk and F Bailey Norwood. "Animal welfare economics". In: *Applied Economic Perspectives and Policy* 33.4 (2011), pp. 463–483.

[162] Luis Macias-Valle et al. "Evaluation of sheep sinonasal endoscopic anatomy as a model for rhinologic research". In: *World journal of otorhinolaryngology-head and neck surgery* 4.04 (2018), pp. 268–272.

[163] Vasile Maciuc et al. "USING THE STATISTIC MODUL WITHIN AN ORIGINAL COMPUTER PROGRAM FOR DAIRY CATTEL FARMS". In: *International Multidisciplinary Scientific GeoConference: SGEM* 17 (2017), pp. 1191–1197.

[164] Abdul Majid et al. "Classification of stomach infections: A paradigm of convolutional neural network along with classical features fusion and selection". In: *Microscopy Research and Technique* 83.5 (2020), pp. 562–576.

[165] Alexandra Manzoli et al. "Volatile compounds monitoring as indicative of female cattle fertile period using electronic nose". In: *Sensors and Actuators B: Chemical* 282 (2019), pp. 609–616.

[166] Qi-Chao Mao et al. "Mini-YOLOv3: real-time object detector for embedded applications". In: *Ieee Access* 7 (2019), pp. 133529–133538.

[167] Mathieu Marsot et al. "An adaptive pig face recognition approach using Convolutional Neural Networks". In: *computers and Electronics in Agriculture* 173 (2020), p. 105386.

[168] Gabriele Mattachini et al. "Monitoring feeding behaviour of dairy cows using accelerometers". In: *Journal of Agricultural Engineering* 47.1 (2016), pp. 54–58.

[169] Krista M McLennan et al. "Development of a facial expression scale using footrot and mastitis as models of pain in sheep". In: *Applied Animal Behaviour Science* 176 (2016), pp. 19–26.

[170] Barbara Helen Meads. *Riding indicator*. US Patent 5,839,390. 1998.

[171] Teweldemedhn Mekonnen and Leul Berhe. "Assessment on artificial insemination service delivery system, challenges and opportunities of artificial insemination services in cattle production in Western zone of Tigray Region, Ethiopia". In: *International Journal of Livestock Production* 11.4 (2020), pp. 135–145.

[172] P Melendez et al. "Uterine involution and fertility of Holstein cows subsequent to early postpartum PGF2α treatment for acute puerperal metritis". In: *Journal of dairy science* 87.10 (2004), pp. 3238–3246.

[173] A Meyling and A Mikél Jensen. "Transmission of bovine virus diarrhoea virus (BVDV) by artificial insemination (AI) with semen from a persistently-infected bull". In: *Veterinary Microbiology* 17.2 (1988), pp. 97–105.

[174] Celine Michaud, Daniel Llerena, and Iragael Joly. "Willingness to pay for environmental attributes of non-food agricultural products: a real choice experiment". In: *European Review of Agricultural Economics* 40.2 (2013), pp. 313–329.

[175] EHAB I Mohamed et al. "Electronic nose technology for the accurate detection of estrus by monitoring changes in perineal odors in dairy cows". In: *Journal of Biophysics and Biomedical Sciences* 2.1 (2009), pp. 128–133.

[176] Toby Mottram et al. "A wireless telemetric method of monitoring clinical acidosis in dairy cows". In: *computers and electronics in agriculture* 64.1 (2008), pp. 45–48.

[177] Shadreck K Mudziwepasi and Mfundo S Scott. "Assessment of a wireless sensor network based monitoring tool for zero effort technologies: a cattle-health and movement monitoring test case". In: *2014 IEEE 6th International Conference on Adaptive Science & Technology (ICAST)*. IEEE. 2014, pp. 1–6.

[178] Markus Nagel et al. "Data-free quantization through weight equalization and bias correction". In: *IEEE Conference on Computer Vision*. 2019, pp. 1325–1334.

[179] L Nagl et al. "Wearable sensor system for wireless state-of-health determination in cattle". In: *Proceedings of the 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (IEEE Cat. No. 03CH37439)*. Vol. 4. IEEE. 2003, pp. 3012–3015.

[180] Abozar Nasirahmadi et al. “Deep Learning and Machine Vision Approaches for Posture Detection of Individual Pigs”. In: *Sensors* 19.17 (2019). ISSN: 1424-8220. URL: https://www.mdpi.com/1424-8220/19/17/3738.

[181] Alam Noor et al. “Automated sheep facial expression classification using deep transfer learning”. In: *Computers and Electronics in Agriculture* 175 (2020), p. 105528.

[182] Chutchada Nusai, Wiruntita Chankeaw, and B Sangkaew. “Dairy cow-vet: A mobile expert system for disease diagnosis of dairy cow”. In: *2015 IEEE/SICE International Symposium on System Integration (SII)*. IEEE. 2015, pp. 690–695.

[183] Peter Leslie Nuthall. “Case studies of the interactions between farm profitability and the use of a farm computer”. In: *Computers and Electronics in Agriculture* 42.1 (2004), pp. 19–30.

[184] MICHAEL L. O’CONNOR. *Heat Detection and Timing of Insemination for Cattle*. PennState Extension, The Pennsylvania State University. 2016. URL: https://extension.psu.edu/heat-detection-and-timing-of-insemination-for-cattle.

[185] H Okada et al. “Development of ultra low power wireless sensor node with piezoelectric accelerometer for health monitoring”. In: *2013 Transducers & Eurosensors XXVII: The 17th International Conference on Solid-State Sensors, Actuators and Microsystems (TRANSDUCERS & EUROSENSORS XXVII)*. IEEE. 2013, pp. 26–29.

[186] Hironao Okada et al. “Wireless sensor system for detection of avian influenza outbreak farms at an early stage”. In: *SENSORS, 2009 IEEE*. IEEE. 2009, pp. 1374–1377.

[187] Geert Opsomer et al. “Risk factors for post partum ovarian dysfunction in high producing dairy cows in Belgium: a field study”. In: *Theriogenology* 53.4 (2000), pp. 841–857.

[188] Michael Overton and John Fetrow. “Economics of postpartum uterine health”. In: *Proc Dairy Cattle Reproduction Council* (2008), pp. 39–44.

[189] Kitsuchart Pasupa and Thanawat Lodkaew. “A new approach to automatic heat detection of cattle in video”. In: *International Conference on Neural Information Processing*. Springer. 2019, pp. 330–337.

[190] Wariston Fernando Pereira et al. “Environmental monitoring in a poultry farm using an instrument developed with the internet of things concept”. In: *Computers and Electronics in Agriculture* 170 (2020), p. 105257.

[191] Hieu Pham et al. “Efficient neural architecture search via parameters sharing”. In: *International conference on machine learning*. PMLR. 2018, pp. 4095–4104.

[192] Konstantin Pogorelov et al. “Efficient disease detection in gastrointestinal videos–global features versus neural networks”. In: *Multimedia Tools and Applications* 76.21 (2017), pp. 22493–22525.

[193] Konstantin Pogorelov et al. “Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection”. In: *Proceedings of the 8th ACM on Multimedia Systems Conference*. 2017, pp. 164–169.

[194] SMC Porto et al. “Localisation and identification performances of a real-time location system based on ultra wide band technology for monitoring and tracking dairy cow behaviour in a semi-open free-stall barn”. In: *Computers and Electronics in Agriculture* 108 (2014), pp. 221–229.

[195] Ashfaqur Rahman et al. “Cattle behaviour classification from collar, halter, and ear tag sensors”. In: *Information processing in agriculture* 5.1 (2018), pp. 124–133.

[196] PJ Rajala and YT Gröhn. “Effects of dystocia, retained placenta, and metritis on milk yield in dairy cows”. In: *Journal of dairy Science* 81.12 (1998), pp. 3172–3181.

[197] Joseph Redmon and Ali Farhadi. “YOLO9000: better, faster, stronger”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 7263–7271.

[198] Joseph Redmon and Ali Farhadi. “Yolov3: An incremental improvement”. In: *arXiv preprint arXiv:1804.02767* (2018).

[199] Joseph Redmon et al. “You only look once: Unified, real-time object detection”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 779–788.

[200] Shaoqing Ren et al. “Faster r-cnn: Towards real-time object detection with region proposal networks”. In: *Advances in neural information processing systems* 28 (2015).

[201] Hamid Rezatofighi et al. "Generalized intersection over union: A metric and a loss for bounding box regression". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 658–666.

[202] Martin Riekert et al. "Automatically detecting pig position and posture by 2D camera imaging and deep learning". In: *Computers and Electronics in Agriculture* 174 (2020), p. 105391.

[203] Jock McDonald Roberts. *Oestrus detector*. US Patent 9,913,703. 2018.

[204] Judith B Roelofs et al. "Pedometer readings for estrous detection and as predictor for time of ovulation in dairy cattle". In: *Theriogenology* 64.8 (2005), pp. 1690–1703.

[205] Agnese Rondoni, Daniele Asioli, and Elena Millan. "Consumer behaviour, perceptions, and preferences towards eggs: A review of the literature and discussion of industry implications". In: *Trends in Food Science & Technology* 106 (2020), pp. 391–401.

[206] Libin Rong and Daoliang Li. "A web based expert system for milch cow disease diagnosis system in China". In: *International Conference on Computer and Computing Technologies in Agriculture*. Springer. 2007, pp. 1441–1445.

[207] Wilbur E Rule and Earl D Smith. *Method and device for detecting period of heat in cows*. US Patent 3,076,431. 1963.

[208] Marie Saint-Dizier and S Chastant-Maillard. "Towards an automated detection of oestrus in dairy cattle". In: *Reproduction in domestic animals* 47.6 (2012), pp. 1056–1061.

[209] Minoru Sakaguchi et al. "Reliability of estrous detection in Holstein heifers using a radiotelemetric pedometer located on the neck or legs under different rearing conditions". In: *Journal of reproduction and development* (2007), pp. 0704050075–0704050075.

[210] Minoru SAKAGUCHI et al. "Reliability of Estrous Detection in Holstein Heifers Using a Radiotelemetric Pedometer Located on the Neck or Legs Under Different Rearing Conditions". In: *Journal of Reproduction and Development* 53.4 (2007), pp. 819–828. DOI: 10.1262/jrd.18099.

[211] Fatima M Salman and Samy S Abu-Naser. "Expert system for COVID-19 diagnosis". In: (2020).

[212] Gerard Salton and Michael J McGill. *Introduction to modern information retrieval*. mcgraw-hill, 1983.

[213] Lamis F Samhan, Amjad H Alfarra, and Samy S Abu-Naser. "Expert System for Knee Problems Diagnosis". In: (2021).

[214] Francis EP Sanderink et al. "Automatic detection of oestrus cows via breath sampling with an electronic nose: A pilot study". In: *Biosystems Engineering* 156 (2017), pp. 1–6.

[215] Heide Schatten and Gheorghe M Constantinescu. *Comparative reproductive biology*. Wiley Online Library, 2007.

[216] Wilfried Schneeweiss et al. "Endoscopic-assisted resection of a pedunculated uterine leiomyoma with maximal tissue preservation in a cow and a mare". In: *Veterinary surgery* 44.2 (2015), pp. 200–205.

[217] Santi Seguí et al. "Generic feature learning for wireless capsule endoscopy analysis". In: *Computers in biology and medicine* 79 (2016), pp. 163–172.

[218] Ramprasaath R Selvaraju et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 618–626.

[219] Md Sumon Shahriar et al. "Detecting heat events in dairy cows using accelerometers and unsupervised learning". In: *Computers and Electronics in Agriculture* 128 (2016), pp. 20–26.

[220] Johnathan Charles Sharpe et al. *Sensor apparatus and associated systems and methods*. US Patent 10,555,504. 2020.

[221] I Martin Sheldon et al. "Defining postpartum uterine disease and the mechanisms of infection and immunity in the female reproductive tract in cattle". In: *Biology of reproduction* 81.6 (2009), pp. 1025–1032.

[222] Weizheng Shen et al. "Design and implementation of livestock house environmental perception system based on wireless sensor networks". In: *International Journal of Smart Home* 10.5 (2016), pp. 69–78.

[223] Hoo-Chang Shin et al. "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning". In: *IEEE transactions on medical imaging* 35.5 (2016), pp. 1285–1298.

[224] Ali Shojaeipour et al. “Automated muzzle detection and biometric identification via few-shot deep transfer learning of mixed breed cattle”. In: *Agronomy* 11.11 (2021), p. 2365.

[225] Connor Shorten and Taghi M Khoshgoftaar. “A survey on image data augmentation for deep learning”. In: *Journal of Big Data* 6.1 (2019), pp. 1–48.

[226] Arne Sieber et al. “Wireless platform for monitoring of physiological parameters of cattle”. In: *Smart Sensing Technology for Agriculture and Environmental Monitoring*. Springer, 2012, pp. 135–156.

[227] Karen Simonyan and Andrew Zisserman. “Very deep convolutional networks for large-scale image recognition”. In: *arXiv preprint arXiv:1409.1556* (2014).

[228] Benjamin L Smarr et al. “Feasibility of continuous fever monitoring using wearable devices”. In: *Scientific reports* 10.1 (2020), pp. 1–11.

[229] Kevin Smith et al. “An integrated cattle health monitoring system”. In: *2006 International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE. 2006, pp. 4659–4662.

[230] Nitish Srivastava et al. “Dropout: a simple way to prevent neural networks from overfitting”. In: *The journal of machine learning research* 15.1 (2014), pp. 1929–1958.

[231] Stevan Stankovski et al. “Dairy cow monitoring by RFID”. In: *Scientia Agricola* 69.1 (2012), pp. 75–80.

[232] Suharjito et al. “Mobile Expert System Using Fuzzy Tsukamoto for Diagnosing Cattle Disease”. In: *Procedia Computer Science* 116 (2017). Discovery and innovation of computer science technology in artificial intelligence era: The 2nd International Conference on Computer Science and Computational Intelligence (ICCSCI 2017), pp. 27–36. ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2017.10.005. URL: https://www.sciencedirect.com/science/article/pii/S1877050917320422.

[233] Christian Szegedy et al. “Going deeper with convolutions”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 1–9.

[234] Christian Szegedy et al. “Inception-v4, inception-resnet and the impact of residual connections on learning”. In: *Thirty-first AAAI conference on artificial intelligence*. 2017.

[235] Christian Szegedy et al. “Rethinking the inception architecture for computer vision”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 2818–2826.

[236] Mingxing Tan and Quoc Le. “Efficientnet: Rethinking model scaling for convolutional neural networks”. In: *International conference on machine learning*. PMLR. 2019, pp. 6105–6114.

[237] M Thibier and H.-G Wagner. “World statistics for artificial insemination in cattle”. In: *Livestock Production Science* 74 (Mar. 2002), pp. 203–212. DOI: 10.1016/S0301-6226(01)00291-3.

[238] Ashwani Kumar Tiwari, Vivek Kanhangad, and Ram Bilas Pachori. “Histogram refinement for texture descriptor based image retrieval”. In: *Signal Processing: Image Communication* 53 (2017), pp. 73–85.

[239] Tran Viet Toan et al. “Cow estrus detection with low-frequency accelerometer sensor by unsupervised learning”. In: *Proceedings of the Tenth International Symposium on Information and Communication Technology*. 2019, pp. 342–349.

[240] Du-Ming Tsai and Ching-Ying Huang. “A motion and image analysis method for automatic detection of estrus and mating behavior in cattle”. In: *Computers and electronics in agriculture* 104 (2014), pp. 25–31.

[241] Yu-Chi Tsai et al. “Assessment of dairy cow heat stress by monitoring drinking behaviour using an embedded imaging system”. In: *biosystems engineering* 199 (2020), pp. 97–108.

[242] Tzutalin. *LabelImg*. Free Software: MIT License. 2015. URL: https://github.com/tzutalin/labelImg.

[243] FJCM Van Eerdenburg, HSH Loeffler, and JH Van Vliet. “Detection of oestrus in dairy cows: a new approach to an old problem”. In: *Veterinary Quarterly* 18.2 (1996), pp. 52–54.

[244] Ashish Vaswani et al. “Attention is all you need”. In: *Advances in neural information processing systems* 30 (2017).

[245] Stefano Viazzi et al. “Image feature extraction for classification of aggressive interactions among pigs”. In: *Computers and Electronics in Agriculture* 104 (2014), pp. 57–62.

[246] Paul Viola and Michael Jones. "Rapid object detection using a boosted cascade of simple features". In: *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001*. Vol. 1. Ieee. 2001, pp. I–I.

[247] S Waiblinger, C Menke, and G Coleman. "The relationship between attitudes, personal characteristics and behaviour of stockpeople and subsequent behaviour and production of dairy cows". In: *Applied Animal Behaviour Science* 79.3 (2002), pp. 195–219. ISSN: 0168-1591. DOI: https://doi.org/10.1016/S0168-1591(02)00155-7. URL: https://www.sciencedirect.com/science/article/pii/S0168159102001557.

[248] Juan Wang et al. "Real-time behavior detection and judgment of egg breeders based on YOLO v3". In: *Neural Computing and Applications* 32.10 (2020), pp. 5471–5481.

[249] Jun Wang et al. "IoT-based measurement system for classifying cow behavior from tri-axial accelerometer". In: *Ciência Rural* 49 (2019).

[250] Wei Wang et al. "High-Resolution Radar Target Recognition via Inception-Based VGG (IVGG) Networks". In: *Computational Intelligence and Neuroscience* 2020 (2020).

[251] Yiqi Wang et al. "Identifying lameness in horses through deep learning". In: *Proceedings of the 36th Annual ACM Symposium on Applied Computing*. 2021, pp. 976–985.

[252] Bartosz Wietrzyk and Milena Radenkovic. "Enabling large scale ad hoc animal welfare monitoring". In: *2009 Fifth International Conference on Wireless and Mobile Communications*. IEEE. 2009, pp. 401–409.

[253] Bartosz Wietrzyk and Milena Radenkovic. "Energy efficiency in the mobile ad hoc networking approach to monitoring farm animals". In: *Sixth International Conference on Networking (ICN'07)*. IEEE. 2007, pp. 1–1.

[254] Bartosz Wietrzyk, Milena Radenkovic, and Ivaylo Kostadinov. "Practical MANETs for pervasive cattle monitoring". In: *Seventh International Conference on Networking (icn 2008)*. IEEE. 2008, pp. 14–23.

[255] Dihua Wu et al. "Lameness detection of dairy cows based on the YOLOv3 deep learning algorithm and a relative step size characteristic vector". In: *Biosystems Engineering* 189 (2020), pp. 150–163.

[256] Dihua Wu et al. “Using a CNN-LSTM for basic behaviors detection of a single dairy cow in a complex environment”. In: *Computers and Electronics in Agriculture* 182 (2021), p. 106016.

[257] Tong Xing et al. “Stress effects on meat quality: A mechanistic perspective”. In: *Comprehensive Reviews in Food Science and Food Safety* 18.2 (2019), pp. 380–401.

[258] Beibei Xu et al. “Evaluation of Deep Learning for Automatic Multi-View Face Detection in Cattle”. In: *Agriculture* 11.11 (2021), p. 1062.

[259] William B Yancey, Linda C Frank, and Bruce E Johnson. *Systems and methods for detecting estrus*. US Patent 9,119,379. 2015.

[260] Aqing Yang et al. “An automatic recognition framework for sow daily behaviours based on motion and image analyses”. In: *Biosystems Engineering* 192 (2020), pp. 56–71.

[261] Aqing Yang et al. “Automated video analysis of sow nursing behavior based on fully convolutional network and oriented optical flow”. In: *Computers and Electronics in Agriculture* 167 (2019), p. 105048.

[262] Boh-Suk Yang et al. *Systems for identification and estrus state detecting in cattle*. US Patent 6,236,318. 2001.

[263] Hong-ying Yang et al. “Color image zero-watermarking based on fast quaternion generic polar complex exponential transform”. In: *Signal Processing: Image Communication* 82 (2020), p. 115747.

[264] Hongying Yang et al. “Robust and discriminative image representation: fractional-order Jacobi-Fourier moments”. In: *Pattern Recognition* 115 (2021), p. 107898.

[265] Linjie Yang et al. “A large-scale car dataset for fine-grained categorization and verification”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 3973–3981.

[266] Shuo Yang et al. “Wider face: A face detection benchmark”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 5525–5533.

[267] Hawar M Zebari, S Mark Rutter, and Emma CL Bleach. “Characterizing changes in activity and feeding behaviour of lactating dairy cows during behavioural and silent oestrus”. In: *Applied Animal Behaviour Science* 206 (2018), pp. 12–17.

[268] Lan Zhang et al. "Solid-state pH sensor prototype for real-time monitoring of the rumen pH value of Japanese cows". In: *Microsystem Technologies* 24.1 (2018), pp. 457–463.

[269] Lei Zhang et al. "Automatic individual pig detection and tracking in pig farms". In: *Sensors* 19.5 (2019), p. 1188.

[270] Yu Zhang et al. "Environment parameters control based on wireless sensor network in livestock buildings". In: *International Journal of Distributed Sensor Networks* 12.5 (2016), p. 9079748.

[271] Yuanqin Zhang et al. "Real-time sow behavior detection based on deep learning". In: *Computers and Electronics in Agriculture* 163 (2019), p. 104884.

[272] Yuanqin Zhang et al. "Real-time sow behavior detection based on deep learning". In: *Computers and Electronics in Agriculture* 163 (2019), p. 104884.

[273] K Zhao et al. "Automatic lameness detection in dairy cattle based on leg swing analysis with an image processing technique". In: *Computers and Electronics in Agriculture* 148 (2018), pp. 226–236.

[274] Chan Zheng et al. "Automatic posture change analysis of lactating sows by action localisation and tube optimisation from untrimmed depth videos". In: *Biosystems Engineering* 194 (2020), pp. 227–250.

[275] Chan Zheng et al. "Automatic recognition of lactating sow postures from depth images by deep learning detector". In: *Computers and electronics in agriculture* 147 (2018), pp. 51–63.

[276] Teng Zhou et al. "Quantitative analysis of patients with celiac disease by video capsule endoscopy: A deep learning method". In: *Computers in biology and medicine* 85 (2017), pp. 1–6.

[277] Xizhou Zhu et al. "Deformable DETR: Deformable Transformers for End-to-End Object Detection". In: *International Conference on Learning Representations*. 2020.

[278] Xunmu Zhu et al. "Automatic recognition of lactating sow postures by refined two-stream RGB-D faster R-CNN". In: *Biosystems Engineering* 189 (2020), pp. 116–132.

[279] Thi Thi Zin et al. "A general video surveillance framework for animal behavior analysis". In: *2016 Third International Conference on Computing Measurement Control and Sensor Network (CMCSN)*. IEEE. 2016, pp. 130–133.

[280] Thi Thi Zin et al. “Image technology based cow identification system using deep learning”. In: *Proceedings of the International MultiConference of Engineers and Computer Scientists*. Vol. 1. 2018, pp. 236–247.

[281] Barret Zoph et al. “Learning transferable architectures for scalable image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 8697–8710.