University of Modena and Reggio Emilia
XXXIII Cycle of the International Doctorate School in
Information and Communication Technologies

Doctor of Philosophy Dissertation in
Computer Engineering and Science

Deep Learning Techniques Applied to Farming

Luca Bergamini

Supervisor: Prof. Simone Calderara
PhD Course Coordinator: Prof. Sonia Bergamaschi

Modena, 2020

Review committee:

Bob Fisher
The University of Edinburgh (UK)

Giovanni Maria Farinella
Università degli Studi di Catania (IT)

Carmen;
even though you don't reckon it's true,
I'd be in much darker places without you

You can't get to the moon by climbing
successively taller trees.

Mo's Law of Evolutionary Development

Abstract

Although Deep Learning (DL) is increasingly being adopted in many sectors, farming is still an almost unscathed niche. This is mainly because of the humongous distance in knowledge between the experts of DL and those of farming itself. It's, first of all, a communication issue, and only in the second place a matter of reluctance to changes. Tackle those issues and you will find that also this sector can greatly benefit from the application of these new technologies. This thesis is a collection of applications of DL to different topics in farming. This has been made possible by the key role of figures who are placed right in the middle and act as intermediaries between the experts to identify targets and measures of success. In our case, this role is covered by Farm4Trade, a startup which is also the main funder of this PhD. The first covered topic is the automatic detection and tracking of pigs in a pen. The goal here is to detect and classify individual behaviours and how they change through time. A collection of state-of-the-art DL techniques has been chained together while each individual piece has been analysed on its own to ensure good final performance. Then, we jump at the end of the production chain to study how to apply DL to slaughtered pigs' carcasses images to detect and segment lungs lesions. These are reliable indicators of a bacterial pathology affecting the animal prior to its death. The second big topic is the automatic cattle re-identification from images and videos. We show how DL methods designed for humans can be adapted to work in a completely different setting. As a feedback loop, methods developed for cattle have been reapplied on people and vehicles with successful results. Results achieved during this PhD show how the whole sector of farming can benefit from the application of artificial intelligence algorithms.

Abstract (Italian)

Nonostante la massiccia diffusione in ormai ogni ambito di tecniche di Deep Learning, il settore dell'allevamento animale rimane una nicchia quasi inesplorata. La distanza siderale tra gli "esperti" di Deep Learning e quelli di questo settore rende quasi impossibile l'applicazione reale di algoritmi e tecnologie. Si tratta innanzitutto di un vero e proprio problema di comunicazione (più volte sperimentato in questa tesi) e solo in seconda istanza di resistenza al cambiamento. Superati questi ostacoli, appare chiaro come anche questo settore possa trarre enorme beneficio dall'applicazione di queste nuove tecnologie. Il presente lavoro è un tentativo di permeare il settore dell'allevamento animale (e dell'intera catena di produzione) da diverse angolazioni. Questo è stato possibile grazie alla fondamentale presenza di figure che si pongono tra i due ambiti (intelligenza artificiale e allevamento animale) e permettono una comunicazione efficace al fine di individuare obbiettivi e misure del successo. Nel caso di questa tesi, la start-up Farm4Trade ha rivestito questo arduo compito. La prima parte si concentra su detection e tracking di maiali da allevamento, con lo scopo di studiare i comportamenti individuali all'evolvere del tempo. Nella sezione successiva l'indagine si sposta alla fine della catena produttiva dove tecniche di segmentazione vengono applicate a carcasse di maiali da macellazione con l'intento di individuare lesioni sulla pleura polmonare. Queste sono infatti un indicatore di patologie batteriche contratte durante l'ultimo periodo di vita dell'animale. Il secondo argomento trattato riguarda invece il riconoscimento automatico dei bovini da immagini e video. Le tecniche sviluppate sono poi state applicate con successo ad altri domini (persone e veicoli), dimostrando come le conoscenze specifiche di un settore (seppur di nicchia) possano facilmente essere rese "globali". Concludendo, alla luce dei risultati ottenuti in questa tesi, emerge quanto l'intero settore dell'allevamento possa trarre largo beneficio dall'applicazione di algoritmi di intelligenza artificiale.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Deep Learning (DL) and Machine Learning (ML) are becoming ubiquitous in today's world and it's expected that they will pervade every aspect of our society in the upcoming years. Something that was until some years ago only academic research is now making its way into industry and society. Information and communication technologies have been the first sector to fully embrace them. There has been almost no resistance from the different stakeholders, as many of the applications they have made possible were almost non-existing prior to the rise of Deep Learning and Machine Learning. Face detection powered by machine learning is a clear example of such applications. Today, top smartphones can detect faces, people and other entities thanks to ML and DL techniques. Tomorrow, even the cheapest device will have enough computational power to run these and many more algorithms. Introducing this technology has not created any frictions as it was not available prior to ML.

Differently, other sectors have seen major social changes following the introduction of ML in the production chain. In the automation industry, for example, robots are becoming smarter and capable of performing jobs that we had assumed would have required a human operator only a couple of years ago. This opens the door to changes that have a much more transformative impact on the society itself. It has already been foreseen that many jobs will disappear in the upcoming years because they simply won't be required anymore (at least not when performed by humans). Bal-

ancing this tipping scale will be one of the biggest challenges for the current and future generations. Investing too much and too abruptly may lead to unemployment, dissatisfaction and even social unrest. Investing too little may preclude mankind incredible opportunities to progress.

Among all the different sectors, agriculture and farming have been left almost unscathed by this revolution as of today. This is not unexpected though, for several different reasons. First, these sectors are very far from immediate applications of DL. Second, there is even less connection between the researchers working on DL and the ones working on the field. This means that investors in ML often don't know or see the available opportunities. Last, there is undoubtedly a reluctance to embrace these new technologies, because workers from this sector have significantly less opportunity to reskill, and therefore face potential unemployment. In these 3 years of PhD, we've seen some examples of this reluctance, especially when working with technologies related to the food chain. Interestingly, it has surfaced more from higher-skilled professional (vets, food safety operators) than from those we often wrongly refer to as "low skilled" workers.

In this thesis, we analyse a few applications of ML and DL to farming and other related sectors. This was the primary interest of Farm4Trade, the start-up which funded the PhD. Although more work has been carried out outside this main thread, we have decided not to include it here, to keep the focus on a single area and present it organically (hopefully).

We start in Section 3.1 by investigating how recent improvements in detection and tracking can help to automatically study individual behaviours in pigs. This work has been carried in collaboration with the University of Edinburgh (UK) and it's of great interest to address the increasing request for animal welfare. Furthermore, performing detection, tracking and identification on livestock (and animals in general) has some key advantages when compared to humans, which will be discussed later.
In section 3.2 we jump to the end of the food chain to study how automatic segmentation can be applied to spot disease-borne lesions on pigs' carcasses. Food control and safety is a sector which could massively improve from using automatic systems, but which has almost been completely neglected for the reasons mentioned above. This work is a first proof that human performance can be obtained when enough data is available. An implementation of the

developed algorithm is currently being studied for future usage in a pig slaughterhouse.
Pigs are not the only livestock of interest of this work though. In Section 3.3 we study whether image-based re-identification (widely applied to human faces) can be applied to cattle. This idea was brought up as a challenge from the company and has grown to become a real project (and a future product). To the best of our knowledge, we were the first ones to present a working system on individual cows re-id from RGB images only using DL algorithms. This could enable a more precise identification (based on something the animal is and not something the animal has such as an ear tag) and has many applications in developing countries where the existing technologies struggle to penetrate effectively.
Finally, the experience on cattle has proven even more fruitful, as some of the approaches applied there have been successfully ported to completely different domains (i.e. people and cars). This emphasises how an approach from a different angle can sometimes lead to improvements in a well known and studied field. In Section 3.4 we illustrate some of these results which have been finalised in a recent publication at a top-tier conference.

Chapter 2

Literature Survey

In the following sections we briefly report a survey of the related literature. We start with a general overview of how DL has been applied to agriculture and farming in recent years. Then, we present other research approaches which are related to the topics tackled in this thesis. We acknowledge that the list could be much longer, but we choose to limit ourselves to the methods which are either most relevant for the community or more strictly related to the proposed algorithms.

2.1 Deep Learning in Agriculture and Farming

One of the biggest areas of application of DL in agriculture has been the so-called *precision agriculture*. This approach involves the constant observation, measurement and actuation of the variability in the crops. With the increasing availability of remote sensing and connectivity, DL can become an effective tool for this practice[53]. Convolutional Neural Networks (CNN) have been successfully employed to spot disease in leaves [131, 80], fruits[35] and vegetables[82]. This can significantly speed up diagnosis times and reduce the risks of marred crops.

They can also be applied to satellite images to automatically analyse, identify and map crops [160, 59, 110]. Organisations could use this technology to control deforestation and enforce crop rotation, which is essential to preserve field yields through time. Again, some works have shown that is feasible to spot weed infestations from satellite images [28, 31]. This can have an immense value for farmers, as it gives a snapshot of the current global situation, which can be used to tailor following interventions. Authors from[45] and [134] have shown how satellite images can also be used to estimate future yield and water request in vineyards.
CNNs can also be employed after the harvest. Among the possible applications fruit and seed counting and classification [56, 75, 157] can significantly speed up the time to market of these product.

Another area with a potential wide impact is the so-called *precision livestock farming*, which applies the same principles to the various types of livestock [53]. This includes monitoring of health indicators like comfort and aggressive behaviours, together with production indices.
In [39] a machine learning algorithm is employed to predict skin, core, and hair-coat temperatures of piglets. In [63] a cascade of CNNs predicts the pose of cows given monocular RGB images. Another area of great interest is the body condition score evaluation and forecasting. This directly correlates with the final yield of the animal and it's still performed manually today. In [51, 152] automatic systems are proposed to directly estimate this value from RGB images.

2.2 Pig Detection and Tracking

Pork is one of the most consumed meats in the world. Therefore, improving the animals' welfare is crucial to ensure the quality of the final product. We present here an overview of how DL has been applied to live pigs to study and monitor their behaviours. We purposely decide not to categorise related works by their specific processes (e.g. tracking) because often tasks are approached in a sequential fashion by the same work (e.g. both detection and tracking can be performed for individual pig behaviour analysis).
In [117] a TinyYOLO [104] architecture is employed to detect pigs from infrared camera. Much focus is placed on execution speed, as the target platform is an embedded device. Images are acquired from a single pen and

the training set includes 2904 images, while the test comprises 1000 images. The authors also approach the same task using traditional computer vision algorithms in [111]. They propose a method to detect pigs under various illumination conditions by combining information from depth and infrared images, using spatio-temporal interpolation.

Authors from [101] take another approach and cast detection as a segmentation task. The targets are not bounding boxes anymore but instead 4 semantic parts of the animal (ears, shoulder and tail) which are detected using a Fully Convolutional Network. The Hungarian algorithm is then employed to link those parts for each individual pig. A dataset with 2000 images from multiple pens is publicly available online. The authors extend their work in [128], where they focus on tracking by leveraging the fixed cardinality of the targets. Their tracker achieves real-time performance and is based on features extracted from a CNN. Also the dataset of this work is publicly available.

Similarly, in [18] the bounding boxes are replaced with ellipses, which are detected through a segmentation network. The intuition is that pigs are much closer to an ellipse in terms of shape when images are acquired from above. The dataset includes 1000 images recorded over a period of 20 days. 13 pigs from a single pen were recorded. An encoder-decoder architecture is trained with multiple losses to segment individual instances, using the notion of outer and inner edge of the animal. In [156] a Single Shot Detector [70] architecture is used to perform detection. A *tag-box* is then extracted from each detected animal to perform tracking using a variation of the MOSSE [15] tracking algorithm. The dataset includes multiple pens and has been acquired over a period of 3 days. In total, 18000 images have been collected and annotated for the training set and 4200 for the test set. Authors from [92] leverage the depth signal from a Microsoft Kinect to fit 3D ellipses in an unsupervised fashion. The pen boundaries need to be annotated only once to define the working area of the following algorithms. Information from normal surfaces is employed to detect the boundaries between the pigs when these are very close. The dataset includes 2.1M frames from 5 consecutive days of a pen with 15 pigs. In [24] detection, tracking and behaviour analysis of individual pigs is performed. First, R-CNN [37] is used to detect bounding boxes, that are then input into two real-time tracker algorithms. Transfer learning is required to accommodate the covariate shift from traditional deep learning dataset. Then, idle and moving behaviours are detected from tracklets.

The dataset includes 1646 annotated images, which are split with 0.5 ratio between training and test set. Authors from [60] focus their attention on the mounting behaviour only, which is identified as a cause of epidermal wounds and fractures. They collect a dataset from a week of acquisitions of a single pen with 4 young male pigs. 1500 frames are annotated with segmentation masks and a mounting/no-mounting behaviour flag. Then, a Mask R-CNN [43] is employed to detect and segment individual pigs. Finally, a multi-dimensional eigenvector is computed from the detected bounding-box and segmentation and classified into the two possible behaviours. Differently, in [155, 61] the behaviour analysis is rephrased as an end-to-end video classification task. A dataset (PBVD-5) of 1000 short clips is collected and annotated with one out of five different behaviours (feeding, lying, walking, scratching and mounting), with 200 videos for each behaviour. Data comes from 4 pens with up to 3 pigs in each. Then, in [155] a two streams architecture employs both RGB and optical flow information to classify snippets and individual frames, and the results are fused using a consensus function. The authors compare the performance of various architecture, including ResNet [44] and Inception [127] networks, as backbones.

2.3 Animal Re-identification

Nowadays, little efforts have been made in the animal Re-identification task, with the noticeable exception of apes, since they share common traits with human beings. [25] achieved 98.7% on facial images coming from 100 red-bellied lemurs (Eulemur rubriventer) of the Ranomafana National Park, Madagascar. The authors employed multi-local binary patterns histograms (MLBPH) and Linear Discrimant Analysis (LDA) on cropped and aligned faces. The authors show how human face recognition algorithms can be adapted to primates, as their faces share the same underlying features. Following their work, [26] expanded the re-identification task to multiple apes species using DCNN. On one hand, the authors tried some traditional baseline such as EigenFaces and LBPH, on the other hand, they exploited two state-of-the art human faces identification networks, namely FaceNet and SphereFace. Using the latter, the authors showed how a CNN trained for solving a human re-identification task may also be adapted to the

primates one, leveraging a fine-tuning strategy. However, they achieved slightly superior performance by means of a smaller CNN (in terms of number of layers and parameters), which is trained from scratch on apes' faces from three different species. In both above mentioned approaches, the underlying assumption lies on the presence of some similarity between the human and the ape faces. This evidence has been corroborated by two surprising results: firstly, the network trained on human faces performed enough well also on apes (showing comparable results with a network trained from scratch on apes [26]) and secondly, it is able to extract and work with facial landmarks.
Other endangered species have also attracted interest in recent years, from zebras [58] to tigers [47].[96] proposed a deep learning technique aiming to automatically identify different wildlife species, as well as counting the occurrences of each species in the image. Differently from us, such methods work on images depicting the entire animal's body, exploiting the characteristic stripe patterns of such animals.
Finally, pets represent an opportunity to build a larger dataset, as they outnumber the above mentioned animals of a large margin, but collecting this data requires huge resources and efforts. As an example, using pictures of two dogs breeds gathered from Flickr, [93] achieved remarkable performances. As dog faces differ from humans ones, the authors developed two Deep CNN trained from scratch on dogs images only, after a pre-processing phase consisting of a tight crop to suppress most of the background.

2.3.1 Cattle Re-identification

Due to their economic value, the literature regarding cattle-identification methods is slowly increasing in recent years. However, to the best of our knowledge, this is the first attempt in doing it using the animal face with Deep Learning techniques.
[5] employed images from Unmanned Aerial Vehicle (UAV) to identify cattle of a single breed using individual stripes and patches. Firstly, the authors gathered a dataset of 89 different cows depicted in 980 RGB images, being captured by a camera placed over the walkway between holding pens and milking stations of a farm. Secondly, the authors presented a CNN trained from images, as well as a complete pipeline involving a Long-Short Term Memory (LSTM) layer to exploit temporal information. They achieved 86.07% identification accuracy on a random train-test split and 99.3%

detection and localisation accuracy.

Similarly, [164] developed a system based on histograms and movements to record images of the backs of 45 cows from a camera placed on the Rotary Milking Parlour, for a period of 22 days. The authors trained a DCNN to perform individual identification, achieving an outstanding 98.97% of accuracy. The collection system was able to correctly detect the back of the cow and crop it from the image. Using this approach, the system was able to record a huge amount of data with a great variation of light condition.

2.4 Human Re-identification

Human Re-identification has a long history of both research and practical uses. Among early methods, EigenFaces [136] has proved to perform well on cropped and aligned faces, such as the Olivetti dataset, where it achieves a re-identification accuracy of 95%. FisherFaces [11] employed a classifier based on the Linear Discriminant Analysis (LDA), exploiting features coming from a preprocessing phase involving a Principal Component Analysis (PCA) stage. In this way, it merges information from multiple views of the same subject in the final classifier. However, both these methods are unable to deal with unaligned faces and suffer from illumination changes.

Since they are widely known for extracting more invariant features, Local Binary Pattern Histograms (LBPH) [97], Histogram of Oriented Gradients (HOG) and Scale Invariant Features Transform (SIFT) [79] descriptors have been widely used. As some of the more extend variations usually occur in illumination changes and scale, these descriptors have been designed to be robust for these applications. However, the human face has more than 15 muscles producing some of the most complex expression in nature, which can alter dramatically the final appearance of a person face. In recent years, DCNN trained on huge datasets, such as [77],[42] and [50] have provided features learned directly from examples with a growing interest from the computer vision and machine learning communities. Among them, [114] and [71] show state-of-the-art performances, as they can handle different facial expressions as well as face aging. Even with some differences, both architectures produce a low-dimensional feature vector, namely an embedding of the input face, efficient to be compared with others using Nearest Neighbours classifiers.

2.4.1 Image-To-Video Re-Identification

The I2V(Image to Video) Re-ID task has been successfully applied to multiple domains. In person Re-ID, [144] frames it as a point-to-set task, where image and video domains are aligned using a single deep network. The authors of [154] exploit time information by aggregating frame features via a Long-Short Term Memory. Eventually, a dedicated sub-network aggregates video features and matches them against single image query ones. Authors of MGAT [10] employ a Graph Neural Network to model relationships between samples from different identities, thus enforcing similarity in the feature space. Dealing with vehicle Re-ID, authors from [74] introduce a large-scale dataset (VeRi-776) and propose PROVID and PROVID-BOT, which combine appearance and number plate information in a progressive fashion. Differently, RAM [72] exploits multiple branches to extract global and local features, imposing a separate supervision on each branch and devising an additional one to predict vehicle attributes. VAMI [163] employs a viewpoint aware attention model to select core regions for different viewpoints. At inference time, they obtain a multiview descriptor through a conditional generative network, inferring information regarding the unobserved viewpoints. Differently, our approach asks the student network to do it implicitly and in a lightweight fashion, thus avoiding the need for additional modules. Similarly to VAMI, [21] predicts the vehicle viewpoint along with appearance features; at inference, the framework provides distances according to the predicted viewpoint.

2.4.2 Knowledge Distillation

Knowledge distillation has been first investigated in [108, 48, 153] for model compression: the idea is to instruct a lightweight model (student) to mimic the capabilities of a deeper one (teacher): as a benefit, one could achieve both an acceleration in inference time as well as a reduction in memory consumption, without experiencing a large drop in performance. In this work, we benefit from the techniques proposed in [48, 135] for a different purpose: we are not primarily engaged in educating a lightweight module, but on improving the original model itself. In this framework – often called *self-distillation* [36, 150] – the transfer occurs from the teacher to a student with the same architecture, with the aim of improving the overall performance at the end of the training. Here, we get a step ahead and

introduce an asymmetry between the teacher and student, which has access to fewer frames. In this respect, our work closely relates to what [13] devises for Video Classification. Besides facing another task, a key difference is: while [13] limits the transfer along the temporal axis, our proposal advocates for distilling many views into fewer ones. On this latter point, we shall show that the teaching signal can be further enhanced when opening to diverse camera viewpoints. In the Re-Identification field, Temporal Knowledge Propagation (TKP) [41] similarly exploits intra-tracklet information to encourage the image-level representations to approach the video-level ones. In contrast with TKP: *i)* we do not rely on matching internal representations but instead their distances solely, thus making our proposal viable for cross-architecture transfer too; *ii)* at inference time, we make use of a single shared network to deal with both image and video domains, thus halving the number of parameters; *iii)* during transfer, we benefit from a larger visual variety, emerging from several viewpoints.

Chapter 3

Deep Learning and Farming

In this chapter, we present the core of the research. We would like to spend a few lines here to talk about two features shared by the vast majority of the work carried out. The first one is the almost complete absence of prior available data. This is not surprising given the traditional lack of interest from researchers towards this area. To address it, we carried out multiple campaigns of data collection and selection. While this always entails a delay to reach the desired target (often by several months), the ownership of such datasets is crucial, especially from a business point of view. When dealing with data-hungry algorithms (as those in DL) data is the real value, often way more than the algorithms themselves. As this PhD is funded by a private company, collecting this data has a potential return of investment. From this optic, data should be kept private and not released to the public. On the other hand, from a research point of view, we would like data to be as open as possible to foster further research on the topic. Meeting these two conditions together is sometimes challenging, but often a good compromise can be found.

We collected a total of three datasets during this PhD. The first one is a collection of cattle images and videos with associated identities. It was collected in multiple locations in Italy and abroad by different operators. This dataset is available to researchers upon request. The second one is a collection of images of pigs' carcasses annotated with pixel-wise segmentation for 7 different classes. We released a subset of it publicly, while the

rest of the dataset is available upon request. The third one is a collection of videos of a group of pigs roaming a pen of the University of Edinburgh. A portion of the dataset has been annotated with detection, tracking and behaviours. This dataset is fully open-sourced. As it can be seen, there is no one size fits all solution, but more likely a set of tools which can be used depending on the situation and the intrinsic value of the data. This last point must be always taken into account to deliver the maximum research impact while balancing the economic value at the same time. More details about the three datasets are included in the next chapters.

The second common feature is the absence of privacy concerns in these applications and how this differs from their human counterparts. In fact, in this work, we discuss detection, tracking and even identification at the individual level. The same applications would pose critical privacy concerns when applied to humans and would require extensive paperwork before being deployed. When working with livestock instead, these issues don't exist. This may not seem a substantial advantage at first, but it critically reduces the time to the marker of any developed system, as well as the regulations required around the acquired data. We strongly believe this difference can be a key to quickly scale and deploy these systems to national and international levels. As an example, the collection of our cattle dataset didn't require any specific permission concerning the privacy of the animals.

We are not suggesting that ML application to livestock are free from any regulations. Indeed, other regulations must be strictly followed. In particular, animal welfare is always a key requirement for all applications and researches in this sector. Even the collection of an image dataset usually has to be reviewed by a designated committee, as in the case of our pig detection and tracking dataset.

The rest of the chapter includes the individual contribution of this thesis. In section 3.1 the detection and tracking of pigs is presented. The analysis of lesions on pigs' carcasses follows in section 3.2. Then, the re-identification of cows and people are presented in sections 3.3 and 3.4 respectively.

3.1 Extracting Accurate Long-Term Behaviour Changes from A Large Pig Dataset

3.1.1 Author's main contributions

The main author's contributions for this research are:

- the design and installation the acquisition system, which was indispensable to collect the data required for the project;
- helping to develop the annotation tools and taking part into the data annotation process;
- the writing of most of the publication text and the literature survey;
- the temporal behaviours analysis

3.1.2 Introduction

Pork is the second most consumed meat[132] across the world behind poultry, and more than 700 million [118] pigs were raised in 2019 alone.

Modern intensive pig farming is highly mechanised, with automation of the environmental temperature and airflow, supply of feed, water and the removal of wastes. Driven by efficiencies of scale, farms have also grown larger, and there has been a reduction in staff time per pig [126]. As an example, in the EU more than half of the pork production comes from large intensive farms [100].

Behaviour analysis could be used by farm staff, vets and scientists to reveal the pigs' state of health and welfare, but on most farms, a typical weaner-grower-finishing pig may only be briefly inspected once or twice a day as part of a large group. There is an increasing interest in using automated methods to monitor pigs' behaviour on farm settings [148, 94]. Aspects of behaviour such as gait, use of different areas and resources in the pen, social clustering, activity can all be valuable information. Changes in behaviour from the expected norm can be used as an early warning sign for behaviour problems such as tail biting [29], social aggression [19], diseases [32], or for production issues such as thermal comfort [23]. The use of cameras and various other sensor technologies in animal agriculture – to gather useful real-time data to guide management decisions – is often referred to as 'precision livestock farming' [143].

In this work, we present a behaviour analysis pipeline built on automatic pig detection and tracking, capable of providing a report of the changes in a set of 5 fundamental individual behaviours (lying, moving, eating, drinking and standing) through time.

While the same topics are of great interest in the scientific community when applied to humans ([124, 125, 139, 33, 6] among many others), less research exists addressing the same tasks in the animal domain. This may sound counter-intuitive at first, given that for some applications, like identification and tracking, working with animals completely lifts any privacy and security concerns. However, there is often a wide gap between the expertise of people working on the techniques (computer vision and machine learning scientists mainly) and those working directly with livestock (veterinary and biology researchers).

Recently, thanks to the democratisation of computer vision and deep learning, numerous works have been presented for livestock and wildlife detection [123, 96, 117, 111, 101], tracking [138, 156, 92], identification [69, 67] and also behaviour analysis [133, 24, 60, 155, 61]. Although techniques are now available, the increasing usage of deep convolutional neural networks has seen the demand for high quality annotated data soaring.

To this end, a contribution of this work is also an unrestricted public pigs dataset, providing both manual and automatic annotation for multiple tasks, including detection, tracking, identification and behaviour analysis.

In summary, the main contributions of our work are:

- A behaviour analysis pipeline that focuses on individual pig behaviours to infer statistics about 5 different individual behaviours and how these change through time;
- Evidence that the behaviour statistics at the aggregated week level are reliable and robust to error in the various steps of the pipeline;
- A public available dataset comprising 7200 fully annotated frames.

3.1.3 Dataset

The dataset was collected between 5 Nov and 11 Dec (2019, 6 weeks) in a single pigpen (5.8m x 1.9m) with 8 growing pigs at SRUC (Scotland's Rural College)'s research pig unit (near Edinburgh, UK). The pigs were mixed

Table 3.1: A comparison of datasets on pigs

Paper	# Frames	# Annotated frames	Annotation types	# Pens	# Acquisition time	# Pigs	Publicly available
[117],[111]	-	3,904	boxes	1	1 day	9	✗
[18]	-	1,000	ellipses	2	4 months	-	✗
[156]	-	22,200	boxes, IDs	1	3 days	9	✗
[92]	2,100,000	-	pen boundaries, feeder, waterer	1	5 days	15	✗
[24]	-	1,646	boxes, IDs	1	-	20	✗
[60]	-	1,500	pigs' contours	1	7 days	4	✗
[155, 61]	156,900	1000 (videos)	5 behaviours	3	80 days	9	✗
[101]	2,000	2,000	4 body parts locations	17	multiple weeks	variable	✔
[128]	135,000	135,000	3 body parts locations, IDs	9	multiple weeks	7-16	✔
Ours	3,429,000	7,200	boxes, IDs, 5 behaviors	1	23 days over 6 weeks	8	✔

intact males and females weighing around 30kg at the start of the study. They were given a 3-space feeder with ad libitum commercial pig feed, two nipple water drinkers and a plastic enrichment device (Porcichew, East Riding Farm Services Ltd, Yorkshire, UK) suspended at pig height from a chain. Pigs were also given straw and shredded paper on a part-slatted floor. Colour image and depth data was collected using an Intel RealSense D435i camera positioned at 2.5 meters from the ground. Both RGB and depth information were acquired at 6fps with a resolution of 1280×720, and the acquisition was limited to daytime (from 7AM to 7PM), due to the absence of artificial light during nighttime.

Figure 3.1: An example of RGB and depth data for the same frame. The depth data presents several artefacts. One of the pigs in front of the feeder has a wide spot with value zero, while one in the rear has both zero and out of distribution (white patch) areas.

The acquired frames were appended into video sequences of fixed size (1800 frames each corresponding to 5 minutes) for both compression efficiency and logical organisation of the data. Fig. 3.1 shows an example of RGB and depth information for the same frame. It is worth noting how the

depth signal proved to be almost completely unreliable due to the presence of heavy non-white noise. Using it as an additional signal in our algorithms not only did not increase performance, but it even hinders it in several trials.

We acquired a total of 3,429,000 frames. Together with the raw data, we also provide manual annotations for different tasks for a subset (12 sequence corresponding to 7200 frames spread over the 6 weeks) of the dataset. These annotations were manually generated by 4 different people using a custom version of VaticJS [14] available at `https://stefanopini.github.io/vatic.js`. In each frame, the annotator:

- Draw a rectangular bounding box around each visible pig;
- Associate each bounding box with one of the 8 pigs using a numeric identifier;
- Select a behaviour among a list of 5 options (lie, move, eat, drink and stand).

The 12 sequences were annotated and split between training and validation to cover the entire time window of the acquisition process. This guarantees that the quality of the supervised algorithms employed in the rest of this work is representative of the full dataset.

Table 3.1 reports statistics for our dataset and compares it with others already published by the scientific community (both publicly and not). Although a bigger dataset [128] is publicly available, it only includes 3 key-points and IDs annotations. Contrarily, ours provides annotations for detection, tracking and behaviour analysis.

3.1.4 Behaviour Analysis Pipeline

Although the main focus of this work is understanding individual pig behaviours, several steps are required to fill the gap between raw data and behaviours. First, pigs need to be individually detected in each frame. This information alone is already enough to identify behaviours that do not require temporal knowledge, such as eating or drinking. In fact, these activities are constrained to very specific locations in the pen where the troughs are. However, as other behaviours require multiple detections of the same pig in consecutive frames (e.g. moving or standing), we use tracking to associate the bounding boxes from consecutive frames into tracklets. A

summary of the employed techniques for detection and tracking is given before focusing on behaviour analysis. For both tasks we report supervised metrics on the annotated evaluation set.

Table 3.2: Metrics from the detector on the validation set. We report results for individual sequences as well as those from the whole validation set

Val Sequence	**AP(%)**	**TP(%)**	**FP(%)**	**Missed(%)**
A	84.63%	89.18%	10.82%	1.16%
B	97.28%	99.59%	0.41%	2.24%
C	100.00%	100.00%	0.00%	0.00%
D	95.75%	96.97%	3.03%	0.85%
E	98.38%	99.45%	0.55%	1.03%
Whole set	95.21%	97.04%	2.96%	1.06%

Detection

Pig detection is treated as a supervised computer vision task, powered by the ground truth annotations. A state-of-the-art deep convolutional neural network is used for multiple object detection, namely YOLO v3 [105]. We pre-train it on the ImageNet dataset [27] and fine tune it for pig detection by replacing the classification grid layer to predict only 2 classes (background and pig). Because the original dataset contains chiefly portrait pictures, we replaced the network anchors with a new set computed on our training set bounding boxes. Furthermore, since the camera depicts also parts of other pigs' pens, we apply a mask on the video frames with the shape of the pen area containing the 8 pigs that we want to track. We set a threshold on the network's confidence scores and we also apply non-maximum suppression using a threshold on the IoU between predicted boxes. We experimented with those two hyper-parameters but found that the default values (0.9 and 0.4 respectively) in practice worked well for the task. However, we include the a-priori knowledge of having a limited known number of entities we want to detect. As such, we always take up to 8 bounding boxes.

Table 3.2 shows results in terms of Average Precision (AP), percentage of true positives (TP), false positives (FP) and missed detections on the

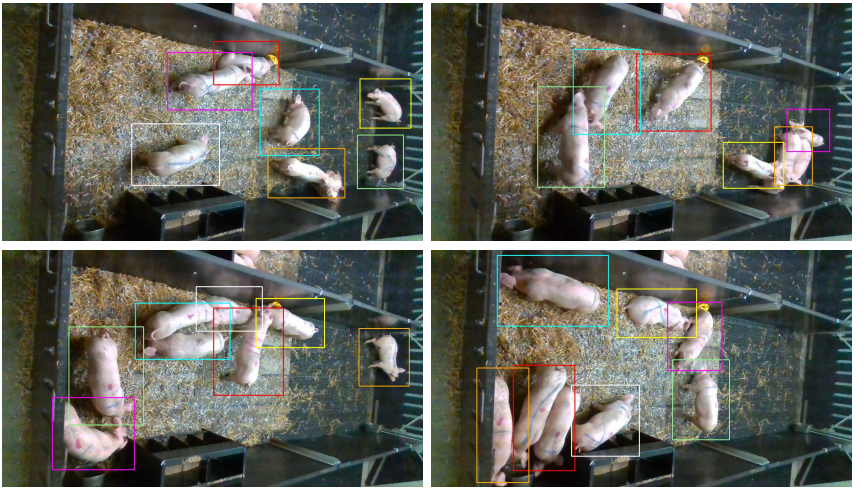


Figure 3.2: Examples of detection failures. A bounding box contains more than a single pig when the pigs are too close (e.g. red bounding box in bottom-right figure). Moreover, even when two separate bounding boxes are successfully generated for close pigs, they sometimes include portions of the other animal (e.g cyan bounding box in top-right figure)

validation set. We report statistics for the individual sequences and the average on the full validation set.

The reported detection metrics are satisfactory. Fig. 3.2 shows some detection failure cases from the validation set. Failures are mainly due to two reasons. First, differently from humans, pigs stay extremely close together most of the time, either while sleeping (usually on top of each other), fighting or just standing. In these conditions, it becomes extremely likely to have more than one pig per cell in the classification layer. Second, bounding box annotations become less reliable when pigs cannot be contained individually by a rectangle.

These two factors pose a great challenge to algorithms designed for detecting humans. While other works use better fitting annotations that partially solve these issues (like ellipses in [18]), these require custom algorithms to be handled and are more expensive to annotate compared to bounding boxes.

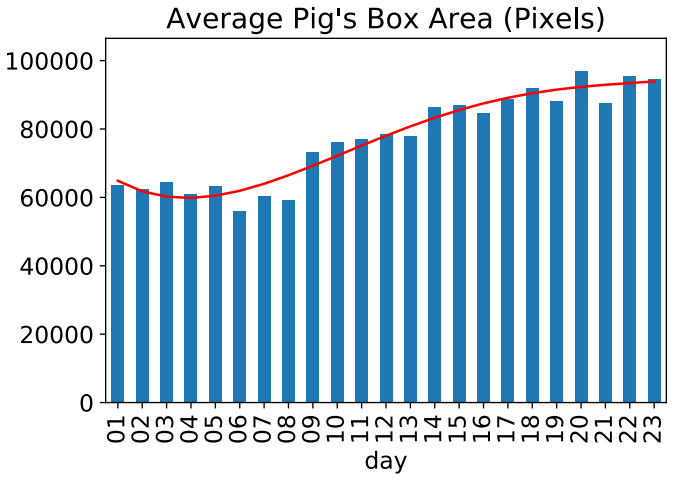


Figure 3.3: Changes of the mean predicted bounding box area on the full dataset through the acquisition days.

Although detection of single instances may be noisy, the amount of available data greatly reduces the noise influence. For example, we report in Fig. 3.3 the detected bounding box area on the full unlabeled dataset averaged by day and pig. It can be observed how the area increases monotonically (by around 45% throughout the entire acquisition window) which is expected when ad libitum food is available and only reduced activity can be performed due to space constraints. The 45% correlates well with the predicted increase in observed area of 67% based on $(W_{after} = 65kgs/W_{before} = 30kgs)^{2/3}$, assuming that weight is proportional to volume.

Tracking

For tracking the pigs, we employ a simple yet effective tracking-by-detection algorithm [99] that groups in tracklets consecutive detections of the same pig. In practice, for each new detection a new tracker is created and initialized. In the following frames, updated trackers and single-frame detections are matched together comparing the Intersection over Union and

Table 3.3: Metrics from the tracker on the validation set. We report results for individual sequences as well as those from the whole validation set

Val Sequence	MOTA (%)	IDF_1 (%)	# Switches	# Fragmentations	# Tracklets	Avg. tracklet length
A	76.78%	55.10%	23	187	24	597
B	97.35%	88.39%	12	13	17	834
C	100%	100.00%	0	0	8	1800
D	92.97%	88.46%	9	43	24	597
E	97.92%	78.29%	12	18	13	1104
Whole set	93.00%	82.05%	11.2	52.2	17.2	986.4

their appearance and finding the best assignments with the Kuhn-Munkres algorithm [57]. If a detection is not matched to any tracker, a new one is initialized while, if a tracker is not matched to a detection for 8 frames, the tracker is removed. As tracker, we employ the MOSSE [15] algorithm.

We evaluate the quality of the tracking algorithm using the following metrics:

- Multiple Object Tracking Accuracy ($MOTA$) [12, 90] combining three sources of errors as:

$$MOTA = 1 - \frac{\sum_t (FN_t + FP_t + IDSW_t)}{\sum_t GT_t} \tag{3.1}$$

  where FN is a tracker hypothesis which is wrongly not generated, FP is a tracker hypothesis generated where there is none in the ground truth, $IDSW$ is a mismatch between the current and previous association and GT is the number of ground truth objects;

- Identification F_1 (IDF_1) score [106] representing the ratio of correctly identified detections over the average number of ground truth and computed detections:

$$IDF_1 = \frac{2IDTP}{2IDTP + IDFP + IDFN} \tag{3.2}$$

  which differs from the $MOTA$ as it performs a 1-to-1 mapping between IDs, without considering identity switches or fragmentation of the predicted trajectories;

- Number of identity switches, occurring when the tracker jump from one identity to another;

- Number of fragmentations, accounting for tracklet switches between missed and not missed states.
- Average tracklet length, which summarises the tracker effectiveness in following the pigs through the sequence (a perfect result would be 8 tracks, each with 1800 frames).

Table 3.3 reports the results on the validation set. While there is some variance between sequences, most of the pigs are tracked for long periods, they are rarely swapped and few false positives occur. In particular, the average tracklet length is more than half a sequence (i.e. more than 2.5 minutes) and the per-sequence number of switches between two pigs is only 11 (i.e. on average, each pig track switches about 1.5 times).

Behaviour Analysis

Behaviour analysis uses the detections and tracklets to predict a behaviour class for each pig in every frame. While it is possible to directly predict the behaviour along with the pig detection, a single-frame approach would struggle to correctly identify behaviours that depend on multiple frames, such as moving or standing. Here, a combination of deep learning based and traditional techniques is used to better fits the different natures of the behaviours of interest.

The first step computes the average movement of the pig, as the movement of bounding box centroid locations in a given time-frame. The average depth inside the bounding box is used to predict the average pig movement in centimeters and compared to a fixed threshold in the same unit. In this way a single threshold can be applied to pigs in any part of the pen. We use a threshold of 2.5 cm over the centre of mass movement in a 2 seconds window and show that it is enough to sufficiently discriminate between unmoving and moving behaviours.

When the initial decision yields unmoving, we identify whether a pig is feeding or drinking by its distance and orientation from the feeder or the drinkers. Because we have collected our data from a single pen, the positions of those items is known. However, because the annotated bounding boxes don't hold any orientation information, identifying the pig orientation is not trivial. In practice, we compute the gray-scale image moments [91] on each bounding box and extract the pig center of mass and angle from a combination of the first and second order central moments, under the

hypothesis of having ellipse-shaped entities, which is a good assumption for pigs [18]. It is worth noting how this approach cannot disambiguate between 2 angles spanned by π like a pig facing or giving its back to the feeder. In practice, we notice the latter happens very rarely, as other pigs are likely to step in to feed frequently.

The remaining behaviours consist of lying and standing. These actions do not depend on specific locations in the pen, and the appearance of the pig must be taken into consideration for choosing between them. Our first approach made use of the depth information but proved unreliable. Therefore, a deep-learning method based on ResNet18 [44] is used to classify the bounding box into one of the classes of interest. The network is trained on the training split and validated on the validation split. We compensate for class imbalance by inverse weighting during training (i.e. samples from the most common classes are weighted less than samples from the uncommon classes).

We report results in terms of accuracy on the validation set for the five behaviours in Fig. 3.4 (left). It is likely that more sophisticated, generic, and accurate behaviour classification methods exist, but we reiterate one claim of the work: the collective behaviour statistics are accurate, even though individual frame-level labels may not always be as accurate (about 73% accurate on average over the unbalanced validation set, of which about 75% of the frames were either standing or lying). To support this claim, we report in Fig. 3.4 (right) two distribution histograms (for the ground-truth and the predicted behaviours, again on the validation set). It can be observed that these two distributions are very similar, with a KL divergence value of 0.014. As another measure of quality, we compute the average global prediction error $\sum_i \mid GT_i - Pred_i \mid / \sum_i GT_i = 0.14$, which shows that the individual errors tend to cancel out to give more accurate collective statistics.

3.1.5 Discussion

We report in this section our observations on the full unlabeled dataset, after applying our pipeline for behaviour understanding. We aggregate the results for the predicted behaviours by days (Fig. 3.5) and by daytime hours (Fig. 3.6). The statistics are computed over approximately 27 million pig detections, which means approximately 1 million detections per day in Fig. 3.5 and 2 million detections per time of day in Fig. 3.6. From the

former we draw the following conclusions:

- **Eating** and **drinking** behaviours do not vary drastically during the observation period. This matches the ad libitum availability of food and water that the animals were provided with. These two actions are performed for a total of around 10% of the whole time. For drinking, the indoor ventilated setting reduces the need of water;

- **Moving**, **standing** and **lying** follow a mirrored pattern. While the first two decrease through time, the latter drastically increases. This matches the expected behavioural pattern of growing pigs in a new environment. The first days are characterised by high levels of activity. This is due to various factors, including the pigs' youth, being in a new environment, the presence of other pigs and not being used to daily inspections among others. After these first days, they rapidly adapt to the new situation while at the same time they begin to grow more quickly (see Fig. 3.3). This eventually results in pigs spending most of the time lying and/or sleeping.

On the other hand, the analysis over the daytime highlights how pigs in these conditions (indoor, artificial light only) are mainly diurnal, where activities (moving, eating and drinking) are performed intensively during the morning and early afternoon and the animals are less active during the late afternoon.

We also visualise the heatmaps for 3 of the 5 behaviours in Fig. 3.7. These are computed by plotting the centroid of the detected bounding boxes for a single behaviour over the 6 weeks period. It is worth noting:

- **Lying** rarely occurs in front of the feeder. This is because pigs keep alternating to eat, making the area crowded. The areas where lying occurs more often are in fact those along the edges, but not at the very far right end where much toilet behaviour occurs due to the slated floor.

- **Standing** is more spread around, with a preference for the left part of the pen;

- **Moving** is focused in two areas mainly. The first one is the right section of the pen. This is where pigs usually run when operators move along the aisle near the left edge of the pen and it's also the

area deemed as toilet. The second area is in front of the feeder, as pigs move here to access and leave the feeder itself.

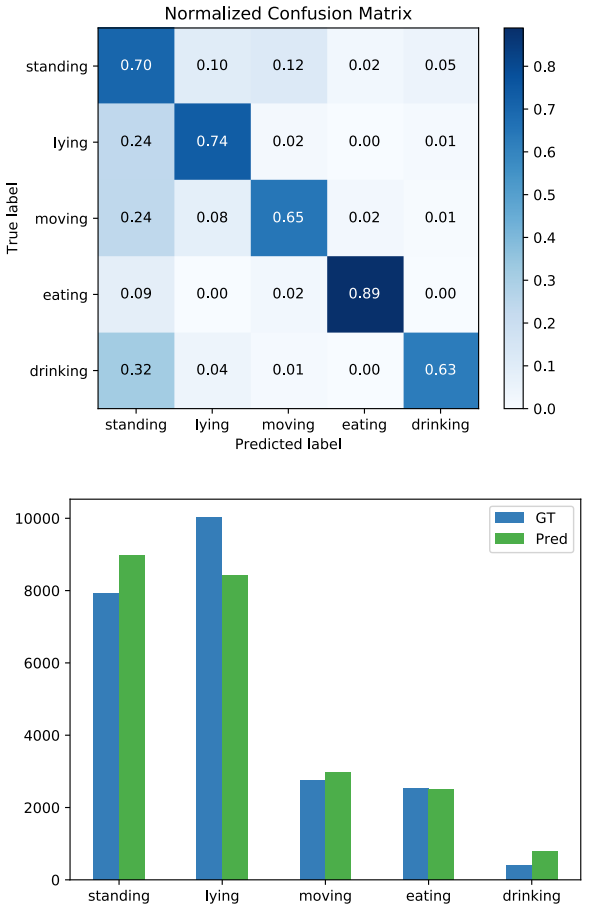


Figure 3.4: Confusion matrix for the 5 behaviours on the validation set (left). Distributions of the GT and predicted behaviours on the validation set (right)

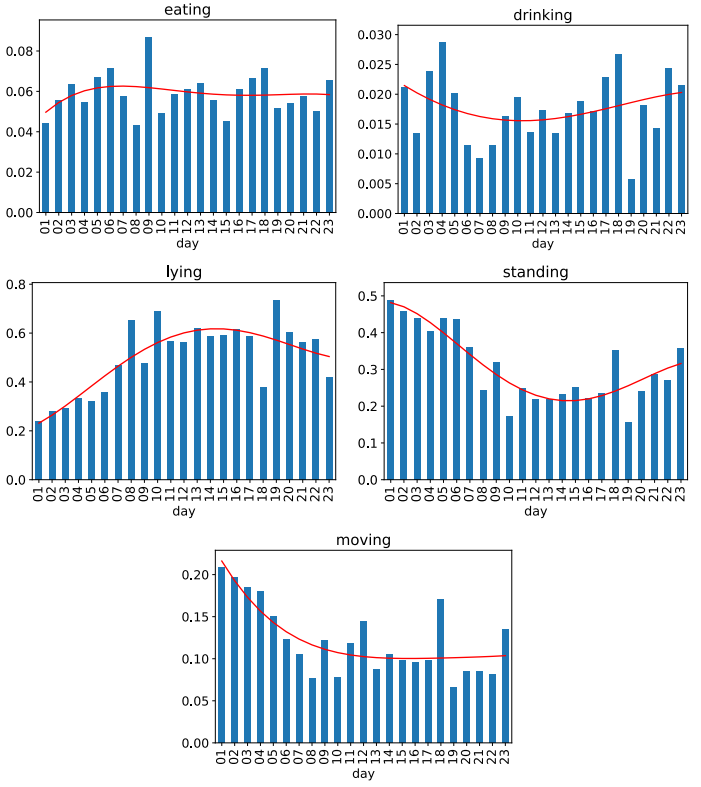


Figure 3.5: Temporal graphs of the behaviour changes aggregated by day on the full dataset. The red solid line shows an interpolation of the data

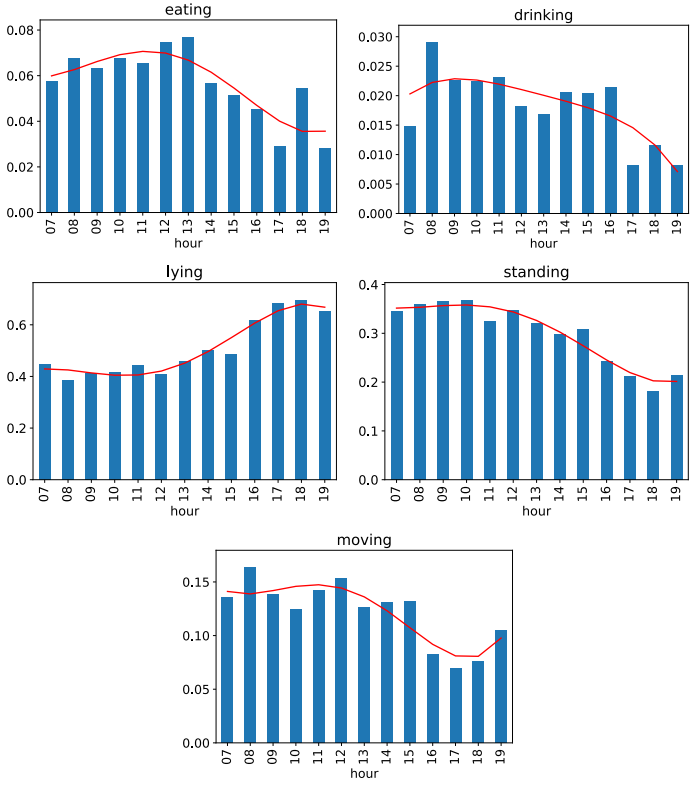


Figure 3.6: Temporal graphs of the behaviour changes aggregated by hour on the full dataset. The red solid line shows an interpolation of the data

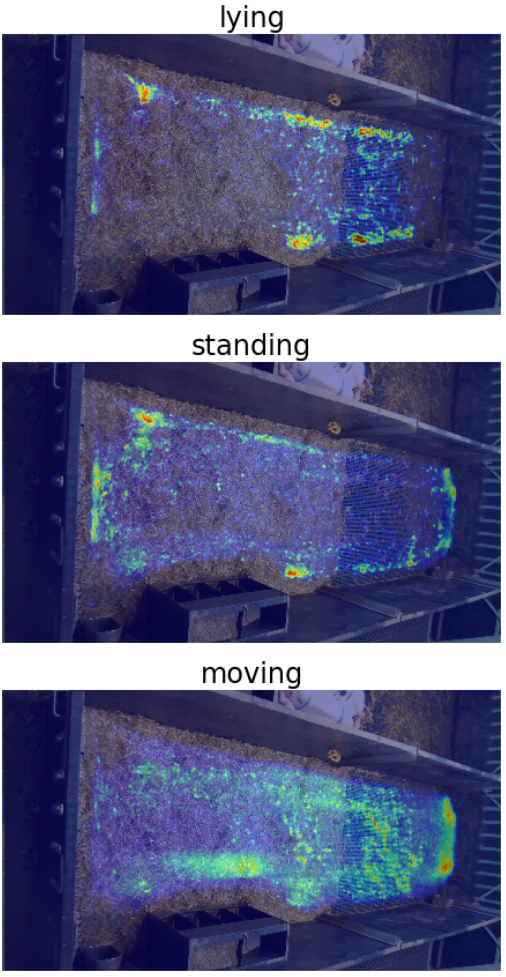


Figure 3.7: Heat-maps for 3 out of the 5 behaviours, computed from the bounding boxes detected on the full dataset. We omit eating and drinking as those behaviours are directly identified using the bounding box location. Best viewed in colour.

3.2 Segmentation Guided Scoring of Pathological Lesions in Swine Through CNNs

3.2.1 Author's main contributions

The main author's contributions for this research are:

- the implementation of the data pipeline to convert data and annotation in training samples;
- the implementation of the deep learning algorithm to score pathological lesions;
- the analysis of the results and the metrics employed.

3.2.2 Introduction

The slaughterhouse can be defined as "an establishment used for slaughtering and dressing animals, the meat of which is intended for human consumption" (Reg. CE 853/2004 of the European Parliament). Therefore, the *postmortem* inspection of slaughtered animals is mainly meant to target public health risks [115]. In addition, the abattoir is also widely recognised as a key checkpoint for monitoring the health status of livestock, representing a useful source of data for epidemiological investigations [122].

Motivations

The systematic analysis of lesions at the abattoir is a helpful feedback to the farm to assess the health status of livestock and to improve herd management: biosecurity measures, welfare, vaccination strategies, the rational administration of antimicrobial drugs. The latter is crucial to limit antibiotic-resistance [86, 141] which has been indicated as a paramount threat to human health in the upcoming future [7]. Moreover, the slaughterhouse can be useful to evaluate the economic impact of diseases, which negatively affect the profitability of farming [9, 40]. This is especially true for pigs, as their "short" productive cycle prevents the full healing of lesions, which are still visible at market weight [122].

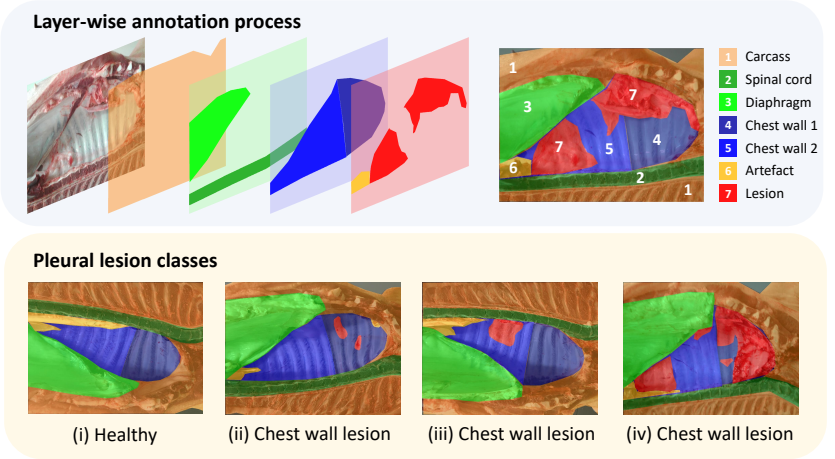


Figure 3.8: Top left: different segmentation layers; some of them were collapsed for clarity. Top right: correspondence between colours and structure names. Bottom: visual examples of the four classes for pleurisy scoring.

Currently, the registration of lesions at the abattoir is performed voluntarily, at the request of the stakeholders, and it is very challenging in terms of time and economic resources. As a consequence, large amounts of informative and relevant data are frequently lost [83].

Pleurisy Scoring Systems

Pleurisy is frequently observed at necropsy or during the *post-mortem* inspection at the abattoir, its prevalence often being close to 50% [89]. Over the years, a number of methods have been developed to score pleurisy [122]. Among these, the Slaughterhouse Pleurisy Evaluation System (SPES [115]) is widely considered as the most informative method under field conditions. Recently, an alternative method (Pleurisy Evaluation on the Parietal Pleura, PEPP [84, 85]) has been developed. The PEPP method provides well matching results when compared with SPES and, differently from SPES, it can be efficiently performed on digital images.

Automatic Swine Analysis

To the best of our knowledge, this is the first work addressing pleural lesion scoring in pigs in an automatic, data-driven fashion. Still, methods have been proposed for tasks related to swine welfare, productivity and diseases prevention. As an example, Yongwha et al. [22] proposes an automatic system for the detection and the recognition of multiple pig wasting diseases analysing audio data. Bin and Hongwei [119] employ computer vision techniques for the assessment of thermal comfort for group-housed pigs.

Contributions

Our contribution is twofold. On the one hand, we introduce a large pixel-level segmentation dataset comprised of more than 4000 images of pig carcasses. Our dataset is annotated by two sectorial experts after agreeing upon the annotation procedure. We believe this data can enable the development of new automatic methods for real-time and systematic analysis of animal health, production and welfare parameters.

On the other hand, we introduce one such method, based on deep learning techniques, which requires only a single RGB image of a pig chest wall to score pleural lesions according to PEPP [84, 85]. This method could be employed in the slaughter chain to provide a systematic, real time and cost effective diagnosis. Moreover, it could produce a continuous stream of data suitable for epidemiological investigations as well as to classify farms according to risk categories.

3.2.3 Dataset

Our dataset is comprised of 4444 high-resolution images of pig carcasses from four independent slaughterhouses. The images taken are from the end of the slaughter chain after the removal of internal organs and washing the carcass. No particular alignment is required in this or the following phases, but images where the whole carcass was not visible were discarded to simplify the annotation process later. Two experts annotate four anatomical structures; namely *carcass*, *spinal cord*, *diaphragm*, *chest wall*[1], as well as two anomalous structures: *artefacts* and *lesions* (see Fig. 3.8 and Fig. 3.9).

[1] According to PEPP [84, 85], we consider the area between the first and the fifth intercostal space as *chest wall 1* and the rest of the chest wall as *chest wall 2*.

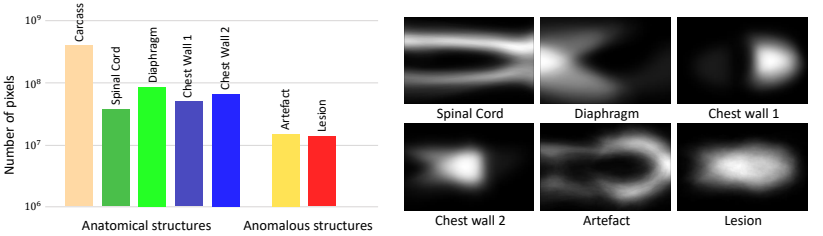


Figure 3.9: Left: number of annotated pixels per category (in log scale). Right: spatial localisation of the segmented classes, computed as the average of the annotations on the training set. A symmetry is introduced by the two halves of the carcass.

The experts follow a depth-wise fashion during the annotation process. In particular, artefacts and lesions are annotated after the anatomical structures, which are thus entirely annotated even when covered by another structure. In the following, we focus on the pleural lesion scoring task only.

Pleurisy Scoring

We simplify the scoring methodology presented in PEPP into a four class problem. Two domain experts classify each dataset image as: i) absence of lesions; ii) lesions on the first chest wall; iii) lesions on the second chest wall; iv) lesions on both chest wall areas. To allow an evenhanded evaluation we provide an independent test set of 200 images. The four classes distribution for the two sets is shown in Table 3.4. It is worth noting how the distribution of the classes in the train set is skewed towards class i (healthy swine). This reflects what is stated in [89]. Contrarily, we artificially ensure an even distribution for the test set to allow for a focus on the different pathological classes.

3.2.4 Method

Here we first discuss four classification baseline methods from state-of-the-art literature for the task of pleural lesion scoring. We then introduce our segmentation-based method, which leads to higher accuracy and interpretability by explicitly exploiting task-specific prior information.

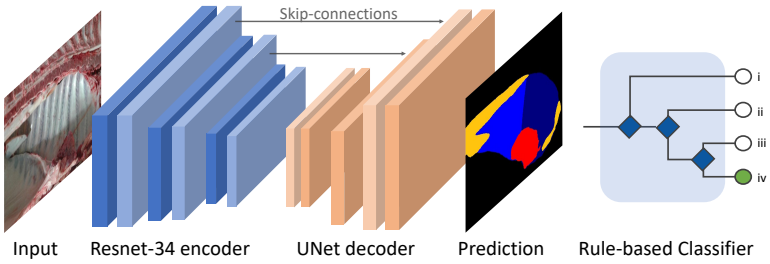


Figure 3.10: Overview of our model for pleural lesion scoring. An encoder-decoder network is trained for the *proxy* task of semantic segmentation; a rule-based classifier is then used to discern the class from the segmentation. We show how this architecture is more interpretable and outperforms competitive end-to-end classification baselines.

End-to-end Classification network

Our first two baselines rely on ResNet-34 architecture [44]. A first network (*ResNet*) is trained from scratch on our dataset. For the second one ($ResNet_{PT}$) we initialise the convolutional layers with pretrained weights from ImageNet [27]. In both networks we replace the last fully connected layer to address a four classes problem.

Deep Features and Shallow Classifiers

Similarly to [66, 88], we train two shallow classifiers in the task of pleural lesion scoring, starting from ResNet-34 activations. The first one (*SVM*) is a Support Vector Machine with Gaussian kernel and C regularisation set to 1. The second one (*RF*) relies on a Random Forest classifier with 50

Table 3.4: Number of images annotated for each class, in each dataset split.

	Class i	Class ii	Class iii	Class iv	total
Train Set	2347	381	498	1018	4244
Test Set	50	50	50	50	200

estimators and entropy split criterion.

Segmentation Guided Network for Pleural lesions

Our proposed method leverages experts' pixel-level annotation as a *proxy* for the pleural lesion scoring. Indeed, according to PEPP [84, 85], the three different pathological classes of pleural lesions (i.e. ii, iii and iv) are discerned by the lesion's location over the chest wall. Consequently, a faithful segmentation always leads to a reliable classification of the image.

With this premise, we propose to train a network for semantic segmentation and to leverage a rule-based classifier to translate the predicted maps into a pleural lesion score, see Fig. 3.10. Besides the performance gain, this framework has the desirable property of enabling an immediate interpretation of the model prediction.

Network

Our network architecture derives from UNet [109]. To ensure consistency with the baselines, we rely on pre-trained ResNet architecture as encoder; the decoder path is based on four Transposed Convolutions blocks [78], each one with the same number of channels of the corresponding ResNet layer in the encoder path. Long skip-connections promote the flow of information between the encoder and the decoder.

For training, we optimise a set of Binary Cross Entropy (BCE) losses. In particular, given the l-th channel of the output map $\tilde{\mathbf{y}}_l \in \mathbb{R}^{N\times W\times H}$ and the corresponding ground truth channel $\mathbf{y}_l \in [0,1]^{N\times W\times H}$:

$$BCE_{(\mathbf{y_1},\tilde{\mathbf{y}}_1)} = -\frac{1}{N}\sum_{i}^{N} \mathbf{y}^i \cdot \log_2(\tilde{\mathbf{y}}^i) + (1-\mathbf{y}^i)\cdot \log_2(1-\tilde{\mathbf{y}}^i) \tag{3.3}$$

Where N is the number of images and W, H are the width and the height of each image respectively. By doing that, we effectively decouple each segmentation label from the others to avoid their competition over a pixel, reflecting the annotation methodology of our dataset. Before feeding the result to the classifier, we perform a a connected component analysis on the segmentation output to identify the different regions.

Classifier

Then, the rule-based classifier C_{RB} assigns images to the different classes following these rules:

- Class i) no lesion component detected;
- Class ii) at least one lesion overlaps the first chest wall; no lesions on the second chest wall;
- Class iii) at least one lesion overlaps the second chest wall; no lesions on the first chest wall;
- Class iv) at lest one lesion on both chest walls.

Due to this small set of rules, it becomes trivial to understand why a certain decision was taken: this in turn leads to a full interpretability of the result.

3.2.5 Experiments

In this section we report quantitative results for all presented methods.

Implementation Details Our model is trained for 40 epochs with batch size 6. We resize the image to 400×300 px and perform extensive data augmentation including random horizontal or vertical flip, translation and rotation. We use Adam [55] optimiser with initial learning rate 0.001 and halve it every 15 epochs.

Results

Classification results are reported in terms of accuracy and confusion matrix. The accuracy is computed using the rules from the classifier presented above. Fig. 3.11 (a) illustrates the test set accuracy scores for the applied methods. Both *RF* and *SVM* perform poorly with respect to end-to-end methods. This suggests that even though ImageNet features may provide a good representation, fine-tuning the features extractor is still required to achieve good results. In fact, end-to-end training gives to $ResNet_{PT}$ a significant performance boost over shallower models. *ResNet* closely follows but loses 2% accuracy, suggesting that even though a domain shift between ImageNet and our dataset is certainly present, features learned on the former still provide a suitable initialisation.

Eventually, we need to steer towards a segmentation guided approach to get an additional significant performance improvement. Since pathological classes are discerned by the position of lesions, segmentation is crucial to preserve this spatial information in the output. Then, the rule based classifier translates this localised information into a discrete label by applying rules that derive from experts' knowledge. As shown in Fig. 3.12 (a), classification accuracy increases steadily as IoU improves. This highlights how segmentation is indeed an optimal *proxy* for pleural lesion scoring.
The confusion matrix for our method is shown in Fig. 3.11 (b). Notably, anti-diagonal entries of the confusion matrix are all zero, highlighting the absence of healthy examples associated with the most severe pathological class and *vice-versa*. Furthermore, very few examples are misidentified even between the two less represented classes (i.e. ii and iii).

Method	Accuracy				
	i	ii	iii	iv	*Avg*
RF	0.92	0.3	0.1	0.44	0.44
SVM	0.8	0.56	0.24	0.46	0.52
ResNet	0.90	0.64	0.58	0.72	0.71
$ResNet_{PT}$	0.92	**0.70**	0.62	0.68	0.73
ours	**0.98**	0.68	**0.78**	**1.00**	**0.86**

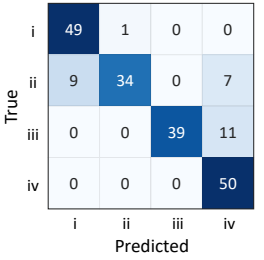


Figure 3.11: (Top) Class-wise accuracy score on the test set for each method. (Bottom) Confusion matrix on the test set for our proposed method (Sec. 3.2.4).

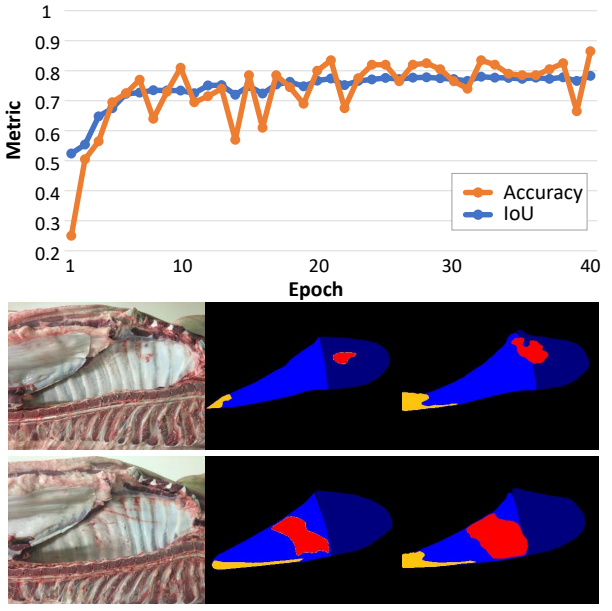


Figure 3.12: (Top) Correlation between accuracy and IoU during training. (Bottom) Input image, ground truth annotation and predicted outputs for two samples. In the first one, a lesion overlapping both chest walls is predicted, while the one in the ground truth is restricted to the second chest wall only. Similarly, in the second one the predicted lesion touches also the second chest wall.

Discussion

Despite the encouraging results, the pleural lesion scoring task is far from being solved. We believe this is due to various factors. On the one hand, the border between the two chest walls is crucial for the final classification. This can be observed from the fourth column of the confusion matrix in Fig. 3.11 (b). The most common source of confusion from the model comes from predicting a wrong border for the chest walls which classifies predicted lesions as type iv. This kind of failure case is shown in Fig. 3.12 (b). On the other hand, organic nature of lesions and artefacts is reflected in a high intra-class appearance variance, including cuts and bruises in tissues and

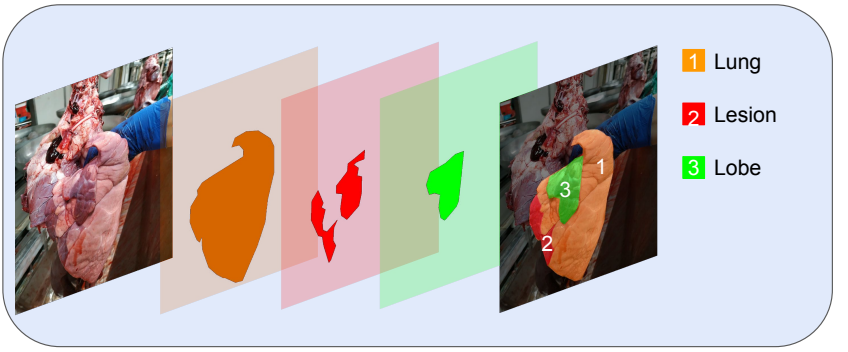


Figure 3.13: Top left: different segmentation layers. Top right: correspondence between colours and structure names.

blood staining: this makes the two classes particularly hard to distinguish.

3.2.6 Extension to Lung Lesions Scoring

In this section we explore the extension of the work done to the anatomical structure of the lungs. Differently from the carcass, lungs are highly irregular and volumetric structures and cannot be easily capture in a single picture. Still, they have seen an increasing interest because post-mortem lesions correlate well with a respiratory infection that significantly affects the final yield of the animal.

The employed approach mimics the one already described above. However, there are some key differences in how results are evaluated. In fact, for lungs we're only interested in estimating how much of the structure is affected by the lesion. This makes it a purely regression task where classification is not needed anymore. Still, a single pixel can belong to multiple structures.

Fig 3.13 shows the classes of interest for this task. While in theory only lung and lesion are requested, lobe is included as the total area of the lung changes accordingly to the presence of lobes (which are in fact the projection of a 3D structure in the image).
Please note that the employed CNN is the same as before, with the only difference of the final rule-base classifier, which is not included for this

regression task. The output of this model is therefore a three-channel image where each value falls in the $[0, 1]$ range. This encodes the probability of each pixel to belong to that specific class. We perform an extensive data augmentation on input/output pairs which includes random translation, rotation, sheer and HSV change.

For this task a new dataset was collected and annotated by two experts. Dataset statistics are reported in Table 3.5. The experts were asked to keep an equilibrium between pathological (lungs affected by a lesion, regardless of the size of it) and healthy (lungs without any lesions) images. The test set is purposely perfectly balanced.

Segmentation results have been scored with the following metric:

- If the image was annotated as healthy, score 1 if and only if the prediction has no pixels assigned to the lesion class;
- If the image was annotated as pathological, score 1 if and only if the IoU between predicted and annotated lesions is greater than 25%;

After 24k iterations our best model scores 85% accuracy according to the above presented metric. This is still an ongoing research and as such these are to be considered preliminary results. Still, it's a promising start for this novel task.

Visual Results Analysis

We report here a detailed analysis of the images from the test set. For all the figures in this section:

- The first two rows present the input images;
- the middle two rows are the targets;

Table 3.5: Number of images annotated for each class (pathological and healthy), in each dataset split.

	Healthy	Pathological	total
Train Set	3113	3884	6997
Test Set	50	50	100

- the last two rows are the predictions.

Figure 3.14 reports results for images from the healthy class only (no lesion label is present the targets). In only a single case (fourth image from the left on last row) out of 16 our model predicts a non-existent lesion. This is located in a lobe which is hidden from the light by the main lung. This dark colour may have triggered the lesion class. It can also be appreciated how the general shape of the lung is correctly predicted, although some holes are still present. A lobe is falsely detected in the third image from right in the last row. In this case, a structure which could be described as a lobe is indeed present. This may be an annotation error.

Figure 3.15 reports results for images from both classes (no lesion label is present on the first two targets). Again, a darker area in the first image triggers a lesion annotation. For the pathological images, a lesion is always correctly predicted and located, although the extension is often overestimated. The lobe classes are almost always correctly found (with the exception of a small one in the fifth image from left in the last row). It can be appreciated how the network can successfully deal with abnormal lung shapes, such as the second and third images in the last row.

Figure 3.16 reports results for images from the pathological class only. Lesions are correctly identified and placed. Lung's areas are in many cases overestimated and fragmented here. Adding a constraint over the learning process to avoid this is a possible future work.

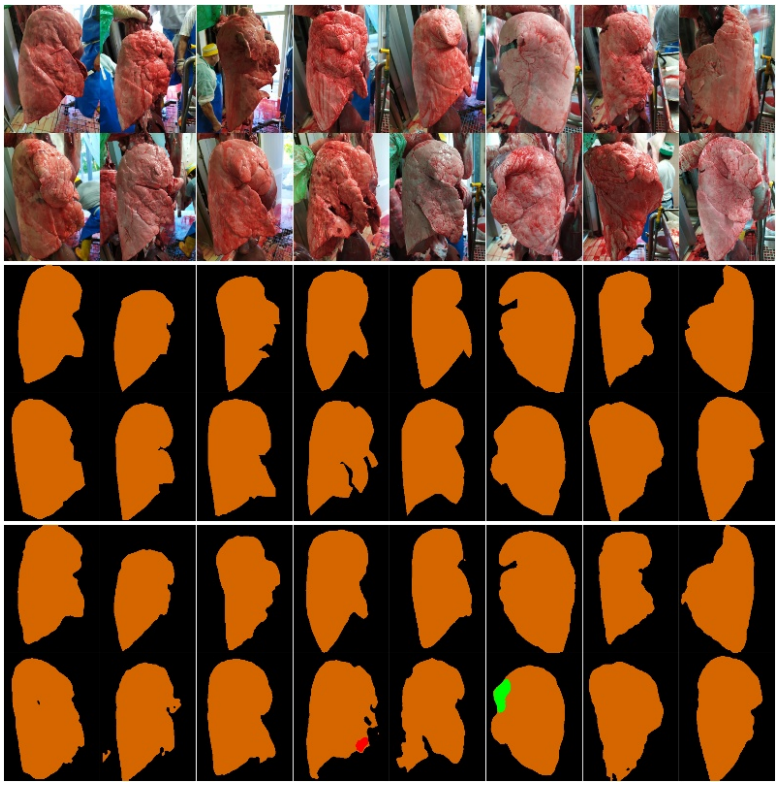


Figure 3.14: Results for lung lesions(I). The input images are shown in the top two rows. Ground truth annotations follow, while predictions are depicted in the bottom two rows.

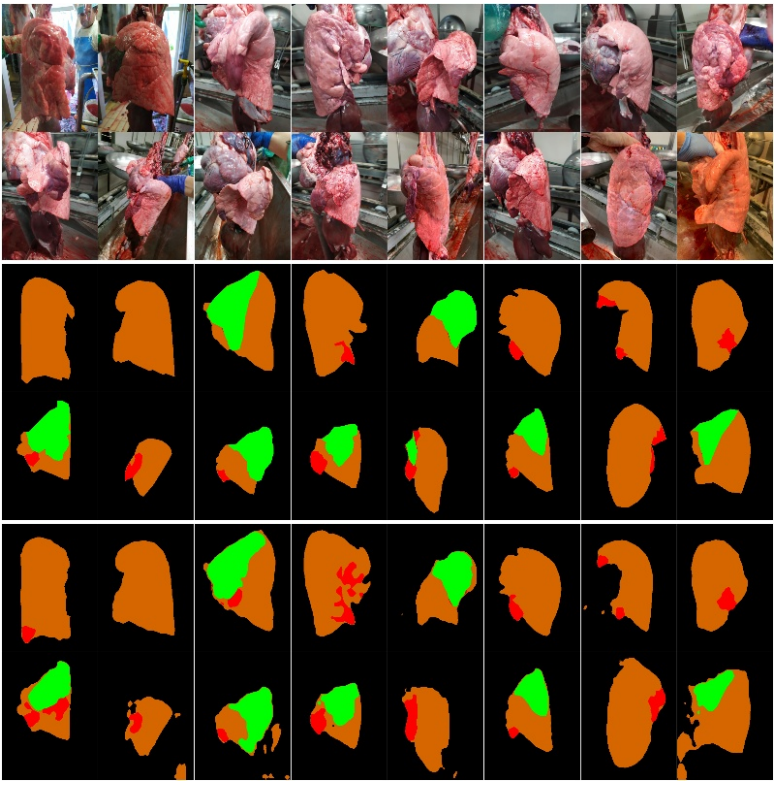


Figure 3.15: Results for lung lesions(II). The input images are shown in the top two rows. Ground truth annotations follow, while predictions are depicted in the bottom two rows.

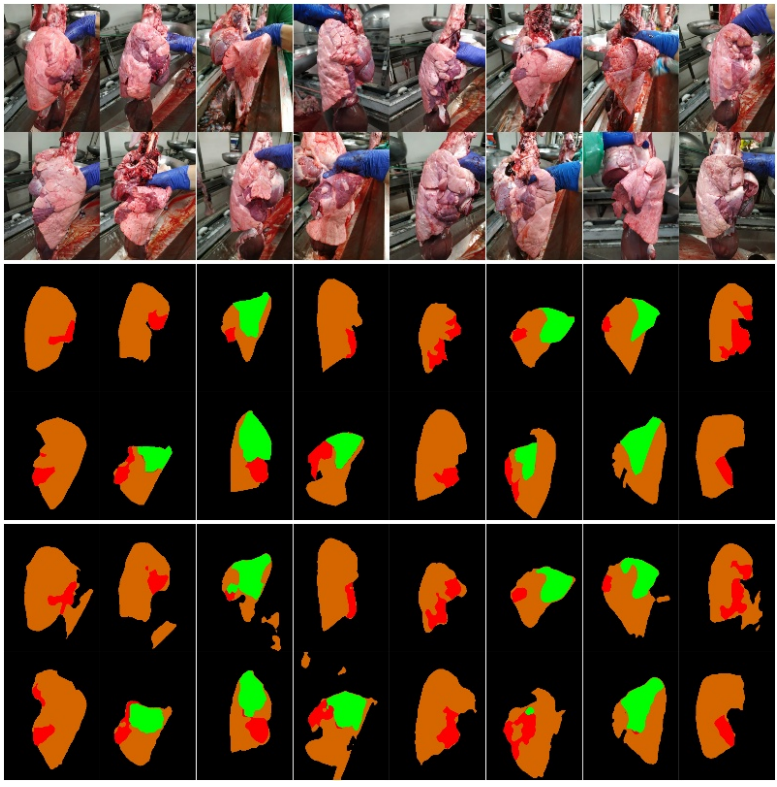


Figure 3.16: Results for lung lesions(III). The input images are shown in the top two rows. Ground truth annotations follow, while predictions are depicted in the bottom two rows.

3.3 Multi-Views Embedding for Cattle Re-identification

3.3.1 Author's main contributions

The main author's contributions for this research are:

- the collection and annotation of a consistent portion of the data;
- the implementation of the data pipeline to convert data and annotation in training samples;
- the implementation of the deep learning algorithm to score pathological lesions;
- the implementation of the competitors' methods on the proposed dataset;

3.3.2 Introduction

Animal Re-identification shares some of the aims of the human task, while also including new challenges. The identification process represents a pillar for national and international trade, especially for animals representing crucial economic assets. Furthermore, it is a method for validating the quality and the "authenticity" of the animal being traded. Similarly, for animals supplying products intended for human consumption, the animal identity and the traceability along the entire value chain are prerequisites for the certification of the quality and the safeness of the product for final consumers. In fact, as some of these animals may host and transmit pathogens, a monitoring system is essential to avoid the spread of such diseases to humans and animals and it is necessary to easily identify and track the origin of infected products. Finally, stock theft represents an issue that often outbreaks into a social challenge in developing countries. As an example, India reported almost ten thousand cattle thefts in 2015 while the number of horse theft has grown past forty thousand world wide [2].

On the other hand, re-identification systems for pets has seen some interest in the computer vision community [93], mainly aiming to retrieving lost "family members".

Finally, re-identification systems may represent an opportunity for safeguarding endangered species, acting as a crucial aid for studying wildlife

and for conservation actions. For such animals, traditional identification systems make use of electronic chips placed in collars and require the animal to be captured and immobilised at least once. Such practice may be unfeasible for aggressive or elusive species, and typically requires GSM or satellite transmitters, the latters being very expensive and often impractical. Again, only few noticeable novel and unobtrusive methods [25], [26], [96] have been proposed, mainly due to the lack of large datasets publicly available.

Cattle Re-identification Motivations

The number of cattle in Europe in 2016 stood at almost 122 million, of which 23.3 million were dairy cows [3], and the number is growing past 1400 million in the world based on the latest surveys [1]. Although efficient identification and re-identification systems already exists, it is mandatory to develop new tools that can support the existing ones not only to ensure milk and meat safeness, but also to avoid kidnappings and counterfeits while improving animal health and animal welfare. Nowadays, the following methods are employed:

- RFID subcutaneous chips or rumen bolus, which can be read using a dedicated electronic reader;
- Ear tags, holding the animal identification number according to the country legislation format;
- Brand code, marked on the animal skin and used as a traditional identification system mainly in developing countries.

Authors from [30] reported further details on RFID and electronic identification system for cattle in the US market, while the reader may refer to [16] for an extensive review of cattle identification systems worldwide.

Machine Learning Motivations and Insights

The methods reported above suffer from major drawbacks. In particular, the use of RFID devices entails a significant cost for farmers of developing countries, because of the installation fees and the need of electronic readers during re-identification procedures. Moreover, they may, sometimes, have a impact on animal welfare. On the other hand, ear tags are cheaper to

buy but can be easily counterfeited or even removed through ear excision, beside being often lost by the animal itself. Consequently, there is room and a strong need to develop new methods with the following requirements:

- Cheap for both installation and maintenance, including the supporting hardware;
- Able to be easily and rapidly used in real scenarios, as instance in the field or in a stable;
- Hard to be counterfeited or removed.

Re-identification based on images holds all these properties, and can be exploited using the latest techniques and advances from Deep Learning and Computer Vision. Differently from traditional machine learning, deep learning techniques do not require any human hand-crafted features as they learn those representations directly from data, identifying features that may be more robust to pose or backgrounds variations as well as illumination changes. This is especially needed for cattle, since it is not easy to obtain images with a predefined pose of the animal, as it tends to move constantly while roaming or eating.

While for humans it is widely recognised that the face holds a great importance for visual identification purposes, in the animal kingdom a similar certainty still lacks, as almost no studies for this specific task have been conducted yet. Cattle present a high inter-breed variance in both body proportions and skin textures. On one hand, this makes fairly easy to distinguish cows of different breeds even for novices, on the other hand, due to the genetic selection made by humans in the past centuries and even more in the last decades, cows present a lower inter-breed variance compared to humans. However, despite this quite high degree of inbreeding, many cow breeds hold a unique texture pattern that is different from animal to animal, while also behaviour and social interaction contribute with marks, scratches and other defects that remain evident on the animal skin.

In this work we used pictures of the head of cows taken from different angles of rotation and inclination, essentially for the following reasons:

- The head of a cow shows a sufficient characteristic set of textures, shapes and patches. Even for texture-less breeds (such as the Bruna Alpina), the presence of horns and their length or the fur colour contributes to this variety. Furthermore, [38] showed how cattle face

muscles are sufficiently developed to exhibit different facial expression, that may be used to distinguish one from another.

- Most of our images have been collected in farms with cows restrained during veterinary procedures, and only the head is easily accessible;

- Pictures of the full cow would introduce variance in both the animal pose and the background and could possibly require even more images of the same animal to perform the re-identification task.

Furthermore, an approach based on "facial" images could be compared with the current literature available on human re-identification.

Main contributions and Novelties

The contributions of this research are two-fold. Firstly, we provide a deep learning based framework for cattle re-identification. Secondly, we demonstrate how the use of multiple views of the same cow leads to superior performance w.r.t. standard approaches, the latter typically using only the front view image of the subject.

3.3.3 Proposed Method

In order to leverage the textures and details of both cattle profiles (i.e. the frontal and one of the two sides), we built an embedding DCNN starting from two images of the same cow.

At a high level, the network takes in input the two images and subsequently outputs a 128 dimensional embedding with unitary L2 norm. More in detail, each of the two images is independently processed by a separate convolutional branch, and their outputs are concatenated to form a single feature vector. It is worth noting that the two branches do not share any parameter, since different features may be required for the two animal profiles.

Our multi-view network has been designed by means of two building blocks:

- **ConvBlock**: A single Convolutional layer followed by InstanceNorm [137] and LReLU activation, reducing the feature maps' spatial resolution by a half;

- **ResBlock**: A residual unit [44] with LReLU activation, preserving the spatial resolution.

A scheme of these blocks is shown in Figure 3.17.

Instead of using the more popular batch normalisation, we address the internal covariate shift problem by means of instance normalisation. Indeed, with the second, we observed improvements in terms of stability during the training phase. The network ends with a single 2D Convolution, with kernel size equals the feature map size and number of filters equals to the desired embedding size (i.e. 128). In this way, each input map is reduced to a single scalar value, leading to a fully convolutional architecture. The overall architecture is presented in Figure 3.18.

We employed the Histogram Loss from [140] as the only loss function of our architecture, which is an alternative to the widely used triplet loss. After a batch of anchors, positives and negatives is embedded into a high-dimensional space by a deep network, the loss computes the histograms of similarities of positive and negative pairs. The integral of the product between the negative distribution and the cumulative density function for the positive distribution is evaluated, corresponding to a probability that a randomly sampled positive pair has smaller similarity than a randomly sampled negative pair. We performed extensive comparison using the triplet loss function described in [114], but found the latter more unstable during the train phase.

3.3.4 Dataset

A potential drawback in the use of Deep Learning methods is that a huge amount of pictures depicting cows' heads is required to achieve reasonable performances on unseen examples. Furthermore, the training set should include a great variety of poses, illumination changes and background for each subject, with the acquisition process spanning potentially in multiple days. However, to the best of our knowledge, such a dataset still does not exist for cows' faces. Thus, we collected pictures and video of cattle from four Italian farms distributed in three regions. We collected videos and extracted images from those for the training process, while employing only images for the test phases. We leveraged the Vatic tool [142] to annotate the cows' faces with a bounding box for each frame. Finally, we discarded some of the extracted frames aiming to ensure a high inter-frames variance.

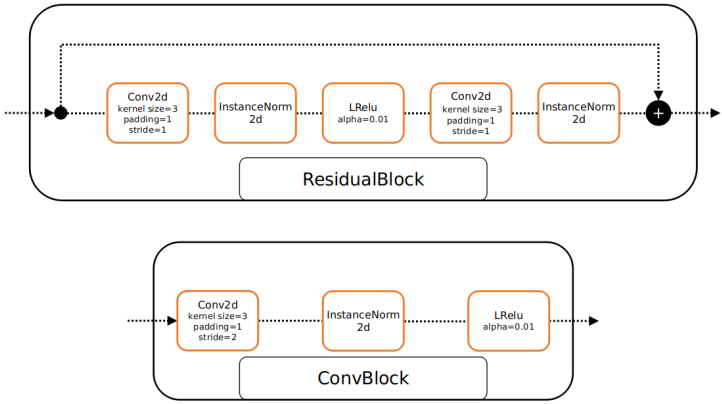


Figure 3.17: Base blocks used in our architectures.

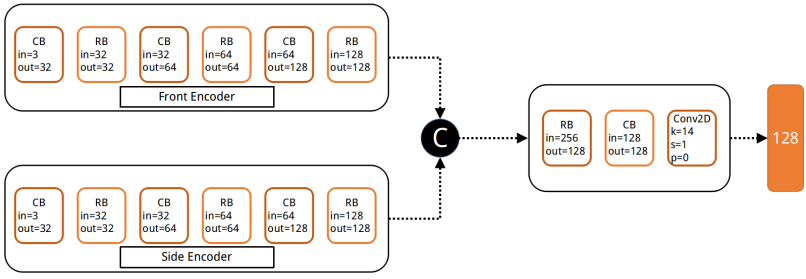


Figure 3.18: Our multi-views architecture. Note that CB stands for ConvBlock, while RB for ResidualBlock, as described in Figure 3.17

Such activity is required since the animal usually moves slowly during video acquisition, introducing a lot of redundancy if all the frames are kept. Moreover, a traditional setting for the re-identification task consists of few different pictures per single identity. Following the widely adopted split strategy for re-identification, we obtained the following splits:

- Train Set; consisting of 12952 pictures from 387 different subjects;
- Database Set; consisting of 4289 pictures from 52 different subjects, recorded during two different days. This set is used to match identity from the test set;
- Test Set; consisting of 561 pictures from 52 different subjects. These cows are the same included in the Database Set;

Some random samples from the last two sets are shown in Table 3.6. It is worth noting that, given an image, one cannot make any assumptions regarding the cow's face location and orientation. Moreover, because of the oblong shape of cow faces, any alignment would lead to part of the cow face being cropped. Finally, only few landmarks detectors work on animal faces [103], but they need to be fine-tuned on the animal domain, thus requiring expensive landmark annotations.

3.3.5 Experiments

Metrics

Following [26] and [71], we test our solution in two different settings:

- **Open-Set**: Identities of the test images are included in the training set.
- **Closed-Set**: Identities of the test images are separated from the training set ones.

Regarding the last, we consider it as more challenging and general, being also able to provide a good estimation of the generalisation capacity of the proposed model. For both settings, we conducted the same experiment, namely the **Identification**. Given images and correspondent ground-truth identities from a test set, the matching strategy returns the k nearest neighbours from the database set. It is worth noting that the above mentioned "matching strategy" can be implemented by every classifier.

Table 3.6: Some randomly drawn samples from our dataset.

Baselines

We include both deep and non-deep baselines to further present and motivate the main challenges of this novel task. For non-deep baselines, we include methods traditionally employed by the computer vision community for the human re-identification task. The reason behind this choice is not only to emphasise the differences between the human related task and the cattle one, but also because the great majority of them are provided as open-source verified software with the open-cv [17] package.

The EigenFaces [136] method consists of a Principal Component Analysis (PCA) applied to images of human faces. The first k eigenvectors, sorted by their respective eigenvalues magnitudes, can be seen as prototypes used to build the data. Test images can then be projected to extract k coefficients, each one describing how much a prototype contributes to the image.

FisherFaces [11] extends Eigenfaces by means of the Fisher Discriminant Analysis (FDA), aiming to force multiple images of the same identities to lie in a nearby region of the subspace. While PCA preserves maximum variance during the projection, FDA attempts to preserves the discrimination capability at the end of such transformation. Indeed, FDA may be consider a supervised embedding algorithm, since it finds a projection that maximises the scatter between classes and minimises scatter within classes. Thanks to this, the Fisherface method shows superior performances w.r.t. Eigenfaces under the presence of variation in lighting and expression.

LBPH[97] has been widely used as a texture description. It builds a circular neighbour with a certain radius for each pixel, and extracts features based on the relationships between each pixel and its neighbours. The latter are used to build a histogram, which is then used to describe the image, and can act as a feature descriptor for a further PCA.

HOG, similarly to LBPH, builds a histogram using neighbours but, instead of pixels values, the spatial derivatives are employed. The histogram takes into account both the magnitude and the orientation of the gradients, with the latter being quantized to contribute to achieving invariance to orientation.

The authors of SphereFace[71] present a DCNN trained on a large human faces dataset, achieving state-of-the-art performances in an open-set setting. The images are firstly aligned and cropped, and a 512-dimensional feature vector is extracted from the second-last layer of the network. A classification loss, named Angular Softmax, is proposed: on one hand,

it requires examples from the same identity to lie nearby on the output landscape. On the other hand, it forces examples from different identities to be spaced by a considerable margin, the latter being in the form of an angle on a hypersphere. As versions of the network pretrained on human faces are available, results with and without a training phase on cattle are reported.

It is worth noting that all methods listed above share the same output representation; in particular a feature vector (embedding) is produced from a given input image and, as such, the "matching strategy" can be the same for all of them.

Implementation Details

As far as it regards the train methodology, it is worth noting that:

- For the **closed set** scenario we train on both the Train and the Database Set, while for the **open set** one we train only on the first;
- We pre-processed the images by scaling them to a fixed size (i.e. 224).
- We performed data augmentation by randomly rotating, cropping and projecting images, while also changing the hue and saturation of the images. We didn't perform horizontal flip, as it causes a noticeable drop in performances. We suspect this is because of the asymmetrical patterns of cows;
- We employed the Histogram Loss using a batch size of 64 triplets and 200 bins for the histograms. K-NN matching was performed using L2 distance in the embedding space;
- We mined both hard positives and negatives during the training phase. The former are positives with an embedding extremely far away from the average embedding of the identity, while the latter are the closer negatives to the identity in the embedding space;

Results

As shown in Tables 3.7 and 3.8, we compare our results with the baselines discussed previously. For each method, the hyper-parameters tuning activity has been conducted using a grid-search strategy. For the identification

	EigenFaces		FisherFaces		LBPH		HOG	
	Open	Close	Open	Close	Open	Close	Open	Close
Top1	0.227	0.229	0.242	0.237	0.263	0.265	0.39	0.4
Top3	0.352	0.354	0.345	0.345	0.406	0.415	0.522	0.532

Table 3.7: Results for the Identification task (I).

	SphereFace		SphereFace(cattle)		Ours	
	Open	Close	Open	Close	Open	Close
Top1	0.139	-	0.367	0.556	0.558	**0.817**
Top3	0.226	-	0.46	0.636	0.742	**0.891**

Table 3.8: Results for the Identification task (II).

task, results are reported in terms of Top1 and Top3 match, for both open and close settings. For every method, KNN (k-nearest neighbours) has been employed as final classifier, as the it requires no hyper-parameters grid-search nor supervised training, while also scaling linearly with the number of identities.

The input of each method is a single image, whereas our proposed method can be provided with two different profiles images. In order to enable a fair comparison and to motivate our design choices, results with only one profile image are reported in subsection 3.3.6.

Looking at results showed in Table 3.8, the following conclusion may be drawn:

- Methods based on local and stationary property (i.e. LBPH, HOG and CNNs) achieve better performance than Eigenfaces and Fisherfaces, which do not implicitly exploit nearby pixel correlations or local pattern's presence.
- Deep models trained on cattle performs better than shallow ones. This may be due to their robustness to different poses and other major source of variations (i.e background or illumination changes).
- SphereFace, when trained on aligned human faces, does not generalise to cattle. Such result show how the cattle re-identification task has nothing to do with the same task in the human domain, differently from what happen with apes.

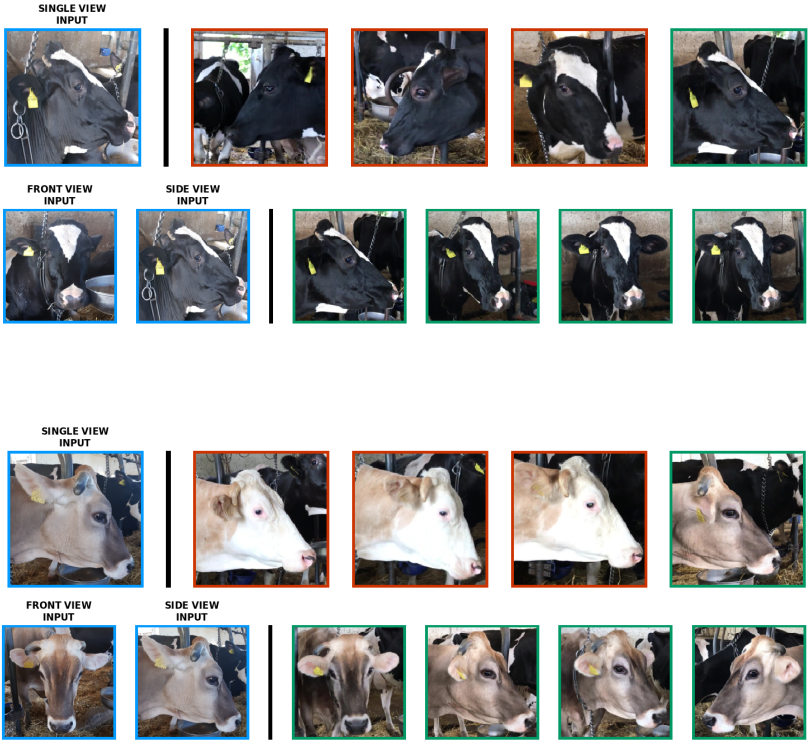


Figure 3.19: Illustration of the K-NN retrievals, computed on two individuals using a single view or a double view architecture. Input images have been labelled with blue contour, while green and red have been used respectively for correct and misclassified examples. Best viewed in colours.

- Our solution outperforms by a consistent margin all the other competitors, including a state-of-the-art human re-identification network as SphereFace (even if trained from scratch on cattle). Such improvement is achieved by leveraging two different cow profiles, which leads to a higher discriminative capability for similar subjects.

	Single-View		Multi-View	
	Open	Close	Open	Close
Top1	0.443	0.688	0.558	**0.817**
Top3	0.575	0.748	0.742	**0.891**

Table 3.9: Comparison between single and multi views methods.

3.3.6 Ablation Study

Multi-view vs Single-view

The results for the Identification task reported in Table 3.9 highlight the superiority of the multi-view approach over the single-view one. 3.19 shows some manually selected example where the single-view model fails, while the multi-view approach effectively fuses features from both views to produce a more representative and robust embedding vector. This justifies the request of both views for a single prediction. Indeed, if two cows may have the same visual appearance under some poses or particular light conditions, on the other hand this possibility has a lower probability under the presence of both profiles.

Moreover, cattle usually do not share a symmetry between left and right profile, as they often present very different spots and patterns. Also for such reason, the use of two profiles instead of one should be considered useful to find a meaningful discrepancy between two cows.

Extended Database

In Table 3.10 we report results, in term of accuracy, showing how the use of the only database set during the test phase leads to better performances with respect to the union of the train and database sets. However, for the close set scenario, the drop of performance between the two settings is much lower w.r.t the open one, highlighting how the network improves its performance if the animal's images are available during the train phase. In this way, such knowledge may be used during the test phase to reject other subjects with similar patterns or characteristics. To further highlight the differences in terms of difficulty between the two settings a random predictor has been included.

	Database Set		Extended Set	
	Open	Close	Open	Close
Random3	0.057	0.057	0.006	0.006
Top1	0.558	**0.817**	0.396	0.732
Top3	0.742	**0.891**	0.583	0.820

Table 3.10: Comparison with different set during the test phase. Random3 stands for a random predictor scoring 1 if the correct identities lies among the first 3 prediction.

3.4 Robust Re-Identification by Multiple Views Knowledge Distillation

subsectionAuthor's main contributions The main author's contributions for this research are:

- the implementation of the data pipeline to convert data and annotation
- the implementation of the teacher network;
- the implementation of the competitors' methods on the available datasets;
- the model explanation and several ablation studies.

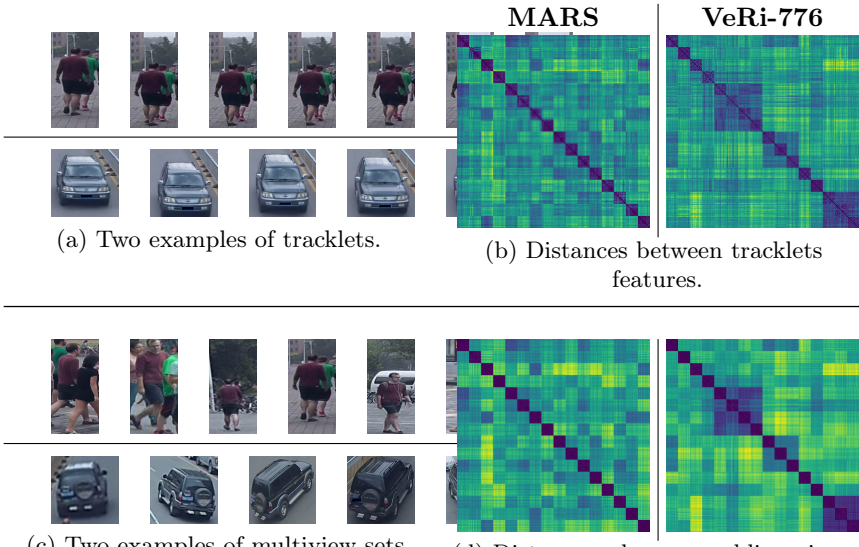
3.4.1 Introduction

Recent advances on Metric Learning [114, 121, 145, 140] give to researchers the foundation for computing suitable distance metrics between data points. In this context, Re-Identification (Re-ID) has greatly benefited in diverse domains [159, 54, 113], as the common paradigm requires distance measures exhibiting robustness to variations in background clutter, as well as different viewpoints. To meet these criteria, various deep learning based approaches leverage videos to provide detailed descriptions for both query and gallery items. However, such a setting – known as Video-To-Video (V2V) Re-ID – does not represent a viable option in many scenarios (e.g. surveillance) [154, 149, 95, 41], where the query comprises a single image (Image-To-Video, I2V).

As observed in [41], a large gap in Re-ID performance still exists between V2V and I2V, highlighting the number of query images as a critical factor in achieving good results. Contrarily, we advise the learnt representation should not be heavily affected when few images are shown to the network (*e.g.* only one). To bridge such a gap, [41, 13] propose a teacher-student paradigm, in which the student – in contrast with the teacher – has access to a small fraction of the frames in the video. Since the student is educated to mimic the output space of its teacher, it will show higher generalisation properties than its teacher when a single frame is available. It is noted that these approaches rely on transferring *temporal* information: as datasets often come with tracking annotation, they can guide the transfer from a tracklet into one of its frames. In this respect, we argue the limits of transferring temporal information: in fact, it is reasonable to assume a high correlation between frames from the same tracklet (Fig. 3.20a), which may potentially underexploit the transfer. Moreover, limiting the analysis to the temporal domain does not guarantee robustness to variation in background appearances.

Here, we take a step forward and consider which information to transfer, shifting the paradigm from *time* to *views*: we argue that more valuable information arises when ensembling diverse views of the same target (Fig. 3.20c). This information often comes for free, as various datasets [158, 147, 73] provide images capturing the same target from different camera viewpoints. To support our claim, Fig. 3.20 (right) reports pairwise distances computed using ResNet-50, when trained on Person and Vehicle Re-ID. In more detail: matrices from Fig. 3.20b visualise the distances when tracklets are provided as input, whereas Fig. 3.20d shows the same for sets of views. As one can see, leveraging different views leads to a more distinctive blockwise pattern: namely, activations from the same identity are more consistent if compared to the ones computed in the tracklet scenario. As shown in [135], this reflects a higher capacity to capture the semantics of the dataset, and therefore a *graceful* knowledge a teacher can transfer to a student.

Based on the above, we propose Views Knowledge Distillation (**VKD**), which transfers the knowledge lying in several views in a teacher-student fashion. VKD devises a two-stage procedure, which pins the visual variety as a teaching signal for a student who has to recover it using fewer views. We remark the following contributions: *i)* the student outperforms its teacher by a large margin, especially in the Image-To-Video setting; *ii)* a



(a) Two examples of tracklets.

(b) Distances between tracklets features.

(c) Two examples of multiview sets.

(d) Distances when ensambling views.

Figure 3.20: Visual comparison between tracklets and viewpoints variety, on person (MARS [158]) and vehicle (VeRi-776 [73]) re-id. Right: pairwise distances computed using features from ResNet-50. Inputs batches comprise 192 sets from 16 different identities, grouped by ground truth identity along each axis.

thorough investigation shows that the student focuses more on the target compared to its teacher and discards uninformative details; *iii)* importantly, we do not limit our analysis to a single domain, but instead achieve strong results on Person, Vehicle and Animal Re-ID.

3.4.2 Method

We purse the aim of learning a function $\mathcal{F}_\theta(\mathcal{S})$ mapping a set of images $\mathcal{S} = (s_1, s_2, ..., s_n)$ into a representative embedding space. Specifically, $\mathcal{S}$ is a sequence of bounding boxes depicting a target (*e.g.* a person or a car), for which we are interested in inferring its corresponding identity. We take advantage of Convolutional Neural Networks (CNNs) for modelling $\mathcal{F}_\theta(\mathcal{S})$.

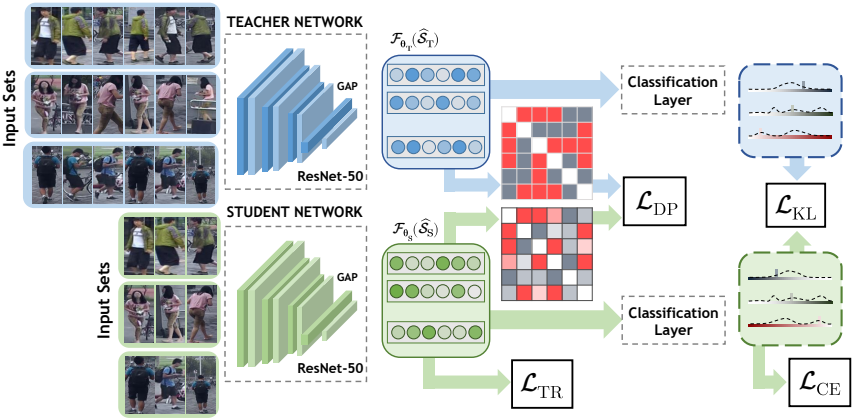


Figure 3.21: An overview of Views Knowledge Distillation (VKD): a student network is optimised to mimic the behaviour of its teacher using fewer views.

Here, we look for two distinctive properties, aspiring to representations that are *i)* invariant to differences in background and viewpoint and *ii)* robust to a reduction in the number of query images. To achieve this, our proposal frames the training algorithm as a two-stage procedure, as follows:

- **First step**: the backbone network is trained for the standard Video-To-Video setting.
- **Second step**: we appoint it as the teacher and freeze its parameters. Then, a new network with the role of the student is instantiated. As depicted in Fig. 3.21, we feed frames representing different views as input to the teacher and ask the student to mimic the same outputs from fewer frames.

Teacher Network

Without loss of generality, we will refer to ResNet-50 [44] as the backbone network, mapping each image s_i from S to a fixed-size representation d_i (in this case $D = 2048$). Following previous works [81, 41], we initialise the network weights on ImageNet and additionally include a few amendments [81]

to the architecture. First, we discard both the last ReLU activation function and final classification layer in favour of the BNNeck one [81] (*i.e.* batch normalisation followed by a linear layer). Second: to benefit from fine-grained spatial details, the stride of the last residual block is decreased from 2 to 1.

Set representation Given a set of images S, several solutions [76, 154, 68] may be assessed for designing the aggregation module, which fuses a variable-length set of representations $d_1, d_2, \ldots, d_n$ into a single one. Here, we naively compute the set-level embedding $\mathcal{F}(\mathcal{S})$ through a temporal average pooling over the representations. While we acknowledge better aggregation modules exist, we do not place our focus on devising a new one, but instead on improving the earlier features extractor.

Teacher optimisation We train the base network - which will be the teacher during the following stage - combining a classification term $\mathcal{L}_{\text{CE}}$ (cross-entropy) with the triplet loss $\mathcal{L}_{\text{TR}}$[2]. The first can be formulated as:

$$\mathcal{L}_{\text{CE}} = -\boldsymbol{y} \log \hat{\boldsymbol{y}} \tag{3.4}$$

where $\boldsymbol{y}$ and $\hat{\boldsymbol{y}}$ represent the one-hot labels (identities) and the output of the softmax respectively. The second term $\mathcal{L}_{\text{TR}}$ encourages distance constraints in feature space, moving closer representations from the same target and pulling away ones from different targets. Formally:

$$\mathcal{L}_{\text{TR}} = \ln(1 + e^{\mathcal{D}\left(\mathcal{F}_\theta(\mathcal{S}_a^i), \mathcal{F}_\theta(\mathcal{S}_p^i)\right) - \mathcal{D}\left(\mathcal{F}_\theta(\mathcal{S}_a^i), \mathcal{F}_\theta(\mathcal{S}_n^j)\right)}), \tag{3.5}$$

where $\mathcal{S}_p^i$ and $\mathcal{S}_n^j$ are the hardest positive and negative for an anchor $\mathcal{S}_a^i$ within the batch. In doing so, we rely on the batch hard strategy [46] and include P identities coupled with K samples in each batch. Importantly, each set $\mathcal{S}^i$ comprises images drawn from the same tracklet [68, 34].

Views Knowledge Distillation (VKD)

After training the teacher, we propose to enrich its representation capabilities, especially when only a few images are made available to the

[2]For the sake of clarity, all the loss terms are referred to one single example. In the implementation, we extend the penalties to a batch by averaging.

model. To achieve this, our proposal bets on the knowledge we can gather from different views, depicting the same object under different conditions. When facing re-identification tasks, one can often exploit camera viewpoints [158, 106, 73] to provide a larger variety of appearances for the target identity. Ideally, we would like to teach a new network to recover such a variety even from a single image. Since this information may not be inferred from a single frame, this can lead to an ill-posed task. Still, one can underpin this knowledge as a supervision signal, encouraging the student to focus on important details and favourably discover new ones.

Views Knowledge Distillation (**VKD**) stresses this idea by forcing a student network $\mathcal{F}_{\theta_S}(\cdot)$ to match the outputs of the teacher $\mathcal{F}_{\theta_T}(\cdot)$. In doing so, we: *i)* allow the teacher to access frames $\hat{\mathcal{S}_T} = (\hat{s}_1, \hat{s}_2, \dots, \hat{s}_N)$ from different viewpoints; *ii)* force the student to mimic the teacher output starting from a subset $\hat{\mathcal{S}_S} = (\hat{s}_1, \hat{s}_2, \dots, \hat{s}_M) \subset \hat{\mathcal{S}_T}$ with cardinality $M < N$ (in our experiments, $M = 2$ and $N = 8$). The frames in $\hat{\mathcal{S}_S}$ are uniformly sampled from $\hat{\mathcal{S}_T}$ without replacement. This asymmetry between the teacher and the student leads to a self-distillation objective, where the latter can achieve better solutions despite inheriting the same architecture of the former.

To accomplish this, VKD exploits the Knowledge Distillation loss [48]:

$$\mathcal{L}_{\text{KD}} = \tau^2 \, \text{KL}(\boldsymbol{y}_T \parallel \boldsymbol{y}_S) \tag{3.6}$$

where $\boldsymbol{y}_T = \text{softmax}(\boldsymbol{h}_T/\tau)$ and $\boldsymbol{y}_S = \text{softmax}(\boldsymbol{h}_S/\tau)$ are the distributions – smoothed by a temperature τ – we attempt to match[3]. Since the student experiences a different task from the teacher one, Eq. 3.6 resembles the regularisation term imposed by [64] to relieve *catastrophic forgetting*. In a similar vein, we intend to *strengthen* the model in the presence of few images, whilst not *deteriorating* the capabilities it achieved with longer sequences.

In addition to fitting the output distribution of the teacher (Eq. 3.6), our proposal devises additional constraints on the embedding space learnt by the student. In details, VKD encourages the student to mirror the pairwise distances spanned by the teacher. Indicating with $\mathcal{D}_T[i,j] \equiv \mathcal{D}(\mathcal{F}_{\theta_T}(\hat{\mathcal{S}_T}[i]), \mathcal{F}_{\theta_T}(\hat{\mathcal{S}_T}[j]))$ the distance induced by the teacher between the i-th and j-th sets (the same notation $\mathcal{D}_S[i,j]$ also holds for the student),

[3]Since the teacher parameters are fixed, its entropy is constant and the objective of Eq. 3.6 reduces to the cross-entropy between $\boldsymbol{y}_T$ and $\boldsymbol{y}_S$.

VKD seeks to minimise:

$$\mathcal{L}_{\text{DP}} = \sum_{(i,j)\in \binom{B}{2}} (\mathcal{D}_T[i,j] - \mathcal{D}_S[i,j])^2, \tag{3.7}$$

where B equals the batch size. Since the teacher has access to several viewpoints, we argue that distances spanned in its space yield a powerful description of corresponding identities. From the student perspective, distances preservation provides additional semantic knowledge. Therefore, this holds an effective supervision signal, whose optimisation is made more challenging since fewer images are available to the student.
Even though VKD focuses on *self-distillation*, we highlight that both $\mathcal{L}_{\text{KD}}$ and $\mathcal{L}_{\text{DP}}$ allow to match models with different embedding size, which would not be viable under the minimisation performed by [41]. As an example, it is still possible to distill ResNet-101 ($D = 2048$) into MobileNet-V2 [112] ($D = 1280$).

Student optimisation The VKD overall objective combines the distillation terms ($\mathcal{L}_{\text{KD}}$ and $\mathcal{L}_{\text{DP}}$) with the ones optimised by the teacher - $\mathcal{L}_{\text{CE}}$ and $\mathcal{L}_{\text{TR}}$ - that promote higher conditional likelihood w.r.t. ground truth labels. To sum up, VKD aims at strengthening the features of a CNN in Re-ID settings through the following optimisation problem:

$$\underset{\theta_S}{\text{argmin}} \quad \mathcal{L}_{\text{VKD}} \equiv \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{TR}} + \alpha\mathcal{L}_{\text{KD}} + \beta\mathcal{L}_{\text{DP}}, \tag{3.8}$$

where α and β are two hyperparameters balancing the contributions to the total loss $\mathcal{L}_{\text{VKD}}$. We conclude with a final note on the student initialisation: we empirically found beneficial to start from the teacher weights θ_T except for the last convolutional block, which is reinitialised according to the ImageNet pretraining. We argue this represents a good compromise between exploring new configurations and exploiting the abilities already achieved by the teacher.

3.4.3 Experiments

Evaluation Protocols

We indicate the query-gallery matching as x2x, where both x terms are features that can be generated by either a single (I) or multiple frames

(V). In the **Image-to-Image (I2I)** setting features extracted from a query set image are matched against features from individual images in the gallery. This protocol – which has been amply employed for person Re-ID and face recognition – has a light impact in terms of resources footprint. However, a single image captures only a single view of the identity, which may not be enough for identities exhibiting multi-modal distributions. Contrarily, the **Video-to-Video (V2V)** setting enables to capture and combine different modes in the input, but with a significant increase in the number of operations and memory. Finally, the **Image-to-Video (I2V)** setting [162, 163, 72, 146, 74] represents a good compromise: building the gallery may be slow, but it is often performed offline. Moreover, matchings perform extremely fast, as a query comprise only a single image. We remark that *i)* We adopt the standard "*Cross Camera Validation*" protocol, not considering examples of the gallery from the same camera of the query at evaluation and *ii)* even if VKD relies on frames from different camera during training, we strictly adhere to the common schema and switch to tracklet-based inputs at evaluation time.

Evaluation Metrics

While settings vary between different datasets, evaluation metrics for Re-Identification are shared by the vast majority of works in the field. In the followings, we report performance in terms of top-k accuracy and Mean Average Precision (mAP). By combining them, we evaluate VKD both in terms of accuracy and ranking performance.

3.4.4 Datasets

Person Re-ID: **MARS** [158] comprises 19680 tracklets from 6 different cameras, capturing 1260 different identities (split between 625 for the training set, 626 for the gallery and 622 for the query) with 59 frames per tracklet on average. MARS has shown to be a challenging dataset because it has been automatically annotated, leading to errors and false detections [159]. The **Duke** [106] dataset was first introduced for multi-target and multi-camera surveillance purposes, and then expanded to include person attributes and identities (414 ones). Consistently with [41, 120, 68, 87], we use the **Duke-Video-ReID** [147] variant, where identities have

been manually annotated from tracking information[4]. It comprises 5534 video tracklets from 8 different cameras, with 167 frames per tracklet on average. Following [41], we extract the first frame of every tracklet when testing in the I2V setting, for both MARS and Duke.

Vehicle Re-ID: **VeRi-776** [73] has been collected from 20 fixed cameras, capturing vehicles moving on a circular road in a 1.0 km^2 area. It contains 18397 tracklets with an average number of 6 frames per tracklet, capturing 775 identities split between train (575) and gallery (200). The query set shares identities consistently with the gallery, but differently from the other two sets it includes only a single image for each couple (id, camera). Consequently, all recent methods perform the evaluation following the I2V setting.

Animal Re-ID: The **Amur Tiger** [62] Re-Identification in the Wild (ATRW) is a recently introduced dataset collected from a diverse set of wild zoos. The training set includes 107 subjects and 17.6 images on average per identity; no information is provided to aggregate images into tracklets. It is possible to evaluate only the I2I setting through a remote http server. As done in [67], we horizontally flip the training images to duplicate the number of identities available, thus resulting in 214 training identities.

Implementation details

Following [46, 68] we adopt the following hyperparameters for MARS and Duke: *i)* each batch contains $P = 8$ identities with $K = 4$ samples each; *ii)* each sample comprises 8 images equally spaced in a tracklet. Differently, for image-based datasets (ATRW and VeRi-776) we increase P to 18 and use a single image at a time. All the teacher networks are trained for 300 epochs using Adam [55], setting the learning rate to 10^{-4} and multiplying it by 0.1 every 100 epochs. During the distillation stage, we feed $N = 8$ images to the teacher and $M = 2$ ones (picked at random) to the student. We found it beneficial to train the student longer: so, we set the number of epochs to 500 and the learning rate decay steps at 300 and 450. We keep fixed $\tau = 10$ (Eq. 3.6), $\alpha = 10^{-1}$ and $\beta = 10^{-4}$ (Eq. 3.8) in all experiments. To improve generalisation, we apply data augmentation as described in [81]. Finally, we put the teacher in training mode during distillation (consequently, batch

[4]In the following, we refer to Duke-Video-ReID simply as Duke. Another variant of Duke named Duke-ReID exists [107], but it does not come with query tracklets.

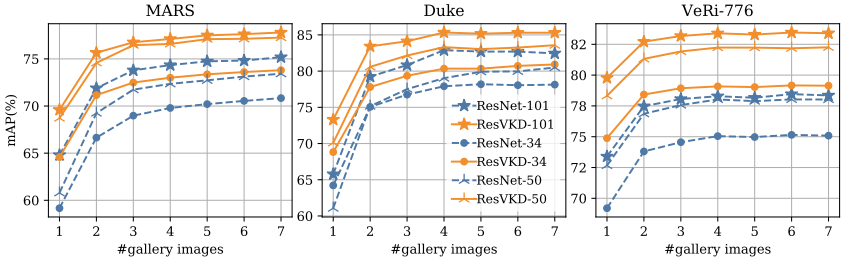


Figure 3.22: Performance (mAP) in the Image-To-Video setting when changing at evaluation time the number of frames in each gallery tracklet.

normalisation [52] statistics are computed on a batch basis): as observed in [8], this provides more accurate teacher labels.

Self-Distillation

In this section we show the benefits of self-distillation for person and vehicle re-id. We indicate the teacher with the name of the backbone (e.g. ResNet-50) and append "VKD" for its student (e.g. ResVKD-50). To validate our ideas, we do not limit the analysis on ResNet-*; contrarily, we test self-distillation on DenseNet-121 [49] and MobileNet-V2 1.0X [112]. Since learning what and where to look represents an appealing property when dealing with Re-ID tasks [34], we additionally conduct experiments on ResNet-50 coupled with Bottleneck Attention Modules [98] (ResNet-50bam).
Table 3.11 reports the comparisons for different backbones: in the vast majority of the settings, *the student outperforms its teacher*. Such a finding is particularly evident when looking at the I2V setting, where the mAP metric gains 4.04% on average. The same holds for the I2I setting on VeRi-776, and in part also on V2V. We draw the following remarks: *i)* in accordance with the objective the student seeks to optimise, our proposal leads to greater improvements when few images are available; *ii)* bridging the gap between I2V and V2V does not imply a significant information loss when more frames are available; on the contrary it sometimes results in superior performance; *iii)* the previous considerations hold true across different architectures. As an additional proof, plots from Figure 3.22

Table 3.11: Self-Distillation results across datasets, settings and architectures.

	MARS				Duke				VeRi-776			
	I2V		V2V		I2V		V2V		I2I		I2V	
	cmc1	mAP	cmc1	mAP	cmc1	mAP	cmc1	mAP	cmc1	mAP	cmc1	mAP
ResNet-34	80.81	70.74	86.67	78.03	81.34	78.70	93.45	**91.88**	92.97	70.30	93.80	75.01
ResVKD-34	**82.17**	**73.68**	**87.83**	**79.50**	**83.33**	**80.60**	**93.73**	91.62	**95.29**	**75.97**	**94.76**	**79.02**
ResNet-50	82.22	73.38	87.88	81.13	82.34	80.19	**95.01**	**94.17**	93.50	73.19	93.33	77.88
ResVKD-50	**83.89**	**77.27**	**88.74**	**82.22**	**85.61**	**83.81**	**95.01**	93.41	**95.23**	**79.17**	**95.17**	**82.16**
ResNet-101	82.78	74.94	88.59	81.66	83.76	82.89	**96.01**	**94.73**	94.28	74.27	94.46	78.20
ResVKD-101	**85.91**	**77.64**	**89.60**	**82.65**	**86.32**	**85.11**	95.44	93.67	**95.53**	**80.62**	**96.07**	**83.26**
ResNet-50bam	82.58	74.11	88.54	81.19	82.48	80.24	94.87	**93.82**	93.33	72.73	93.80	77.14
ResVKD-50bam	**84.34**	**78.13**	**89.39**	**83.07**	**86.18**	**84.54**	**95.16**	93.45	**96.01**	**78.67**	**95.71**	**81.57**
DenseNet-121	82.68	74.34	89.75	81.93	82.91	80.26	93.73	91.73	91.24	69.24	91.84	74.52
DenseVKD-121	**84.04**	**77.09**	**89.80**	**82.84**	**86.47**	**84.14**	**95.44**	**93.54**	**94.34**	**76.23**	**93.80**	**79.76**
MobileNet-V2	78.64	67.94	85.96	77.10	78.06	74.73	93.30	91.56	88.80	64.68	89.81	69.90
MobileVKD-V2	**83.33**	**73.95**	**88.13**	**79.62**	**83.76**	**80.83**	**94.30**	**92.51**	**92.85**	**70.93**	**92.61**	**75.27**

draw a comparison between models before and after distillation. VKD improves metrics considerably on all three datasets, as highlighted by the bias between the teachers and their corresponding students. Surprisingly, this often applies when comparing lighter students with deeper teachers: as an example, ResVKD-34 scores better than even ResNet-101 on VeRi-776, regardless of the number of images sampled for a gallery tracklet.

Comparison with State-Of-The-Art

Image-To-Video Tables 3.12, 3.13 and 3.14 report a thorough comparison with current state-of-the-art (SOTA) methods, on MARS, Duke and VeRi-776 respectively. As common practice [41, 10, 102], we focus our analysis on ResNet-50, and in particular on its distilled variants ResVKD-50 and ResVKD-50bam. Our method clearly outperforms other competitors, with an increase in mAP w.r.t. top-scorers of 6.3% on MARS, 8.6% on Duke and 5% on VeRi-776. This result is totally in line with our goal of conferring robustness when just a single image is provided as query. In doing so, we do not make any task-specific assumption, thus rendering our proposal easily applicable to both person and vehicle Re-ID.

Table 3.12: MARS **I2V**

Method	top_1	top_5	mAP
P2SNet[144]	55.3	72.9	-
Zhang[154]	56.5	70.6	-
XQDA[65]	67.2	81.9	54.9
TKP[41]	75.6	87.6	65.1
STE-NVAN[68]	80.3	-	68.8
NVAN[68]	80.1	-	70.2
MGAT[10]	81.1	92.2	71.8
ResVKD-50	83.9	93.2	77.3
ResVKD-50bam	**84.3**	**93.5**	**78.1**

Table 3.13: Duke **I2V**

Method	top_1	top_5	mAP
STE-NVAN[68]	42.2	-	41.3
TKP[41]	77.9	-	75.9
NVAN[68]	78.4	-	76.7
ResVKD-50	85.6	93.9	83.8
ResVKD-50bam	**86.2**	**94.2**	**84.5**

Video-To-Video Analogously, we conduct experiments on the V2V setting and report results in Table 3.15 (MARS) and Table 3.16 (Duke)[5]. Here, VKD yields the following results: on the one hand, on MARS it pushes a baseline architecture as ResVKD-50 close to NVAN and STE-NVAN [68], the latter being tailored for the V2V setting. Moreover – when exploiting spatial attention modules (ResVKD-50bam) – it establishes new SOTA results, suggesting that a positive transfer occurs when matching tracklets also. On the other hand, the same does not hold true for Duke, where exploiting video features as in STA [34] and NVAN appears rewarding. We leave the investigation of further improvements on V2V to future works. As of today, our proposal is the only one guaranteeing consistent and stable

[5]Since VeRi-776 does not include any tracklet information in the query set, following all other competitors we limit experiments to the I2V setting only.

Table 3.14: VeRi-776 **I2V**

Method	top_1	top_5	mAP
PROVID[74]	76.8	91.4	48.5
VFL-LSTM[4]	88.0	94.6	59.2
RAM[72]	88.6	-	61.5
VANet[21]	89.8	96.0	66.3
PAMTRI[129]	92.9	92.9	71.9
SAN[102]	93.3	97.1	72.5
PROVID-BOT[74]	**96.1**	97.9	77.2
ResVKD-50	95.2	**98.0**	**82.2**
ResVKD-50bam	95.7	98.0	81.6

Table 3.15: MARS **V2V**

Method	top_1	top_5	mAP
DuATN[120]	81.2	92.5	67.7
TKP[41]	84.0	93.7	73.3
CSACSE+OF[20]	86.3	94.7	76.1
STA[34]	86.3	95.7	80.8
STE-NVAN[68]	88.9	-	81.2
NVAN[68]	**90.0**	-	82.8
ResVKD-50	88.7	96.1	82.2
ResVKD-50bam	89.4	**96.8**	**83.1**

results under both I2V and V2V settings.

Analysis of VKD

In the absence of camera information.

Here, we address the setting where we do not have access to camera information. As an example, when dealing with animal re-id this information is often lacking and datasets come with images and labels solely: can VKD still provide any improvement? We think so, as one can still exploit the visual diversity lying in a bag of randomly sampled images. To demonstrate

Table 3.16: Duke **V2V**

Method	top_1	top_5	mAP
DuATN[120]	81.2	92.5	67.7
Matiyali[87]	89.3	98.3	88.5
TKP[41]	94.0	-	91.7
STE-NVAN[68]	95.2	-	93.5
STA[34]	96.2	**99.3**	94.9
NVAN[68]	**96.3**	-	**94.9**
ResVKD-50	95.0	98.9	93.4
ResVKD-50bam	95.2	98.6	93.5

Table 3.17: ATRW **I2I**

Method	top_1	top_5	mAP
PPbM-a [62]	82.5	93.7	62.9
PPbM-b [62]	83.3	93.2	60.3
NWPU [151]	94.7	96.7	75.1
BRL [69]	94.0	96.7	77.0
NBU [67]	**95.6**	**97.9**	**81.6**
ResNet-101	92.3	93.5	75.7
ResVKD-101	92.0	96.4	77.2

our claim, we test our proposal on Amur Tigers re-identification (ATRW), which was conceived as an Image-To-Image dataset. During comparisons: *i)* since other works do not conform to a unique backbone, here we opt for ResNet-101; *ii)* as common practice in this benchmark [67, 69, 151], we leverage re-ranking [161]. Table 3.17 compares VKD against the top scorers in the "Computer Vision for Wildlife Conservation 2019" competition. Importantly, the student ResVKD-101 improves over its teacher (1.5% on mAP and 2.9% on top_5) and places second behind [67], confirming its effectiveness in a challenging scenario. Moreover, we remark that the top-scorer requires additional annotations - such as body parts and pose information - which we do not exploit.

Distilling viewpoints *vs* time.

Figure 3.23 shows results of distilling knowledge from multiple views against time (*i.e.* multiple frames from a tracklet). On one side, as multiple views hold more "*visual variety*", the student builds a more invariant representation for the identity. On the opposite, a student trained with tracklets still considerably outperforms the teacher. This shows that, albeit the visual variety is reduced, our distillation approach still successfully exploits it.

VKD reduces the camera bias.

As pointed out in [130], the appearance encoded by a CNN is heavily affected by external factors surrounding the target object (*e.g.* different backgrounds, viewpoints, illumination ...). In this respect, is our proposal effective for reducing such a bias? To investigate this aspect, we perform a camera classification test on both the teacher (*e.g.* ResNet-34) and the student network (*e.g.* ResVKD-34) by fitting a linear classifier on top of their features, with the aim of predicting the camera the picture is taken from. We freeze all backbone layers and train for 300 epochs ($\text{lr} = 10^{-3}$ and halved every 50 epochs). Table 3.18 reports performance on the gallery set for different teachers and students. To provide a better understanding, we include a baseline that computes predictions by sampling from the cameras' prior distribution. As expected: *i)* the teacher outperforms the baseline, suggesting it is in fact biased towards background conditions; *ii)* the student consistently reduces the bias, confirming VKD encourages the student to focus on identity features and drops viewpoint-specific information. Finally, it is noted that time-based distillation does not yield the bias reduction we observe for VKD.

Can performance of the student be obtained without distillation?

To highlight the advantages of the two-stage procedure above discussed, we here consider a teacher (ResNet-50) trained using few frames ($N = 2$) only. The first two rows of Table 3.19 show the performance achieved by this baseline (using tracklets and views respectively). Results show that major improvements come from the teacher-student paradigm we devise (third row), instead of simply reducing the number of input images available to the teacher.

	MARS	Duke	VeRi-776
Prior Class.	0.19	0.14	0.06
ResNet-34	**0.61**	**0.73**	**0.55**
ResVKD-34	0.40	0.67	0.51
ResNet-101	**0.71**	**0.72**	**0.73**
ResVKD-101	0.51	0.70	0.68

Table 3.18: Analysis on camera bias, in terms of viewpoint classification accuracy.

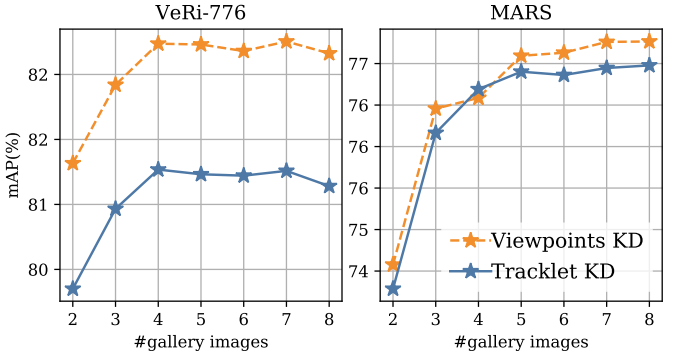


Figure 3.23: Comparison between time and viewpoints distillation.

Student explanation.

To further assess the differences between teachers and students, we leverage GradCam [116] to highlight the input regions that have been considered paramount for predicting the identity. Figure 3.24 depicts the impact of VKD for various examples from MARS, VeRi-776 and ATRW. In general, the student network pays more attention to the subject of interest compared to its teacher. For person and animal Re-ID, background features are suppressed (third and last columns) while attention tends to spread to the whole subject (first and fourth columns). When dealing with vehicle Re-ID, one can appreciate how the attention becomes equally distributed on symmetric parts, such as front and rear lights (second, seventh and last columns).

Table 3.19: Analysis on different modalities for training the teacher.

	Input Bags	MARS				Duke			
		I2V		V2V		I2V		V2V	
		cmc1	mAP	cmc1	mAP	cmc1	mAP	cmc1	mAP
ResNet-50	Viewpoints ($N = 2$)	80.05	71.16	84.70	76.99	77.21	75.19	89.17	87.70
ResNet-50	Tracklets ($N = 2$)	82.32	73.69	87.32	79.91	81.77	80.34	93.73	92.88
ResVKD-50	Viewpoints ($N = 2$)	**83.89**	**77.27**	**88.74**	**82.22**	**85.61**	**83.81**	**95.01**	**93.41**

Table 3.20: Ablation study questioning the impact of each loss term.

	$\mathcal{L}_{\text{CE}}$	$\mathcal{L}_{\text{TR}}$	$\mathcal{L}_{\text{KL}}$	$\mathcal{L}_{\text{DP}}$	MARS				Duke				VeRi-776			
					I2V		V2V		I2V		V2V		I2I		I2V	
					cmc1	mAP	cmc1	mAP	cmc1	mAP	cmc1	mAP	cmc1	mAP	cmc1	mAP
ResNet-50 (teacher)					82.22	73.38	87.88	81.13	82.34	80.19	95.01	94.17	93.50	73.19	93.33	77.88
ResVKD-50 (students)	✓	✓	✗	✗	80.25	71.26	85.71	77.45	82.62	81.03	94.73	93.29	92.61	70.06	92.31	74.82
	✗	✗	✓	✓	84.09	**77.37**	88.33	82.06	84.90	83.56	95.30	93.79	95.29	**79.35**	**95.29**	**82.26**
	✓	✓	✓	✗	83.54	75.18	88.43	80.77	83.90	82.34	94.30	92.97	**95.41**	78.01	95.17	81.32
	✓	✓	✗	✓	**84.29**	76.82	88.69	81.82	85.33	83.45	**95.44**	**93.90**	94.40	77.41	94.87	80.93
	✓	✓	✓	✓	83.89	77.27	**88.74**	**82.22**	**85.61**	**83.81**	95.01	93.41	95.23	79.17	95.17	82.16

Cross-Distillation.

Differently from other approaches [13, 41], VKD is not confined to self-distillation, but instead allows the knowledge transfer from a complex architecture (e.g. ResNet-101) into a simpler one, such as MobileNet-V2 or ResNet-34 (*cross-distillation*). Here, drawing inspirations from the model compression area, we attempt to reduce the network complexity but, at the same time, increase the profit we already achieve through self-distillation. In this respect, Table 3.21 shows results of cross-distillation, for various combinations of a teacher and a student. It appears that *better the teacher, better the student*: as an example, ResVKD-34 gains an additional 3% mAP on Duke when educated by ResNet-101 rather than "itself".

On the impact of loss terms.

We perform a thorough ablation study (Table 3.20) on the student loss (Eq. 3.8). It is noted that leveraging ground truth solely (second row) hurts performance. Differently, best performance for both metrics are obtained exploiting teacher signal (from the third row onward), with particular emphasis to $\mathcal{L}_{\text{DP}}$, which proves to be a fundamental component.

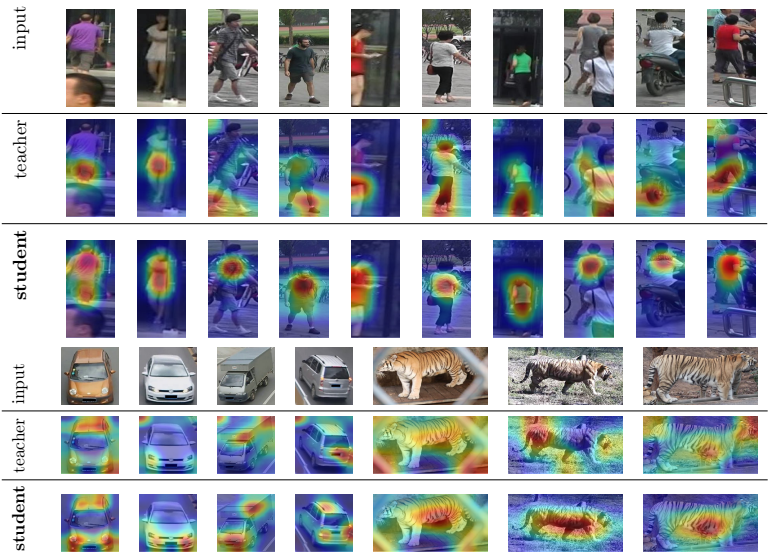


Figure 3.24: Model explanation via GradCam[116] on ResNet-50 (teacher) and ResVKD-50 (student). The student favours visual details characterising the target, discarding external and uninformative patterns.

Distilling viewpoints *vs* time: impact on camera bias.

As previously discussed, limiting the teacher-student transfer to the temporal axis does not explicitly encourage invariance and robustness to different viewpoints. This can be investigated with the same approach presented above. To further prove such a claim, we again measure the camera bias lying in high-level features, in the same manner as described before. This time, though, we focus on a student accessing fewer frames from the same tracklet, thus being educated to capture time information solely. Table 3.22 compares this strategy (third row) with our proposal (fourth row), which instead forces the transfer at viewpoint level. As expected: *i)* time-based distillation performs similarly to the teacher, confirming its poor ability to confer robustness to shifts in background appearance; *ii)* as advocated by our work, a student shows a lower camera bias when trained on different viewpoints instead of using temporal information only.

Table 3.21: Measuring the benefit of VKD for cross-architecture transfer.

Student (#params)	Teacher (#params)	MARS		Duke		VeRi-776	
		I2V		I2V		I2V	
		cmc1	mAP	cmc1	mAP	cmc1	mAP
ResNet-34 (21.2M)	ResNet-34 (21.2M)	82.17	73.68	83.33	80.60	94.76	79.02
	ResNet-50 (23.5M)	83.08	75.45	84.05	82.61	**95.05**	80.05
	ResNet-101 (42.5M)	**83.43**	**75.47**	**85.75**	**83.65**	94.87	**80.41**
ResNet-50 (23.5M)	ResNet-50 (23.5M)	83.89	77.27	85.61	83.81	95.17	82.16
	ResNet-101 (42.5M)	**84.49**	**77.47**	**85.90**	**84.34**	**95.41**	**82.99**
MobileNet-V2 (2.2M)	MobileNet-V2 (2.2M)	83.33	73.95	**83.76**	80.83	92.61	75.27
	ResNet-101 (42.5M)	**83.38**	**74.72**	**83.76**	**81.36**	**93.03**	**76.38**

Student explanation - other examples.

Previously, we investigated which regions the student focuses on, showing that it pays higher attention to foreground details when compared to its teacher. We observe that this happens systematically, especially when dealing with person Re-ID. Figure 3.25 reports additional comparisons between the explanations provided by the teacher and its student on Duke-Video-ReID [147].

Error Analysis

We provide here some visual examples of the errors of our method and try to investigate their nature. With reference to the Video-To-Video setting on MARS [158], our model (ResVKD-50) misidentifies 223 out of 1980 top-1 matchings. From an analysis computed on top of these 223 cases,

Table 3.22: Analysis on camera bias – in terms of viewpoint classification accuracy – for different methods. We indicate with "ResTKD-50" a student restricted to time information solely.

	MARS	Duke
Prior Class.	0.19	0.14
ResNet-50 (teacher)	**0.74**	**0.76**
ResTKD-50 (time-based distillation)	0.69	**0.76**
ResVKD-50 (viewpoints-based distillation)	0.49	0.69

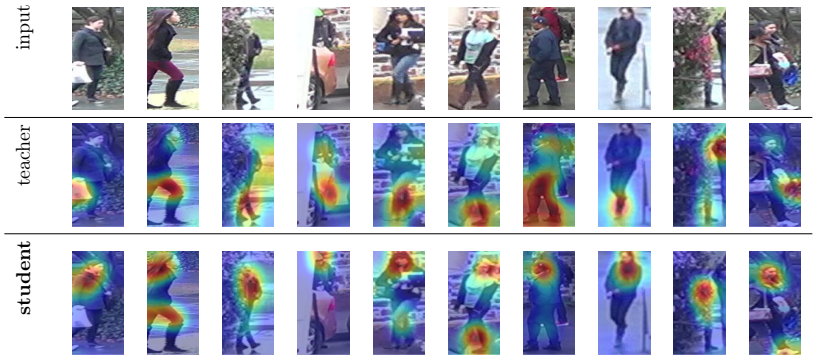


Figure 3.25: Model explanation (Duke-Video-ReID) on ResNet-50 (teacher) and ResVKD-50 (student).

we identify four different categories of errors. We also asked two external researchers to annotate the errors according to these four classes as follows:

- **True errors**: the network associates the query to a wrong identity from the gallery set (Figure 3.26a). This often happens when similar clothes and appearances between the two identities fool the network. Out of 223, 103 (**46.2%**) were identified as true errors;

- **Wrong ID Annotations**: the ground truth indicates that the network associates the query to a wrong identity from the gallery set. However – for a limited set of queries – this does not hold true when visually inspecting the gallery identity. This is due to annotation errors, probably caused by a drift in the tracker (Figure 3.26b). Out of 223, 29 (**13.0%**) were identified as true errors;

- **Couples of People**: some crops depict more than one subject (*e.g.* two) but only one can be associated with the tracklet id (Figure 3.26c). Out of 223, 37 (**16.6%**) were identified as errors involving frames with more than one person;

- **Misleading Distractors**: cases in which the subject has been correctly identified, but the gallery tracklet was erroneously indicated as a distractor. Again, because this set has not been manually checked,

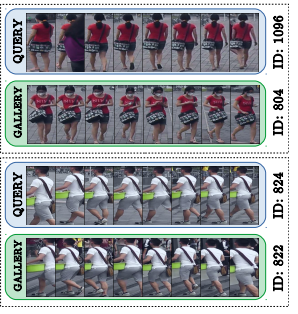
some distractors are valid as they depict people (Figure 3.26d). Out of 223, 54 (**24.2%**) were identified as misleading distractors;

It is worth noting that the presence of the last three types of errors places a limit on the maximum score a method can obtain.

3.4.5 Bridging the domain gap

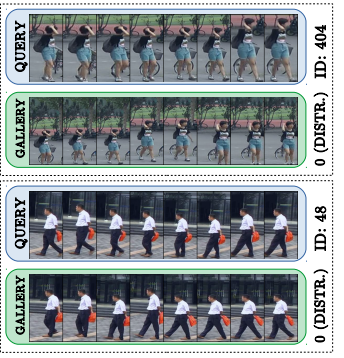
This research was inspired by the results obtained on cattle's re-id. However, it has evolved to include a strong distillation technique which could be used also in the original domain. An approach similar to the one presented here could be applied on cattle re-id to further improve the performance of the model presented in Section3.3. Furthermore, this approach overcomes the limit of using a pair of view and can be applied to the more general setting of a set. Again, this feature is not limited to the domains presented in this chapter but it can instead generalise to others, provided enough images per identity are available.

(a) True Errors.



(b) Wrong ID Annotations.

(c) Couples of People.



(d) Misleading Distractors.

Figure 3.26: Different categories of errors on MARS. While almost half of them can be attributed to our method misidentifying between similar appearances (a), the other half are due to the automatic annotation process. In particular, wrong annotation caused by tracking drift (b), more than one identity in the same tracklet (c) and misleading distractors (d).

Chapter 4

Conclusions and Future Works

In this thesis, we aimed to apply Deep Learning approaches and techniques to different topics in livestock farming and the food production chain. In general, when enough data is available (or is available to be collected) research has been proved fruitful. While the research impact of this work is noticeable, the industrial application still lags. This is due to both the long manual data collection time and the lack of expertise in bringing the developed solution into a production environment. Still, the house's foundations are now well in place, which opens the doors to a set of potential successful and remunerative applications. In the following, we include a list of possible future works, grouped by topic. Some of them have a pronounced academia scope, while others are more practical and industry oriented.

4.1 Pigs Detection and Tracking

We present a detection-tracking-behaviour pipeline for long-term behaviour changes of individual pigs in an indoor pen. This analysis is powered by our new large pig dataset which includes annotations for various tasks. The conclusions drawn from the aggregated data match the expectations of experts and prove our claim that collective behaviour statistics are accurate, even though individual frame-level labels may not always be as accurate.

This is valid not only for actions performed frequently (e.g. lying) but also for those occurring less often (e.g. eating or drinking). Future improvements can be envisioned for this task. On the one hand, single components (e.g. the detection algorithm) could be specialised for the setting. On the other hand, given that the errors in the different stages of the pipeline compound, a single end-to-end method for detection-tracking-behaviour is also a possible future outcome. Another direction for extensions is increasing the number or breakdown of the behaviour classifications. An even more challenging future improvement is the analysis of collective behaviours. This requires high accuracy on all the remaining pieces of the system, to avoid errors compounding in this final stage. Still, these behaviours are of great interest. On the industry side, the environment poses some unique challenges to the integration of the system. Vibrations, light changes and dust can all lower the quality of the incoming video stream. This can either be solved on the hardware side (adaptive lens and stabilisation systems), on the software side (equalisation and stabilisation algorithms) or by the DL methods themselves (e.g. through data augmentation or synthetic data generation). The final system could be easily scaled to multiple pens of a farm to give an overall snapshot of the animals' welfare.

4.2 Lesion Scoring

We present a large visual dataset for the pixel-level segmentation of swine carcasses. This dataset includes 4444 images annotated by domain experts with four anatomical and two anomalous structures. Furthermore, we show how these annotations can be leveraged to tackle the *post-mortem* diagnosis and scoring of pleurisy. Our segmentation-guided method enables much higher interpretability of the output and outperforms competitive baselines on an independent test set by a large margin.

Future works on the research side could exploit the same dataset for the analysis of other diseases of pigs and focus on applying the same methodology to pathological conditions affecting other anatomical structures in swine and other livestock. More recent approaches (e.g. based on the popular Transformer architecture) could also be considered to reduce the inference time and improve performance.

On the industry side, a production-ready system for lung lesions scoring is being developed and will be installed in a slaughterhouse in the upcoming

future. This poses several new challenges. An average slaughterhouse can process thousands of carcasses per day. This means that not only the acquisition and processing components must cope with this speed, but also the reports must be produced with a daily schedule at least. To this end, recent development in deep network compression and quantisation could be used to drastically reduce the inference time, without sacrificing too much on performance.

4.3 Cattle Re-Identification

We propose a Deep Learning base method for cattle re-identification in unconstrained environments from single and multi-views. We present extended baseline comparisons both with non-deep and deep methods. We show that human and cattle re-identification are slightly similar tasks, but present important and significant differences. Finally, we highlight how a multi-views method (i.e. a method combining information from multiple profiles) clearly outperforms both baseline and single-view methods.

On the research side, the proposed method can be improved to work with a variable number of views by design. This would enable it to scale natively to video streams (which is a desirable feature for a production-ready system). Again, segmentation could be introduced to remove background noise or to make the network more robust to the background's variations. On the industry side, quantisation techniques could be considered to deliver and scale the system. When dealing with multiple concurrent requests, the number of operations in the network could become a bottleneck. Along with quantisation, the system could be rendered easy to be deployed in a cluster environment by leveraging containerisation technologies (e.g. Docker). The deployed application could then run on mobile devices by sending requests (either pictures or videos) to a cluster of GPU-powered nodes to be processed. The proposed network would be employed to retrieve the identity of the animal and provide the client with information about it.

4.4 People and Vehicles Re-Identification

An effective Re-ID method requires visual descriptors robust to changes in both background appearances and viewpoints. Moreover, its effectiveness should be ensured even for queries composed of a single image.

To accomplish these, we propose Views Knowledge Distillation (VKD), a teacher-student approach where the student observes only a small subset of input views. This strategy encourages the student to discover better representations: as a result, it outperforms its teacher at the end of the training. Importantly, VKD shows robustness on diverse domains (person, vehicle and animal), surpassing by a wide margin the state of the art in I2V. Thanks to extensive analysis, we highlight that the student presents a stronger focus on the target and reduces the camera bias.
The major future work here is to export the teacher-student approach to new topics. This method is flexible enough to be applied to completely different topics. The only request is that the input can be degraded to a subset of itself to provide a lesser version for the student. An example could be time-series, where the student can learn to mimic the original series starting from subsampled ones. On the industry side, testing this technology in a real-world setting presents numerous privacy challenges when applied to people (and vehicles, since plate numbers can often be associated with individuals).

Acknowledgements

This thesis has been made possible thanks to the help and contributions of many people. We do our best to acknowledge them all here.

The author would like to acknowledge Simone, who has been the best supervisor one could have asked for. This is not a farewell.

The author would like to acknowledge Stefano, Andrea, Angelo, Davide and every other colleague from the AImageLab.

The author would like to acknowledge Farm4Trade for its financial and technical support through these past 3 years. In particular, the author would like to acknowledge Andrea for his tireless striving for success. I'm sure it will be rewarded.

The author would like to acknowledge Allevamento Martin, Allevamento Costantini, Cooperativa Venditti, Allevamento Hombre, Malga Zeledria, Malga Boch and Malga Ritort for providing access to their farms during

the dataset acquisition. Moreover, the author would like to acknowledge Francesco di Tondo, Lisa Leonzi, Giuseppe Carolla and Carmen Sabia for their precious help with data acquisition.

The author would like to acknowledge Rosie, Jasmine and Giuseppe from UNITE for their contribution on pleural lesion scoring.

The author would like to acknowledge Bob, who kept collecting data from the farm regardless of the smell.

The author would like to acknowledge the SRUC technician Mhairi Jack, and farm staff Peter Finnie and Phil O'Neill. SRUC's contribution to the work on pig detection and tracking was funded by the Rural and Environment Science and Analytical Services Division of the Scottish Government.

Finally, the author would like to acknowledge his family, who has been a constant support throughout this PhD.

Appendix A

List of publications

In this section we briefly report the research papers published during the PhD period (including preprint if proceeding not available).
Content and experimental results published in some of these papers has been included, even *verbatim*, in the previous chapters.

- Palazzi, A., Bergamini, L., Calderara, S. and Cucchiara, R., 2018. End-to-end 6-DoF Object Pose Estimation through Differentiable Rasterization. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 0-0).

- Antonio Marin-Reyes, P., Palazzi, A., Bergamini, L., Calderara, S., Lorenzo-Navarro, J. and Cucchiara, R., 2018. Unsupervised vehicle re-identification using triplet networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (pp. 166-171).

- Bergamini, L., Porrello, A., Dondona, A. C., Del Negro, E., Mattioli, M., D'alterio, N., and Calderara, S. (2018, November). Multi-views embedding for cattle re-identification. In 2018 14th International Conference on Signal-Image Technology and Internet-Based Systems (SITIS) (pp. 184-191). IEEE.

- Bergamini, L., Sposato, M., Pellicciari, M., Peruzzini, M., Calderara, S., and Schmidt, J. (2020). Deep learning-based method for vision-

guided robotic grasping of unknown objects. Advanced Engineering Informatics, 44, 101052.

- Palazzi, A., Bergamini, L., Calderara, S. and Cucchiara, R., 2019. Warp and Learn: Novel Views Generation for Vehicles and Other Objects, arXiv preprint

- Ferrari, A., Bergamini, L., Guerzoni, G., Calderara, S., Bicocchi, N., Vitetta, G., ... and Ferrari, A. (2019). Gait-Based Diplegia Classification Using LSMT Networks. Journal of healthcare engineering, 2019.

- Houston, J., Zuidhof, G., Bergamini, L., Ye, Y., Jain, A., Omari, S., ... and Ondruska, P. (2020). One Thousand and One Hours: Self-driving Motion Prediction Dataset. arXiv preprint arXiv:2006.14480.

- Bergamini, L., Calderara, S., Bicocchi, N., Ferrari, A., and Vitetta, G. (2017, September). Signal Processing and Machine Learning for Diplegia Classification. In International Conference on Image Analysis and Processing (pp. 97-108). Springer, Cham.

- Trachtman, A. R., Bergamini, L., Palazzi, A., Porrello, A., Capobianco Dondona, A., Del Negro, E., ... and Marruchella, G. (2020). Scoring pleurisy in slaughtered pigs using convolutional neural networks. Veterinary Research, 51, 1-9.

- Porrello, A., Bergamini, L., and Calderara, S. (2020). Robust Re-Identification by Multiple Views Knowledge Distillation. arXiv preprint arXiv:2007.04174. Simoni, A., Bergamini, L., Palazzi, A., Calderara, S., and Cucchiara, R. (2020). Future Urban Scenes Generation Through Vehicles Synthesis. arXiv preprint arXiv:2007.00323.

- Bergamini, L., Trachtman, A.R., Palazzi, A., Del Negro, E., Dondona, A.C., Marruchella, G. and Calderara, S., 2019, September. Segmentation Guided Scoring of Pathological Lesions in Swine Through CNNs. In International Conference on Image Analysis and Processing (pp. 352-360). Springer, Cham.

Appendix B

Activities carried out during the PhD

Here we report research activites carried out during the 3 years of PhD.

Foreign collaborations

- Research Internship at Bayes Centre, the University of Edinburgh (UK), April - October 2019.
- Research internship at Lyft L5. London (UK), March - September 2020.

Conferences, courses, seminars attended

Conferences

- International Conference on Computer Vision - ICCV, Venice, 2017
- International Conference on Signal-Image Technology and Internet-Based Systems - SITIS, Las Palmas, 2018
- European Conference on Computer Vision - ECCV, Glasgow, 2020

Courses and seminars

- Algoritmi Avanzati - Dr. Mauro Leoncini - September 2017
- Security and quantistic technology: potential uses and risks - Dr. Enrico Prati, CNR - November 21st, 2017.
- Deep learning technologies: from hardware components to vertical frameworks - Dr. Piero Altoè, NVIDIA - November 29th, 2017.
- Deep Learning for Fault Prediction - Prof. Roberto Paredes Palacios - February 2018
- Regularization Methods for Machine Learning Summer School (RegML) - Prof. Lorenzo Rosasco - June 2018
- Deep Learning for Fault Prediction - Prof. Roberto Paredes Palacios - February 2018
- Academic English Workshop I - Dr. Silvia Cavalieri - February 2019
- Academic English Workshop II - Dr. Silvia Cavalieri - February 2018

Bibliography

[1] Cattle fao world statistics. http://www.fao.org/faostat/en/. Accessed: 2018-09-316. 47

[2] Cattle theft statistics. https://www.nytimes.com/2013/05/27/world/asia/cow-thefts-on-the-rise-in-india.html. Accessed: 2018-09-316. 46

[3] Dairy cattle in europe statistics. https://dairy.ahdb.org.uk/market-information/farming-data/cow-numbers/eu-cow-numbers/#.W6aLCRyxVhF. Accessed: 2018-09-316. 47

[4] Saghir Ahmed Saghir Alfasly, Yongjian Hu, Tiancai Liang, Xiaofeng Jin, Qingli Zhao, and Beibei Liu. Variational representation learning for vehicle re-identification. In *IEEE International Conference on Image Processing*, 2019. 71

[5] William Andrew, Colin Greatwood, and Tilo Burghardt. Visual localisation and individual identification of holstein friesian cattle via deep learning. In *Proc. IEEE International Conference on Computer Vision (ICCV), Venice, Italy*, pages 22–29, 2017. 9

[6] Mykhaylo Andriluka, Stefan Roth, and Bernt Schiele. People-tracking-by-detection and people-detection-by-tracking. In *2008 IEEE Conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2008. 16

[7] Bilal Aslam, Wei Wang, Muhammad Imran Arshad, Mohsin Khurshid, Saima Muzammil, Muhammad Hidayat Rasool, Muhammad Atif Nisar, Ruman Farooq Alvi, Muhammad Aamir

Aslam, Muhammad Usman Qamar, et al. Antibiotic resistance: a rundown of a global crisis. *Infection and Drug Resistance*, 11:1645, 2018. 31

[8] Hessam Bagherinezhad, Maxwell Horton, Mohammad Rastegari, and Ali Farhadi. Label refinery: Improving imagenet classification through label progression. *arXiv preprint arXiv:1805.02641*, 2018. 68

[9] Chun-tong Bao, Jia-meng Xiao, Bai-jun Liu, Jian-fang Liu, Ri-ning Zhu, Peng Jiang, Lei Li, Paul Richard Langford, and Lian-cheng Lei. Establishment and comparison of actinobacillus pleuropneumoniae experimental infection model in mice and piglets. *Microbial pathogenesis*, 2019. 31

[10] Liqiang Bao, Bingpeng Ma, Hong Chang, and Xilin Chen. Masked graph attention network for person re-identification. In *IEEE International Conference on Computer Vision and Pattern Recognition Workshops*, 2019. 11, 69, 70

[11] Peter N Belhumeur, João P Hespanha, and David J Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. Technical report, Yale University New Haven United States, 1997. 10, 54

[12] Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: the clear mot metrics. *EURASIP Journal on Image and Video Processing*, 2008:1–10, 2008. 22

[13] Shweta Bhardwaj, Mukundhan Srinivasan, and Mitesh M Khapra. Efficient video classification using fewer frames. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2019. 12, 60, 75

[14] Dinesh Bolkensteyn. Vaticjs. `https://github.com/dbolkensteyn/vatic.js`, 2016. 18

[15] David S Bolme, J Ross Beveridge, Bruce A Draper, and Yui Man Lui. Visual object tracking using adaptive correlation filters. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 2544–2550. IEEE, 2010. 7, 22

[16] MB Bowling, DL Pendell, DL Morris, Y Yoon, K Katoh, KE Belk, and GC Smith. Identification and traceability of cattle in selected countries outside of north america. 2008. 47

[17] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000. 54

[18] Johannes Brünger, Maria Gentz, Imke Traulsen, and Reinhard Koch. Panoptic instance segmentation on pigs. *arXiv preprint arXiv:2005.10499*, 2020. 7, 17, 20, 24

[19] Chen Chen, Weixing Zhu, Changhua Ma, Yizheng Guo, Weijia Huang, and Chengzhi Ruan. Image motion feature extraction for recognition of aggressive behaviors among group-housed pigs. *Computers and Electronics in Agriculture*, 142:380–387, 2017. 15

[20] Dapeng Chen, Hongsheng Li, Tong Xiao, Shuai Yi, and Xiaogang Wang. Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2018. 71

[21] Ruihang Chu, Yifan Sun, Yadong Li, Zheng Liu, Chi Zhang, and Yichen Wei. Vehicle re-identification with viewpoint-aware metric learning. In *IEEE International Conference on Computer Vision*, 2019. 11, 71

[22] Yongwha Chung, Seunggeun Oh, Jonguk Lee, Daihee Park, Hong-Hee Chang, and Suk Kim. Automatic detection and recognition of pig wasting diseases using sound data in audio surveillance systems. *Sensors*, 13(10):12929–12942, 2013. 33

[23] Annamaria Costa, Gunel Ismayilova, Federica Borgonovo, Stefano Viazzi, Daniel Berckmans, and Marcella Guarino. Image-processing technique to measure pig activity in response to climatic variation in a pig barn. *Animal Production Science*, 54(8):1075–1083, 2014. 15

[24] Jake Cowton, Ilias Kyriazakis, and Jaume Bacardit. Automated individual pig localisation, tracking and behaviour metric extraction using deep learning. *IEEE Access*, 7:108049–108060, 2019. 7, 16, 17

[25] David Crouse, Rachel L Jacobs, Zach Richardson, Scott Klum, Anil Jain, Andrea L Baden, and Stacey R Tecot. Lemurfaceid: a face recognition system to facilitate individual identification of lemurs. *BMC Zoology*, 2(1):2, 2017. 8, 47

[26] Debayan Deb, Susan Wiper, Alexandra Russo, Sixue Gong, Yichun Shi, Cori Tymoszek, and Anil Jain. Face recognition: Primates in the wild. *arXiv preprint arXiv:1804.08790*, 2018. 8, 9, 47, 52

[27] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 19, 35

[28] Alessandro dos Santos Ferreira, Daniel Matte Freitas, Gercina Gonçalves da Silva, Hemerson Pistori, and Marcelo Theophilo Folhes. Weed detection in soybean crops using convnets. *Computers and Electronics in Agriculture*, 143:314–324, 2017. 6

[29] Richard B D'Eath, Mhairi Jack, Agnieszka Futro, Darren Talbot, Qiming Zhu, David Barclay, and Emma M Baxter. Automatic early warning of tail biting in pigs: 3d cameras can detect lowered tail posture before an outbreak. *PloS one*, 13(4):e0194524, 2018. 15

[30] John Evans. Livestock identification. 2005. 47

[31] Adnan Farooq, Jiankun Hu, and Xiuping Jia. Analysis of spectral bands and spatial resolutions for weed classification via deep convolutional neural network. *IEEE Geoscience and Remote Sensing Letters*, 16(2):183–187, 2018. 6

[32] Eduardo Fernández-Carrión, Marta Martínez-Avilés, Benjamin Ivorra, Beatriz Martínez-López, Ángel Manuel Ramos, and José Manuel Sánchez-Vizcaíno. Motion-based video monitoring for early detection of livestock diseases: The case of african swine fever. *PloS one*, 12(9):e0183793, 2017. 15

[33] Francois Fleuret, Jerome Berclaz, Richard Lengagne, and Pascal Fua. Multicamera people tracking with a probabilistic occupancy map. *IEEE transactions on pattern analysis and machine intelligence*, 30(2):267–282, 2007. 16

[34] Yang Fu, Xiaoyang Wang, Yunchao Wei, and Thomas Huang. Sta: Spatial-temporal attention for large-scale video-based person re-identification. In *AAAI Conference on Artificial Intelligence*, 2019. 63, 68, 70, 71, 72

[35] Alvaro Fuentes, Sook Yoon, Sang Cheol Kim, and Dong Sun Park. A robust deep-learning-based detector for real-time tomato plant diseases and pests recognition. *Sensors*, 17(9):2022, 2017. 5

[36] Tommaso Furlanello, Zachary C Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born again neural networks. *International Conference on Machine Learning*, 2018. 11

[37] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014. 7

[38] Karina Bech Gleerup, Pia Haubro Andersen, Lene Munksgaard, and Björn Forkman. Pain evaluation in dairy cattle. *Applied Animal Behaviour Science*, 171:25–32, 2015. 48

[39] Michael T Gorczyca, Hugo Fernando Maia Milan, Alex Sandro Campos Maia, and Kifle G Gebremedhin. Machine learning algorithms to predict core, skin, and hair-coat temperatures of piglets. *Computers and Electronics in Agriculture*, 151:286–294, 2018. 6

[40] M Gottschalk. *Diseases of Swine (10th edition)*, volume 2012, pages 653–669. Wiley-Blackwell Oxford, UK, 2012. 31

[41] Xinqian Gu, Bingpeng Ma, Hong Chang, Shiguang Shan, and Xilin Chen. Temporal knowledge propagation for image-to-video person re-identification. In *IEEE International Conference on Computer Vision*, 2019. 12, 59, 60, 62, 65, 66, 67, 69, 70, 71, 72, 75

[42] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European Conference on Computer Vision*, pages 87–102. Springer, 2016. 10

[43] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 8

[44] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2016. 8, 24, 35, 50, 62

[45] Kai Heinrich, Andreas Roth, Lukas Breithaupt, Björn Möller, and Johannes Maresch. Yield prognosis for the agrarian management of vineyards using deep learning for object counting. 2019. 6

[46] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 63, 67

[47] Lex Hiby, Phil Lovell, Narendra Patil, N Samba Kumar, Arjun M Gopalaswamy, and K Ullas Karanth. A tiger cannot change its stripes: using a three-dimensional model to match images of living tigers and tiger skins. *Biology letters*, pages rsbl–2009, 2009. 9

[48] Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. In *NeurIPS Deep Learning and Representation Learning Workshop*, 2015. 11, 64

[49] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2017. 68

[50] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007. 10

[51] Xiaoping Huang, Zelin Hu, Xiaorun Wang, Xuanjiang Yang, Jian Zhang, and Daoling Shi. An improved single shot multibox detector method applied in body condition score for dairy cows. *Animals*, 9(7):470, 2019. 6

[52] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, 2015. 68

[53] Andreas Kamilaris and Francesc X Prenafeta-Boldú. Deep learning in agriculture: A survey. *Computers and electronics in agriculture*, 147:70–90, 2018. 5, 6

[54] Sultan Daud Khan and Habib Ullah. A survey of advances in vision-based vehicle re-identification. *Computer Vision and Image Understanding*, 2019. 59

[55] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 2015. 37, 67

[56] A Koirala, KB Walsh, Z Wang, and C McCarthy. Deep learning for real-time fruit detection and orchard fruit load estimation: Benchmarking of 'mangoyolo'. *Precision Agriculture*, 20(6):1107–1135, 2019. 6

[57] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955. 22

[58] Mayank Lahiri, Chayant Tantipathananandh, Rosemary Warungu, Daniel I Rubenstein, and Tanya Y Berger-Wolf. Biometric animal databases from field photographs: identification of individual zebra in the wild. In *Proceedings of the 1st ACM international conference on multimedia retrieval*, page 6. ACM, 2011. 9

[59] Mykola Lavreniuk, Nataliia Kussul, and Alexei Novikov. Deep learning crop classification approach based on coding input satellite data into the unified hyperspace. In *2018 IEEE 38th International Conference on Electronics and Nanotechnology (ELNANO)*, pages 239–244. IEEE, 2018. 6

[60] Dan Li, Yifei Chen, Kaifeng Zhang, and Zhenbo Li. Mounting behaviour recognition for pigs based on deep learning. *Sensors*, 19(22):4924, 2019. 8, 16, 17

[61] Dan Li, Kaifeng Zhang, Zhenbo Li, and Yifei Chen. A spatiotemporal convolutional network for multi-behavior recognition of pigs. *Sensors*, 20(8):2381, 2020. 8, 16, 17

[62] Shuyuan Li, Jianguo Li, Weiyao Lin, and Hanlin Tang. Amur tiger re-identification in the wild. *arXiv preprint arXiv:1906.05586*, 2019. 67, 72

[63] Xiangyuan Li, Cheng Cai, Ruifei Zhang, Lie Ju, and Jinrong He. Deep cascaded convolutional models for cattle pose estimation. *Computers and Electronics in Agriculture*, 164:104885, 2019. 6

[64] Zhizhong Li and Derek Hoiem. Learning without forgetting. In *European Conference on Computer Vision*, 2016. 64

[65] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li. Person re-identification by local maximal occurrence representation and metric learning. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2015. 70

[66] Yen Khye Lim, Zukang Liao, Stavros Petridis, and Maja Pantic. Transfer learning for action unit recognition. *arXiv preprint arXiv:1807.07556*, 2018. 35

[67] Cen Liu, Rong Zhang, and Lijun Guo. Part-pose guided amur tiger re-identification. In *IEEE International Conference on Computer Vision Workshops*, 2019. 16, 67, 72

[68] Chih-Ting Liu, Chih-Wei Wu, Yu-Chiang Frank Wang, and Shao-Yi Chien. Spatially and temporally efficient non-local attention network for video-based person re-identification. In *British Machine Vision Conference*, 2019. 63, 66, 67, 70, 71, 72

[69] Ning Liu, Qijun Zhao, Nan Zhang, Xinhua Cheng, and Jianing Zhu. Pose-guided complementary features learning for amur tiger re-identification. In *IEEE International Conference on Computer Vision Workshops*, 2019. 16, 72

[70] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016. 7

[71] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, page 1, 2017. 10, 52, 54

[72] Xiaobin Liu, Shiliang Zhang, Qingming Huang, and Wen Gao. Ram: a region-aware deep model for vehicle re-identification. In *IEEE International Conference on Multimedia and Expo (ICME)*, 2018. 11, 66, 71

[73] Xinchen Liu, Wu Liu, Tao Mei, and Huadong Ma. A deep learning-based approach to progressive vehicle re-identification for urban surveillance. In *European Conference on Computer Vision*, 2016. 60, 61, 64, 67

[74] Xinchen Liu, Wu Liu, Tao Mei, and Huadong Ma. Provid: Progressive and multimodal vehicle reidentification for large-scale urban surveillance. *IEEE Transactions on Multimedia*, 2017. 11, 66, 71

[75] Xu Liu, Steven W Chen, Shreyas Aditya, Nivedha Sivakumar, Sandeep Dcunha, Chao Qu, Camillo J Taylor, Jnaneshwar Das, and Vijay Kumar. Robust fruit counting: Combining deep learning, tracking, and structure from motion. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1045–1052. IEEE, 2018. 6

[76] Yu Liu, Yan Junjie, and Wanli Ouyang. Quality aware network for set to set recognition. In *IEEE International Conference on Computer Vision*, 2017. 63

[77] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015. 10

[78] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 36

[79] David G Lowe. Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE*

international conference on, volume 2, pages 1150–1157. Ieee, 1999. 10

[80] Yang Lu, Shujuan Yi, Nianyin Zeng, Yurong Liu, and Yong Zhang. Identification of rice diseases using deep convolutional neural networks. *Neurocomputing*, 267:378–384, 2017. 5

[81] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *IEEE International Conference on Computer Vision and Pattern Recognition Workshops*, 2019. 62, 63, 67

[82] Juncheng Ma, Keming Du, Feixiang Zheng, Lingxian Zhang, Zhihong Gong, and Zhongfu Sun. A recognition method for cucumber diseases using leaf symptom images based on deep convolutional neural network. *Computers and electronics in agriculture*, 154:18–24, 2018. 5

[83] Paolo Stefano Marcato. *Patologia Suina: Testo E Atlante; a Colour Atlas of Pathology of the Pig*. Edagricole, 1998. 32

[84] Giuseppe Marruchella, Michael Odintzov Vaintrub, Andrea Di Provvido, Elena Farina, Giorgio Fragassi, and Giorgio Vignola. Alternative scoring method of pleurisy in slaughtered pigs: Preliminary investigations. In *Proceedings of SIPAS*, pages 375–380, 2018. 32, 33, 36

[85] Giuseppe Marruchella, Michael Odintzov Vaintrub, Andrea Di Provvido, Elena Farina, and Giorgio Vignola. Scoring pleurisy in slaughtered pigs. In *Proceedings of SISVet*, pages 238–239, 2018. 32, 33, 36

[86] Alan G Mathew, Robin Cissell, and S Liamthong. Antibiotic resistance in bacteria associated with food animals: a united states perspective of livestock production. *Foodborne pathogens and disease*, 4(2):115–133, 2007. 31

[87] Neeraj Matiyali and Gaurav Sharma. Video person re-identification using learned clip similarity aggregation. In *The IEEE Winter Conference on Applications of Computer Vision*, 2020. 66, 72

[88] Patrick McAllister, Huiru Zheng, Raymond Bond, and Anne Moorhead. Combining deep residual neural network features with supervised machine learning algorithms to classify diverse food image datasets. *Computers in biology and medicine*, 95:217–233, 2018. 35

[89] G Merialdi, M Dottori, P Bonilauri, A Luppi, S Gozio, P Pozzi, B Spaggiari, and P Martelli. Survey of pleuritis and pulmonary lesions in pigs at abattoir with a focus on the extent of the condition and herd risk factors. *The Veterinary Journal*, 193(1):234–239, 2012. 32, 34

[90] Anton Milan, Laura Leal-Taixé, Ian Reid, Stefan Roth, and Konrad Schindler. Mot16: A benchmark for multi-object tracking. *arXiv preprint arXiv:1603.00831*, 2016. 22

[91] Ming-Kuei Hu. Visual pattern recognition by moment invariants. *IRE Transactions on Information Theory*, 8(2):179–187, 1962. 23

[92] Mateusz Mittek, Eric T Psota, Jay D Carlson, Lance C Pérez, Ty Schmidt, and Benny Mote. Tracking of group-housed pigs using multi-ellipsoid expectation maximisation. *IET Computer Vision*, 12(2):121–128, 2017. 7, 16, 17

[93] Thierry Pinheiro Moreira, Mauricio Lisboa Perez, Rafael de Oliveira Werneck, and Eduardo Valle. Where is my puppy? retrieving lost dogs by facial features. *Multimedia Tools and Applications*, 76(14):15325–15340, 2017. 9, 46

[94] Abozar Nasirahmadi, Sandra A Edwards, and Barbara Sturm. Implementation of machine vision for detecting behaviour of cattle and pigs. *Livestock Science*, 202:25–38, 2017. 15

[95] Thuy-Binh Nguyen, Thi-Lan Le, Dinh-Duc Nguyen, and Dinh-Tan Pham. A reliable image-to-video person re-identification based on feature fusion. In *Asian conference on intelligent information and database systems*, 2018. 59

[96] Mohammad Sadegh Norouzzadeh, Anh Nguyen, Margaret Kosmala, Alexandra Swanson, Meredith S Palmer, Craig Packer, and Jeff Clune. Automatically identifying, counting, and describing wild animals in

camera-trap images with deep learning. *Proceedings of the National Academy of Sciences*, page 201719367, 2018. 9, 16, 47

[97] Timo Ojala, Matti Pietikainen, and David Harwood. Performance evaluation of texture measures with classification based on kullback discrimination of distributions. In *Pattern Recognition, 1994. Vol. 1-Conference A: Computer Vision & Image Processing., Proceedings of the 12th IAPR International Conference on*, volume 1, pages 582–585. IEEE, 1994. 10, 54

[98] Jongchan Park, Sanghyun Woo, Joon-Young Lee, and In So Kweon. BAM: bottleneck attention module. In *British Machine Vision Conference*, 2018. 68

[99] Stefano Pini, Marcella Cornia, Federico Bolelli, Lorenzo Baraldi, and Rita Cucchiara. M-VAD Names: a Dataset for Video Captioning with Naming. *Multimedia Tools and Applications*, 78(10):14007–14027, 2019. 21

[100] Roberta Forti Pol Marquer, Teresa Rabade. *Pig farming in the European Union: considerable variations from one Member State to another*, 2020. https://ec.europa.eu/eurostat/statistics-explained/pdfscache/3688.pdf. 15

[101] Eric T Psota, Mateusz Mittek, Lance C Pérez, Ty Schmidt, and Benny Mote. Multi-pig part detection and association with a fully-convolutional network. *Sensors*, 19(4):852, 2019. 7, 16, 17

[102] Jingjing Qian, Wei Jiang, Hao Luo, and Hongyan Yu. Stripe-based and attribute-aware network: A two-branch deep model for vehicle re-identification. *arXiv preprint arXiv:1910.05549*, 2019. 69, 71

[103] Maheen Rashid, Xiuye Gu, and Yong Jae Lee. Interspecies knowledge transfer for facial keypoint detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, 2017. 52

[104] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 6

[105] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 19

[106] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *European Conference on Computer Vision*, pages 17–35. Springer, 2016. 22, 64, 66

[107] Ergys Ristani and Carlo Tomasi. Features for multi-target multi-camera tracking and re-identification. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2018. 67

[108] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *International Conference on Learning Representations*, 2015. 11

[109] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 36

[110] Marc Rußwurm and M Körner. Multi-temporal land cover classification with long short-term memory neural networks. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 42:551, 2017. 6

[111] Jaewon Sa, Younchang Choi, Hanhaesol Lee, Yongwha Chung, Daihee Park, and Jinho Cho. Fast pig detection with a top-view camera under various illumination conditions. *Symmetry*, 11(2):266, 2019. 7, 16, 17

[112] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2018. 65, 68

[113] Stefan Schneider, Graham W Taylor, Stefan Linquist, and Stefan C Kremer. Past, present and future approaches using computer vision for animal re-identification from camera trap data. *Methods in Ecology and Evolution*, 2019. 59

[114] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 10, 50, 59

[115] Annalisa Scollo, Flaviana Gottardo, Barbara Contiero, Claudio Mazzoni, Philippe Leneveu, and Sandra A Edwards. Benchmarking of pluck lesions at slaughter as a health monitoring tool for pigs slaughtered at 170 kg (heavy pigs). *Preventive veterinary medicine*, 144:20–28, 2017. 31, 32

[116] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2017. 74, 76

[117] Jihyun Seo, Hanse Ahn, Daewon Kim, Sungju Lee, Yongwha Chung, and Daihee Park. Embeddedpigdet—fast and accurate pig detection for embedded board implementations. *Applied Sciences*, 10(8):2878, 2020. 6, 16, 17

[118] M. Shahbandeh. *Number of pigs worldwide from 2012 to 2020*, 2020. `https://www.statista.com/statistics/263963/number-of-pigs-worldwide-since-1990`. 15

[119] Bin Shao and Hongwei Xin. A real-time computer vision assessment and control of thermal comfort for group-housed pigs. *Computers and electronics in agriculture*, 62(1):15–21, 2008. 33

[120] Jianlou Si, Honggang Zhang, Chun-Guang Li, Jason Kuen, Xiangfei Kong, Alex C Kot, and Gang Wang. Dual attention matching network for context-aware feature sequence based person re-identification. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2018. 66, 71, 72

[121] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *Neural Information Processing Systems*, 2016. 59

[122] V. Sorenson, S.E. Jorsal, and Mousing J. *Diseases of Swine (9th edition)*, pages 149–178. Blackwell Pubishing, 2006. 31, 32

[123] Concetto Spampinato, Yun-Heh Chen-Burger, Gayathri Nadarajan, and Robert B Fisher. Detecting, tracking and counting fish in low quality unconstrained underwater videos. *VISAPP (2)*, 2008(514-519):1, 2008. 16

[124] Luciano Spinello and Kai O Arras. People detection in rgb-d data. In *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3838–3843. IEEE, 2011. 16

[125] Russell Stewart, Mykhaylo Andriluka, and Andrew Y Ng. End-to-end people detection in crowded scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2325–2333, 2016. 16

[126] Michael K. Swan. *Swine Human Resources: Managing Employees*, 2020. `https://swine.extension.org/swine-human-resources-managing-employees`. 15

[127] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. 8

[128] Eric T Psota, Ty Schmidt, Benny Mote, and Lance C Pérez. Long-term tracking of group-housed livestock using keypoint detection and map estimation for individual animal identification. *Sensors*, 20(13):3670, 2020. 7, 17, 18

[129] Zheng Tang, Milind Naphade, Stan Birchfield, Jonathan Tremblay, William Hodge, Ratnesh Kumar, Shuo Wang, and Xiaodong Yang. Pamtri: Pose-aware multi-task learning for vehicle re-identification using highly randomized synthetic data. In *IEEE International Conference on Computer Vision*, 2019. 71

[130] Maoqing Tian, Shuai Yi, Hongsheng Li, Shihua Li, Xuesen Zhang, Jianping Shi, Junjie Yan, and Xiaogang Wang. Eliminating background-bias for robust person re-identification. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2018. 73

[131] Yunong Tian, Guodong Yang, Zhe Wang, En Li, and Zize Liang. Detection of apple lesions in orchards based on deep learning methods of cyclegan and yolov3-dense. *Journal of Sensors*, 2019, 2019. 5

[132] Transparency Research. *Hog Production and Pork Market*, 2019. https://www.transparencymarketresearch.com/hog-production-pork-market.html. 15

[133] Matthew Tscharke and Thomas M Banhazi. A brief review of the application of machine vision in livestock behaviour analysis. *Agrárinformatika/Journal of Agricultural Informatics*, 7(1):23–42, 2016. 16

[134] David Tseng, David Wang, Carolyn Chen, Lauren Miller, William Song, Joshua Viers, Stavros Vougioukas, Stefano Carpin, Juan Aparicio Ojea, and Ken Goldberg. Towards automating precision irrigation: Deep learning to infer local soil moisture conditions from synthetic aerial agricultural images. In *2018 IEEE 14th International Conference on Automation Science and Engineering (CASE)*, pages 284–291. IEEE, 2018. 6

[135] Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *IEEE International Conference on Computer Vision*, 2019. 11, 60

[136] Matthew Turk and Alex Pentland. Eigenfaces for recognition. *Journal of cognitive neuroscience*, 3(1):71–86, 1991. 10, 54

[137] Dmitry Ulyanov, Andrea Vedaldi, and Victor S. Lempitsky. Instance normalization: The missing ingredient for fast stylization. *CoRR*, abs/1607.08022, 2016. 49

[138] James P Underwood, Mark Calleija, Juan Nieto, Salah Sukkarieh, Cameron EF Clark, Sergio C Garcia, Kendra L Kerrisk, and Greg M Cronin. A robot amongst the herd: Remote detection and tracking of cows. In *NEW ZEALAND SPATIALLY ENABLED LIVESTOCK MANAGEMENT SYMPOSIUM*, page 45, 2013. 16

[139] Raquel Urtasun, David J Fleet, and Pascal Fua. 3d people tracking with gaussian process dynamical models. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 1, pages 238–245. IEEE, 2006. 16

[140] Evgeniya Ustinova and Victor Lempitsky. Learning deep embeddings with histogram loss. In *Advances in Neural Information Processing Systems*, pages 4170–4178, 2016. 50, 59

[141] Thomas P Van Boeckel, Emma E Glennon, Dora Chen, Marius Gilbert, Timothy P Robinson, Bryan T Grenfell, Simon A Levin, Sebastian Bonhoeffer, and Ramanan Laxminarayan. Reducing antimicrobial use in food animals. *Science*, 357(6358):1350–1352, 2017. 31

[142] Carl Vondrick, Donald Patterson, and Deva Ramanan. Efficiently scaling up crowdsourced video annotation. *International Journal of Computer Vision*, pages 1–21. 10.1007/s11263-012-0564-1. 50

[143] Erik Vranken and Dries Berckmans. Precision livestock farming for pigs. *Animal Frontiers*, 7(1):32–37, 2017. 15

[144] Guangcong Wang, Jianhuang Lai, and Xiaohua Xie. P2snet: can an image match a video for person re-identification in an end-to-end way? *IEEE Transactions on Circuits and Systems for Video Technology*, 2017. 11, 70

[145] Jian Wang, Feng Zhou, Shilei Wen, Xiao Liu, and Yuanqing Lin. Deep metric learning with angular loss. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2017. 59

[146] Zhongdao Wang, Luming Tang, Xihui Liu, Zhuliang Yao, Shuai Yi, Jing Shao, Junjie Yan, Shengjin Wang, Hongsheng Li, and Xiaogang Wang. Orientation invariant feature embedding and spatial temporal regularization for vehicle re-identification. In *IEEE International Conference on Computer Vision*, 2017. 66

[147] Yu Wu, Yutian Lin, Xuanyi Dong, Yan Yan, Wanli Ouyang, and Yi Yang. Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2018. 60, 66, 77

[148] Kaitlin Wurtz, Irene Camerlink, Richard B D'Eath, Alberto Peña Fernández, Tomas Norton, Juan Steibel, and Janice Siegford. Recording behaviour of indoor-housed farm animals automatically us-

ing machine vision technology: A systematic review. *PloS one*, 14(12):e0226669, 2019. 15

[149] Zhongwei Xie, Lin Li, Xian Zhong, Luo Zhong, and Jianwen Xiang. Image-to-video person re-identification with cross-modal embeddings. *Pattern Recognition Letters*, 2019. 59

[150] Chenglin Yang, Lingxi Xie, Siyuan Qiao, and Alan Yuille. Knowledge distillation in generations: More tolerant teachers educate better students. *arXiv preprint arXiv:1805.05551*, 2018. 11

[151] Jiwen Yu, Haibo Su, Junnan Liu, Zhizheng Yang, Zhouyangzi Zhang, Yixin Zhu, Lu Yang, and Bingliang Jiao. A strong baseline for tiger re-id and its bag of tricks. In *IEEE International Conference on Computer Vision Workshops*, 2019. 72

[152] Sun Yukun, Huo Pengju, Wang Yujie, Cui Ziqi, Li Yang, Dai Baisheng, Li Runze, and Zhang Yonggen. Automatic monitoring system for individual dairy cows based on a deep learning framework that provides identification via body parts and estimation of body condition score. *Journal of dairy science*, 102(11):10140–10151, 2019. 6

[153] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *International Conference on Learning Representations*, 2017. 11

[154] Dongyu Zhang, Wenxi Wu, Hui Cheng, Ruimao Zhang, Zhenjiang Dong, and Zhaoquan Cai. Image-to-video person re-identification with temporally memorized similarity learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 2017. 11, 59, 63, 70

[155] Kaifeng Zhang, Dan Li, Jiayun Huang, and Yifei Chen. Automated video behavior recognition of pigs using two-stream convolutional networks. *Sensors*, 20(4):1085, 2020. 8, 16, 17

[156] Lei Zhang, Helen Gray, Xujiong Ye, Lisa Collins, and Nigel Allinson. Automatic individual pig detection and tracking in pig farms. *Sensors*, 19(5):1188, 2019. 7, 16, 17

[157] Yu-Dong Zhang, Zhengchao Dong, Xianqing Chen, Wenjuan Jia, Sidan Du, Khan Muhammad, and Shui-Hua Wang. Image based fruit category classification by 13-layer deep convolutional neural network and data augmentation. *Multimedia Tools and Applications*, 78(3):3613–3632, 2019. 6

[158] Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, Shengjin Wang, and Qi Tian. Mars: A video benchmark for large-scale person re-identification. In *European Conference on Computer Vision*, 2016. 60, 61, 64, 66, 77

[159] Liang Zheng, Yi Yang, and Alexander G Hauptmann. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*, 2016. 59, 66

[160] Liheng Zhong, Lina Hu, and Hang Zhou. Deep learning based multi-temporal crop classification. *Remote sensing of environment*, 221:430–443, 2019. 6

[161] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2017. 72

[162] Yi Zhou, Li Liu, and Ling Shao. Vehicle re-identification by deep hidden multi-view inference. *IEEE Transactions on Image Processing*, 2018. 66

[163] Yi Zhou and Ling Shao. Aware attentive multi-view inference for vehicle re-identification. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2018. 11, 66

[164] Thi Thi Zin, Cho Nilar Phyo, Pyke Tin, Hiromitsu Hama, and Ikuo Kobayashi. Image technology based cow identification system using deep learning. In *Proc. of the International MultiConference of Engineers and Computer Scientists IMECS 2018, Hong Kong*, volume 1, 2018. 10