

Alma Mater Studiorum Università di Bologna  
Archivio istituzionale della ricerca

A computer vision approach based on deep learning for the detection of dairy cows in 2 free stall barn

This is the final peer-reviewed author's accepted manuscript (postprint) of the following publication:

*Published Version:*

Tassinari, P., Bovo, M., Benni, S., Franzoni, S., Poggi, M., Mammi, L.M.E., et al. (2021). A computer vision approach based on deep learning for the detection of dairy cows in 2 free stall barn. COMPUTERS AND ELECTRONICS IN AGRICULTURE, 182, 1-15 [10.1016/j.compag.2021.106030].

*Availability:*

This version is available at: <https://hdl.handle.net/11585/800275> since: 2021-02-16

*Published:*

DOI: <http://doi.org/10.1016/j.compag.2021.106030>

*Terms of use:*

Some rights reserved. The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (<https://cris.unibo.it/>).  
When citing, please refer to the published version.

(Article begins on next page)

# A computer vision approach based on deep learning for the detection of dairy cows in free stall barn

Patrizia Tassinari<sup>a</sup>, Marco Bovo<sup>a\*</sup>, Stefano Benni<sup>a</sup>, Simone Franzoni<sup>b</sup>, Matteo Poggi<sup>b</sup>, Ludovica Maria Eugenia Mammi<sup>c</sup>, Stefano Mattoccia<sup>b</sup>, Luigi Di Stefano<sup>b</sup>, Filippo Bonora<sup>a</sup>, Alberto Barbaresi<sup>a</sup>, Enrica Santolini<sup>a</sup>, Daniele Torreggiani<sup>a</sup>

<sup>a</sup> Department of Agricultural and Food Sciences, University of Bologna, Bologna, Italy

<sup>b</sup> Department of Computer Science and Engineering, University of Bologna, Bologna, Italy

<sup>c</sup> Department of Veterinary Medical Sciences, University of Bologna, Bologna, Italy

\* Corresponding author. Email: [marco.bovo@unibo.it](mailto:marco.bovo@unibo.it)

## Abstract

Precision Livestock Farming relies on several technological approaches to acquire in the most efficient way precise and up-to-date data concerning individual animals. In dairy farming, particular attention is paid to the automatic cow detection and tracking, as such information is closely related to animal welfare and thus to possible health issues. Computer vision represents a suitable and promising method for this purpose.

This paper describes the first step for the development of a computer vision system, based on deep learning, aiming to recognize in real-time the individual cows, detect their positions, actions and movements and record the time history outputs for each animal.

Specifically, a neural network based on deep learning techniques has been trained and validated on a case study farm, for the automatic recognition of individual cows in videos recorded in the barn. Four cows were selected to train and validate a YOLO neural network able to recognize a cow starting from the coat pattern. Then, precision-recall curves of the identification of individual cows were elaborated for both the specific target classes and the whole dataset in order to assess the performances of the network.

By means of data augmentation techniques, an enlarged dataset has been created and considered in order to improve the performance of the network and to provide indications to increase detection efficiency in those cases where data acquisition is not easy to be carried out for long periods. The mean average precision of the detection, ranging from 0.64 to 0.66, showed that it is possible to properly identify individual cows based on their morphological appearance and that the piebald spotting pattern of a cow's coat represents a clearly distinguishable object for a computer vision network. The results also led to obtain indications about the quantity and the characteristics of the images to be used for the network training in order to achieve efficient detections when facing with applications involving animals.

**Keywords:** precision livestock farming; computer vision; deep learning; dairy cow; herd management

Symbol	Description
$A_t$	Average Area of the bounding boxes in training phase
$A_v$	Average Area of the bounding boxes in validation phase
$O_t$	number of Occurrences in training phase
$O_v$	number of Occurrences in validation phase
tp	true positive
tn	true negative
fp	false positive
fn	false negative
P	Precision
R	Recall
$F_1$	$F_1$ -score
Pr	detection Probability
C	Confidence score
AP	Average Precision
mAP	mean Average Precision
IoU	Intersection over Union
AIoU	Average Intersection over Union

## 37 1. Introduction

38 The implementation of precision livestock farming (PLF) techniques in animal husbandry involves many fields  
39 of the technological innovation and several researchers are currently seeking to apply new methodologies and  
40 algorithms for both commercial and research purposes (Tullo et al., 2019). In particular, with reference to  
41 innovative applications in the livestock sector, the animals are increasingly being analyzed and studied with  
42 the help of informatics tools such as support vector machines, random forests techniques, neural networks,  
43 machine learning approaches etc. (Kamilaris and Prenafeta-Boldú, 2018; Tsai and Huang, 2014; Li et al., 2017;  
44 Okura et al., 2019). In this context, multiple challenges involve the dairy cattle sector, where the methods  
45 currently available are often unsuitable to manage the collected data and to fully extrapolate the potential  
46 informative content.

47 In fact, weather stations monitoring barn temperature and humidity, robotic milking systems helping the daily  
48 work of the farmers, collars and pedometers controlling animals' activities and positions (Berckmans, 2014)  
49 are capable to collect huge amounts of real-time data currently used to support the herd management.

50 PLF has thus contributed to switch the analysis framework of a farm from a data-poor to a data-rich situation:  
51 the research key challenge is actually to turn those data into knowledge able to provide decision-maker with  
52 real-time support for the cattle farm optimization (Barkema et al., 2015; Bewley et al., 2017; Fournel et al.,  
53 2017; Van Hertem et al., 2014; Halachmi et al., 2013; Guzhva et al., 2016; Martinez-Ortiz et al., 2013).  
54 Several researches have pointed out the opportunity to develop both algorithms suitable to provide early  
55 warning and control systems able to optimize animal welfare and productivity based on data collected through  
56 Information Communication Technology (ICT) systems (Alsaad et al., 2019; Jaeger et al., 2019; Cowley et al.,  
57 2015). In this context, computer vision together with numerical analysis methodologies proved to have  
58 fundamental importance (Van Hertem et al., 2018) . Computer Vision techniques are meant to be applied in  
59 this research field in order to automate actions normally carried out by the human visual system (Taigman et  
60 al., 2014). The aim of these algorithms is to “teach” a computer to apprehend from images and videos in order  
61 to simulate the human vision and substitute the human beings in repetitive or complex actions. They have  
62 already been tested on animals with promising results (Norouzzadeh et al., 2017; Trnovszky et al., 2017), also  
63 in the livestock farming area (Aydin, 2017; Van Hertem et al., 2013), but recognizing each individual cow within  
64 the herd is still representing a challenging issue.

65 The monitoring of position and movements of the individual animals may be necessary to quantify the main  
66 indices related to animal welfare (Song et al., 2008; Jiang et al., 2019) and behavior (Porto et al., 2013, 2015),  
67 as well as to identify any preferences of the cows regarding different zones of the barn.

68 Moreover, video monitoring of the herd, together with the adoption of quantitative criteria to control animal  
69 welfare by means of computer vision, may represent a tool to improve citizens’ consciousness about rearing  
70 conditions and increase their knowledge about feeding and housing practices. A recent study found that fresh  
71 food and water, pasture access, gentle handling, space, shelter, hygiene, fresh air and sunshine, social  
72 companions, absence of stress, health and safety from predators are considered by citizens as necessary  
73 requirement for dairy cattle “good life” (Ventura et al., 2016). These results suggested that a transparent  
74 exposure of livestock farming to the public may resolve some concerns, and video recording appear to be a  
75 powerful tool for an effective and widespread information.

76 Therefore, the monitoring requirements can be considered according to two main levels of information and  
77 complexity. The first one concerns the identification at regular time intervals of the number of animals that  
78 are in a certain position, for example lying in a cubicle, standing at the manger etc. The second one deals with  
79 the identification, instant by instant, of the behavior of the individual cows, with the possibility of calculating,  
80 for each animal, the time spent in each position and the temporal sequence of its positions, including the  
81 trajectory of its movements.

82 The first level of information makes it possible to quantify the aforementioned indices and to have an overview  
83 of the performances of the general animal welfare conditions in the farm, by appropriately integrating these  
84 indices with information on feeding and productivity of the animals, which can be deduced from other sources,  
85 such as the milking robots and mixed ration delivery systems. The second level of information allows to have  
86 specific information on the individual cow welfare condition and on the use of the various areas of the barn  
87 and represents a challenge for innovation in PLF. A computer vision system implemented through deep  
88 learning constitutes a suitable approach to achieve the latter objective.

89 The aim of this study is to develop and test the reliability level of a computer vision system, based on deep  
90 learning techniques, for the automatic recognition of individual cows within images representing their  
91 position. In particular, whilst developing a new software framework lies outside the scope of this paper, the  
92 paper focuses on methodological aspects related to a more efficient application of ICT in the dairy cows  
93 monitoring field. In fact, while object detection is already applied in commercial applications in various  
94 contexts, individual cow recognition still represents an open issue for both the research and commercial fields,  
95 since no consolidated approach still exists. Actually, some of these blocks of information are collected by  
96 means of different very expensive sensors (ALLFLEX, 2020; DELAVAL, 2020; AFIMILK, 2020), the most of them  
97 to be worn by the animals but with the system proposed here, it will be possible to collect wider information,  
98 with a less expensive technology and finally yet importantly using a system that avoids the problems and labor  
99 due to wearable sensors.

100 Therefore, the paper focuses on the selection, validation and performance assessment of a neural network, in  
101 terms of speed and accuracy, and at the same time outlines the key issues of the broader process, which is

102 strictly depending and related to the enabling steps presented in the paper. Further testing and validating cow  
103 detection procedures in various contexts, in different types of livestock structures, and in different operating  
104 conditions, is an important research field, contributing to the definition of consolidated approaches enabling  
105 cow recognition, displacement tracking and cow action/behavior (eating, drinking, lying down, standing etc.)  
106 recognition systems. This system could represent an innovative and useful tool to support the farmer in the  
107 daily management and decision-making. In fact, by means of the technology proposed here, it will be possible  
108 to calculate, for example, the indices connected to animal welfare but also to check the time spent for  
109 nutrition, drinking and walking. The outcomes of this monitoring could be effectively used by the farmers to  
110 identify potential anomalies or diseases and promptly apply specific corrective actions (e.g. on the fans  
111 controlling the barn environmental conditions, water supply, etc.). In addition to the technological system and  
112 the experimental setup adopted, the paper describes an innovative algorithm for the detection of the cows  
113 implemented and tested on a case study farm with computer vision procedures. The promising results  
114 reported here represents a first preliminary contribution to the progress of the computer vision for herd  
115 monitoring applications and could be an important support for the following study of the movements of the  
116 cows in the barn and for the analysis of their actions and behavior.

117

## 118 2. Materials and methods

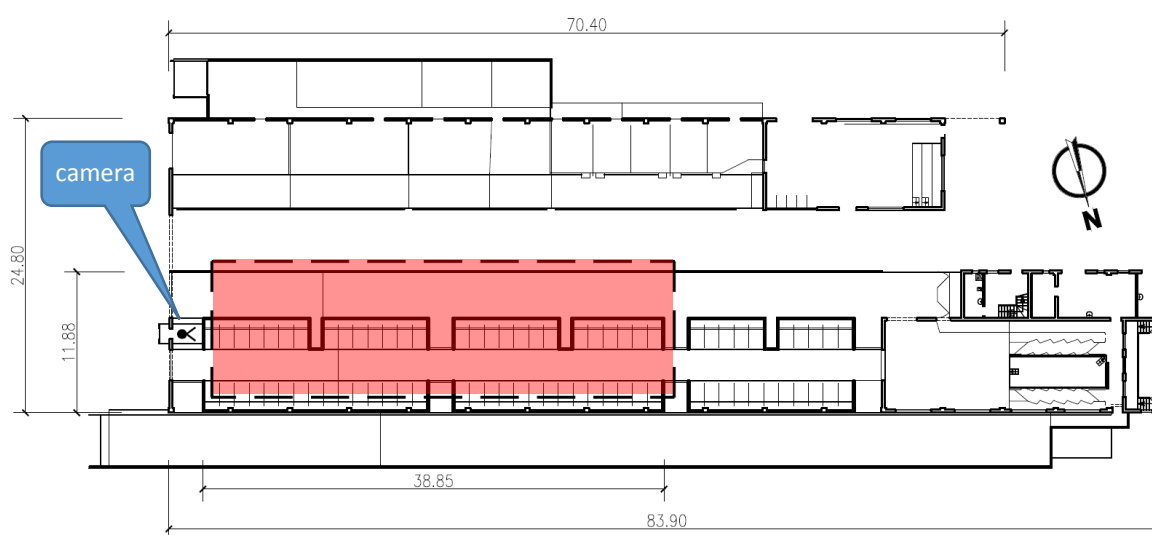
119

### 120 2.1 Study case

121 The case study considered in this work is the experimental dairy cattle farm of the University of Bologna,  
122 located in Ozzano Emilia, Bologna, in the North-East of Italy, where Holstein Friesian cattle is reared. This farm  
123 is managed by the Department of Veterinary Medical Science and represents a unique reality in the national  
124 context, thanks to its equipment and monitoring systems, which will allow to carry out the integrated analyses  
125 aimed at the definition of models for the interpretation of milk production, reproductive, environmental and  
126 management data. The building is a free stall experimental barn hosting about 150 animals, including 80 cows  
127 (72 milking and 8 dry cows, and 70 among calves and heifers). The barn has a resting area in bedding material  
128 for dry cows and litter cubicles for lactating cows (see Figure 1). A layout of the building is showed in Figure 1.

129 The barn is provided with cooling ventilation systems based on the Temperature-Humidity Index (THI) value.  
 130 The cows are milked twice a day and for each cow, the behavioral, productive and health parameters are daily  
 131 recorded automatically. The animals are fed with total mixed ration and auto-feeders for concentrate  
 132 supplementation.

133



134 Figure 1. Layout of the experimental barn. The dashed line delimitates the area framed by the camera (red  
 135 colored).  
 136  
 137

## 138 2.2 *Software and neural network*

### 139 2.2.1 *Tagging software*

140 VoTT (Visual object Tagging Tool) (Microsoft, 2018) was the software selected for the tagging tasks. It is an  
 141 open source annotation and labelling tool for image and video assets, it is a React + Redux Web application  
 142 written in Typescript. This software has been selected because it has several features, useful for the application  
 143 in the field of the present work. For example, it has an extensible model for importing data from local or cloud  
 144 storage providers and an extensible model for exporting labelled data to local or cloud storage providers.  
 145 Moreover, it has been developed not only to analyze images, but also video frames. So, it is possible to decide  
 146 the interval of frames per second and VoTT automatically transforms a video in a set of pictures. VoTT was  
 147 programmed following the 'Bring Your Own Data' (BYOD) approach and in VoTT, connections are used to  
 148 configure and manage source, i.e. the assets to label, and the target, i.e. the location to which labels should  
 149 be exported. Then, for the development of the computer vision technology in the field of animal monitoring,

150 VoTT results very practical to set up and facilitate an end-to-end machine learning pipeline. In this work the  
151 version 1.0.8 has been used.

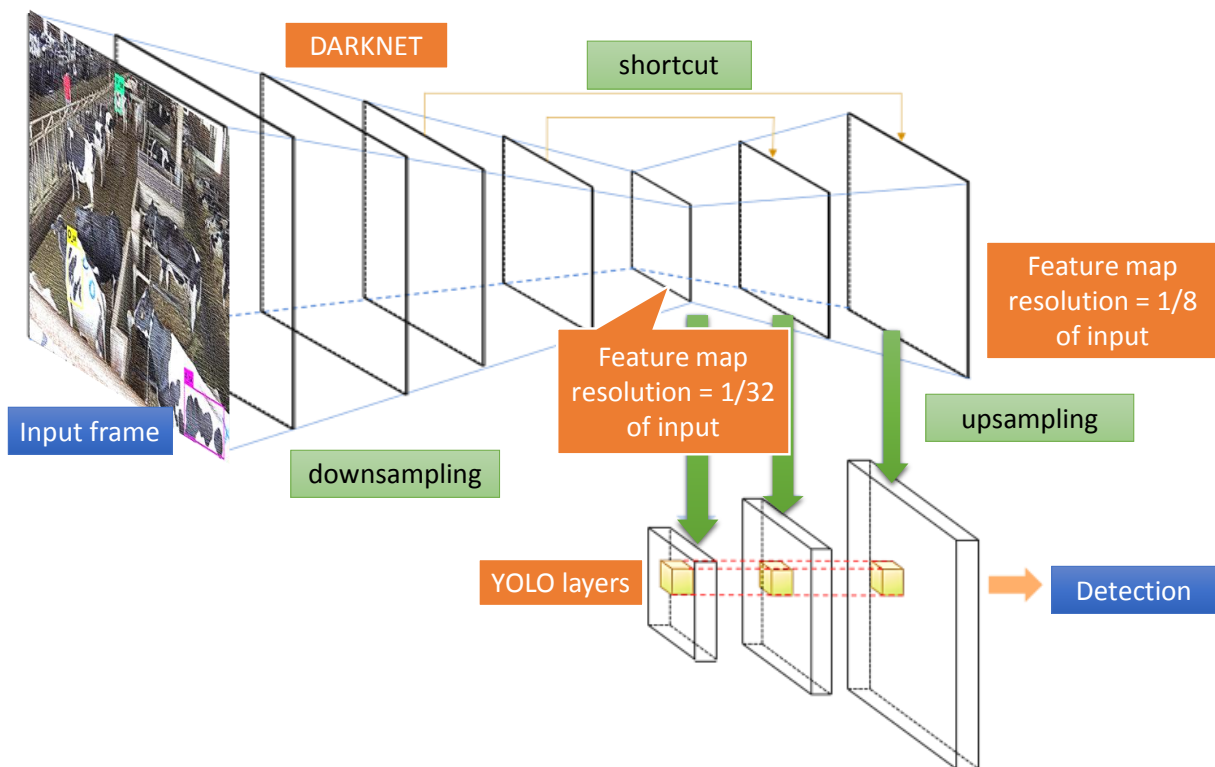
152

### 153 2.2.2 *Detection algorithm and deep learning network*

154 At the state-of-the-art, in the field of computer vision for object detection, different more or less efficient  
155 algorithms could be considered. For example, region-based convolutional neural networks (R-CNNs) (Girshick  
156 et al., 2014) have been a pioneering approach that applies deep models to object detection. Actually, Faster  
157 R-CNN (Ren et al., 2017) has been proved to be one of the most performing of its class, if compared with R-  
158 CNN and Fast R-CNN (Girshick, 2015) and it has been already applied for cow detection for instance by (Bezen  
159 et al., 2020). Different is the approach implemented in YOLO - You Only Look Once (Redmon et al., 2016),  
160 where predictions are made with a single network evaluation, thus making the process much faster than  
161 region-based convolutional neural networks (R-CNN) which require thousands of evaluations for a single  
162 image. YOLO presents a totally different approach from prior detection systems, which proposes classifiers or  
163 localizers to perform detection. It is one of the state-of-the-art detectors which are capable of localizing and  
164 classifying multiple objects in images. In particular, the detector is faster than two-stage detectors: while they  
165 propose object regions first and investigate the regions for object localization and classification, YOLO  
166 combines the two stages into one neural network. Therefore, instead of applying the model to an image at  
167 multiple locations and scales, so that high-score regions of the image are considered detections, Yolo applies  
168 a single neural network to the full image. This network divides the image into regions and predicts bounding  
169 boxes and probabilities for each region and these bounding boxes are weighted by the predicted probabilities.  
170 The improvements in YOLO v3 (Redmon and Farhadi, 2018) made the algorithm even faster and suitable for  
171 problems like those investigated here. In fact, the future real applications of this system will face with the  
172 detection of large number of cows in a herd, during the time, and should be able to recognize the possible  
173 action the cow is doing in real-time. For this reason a fast algorithm is necessary and this has driven the choice  
174 of the authors towards the framework Darknet (Redmon, 2013), an open source framework for neural network  
175 development including YOLO v3. The algorithm has the advantage to look at the whole image at test time, so



176 that predictions are informed by the global context of the image (see Figure 2). More specifically, YOLO v3  
 177 predicts an objectness score for each bounding box using logistic regression. This score should be 1 if the  
 178 bounding box prior overlaps a ground truth object by more than any other bounding box prior. Each box  
 179 predicts the classes the bounding box may contain using multi-label classification. At the current state of the  
 180 research the authors selected and adopted YOLO v3 as detection system but the evaluation of the  
 181 performances of different detection algorithms, for the problem investigated here, will be object of future  
 182 investigations.  
 183



184  
 185 Figure 2. Architecture of YOLO v3.

186  
 187 **2.3 Data collection**

188 As the performance of the neural network are strictly connected to the quantity and the quality of the training  
 189 dataset, particular attention was paid to the collection of a suitable dataset of video frames. The videos were  
 190 registered by a HDR-CX115E (Sony) camera (see Figure 3) in a high quality standard (HD resolution, 25 frames

191 per second). The recording has been conducted on a tripod positioned 2 meters above the barn floor, so the  
192 total height for the recording was about 3.50 meters.  
193 The section recorded by the camera focused on the feeding area, including the rack (on the left) and the  
194 cubicles (on the center and on the right of the frames). A limitation of this position is that the images of the  
195 cows up in the trough are greater for quality and quantity, so the dataset is composed by more photos of the  
196 left hips than the right ones.



197

198 Figure 3. The HDR-CX115E camera positioned for video recording.

199

## 200 2.4 Research method

201 The research has been structured in the following main phases:

- 202 1) random selection in the herd of a sample of 10% of the cows' population in the free stall area monitored  
203 by the camera (i.e. about 40 cows) during the recording phase. The sample is suitable to verify if the  
204 network is capable to recognize a cow among others and to distinguish between more animals. Then the  
205 animals have been marked, with a specific blue paint, with a letter only to help the identification of the  
206 cows in the different frames. It is worth to notice this aspect since the neural network has not been

207 trained based on these symbols but the bounding boxes considered only pelt portions with natural  
208 pattern of the cows;

209 2) recording of the videos after selecting the more suitable position for the camera;

210 3) creation of the dataset for the training phase;

211 4) training of the neural network, i.e. definition of the parametric weights for object recognition;

212 5) validation test of the neural network with the weights defined in phase 4 and scoring the results by means  
213 of the performance indicators defined before;

214 6) creation of an augmented “virtual” dataset aiming to improve the detection performances of the network  
215 for the classes poorly represented in the frames used for the network training;

216 7) repetition of the phases from 3 to 5 in order to assess the improvement in the detection performances  
217 after the manipulations operated to the frame dataset.

218

## 219 **2.5 Experimental setup and tests**

220 Four cows have been selected to train and test the neural network adopted in the study. The four letters X, V  
221 O and I have been used only to identify a specific cow and the letter corresponding to each cow was drawn on  
222 both right and left hips and on the forehead of the various cows. Two types of classes have been defined for  
223 the identification of each cow corresponding to the pelt of the hips of the four cows. A total number of 8  
224 classes (4 cows × 2 hips) has been adopted for the neural network training/validation in order to recognize the  
225 cows by the black-white pattern of each specific pelt. Therefore, each class was identified by the capital letter  
226 indicating the cow followed by right (r) or left (l) for identify the two different hips. The eight classes considered  
227 in the study are: X<sub>left</sub>, X<sub>right</sub>, V<sub>left</sub>, V<sub>right</sub>, O<sub>left</sub>, O<sub>right</sub>, I<sub>left</sub> and I<sub>right</sub>. For example, Figure 4 illustrates the internal view  
228 of the barn with a detail of the blue paint on the right hip of the cow labelled “V”, thus labelled as class V<sub>right</sub>.



Figure 4. Example of a frame with detail of the blue paint on the right hip of the cow labelled “V” so representing the class  $V_{right}$ .

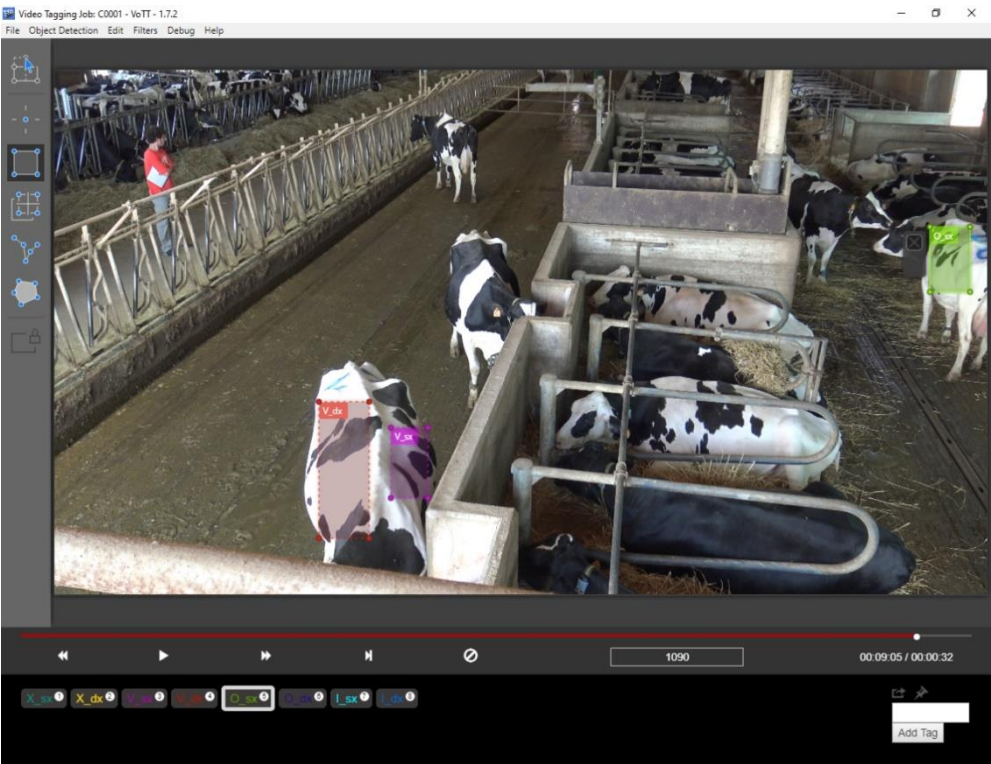
The videos were recorded on July 2019, collecting a total duration of 210 min to carry out the study. The frame tagging phase was performed through the abovementioned VoTT software, which allowed to sample the videos at a chosen frequency and to make tags (with bounding boxes) on all the frames. In this study, the sampling frequency for the selection of the frame dataset was chosen equal to 2 frames per second because the scenario does not change in a fast way and so a higher frequency would have been redundant. Therefore, about 25200 frames were sampled. The bounding boxes used for tagging the frames were rectangular, rather than square, because the objects to be tagged, i.e. the pelt of the cows, had generally different horizontal and vertical dimensions. The coat area selected in the tagging bounding box has been the biggest rectangular area (with horizontal orientation) identifiable with continuity within the image of the hip of the cow.

Thus, at the end of the tagging phase we obtained:

- a collection of graphical files corresponding to every sampled frame
- for each frame, a text data file like the one in Figure 5 containing the class number recognized in the frame (if any), the coordinates of the centroid of the corresponding box and the sizes of the box.



246 A total of 11754 frames showing at least one of the classes were identified and labelled.



(a)

	File	Modifica	Formato	Visualizza	?
2	0.446484	0.105787	0.034115	0.065278	
0	0.876302	0.913889	0.193750	0.170370	
4	0.224479	0.726620	0.079167	0.159722	

Callouts from the table:

- Class number (points to the first column)
- Coordinates of the centroid of the bounding box (points to the second and third columns)
- Width and height of the bounding box (points to the fourth and fifth columns)

(b)

251 Figure 5. Tagging phase through VoTT. (a) Example of graphical file of a frame and (b) example of the data  
252 acquired by the tagging process. The coordinates of the centroid and the dimensions of the box are expressed  
253 as ratios of the dimensions of the frame.

254 The data acquired through the tagging phase were then analyzed to quantify the occurrences of the various  
255 target classes. Moreover, the sizes (i.e. width and height) of the bounding boxes were computed as an  
256 indicator of the visibility of the target cow within the frame. In fact, a small box may indicate either a position  
257 of the cow far from the recording viewpoint, or a partial coverage of the animal by an object in the foreground.  
258

The data about the occurrences of the target classes were considered for a proper definition of the split of the dataset in a training set and a validation set. The criterion adopted for the split was the selection of sets of consecutives frames accounting for about 80% of the occurrences for each class in the training set and the remaining 20% for the validation set. 10105 frames for training, 1649 frames for validation. The data resulting for each class are summarized in Table 1.

Table 1  
Data resulting from the analysis of the tagging process.

Class code	Class #	Occurrences in the whole dataset
X <sub>left</sub>	0	1649
X <sub>right</sub>	1	1575
V <sub>left</sub>	2	3249
V <sub>right</sub>	3	839
O <sub>left</sub>	4	1113
O <sub>right</sub>	5	649
I <sub>left</sub>	6	2524
I <sub>right</sub>	7	771

The training and validation of the network was performed by a Nvidia GTX GeoForce 12GB Titan GPU. The already trained weights downloaded from the Darknet repository (Redmon and Farhadi, 2018) were used as initial weights for training and 10 000 iterations were performed to obtain the final (adjusted) weights.

## 2.6 Results assessment

In this subsection the metrics adopted for the evaluation of the performance of the system are presented. The first metric to introduce is the intersection-over-union (IoU) index, also known as Jaccard index (Rezatofighi et al., 2019), maybe the most commonly used metric for comparing the similarity between two general images. IoU encodes the properties of the items under comparison (e.g. widths, heights, locations of bounding boxes) and then calculates a normalized measure reported in Eq. (1) as the ratio between the area of the intersection divided by the union of the two bounding boxes (i.e. the predicted and the ground truth bounding boxes).

$$IoU = \frac{\text{Area of ground truth box} \cap \text{Area of predicted box}}{\text{Area of ground truth box} \cup \text{Area of predicted box}} \quad (1)$$

281 IoU results invariant to the problem scale and thanks to this feature the most of the performance measures in  
 282 segmentation, object detection, and tracking are based on this metric (Rezatofighi et al., 2019).  
 283 In pattern recognition applications, the precision ( $P$ ) is the fraction of relevant instances among the retrieved  
 284 instances, while the recall ( $R$ ) is the fraction of relevant instances that have been retrieved over the total  
 285 amount of relevant instances. Both precision and recall are therefore based on an understanding and measure  
 286 of the relevance. Precision ( $P$ ) and recall ( $R$ ) can be expressed, in analytical form, by means of the following  
 287 expressions:

$$288 \quad P = \frac{tp}{tp+fp} \quad (2)$$

$$289 \quad R = \frac{tp}{tp+fn} \quad (3)$$

290 Where:  $tp$  represents true positive, i.e., the number of cases that the detector successfully detects a class in  
 291 an image with IoU greater than a prescribed threshold;  $fp$  is false positive, i.e., the number of cases that the  
 292 detector reports other objects as a target class in an image, or IoU is less than a prescribed threshold;  $fn$  is  
 293 false negative, i.e. the number of cases that the detector fails to detect a target class in an image. In the present  
 294 work, the specific threshold has been fixed equal to 0.5. Precision is also known as “positive predictive value”,  
 295 while recall is also called “true positive rate” or “sensitivity” and this last represents the proportion of actual  
 296 positives are correctly identified.

297 In the interpretation of computer vision results, the balanced  $F_1$ -score (also F-score or F-measure) is a metric  
 298 that combines both  $P$  and  $R$  and represent the harmonic mean (Nie et al., 2019):

$$299 \quad F_1 = 2 \cdot \frac{P \cdot R}{P + R} = \left( \frac{P^{-1} + R^{-1}}{2} \right)^{-1} \quad (4)$$

300 This metric coincides with the square of the geometric mean divided by the arithmetic mean of precision and  
 301 recall and is clearly close to the arithmetic mean of the two when  $P$  and  $R$  have similar values.  $F_1$  reaches its  
 302 best value at 1.0 and the worst at 0.0.

303 Moreover, the confidence score  $C$  (%) has been considered, which quantifies the reliability of the recognition  
 304 of a given object within a frame. Confidence score can be calculated using the formula:

$$305 \quad C = Pr \times IoU \quad (5)$$

where:  $Pr$  represents the detection probability assessed by the network that the object at hand belongs to the class attributed to it.

Precision and recall have been computed for each class based on different confidence thresholds ranging from 0 to 100%, and the precision-recall curves have been drawn for each class. Besides, AP (average precision) is a popular metric measuring the accuracy of object detectors. AP computes the average precision value for recall value from 0 to 1 for a specific class (Szeliski, 2011). Therefore, AP can be computed as the area under the P-R curve of such class. Then, the mean average precision (mAP) of the network was assessed as the mean of the AP values of the different classes. With analogous criteria is possible to define the average IoU (AloU) for a specific class as the average of all the IoU values of the occurrences of the same class.

315

### 3. Results and Discussion

#### 3.1 With original data frames

This subsection deals with the main results obtained by considering the original set of frames described in the previous section. The set is constituted by 11754 frames, 10105 (about 85% of the total dataset) have been used for the network training whereas 1649 (about 15%) for the network validation test. The frames to be used in the validation phase were carefully selected in order to guarantee that all the 8 considered classes were adequately represented in the frames. The occurrences of each class are reported in Table 2, for both training and validation phases. Obviously, the sum of the occurrences of all the classes, i.e. 10167 and 2202 respectively for training and validation phases, are bigger than the total frame number since some frames includes multiple cows belonging to different classes.

326

Table 2

Number of occurrences and average area of the bounding box for each class and for both training and validation original datasets.

Training dataset									
	$X_{left}$	$X_{right}$	$V_{left}$	$V_{right}$	$O_{left}$	$O_{right}$	$I_{left}$	$I_{right}$	Sum
$O_t$	1325	1273	2599	673	904	512	2270	611	10167
$A_t$	0.01063	0.00920	0.00540	0.01688	0.00538	0.01567	0.00447	0.00502	-
$O_t \times A_t$	14.08	11.71	14.03	11.36	4.86	8.02	10.15	3.07	-
Validation dataset									



	$X_{left}$	$X_{right}$	$V_{left}$	$V_{right}$	$O_{left}$	$O_{right}$	$I_{left}$	$I_{right}$	Sum
$O_v$	324	302	650	166	209	137	254	160	2202
$A_v$	0.02748	0.01540	0.00356	0.01306	0.01542	0.01364	0.01552	0.01275	-
$O_v \times A_v$	8.90	4.65	2.31	2.17	3.22	1.87	3.94	2.04	-

330

331 The validation of the computer vision detection could be carry out from a visual (or graphical) point of view,  
332 by looking if in one specific frame the classes object of the test (in this case the 8 hips of the 4 cows) are  
333 properly identified by the neural network. For example, Figure 6 shows in the yellow box, in the magenta box  
334 and in the green box, the identification of the left hip respectively of the cow O, cow X and cow V. The accuracy  
335 of the detection, reported in the rectangle at top-right of the figure, represent the confidence score C for each  
336 class, as calculated by YOLO, and could be correlated to the probability of finding the cow in the bounding box.  
337 For the frame in the Figure 6, for example, the detection results very good since the class  $O_{left}$  and  $V_{left}$  have C  
338 of 100% whereas  $X_{left}$  has C about 98%.



339

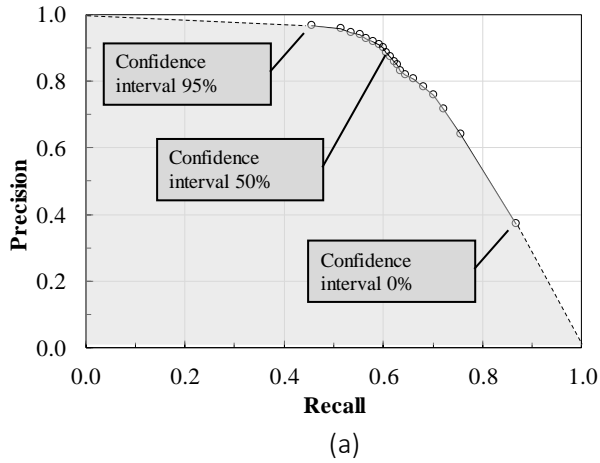
340 Figure 6. Example of visual validation of the classes in a frame. The yellow box ( $O_{sx}$ ) is the identified left hip  
341 of the cow marked with the letter O; the magenta box ( $X_{sx}$ ) is the identified left hip of the cow X and the  
342 green box ( $V_{sx}$ ) is the identified left hip of cow V.

343

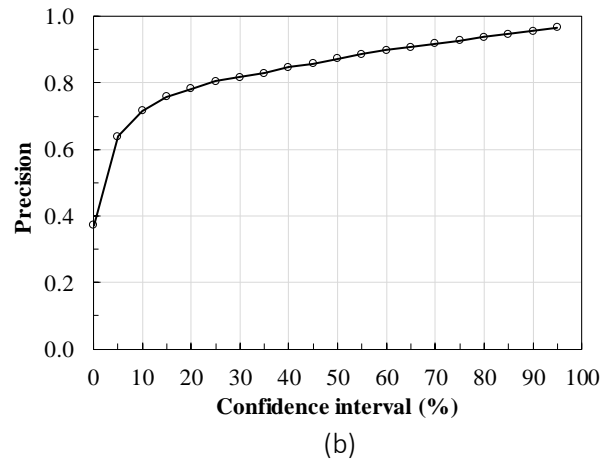
344 In addition, the validation of the neural network can be performed from a mathematical point of view on the  
345 basis of global P and R scores (i.e. considering the whole data set coming from the various classes). As an

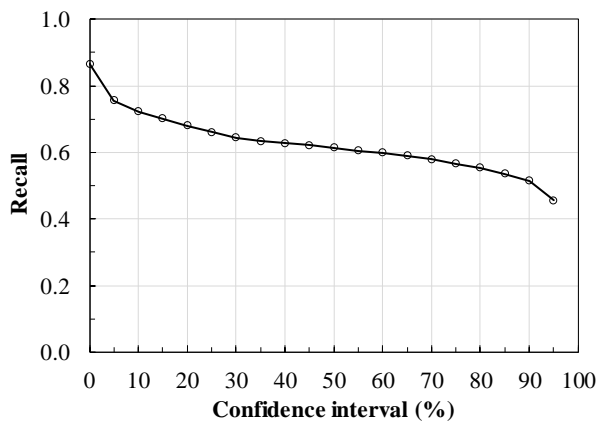
346 example, Figure 7(a) shows the global P-R graph (i.e. considering the dataset of the all 8 classes considered in  
 347 the study) of the validation test. The trend of the global P-R graph was obtained by considering 20 different  
 348 confidence interval (CI) with increasing confidence level, from 0.0% to 95% with step 5%, for the assessment  
 349 of the detections based on the IoU of the single detection. As a general trend, the lower confidence interval  
 350 produces points with low P and high R values. The opposite for high confidence interval. It seems useful to  
 351 remember that the optimal graph trend should be that presenting high level of precision (i.e. higher than 0.8)  
 352 all along the R value. Figures 7(b) and 7(c) respectively show the trend of P and R for the different CI values.  
 353 For the case at hand the P value is adequate (i.e. higher than 0.8) for CI higher than 20%. For CI equal to 20%  
 354 the R value is about 0.7 and it means the neural network is able to detect about the 70% of the “real”  
 355 occurrences of the different classes for the various frames. Table 3 reports the main data resulting from the  
 356 validation phase and related to the whole dataset. Moreover, same table reports the global  $F_1$ -score and IoU  
 357 values for each CI, also depicted in Figure 7d. IoU values go from 0.75 to 0.81 with an AIOU equal to 0.78. In  
 358 the object detection field values higher than 0.7 until 1.0, commonly, identified detections good to excellent,  
 359 then we reached, on average, a rather good detection from the network.

360

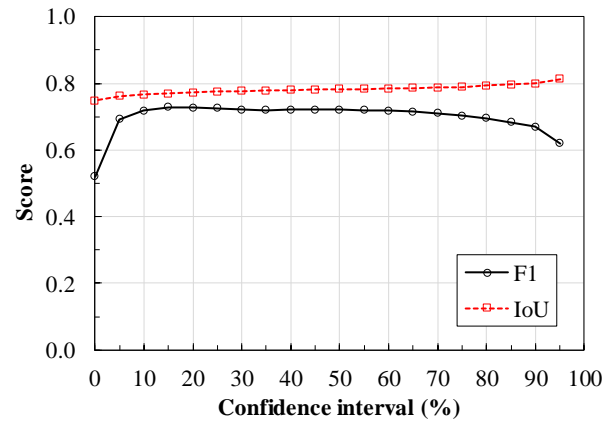


361  
 362





(c)



(d)

Figure 7. Principal trends obtained from the validation test by considering all the occurrences dataset and reported for different confidence interval (with original data frames). (a) P-R curve. (b) P trend; (c) R trend; (d)  $F_1$  and IoU score for different confidence interval.

Table 3  
Main results from the validation test by considering all the occurrences dataset and reported for different confidence interval (with original data frames).

Confidence interval	True positive*	Ground truth**	Precision	Recall	F <sub>1</sub> -score	IoU
0.95	1004	1039	0.9663	0.4559	0.6196	0.8126
0.90	1134	1186	0.9562	0.5150	0.6694	0.8000
0.85	1178	1245	0.9462	0.5350	0.6835	0.7960
0.80	1218	1299	0.9376	0.5531	0.6958	0.7928
0.75	1247	1346	0.9264	0.5658	0.7026	0.7891
0.70	1277	1390	0.9187	0.5795	0.7107	0.7874
0.65	1301	1434	0.9073	0.5904	0.7153	0.7853
0.60	1319	1467	0.8991	0.5985	0.7187	0.7838
0.55	1333	1505	0.8857	0.6049	0.7189	0.7833
0.50	1353	1551	0.8723	0.6140	0.7207	0.7821
0.45	1370	1596	0.8584	0.6217	0.7211	0.7811
0.40	1382	1631	0.8473	0.6272	0.7208	0.7803
0.35	1398	1686	0.8292	0.6344	0.7188	0.7790
0.30	1422	1740	0.8172	0.6449	0.7209	0.7774
0.25	1458	1812	0.8046	0.6612	0.7259	0.7752
0.20	1500	1918	0.7821	0.6803	0.7276	0.7723
0.15	1548	2040	0.7588	0.7012	0.7289	0.7700
0.10	1598	2234	0.7153	0.7221	0.7187	0.7665
0.05	1681	2632	0.6387	0.7566	0.6926	0.7617
0.00	2011	5409	0.3718	0.8665	0.5203	0.7480

\* : is the number of true positive occurrences detected from the neural network

\*\* : is the number of “real” occurrences in the dataset as resulting from the visual detection performed by the operator.

As far as the single class is concerned, Figure 8 shows the Precision-Recall graphs for every considered class and in Table 4 the most important parameters are collected, useful to judge the detection quality of each single class. In fact, if in some contexts an “on-average” detection score is sufficient (Szeliski, 2011). In the present applications, it is not enough being the single class detection score, important as much the “on-average” score for practical PLF purposes. Then, it is possible to identify in Figure 8, and evaluate from Table 4, the classes with better/worse detection scores. E.g., from the table, the classes with better AP are  $V_{left}$ ,  $X_{right}$  and  $O_{left}$ . Instead, the classes with the worst AP are  $O_{right}$ ,  $I_{right}$  and  $I_{left}$ .

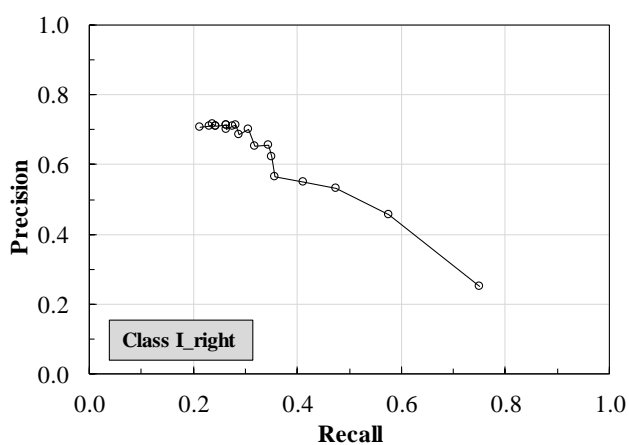
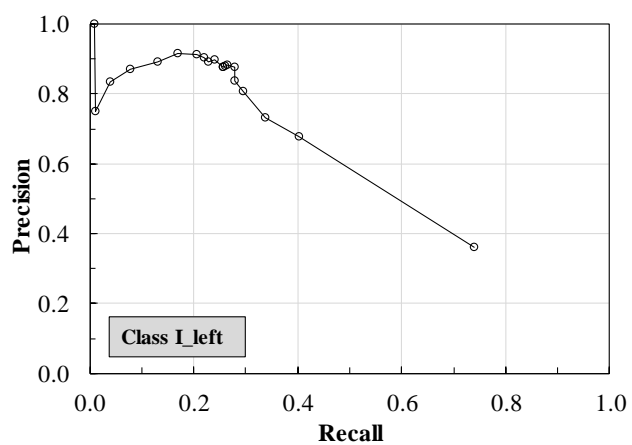
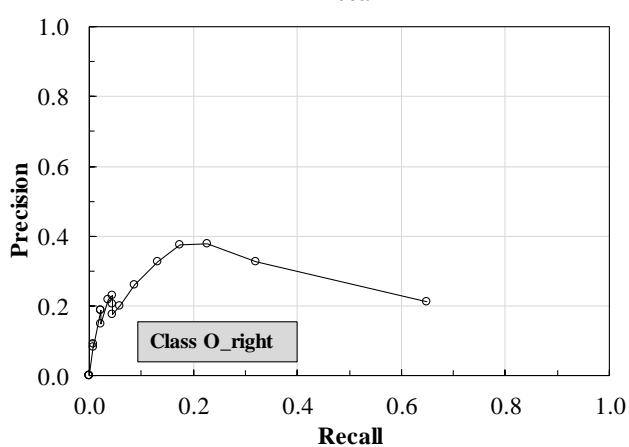
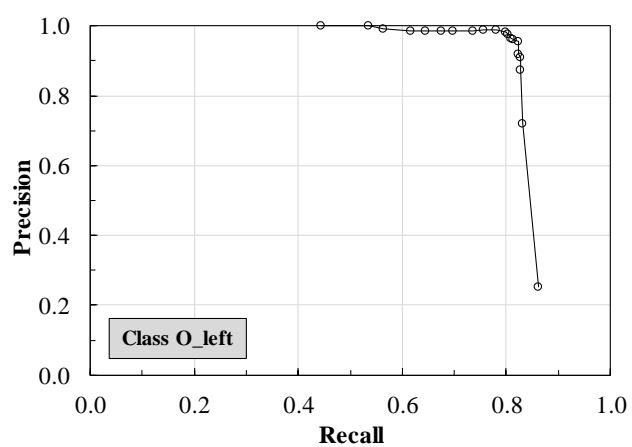
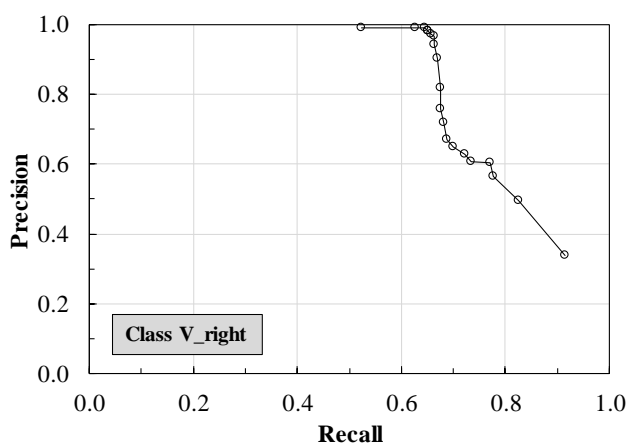
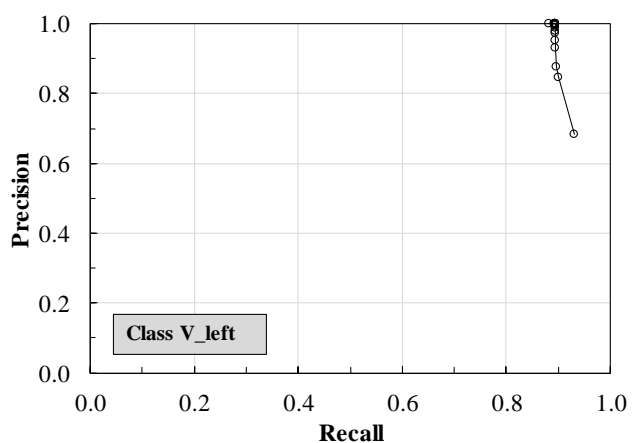
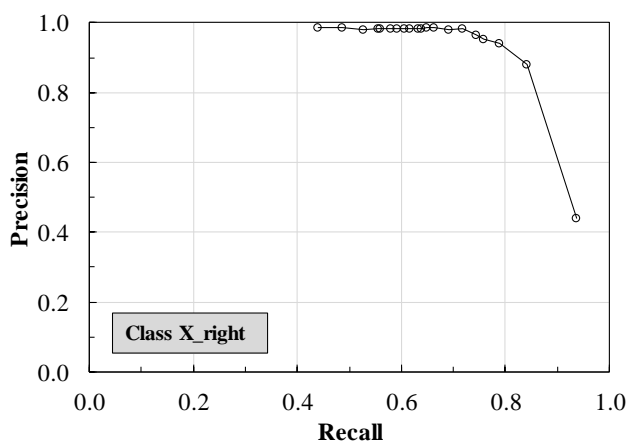
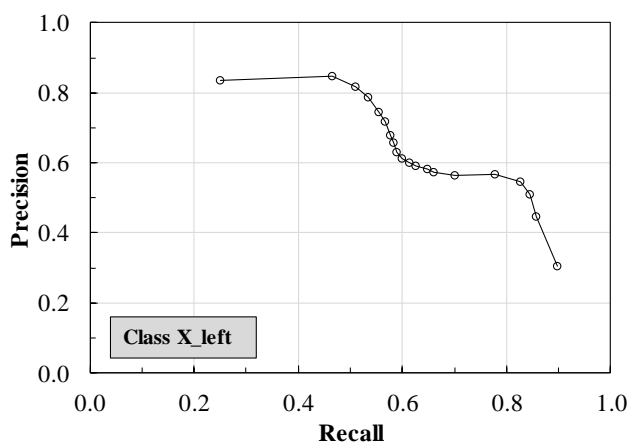


Figure 8. Precision-Recall diagram of the computer vision detection for each one of the 8 target classes by using the original data frames.

Even if the AP of the total dataset is 0.7356, some classes have AP value also rather small (i.e.  $O_{right}$ ,  $I_{right}$  and  $I_{left}$  with AP values respectively of 0.17, 0.40 and 0.45). The low values for the three worst classes are confirmed by the unusual trends in Figure 8. This analysis should drive the future investigations, oriented them to define the main causes of the low scores and to look for the adequate corrective actions. Differently from the previous discussed scores, the AIoU values are quite similar for the various classes are the AIoU-based identification of the worst classes is more difficult. This confirms some well-known weaknesses and limitations of this metric (Szeliski, 2011) useful to decide, with regards to an object, whether a prediction is correct or not, but it is not suitable to describe the precision of the prediction. For the sake of completeness, the conducted validation test provides a value of  $mAP=0.6350$  obtained as mean value among the 8 classes.

Finally, in order to establish possible correlations between the main features of the dataset and the outcomes from the validation test, the correlation matrix, reported in Figure 9, has been realized and adopted for the evaluation. Six independent variables, numerically quantifiable, have been selected (i.e. Occurrences, Average Area of the bounding boxes and the product Occurrences  $\times$  Average Area of the bounding boxes for both training and validation datasets). Two dependent variables have been selected among the metrics adopted to evaluate the reliability of the detections (i.e. AP and AIoU).

Then the correlation matrix has dimension  $8 \times 8$ . The numerical values adopted for the creation of the correlation matrix are those reported in Table 2 and Table 4.

Table 4  
Summary of the results from the validation test for each class by considering the original data frames.

	$X_{left}$	$X_{right}$	$V_{left}$	$V_{right}$	$O_{left}$	$O_{right}$	$I_{left}$	$I_{right}$	Total
AP	0.6485	0.8627	0.9202	0.7838	0.8336	0.1698	0.4583	0.4027	0.7356
AIoU	0.6625	0.7600	0.8547	0.7691	0.7553	0.5232	0.6999	0.7333	0.7812

The main aspects are the following:

- (see column #1, row #4,) the occurrences of the different classes populating training and validation datasets, shows good correlations confirming a proper subdivision of the frames into the training group and validation group (the slope of the linear regression represent the ratio  $20\%/80\%=0.25$  of used for the subdivision of the whole frame dataset);

- 415 • (see column #1, rows #7, #8 and #9) a characteristic trend exists between number of occurrences in the  
416 training dataset and the metrics used for the evaluation of the detection quality (i.e. AP and AIoU). By  
417 increasing the number of occurrences until a certain “threshold” value it increases in a considerable way  
418 the metric values, but after this threshold, the trend (see dashed red line in the subfigures) presents a  
419 knee characterized by a second branch with low slope. Then, it is like to say that after certain threshold  
420 (occurrences) value a considerable augment of the number of occurrences produces almost negligible  
421 improvement in the detections;
- 422 • (see column #4, rows #7, #8 and #9) the same evaluation is valid also for the relation between number of  
423 occurrences in the validation dataset and the metrics used for the evaluation of the detection quality;
- 424 • the area of the bounding boxes shows no significant correlation with the metrics of reliability of the  
425 detections (see columns #2 and #5). Nevertheless, the features obtained by the product “Occurrences ×  
426 Average Area of the bounding boxes” (see columns #3 and #6) shows a positive correlation with the  
427 metrics, although it is not possible to identify a clear regression curve. However, if we exclude the two  
428 classes with values of the product  $O_t \times A_t$  of the training phase significantly smaller than all the other  
429 classes, i.e.  $O_{left}$  and  $I_{right}$ , a quadratic regression curve of the relationship between AP and  $O_t \times A_t$  is  
430 recognizable, with  $R^2=0.837$  and the following equation:

$$431 \quad y = -0.0272 x^2 + 0.7111 x - 3.8179 \quad (6)$$

432 This indicates that for the classes having  $O_t \times A_t > 8$ , AP rapidly increases up to  $O_t \times A_t$  near to 12, then it  
433 becomes almost constant. This result shows that is possible to identify optimal values for the occurrence  
434 number and the bounding box areas in the training phase, which could be very useful to efficiently plan  
435 video acquisition to use for train the deep learning network. Therefore, this aspect deserves further and  
436 more in-depth investigations that will be carried out in future experimental campaigns carried out with  
437 additional video shooting.

- 438 • lastly (see columns #7, row #8) a positive correlation is confirmed between AP and AIoU.

439 These results provide useful indications for both the selection of the strategies more convenient in order to  
440 improve the detection quality and the development of optimal datasets for the application investigated in this  
441 paper.

442 First of all, if some classes are detected with poor precision, increasing the occurrences in the training dataset  
443 the detection accuracy is expected to rise. It makes no sense to increase the occurrences of all classes,  
444 especially for those classes that already have suitable metric values, because this would increase the costs of  
445 the labelling phase and the computational time of the training phase without producing considerable  
446 improvements.

447 Above a certain threshold value of the metrics, it also seems that increasing even the occurrences of the  
448 training dataset does not produce considerable benefits. So in such cases, probably alternative solutions are  
449 to be sought, such as the replacement of some data frames with others more informative for the network.

450 Finally, it seems that the average area of bounding boxes used for animal detection is not so influential, while  
451 the product of number of occurrences and box area is positively correlated with the average precision. This  
452 from a practical point of view has considerable advantages for the type of application investigated here, since  
453 typically in facilities such as cattle barns the videos are taken from considerable distance (even several tens of  
454 meters) due to logistical and security reasons of the cameras. On the other hand, the possibility to record  
455 videos from far away means that few cameras could be sufficient to cover large areas typical of cattle farms.

456 The evaluation and the development of these further steps in the process of developing the applications of  
457 computer vision systems to the dairy cow sector will be the first objectives of future research work.



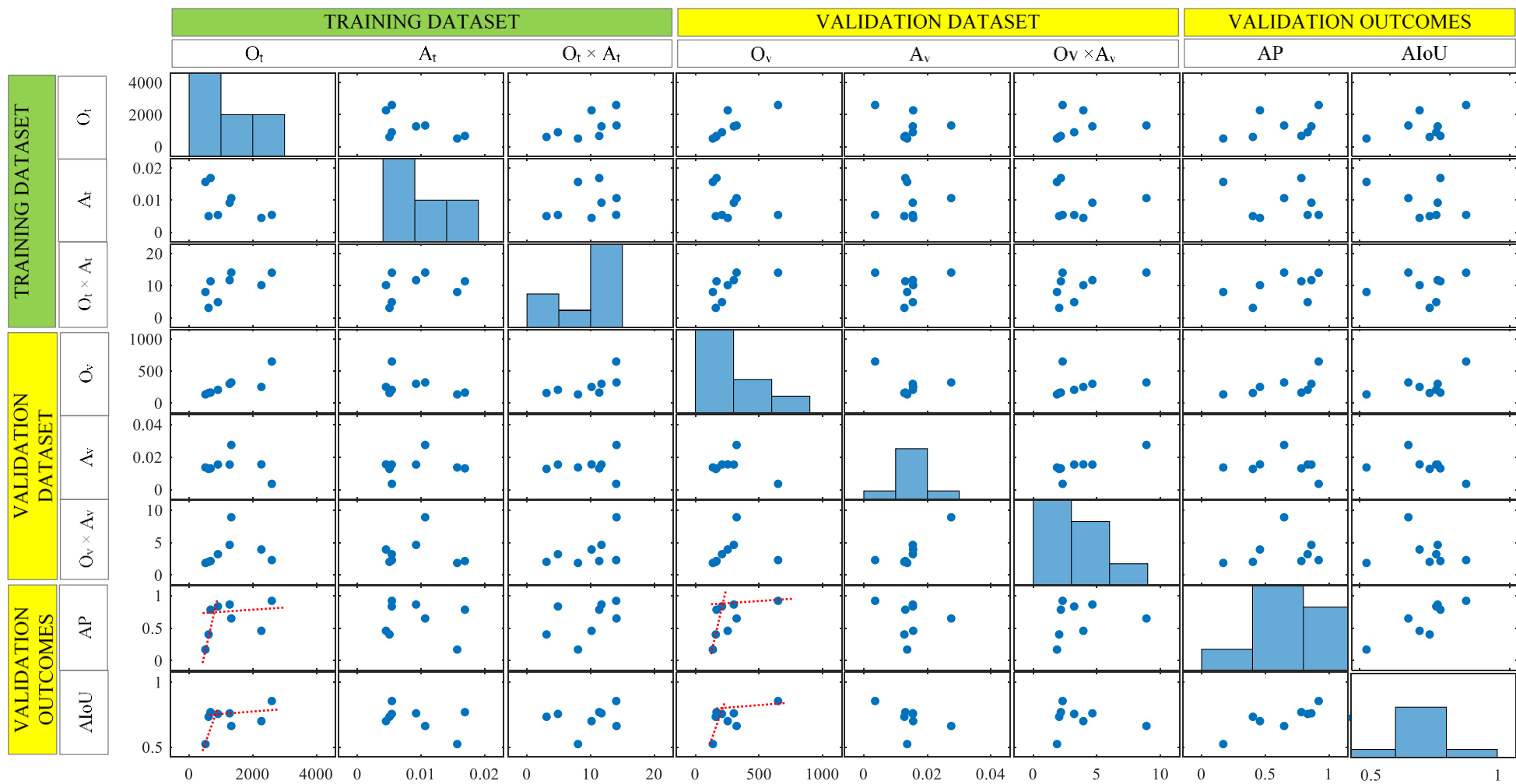


Figure 9. Correlation matrix between some features of the training dataset (i.e.  $O_t$ ,  $A_t$  and  $O_t \times A_t$ ) and validation dataset (i.e.  $O_v$ ,  $A_v$  and  $O_v \times A_v$ ) selected as independent variables and some metric outcomes of the validation phase selected as dependent variables.

### 462 3.2 *With augmented data frames*

463 This subsection presents a first preliminary attempt to increase the detection quality of some classes. In this  
464 case, the four classes with lowest total occurrence number (i.e.  $V_{right}$ ,  $O_{left}$ ,  $O_{right}$  and  $I_{right}$ ) have been selected  
465 and by adopting a procedure of data augmentation, their total occurrence number have been increased. The  
466 two main objectives of the data augmentation test were:

- 467 i) to understand if simple alteration of the original frames can constitute a viable method to increase the  
468 available frame datasets in the context of cow detection and, following authors' knowledge, this represent  
469 the first application of this type of methods in the herd monitoring research field;  
470 ii) to estimate the improvement of the detection performances of the network, in terms of AP, connected to  
471 the augment of the number of occurrences.

472 In order to judge the possible relation between augment of the number of occurrences and AP improvement,  
473 different threshold values have been investigated for the four classes. The following parameter  $\Delta_o$  (%) has  
474 been adopted for the identification of the augment of the number of occurrences:

$$475 \Delta_o (\%) = \text{Augment}_t / O_t \times 100 \quad (7)$$

476 where:  $\text{Augment}_t$  is the increase of the occurrence number respect to the original dataset in the frames  
477 used for the training;  $O_t$  is the occurrence number in the frames of the original dataset used in the training.

478 Four different target ranges have been considered for  $\Delta_o$  (three below 50%, i.e. 10-20%; 20-30%; 30-40% to  
479 test the low rates of increase and one above, i.e. 60-80%) and every class has been associated to one of the  
480 ranges in order to increase the number of occurrences of the different classes in different ways. In this way  
481 has been possible to estimate a correlation between increase of occurrences and increase of detection  
482 performances when artificial frames are added in the dataset.

483 The data augmentation procedure has selected some frames, randomly extracted among those containing the  
484 four classes indicated above, and artificially have produced a modified copy of every selected frames. The  
485 modified copy has been obtained by changing the brightness level of the original frame so simulating possible  
486 different light conditions. This procedure, performed with the software XnConvert (Allan et al., 2019), creates  
487 a series of modified frames that could really occur in the stable. As an example, Figure 10 shows the

comparison between the original and modified frame created by means of the described procedure. The modified frames have been added to the original dataset (with 10105 frames) in order to constitute the augmented dataset (with 11024 frames in total).



(a)

(b)

Figure 10. Comparison between (a) the original frame as recorded and (b) the modified frame created modifying the brightness of the image.

Also in this case the frames to be used in the validation phase were carefully selected in order to guarantee that all the 8 considered classes were adequately represented in the frames. The occurrences of each class, in the augmented dataset have been reported in Table 5, for both training and validation phases. The table also collects the augment of the occurrences, for each class, with respect to the original dataset. The augment of occurrences have generated  $\Delta_0$  values equal to 29.0%, 15.9%, 71.3% and 37.5% for  $V_{right}$ ,  $O_{left}$ ,  $O_{right}$  and  $I_{right}$  respectively. In total, the augments of occurrences have been 1048, 158 and 1206 respectively for training, validation and total datasets. It is worth to highlight that original dataset, in terms of occurrences, has been augmented of about 10%, with the major increases related to training dataset of the classes  $V_{right}$ ,  $O_{left}$ ,  $O_{right}$  and  $I_{right}$ .

507 Table 5

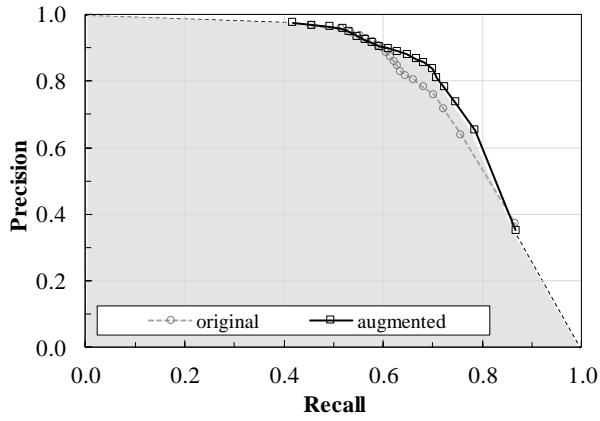
508 Number of occurrences for each class for both training and validation augmented datasets. Augment is the  
 509 increase on the occurrence number respect to the original datasets.

Training dataset									
	X <sub>left</sub>	X <sub>right</sub>	V <sub>left</sub>	V <sub>right</sub>	O <sub>left</sub>	O <sub>right</sub>	I <sub>left</sub>	I <sub>right</sub>	Sum
O <sub>t</sub>	1383	1384	2668	868	1048	877	2147	840	11215
Augment <sub>t</sub>	58	111	69	195	144	365	-123	229	1048
Validation dataset									
	X <sub>left</sub>	X <sub>right</sub>	V <sub>left</sub>	V <sub>right</sub>	O <sub>left</sub>	O <sub>right</sub>	I <sub>left</sub>	I <sub>right</sub>	Sum
O <sub>v</sub>	331	302	656	176	218	139	377	161	2360
Augment <sub>v</sub>	7	0	6	10	9	2	123	1	158

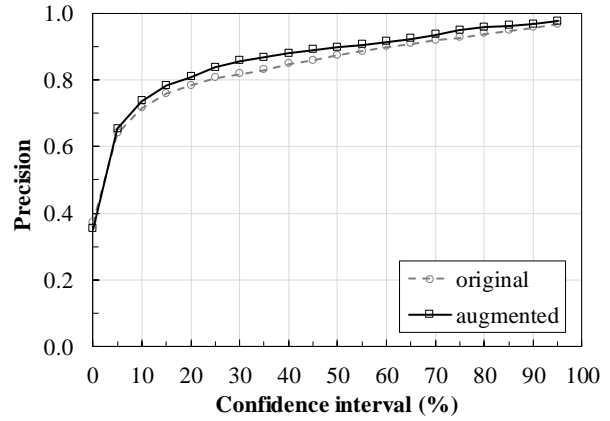
510

511 Then, the process follows the same steps used for the analysis on the original dataset, and for the sake of  
 512 comparison with the previous case, analogous graphs and tables are reported in the following in order to  
 513 summarize the main results. Figure 11(a) reports the global P-R graphs of the validation test obtained for both  
 514 original and augmented datasets by considering all the 8 classes. From the comparison between the two  
 515 curves it emerges that also the introduction of very similar frames, which differ only in brightness from the  
 516 original, can improve both the precision (P) and the network's detection quality. In fact, for the present  
 517 dataset, the precision improves in the CI range from 10% to 50% (see Figure 11(b)). Conversely, the recall (R)  
 518 has an anti-symmetric trend with respect to CI of 50%. It improves in the CI from 0% to 50% and slightly  
 519 deteriorates for CI higher than 50% (see Figure 11(c)). Same trend is obtained for the F<sub>1</sub>-score (see Figure  
 520 11(d)). As far as the IoU metrics is concerned, it decreases about 8-9% all along the CI set. For the case at hand,  
 521 the P value is adequate (i.e. higher than 0.8) for CI higher than 15%. For CI equal to 15% the R value is about  
 522 0.75. Table 6 collects all the results graphically reported in Figure 11.

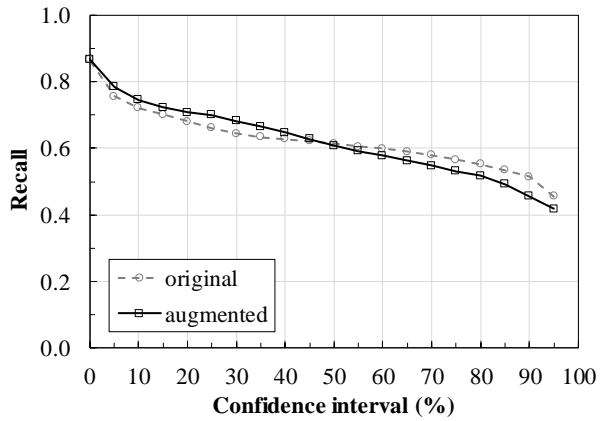
523



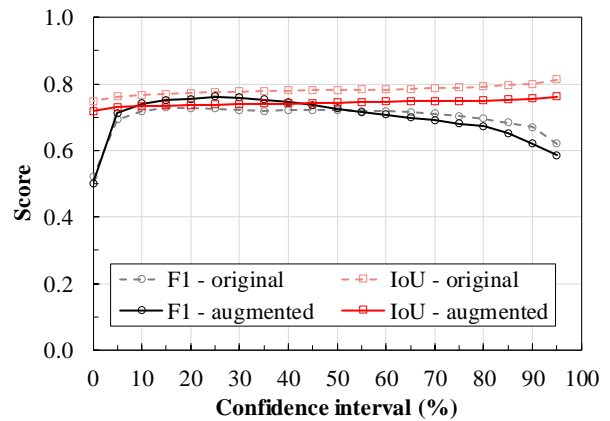
(a)



(b)



(c)



(d)

Figure 11. Comparison between the principal trends obtained from the validation test by considering all the occurrences dataset and reported for different confidence interval (with both original and augmented data frames). (a) P-R curve. (b) P trend; (c) R trend; (d)  $F_1$  and IoU score for different confidence interval.

As far as the single class is concerned, in Table 7 are collected the parameters used to judge the detection quality of the various classes and Figure 12 shows the Precision-Recall graphs of each single class. From the table, the classes with better AP are  $V_{\text{left}}$ ,  $O_{\text{left}}$  and  $V_{\text{right}}$ . Instead, the classes with the worst AP are  $O_{\text{right}}$ ,  $I_{\text{right}}$  and  $X_{\text{left}}$ .

537 Table 6

538 Main results from the validation test by considering all the occurrences dataset and reported for different  
 539 confidence interval (with augmented data frames).

Confidence interval	True positive*	Ground truth**	Precision	Recall	F <sub>1</sub> -score	IoU
0.95	986	1011	0.9753	0.4178	0.5850	0.7615
0.90	1076	1113	0.9668	0.4559	0.6196	0.7563
0.85	1162	1207	0.9627	0.4924	0.6515	0.7533
0.80	1223	1278	0.9570	0.5182	0.6723	0.7506
0.75	1253	1322	0.9478	0.5309	0.6806	0.7489
0.70	1293	1383	0.9349	0.5479	0.6909	0.7484
0.65	1328	1438	0.9235	0.5627	0.6993	0.7477
0.60	1363	1491	0.9142	0.5775	0.7079	0.7463
0.55	1396	1544	0.9041	0.5915	0.7152	0.7454
0.50	1437	1601	0.8976	0.6089	0.7256	0.7439
0.45	1483	1668	0.8891	0.6284	0.7363	0.7424
0.40	1529	1739	0.8792	0.6479	0.7460	0.7410
0.35	1571	1812	0.8670	0.6657	0.7531	0.7403
0.30	1609	1880	0.8559	0.6818	0.7590	0.7392
0.25	1652	1976	0.8360	0.7000	0.7620	0.7377
0.20	1671	2067	0.8084	0.7081	0.7549	0.7371
0.15	1707	2183	0.7820	0.7233	0.7515	0.7351
0.10	1760	2391	0.7361	0.7458	0.7409	0.7329
0.05	1854	2843	0.6521	0.7852	0.7125	0.7296
0.00	2090	5927	0.3526	0.8669	0.5013	0.7184

540 \* : is the number of true positive occurrences detected from the neural network

541 \*\* : is the number of “real” occurrences in the dataset as resulting from the visual detection performed  
 542 by the operator.

544 Table 7

545 Summary of the results from the validation test for each class by considering the augmented data frames.

	X <sub>left</sub>	X <sub>right</sub>	V <sub>left</sub>	V <sub>right</sub>	O <sub>left</sub>	O <sub>right</sub>	I <sub>left</sub>	I <sub>right</sub>	Total
AP	0.5489	0.8755	0.9405	0.8815	0.8954	0.2405	0.7618	0.5391	0.7559
AIoU	0.7034	0.7593	0.7481	0.8153	0.7221	0.6392	0.6635	0.7865	0.7428

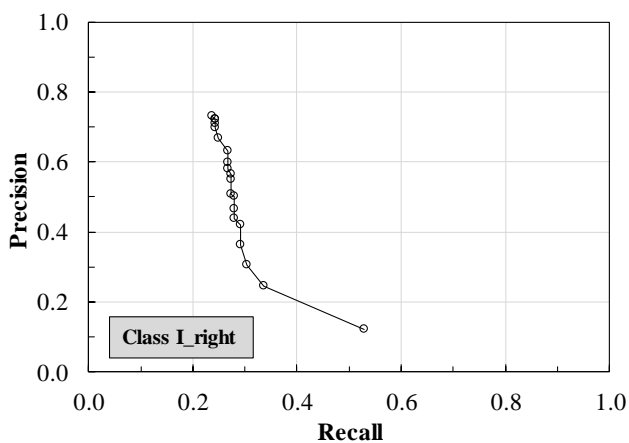
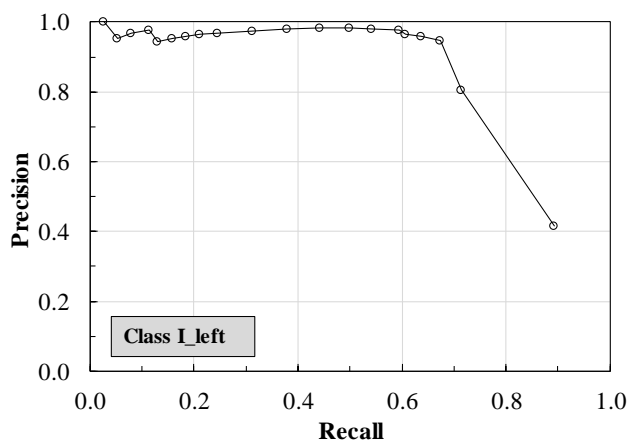
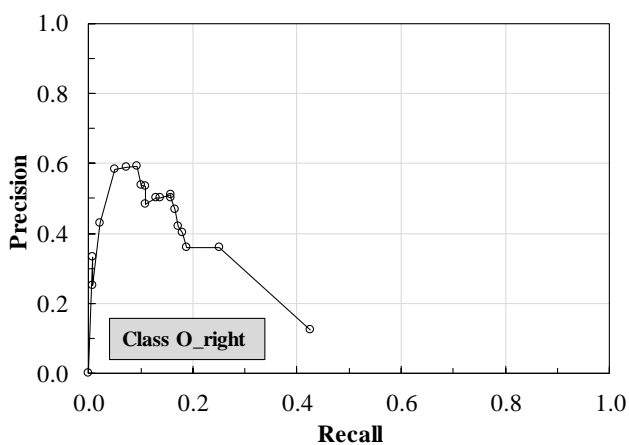
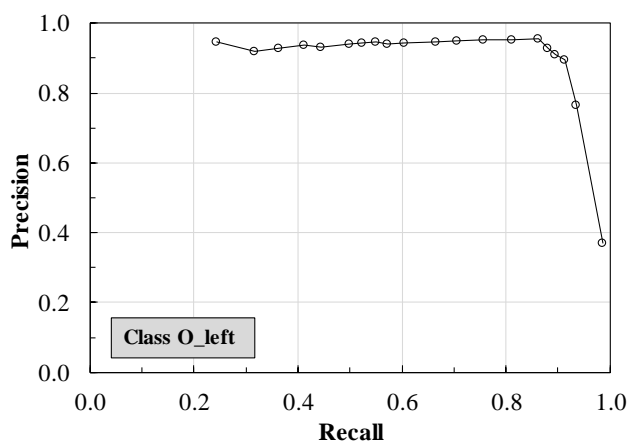
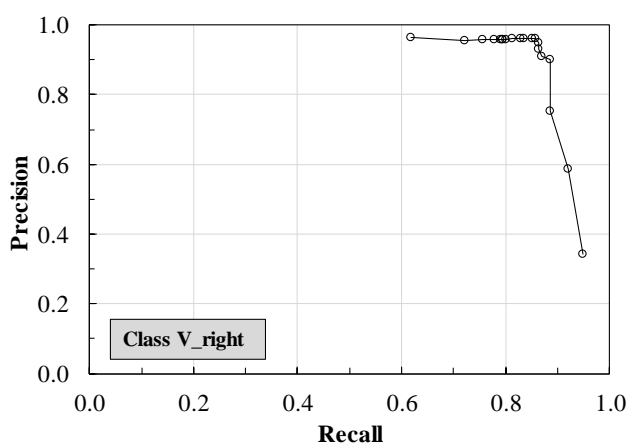
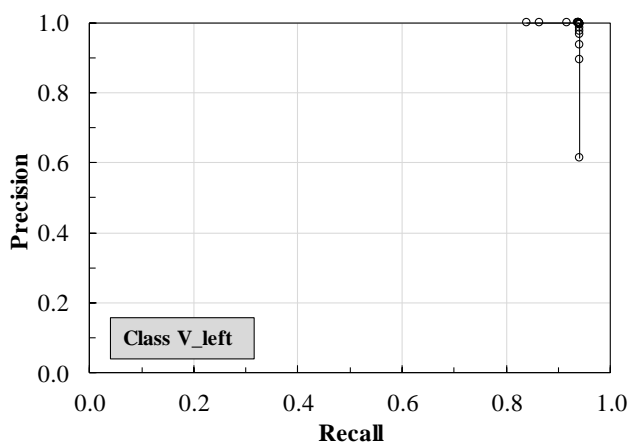
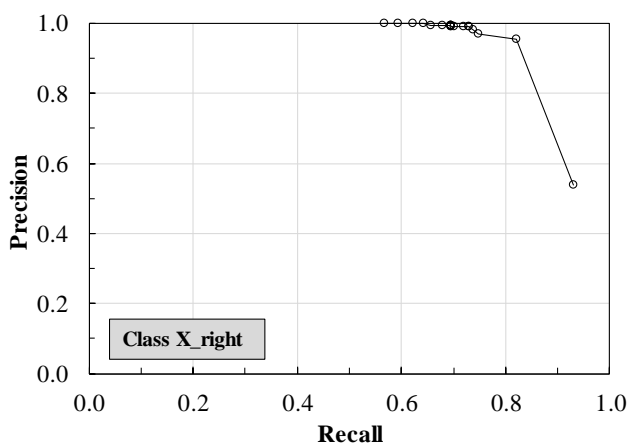
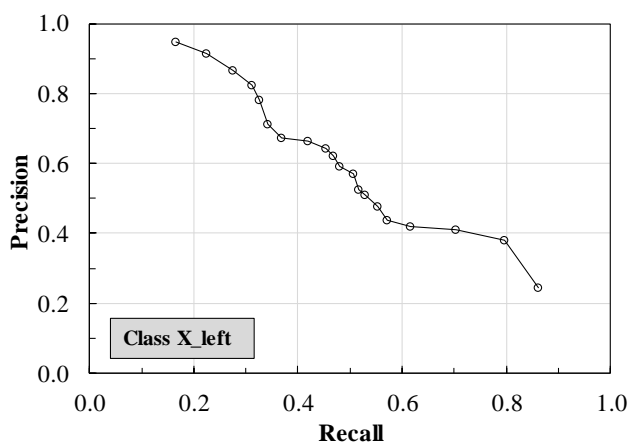


Figure 12. Precision-Recall diagram of the computer vision detection for each one of the 8 target classes by using the augmented data frames.

In order to compare the network detection performances of the original and augmented data frame cases, the following parameter  $\Delta_{AP}$  (%) has been defined and calculated for the four classes majorly influenced by the data augmentation procedure:

$$\Delta_{AP} (\%) = (AP_{augmented} - AP_{original}) / AP_{original} \times 100 \quad (8)$$

where:  $AP_{augmented}$  is the average precision obtained using the augmented dataset;  $AP_{original}$  is the average precision obtained using the original dataset.

The values of  $\Delta_O$  and  $\Delta_{AP}$ , representing a (percentage) relative difference of occurrences and average precision respectively, are reported in Table 8 for the four classes of interest.

564

Table 8

Percentage relative difference of occurrences ( $\Delta_O$ ) and average precision ( $\Delta_{AP}$ ) for the four classes mainly influenced by the data augmentation procedure.

	$V_{right}$	$O_{left}$	$O_{right}$	$I_{right}$
$\Delta_O$ (%)	29.0	15.9	71.3	37.5
$\Delta_{AP}$ (%)	12.5	7.4	41.6	33.9

568

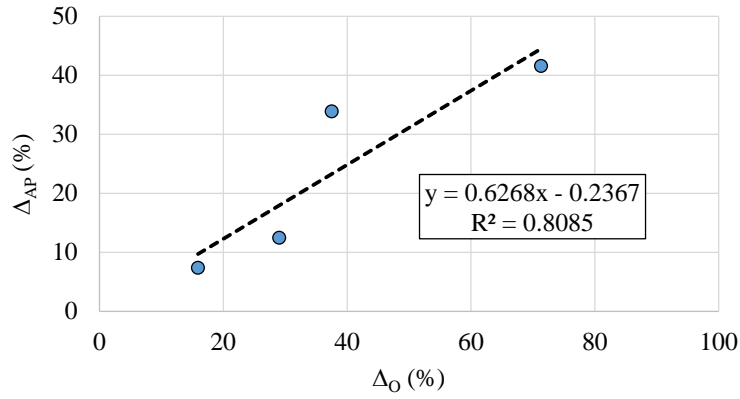
By the analysis of the  $\Delta_{AP}$  values in Table 8, it emerges that the augmented dataset is able, in general, to provide an improvement of the metric values, or i.e. is able to increase also the single class detection quality of the neural network, especially for those classes characterized by a considerable augment of occurrences. This was probably expected, but the most important outcomes may be that, also for the single classes, the introduction of artificially-obtained frames is suitable to increase the detection scores. This aspect has important practical implications since in applications similar to the one studied here it is not always possible to record videos for long periods.

Figure 13 graphically shows the position of the values of  $\Delta_O$  and  $\Delta_{AP}$  reported in Table 8. The further emerging outcome is that exist a clear direct relation between the increase of the occurrences number and the detection performances of the network and the relation seems almost linear, at least for the investigated ranges.

579

580





581

582 Figure 13. Percentage relative difference of occurrences ( $\Delta_O$ ) Vs. average precision ( $\Delta_{AP}$ ) for the four classes  
583 mainly influenced by the data augmentation procedure.

584

585 These outcomes confirm that for the situations in which the number of frames is not sufficient to provide a  
586 suitable occurrence number for some classes, in order to increase the detection performance of the network,  
587 a reliable strategy could be augment the dataset by adopting “virtual” frames made in-house and ad-hoc for  
588 the single classes not adequately represented in the acquired videos. This will be object of future investigations  
589 and further details are not reported here since they are beyond the scope of this paper. For the sake of  
590 completeness, the conducted validation test provides a value of mAP=0.6604, higher than the value obtained  
591 for the original dataset.

592 The developed system is suitable to implement various future extensions performing tracking and  
593 identification of anomalous behaviours. Cows tracking can be achieved by the integration of algorithms  
594 capable to compute the IoU between chronologically subsequent frames. IoU rate can be interpreted as  
595 displacement of the animal if it overcomes a predefined threshold. This will allow to record the trajectories of  
596 individual animals and also to compute the time spent in different positions, thus providing an accurate proxy  
597 of the time budget of every cow. Furthermore, this approach can be also used to detect standing time and  
598 lying bouts, which can be used to assess welfare indices of the herd, or different groups or even individual  
599 animals. Moreover, the evaluation of a different detection algorithm will be also object of future investigations.

600

601

#### 602 4. Conclusions

603 This study represents the first step for the development of a detection system aiming to recognize individual  
604 cows, evaluate their position, understand the action the cow is carrying out and finally tracking the cow  
605 movements in the barn. A computer vision system based on deep learning models for the automatic detection  
606 of individual cow based on the pelt pattern, within images using an HD resolution camera, was designed and  
607 implemented in a case study barn.

608 The global detection performances of the network, reported in terms of precision-recall curves, have been  
609 proved to be good for some classes and excellent for others since the AIoU is about 0.78 and the IoU for the  
610 different classes ranges from 0.75 to 0.81. The performances of the network is confirmed by the  $F_1$ -score  
611 ranging from 0.67 to 0.73 for common confidence interval from 5% to 90%. The outcomes proved that the  
612 natural pattern of the cow pelt investigated in the study is suitable for the animal detection, all this  
613 representing a necessary step prior to understand the cow action. Moreover, in this study a useful still simple  
614 equation has been proposed for the evaluation of the optimal values for the occurrence number and the  
615 bounding box areas in the training phase. The regression procedure for the equation calibration showed that  
616 a quadratic relation exists between AP and  $O_t \times A_t$ . This proposal could be very useful to efficiently plan the  
617 acquisition to use for train the network in this type of context. Finally, the application of a very simple data  
618 augmentation technique, changing the frame brightness level, has been confirmed to be an effective strategy  
619 to adopt in order to improve the performances of the network, in case of insufficient occurrences.

620 The promising results reported here present the first phase of the work for the definition of a computer vision-  
621 based system for herd monitoring applications devoted to the study of movements, actions and behavior of  
622 the cows in a barn.

623

624    **Acknowledgements**

625    The authors wish to thank Prof. Andrea Formigoni, Scientific Supervisor of the Experimental and Didactic Dairy  
626    Cows Unit of the University of Bologna, where all the video recordings, images and data necessary for carrying  
627    out the research have been acquired.

628

629    **Funding**

630    The activity presented in the paper is part of the research project PRIN 2017 “Smart dairy farming: innovative  
631    solutions to improve herd productivity” funded by the Italian Ministry of Education, University and Research  
632    [20178AN8NC].

633

634

## 635 References

- 636 AFIMILK, 2020. Cow monitoring. website. URL <https://www.afimilk.com/cow-monitoring> (accessed  
637 5.15.20).
- 638 Allan, E.L., Livermore, L., Price, B.W., Shchedrina, O., Smith, V.S., 2019. A Novel Automated Mass  
639 Digitisation Workflow for Natural History Microscope Slides. *Biodivers. data J.* 7, e32342–e32342.  
640 <https://doi.org/10.3897/BDJ.7.e32342>
- 641 ALLFLEX, 2020. Herd monitoring. website. URL <https://westgen.com/products/scr-herd-monitoring>  
642 (accessed 5.20.20).
- 643 Alsaaod, M., Fadul, M., Steiner, A., 2019. Automatic lameness detection in cattle. *Vet. J.*  
644 <https://doi.org/10.1016/j.tvjl.2019.01.005>
- 645 Aydin, A., 2017. Development of an early detection system for lameness of broilers using computer vision.  
646 *Comput. Electron. Agric.* 136, 140–146. <https://doi.org/10.1016/j.compag.2017.02.019>
- 647 Barkema, H.W., von Keyserlingk, M.A.G., Kastelic, J.P., Lam, T.J.G.M., Luby, C., Roy, J.-P., LeBlanc, S.J., Keefe,  
648 G.P., Kelton, D.F., 2015. Invited review: Changes in the dairy industry affecting dairy cattle health and  
649 welfare. *J. Dairy Sci.* 98, 7426–7445.
- 650 Berckmans, D., 2014. Precision livestock farming technologies for welfare management in intensive  
651 livestock systems. *Rev. sci. tech. Off. int. Epiz* 33, 189–196. <https://doi.org/10.20506/rst.33.1.2273>
- 652 Bewley, J.M., Robertson, L.M., Eckelkamp, E.A., 2017. A 100-Year Review: Lactating dairy cattle housing  
653 management. *J. Dairy Sci.* <https://doi.org/10.3168/jds.2017-13251>
- 654 Bezen, R., Edan, Y., Halachmi, I., 2020. Computer vision system for measuring individual cow feed intake  
655 using RGB-D camera and deep learning algorithms. *Comput. Electron. Agric.* 172, 105345.  
656 <https://doi.org/10.1016/j.compag.2020.105345>
- 657 Cowley, F.C., Barber, D.G., Houlihan, A. V, Poppi, D.P., 2015. Immediate and residual effects of heat stress  
658 and restricted intake on milk protein and casein composition and energy metabolism. *J. Dairy Sci.* 98,  
659 2356–68. <https://doi.org/10.3168/jds.2014-8442>
- 660 DELAVAL, 2020. Sensors for herd. website. URL <https://www.delaval.com> (accessed 5.12.20).
- 661 Fournel, S., Rousseau, A.N., Laberge, B., 2017. Rethinking environment control strategy of confined animal  
662 housing systems through precision livestock farming. *Biosyst. Eng.* 155, 96–123.  
663 <https://doi.org/10.1016/j.biosystemseng.2016.12.005>
- 664 Girshick, R., 2015. Fast R-CNN, in: *Proceedings of the IEEE International Conference on Computer Vision*. pp.  
665 1440–1448. <https://doi.org/10.1109/ICCV.2015.169>
- 666 Girshick, R., Donahue, J., Darrell, T., Malik, J., 2014. Rich feature hierarchies for accurate object detection  
667 and semantic segmentation, in: *Proceedings of the IEEE Computer Society Conference on Computer*  
668 *Vision and Pattern Recognition*. pp. 580–587. <https://doi.org/10.1109/CVPR.2014.81>
- 669 Guzhva, O., Ardö, H., Herlin, A., Nilsson, M., Åström, K., Bergsten, C., 2016. Feasibility study for the

670 implementation of an automatic system for the detection of social interactions in the waiting area of  
671 automatic milking stations by using a video surveillance system. *Comput. Electron. Agric.* 127, 506–  
672 509. <https://doi.org/10.1016/j.compag.2016.07.010>

673 Halachmi, I., Klopčič, M., Polak, P., Roberts, D.J., Bewley, J.M., 2013. Automatic assessment of dairy cattle  
674 body condition score using thermal imaging. *Comput. Electron. Agric.* 99, 35–40.  
675 <https://doi.org/10.1016/j.compag.2013.08.012>

676 Jaeger, M., Brügemann, K., Brandt, H., König, S., 2019. Associations between precision sensor data with  
677 productivity, health and welfare indicator traits in native black and white dual-purpose cattle under  
678 grazing conditions. *Appl. Anim. Behav. Sci.* <https://doi.org/10.1016/j.applanim.2019.01.008>

679 Jiang, B., Song, H., He, D., 2019. Lameness detection of dairy cows based on a double normal background  
680 statistical model. *Comput. Electron. Agric.* 158, 140–149.  
681 <https://doi.org/10.1016/j.compag.2019.01.025>

682 Kamilaris, A., Prenafeta-Boldú, F.X., 2018. Deep learning in agriculture: A survey. *Comput. Electron. Agric.*  
683 <https://doi.org/10.1016/j.compag.2018.02.016>

684 Li, W., Ji, Z., Wang, L., Sun, C., Yang, X., 2017. Automatic individual identification of Holstein dairy cows  
685 using tailhead images. *Comput. Electron. Agric.* 142, 622–631.  
686 <https://doi.org/10.1016/j.compag.2017.10.029>

687 Martinez-Ortiz, C.A., Everson, R.M., Mottram, T., 2013. Video tracking of dairy cows for assessing mobility  
688 scores, in: *Joint European Conference on Precision Livestock Farming*, 10 – 12 September 2013,  
689 Leuven, Belgium, p. 8.

690 Microsoft, 2018. VoTT: Visual Object Tagging Tool. GitHub Repos.

691 Nie, X., Yang, M., Liu, R.W., 2019. Deep Neural Network-Based Robust Ship Detection Under Different  
692 Weather Conditions, in: *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*. pp. 47–52.  
693 <https://doi.org/10.1109/ITSC.2019.8917475>

694 Norouzzadeh, M.S., Nguyen, A., Kosmala, M., Swanson, A., Palmer, M., Packer, C., Clune, J., 2017.  
695 Automatically identifying, counting, and describing wild animals in camera-trap images with deep  
696 learning, in: *Proceedings of the National Academy of Sciences of the United States of America*. pp. 1–  
697 17. <https://doi.org/10.1073/pnas.1719367115>

698 Okura, F., Ikuma, S., Makihara, Y., Muramatsu, D., Nakada, K., Yagi, Y., 2019. RGB-D video-based individual  
699 identification of dairy cows using gait and texture analyses. *Comput. Electron. Agric.* 165, 104944.  
700 <https://doi.org/10.1016/j.compag.2019.104944>

701 Porto, S.M.C., Arcidiacono, C., Anguzza, U., Cascone, G., 2015. The automatic detection of dairy cow feeding  
702 and standing behaviours in free-stall barns by a computer vision-based system. *Biosyst. Eng.* 133, 46–  
703 55. <https://doi.org/10.1016/j.biosystemseng.2015.02.012>

704 Porto, S.M.C., Arcidiacono, C., Anguzza, U., Cascone, G., 2013. A computer vision-based system for the

705 automatic detection of lying behaviour of dairy cows in free-stall barns. *Biosyst. Eng.* 115, 184–194.  
706 <https://doi.org/10.1016/j.biosystemseng.2013.03.002>

707 Redmon, J., 2013. 2016. Darknet: Open Source Neural Networks in C [WWW Document]. website. URL  
708 <https://pjreddie.com/darknet/> (accessed 5.10.20).

709 Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2016. You only look once: Unified, real-time object  
710 detection, in: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern*  
711 *Recognition*. IEEE, pp. 779–788. <https://doi.org/10.1109/CVPR.2016.91>

712 Redmon, J., Farhadi, A., 2018. YOLOv3: An Incremental Improvement. *arXiv Prepr.* 1804.02767.

713 Ren, S., He, K., Girshick, R., Sun, J., 2017. Faster R-CNN: Towards Real-Time Object Detection with Region  
714 Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* 39, 1137–1149.  
715 <https://doi.org/10.1109/TPAMI.2016.2577031>

716 Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese, S., 2019. Generalized intersection over  
717 union: A metric and a loss for bounding box regression. *Proc. IEEE Comput. Soc. Conf. Comput. Vis.*  
718 *Pattern Recognit.* 2019-June, 658–666. <https://doi.org/10.1109/CVPR.2019.00075>

719 Song, X., Leroy, T., Vranken, E., Maertens, W., Sonck, B., Berckmans, D., 2008. Automatic detection of  
720 lameness in dairy cattle-Vision-based trackway analysis in cow's locomotion. *Comput. Electron. Agric.*  
721 64, 39–44. <https://doi.org/10.1016/j.compag.2008.05.016>

722 Szeliski, R., 2011. *Computer Vision*, I. ed. Springer-Verlag London, London. [https://doi.org/10.1007/978-1-](https://doi.org/10.1007/978-1-84882-935-0)  
723 84882-935-0

724 Taigman, Y., Ming, Y., Ranzato, M., Wolf, L., 2014. DeepFace: Closing the Gap to Human-Level Performance  
725 in Face Verification, in: *Conference on Computer Vision and Pattern Recognition (CVPR)*.  
726 <https://doi.org/10.1109/CVPR.2014.220>

727 Trnovszky, T., Kamencay, P., Orjeseck, R., Benco, M., Sykora, P., 2017. Animal recognition system based on  
728 convolutional neural network. *Adv. Electr. Electron. Eng.* 15, 517–525.  
729 <https://doi.org/10.15598/aelee.v15i3.2202>

730 Tsai, D.M., Huang, C.Y., 2014. A motion and image analysis method for automatic detection of estrus and  
731 mating behavior in cattle. *Comput. Electron. Agric.* 104, 25–31.  
732 <https://doi.org/10.1016/j.compag.2014.03.003>

733 Tullo, E., Finzi, A., Guarino, M., 2019. Review: Environmental impact of livestock farming and Precision  
734 Livestock Farming as a mitigation strategy. *Sci. Total Environ.*  
735 <https://doi.org/10.1016/j.scitotenv.2018.10.018>

736 Van Hertem, T., Alchanatis, V., Antler, A., Maltz, E., Halachmi, I., Schlageter-Tello, A., Lokhorst, C., Viazzi, S.,  
737 Romanini, C.E.B., Pluk, A., Bahr, C., Berckmans, D., 2013. Comparison of segmentation algorithms for  
738 cow contour extraction from natural barn background in side view images. *Comput. Electron. Agric.*  
739 91, 65–74. <https://doi.org/10.1016/j.compag.2012.12.003>

740 Van Hertem, T., Schlageter Tello, A., Viazzi, S., Steensels, M., Bahr, C., Romanini, C.E.B., Lokhorst, K., Maltz,  
 741 E., Halachmi, I., Berckmans, D., 2018. Implementation of an automatic 3D vision monitor for dairy cow  
 742 locomotion in a commercial farm. Biosyst. Eng. <https://doi.org/10.1016/j.biosystemseng.2017.08.011>  
 743 Van Hertem, T., Steensels, M., Viazzi, S., Romanini, E.C.B., Bahr, C., Berckmans, D., Schlageter Tello, A.,  
 744 Lokhorst, K., Maltz, E., Halachmi, I., 2014. Improving a computer vision lameness detection system by  
 745 adding behaviour and performance measures, in: International Conference of Agricultural  
 746 Engineering. pp. 1–8.  
 747 Ventura, B.A., Von Keyserlingk, M.A.G., Wittman, H., Weary, D.M., 2016. What difference does a visit  
 748 make? Changes in animal welfare perceptions after interested citizens tour a dairy farm. PLoS One.  
 749 <https://doi.org/10.1371/journal.pone.0154733>  
 750

# CRediT author statement

**Patrizia Tassinari:** Conceptualization, Funding acquisition, Project administration, Resources, Supervision.

**Marco Bovo:** Conceptualization, Data curation, Formal analysis, Methodology, Validation, Visualization, Writing—original draft, Writing—review & editing.

**Stefano Benni:** Conceptualization, Formal analysis, Investigation, Methodology, Supervision, Validation, Visualization, Writing—original draft.

**Simone Franzoni:** Software

**Matteo Poggi:** Software

**Ludovica Maria Eugenia Mammi:** Investigation

**Stefano Mattoccia:** Conceptualization, Software, Validation

**Luigi Di Stefano:** Conceptualization, Software, Validation

**Filippo Bonora:** Data curation

**Alberto Barbaresi:** Visualization, Writing—review & editing

**Enrica Santolini:** Visualization, Writing—review & editing

**Daniele Torreggiani:** Conceptualization, Project administration, Resources, Supervision, Writing—review & editing.