Detection of Cow Face and Subregions Using a New Annotated Dataset

Mathews Edwirds Gomes Almeida[1], Pedro de Araújo Bhering Bittencourt[1], Brian Luís Coimbra Maia[1], Lucas Silva Santana[1], João Vítor de Castro Martins Ferreira Nogueira[1], Daniel Muller Rezende[1], Gabriel Rezende da Silva[1], Thais de Souza Marins[1], Luiz Maurílio da Silva Maciel[1(✉)], Marcelo Bernardes Vieira[1], Saulo Moraes Villela[1], and Bruno Campos de Carvalho[2]

[1] Federal University of Juiz de Fora, Juiz de Fora, MG, Brazil
{mathews.edwirds,pedro.bittencourt,brian.maia,lucas.santana,joao.nogueira,daniel.rezende,rezende.gabriel,thais.marins}@estudante.ufjf.br,
{luiz.maciel,marcelo.bernardes,saulo.moraes}@ufjf.br
[2] Embrapa Dairy Cattle,Juiz de Fora, MG, Brazil
bruno.carvalho@embrapa.br

Abstract. This study proposes a novel methodology for detecting cattle faces and their key facial subregions – ears, eyes, and muzzle – through a two-stage object detection pipeline based on deep convolutional neural networks. To address the lack of annotated datasets in this domain, we introduce the Annotated Cattle Images (ACI) dataset, composed of 7,146 images and 54,111 labeled instances, collected from both public repositories and real-world farming environments. This dataset was essential for training our detection models, and we made it publicly available for future research. Our detection approach employs a two-stage detection strategy: initially locating the bovine face in the entire image and subsequently detecting subregions within the localized area, effectively reducing background noise and improving accuracy for small features. Experiments using YOLOv8-X, YOLOv9-E, and YOLOv10-X show that the YOLOv9-E model achieved the best performance. Comparative analyses confirm the advantage of the two-stage approach, which improved subregion detection over the single-stage baseline. We also implemented a dedicated post-processing module to remove duplicate predictions and further enhance the reliability of the models. These results show the viability of the proposed solution for real-time, non-invasive livestock monitoring, advancing the applicability of computer vision in precision livestock farming.

Keywords: Object Detection · Dataset · Computer Vision · Precision Livestock Farming

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026, LNCS 16767, pp. 561–579, 2027.
https://doi.org/10.1007/978-3-032-30491-9_35

1 Introduction

Livestock farming constitutes a pivotal component of the global agricultural economy, serving as a primary source of food and animal-derived raw materials. Globally, livestock farming is a key element of economic development, with the cattle herd reaching approximately 1.5 billion in recent years [1]. The expansion of cattle herds reflects the growing global appetite for meat and animal products, despite ongoing challenges related to environmental sustainability. This emphasizes the critical need to enhance the efficiency and sustainability of livestock production, especially in the face of growing environmental challenges and increasing social demands for ethical animal husbandry practices.

In response to these ecological and livestock welfare concerns, the concept of precision livestock farming has gained traction as a transformative approach. This approach integrates advanced Information and Communication Technologies (ICTs) to optimize livestock production through the continuous and detailed monitoring of animals health, welfare, and environmental conditions. The deployment of sensors, cameras, and mobile technologies facilitates real-time data acquisition and analysis, thereby enabling data-driven decision-making to enhance productivity and sustainability [2].

In the context of precision livestock farming, computer vision allows to recognize and analyze patterns in images or videos. This technology offers a wide range of applications, from animal detection to monitoring livestock health and welfare. Specifically, object detection has shown substantial potential in identifying and classifying specific animal features, such as the location of the face, eyes, ears, and muzzle. These detections are particularly valuable for developing automated monitoring systems that operate in real time, reducing the need for human intervention and increasing the accuracy of animal health assessments. For instance, analyzing the position and movement of bovine ears, eyes, and muzzles can provide valuable insights into their emotional states and levels of discomfort, facilitating early intervention and enhanced welfare management [3].

Despite numerous fields have being benefited from extensive research and abundant datasets, the detection of facial subregions in cattle presents significant challenges due to the limited availability of high-quality annotated datasets in the literature. The scarcity of representative datasets that encompass diverse environments, variability in animal characteristics, and varying image capture conditions makes the development of effective precision livestock systems in uncontrolled settings difficult [4]. This disparity underscores the need for specialized resources tailored to livestock. To address these challenges and improve model performance in varied conditions, one of the main contributions of this study is the creation of a dataset with annotations dedicated to detecting the face and its facial subregions of interest in cattle images.

The present study aims to develop object detection models tailored to identify key facial subregions in cattle within non-standardized, real-world environments. Thus, our proposal explores a two-stage detection approach: we first detect the cattle faces and then detect the subregions, which effectively eliminates background noise and improves model results. We employed our proposed dataset

to train these models to deliver precise and reliable detection performance even under varying conditions of illumination, angle, and background complexity. The integration of such models into real-time monitoring systems has the potential to significantly enhance the assessment of animal health and welfare in precision livestock farming, thereby contributing to the broader objectives of improving productivity, sustainability, and animal well-being in the sector.

2 Related Work

In recent years, some studies have addressed the problem of facial subregions detection in animals for different applications. Billah et al. [5] conducted a facial subregions detection process in goats using convolutional neural networks (CNNs) to improve and facilitate the animal management. Furthermore, the work of Jaddoa et al. [6] proposed a method to monitor the body temperature of cattle to identify certain risks to the animal. They analyze the animal's body temperature by identifying the eye region in infrared thermal images using the Infrared Thermography Technique (IRT). However, we note a scarcity of works addressing facial subregions detection. Particularly, to the best of our knowledge, we did not find works that propose methods to detect the facial subregions.

Some works propose to detect the entire animal face, without separating the subregions. A primary objective in many studies involving face detection is the recognition of individuals [7–9]. Kawagoe et al. [10], in addition to individual recognition, conducted a feeding time estimation. Another common motivation is facial expression recognition [11]. Mao and Liu [11] focused on pet dogs, in contrast to other related works that studied livestock animals. The YOLO series algorithms and their variations are the most common methods for solving face detection problems.

The detection of other parts of the animal's body for individual identification has also been studied in several works. Tassinari et al. [12] conducted a real-time identification of cows based on coat pattern, to monitor the position, actions, and movements of each cow. Smink et al. [13] focused on detecting, tracking, and reading ear tags in videos to identify animals with diseases on large-scale cattle farms. Some studies conducted the detection of the animal's entire body. For instance, Gupta et al. [14] detected and classified cows into eight different breeds.

Despite the related works presented, in this study, we introduce a methodology for detecting faces and subregions (ears, eyes, and muzzle) in cattle. We propose an annotated dataset specifically for this task, which was used to train and evaluate the performance of detection models.

3 Deep Convolutional Neural Networks for Object Detection

Modern object detection methods are predominantly based on CNNs. These approaches follow either a two-stage strategy – where region proposals are generated before classification – or a one-stage strategy, which performs detection

and classification in a single step. One-stage detectors, despite a slight reduction in accuracy, offer significantly better computational performance and are thus more suitable for real-time applications [15]. In this work, we adopted one-stage models for detecting cattle facial regions, with a focus on different versions of the You Only Look Once (YOLO) architecture.

YOLO was introduced by Redmon [16] as a unified, real-time object detection framework. Since then, successive versions have improved performance, robustness, and usability. YOLO's core principle is to treat detection as a regression problem: it divides the input image into an $S \times S$ grid, where each cell predicts bounding boxes, class probabilities, and objectness scores. The original architecture comprised 24 convolutional layers and 2 fully connected layers, using 1×1 convolutions to reduce feature dimensions. Outputs are refined using Non-Maximum Suppression (NMS) to eliminate redundant predictions.

Recent YOLO versions have adopted significant structural and algorithmic innovations to enhance detection accuracy and computational efficiency. YOLOv8, released by Ultralytics in 2023, adopts an anchor-free approach for bounding box prediction and integrates techniques like mosaic augmentation and an improved neck for better feature fusion. This version introduced a suite of pre-trained models with varying sizes to balance accuracy and speed. YOLOv9, proposed in 2024 by an independent team, incorporates Programmable Gradient Information (PGI) and Generalized Efficient Layer Aggregation Network (GELAN), addressing the information bottleneck and improving gradient flow. These features enhance generalization, particularly in low-data regimes and multitask scenarios. YOLOv10, released in mid-2024, eliminates the need for NMS via consistent double assignments and leverages large-kernel convolutions and partial self-attention to improve accuracy with minimal computational cost [17].

In addition to the different versions of YOLO, there are variations in the sizes of each model based on the parameter count and complexity. In our experiments, we employed the extra-large version (X) of YOLOv8, which has 68.2 million parameters. We also use the extended version (E) of YOLOv9, with 58.1 million parameters and the extra-large version (X) of YOLOv10, with 29.5 million parameters. These are the biggest versions of each network available. We intend to evaluate the performance of our detection approach using different network versions.

4 ACI Dataset

Due to the lack of available datasets including bovine facial subregions, we created a new dataset to meet the specific needs of this study, named "Annotated Cattle Images" (ACI) dataset. Its construction involved several critical steps: data collection, filtering, and labeling. Data collection employed specific techniques to ensure the dataset's diversity and representativeness. Filtering was applied to remove irrelevant and low-quality data. Labeling was essential for identifying and categorizing key features, such as facial subregions, to train and evaluate models effectively. These steps were crucial for building a reliable dataset for bovine facial detection models.

To support further research and development in the area, we made the training dataset publicly available at Almeida et al. [18]. This provision allows multiple studies and institutions to utilize the data for further research and model development. In the following sections, we describe the steps performed to create the proposed dataset.

4.1 Data Collection

To develop a dataset for advancing bovine detection and classification methodologies, we implemented a data collection strategy that combined curated images from the internet with captured images. Initially, we employed a tool to assist in downloading a diverse batch of bovine images from selected open-access platforms, including Pexels,[1] Pixabay,[2] Creative Commons' Openverse,[3] and Unsplash.[4] These platforms align with our non-commercial research objectives and guarantee adherence to ethical guidelines concerning copyrighted materials. We meticulously filtered the images to include only those explicitly designated as public domain or available under licenses that do not require specific attribution to the original authors.

To capture a broad range of bovine breeds, poses, and environmental contexts, we utilized a variety of search terms in Portuguese (e.g., *vaca*, *bezerro*, *novilho*, *boi*, *touro*, *bovino*, *gado*), Spanish (e.g., *vaca*, *becerro*, *buey*, *toro*, *bovino*, *vacuno*, *ganado*), and English (e.g., *cow*, *calf*, *bull*, *taurus*, *bovine*, *cattle*, *livestock*). This multilingual approach enriched the diversity of our dataset by incorporating images from various regional sources.

Furthermore, manual curation was an integral part of our data collection process, as we did not use any scripts or web scraping libraries to filter the images before downloading them from open-access websites. Given that many downloaded images lacked complete metadata, manual verification was essential to assess their suitability for inclusion in our dataset.

After compiling the initial set of images, we performed a post-filtering process to improve the general consistency of the dataset. Firstly, we eliminated images in which the bovine's face occupied less than 10% of the image area, ensuring that the faces were sufficiently prominent for facial analysis. In addition, we removed images that were excessively blurry, distorted, or had severe lighting problems, as these challenges are not in the scope of our work. Images that did not contain bovines or showed significant occlusions were also excluded to maintain the relevance and reliability of the dataset.

Finally, we complemented our dataset through direct image capture conducted in collaboration with Embrapa Dairy Cattle. This approach involved photographing bovines in their natural farm environments, providing practical and realistic examples of appearances under real-world conditions. By incorporating images of diverse breeds, poses, and environmental conditions through

[1] https://www.pexels.com.
[2] https://www.pixabay.com.
[3] https://openverse.org.
[4] https://www.unsplash.com.

direct capture, we significantly improved the dataset's applicability for detection systems and practical deployment scenarios.

4.2 Annotation

An annotation process was implemented to label bovine facial subregions meticulously. This methodical approach aimed to achieve both precision and consistency in labeling. The annotation adhered to the following principles:

1. **Bounding Box Accuracy**: Bounding boxes were meticulously adjusted to fit tightly around each object of interest, ensuring central placement and minimizing background inclusion to enhance the focus on relevant facial subregions and features.
2. **Labeling Key Regions**: Each subregion of interest was tagged with specific labels such as "cow" for the bovine face, "left ear," "right ear," "left eye," "right eye," and "muzzle" to accurately categorize facial components.
3. **Perspective Consideration**: Labels were assigned carefully to the animal's perspective, ensuring that "left" and "right" features were represented correctly based on the animal's viewpoint.
4. **Partial Occlusions**: All visible regions of interest were annotated, including those partially occluded.
5. **Special Cases**: In instances where we observed specific features, such as closed eyes or an open jaw, annotations were made approximately around the region.
6. **Image Quality Considerations**: Instances where objects were difficult to recognize due to blurriness or small dimensions were ignored.
7. **Avoiding Mislabeling**: We avoided overly tight bounding boxes around features like eyes, which could lead to confusion with background patterns. We adjusted the boxes to ensure a clear distinction from the surrounding area.

The annotation process was performed using LabelImg (version 1.8.6),[5] a tool that offers an intuitive interface and supports various output formats, such as PASCAL VOC and YOLO. The tool's features, including simultaneous image and label visualization, facilitated error detection and correction.

4.3 Dataset Overview and Statistics

The ACI dataset consists of bovine facial images collected from multiple sources, ensuring a diverse representation of breeds, poses, and environmental conditions. However, the images exhibit a wide range of resolutions, with many occupying excessive storage space. This variability not only increased storage demands but also slowed down the preprocessing stage before model training.

[5] https://github.com/heartexlabs/labelImg.

To address this, we applied a dataset adaptation strategy aimed at getting image sizes more compatible with deep learning models while preserving annotation integrity and image diversity. As part of this strategy, we established a target resolution A_{target} of 15 megapixels and resized the images that exceeds this threshold. Specifically, given an image with area $A = H \times W$ (in pixels), if $A > A_{\text{target}}$, the scaling factor $s = \sqrt{A_{\text{target}}/A}$ was applied uniformly to both dimensions. This ensures that the resized image approximates the target resolution while preserving the original aspect ratio. This dynamic resizing approach effectively reduced storage demands while maintaining visual and annotation consistency. Importantly, this resizing procedure was applied exclusively to the training set, leaving the test set untouched.

Although very high-resolution images do not negatively impact training, they also do not provide significant benefits for standard object detection tasks. All images were resized to align with the typical input dimensions expected by YOLO and other modern detection architectures, which commonly operate on relatively low-resolution inputs (e.g., 640×640). This resizing process was not primarily intended to reduce storage usage – although that became a positive side effect – but rather to ensure architectural compatibility and consistency throughout the dataset. Importantly, this adjustment did not alter the dataset's distribution or introduce any biases that could affect model performance. In fact, most detection models in literature use input sizes smaller than those resulting from our resizing procedure, and only a small fraction of the original images had very high resolution. By implementing this adjustment, we obtained lower storage space and computational feasibility.

The dataset final structure is summarized in Table 1, which presents key attributes of the training and test sets, including the number of images, resolution details, and overall dataset size. Specifically, images sourced from the internet were used for the training set, while images collected by our team were designated for the test set. Notably, the training images were sourced from the internet without prior labeling, making it challenging to determine the exact number of individuals and breeds within this subset. Additionally, the average resolution of images in the training set is lower than in the test set. The image distribution between the sets is approximately 86.51% for training and 13.49% for testing.

The analysis of the images from both the training and test sets reveals significant variability in many aspects, including proximity, angle, background, lighting, and sharpness. This variability is crucial for understanding the dataset's scope and challenges. For instance, the presence of occlusions, multiple faces, diverse categories (such as calf, cow, or other bovine classifications), poses, and breeds adds complexity to detecting cow facial subregions. Moreover, some images feature multiple facial subregions, even when the bovine face is not fully visible.

The diversity of environments in which the images were captured further highlights the dataset's variability. The training set includes bovines captured in various scenarios, such as urban areas, which contrasts with the test set, which is

Table 1. Detailed information on the training and test sets.

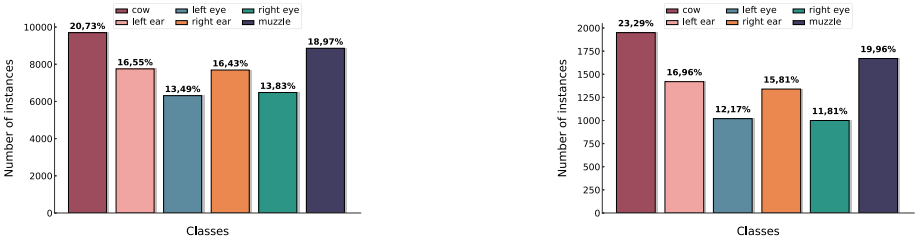
Attribute	Training set	Test set
Number of Images	6182	964
Minimum Resolution	136 × 235	2974 × 2918
Maximum Resolution	7360 × 4912	4624 × 3468
Average Resolution	1982 × 1694	4317 × 3243
Minimum Size (MB)	0.009	1.82
Maximum Size (MB)	9.77	25.72
Average Size (MB)	1.24	13.33
Total Size (GB)	7.45	12.5
Number of Individuals	-	165
Breeds	-	Girolando and Holstein
Colored	Yes	Yes
Format	JPG	JPG

composed exclusively of images from rural production environments, including pens, calving areas, pastures, and confinement systems. This distinction is visually represented in Fig. 1, which showcases representative examples from each dataset.



Fig. 1. Examples of images from the proposed dataset. Fist row: training set. Second row: test set.

Figure 2 presents the class distribution of the annotated objects in both the training and test sets. Larger facial regions, such as the full cow head and the muzzle, dominate the distribution. In contrast, smaller object classes, particularly ears and eyes, are consistently underrepresented—a disparity likely caused by the inherent difficulty in annotating small-scale features and by occlusions. This imbalance not only affects annotation efforts but may also hinder the detection performance for these classes.



(a) Training set. (b) Test set.

Fig. 2. Distribution of labels by class and by set.

5 Proposed Detection Method

We used the proposed dataset to train a system that consists of two detection models for distinct purposes, as illustrated in Fig. 3. The first model is designed to detect the animal's face, isolating this area from the image background. Next, the second model only detects the facial subregions within the facial area. This two-stage approach aims to obtain precise results for larger, more easily identifiable objects, such as faces, and for identifying smaller objects, such as cattle ears, eyes, and muzzles.

To implement the proposed approach, it was essential to prepare the dataset described in Sect. 4.3 using the ground truth labels. At this stage, specific classes in the original dataset were selected, creating two complementary subdatasets for training two independent models. In the first subdataset, named "Annotated Cattle Face Images" (ACFI) subdataset, we kept the entire images but only the annotations related to the animal face. In the second subdataset, designated "Annotated Cattle Facial Subregions Images" (ACFSI) subdataset, faces were ignored and only the annotations of ears, eyes, and muzzle were kept. The original image was cropped to contain only these subregions. Initially, ground truth annotations guided the cropping process, after which a careful manual review was conducted to verify that each crop contained the correct number of features, avoiding any inconsistencies (e.g., no crop showing more than two ears or missing critical features). This thorough filtering process was essential to exclude non-target regions or background elements, focusing solely on the relevant subregions for facial recognition. Additionally, each cropped face was adjusted to maintain a 1:1 aspect ratio, achieved through resizing or padding as necessary. This standardization ensured uniform input dimensions, avoiding image distortion for the model training and evaluation.

Each one of the subdatasets was subdivided into training and test sets based on the original dataset structure to facilitate the performance evaluation. This organization of the original dataset in two new subdatasets is essential to enable specialized training, which addresses the detection of complete faces and the identification of specific features within those faces.

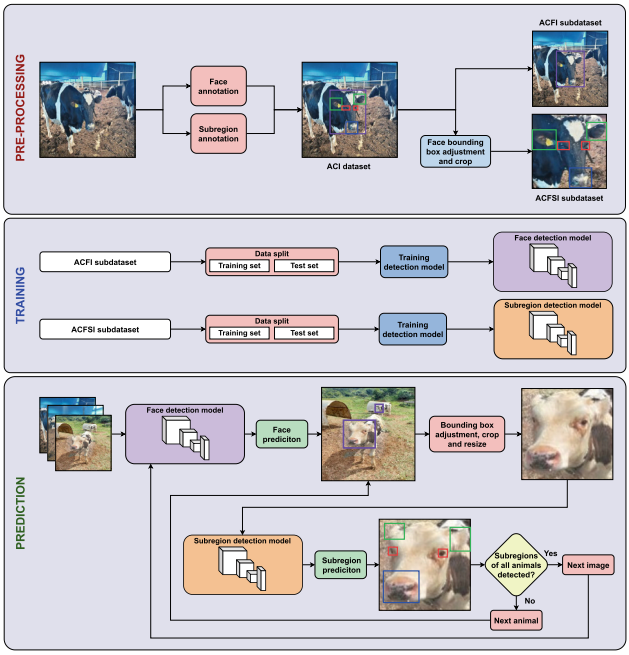


Fig. 3. Proposed detection method overview.

In the next step, model training, two different models were produced. The first one, responsible for face detection, was trained using the ACFI subdataset, enabling it to learn how to automatically identify the presence and location of bovine faces in any cattle images. In parallel, a second model, focused on subregion detection, was trained using the ACFSI subdataset. This second model was trained to detect specific subregions within the face area but with a reduced influence of background noise.

With both models trained, we can apply them to perform the prediction of new images. The prediction system receives an input image and uses the face detection model to locate all faces on the complete image. Given a detected face, its bounding box is adjusted to keep a 1:1 aspect ratio. Then, the face is cropped and resized to the input size of the subregion detection model. This model perform the detection of the subregions on the face image. If the face detection model identifies multiple faces on an image, the system extracts the next face to be processed by the subregion detection model. Otherwise, the face detection model process the following image.

6 Experiments and Results

6.1 Experimental Setup

In this subsection, we present the results of the experiments conducted to evaluate the proposed methods for the problem of detecting the faces of cattle and

their subregions. We conducted four different types of training, where two of them are integral to our proposed method: one dedicated to faces detection in the entire image and another focused on detecting subregions within the cropped face image, i.e. with the background removed, as detailed in the previous section. Additionally, two other approaches were trained for comparison and validation purposes: one to detect both the face and subregions in the entire image, and another to detect only the subregions in the entire image.

The training step were performed in an Intel(R) Core(TM) i7 CPU 870 @ 2.93 GHz processor, 8 GB of single-channel RAM, a Quadro M5000 GPU, and running Ubuntu 22.04 LTS as the operating system. Furthermore, we utilized the implementations provided by Ultralytics for the YOLO models, ensuring that the dependencies required for their usage are consistent with the official releases.

We trained all networks with an input resolution of 640 $\times$ 640 pixels, a batch size of 2, for 40 epochs and with the SGD optimizer in all cases. These settings were determined by the limitations of the available hardware resources. Additionally, we kept the default data augmentation for all networks, except for the horizontal flip. This operation was disabled (set to 0.0) because it inverts the orientation of the face. For the YOLOv8-X we set, in all cases, the initial learning rate at 10^{-3} and linearly decay it to 10^{-4}.

Furthermore, for the all classes, detection problem and for the face detection problem, we set the initial learning rate to 10^{-3} and decay it to 10^{-4} for YOLOv9-E network. For the subregions detection problem with the entire image we set the initial learning rate to 10^{-3} and decay it to 10^{-5} and for the subregions detection problem with cropped image we set the initial learning rate to 10^{-4} and decay it to 10^{-5}.

For YOLOv10-X we set initial learning rate to 10^{-3} and decay it to 10^{-5} in all problems. Furthermore, we enabled the cosine learning rate scheduler, ensuring a smooth and gradual reduction of the learning rate throughout the training process. Additionally, we employed a dropout regularization of 0.5 to help prevent overfitting by randomly deactivating half of the units in each layer. The minimum confidence threshold was set to 0.05, which is less than the default value of 0.25. This adjustment is particularly effective in YOLOv10 since, unlike previous versions such as YOLOv8, it does not use non-maximum suppression (NMS). As a result, lowering the confidence threshold does not negatively affect inference performance, enabling the detection of smaller or more distant objects, contributing, in our case, to the detection of eyes for example. The half-precision floating-point option was enabled to improve computational efficiency.

6.2 Training Results

The results are computed on the test set for each architecture, based on the model of the last epoch of each training. We evaluated the results for precision, recall, mAP50, and mAP50-95. The precision represents the fraction of correctly identified positive instances, called True Positives (TP), out of the total number of predicted positive instances, that is, the sum of TP and False Positives (FP). In contrast, recall represents the proportion of positive instances correctly identified

relative to the total number of actual positive instances, including both TP and False Negatives (FN). The mean Average Precision (mAP) is calculated as follows:

$$\text{mAP} = \frac{1}{N} \sum_{i=1}^{N} \text{AP}_i, \tag{1}$$

that is, for each class i out of a total of N classes, the area under the Precision-Recall curve is computed. Therefore, mAP50 corresponds to the average of the AP values using an IoU threshold of 0.5, while mAP50-95 is computed using multiple IoU thresholds ranging from 0.5 to 0.95 in increments of 0.05.

We divided the results into four tables, one for each problem. Table 2 refers to the problem of detecting all classes in the entire image, Table 3 presents the results for the problem of detecting only the face in the entire image, Table 4 refers to the problem of detecting subregions in the entire image, and finally, Table 5 shows the values achieved for the problem of detecting subregions in cropped images. The best results in each table are highlighted in bold.

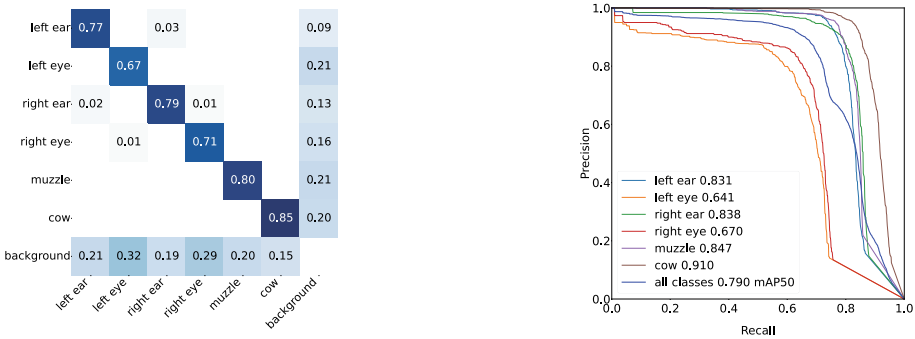
For each of these scenarios, right after the respective table, we present a confusion matrix and a precision-recall (PR) curve corresponding to the model that achieved the best results (except for the face-detection case, which is single-class. In that case, we report only the numerical metrics (Table 3) without confusion matrix or PR curve). These are shown in the following figures: Fig. 4 for the all-classes detection problem, Fig. 5 for subregions detection in entire images, and Fig. 6 for subregions detection in cropped images.

Table 2. Results on the test set for each CNN for the problem of all classes detection in the entire image.

Model	Precision	Recall	mAP50	mAP50-95
YOLOv8-X	0.862	0.689	0.743	0.434
YOLOv9-E	**0.882**	**0.734**	**0.790**	**0.466**
YOLOv10-X	0.864	0.693	0.751	0.429

Table 2 shows that the YOLOv9-E model obtained the best results for the problem of all classes detection in the entire image. In Fig. 4a, we observe that the class cow achieved the best detection accuracy. This result indicates that the face is more easily detected in the entire image compared to the subregions. On the other hand, the classes of eyes present lower accuracy values, evidencing the difficulty of detecting these subregions in the entire image. The behavior of the PR curves in Fig. 4b reinforce these conclusions; the class cow presents high precision values even for high recall values, while the eyes classes curves show a decrease of precision when the recall increases.

In Table 3, we note that YOLOv9-E achieved the best results, except for precision. Compared to Table 2, we observe greater values for all criteria. This fact suggests that face detection is a relatively more straightforward problem.



(a) Confusion matrix. (b) Precision-Recall curve.

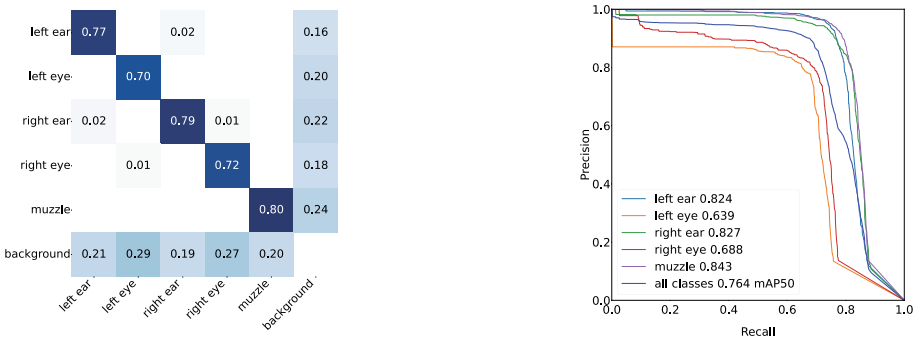
Fig. 4. Detection of all classes in entire images.

Table 3. Results on the test set for each CNN for the problem of face detection.

Model	Precision	Recall	mAP50	mAP50-95
YOLOv8-X	**0.928**	0.795	0.885	0.748
YOLOv9-E	0.919	**0.820**	**0.894**	**0.755**
YOLOv10-X	0.926	0.771	0.869	0.722

Table 4. Results on the test set for each CNN for the problem of subregions detection with entire images.

Model	Precision	Recall	mAP50	mAP50-95
YOLOv8-X	0.855	0.617	0.736	0.372
YOLOv9-E	**0.867**	**0.728**	**0.764**	**0.388**
YOLOv10-X	0.838	0.688	0.732	0.372



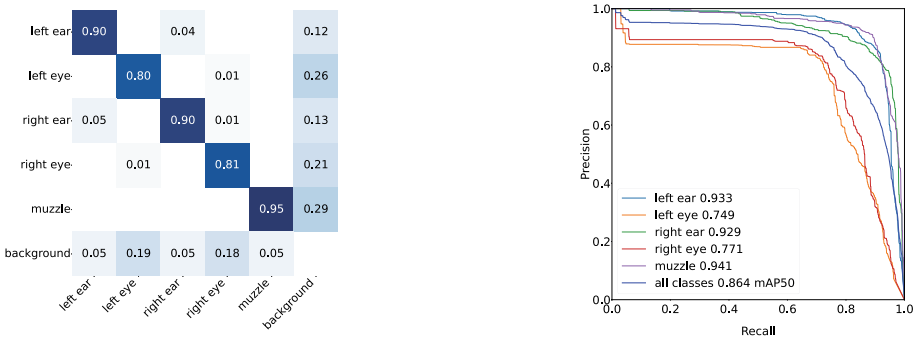
(a) Confusion matrix. (b) Precision-Recall curve.

Fig. 5. Detection of the subregions in entire images.

By the analysis of Table 4, we observe that YOLOv9-E model also presents the best results. We note a decrease in the values compared to Table 2, which highlights the influence of the class cow. The confusion matrix in Fig. 5a also reveals worse results for the eyes classes since these are the smallest subregions. Figure 5b presents the same behavior of Fig. 4b, with the eye classes presenting a sustantial decrease of precision when the recall increases.

Table 5. Results on the test set for each CNN for the problem of subregions detection with cropped images.

Model	Precision	Recall	mAP50	mAP50-95
YOLOv8-X	0.833	0.825	0.842	0.449
YOLOv9-E	**0.847**	**0.826**	**0.864**	**0.472**
YOLOv10-X	0.843	0.816	0.851	0.440



(a) Confusion matrix.

(b) Precision-Recall curve.

Fig. 6. Detection of the subregions in cropped images.

Similar to the other problems, Table 5 shows that YOLOv9-E provided the best results for the problem of detecting subregions in images cropped only with the animal face. We note that the values are better than those in Table 4, which confirms that detecting subregions after cropping the animal face is an advantageous strategy. In Fig. 6a, we note an increase in the accuracy of all classes, although the eyes remain the more difficult classes. The PR curve (Fig. 6b) presents a better behavior, with the recall values presenting a decay only for high precision values.

Thus, through the analysis of the four result tables, it is evident that YOLOv9-E achieved the best mAP50-95 results, which is the most challenging criterion. For comparison, Table 6 presents the results obtained by YOLOv9-E

model across all the addressed problems. We notice that the inclusion of face detection in the problem of all classes detection had a significant impact on the results. The results dropped considerably when this step is removed in the problem of detecting subregions in the entire image. We argue that our two-stage method tends to provide superior results by separating the face detection and the subregions detection.

Table 6. Results on the test set of YOLOv9-E model for all problems described.

Problem	Precision	Recall	mAP50	mAP50-95
All classes	0.882	0.734	0.790	0.466
Face	0.919	0.820	0.894	0.755
Subregions (entire images)	0.867	0.728	0.764	0.388
Subregions (cropped images)	0.847	0.826	0.864	0.472

6.3 Evaluation and Analysis of the Facial and Subregion Detection Method

We developed an application that performs detections according to the method described in Sect. 5. It is applied on the test set. Initially, the algorithm identifies faces using the model trained with the ACFI subdataset. Then, detected faces are first expanded and centered by adding uniform margins so that each region becomes a square (1:1 aspect ratio). Next, these square regions are cropped, resized to 640 $\times$ 640 pixels, and saved. Thus, the face images are delivered to the model trained on the ACFSI subdataset, which detects the subregions. To evaluate the method, we compared the generated annotations with those available in the test set. The comparison is carried out with the Intersection over Union (IoU), where each detected subregion is compared with its corresponding annotated subregion for a given image.

The application operates in two detection modes:

1. **Most Prominent Face:** This mode detects only the largest (most prominent) face in the image along with its subregions. In this mode, 964 faces were detected in 964 images, meaning that one face was successfully detected in every image.
2. **All Faces:** In this mode, the application detects all faces in the image and the subregions within each face. A total of 1,730 faces were detected from 964 images and the dataset annotations include 1,934 faces. Consequently, some faces (typically those less prominent and more challenging for the network to detect) were missed.

We also observed cases in which the same subregion appeared multiple times within a single face. This phenomenon can occur when animals are close to

each other and the cropped face inadvertently includes a subregion belonging to another animal, resulting in duplicate detections of the same subregion. In the **Most Prominent Face** mode, we found 76 duplicate subregion occurrences in 70 images, while in the **All Faces** mode, 116 duplicate subregions were found in 110 images.

We developed two approaches to eliminate these duplicated detections:

1. The more straightforward approach named **Max Confidence** retains only the instance with the highest confidence among the repeated detections.
2. The second approach named **Advanced Filter** verifies for each subregion the intersection with the detected face with a tolerance of 0.8. This allows to identify and eliminate cases where the subregions actually belong to other animals adjacent to the target one. If elimination does not occur through this method, we check whether these detections overlap another subregion with an IoU threshold of 0.8, for every part with multiple detections. This additional step helps to remove further erroneous detections. Finally, if more than one instance remains for the same class, we retain only the detection with the highest confidence.

Table 7 presents the accuracies for each class using the most prominent face and the all faces detection modes. Two result sets are presented: one uses the max confidence strategy to remove duplicates and the other uses the advanced duplication filter. We calculated the Intersection over Union (IoU) between the detected regions and the respective ground truth. We considered a detection as corrected if the IoU was greater than the threshold of 0.5. The proposed pipeline operates with an inference time of approximately 500 ms per stage, amounting to about 1 s per image.

Table 7. Accuracy (%) by subregion and approach.

Subregion	Max Confidence		Advanced Filter	
	Most Prominent	All Faces	Most Prominent	All Faces
Muzzle	94.11	83.60	94.32	83.99
Left Ear	89.50	79.00	89.74	79.22
Right Ear	88.49	83.26	88.49	83.26
Left Eye	73.76	70.86	73.76	70.86
Right Eye	77.39	73.22	77.39	73.22
Overall	85.67	78.93	85.77	79.02

The results show that the Advanced Filter proved advantageous both for the most prominent individual and for the detection of all faces. Additionally, the detection results for the most prominent individual performed better. One explanation is that, in the case of detecting all faces, there are several instances where the network had difficulty in identifying the subregions, as these faces are

relatively smaller and less prominent in the images. Besides, we observe that the "Muzzle" class has higher accuracy, while "Left Eye" and "Right Eye" classes show poorer performances. This may be due to the difficulty in detecting these subregions, as they are less salient compared to the others.

7 Conclusion

This work aimed to apply neural networks on images to detect the faces of cattle as well as their regions of interest, such as eyes, ears, and muzzle, under various conditions of lighting, angle, position, and others. To this end, we proposed a dataset comprising images of cattle faces and their corresponding subregions of interest. We have made this dataset publicly available, which is a contribution of our work. This dataset may be used for future research that requires cattle images with annotations of subregions. Based on this dataset, we trained and evaluated neural networks for object detection.

To avoid the influence of the background in the detection, we developed a two-stage approach. First, we detect the animal faces in the entire image. Then, the face is cropped and the subregions are detected. Through a quantitative analysis of the models, we observed that YOLOv9-E achieved the best performance according to the mAP50-95 evaluation criterion. The network showed satisfactory results for the classes cow, muzzle, and ears (left and right), likely due to their greater salience and larger number of training samples. However, for the eye classes, the overall performance was lower, possibly due to underrepresentation in the dataset and smaller object size compared to other classes.

To validate our strategy, we developed an application that performs facial and subregion detection using the proposed method. The evaluation was conducted on a test set, using IoU. Two detection modes were explored: detecting only the most prominent face or detecting all faces in the image. Although the "All Faces" mode resulted in a larger number of detections, it also introduced more erroneous responses, particularly due to missed detections of smaller or less prominent faces. Furthermore, this mode increased instances of duplicated subregions per face.

To address the face duplication issue, we implemented two filtering approaches: Max Confidence and Advanced Filter. The latter, which incorporates geometric constraints and IoU-based filtering, achieved better overall accuracy by effectively removing erroneously duplicated subregions that were detected. Among the subregions, the "muzzle" class consistently presented the highest detection accuracy, while the "eyes" classes performed less accurately, likely due to their size and lower prominence in the images. The results reinforce the effectiveness of the two-stage detection strategy.

In summary, the results emphasize the strong potential of deep learning-based object detection networks in identifying cattle faces and their key subregions under real-world conditions. As future works we may improve the detection of the eyes by using models for Tiny Object Detection (TOD). Another possibility is to combine different detectors in an ensemble to improve the detection performance.

Acknowledgments. We thank the Minas Gerais State Research Foundation - FAPEMIG (APQ-02230-23) for the financial support in this research.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Food and Agriculture Organization of the United Nations (FAO): Global cattle distribution and production trends (2022). https://www.fao.org/livestock-systems/global-distributions/cattle/en/. Accessed 02 Sept 2024
2. Monteiro, A., Santos, S., Gonçalves, P.: Precision agriculture for crop and livestock farming–brief review. Animals **11**(8), 2345 (2021)
3. Ginger, L., et al.: A six-step process to explore facial expressions performances to detect pain in dairy cows with lipopolysaccharide-induced clinical mastitis. Appl. Animal Behav. Sci. **264**, 105951 (2023)
4. Kumar, P., Luo, S., Shaukat, K.: A comprehensive review of deep learning approaches for animal detection on video data. Int. J. Adv. Comput. Sci. Appl. **14**(11) (2023)
5. Billah, M., Wang, X., Yu, J., Jiang, Y.: Real-time goat face recognition using convolutional neural network. Comput. Electron. Agric. **194**, 106730 (2022)
6. Jaddoa, M.A., Gonzalez, L., Cuthbertson, H., Al-Jumaily, A.: Multiview eye localisation to measure cattle body temperature based on automated thermal image processing and computer vision. Infrared Phys. Technol. **119**, 103932 (2021)
7. Xu, B., et al.: Cattlefacenet: a cattle face identification approach based on retinaface and arcface loss. Comput. Electron. Agric. **193**, 106675 (2022)
8. Bergman, N., Yitzhaky, Y., Halachmi, I.: Biometric identification of dairy cows via real-time facial recognition. Animal **18**(3), 101079 (2024)
9. Zhang, X., Xuan, C., Xue, J., Chen, B., Ma, Y.: LSR-YOLO: a high-precision, lightweight model for sheep face recognition on the mobile end. Animals **13**(11), 1824 (2023)
10. Kawagoe, Y., Kobayashi, I., Zin, T.T.: Facial region analysis for individual identification of cows and feeding time estimation. Agriculture **13**(5), 1016 (2023)
11. Mao, Y., Liu, Y.: Pet dog facial expression recognition based on convolutional neural network and improved whale optimization algorithm. Sci. Rep. **13**(1), 3314 (2023)
12. Tassinari, P., et al.: A computer vision approach based on deep learning for the detection of dairy cows in free stall barn. Comput. Electron. Agric. **182**, 106030 (2021)
13. Smink, M., Liu, H., Döpfer, D., Lee, Y.J.: Computer vision on the edge: individual cattle identification in real-time with readmycow system. In: IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 7056–7065 (2024)
14. Gupta, H., et al.: Computer vision-based approach for automatic detection of dairy cow breed. Electronics **11**(22), 3791 (2022)
15. Zou, Z., Chen, K., Shi, Z., Guo, Y., Ye, J.: Object detection in 20 years: a survey. Proc. IEEE **111**(3), 257–276 (2023)
16. Redmon, J.: You only look once: unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016)

17. Hussain, M., Khanam, R.: In-depth review of yolov1 to YOLOv10 variants for enhanced photovoltaic defect detection. In: Solar, vol. 4, pp. 351–386. MDPI (2024)
18. Almeida, M.E.G., de Souza Marins, T., da Silva Maciel, L.M., Villela, S.M., Vieira, M.B.: Annotated Cattle Images (ACI) dataset (2025). https://doi.org/10.17632/rzyc5ssvzx.2, https://data.mendeley.com/datasets/rzyc5ssvzx/2