

# When Language Model Guides Vision: Grounding DINO for Cattle Muzzle Detection

Rabin Dulal<sup>1,2</sup>, Lihong Zheng<sup>1,2</sup>, and Muhammad Ashad Kabir<sup>1,2</sup>

<sup>1</sup> School of Computing, Mathematics and Engineering, Charles Sturt University,  
NSW 2678, Australia

<sup>2</sup> Food Agility CRC Ltd, Sydney, NSW 2000, Australia  
`{rdulal, lzheng, akabir}@csu.edu.au`

**Abstract.** Muzzle patterns are among the most effective biometric traits for cattle identification. Fast and accurate detection of the muzzle region as the region of interest is critical to automatic visual cattle identification.. Earlier approaches relied on manual detection, which is labor-intensive and inconsistent. Recently, automated methods using supervised models like YOLO have become popular for muzzle detection. Although effective, these methods require extensive annotated datasets and tend to be trained data-dependent, limiting their performance on new or unseen cattle. To address these limitations, this study proposes a zero-shot muzzle detection framework based on Grounding DINO, a vision-language model capable of detecting muzzles without any task-specific training or annotated data. This approach leverages natural language prompts to guide detection, enabling scalable and flexible muzzle localization across diverse breeds and environments. Our model achieves a mean Average Precision (mAP)@0.5 of 76.8%, demonstrating promising performance without requiring annotated data. To our knowledge, this is the first research to provide a real-world, industry-oriented, and annotation-free solution for cattle muzzle detection. The framework offers a practical alternative to supervised methods, promising improved adaptability and ease of deployment in livestock monitoring applications.

**Keywords:** Cattle identification · CNN · Transformer · Zero shot detection · Grounding DINO · Deep learning

## 1 Introduction

Accurate cattle identification is essential for biosecurity and livestock management. Traditional methods like RFID tags face issues such as tag loss and tampering [10, 26, 6]. Biometric approaches using features like the retina, iris, and muzzle offer more reliable alternatives [10]. Muzzle patterns are particularly effective due to their uniqueness and ease of capture.

Deep learning models, especially CNNs and transformers, have improved muzzle-based identification but depend on accurate muzzle detection. Early approaches relied on ink marks and manual cropping, which was labor-intensive and prone to inconsistency. While recent object detectors such as YOLO [23] and its

variants have automated this task [26,7], they require annotated datasets and retraining for new cattle [21,3]. Although these models improve accuracy and efficiency, they require annotated datasets with labeled muzzle regions for supervised training [21]. However, collecting such large-scale labeled data is often challenging, expensive, and also requires domain-specific knowledge. To address this, we explore zero-shot muzzle detection as a scalable alternative.

Zero-shot object detection (ZSD) addresses the challenge of detecting objects, such as cattle muzzles, without annotated data [21,3]. It mimics the human ability to generalize from known to unseen categories using shared semantic information. Early ZSD models, such as DeVISE and ALE, used fixed semantic embeddings (e.g., Word2Vec, GloVe) and aligned them with visual features via metric learning. However, their performance was limited in complex visual scenes and lacked contextual understanding.

Recent models like RegionCLIP [34] and Grounding DINO [16] leverage transformer-based architectures and large-scale image-text pairs to learn joint representations. These systems support fine-grained alignment between visual and textual features, improving generalization to unseen classes. In ZSD, models are trained on annotated seen classes with text descriptions and can detect unseen objects at inference by matching visual regions to text embeddings, without retraining [21,16]. State-of-the-art ZSD methods include OWL-ViT [19], ViLD [8], RegionCLIP, OV-DETR [32], DetCLIP [31], OmDet [33], and Grounding DINO.

This research adopts Grounding DINO, a language-guided zero-shot object detection model, for the task of cattle muzzle detection. Inspired by the human ability to generalize from known to unknown concepts by understanding semantics, we explore models that align visual features with semantic descriptions. This approach does not require training the model, as it leverages pre-trained vision-language alignment to perform detection based solely on textual prompts such as “cattle muzzle” or “nose of a cattle”. By eliminating the reliance on manually annotated training data and the necessity for model retraining, the proposed method facilitates scalable, adaptable, and efficient muzzle detection in real-world applications. The principal contributions of this study are as follows:

- This study presents, for the first time to our knowledge, a muzzle detection model utilizing a zero-shot muzzle detection approach.
- We conducted a comprehensive evaluation of seven state-of-the-art zero-shot object detection models that combine language and vision for the task of muzzle detection. This comparison establishes a benchmark to guide and support future research in this domain.

## 2 Related Works

This section outlines the research efforts focused on the development and evaluation of muzzle detection methods.

Research [11] developed a biometric identification system using Haar cascade classifiers [14] to detect cattle faces, followed by segmentation to isolate the

muzzle region. Deep learning further enhanced this process. Research [26] applied YOLOv3 [24] to detect muzzle regions, while research [30] used an improved YOLOv5 [27] model. Similarly, other studies [12,2,1] have effectively utilized YOLO-based models for precise and automated muzzle extraction.

Detecting the muzzle using deep learning models poses challenges due to the need for large amounts of labeled data, which increases the time, cost, and computational resources required for model training and deployment. Unlike existing studies that depend heavily on large, labeled datasets of cattle muzzle images, which can be costly and time-consuming to collect and annotate, this research adopts a zero-shot muzzle detection approach. Zero-shot detection enables the model to identify and localize the muzzle region without requiring any prior labeled examples of muzzle data. This approach reduces the dependency on extensive manual annotation, thereby saving time and resources while maintaining effective detection performance. Most of the prior works employed fine-tuning strategies using pre-trained object detection models such as YOLOv3 [24], YOLOv5 [27], YOLOv7 [29], and YOLOv8 [28], adapting them to specific datasets through additional training. Traditional machine learning methods, such as the Haar cascade classifier, did not involve either fine-tuning or zero-shot learning. In contrast, this study uses a zero-shot detection approach with Grounding DINO. This method does not require retraining on specific tasks and can detect objects without having seen the target classes during training.

### 3 Methodology

The methodology of this research is illustrated in Fig. 1. It begins with input data, which is a pair of an image and a text prompt. Seven different ZSD models are selected for this research. The outputs are evaluated using appropriate evaluation metrics, such as mean Average Precision (mAP), and the best-performing model is subsequently selected.

#### 3.1 Datasets

To conduct zero-shot muzzle detection, we captured cattle images (Data1) at Charles Sturt University’s Global Digital Farm in Wagga Wagga, Australia, under the approved animal ethics protocol ID A21414. Additionally, publicly available datasets from UNE [26] and NUACES [1] were used. Table 1 presents a summary of the datasets that are used. All datasets are collected in different geographical locations, conditions, devices, and breeds.

#### 3.2 Data Preprocessing

To evaluate the performance of ZSD models for muzzle detection, a ground truth dataset was required for objective comparison. Although ZSD models do not require annotated data for the target class, labeled data is essential for model evaluation. In this study, we used Roboflow, a widely adopted annotation and

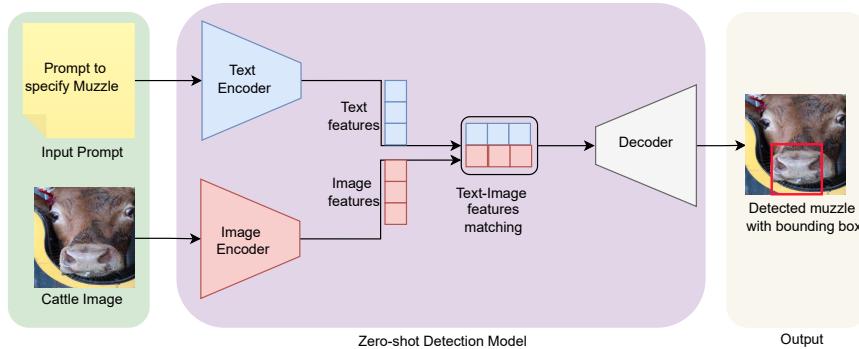


Fig. 1: A general framework for zero-shot muzzle detection using a zero-shot detection model guided by text prompts that describe the muzzle region in cattle images.

Table 1: Summary of datasets used for zero-shot muzzle detection.

Dataset	Image Count	Breed	Location	Device
Data1	163	Angus	Wagga Wagga, Australia	Nikon D7200
UNE	2632	Angus, Simmental, Hereford Charolais	Armidale, Australia	Canon D800
NUCES	2893	Not mentioned	Lahore, Pakistan	Canon D80, Oppo A76, OnePlus 9 Pro

dataset management platform, to draw bounding boxes around the muzzle region of cattle head images. The annotated images were exported in the COCO format, containing image metadata along with the corresponding bounding box coordinates and class labels. These annotations served exclusively for the evaluation, enabling us to assess detection performance.

### 3.3 Zero-shot Detection Models

The selected seven ZSD models are OWL-ViT, ViLD, RegionCLIP, OV-DETR, DetCLIP, OmDet, and Grounding DINO. These models were chosen based on two main criteria. First, each model supports text prompts in sentence format, allowing the use of detailed natural language descriptions to guide detection. This is particularly important for detecting specific and subtle features like the cattle muzzle, where a sentence (e.g., “the front part of a cattle’s face including the nostrils and mouth”) can convey richer semantic information than a single word. Second, all selected models have demonstrated strong performance on standard benchmark datasets such as COCO [15], LVIS [9], and ODinW [13], reflecting their robustness and generalizability across diverse object categories. By leveraging both visual features and semantic text embeddings, these models offer a promising approach to zero-shot muzzle detection tasks.

Table 2 presents a summary of the selected zero-shot detection models, categorizing them based on their base architecture, open-vocabulary strategy, and detection approach. The base architecture refers to the components used for extracting features from both images and texts. It typically includes a visual backbone like ViT [5], Faster R-CNN [25], ConvNeXt [18], DETR [4], and Swin [17] for image feature extraction, and a language encoder like CLIP [22] for processing text descriptions or prompts. The open-vocabulary strategy describes how the model integrates visual features with language supervision to enable the detection of previously unseen categories. Techniques vary from CLIP-guided classification and region-level contrastive learning to query-based grounding and multi-source semantic distillation. The detection approach indicates the model’s detection paradigm, whether it requires a region proposal network followed by classification, or end-to-end, where object queries are predicted directly using transformers without separate proposals. Most recent models like OWL-ViT, OV-DETR, DetCLIP, OmDet, and Grounding DINO adopt end-to-end frameworks for better language-vision integration and zero-shot generalization, whereas ViLD and RegionCLIP retain a proposal-based two-stage design for compatibility with traditional detectors.

Table 2: Summary of the selected Zero-shot Detection Models.

Model	Base Architecture	Detection Approach	Open-Vocab Strategy
OWL-ViT [19]	ViT+CLIP	End-to-end	CLIP-guided transformer with prompt-based classification
ViLD [8]	Faster R-CNN+CLIP	Proposal-based	Region-level CLIP knowledge distillation
RegionCLIP [34]	Faster R-CNN+CLIP	Proposal-based	Region-text contrastive learning
OV-DETR [32]	DETR+CLIP	End-to-end	CLIP embeddings as query tokens in DETR
DetCLIP [31]	Swin+CLIP	End-to-end	Concept dictionary, multi-source contrastive learning
OmDet [33]	ConvNeXt + CLIP	End-to-end	General open-world detection framework
Grounding DINO [16]	DINO+encoder	End-to-end	Language-guided queries with cross-modal decoder

Based on the methodological framework described above, we conducted a series of experiments to assess the effectiveness of the selected models. The following section reports the results of these evaluations.

## 4 Experiments and Results

All experiments were conducted on an HP Victus laptop equipped with an Intel Core i7 processor and an NVIDIA GeForce RTX 4070 Laptop GPU. The im-

Table 3: Detection performance (mAP@0.5) of different prompt variations used with Grounding DINO. The **bold** value represents the best performance.

Prompt No.	Prompt Description	mAP@0.5
1	cattle muzzle	0.242
2	[Prompt 1], the nose and mouth of a cattle	0.386
3	[Prompt 2], the lower front part of a cattle's face	0.536
4	[Prompt 3], the snout of a cattle	0.713
5	[Prompt 4], <b>the area around the nostrils and lips of a cattle</b>	<b>0.768</b>
6	[Prompt 5], the fleshy soft rounded part of a cattle's face used for eating and smelling	0.755
7	[Prompt 6], cattle's face with visible nasal cavities	0.682

plementation was carried out using Python 3.13 and the PyTorch deep learning framework. All selected models were used with their pre-trained weights for zero-shot detection. We initially conducted a series of experiments using Grounding DINO to identify the most effective text prompt for the muzzle detection task. Multiple prompt variations were tested, and their detection performances were compared. The prompt that yielded the highest accuracy was selected as the optimal one. Following this, the selected prompt was uniformly applied across all seven zero-shot learning models to evaluate and compare their detection capabilities under consistent textual guidance. This approach allowed us to isolate model performance from prompt variability and ensured a fair comparison across different architectures. The first two experiments were evaluated by combining the Data1 and UNE datasets. Finally, the performance of the selected ZSD model is compared with other existing muzzle detector models YOLOv3 [26], YOLOv5 [30], YOLOv7 [1], and YOLOv8 [12], with varying training samples of each dataset.

#### 4.1 Selection of Prompts

This section presents an analysis of the different prompts to identify the most effective prompt for muzzle detection. It was observed that customized prompts providing detailed or descriptive information yielded better performance compared to generic prompts commonly used by zero-shot learning models [20]. Based on this observation, we began with a basic, generic prompt and progressively added descriptive phrases or sentences to enhance the specificity of the muzzle. The detection performance of each prompt was evaluated using mAP@0.5, and the results are summarized in Table 3.

In this experiment, we began with a generalized and minimal prompt (“cattle muzzle”) and progressively added more descriptive textual elements to enhance the semantic understanding of the target muzzle. Each subsequent prompt built upon the previous one by appending additional context or anatomical details, such as “the nose and mouth of a cattle” and “the area around the nostrils and lips of a cattle”. This incremental refinement was aimed at improving the alignment between the visual features and the text input used by the Grounding DINO model. As shown in Table 3, this strategy led to improved detection performance, with the highest mAP@0.5 achieved when using a prompt that

balanced specificity and conciseness (Prompt 5). However, adding excessive or overly detailed phrases beyond this point (e.g., Prompt 6 and 7) slightly decreased performance, suggesting that overly verbose prompts may dilute the model’s focus.

#### 4.2 Performance of ZSD Models

The primary objective of this research is to investigate the feasibility of applying ZSD models for cattle muzzle detection. To facilitate a comparative performance evaluation, we report the results of selected models using standard metrics. All models were evaluated using their respective base variants to maintain consistency in performance comparison. Specifically, we present the muzzle detection performance in terms of mAP@0.5, mAP@0.75, and mAP@[0.50:0.95]. These results are summarized in Table 4, enabling a comprehensive comparison of each model’s accuracy. Among all models, Grounding DINO clearly achieves

Table 4: Detection performance (mAP) for ZSD models on cattle muzzle detection. Results show mean  $\pm$  95% confidence interval (CI) over five runs. The **bold** value represents the best performance.

Model	mAP		
	0.50:0.95	0.5	0.75
OWL-ViT	0.103 $\pm$ 0.119	0.298 $\pm$ 0.026	0.032 $\pm$ 0.011
ViLD	0.0852 $\pm$ 0.014	0.2178 $\pm$ 0.018	0.0782 $\pm$ 0.016
RegionCLIP	0.160 $\pm$ 0.017	0.489 $\pm$ 0.028	0.104 $\pm$ 0.021
OV-DETR	0.0985 $\pm$ 0.012	0.368 $\pm$ 0.027	0.0792 $\pm$ 0.013
DetCLIP	0.132 $\pm$ 0.015	0.386 $\pm$ 0.022	0.102 $\pm$ 0.020
OmDet	0.227 $\pm$ 0.019	0.538 $\pm$ 0.030	0.116 $\pm$ 0.018
Grounding DINO	<b>0.340 <math>\pm</math> 0.021</b>	<b>0.768 <math>\pm</math> 0.025</b>	<b>0.180 <math>\pm</math> 0.021</b>

the best overall performance across all thresholds, with scores of  $0.340 \pm 0.031$  (mAP@[0.50:0.95]),  $0.768 \pm 0.025$  (mAP@0.5), and  $0.180 \pm 0.021$  (mAP@0.75). This indicates that the model not only detects muzzle regions reliably but also localizes them with high precision. Its superior performance can be attributed to the integration of a robust image-text alignment module using CLIP [22] with a transformer-based object detector, enabling fine-grained grounding even without task-specific fine-tuning.

OmDet’s performance comes second with scores of  $0.227 \pm 0.019$  (mAP@[0.50:0.95]),  $0.538 \pm 0.030$  (mAP@0.5), and  $0.116 \pm 0.018$  (mAP@0.75). While it performs relatively well at lower IoU thresholds, its score at mAP@0.75 indicates a moderate decline in localization accuracy. This gap suggests that OmDet’s detection head is less precise in tight object localization compared to Grounding DINO, though its dynamic prompt learning contributes to better generalization.

RegionCLIP and DetCLIP deliver competitive mid-range performance. RegionCLIP achieves  $0.160 \pm 0.017$  (mAP@[0.50:0.95]),  $0.489 \pm 0.028$  (mAP@0.5), and

$0.104 \pm 0.021$  (mAP@0.75), while DetCLIP scores  $0.132 \pm 0.015$ ,  $0.386 \pm 0.022$ , and  $0.102 \pm 0.020$  respectively. Both models leverage CLIP-based representations enriched with region-level features and contrastive training, resulting in reasonable detection and moderate localization performance. However, their lower mAP@0.75 values reflect a drop in precision for tightly localized predictions.

OV-DETR performs lower with scores of  $0.0985 \pm 0.012$  (mAP@0.50:0.95),  $0.368 \pm 0.027$  (mAP@0.5), and  $0.0792 \pm 0.013$  (mAP@0.75). It outperforms ViLD and OWL-ViT, but its lack of fine-grained localization capabilities, due to the absence of a strong grounding mechanism, limits its usefulness for tasks like muzzle detection.

ViLD and OWL-ViT performed the poorest. ViLD achieves  $0.0852 \pm 0.014$  (mAP@0.50:0.95),  $0.2178 \pm 0.018$  (mAP@0.5), and  $0.0782 \pm 0.016$  (mAP@0.75), while OWL-ViT records  $0.103 \pm 0.119$ ,  $0.298 \pm 0.026$ , and only  $0.032 \pm 0.011$  at mAP@0.75. Their relatively low scores across all thresholds, especially at higher IoU levels, indicate poor localization ability, likely due to their reliance on coarse-level CLIP features and the absence of dedicated detection heads or spatial reasoning modules.

### 4.3 Performance Comparison with Existing Models

The objective of this experiment is to conduct a comparative study between Grounding DINO and existing muzzle detection models, with a focus on determining the minimum number of labeled muzzle images required for those models to achieve accuracy comparable to that of Grounding DINO. The existing deep learning based muzzle detection models are YOLOv3, YOLOv5, YOLOv7, and YOLOv8. Haar cascade is a classical machine learning model that relies on handcrafted features and does not support automatic learning by extracting features from the input images. Therefore, this study considers existing deep learning muzzle detection models. All three datasets were trained using different numbers of labeled training images: 10, 20, 40, 80, and 160. All experiments were conducted using an image size of  $640 \times 640$  pixels and were trained for 100 epochs with a batch size of 32. All models were evaluated based on the mean Average Precision at IoU threshold 0.5 (mAP@0.5), and the results are presented in Table 5. Results indicate that Grounding DINO achieves strong zero-shot performance, with mAP@0.5 scores of 0.753 (Data1), 0.789 (UNE), and 0.758 (NUCES), without requiring any fine-tuning or training. In contrast, all YOLO variants require between 40 and 80 labeled training samples to achieve comparable performance levels, highlighting the effectiveness of Grounding DINO in low-data scenarios.

Grounding DINO offers significant practical advantages. Fine-tuning YOLO-based models (YOLOv3, YOLOv5, YOLOv7, YOLOv8) demands a considerable amount of time, computational resources, and manual effort. The process involves dataset preparation, manual annotation, training on GPU-enabled systems, and extensive hyperparameter tuning. Furthermore, the process must be repeated for each target dataset, increasing both the time and cost. In contrast, Grounding DINO provides a cost-effective and scalable alternative, capable of

Table 5: Comparison of Muzzle Detection Models with Varying Number of Training Samples. The **bold** value represents the Grounding DINO’s performance. Light green cells highlight cases where other models achieve lower scores than Grounding DINO for the same training sample size.

Model (Approach)	Training Samples	mAP@0.5 on Dataset		
		Data1	UNE	NUCES
YOLOv3 (Fine tuned)	160	0.829	0.985	0.988
	80	0.834	0.887	0.954
	40	<b>0.712</b>	<b>0.697</b>	<b>0.661</b>
	20	<b>0.661</b>	<b>0.635</b>	<b>0.596</b>
	10	<b>0.491</b>	<b>0.365</b>	<b>0.040</b>
YOLOv5 (Fine tuned)	160	0.874	0.975	0.995
	80	0.848	0.975	0.995
	40	<b>0.738</b>	<b>0.655</b>	<b>0.682</b>
	20	<b>0.637</b>	<b>0.629</b>	<b>0.437</b>
	10	<b>0.443</b>	<b>0.464</b>	<b>0.039</b>
YOLOv7 (Fine tuned)	160	0.968	0.995	0.995
	80	0.968	0.995	0.995
	40	<b>0.765</b>	<b>0.729</b>	<b>0.592</b>
	20	<b>0.689</b>	<b>0.581</b>	<b>0.441</b>
	10	<b>0.431</b>	<b>0.419</b>	<b>0.189</b>
YOLOv8 (Fine tuned)	160	0.985	0.995	0.995
	80	0.989	0.995	0.995
	40	<b>0.731</b>	<b>0.728</b>	<b>0.647</b>
	20	<b>0.558</b>	<b>0.493</b>	<b>0.227</b>
	10	<b>0.139</b>	<b>0.256</b>	<b>0.045</b>
Grounding DINO (Zero-shot)	0	<b>0.753</b>	<b>0.789</b>	<b>0.758</b>

producing competitive results without the need for any supervised training. Its ability to generalize across datasets without re-training makes it particularly suitable for applications with limited data availability or constrained computational budgets.

## 5 Discussion

As presented in Table 5, the recent studies have utilized supervised deep learning models such as YOLOv3, YOLOv5, YOLOv7, and YOLOv8, achieving high accuracy, up to 99.5% in some cases. However, a major limitation of supervised models is their dependency on the specific dataset they are trained on. These models are typically optimized for particular cattle breeds or controlled environments. If a new dataset is introduced or if the breed changes, the entire dataset must be annotated and retrained from scratch, which is resource-intensive and impractical for real-time or large-scale applications.

In contrast, the zero-shot detection method explored in this study using Grounding DINO does not require task-specific retraining. It leverages natural language prompts (e.g., “cattle muzzle” or “snout of a cattle”) to detect the target object, making it more flexible and generalizable across different cattle breeds and unseen environments. Although the accuracy (76.8%) is lower than that of supervised models, this trade-off is offset by the method’s adaptability, annotation-free deployment, and potential to scale across breeds and geographies. Therefore, zero-shot detection presents a promising direction for practical muzzle detection applications.

Future work can focus on further improving the accuracy of zero-shot muzzle detection through advanced vision-language models and refined prompt design. Using detailed, domain-specific prompts can lead to more accurate and reliable results. A promising direction is enabling natural language queries, such as “How many cattle have a white muzzle?” or “What is the medication history of cattle with a white muzzle with red marks?”, to support intuitive interaction and enhance traceability. This would allow farmers and livestock managers to monitor and search for specific cattle using simple descriptions. Further evaluation across breeds and conditions, along with optimization for edge deployment, will support practical and scalable use in real-world farming environments.

## 6 Conclusion

In this study, we selected seven ZSD models capable of handling long and descriptive prompts to detect cattle muzzles. We explored various custom natural language prompts to describe muzzle characteristics and identified the most effective ones for detection performance. Through a comprehensive evaluation, we assessed the performance of all seven ZSD models and found that Grounding DINO outperformed the others in terms of accuracy and consistency. Unlike traditional supervised methods, Grounding DINO offers the advantage of detecting muzzles without requiring task-specific training data, making it more adaptable across breeds and environments. These findings demonstrate the potential of prompt-driven ZSD models for practical and scalable livestock monitoring. Nevertheless, certain limitations should be acknowledged. Large language models remain susceptible to hallucination, which may compromise the reliability of generated guidance. Furthermore, the approach is sensitive to prompt design, as variations in phrasing can influence both the consistency and quality of results.

## Acknowledgment

This project was supported by funding from Food Agility CRC Ltd, funded under the Commonwealth Government CRC Program. The CRC Program supports industry-led collaborations between industry, researchers, and the community. We also thank Dr Shawn McGrath, Mr. Jonathan Medway, and Prof. Dave Swain for their assistance with the project.

## References

1. Ahmed, S.U., Frnda, J., Waqas, M., Khan, M.H.: Dataset of cattle biometrics through muzzle images. Data in Brief p. 110125 (2024)
2. Anitha, J., Avanthika, R., Kavipriya, B., Vishnupriya, S.: Cattle identification using muzzle images. In: Applied Data Science and Smart Systems, pp. 370–377 (2024)
3. Cao, W., Yao, X., Xu, Z., Liu, Y., Pan, Y., Ming, Z.: A survey of zero-shot object detection. Big Data Mining and Analytics **8**(3), 726–750 (2025)
4. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: Proceedings of the European Conference on Computer Vision. pp. 213–229. Springer (2020)
5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv:2010.11929 (2020)
6. Dugal, R., Zheng, L., Kabir, M.A.: Mhaff: Multihead attention feature fusion of cnn and transformer for cattle identification. IEEE Transactions on AgriFood Electronics pp. 1–12 (2025)
7. Dugal, R., Zheng, L., Kabir, M.A., McGrath, S., Medway, J., Swain, D., Swain, W.: Automatic cattle identification using yolov5 and mosaic augmentation: A comparative analysis. In: International Conference on Digital Image Computing: Techniques and Applications. pp. 1–8. IEEE (2022)
8. Gu, X., Lin, T.Y., Kuo, W., Cui, Y.: Open-vocabulary object detection via vision and language knowledge distillation. arXiv:2104.13921 (2021)
9. Gupta, A., Dollar, P., Girshick, R.: Lvvis: A dataset for large vocabulary instance segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5356–5364 (2019)
10. Hossain, M.E., Kabir, A., Zheng, L., Swain, D., McGrath, S., Medway, J.: A systematic review of machine learning techniques for cattle identification: Datasets, methods and future directions. Artificial Intelligence in Agriculture (2022)
11. Kusakunniran, W., Wiratsudakul, A., Chuachan, U., Kanchanapreechakorn, S., Imaromkul, T., Suksriupatham, N., Thongkanchorn, K.: Biometric for cattle identification using muzzle patterns. International journal of pattern recognition and artificial intelligence **34**(12), 2056007 (2020)
12. Lee, T., Na, Y., Kim, B.G., Lee, S., Choi, Y.: Identification of individual hanwoo cattle by muzzle pattern images through deep learning. Animals **13**(18) (2023)
13. Li, L.H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.N., et al.: Grounded language-image pre-training. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 10965–10975 (2022)
14. Lienhart, R., Maydt, J.: An extended set of haar-like features for rapid object detection. In: Proceedings. international conference on image processing. vol. 1, pp. I–I. IEEE (2002)
15. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Proceedings of the European Conference on Computer Vision. pp. 740–755. Springer (2014)
16. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In: European Conference on Computer Vision. pp. 38–55. Springer (2024)

17. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 10012–10022 (2021)
18. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE Computer Vision and Pattern Recognition. pp. 11976–11986 (2022)
19. Minderer, M., Gritsenko, A., Stone, A., Neumann, M., Weissenborn, D., Dosovitskiy, A., Mahendran, A., Arnab, A., Dehghani, M., Shen, Z., et al.: Simple open-vocabulary object detection. In: European Conference on Computer Vision. pp. 728–755. Springer (2022)
20. Nawaz, U., Awais, M., Gani, H., Naseer, M., Khan, F., Khan, S., Anwer, R.M.: Agriclip: Adapting clip for agriculture and livestock via domain-specialized cross-model alignment. arXiv:2410.01407 (2024)
21. Pourpanah, F., Abdar, M., Luo, Y., Zhou, X., Wang, R., Lim, C.P., Wang, X.Z., Wu, Q.J.: A review of generalized zero-shot learning methods. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(4), 4051–4070 (2022)
22. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: Proceedings of the International Conference on Machine Learning. pp. 8748–8763. PMLR (2021)
23. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 779–788 (2016)
24. Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. arXiv:1804.02767 (2018)
25. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in Neural Information Processing Systems **28** (2015)
26. Shoaieiipour, A., Falzon, G., Kwan, P., Hadavi, N., Cowley, F.C., Paul, D.: Automated muzzle detection and biometric identification via few-shot deep transfer learning of mixed breed cattle. Agronomy **11**(11), 2365 (2021)
27. Ultralytics: The state-of-the-art yolo model. <https://github.com/ultralytics/yolov5>, (accessed: 15.01.2023)
28. Ultralytics: The state-of-the-art yolo model. <https://ultralytics.com/yolov8>, (accessed: 15.01.2023)
29. Wang, C.Y., Bochkovskiy, A., Liao, H.Y.M.: Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. arXiv:2207.02696 (2022)
30. Xu, X., Wang, Y., Shang, Y., Yang, G., Hua, Z., Wang, Z., Song, H.: Few-shot cow identification via meta-learning. Information Processing in Agriculture (2024)
31. Yao, L., Han, J., Wen, Y., Liang, X., Xu, D., Zhang, W., Li, Z., Xu, C., Xu, H.: Detclip: Dictionary-enriched visual-concept paralleled pre-training for open-world detection. Advances in Neural Information Processing Systems **35**, 9125–9138 (2022)
32. Zang, Y., Li, W., Zhou, K., Huang, C., Loy, C.C.: Open-vocabulary detr with conditional matching. In: European Conference on Computer Vision. Springer (2022)
33. Zhao, T., Liu, P., Lee, K.: Omdet: Large-scale vision-language multi-dataset pre-training with multimodal detection network. IET Computer Vision **18**(5) (2024)
34. Zhong, Y., Yang, J., Zhang, P., Li, C., Codella, N., Li, L.H., Zhou, L., Dai, X., Yuan, L., Li, Y., et al.: Regionclip: Region-based language-image pretraining. In: Proceedings of the IEEE Computer Vision and Pattern Recognition. pp. 16793–16803 (2022)