YOLOv5-VF-W3: A novel cattle body detection approach for precision livestock farming

Wangli Hao[1], Chao Ren[1], Meng Han[1], Fuzhong Li[1], Zhenyu Liu[2*]

(1. *School of Software, Shanxi Agricultural University, Shanxi 030801, China*;
2. *College of Engineering, Shanxi Agricultural University, Shanxi 030801, China*)

Abstract: Accurate cattle body detection can significantly enhance the efficiency and quality of animal husbandry production. Traditional manual observation approaches are not only inefficient but also lack objectivity, while computer vision-based methods demand prolonged training periods and present challenges in implementation. To address these issues, this paper develops a novel precise cattle body detection solution, namely YOLOv5-VF-W3. By introducing the Varifocal loss, the YOLOv5-VF-W3 model can handle imbalanced samples and focus more attention on difficult-to-recognize instances. Additionally, the introduction of the WIoUv3 loss function provides the model with a wise gradient gain allocation strategy. This strategy reduces the competitiveness of high-quality anchor boxes while mitigating harmful gradients produced by low-quality anchor boxes, thereby emphasizing anchor boxes of ordinary quality. Through these enhancements, the YOLOv5-VF-W3 model can accurately detect cattle bodies, improving the efficiency and quality of animal husbandry production. Numerous experimental results have demonstrated that the proposed YOLOv5-VF-W3 model achieves superior cattle body detection results in both quantitative and qualitative evaluation criteria. Specifically, the YOLOv5-VF-W3 model achieves an mAP of 95.2% in cattle body detection, with individual cattle detection, leg detection, and head detection reaching 95.3%, 94.8%, and 95.4%, respectively. Furthermore, in complex scenarios, especially when dealing with small targets and occlusions, the model can accurately and efficiently detect individual cattle and key body parts. This brings new opportunities for the development of precision livestock farming.

Keywords: cattle body detection, varifocal loss, key body parts, WIoUv3 loss
DOI: 10.25165/j.ijabe.20251802.9107

Citation: Hao W L, Ren C, Han M, Li F Z, Liu Z Y. YOLOv5-VF-W3: A novel cattle body detection approach for precision livestock farming. Int J Agric & Biol Eng, 2025; 18(2): 269–277.

1 Introduction

In modern animal husbandry, monitoring the health of cattle is crucial. By observing the body posture, overall condition, eyes, hair, and other relevant factors of cattle, their health status can be assessed. Therefore, to ensure the effectiveness of cattle health monitoring, it is crucial to implement more precise and efficient methods for cattle body detection.

Traditional approaches to cattle health monitoring encompass direct observation or physical contact as means to evaluate their health status. This qualitative technique is adept at identifying early indicators of health problems, thereby enabling timely measures to be implemented[1]. Nonetheless, owing to gaps in the frequency of monitoring, this method risks overlooking crucial health conditions or behavioral patterns that cattle may display at particular instances. Additionally, it demands a considerable investment of human resources and labor.

With the development of Radio Frequency Identification (RFID) technology and wearable devices, physical devices have played a significant role in the automated collection and analysis of individual information and health data of cattle, effectively reducing the pressure of manual monitoring[2]. RFID technology is applied by installing devices such as ear tags on the cattle's ears[3]. Although this method has been widely adopted, it does not fully comply with the requirements of animal welfare. In addition, RFID devices also have some limitations, such as tag detachment, loss, malfunction, or duplication, all of which can affect the accuracy of identification[4]. Meanwhile, wearable devices have been proven to significantly improve the precision of cattle positioning and behavior monitoring[5,6]. However, most of these devices lack the capability to visually capture and validate the movements of cattle through visual imagery. Therefore, in some cases, behavior classification based on wearable devices may not accurately reflect the actual behavior of the cattle, leading to false reports.

With the rapid advancement of computer vision technology, utilizing visual features for cattle inspection has gradually become the core focus of research in this field. Computer vision is utilized to automatically identify individual cattle as well as monitor their behavior and health status, which is of great significance for improving breeding efficiency, disease prevention, and health management[7]. Bercovich et al.[8] and Zhao et al.[9] respectively designed different tools based on computer vision, with the former achieving automatic scoring of cow body condition and the latter achieving a recognition accuracy of 96.72% for cows. Gao et al.[10] utilized a method of multiple feature fusion to extract features of

Received date: 2024-05-30 **Accepted date:** 2025-03-04
Biographies: Wangli Hao, PhD, Associate Professor, research interest: computer vision, pattern recognition, and machine learning, Email: haowangli@sxau.edu.cn; **Chao Ren**, Master candidate, research interest: smart agriculture, Email: sxaurc@stu.sxau.edu.cn; **Meng Han**, PhD candidate, research interest: distributed and parallel computing, distributed machine learning, Email: hanm@hdu.edu.cn; **Fuzhong Li**, PhD, Professor, Doctoral supervisor, the Dean of the Software College of Shanxi Agricultural University, Email: lifuzhong@sxau.edu.cn.

***Corresponding author: Zhenyu Liu**, Professor, Doctoral supervisor, Postdoctoral Researcher. College of Engineering, Shanxi Agricultural University, Shanxi 030801, China. research interest: electromagnetic properties of agricultural materials and livestock informatization. Tel: +86-354-6287098, Email: lzysyb@126.com.

cows and employed a classifier trained with the Gentle Adaboost algorithm for cattle body detection, achieving a detection accuracy of 97.3%. Liu et al.[11] proposed a two-class classification algorithm based on chromatic distortion and brightness distortion, achieving realtime extraction of cow targets under complex background and environmental conditions. Kaur et al.[12] achieved an 83.35% improvement in cattle recognition accuracy by combining the use of a random forest classifier with image feature extraction methods such as SIFT and SURF. The above mentioned traditional computer vision and machine learning-based methods for cattle health monitoring often require redesigning and adjusting parameters when faced with new datasets or tasks. This results in relatively weak model generalization ability.

In order to further improve the performance of cattle body detection models, deep learning-based methods are increasingly gaining attention from researchers. These methods fully leverage the advantages of deep learning technology, such as powerful feature learning, classification, and generalization capabilities, thereby greatly improving the accuracy and efficiency of cattle body detection. Tassinari et al.[13], Lodkaew et al.[14], and Xiao et al.[15] respectively employed YOLOv3[16], YOLOv4[17], and an improved Mask-RCNN[18] model to label and train data collected from indoor Holstein cows with fixed cameras. In their studies, the average precision of cattle body detection reached 64.0%, 90.0%, and 97.39%, respectively. Shao et al.[19], Weber et al.[20], Andrew et al.[21], and Xu et al.[22] respectively utilized YOLOv2[23], YOLOv4, Faster-RCNN[24], and Mask-RCNN models to label and train datasets of cattle herds captured by drones, achieving cattle body detection accuracies of 95.0%, 98.0%, 99.6%, and 96.0%, respectively. These excellent detection results are due to the significant differences in the color of the cattle's body and the pasture, making them easy to distinguish. Additionally, it is worth noting that Faster-RCNN and Mask-RCNN belong to two-stage object detectors, which typically require longer training and detection times, making them unsuitable for realtime detection in cattle farms.

Despite the certain development of cattle body detection technology, especially its outstanding performance in application scenarios such as automatic counting[25] and behavior monitoring[26], relying solely on the detection of the entire cattle body is still insufficient for in-depth analysis of cattle behavior patterns and health status. By accurately identifying the key parts of the cattle body, this approach can not only significantly improve the accuracy of disease diagnosis but also further refine behavior analysis. This includes, but is not limited to, monitoring cattle for lameness[27], identifying rumination behavior[28], and individual identification through facial features of cattle[29,30]. Therefore, the detection of the key parts can facilitate a deeper understanding and management of the cattle herd, thereby optimizing breeding efficiency and animal welfare.

Although the aforementioned methods have demonstrated significant efficacy in cattle body detection tasks, they still exhibit notable limitations in detecting small targets within images, particularly when confronted with complex scenarios involving occlusion.

YOLOv5 (You Only Look Once)[31] is recognized as having significant advantages over its previous versions in object detection tasks. However, it has been observed that YOLOv5 tends to miss the key small-sized parts of cattle, such as the head and legs, and its detection performance is not ideal in scenes where cattle are obstructed.

To enhance the detection capability in such challenging scenarios, this study leverages the strengths of Varifocal loss[32] and WIoUv3[33]. Varifocal loss is a classification loss function that introduces an adjustable parameter to adaptively adjust the weights of positive and negative samples. This mechanism enables the model to focus more on difficult-to-classify samples, such as small or partially occluded targets, thereby improving its learning ability for challenging cases and enhancing the overall performance and robustness of the model.

WIoUv3, on the other hand, is a localization loss function that proposes a dynamic non-monotonic focusing mechanism. This mechanism employs "outlier" to evaluate the quality of anchor boxes instead of relying solely on IoU. It provides a wise gradient gain allocation strategy, which reduces the competitiveness of high-quality anchor boxes while mitigating the harmful gradients produced by low-quality anchor boxes. This approach allows the model to focus more on anchor boxes of ordinary quality, which are more likely to represent partially occluded targets, thereby improving the generalization ability and performance of the model in complex detection scenes.

To address the degraded detection accuracy of small targets in occluded cattle body scenarios, this paper innovatively integrates the Varifocal loss function and the WIoUv3 regression loss function into the YOLOv5 framework.

Overall, the contribution of this paper is summarized as follows:

1) This paper proposes a novel YOLOv5-VF-W3 model designed to efficiently detect cattle bodies. The model addresses the issue of sample imbalance by adjusting the sample weights. Additionally, it improves detection performance in scenarios with small targets and occlusions by focusing on ordinary anchor boxes.

2) Two loss functions, Varifocal and WIoUv3, are first leveraged in the cattle body detection framework. The Varifocal loss function adjusts the weight of positive and negative samples, emphasizing more difficult samples. The WIoUv3 loss improves anchor box quality by incorporating a non-monotonic focus coefficient, which reduces the impact of high-quality anchors and mitigates harmful gradients from low-quality ones. Together, these losses enable the model to better focus on challenging samples and ordinary quality anchor boxes, thereby enhancing both performance and robustness.

3) Through numerous ablation studies, the superior performance of the YOLOv5-VF-W3 model in cattle key body detection tasks is demonstrated in the experiment. Specifically, the model achieves a mean Average Precision (mAP) of 95.2% in cattle body detection, with individual detection scores for the cattle body, legs, and head reaching 95.3%, 94.8%, and 95.4%, respectively. These results emphasize its high efficiency and practical applicability.

2 Materials and methods

2.1 Datasets

This paper aims to develop an efficient detection model for Jinnan cattle, based on image data collected from the Jinnan Cattle Genetic Resource Gene Protection Center in Yuncheng City. The work covered a series of data collection activities on healthy Jinnan cattle from July to October 2021. To ensure the diversity and complexity of the dataset, Canon EOS 1300D cameras and SEA-AL10 smartphones were employed to capture images of the cattle from different angles and under various weather conditions between 7:00 AM and 8:00 PM daily. The cattle were divided into three age groups: calves from birth to six months, young cattle from six

months to two years, and adult cattle older than two years, with all images taken in natural environments, as shown in Figure 1.

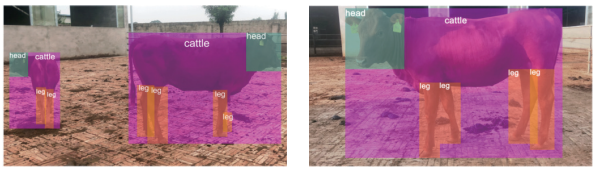
a. Under 6 months

b. 6 months to 2 years old

c. 2 years old and older

Figure 1 The collected data samples for cattle body detection

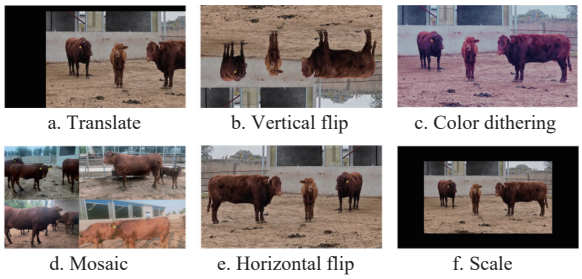
In the process of building the dataset, highly repetitive or blurry images were first manually filtered out from the sequentially captured images to ensure data quality. Subsequently, the Labelimg tool was employed to precisely annotate individual cattle, heads, and legs in the images, with the annotated results saved as TXT format files. Through this series of steps, a dataset containing 8024 images, including 113 images of Jinnan cattle and their annotated files, was ultimately obtained. Some example images from the dataset are shown in Figure 2.



a. Multi-sample b. Single-sample

Figure 2 Some examples of annotated cattle images

To comprehensively evaluate the performance of the developed model, the dataset was randomly divided into training, validation, and test sets at a ratio of 7:2:1. Specifically, the training set contains 5617 samples, the validation set contains 1604 samples, and the test set contains 803 samples. Several instances of cattle data augmentation are shown in Figure 3.



a. Translate b. Vertical flip c. Color dithering

d. Mosaic e. Horizontal flip f. Scale

Figure 3 Several instances of cattle data augmentation

2.2 Technical roadmap

The technical roadmap of the cattle body detection model proposed in this paper is shown in Figure 4. To minimize the impact of data noise on model training, video frames of cattle are randomly sampled and preprocessed to generate images. Furthermore, to enhance the diversity of the data and simulate different scenarios, various data augmentation techniques, including translate, vertical flip, color dithering, mosaic, horizontal flip, and scale, were employed. The application of these methods aims to improve the generalization ability and accuracy of the model under various conditions. The data is then accurately annotated to identify the key body parts of the cattle, providing precise supervisory signals for model training. Next, the YOLOv5-VF-W3 model is utilized for training, enabling the model to learn to recognize and detect the key body parts of the cattle. Finally, the trained model is evaluated, and the accuracy and effectiveness of the detection results are analyzed to ensure the practicality and reliability of the detection model.

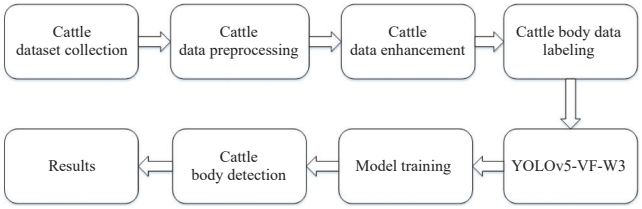


Figure 4 The technical roadmap of cattle body detection model

2.3 YOLOv5-VF-W3 model

The novel YOLOv5-VF-W3 model is proposed by ingeniously integrating the Varifocal and WIoUv3 loss functions into YOLOv5. Specifically, the loss contribution of negative samples is reduced by the Varifocal loss function, thereby improving the classification performance of the model. Additionally, a non-monotonic focusing mechanism is introduced by the WIoUv3 loss function, enabling a gradient gain allocation strategy that best fits the current situation for anchor boxes of different qualities during training, which in turn enhances the localization performance of the model. As a result, this ingenious design of loss significantly enhances the detection performance of the model. Moreover, this design strengthens the robustness of the model, allowing the network to learn more comprehensive features. In this paper, our primary innovation lies in the application of loss functions. Therefore, for the network of the model, please refer to YOLOv5[31].

2.3.1 The loss function

In the task of cattle body detection, the choice of loss function has a significant impact on the training and performance of the model. Different loss functions can guide the model to learn different features and representations, affecting the accuracy and stability of the model in the object detection task. The loss function of the proposed YOLOv5-VF-W3 model consists of three terms, including the confidence loss (L_{conf}), the localization loss (L_{loc}), and the classification loss (L_{cls}), respectively.

$$\text{Loss} = L_{\text{conf}} + L_{\text{loc}} + L_{\text{cls}} \tag{1}$$

In the following, the detailed definition of these loss terms will be given.

2.3.2 The confidence loss

The confidence loss L_{conf} quantifies the accuracy of the model in predicting the presence or absence of a target. Thus, it ensures that the model can accurately assess whether cattle are present in the image.

The confidence loss employed in this paper shares the same definition of that in YOLOv5[31], and detailed information can be found in the corresponding paper.

2.3.3 The localization loss: WIoUv3 loss

The localization loss L_{WIoUv3} is formulated as follows:

$$L_{\text{loc}} = L_{\text{WIoUv3}} = r \cdot R_{\text{WIoU}} \cdot L_{\text{IoU}}, \quad r = \frac{\beta}{\delta \cdot \alpha^{\beta-\delta}} \tag{2}$$

where α and δ are configured as hyperparameters; β defines the outlier degree, as defined in Equation 3; r represents a non-monotonic focus coefficient; L_{IoU} indicates the degree of mismatch

or dissimilarity between the ground truth box and the predicted box, presented in Equation 4; R_{WIoU} denotes a loss function constructed based on a distance metric, as shown in Equation 6; and L_{WIoUv3} defines a loss function featuring a dynamic non-monotonic focusing mechanism and incorporating geometric constraints such as distance and overlap.

$$\beta = \frac{L_{\mathrm{IoU}}^{*}}{\overline{L_{\mathrm{IoU}}}} \in [0, +\infty) \tag{3}$$

where, L_{IoU}^{*} represents the L_{IoU} of the current predicted box, and $\overline{L_{\mathrm{IoU}}}$ is the mean L_{IoU} over a set of predicted boxes.

$$L_{\mathrm{IoU}} = 1 - \mathrm{IoU} \tag{4}$$

where, IoU represents the ratio of the area of overlap between the predicted box and the ground truth box to the area of their union, as defined in Equation 5.

$$\mathrm{IoU} = \frac{w_i \cdot h_i}{w \cdot h + w_{\mathrm{gt}} \cdot h_{\mathrm{gt}} - w_i \cdot h_i} \tag{5}$$

where, w and h represent the width and height of the predicted box, w_{gt} and h_{gt} denote the width and height of the ground truth box, and w_i and h_i represent the width and height of the intersecting rectangle between the predicted box and the ground truth box, as shown in Figure 5.

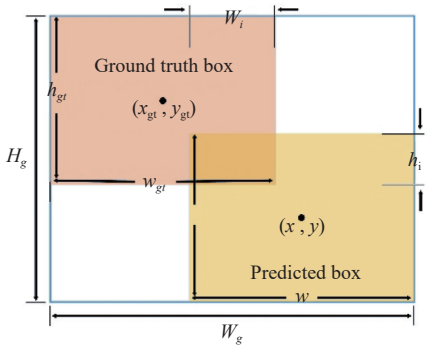


Figure 5 Schematic diagram of localization loss function

$$R_{\mathrm{WIoU}} = \exp\left(\frac{(x - x_{\mathrm{gt}})^2 + (y - y_{\mathrm{gt}})^2}{(W_g^2 + H_g^2)}\right) \tag{6}$$

where, $(x_{\mathrm{gt}}, y_{\mathrm{gt}})$ and (x, y) represent the center points of the ground truth box and the predicted box, and W_g and H_g define the width and height of the smallest enclosing box, as shown in Figure 5.

2.3.4 The classification loss: Varifocal loss

The classification loss $L_{\mathrm{Varifocal}}$ is formulated as follows:

$$L_{\mathrm{cls}} = L_{\mathrm{Varifocal}} = \begin{cases} -q \cdot (q \cdot \log p + (1 - q) \cdot \log(1 - p)), & q > 0 \\ \eta \cdot p^{\lambda} \cdot \log(1 - p), & q = 0 \end{cases} \tag{7}$$

where, p represents the predicted IoU-Aware classification score, q defines the target score, η represents the sample balance coefficient, and λ denotes decay factor.

By adjusting the decay factor λ, the overall contribution of negative samples to the loss is reduced.

2.4 Evaluation metric

In this paper, various evaluation metrics were employed to assess the detection performance of the model, including P (Precision), R (Recall), F1_score, AP (Average Precision), and mAP (mean Average Precision).

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{8}$$

where, TP (True Positive) represents the number of samples correctly predicted as positive, FP (False Positive) represents the number of samples incorrectly predicted as positive, and P represents the proportion of true positive predictions among all predicted positives.

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{9}$$

where, FN (False Negative) represents the number of samples incorrectly predicted as negative but that are actually positive, and R defines the proportion of true positive predictions among all actual positives.

$$\mathrm{F1\_score} = \frac{2 \cdot P \cdot R}{P + R} \tag{10}$$

where, F1_Score represents the harmonic mean of P and R.

$$\mathrm{AP} = \int_0^1 P(R)\mathrm{d}R \tag{11}$$

where, AP defines the average of P at different R levels.

$$\mathrm{mAP} = \frac{\sum_{1}^{n}(\mathrm{AP})}{n} \tag{12}$$

where, mAP represents the average of AP across n classes.

In addition, Grad-CAM[34] (Gradient-weighted Class Activation Mapping) heatmaps can be leveraged for visualizing deep convolutional neural network models. Here, it was utilized for facilitating the understanding of the decision-making process for the model in cattle body detection.

2.5 Experimental setup

In this paper, the experimental environment configurations are presented in the following:

The operating system was Linux Ubuntu 18.04, and the software framework was PyTorch. The hardware configuration consisted of an Intel Core i7 7800X processor, NVIDIA GeForce GTX TITAN XP Graphic Processing Unit, and 128GB of memory. Furthermore, the training epochs of all experiments were set as 100, the batch size was 16, the learning rate was set as 0.01, the momentum was 0.937, and input image resolutions for all experiments were 640×640 pixels.

3 Experimental results and analysis

To thoroughly and precisely evaluate the effectiveness of the proposed model YOLOv5-VF-W3, a comprehensive series of experiments have been designed and executed. The subsequent sections detail the specific experimental results and their corresponding analyses.

3.1 Validating the effectiveness of the YOLOv5-VF-W3 model

To comprehensively evaluate the performance of the YOLOv5-VF-W3 model, comparative experiments were performed with multiple state-of-the-art detection architectures, including both single-stage detectors (YOLOv2[23], YOLOv3[16], YOLOv4[17], YOLOv5[31], and SSD[35]) and the two-stage detector Faster-RCNN[24]. The comparative results are summarized in Table 1.

The comparison results presented in Table 1 demonstrate the superior performance of the YOLOv5-VF-W3 model across various evaluation metrics. Specifically, YOLOv5-VF-W3 achieved a precision of 95.0%, surpassing SSD (2.93%), Faster-RCNN (26.33%), YOLOv2 (20.25%), YOLOv3 (6.74%), YOLOv4 (7.95%), and baseline YOLOv5 (1.39%). The model obtained a recall of 90.7%, exceeding SSD (7.72%), YOLOv2 (17.79%), YOLOv3 (6.71%), YOLOv4 (4.25%), and YOLOv5 (1.91%). YOLOv5-VF-W3 achieved an F1 score of 92.8%, showing

improvements over SSD (5.45%), Faster-RCNN (11.94%), YOLOv2 (18.97%), YOLOv3 (6.67%), YOLOv4 (6.67%), and YOLOv5 (1.64%). With an mAP of 95.2%, the YOLOv5-VF-W3 model outperformed the SSD, Faster-RCNN, YOLOv2, YOLOv3, YOLOv4, and YOLOv5 models by margins of 3.25%, 6.61%, 14.56%, 4.27%, 3.14%, and 1.17%, respectively. These consistent performance advantages across precision, recall, F1-score, and mAP metrics collectively validate the enhanced detection capabilities of this YOLOv5-VF-W3 architecture compared to conventional single-stage and two-stage detectors.

Table 1 Comparative analysis of object detection model performance

Model	Precision/%	Recall/%	F1 Score/%	mAP/%
SSD	92.3	84.2	88.0	92.2
Faster-RCNN	75.2	**92.3**	82.9	89.3
YOLOv2	79.0	77.0	78.0	83.1
YOLOv3	89.0	85.0	87.0	91.3
YOLOv4	88.0	87.0	87.0	92.3
YOLOv5	93.7	89.0	91.3	94.1
YOLOv5-VF-W3	**95.0**	90.7	**92.8**	**95.2**

It is worth noting that Faster-RCNN, being a two-stage detector, produces high-quality candidate regions via its Region Proposal Network (RPN), which is finely tuned to boost recall. Consequently, in this experiment, YOLOv5-VF-W3's recall is slightly lower than that of Faster-RCNN. Nevertheless, YOLOv5-VF-W3 outperforms the Faster-RCNN model in terms of mAP, precision, and F1 score. Overall, the YOLOv5-VF-W3 model demonstrates superior detection performance when compared to the other models.

These results validate the effectiveness of the proposed improvements in enhancing the performance of cattle body detection.

3.2 Comparison of different models on key body parts of cattle

To thoroughly verify the efficacy of the YOLOv5-VF-W3 model in detecting key parts of the cattle body, this study performed an in-depth analysis of its performance on individual cattle as well as specific body parts, namely the legs and head. By conducting a comparative analysis, this study identified the specific differences in performance among different models and summarized these results in Table 2.

Table 2 Comparison of different models based on mAP across cattle body detection

Model	Cattle/%	Leg/%	Head/%
SSD	88.2	92.7	95.4
Faster-RCNN	89.3	88.1	90.5
YOLOv2	88.1	72.0	89.3
YOLOv3	92.5	89.2	92.2
YOLOv4	93.4	90.7	92.7
YOLOv5	94.8	93.4	94.0
YOLOv5-VF-W3	**95.3**	**94.8**	**95.4**

Samples 1-3 in the single-sample and double-sample detection examples are free from occlusion and overlap, while samples 4-7 in the multi-sample detection example exhibit varying degrees of occlusion and overlap.

Table 2 illustrates that the YOLOv5-VF-W3 model exhibits better performance in detecting individual cattle and key body parts across most evaluation metrics. Specifically, the YOLOv5-VF-W3 model achieved an impressive mAP of 95.3% in individual cattle detection, outperforming the SSD, Faster-RCNN, YOLOv2, YOLOv3, YOLOv4, and YOLOv5 models by 8.05%, 6.72%, 8.17%, 3.03%, 2.03%, and 0.53%, respectively. Furthermore, in leg object detection, the YOLOv5-VF-W3 model attained an mAP of 94.8%, surpassing the aforementioned models by 2.27%, 7.60%, 31.67%, 6.28%, 4.52%, and 1.50%, respectively. Additionally, the YOLOv5-VF-W3 model excelled in head object detection with an mAP of 95.4%, outperforming the Faster-RCNN, YOLOv2, YOLOv3, YOLOv4, and YOLOv5 models by 5.41%, 6.83%, 3.47%, 3.58%, 2.91%, and 1.49%, respectively. Table 2 indicates that the YOLOv5-VF-W3 model outperforms all other models on the mAP metric across various body parts.

To assess the detection performance of the model across different complex scenarios, comparative experiments were carried out using single-sample and double-sample data without occlusion, as well as multi-sample data with occlusion. The qualitative detection results are presented in Figure 6.

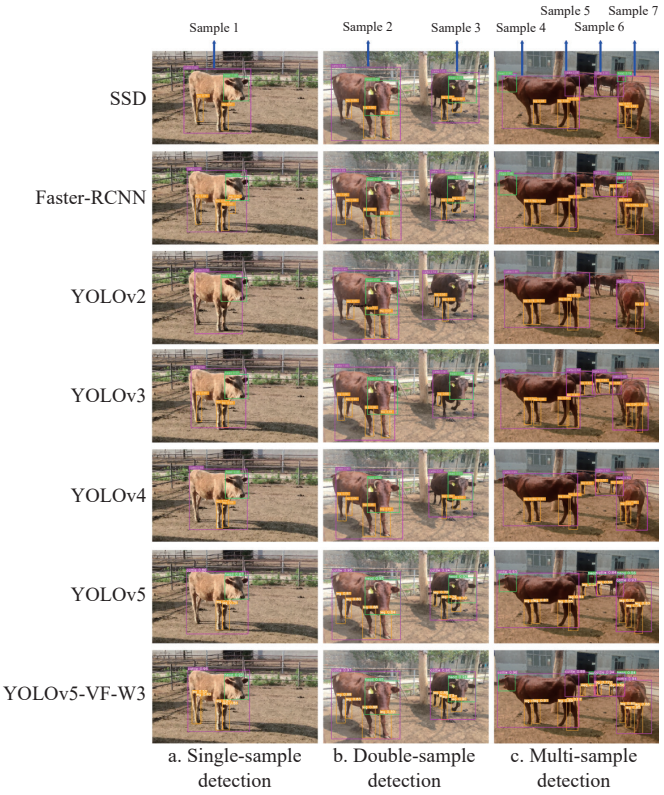


Figure 6 Qualitative comparison of detection performance across different models

Furthermore, to enhance the clarity of missed cattle body detection counts originally presented in Figure 6, this study provides explicit quantitative results through comparative visualization in Figure 7.

As illustrated in Figure 6 and Figure 7, the YOLOv5-VF-W3 model achieves better detection performance. Additionally, it obtains the lowest missed detection number for cattle bodies, highlighting its outstanding detection performance. This further confirms the effectiveness of the YOLOv5-VF-W3 model.

The superior performance of YOLOv5-VF-W3 arises from the synergistic incorporation of two critical components: the Varifocal loss function and the WIoUv3 loss function. The Varifocal loss function dynamically adjusts the weights of positive and negative samples, enabling the model to focus more on challenging samples,

such as small or partially occluded cattle targets, during training. Meanwhile, WIoUv3 focuses on anchor boxes with ordinary quality, enhancing the localization accuracy of the model in complex scenarios involving occlusion. Together, these improvements enable YOLOv5-VF-W3 to achieve better detection performance in the challenging task of cattle body detection.

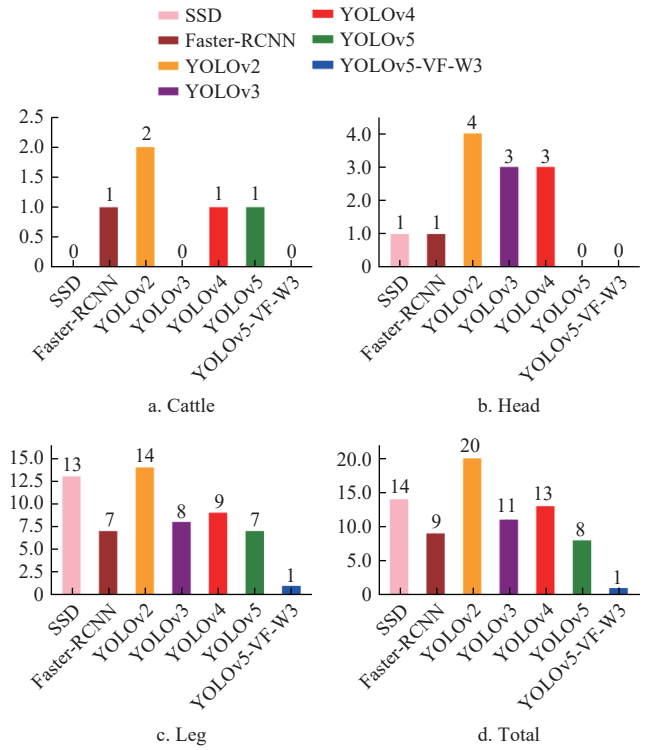


Figure 7 Comparison of missed detection number for cattle body across different models

3.3 Comparison of models with different localization loss functions

To assess the performance of the localization loss function WIoUv3, this study compared models incorporating different alternative formulations (including CIoU, WIoUv1, and WIoUv2), with results reported in Table 3.

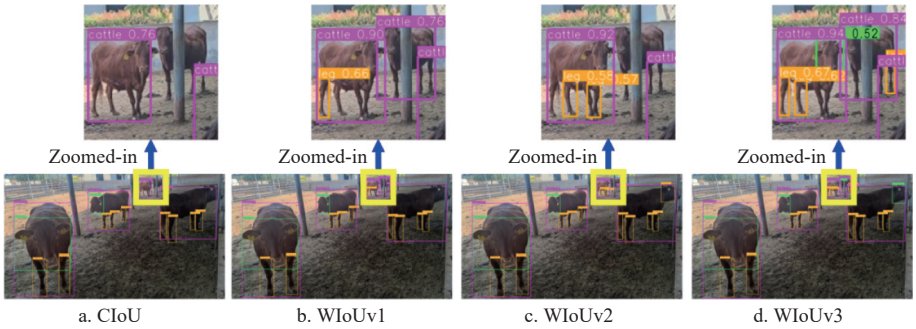
Table 3 showcases that the model employing the WIoUv3 loss function demonstrates superior performance in detecting individual cattle and their key body parts. Specifically, the model attained an mAP of 95.0% for all object detections, marking improvements of 0.96%, 0.11%, and 0.11% over models using CIoU, WIoUv1, and WIoUv2 loss functions, respectively. In the detection of individual cattle, the mAP reached 95.1%, representing enhancements of 0.32%, 0.32%, and 0.21% compared to models utilizing CIoU, WIoUv1, and WIoUv2 loss functions, respectively. For the detection of cattle legs, the mAP was 94.7%, showing improvements of 1.39%, 0.21%, and 0.32% over models employing CIoU, WIoUv1, and WIoUv2 loss functions, respectively. In the detection of cattle heads, the mAP reached 95.4%, indicating an improvement of 1.49% compared to the model using the CIoU loss function.

Table 3 Comparative performance of localization loss functions for cattle key body parts

Loss	All/%	Cattle/%	Leg/%	Head/%
CIoU	94.1	94.8	93.4	94.0
WIoUv1	94.9	94.8	94.5	95.4
WIoUv2	94.9	94.9	94.4	95.4
WIoUv3	**95.0**	**95.1**	**94.7**	**95.4**

To showcase the performance enhancements from the WIoUv3 localization loss function, Figure 8 provides a detailed visual comparison of detection results across various loss formulations. Notably, challenging scenarios involving distant and occluded instances are emphasized through zoomed-in insets, positioned above the corresponding subfigures for easier cross-method evaluation.

As evidenced in Figure 8, the model employing the WIoUv3 localization loss achieves optimal detection performance. Specifically, in challenging scenarios involving distant and occluded instances, the magnified subfigures reveal distinct advantages of the WIoUv3-enhanced model. It demonstrates superior detection quantity, improved localization accuracy, and higher confidence scores compared to alternative loss functions.



Note: a. CIoU-based model, b. WIoU1-enhanced model, c. WIoU2-optimized model, and d. WIoU3-enhanced model.

Figure 8 Qualitative performance comparison of models with diverse localization loss functions

The WIoUv3-optimized model achieves superior detection performance in challenging scenarios, particularly for distant and occluded instances, due to three key enhancements over CIoU, WIoUv1, and WIoUv2. Firstly, its dynamic gradient modulation mechanism enhances learning for partially visible and distant targets. Secondly, the curvature-sensitive formulation improves boundary localization for occluded objects. Thirdly, the spatial attention mechanism maintains stable gradient propagation for small targets. These advancements enable precise detection across varying scales, from cattle bodies to leg features.

3.4 Comparison of models with different classification loss functions

To verify the effectiveness of the Varifocal loss function, models utilizing various alternative formulations were compared, namely Cross-Entropy loss (abbreviated as CEL) and Focal loss (abbreviated as FL). The results of this comparison are presented in

Table 4. For brevity, the Varifocal loss is referred to as VFL.

Table 4 reports that the model employing the VFL loss function demonstrates optimal performance in detecting individual cattle and their key body parts. Concretely, the model with VFL attained an mAP of 94.8% for all object detections, marking improvements of 0.21% and 0.74% over models using FL and CEL loss functions, respectively. In the detection of individual cattle, the mAP reached 95.0%, representing enhancements of 0.21% and 0.21% compared to models utilizing FL and CEL loss functions, respectively. For the detection of cattle legs, the mAP was 94.6%, showing improvements of 0.11% and 1.28% over models employing FL and CEL loss functions, respectively. In the detection of cattle heads, the mAP reached 94.8%, indicating improvements of 0.21% and 0.85% compared to the models using FL and CEL loss functions.

To illustrate the performance improvements obtained by the VFL classification loss function, Figure 9 offers a thorough visual comparison of detection results achieved using various classification loss formulations. Specifically, challenging scenarios featuring distant and occluded instances are emphasized by including zoomed-in insets, which are displayed above the corresponding subfigures to enable direct comparisons across different models.

Table 4 Comparative performance of different loss functions for cattle key body parts

Loss	All/%	Cattle/%	Leg/%	Head/%
CEL	94.1	94.8	93.4	94.0
FL	94.6	94.8	94.5	94.6
VFL	**94.8**	**95.0**	**94.6**	**94.8**

a. CEL b. FL c. VFL

Figure 9 Comparative performance of different loss functions for cattle key body parts

The Varifocal loss-optimized model demonstrates superior classification performance in challenging scenarios due to several reasons. Firstly, its asymmetric weighting mechanism emphasizes positive samples while suppressing excessive negative gradients, effectively handling partial visibility and low-confidence distant targets. Secondly, the continuous IoU-aware score prediction maintains better calibration between classification and localization tasks.

3.5 Comparison of models with different loss function combinations

To validate the effectiveness of the proposed localization (WIoUv3) and classification (VFL) losses in cattle key body detection tasks, this study conducted comparative experiments with models configured with different loss combinations. Specifically, this study evaluated four model variants: 1) baseline without either proposed loss, 2) model with only the new localization loss, 3) model with only the new classification loss, and 4) model incorporating both proposed losses. The comparative results are systematically presented in Tables 5 and 6.

Table 5 Comparison of models with different loss function combinations

Index	WIoUv3	Varifocal	Precision/%	Recall/%	F1 Score/%	mAP/%
1	×	×	93.7	89.0	91.3	94.1
2	√	×	94.8	90.7	92.8	95.0
3	×	√	94.9	90.3	92.5	94.8
4	√	√	**95.0**	**90.7**	**92.8**	**95.2**

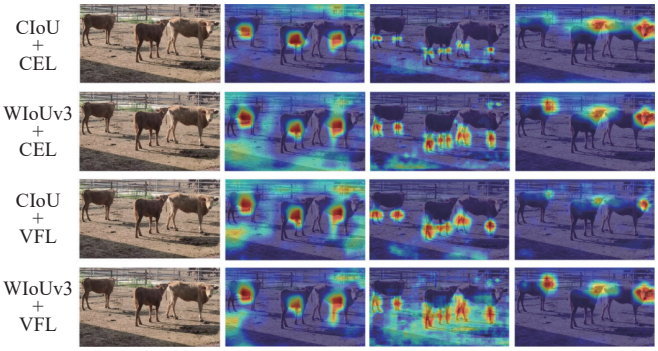
Tables 5 and 6 demonstrate that models incorporating the new loss functions (WIoUv3, VFL) consistently outperform their counterparts. Specifically, models with either the new localization or classification loss achieve superior performance compared to the baseline without these components, while models combining both proposed losses exhibit further performance enhancements. These experimental results validate the superiority of the proposed loss functions in cattle detection tasks. To further validate the effectiveness of the proposed loss functions, a comparative visualization of Grad-CAM heatmaps across models with different loss combinations was conducted, as presented in Figure 10. The results demonstrate that the model incorporating both the new localization and classification losses exhibits the most human-perception-aligned attention distribution among all evaluated configurations.

Table 6 Performance metrics (mAP) for cattle body across different loss function combination models

Index	WIoUv3	Varifocal	All/%	Cattle/%	Leg/%	Head/%
1	×	×	94.1	94.8	93.4	94.0
2	√	×	95.0	95.1	94.7	95.4
3	×	√	94.8	95.0	94.6	94.8
4	√	√	**95.2**	**95.3**	**94.8**	**95.4**

Specifically, from Figure 10, the following visualization results can be observed. For individual cattle detection, while all models demonstrate competent attention distribution, the VFL and WIoUv3-enhanced model achieves more centralized and focused attention, facilitating superior detection accuracy.

In cattle leg detection, baseline models without VFL and WIoUv3 losses primarily focus on knee regions, while models incorporating either VFL or WIoUv3 demonstrate extended attention coverage across entire leg areas. Notably, WIoUv3-enhanced models maintain precise attention localization even for small-scale leg targets in distant image regions, as evidenced in the top-right corner of the figure.



Note: The areas with deeper colors indicate that the model pays more attention during decision making.

Figure 10 The Grad-CAM heatmaps of the models using different combinations of loss functions

For cattle head detection, WIoUv3-enhanced models demonstrate significantly improved attention localization on the leftmost head region.

4 Conclusions

This paper develops a novel precise cattle key body detection solution, YOLOv5-VF-W3, which significantly enhances the efficiency and quality of animal husbandry production. By incorporating the Varifocal loss, the model effectively addresses the issue of imbalanced samples and focuses more on difficult-to-recognize instances. Furthermore, the introduction of the WIoUv3 loss function provides a wise gradient gain allocation strategy, reducing the competitiveness of high-quality anchor boxes while effectively mitigating the adverse gradients stemming from low-quality anchor boxes, thereby emphasizing anchor boxes of ordinary quality. Experimental results demonstrate that the YOLOv5-VF-W3 model achieves superior cattle body detection results, with an mAP of 95.2%. Specifically, the model excels in individual cattle detection, leg detection, and head detection, reaching accuracies of 95.3%, 94.8%, and 95.4%, respectively. Moreover, the model performs accurately and efficiently in complex detection scenarios, especially when dealing with small targets and occlusions.

In future research, various data augmentation techniques will be explored to further enhance the robustness of the model. Additionally, comparative experiments with different loss functions are planned to more precisely control the contributions of samples and anchor boxes with varying qualities, with the aim of achieving better performance. Through these efforts, the advancement of precision livestock farming is expected to be driven, providing more effective and intelligent technological support for agricultural production.

Acknowledgements

This work was supported by the Shanxi Province Basic Research Program (Grant No. 202203021212444); the GuangHe Fund D (Grant No. ghfund202407042032); Shanxi Agricultural University Science and Technology Innovation Enhancement Project (Grant No. CXGC2023045); Shanxi Postgraduate Education and Teaching Reform Project Fund (Grant No. 2022YJJG094); Shanxi Agricultural University Doctoral Research Start-up Project (Grant No. 2021BQ88); Shanxi Agricultural University Academic Restoration Research Project (Grant No. 2020xshf38); Young and Middle-aged Top-notch Innovative Talent Cultivation Program of the Software College, Shanxi Agricultural University (Grant No. SXAUKY2024005); and the Key Research and Development Program of Zhejiang Province under Grand (Grant No. 2024C01104, 2024001026).

[References]

[1] Guo H, Ma X D, Ma Q, Wang K, Su W, Zhu D H, et al. An interactive 3D point clouds analysis software for body measurement of livestock with similar forms of cows or pigs. Computers and Electronics in Agriculture, 2017; 138: 60–68.

[2] Mao A, Huang E D, Wang X S, Liu K. Deep learning-based animal activity recognition with wearable sensors: Overview, challenges, and future directions. Computers and Electronics in Agriculture, 2023; 211: 108043.

[3] Awad A I. From classical methods to animal biometrics: A review on cattle identification and tracking. Computers and Electronics in Agriculture, 2016; 123: 423–435.

[4] Roberts C M. Radio frequency identification (RFID). Computers and Security, 2006; 25(1): 18–26.

[5] Islam M A, Lomax S, Doughty A K, Islam M R, Thomson P C, Clark C E F, et al. Revealing the diversity in cattle behavioural response to high environmental heat using accelerometer-based ear tag sensors. Computers and Electronics in Agriculture, 2021; 191: 106511.

[6] Dutta D, Natta D, Mandal S, Ghosh N. MOOnitor: An IoT based multi-sensory intelligent device for cattle activity moni-toring. Sensors and Actuators A: Physical, 2022; 333: 113271.

[7] Hossain M E, Kabir M A, Zheng L H, Swain D L, McGrath S, Medway J. A systematic review of machine learning techniques for cattle identification: Datasets, methods and future directions. Artificial Intelligence in Agriculture, 2022; 6: 138–155.

[8] Bercovich A, Edan Y, Alchanatis V, Moallem U, Parmet Y, Honig H, et al. Development of an automatic cow body condition scoring using body shape signature and Fourier descriptors. Journal of Dairy Science, 2013; 96(12): 8047–8059.

[9] Zhao K X, Jin X, Ji J T, Wang J, Ma H, Zhu X F. Individual identification of Holstein dairy cows based on detecting and matching feature points in body images. Biosystems Engineering, 2019; 181: 128–139.

[10] Gao T. Detection and tracking cows by computer vision and image classification methods. International Journal of Security and Privacy in Pervasive Computing, 2021; 13(1): 45.

[11] Liu D, Zhao K X, He D J. Real-time target detection for moving cows based on gaussian mixture model. Transactions of the CSAM, 2016; 47(5): 288–294. (in Chinese)

[12] Kaur A, Kumar M, Jindal M K. Shi-Tomasi corner detector for cattle identification from muzzle print image pattern. Ecological Informatics, 2022; 68: 101549.

[13] Tassinari P, Bovo M, Benni S, Franzoni S, Poggi M, Mammi L M E, et al. A computer vision approach based on deep learning for the detection of dairy cows in free stall barn. Computers and Electronics in Agriculture, 2021; 182: 106030.

[14] Lodkaew T, Pasupa K, Loo C K. CowXNet: An automated cow estrus detection system. Expert Systems with Applications, 2023; 211: 118550.

[15] Xiao J X, Liu G, Wang K J, Si Y S. Cow identification in free-stall barns based on an improved Mask R-CNN and an SVM. Computers and Electronics in Agriculture, 2022; 194: 106738.

[16] Redmon J, Farhadi A. YOLOv3: An Incremental Improvement. arXiv: 1804.02767, 2018; In press. doi: 10.48550/arXiv.1804.02767.

[17] Bochkovskiy A, Wang C Y, Liao H Y M. YOLOv4: Optimal speed and accuracy of object detection. arXiv, 2020; doi: 10.48550/arXiv.2004.10934.

[18] He K, Kioxari G G, Dollar P, Girshick R. Mask R-CNN. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020; 42(2): 386–397.

[19] Shao W, Kawakami R, Yoshihashi R, You S, Kawase H, Naemura T. Cattle detection and counting in UAV images based on convolutional neural networks. International Journal of Remote Sensing, 2020; 41(1): 31–52.

[20] Weber F d L, Weber V A d M, Moraes P H d, Matsubara E T, Paiva D M B, Gomes M d N B, et al. Counting cattle in UAV images using convolutional neural networks. Remote Sensing Applications: Society and Environment, 2023; 29: 100900.

[21] Andrew W, Gao J, Mullan S, Campbell N W, Dowsey A W, Burghardt T,

et al. Visual identification of individual Hol-stein-Friesian cattle via deep metric learning. Computers and Electronics in Agriculture, 2021; 185: 106133.

[22] Xu B, Wang W, Falzon G, Kwan P, Schneider D. Automated cattle counting using Mask R-CNN in quadcopter vision system. Computers and Electronics in Agriculture, 2020; 171: 105300.

[23] Redmon J, Farhadi A. YOLO9000: Better, Faster, Stronger. In proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA: IEEE, 2017; pp.726–727. doi: 10.1109/CVPR.2017.690.

[24] Ren S, He K, Girshick R, Sun J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017; 39(6): 1137–1149.

[25] Soares V H A, Ponti M A, Gonçalves R A, Campello R J G B. Cattle counting in the wild with geolocated aerial images in large pasture areas. Computers and Electronics in Agriculture, 2021; 189: 106354.

[26] Shang C, Wu F, Wang M L, Gao Q. Cattle behavior recognition based on feature fusion under a dual attention mechanism. Journal of Visual Communication and Image Representation, 2022; 85: 103524.

[27] Wu D H, Wu Q, Yin X Q, Jiang B, Wang H, He D J, Song H b. Lameness detection of dairy cows based on the YOLOv3 deep learning algorithm and a relative step size characteristic vector. Biosystems Engineering, 2020; 189: 150–163.

[28] Beauchemin K A. Invited review: Current perspectives on eating and rumination activity in dairy cows. Journal of Dairy Science, 2018; 101: 4762–4784.

[29] Wang Z, Meng F S, Liu S Q, Zhang Y, Zheng Z Q, Gong C L, et al. Cattle face recognition based on a Two-Branch convolutional neural network. Computers and Electronics in Agriculture, 2022; 196: 106871.

[30] Shen W Z, Hu H Q, Dai B S, Wei X L, Sun J, et al. Individual identification of dairy cows based on convolutional neural networks. Multimedia Tools and Applications, 2020; 75: 14711–14724.

[31] Gladstone J. YOLOv5: Better speed and accuracy. arXiv, 2021; In press

[32] Zhang H Y, Wang Y, Dayoub F, Sunderhauf N. VarifocalNet: An IoU-aware dense object detector. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, TN, USA: IEEE, 2020; pp.8510-8519.

[33] Tong Z J, Chen Y H, Xu Z W, Yu R. Wise-IoU: Bounding box regression loss with dynamic focusing mechanism. arXiv, 2023; In press.

[34] Selvaraju R R, Cogswell M, Das A, Vedantam R, Parikh D, Batra D, et al. Grad-CAM: Visual explanations from deep networks via Gradient-Based localization. IEEE International Conference on Computer Vision, 2019; 128: 336–359.

[35] Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C Y, et al. SSD: Single shot MultiBox detector. In: Computer Vision - ECCV 2016, Springer, 2015. doi: 10.1007/978-3-319-46448-0_2