applied sciences

MDPI

Article

Enhancing Livestock Detection: An Efficient Model Based on YOLOv8

Chengwu Fang, Chunmei Li *, Peng Yang, Shasha Kong, Yaosheng Han, Xiangjie Huang and Jiajun Niu

College of Computer Technology and Applications, Qinghai University, Xining 810016, China; chengwu_fang@126.com (C.F.); ys220854100375@qhu.edu.cn (P.Y.); ys220854100367@qhu.edu.cn (S.K.); hanyao_sheng@163.com (Y.H.); 17673452204@163.com (X.H.); jedino1003@outlook.com (J.N.)
* Correspondence: li_chm@qhu.edu.cn

Abstract: Maintaining a harmonious balance between grassland ecology and local economic development necessitates effective management of livestock resources. Traditional approaches have proven inefficient, highlighting an urgent need for intelligent solutions. Accurate identification of livestock targets is pivotal for precise livestock farming management. However, the You Only Look Once version 8 (YOLOv8) model exhibits limitations in accuracy when confronted with complex backgrounds and densely clustered targets. To address these challenges, this study proposes an optimized CCS-YOLOv8 (Comprehensive Contextual Sensing YOLOv8) model. First, we curated a comprehensive livestock detection dataset encompassing the Qinghai region. Second, the YOLOv8n model underwent three key enhancements: (1) incorporating a Convolutional Block Attention Module (CBAM) to accentuate salient image information, thereby boosting feature representational power; (2) integrating a Content-Aware ReAssembly of FEatures (CARAFE) operator to mitigate irrelevant interference, improving the integrity and accuracy of feature extraction; and (3) introducing a dedicated small object detection layer to capture finer livestock details, enhancing the recognition of smaller targets. Experimental results on our dataset demonstrate the CCS-YOLOv8 model's superior performance, achieving 84.1% precision, 82.2% recall, 84.4% mAP@0.5, 60.3% mAP@0.75, 53.6% mAP@0.5:0.95, and 83.1% F1-score. These metrics reflect substantial improvements of 1.1%, 7.9%, 5.8%, 6.6%, 4.8%, and 4.7%, respectively, over the baseline model. Compared to mainstream object detection models, CCS-YOLOv8 strikes an optimal balance between accuracy and real-time processing capability. Its robustness is further validated on the VisDrone2019 dataset. The CCS-YOLOv8 model enables rapid and accurate identification of livestock age groups and species, effectively overcoming the challenges posed by complex grassland backgrounds and densely clustered targets. It offers a novel strategy for precise livestock population management and overgrazing prevention, aligning seamlessly with the demands of modern precision livestock farming. Moreover, it promotes local environmental conservation and fosters sustainable development within the livestock industry.

Keywords: precision livestock farming; CCS-YOLOv8; object detection

Citation: Fang, C.; Li, C.; Yang, P.; Kong, S.; Han, Y.; Huang, X.; Niu, J. Enhancing Livestock Detection: An Efficient Model Based on YOLOv8. *Appl. Sci.* **2024**, *14*, 4809. https://doi.org/10.3390/app14114809

Academic Editor: Andrea Prati

Received: 18 April 2024
Revised: 23 May 2024
Accepted: 30 May 2024
Published: 2 June 2024

1. Introduction

Grassland ecosystems are vital to terrestrial ecosystems because they offer a range of uncommon wild plants and animals, unique habitats, and basic means of production. They are also a vital resource that pastoral people rely on to survive [1,2]. Many ecological processes, including soil and water conservation, wind and sand control, air purification, environmental beautification, and biodiversity protection, are carried out by grasslands, which serve as an essential ecological barrier. However, several causes, such as climate change and human activities, have made the deterioration of grasslands worse recently. This hurts the growth of grassland livestock farming as well as the ecosystems' ability to perform their service role [3,4]. Overgrazing is one of the primary anthropogenic factors contributing to grassland degradation. Improving grazing management is crucial for

maintaining the multifunctionality of grassland ecosystems [5]. Controlling overgrazing necessitates effective management of herders' livestock numbers to maintain the pasture's healthy condition and preclude further grassland degradation. At the moment, the primary technique for determining grassland stocking capacity is manual field surveys carried out by specialists in plant protection and grassland management [6]. The typical manual survey, however, has numerous drawbacks, including low representativeness, poor timeliness, and high subjectivity. It is also time-consuming and difficult, has restricted coverage, and is hampered by the vastness and complexity of the grassland area [7–9]. Therefore, it is crucial to advance the intelligent development and automation of livestock detection and early warning systems.

The use of various intelligent gadgets in contemporary livestock farming together with computer vision technology in livestock production has become more commonplace in recent years and has become a hotspot for livestock farming research [10–12]. The utilization of these intelligent devices has markedly enhanced livestock-raising accuracy and efficiency, reduced dependence on labor and resources, lowered livestock-raising costs, and fostered sustainable growth in the livestock industry. Computer vision technology can be employed to analyze livestock through images or videos for disease detection, behavior monitoring, and individual identification. Among these applications, the identification and localization of individual livestock using target detection technology constitutes an important research direction.

Currently, deep learning target recognition methods are categorized into two groups based on the number of detection stages [13]. Two-stage object detectors entail a detection process segmented into two stages. In the first stage, regions that potentially contain targets are selected from the input image, while in the second stage, targets are identified and located within these candidate regions. Representative methods encompass R-CNN [14], Fast-RCNN [15], and Faster-RCNN [16] proposed by Girshick et al. While this type of target detector typically achieves high accuracy, its substantial computational demands and intricate model structure necessitate more advanced hardware configurations and longer computation times. By comparison, one-stage models have superior real-time performance due to their ability to recognize and locate items inside the picture. While their accuracy may be somewhat lower than that of two-stage models, they can detect things much faster [17].

Representing single-stage models, the YOLO series achieves rapid and highly accurate detection through the rational design and modification of its network structure [18]. The YOLO series models have shown significant promise in livestock detection applications [19]. Du and Qi et al. enhanced the YOLOV4 model by incorporating a novel composite multi-channel attention mechanism. This innovation significantly improved the model's performance in livestock detection tasks within agricultural environments, achieving an mAP of 89.77%. However, it is important to note that the model still faces challenges with missed detection of small targets, especially when livestock are lying down or in similar positions. This issue is largely attributed to the horizontal shooting angles commonly employed in the dataset [20]. Pu and Yu et al. introduced an enhanced Chengdu horse goat detection algorithm based on TPH-YOLOv5, integrating BiFPN instead of PANet. This technique successfully identifies Chengdu ma goats in actual indoor cowshed breeding settings, laying the groundwork for precision livestock feeding according to age and gender. However, its limitation lies in its applicability being restricted to indoor breeding scenarios and its focus on a single livestock species [21]. Kurniadi and Setianingsih et al. combined YOLOv5 with UAV imagery and videography to achieve recognition and localization of free-grazing dairy cattle. However, the model's accuracy significantly decreases at higher altitudes, making it unsuitable for detecting livestock groups that are easily startled [22]. Zhang and Xuan et al. integrated a Dyhead module with the detecting head of the YOLOV7 model. They improved the model via knowledge distillation, enhancing accuracy while decreasing identification time, furthering the application of sheep facial recognition technology in real-world applications [23].

YOLO series models can accurately identify and locate animals and can distinguish them from the background. Remote sensing using UAVs can gather target imagery rapidly and comprehensively [24]. The combination of UAV-acquired image data and YOLO target detection provides a new solution for livestock detection in large-scale pasture grasslands. This study's goal is to gather and build an image collection of popular livestock species (cattle, sheep, and yak) in the Qinghai region while also promoting relevant computer vision research and applications. Building upon the original YOLOv8n model, this study proposes improvements and optimizations, resulting in the CCS-YOLOv8 model, which is aimed at enhancing the detection performance for livestock targets. The following are the particular optimization techniques:

1. A CBAM is integrated into the C2f module of YOLOv8n to analyze image information more effectively and emphasize salient features. By enhancing the representational capacity of the output feature information, the model's detection accuracy is consequently improved.
2. The lightweight up-sampling operator CARAFE is introduced to solve the shortcomings of conventional up-sampling operators, which have small receptive fields and disregard the semantic content of feature maps.
3. To mitigate the loss of small target feature information, an additional small object detection layer is incorporated into the YOLOv8n neck structure. This layer facilitates the extraction of livestock characteristics and details across multiple receptive fields.

2. Materials and Methodology

2.1. Obtaining Images and Creating Datasets

2.1.1. Data Gathering

The research region is situated in southern Qinghai Province in China's Hainan Tibetan Autonomous Prefecture. The area is primarily mountainous and is situated at the center of the Tibetan Plateau, and it experiences a typical plateau continental climate, which is part of the Tibetan Plateau climatic system. Animal species that are widely distributed and raised in the alpine grasslands of the Hainan Tibetan Autonomous Prefecture include yaks, sheep, and cows. These animals are well-adapted to the local natural habitat and climatic circumstances. Data collecting was done in September and December of 2023 to guarantee data variety. The data collection targeted yaks, cattle, and sheep using a DJI Air 3 drone (DJI, Shenzhen, China). Livestock data were gathered using drone images at various times, places, perspectives, and elevations to guarantee the images' applicability to the actual world and the model's capacity for generalization.

2.1.2. Preparing Images and Building Datasets

This study used OpenCV (version 4.8.0) to process UAV-captured livestock films, and pictures were retrieved at 90-frame intervals. After OpenCV processing and discarding highly repetitive or blurred images, a total of 8750 valid images were obtained. A subsequent step involved utilizing the LabelMe data annotation tool (version 5.3.0) to annotate the livestock in the images. To fulfill the prerequisites of the subsequent study, the various livestock species were divided into two distinct groups: juveniles (from birth to 1 year old) and adults (over 1 year old). The YOLO standard format was used to record the annotation results in TXT files. These files contained the following information: category, relative center coordinates, relative width, and relative height. Figure 1 provides an example of a few of the labeled photographs out of the 105,852 livestock instances that were labeled in total. Ultimately, a ratio of 8:1:1 was employed to partition all images into training, validation, and test sets, respectively.

Figure 1. Some annotated samples for livestock detection.

2.2. Network Architecture of CCS-YOLOv8

2.2.1. Model of YOLOv8 Network

The same design team that created YOLOv5 is also working on YOLOv8. Owing to YOLOv5's popularity, YOLOv8 adds further enhancements and functionality [25]. Figure 2 depicts the YOLOv8 network's design.

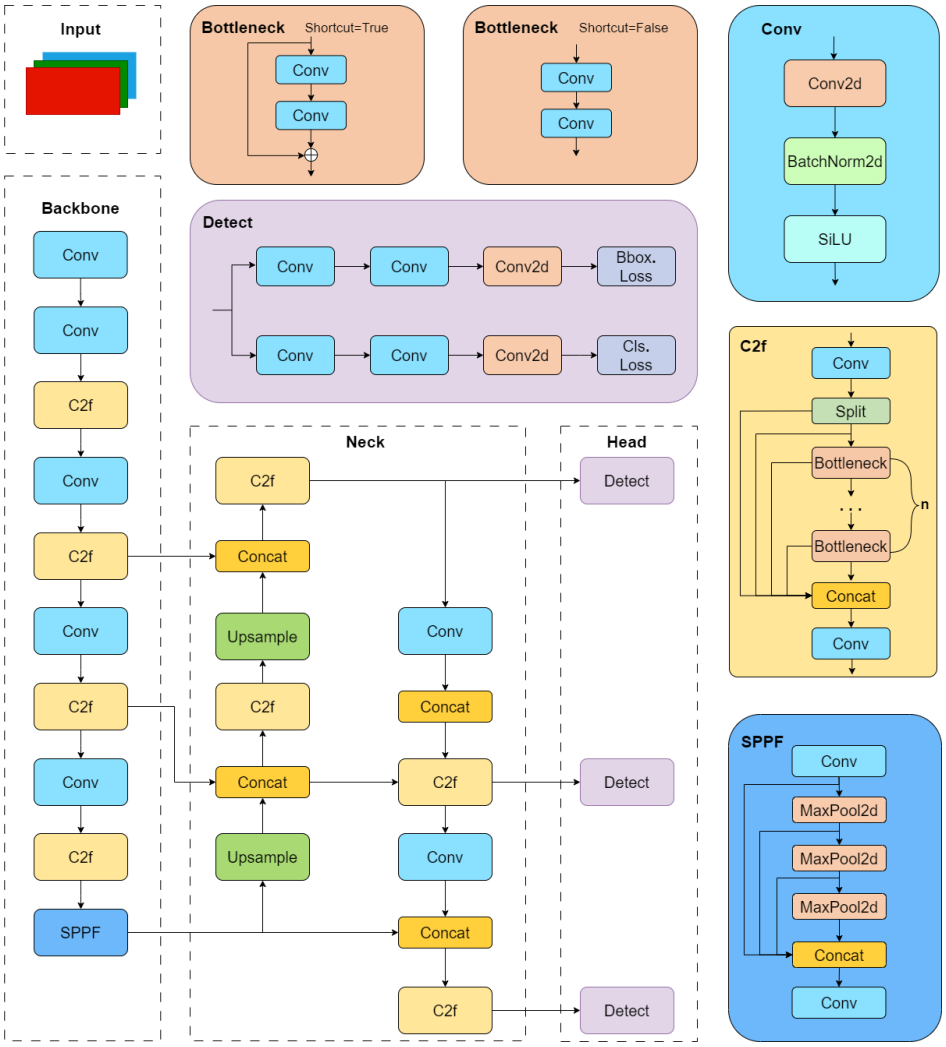

Figure 2. Architecture of YOLOv8 network.

In the preprocessing stage of the image, the input side scales the image to the 640 × 640 resolution required by YOLOv8. In addition, YOLOv8 uses data augmentation techniques such as mosaics to enhance training effectiveness during training.

The CSP architectural concept is carried over into the YOLOv8 backbone network [26]. Five down-samplings of the input characteristics allow for the acquisition of visual features at five distinct scales. In contrast to the C3 module [27] used in YOLOv5, YOLOv8 employs the C2f module. The traditional C3 module Dodule is limited to processing neighboring Bottleneck structures, resulting in extracted features that lack cross-layer information transfer. The C2f module introduces cross-layer feature transfer, eliminates the branching convolution, and adds more splitting operations by combining the architecture of the C3 and ELAN modules from YOLOv7 [28]. These modifications allow YOLOv8 to capture richer gradient flow information. The use of the SPPF module enhances YOLOv8's ability to identify small, obstructed, and fuzzy objects compared to the SPP module [29].

YOLOv8's neck network combines the FPN [30] and PAN [31] architectures to provide multi-scale picture feature fusion. In contrast to earlier iterations, YOLOv8 eliminates convolutional structures from PAN-FPN during the sampling phase, guaranteeing the model's lightweight and effective operation. Up-sampling is done from top to bottom via the FPN structure, combining low-level detail information with high-level semantic characteristics. However, target location data are lacking if FPN is used alone. The PAN structure down-samples from the top to the bottom and integrates feature maps of different levels through convolutional layers in order to precisely maintain the target's spatial position information. Through the generation of complementing semantic and positional information, the both-direction fusion of FPN and PAN improves the accuracy of target recognition in pictures of different sizes.

YOLOv8 adopts a decoupled head structure, meaning that each scale has an independent detector, and each detector is responsible for separately predicting the bounding box for that scale. The detector consists of a set of convolutional and fully connected layers, which are used to predict the bounding box for the corresponding scale. The convolutional and fully connected layers are employed for bounding box regression and target classification tasks. For the target classification task, the BCE loss is utilized, incorporating the asymmetric weighting scheme of VFL [32]. For bounding box regression, the Bbox loss function is employed, which combines the CIoU with the DFL [33]. YOLOv8 introduces an anchor-free detection head—no longer relying on anchor boxes—thus providing greater flexibility to better adapt to various target shapes and sizes.

Five models with varied sizes are available in YOLOv8 to achieve a balance between accuracy and speed in different environments. YOLOv8n is particularly well-suited for field applications due to its portability and simplicity compared to the other variants. Consequently, we selected YOLOv8n as our baseline model.

2.2.2. CBAM Attention Mechanism

By focusing on key characteristics of the detected object, the attention process may be viewed as an allocation mechanism that improves target localization and classification accuracy. The central concept is to give the raw data varying weights to identify underlying relationships and highlight important information [34]. Since the majority of the livestock targets to be recognized are small- and medium-sized, and the background makes up a sizable component of the information that is used; it is especially advantageous to incorporate an attention mechanism. To improve the model's representation of livestock features, we integrated CBAM [35] into the C2f module of YOLOv8n. CBAM is a lightweight and end-to-end attention mechanism. For each feature map, CBAM sequentially applies attention to the channel and spatial dimensions, achieving comprehensive attention in both aspects.

For the input feature map F1, CBAM initially refines it using the channel attention module to derive the intermediate feature map F2. Subsequently, F2 undergoes further refinement through the spatial attention module to produce the final feature map F3, as Figure 3 illustrates.

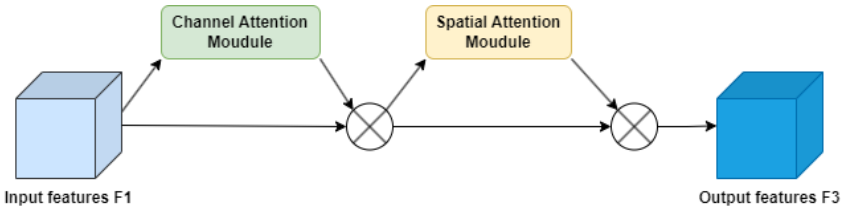

Figure 3. Overview of the CBAM.

In the channel attention module, important feature information is extracted from the input feature map through GAP and GMP. The channel attention weights are then determined using an MLP, and normalization is accomplished using the sigmoid function. Ultimately, channel-wise multiplication is used to apply the acquired weights to the original input feature maps. This process may be stated as follows:

$$\begin{aligned} \mathcal{M}_c(\mathcal{F}) &= \phi(MLP_1(AvgPool(\mathcal{F})) + MLP_2(MaxPool(\mathcal{F}))) \\ &= \phi(\mathcal{W}_3(\mathcal{W}_2(\mathcal{F}^c_{avg})) + \mathcal{W}_3(\mathcal{W}_2(\mathcal{F}^c_{max}))), \end{aligned} \tag{1}$$

The spatial attention module mainly performs GAP and GMP in the spatial dimension. This mechanism captures the correlation between spatial features through convolution while maintaining the input and output dimensions, facilitating the extraction of salient livestock features. The specific calculations are as follows:

$$\begin{aligned} \mathcal{S}_s(\mathcal{F}) &= \psi(g^{3\times3}([AvgPool(\mathcal{F}); MaxPool(\mathcal{F})])) \\ &= \psi(g^{3\times3}([\mathcal{F}^s_{avg}; \mathcal{F}^s_{max}])), \end{aligned} \tag{2}$$

The newly constructed C2fCBAM module replaces all of the Bottleneck modules with the BottleneckCBAM module, as shown in Figure 4. To increase the diversity of learned features at various network levels, this structure combines a cross-stage feature fusion technique with a truncated gradient flow approach. Consequently, it lessens the effect of redundant gradient information and enhances the network's capacity for learning. The C2fCBAM module is included in the model to improve output feature representation, which raises the model's detection accuracy and boosts algorithmic performance as a whole.

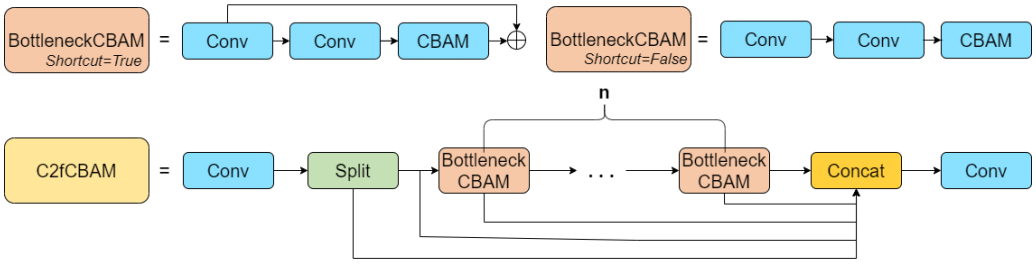

Figure 4. Details of the C2fCBAM.

2.2.3. CARAFE

Conventional up-sampling techniques often fail to incorporate semantic information, leading to the production of feature maps that may lack realism, naturalness, and fidelity to the original image content. We include the CARAFE [36] lightweight up-sampling operator into the YOLOv8n baseline model's feature fusion network. The CARAFE operator adaptively recombines the up-sampled feature maps, increasing the perceptual range based on the content and structure of the feature maps, helping the model better capture global information in the image, thereby improving the accuracy of the reconstruction and alleviating the inherent limitations of traditional up-sampling techniques.

As illustrated in Figure 5, the CARAFE operator accomplishes efficient feature up-sampling and reorganization through two principal steps: First, the up-sampling kernel

prediction module determines the attention weight of each up-sampled site based on the mapping relationship between the up-sampled sites and the down-sampled feature map. The input feature map is rearranged in the second stage by the content-aware reorganization module using the resulting up-sampling kernel to maintain contextual information and spatial details while enhancing multi-scale target identification performance.

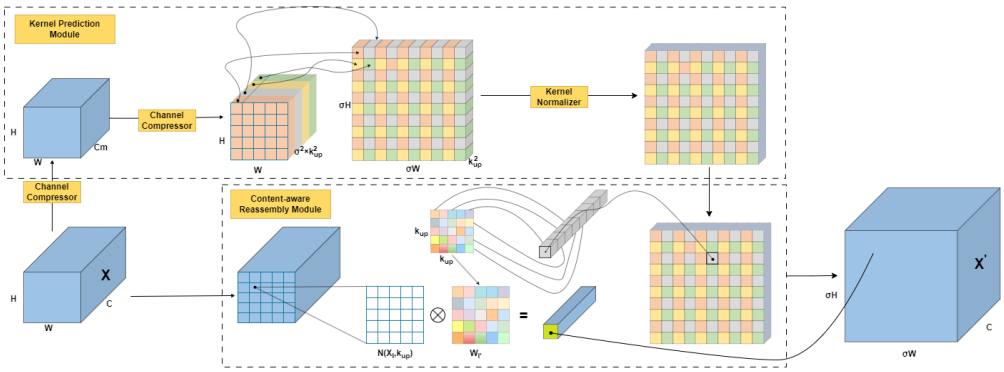

Figure 5. The general structure of CARAFE. The up-sampling factor in the figure is 2.

Define $N(\mathcal{X}_l, k)$ as the $k \times k$ neighborhood region of $\mathcal{X}$ centered at position l. The kernel prediction module predicts a location kernel $\mathcal{W}_{l'}$ based on $N(\mathcal{X}_l, kup)$, represented as:

$$\mathcal{W}_{l'} = \psi(N(\mathcal{X}_l, k_{up})), \tag{3}$$

where ψ is the kernel prediction module.

Subsequently, the content-aware reorganization module recombines the neighborhood region $N(X_l, k_{up})$ of $\mathcal{X}_l$ with the predicted kernel $\mathcal{W}_{l'}$ to obtain the up-sampled target feature $\mathcal{X}'_{l'}$:

$$\mathcal{X}'_{l'} = \phi(N(X_l, k_{up}), \mathcal{W}_{l'}), \tag{4}$$

where ϕ is the content-aware reorganization module.

The CARAFE operator offers multiple advantages over traditional up-sampling methods and other decomposition methods. It possesses a large sensory field, which can better capture the image's semantic information during the feature fusion process. Furthermore, the CARAFE operator provides a potent tool for multi-scale image processing, which is crucial for enhancing the precision and efficiency of the target detection task. It does this by amplifying the multi-scale target detection effect following multi-level feature fusion without adding excessive parameters or calculations. Experiments have demonstrated that the CARAFE operator can reduce unnecessary interference information and works effectively with datasets that contain background noise.

2.2.4. Small Object Detection Layer

The YOLOv8 network uses a three-layer proportional feature map architecture in its neck structure. Three distinct feature maps can be produced following the fusing of features. Smaller receptive fields and greater information regarding target positions and nearby features are characteristics of larger-scale feature maps, which make them appropriate for small target detection. Smaller-scale feature maps, on the other hand, are better at recognizing huge objects because they contain richer semantic information and broader receptive fields, but they also lack clear local details. The YOLOv8 network may not fully meet the requirements for detecting young animals due to its limited 80×80 maximum feature map size.

To improve the recognition performance of young and small target livestock, we added a small object detection layer to the YOLOv8n network during the neck feature fusion step. With this improvement, the YOLOv8n network can identify objects in a narrower receptive field, which improves its capacity to capture the distinctive characteristics and intricate

details of young livestock. YOLOv8n takes advantage of the four distinct detection layer scales for feature fusion, which allows the system to efficiently utilize semantic information and fine-grained details at every level, which reduces the possibility of misidentifying or omitting small targeted livestock and promotes more precise livestock identification and localization.

2.2.5. CCS-YOLOv8 Algorithm

As seen in Figure 6, we propose an enhanced model called CCS-YOLOv8, which incorporates the CBAM attention mechanism, the CARAFE operator, and a small object detection layer. By integrating these elements, the CCS-YOLOv8 model enhances feature expression capabilities, reduces unnecessary interference, and effectively highlights important elements in the images. The detection performance of the improved CCS-YOLOv8 is significantly improved compared to before, even when dealing with small and distant objects.

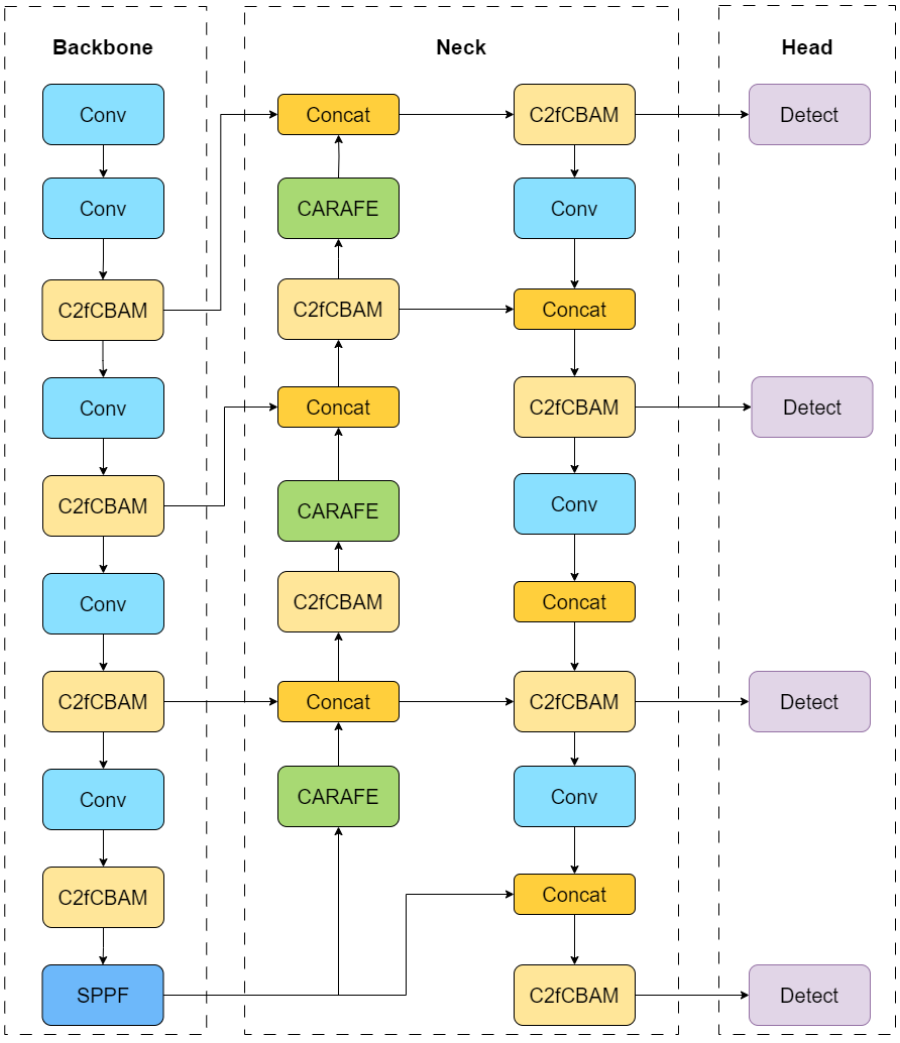

Figure 6. The improved CCS-YOLOv8 model.

2.3. Experimental Design

All experiments were conducted on the Ubuntu 20.04 operating system with two NVIDIA RTX 1080TI GPUs. The GPUs were accelerated using CUDAv11.7 and CUDNNv8.5, and training was based on the deep learning framework Python 2.0.0. The training parameter settings are shown in Table 1.

Table 1. Training parameters settings.

Parameter	Setting	Parameter	Setting
optimizer	SGD	epochs	300
momentum	0.937	batch	16
seed	0	workers	8
imgsz	640	close_mosaic	10
lr0	0.01	lr1	0.01

2.4. Indicators for Model Evaluation

To evaluate the performance of the model at different IOU thresholds, we utilize several metrics, including precision, recall, and mAP, at different IOU thresholds, such as 0.5, 0.75, and across the range of 0.5 to 0.95, as well as calculate the F1-score. These metrics provide insights into the accuracy, completeness, and overall performance of the model in detecting and localizing objects. Simultaneously, we also consider the number of parameters and the amount of computation (FLOPs) of the model as auxiliary indices to evaluate model complexity when comparing it with other mainstream models.

Precision is computed by dividing the number of accurately recognized livestock instances by the total number of livestock instances detected by the model. It represents the ratio of correctly identified instances out of all the instances predicted as livestock. Recall, on the other hand, is determined by dividing the total number of correctly tagged livestock instances in the dataset by the ratio of properly recognized livestock examples. It represents the proportion of correctly identified instances out of all the actual livestock instances present in the dataset. They are calculated as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{5}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{6}$$

The number of livestock cases the YOLOv8 network model successfully identified is shown by the acronym TP. The number of livestock cases that the model misidentified as livestock is shown by the acronym FP. The number of livestock cases that the model overlooked or could not identify is shown by the acronym FN.

The area under the PR curve represents AP, which is the average precision for various recall settings. The average accuracy for a class is represented by the symbol AP@0.5 when the IOU threshold for confusion matrices is set to 0.5. To clarify, a predicted bounding box is only regarded as a true positive and included in the precision computation if it surpasses an IoU of 0.5 with the matching ground truth bounding box. AP@0.75 is calculated similarly, yet with a more stringent IoU threshold of 0.75. This imposes greater demands on detection accuracy, requiring that predicted bounding boxes achieve an IoU of no less than 0.75 with the ground truth to be acknowledged as correct detections and to be counted towards the true positive tally. Elevating the IoU threshold from 0.5 to 0.75 introduces a stricter criterion for classifying a detection as accurate, thus offering a more refined estimate of the model's detection efficacy. An extensive analysis of the trade-off between overall detection precision and localization accuracy is provided via assessments across several IoU levels. The different AP values are calculated as follows:

$$\text{AP@0.5} = \int_0^1 \mathit{Precision}(\mathit{Recall})\mathrm{d}(\mathit{Recall}), \text{ where IoU} \geq 0.5 \tag{7}$$

$$\text{AP@0.75} = \int_0^1 \mathit{Precision}(\mathit{Recall})\mathrm{d}(\mathit{Recall}), \text{ where IoU} \geq 0.75 \tag{8}$$

The mAP is the mean of the AP values for all categories. Greater mAP values signify a model's elevated average detection precision across all target categories. The mAP values are calculated as follows:

$$\text{mAP} = \frac{1}{Q}\sum_{q=1}^{Q}\text{AP}(q) \tag{9}$$

$$\text{mAP@0.5:0.95} = \frac{1}{10}\sum_{t=0.5}^{0.95}\text{mAP@t} \tag{10}$$

An important metric for evaluating the efficacy of a binary classification model is the F1-score. It is an average of precision and recall that is harmonized and takes both into account. The range of values for the F1-score is 0 to 1. It is calculated as follows:

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{11}$$

3. Experimental Results and Analysis

3.1. Improved CCS-YOLOv8 Model

3.1.1. Changes in Losses

Figure 7 shows the variation curves of box_loss, dfl_loss, and cls_loss of the proposed CCS-YOLOv8 model on the training and validation sets of common livestock detection datasets in the Qinghai region, which are used to validate the convergence performance of the CCS-YOLOv8 model.

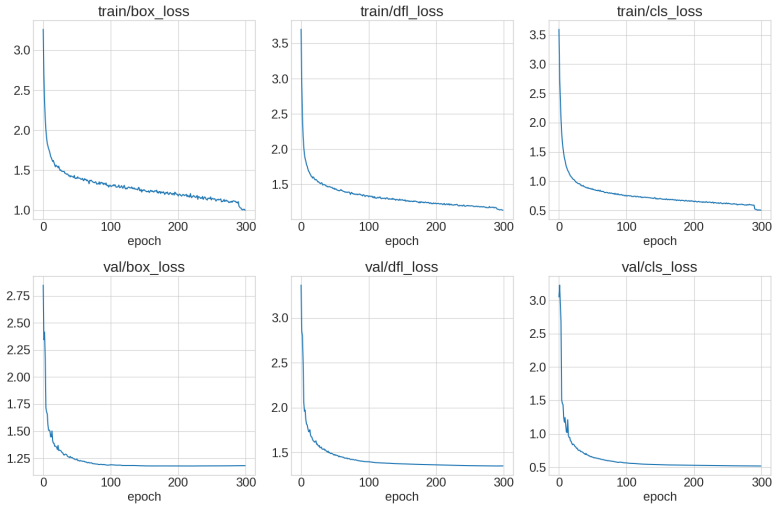

Figure 7. Convergence curve of loss function of CCS-YOLOv8.

At first, the loss values drop off quickly, showing that the model is learning and adapting quickly. Nevertheless, the loss numbers tend to level out and vary within a small range as the exercise continues. This suggests that the optimization process has converged and the model has achieved a reasonably stable state. The loss values' oscillations imply that the model is adjusting its parameters to strike the best possible balance between generalization and precision. Overall, there are no overfitting or underfitting problems with the CCS-YOLOv8 model's generalization capabilities in the livestock target identification test.

3.1.2. Changes in Performance

To evaluate the performance increase for the livestock target recognition task, we used a set of test images as examples and performed target identification on them using the YOLOv8n baseline model and the upgraded CCS-YOLOv8 model. Due to the presence of densely aggregated livestock targets in the example images, directly displaying the

classification labels and confidence values would lead to visual clutter and affect the comparison. To evaluate the detection performance of the two models more objectively, we present the detection results with only the target detection boxes retained while omitting the confidence value and category label information, as shown in Figure 8.

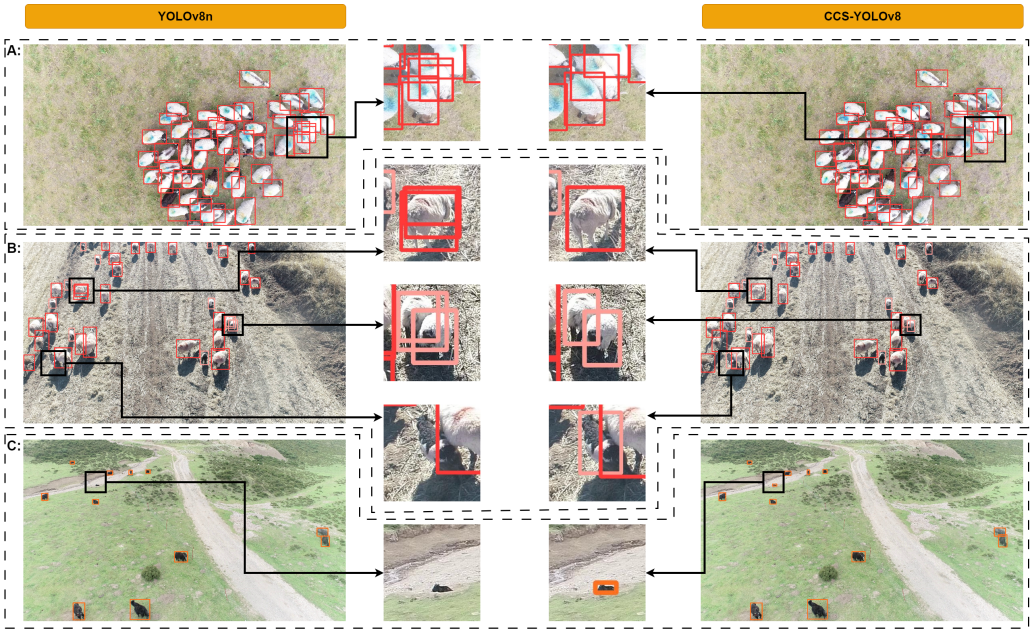

Figure 8. Exemplary detection results comparing YOLOv8n and CCS-YOLOv8 on the common livestock detection dataset from Qinghai. (**A**) The YOLOv8n model misdetects two of the sheep, counting them as four in the densely distributed group, whereas the CCS-YOLOv8 model correctly identifies these sheep targets. (**B**) The YOLOv8n model struggles with missed detections of lambs and repeated detections of sheep, while the CCS-YOLOv8 model effectively avoids these errors. (**C**) The CCS-YOLOv8 model accurately detects a small yak in the distance, a target missed by the YOLOv8n model.

- In group A images, the YOLOv8n baseline model misdetects two of the sheep as four for the densely distributed sheep in the overhead view, while the CCS-YOLOv8 model correctly recognizes these sheep targets. Compared to YOLOv8n, CCS-YOLOv8 demonstrates better detection performance when detecting densely distributed sheep targets with occlusion.
- In group B images, the YOLOv8n model suffers from the problems of missed detection of lamb targets and repeated detection of adult sheep when dealing with situations where adult sheep and lambs are in the same frame, while the CCS-YOLOv8 model effectively avoids such errors. This result indicates that, compared to the baseline model, CCS-YOLOv8 has a stronger ability to detect lamb targets and significantly reduces the risk of missed and repeated detections.
- In group C images, in the oblique side view, the CCS-YOLOv8 model accurately detects a small yak in the distance, while the YOLOv8n model misses the detection. This confirms that the CCS-YOLOv8 model demonstrates better performance in detecting small targets.

3.2. Ablation Experiment

We performed eight sets of ablation tests to determine the improvement due to each module for the YOLOv8n model's livestock target detection capability. The purpose of these studies was to assess how various module integrations, either separate or in combination, affect the model's overall performance. These tests aimed to ascertain the optimization contributions of each module and their efficacy in enhancing the YOLOv8n model's livestock target identification capabilities.

Table 2 displays the ablation experiment results. After integrating the small object detection layer, the YOLOv8 model demonstrated improvements in precision, recall, mAP@0.5, mAP@0.75, mAP@0.5:0.95, and F1-score, achieving 84.9%, 77.9%, 82.7%, 59%, 52.7%, and 80.4%, respectively: marking an enhancement of 1.9%, 3.6%, 4.1%, 5.3%, 3.9%, and 2%, respectively, over the baseline YOLOv8n. This discovery implies that the YOLOv8 network's capacity to acquire livestock's finer details is improved by lowering the receptive field and adding the tiny target detection layer. As a result, the model becomes more effective at reducing missed detections and false detections of small target livestock. The addition of the small object detection layer allows the network to focus on and accurately identify fine details, enabling improved performance for detecting and localizing small livestock instances. Relative to the baseline YOLOv8n, the model's performance is enhanced after incorporating the C2fCBAM module, with metrics reaching 85.2%, 74.8%, 79.7%, 55.2%, 49.8%, and 79.4% for precision, recall, mAP@0.5, mAP@0.75, mAP@0.5:0.95, and F1-score, respectively. In particular, the precision was enhanced by 2.2% relative to the baseline. The C2fCBAM module allows for more precise focusing on the important features within the image. This module enhances the quality and diversity of feature extraction, leading to improved precision in the model's predictions. By selectively attending to relevant image regions, the C2fCBAM module helps the model capture and emphasize crucial features, resulting in more accurate and precise detection of livestock instances. After integrating the CARAFE operator, the YOLOv8 model exhibits metrics of 82.6%, 76.9%, 79.7%, 55.2%, 50%, and 79.2% in terms of precision, recall, mAP@0.5, mAP@0.75, mAP@0.5:0.95, and F1-score, respectively. The CARAFE operator demonstrates a notable enhancement in mAP, and recall improved by 2.6%. This signifies that the model can leverage contextual information surrounding the target, which is crucial for understanding the target's background environment. By incorporating contextual information, the model can better analyze the relationships and dependencies between the target and its surroundings. This contextual understanding aids with improving the accuracy of livestock detection by reducing false detections and enhancing the model's ability to discriminate between livestock and other objects or backgrounds. By considering the broader context, the model can make more informed decisions about the presence of livestock based on the surrounding visual cues. This helps to reduce false positives and improve the overall reliability of the model's predictions.

Table 2. Ablation experiment results.

Model	Precision (%)	Recall (%)	mAP@0.5 (%)	mAP@0.75 (%)	mAP@0.5:0.95 (%)	F1-Score (%)
YOLOv8n	83	74.3	78.6	53.7	48.8	78.4
+C2fCBAM	85.2	74.8	79.7	55.2	49.8	79.4
+CARAFE	82.6	76.9	79.7	55.2	50	79.2
+small object detection layer	84.9	77.9	82.7	59	52.7	80.4
+C2fCBAM+CARAFE	82.8	77.2	80.2	56.2	51	79.4
+C2fCBAM+small object detection layer	84.9	79.3	82.4	59.8	52.8	82
+CARAFE+small object detection layer	83.9	80.6	82.6	59.6	52.7	81.6
CCS-YOLOv8	84.1	82.2	84.4	60.3	53.6	83.1

The CCS-YOLOv8 model demonstrates the best detection performance. The mAP and PR curve comparison results can be found in Figure 9. Relative to the baseline model, the CCS-YOLOv8 model demonstrates significant improvements across all detection metrics. The CCS-YOLOv8 model attained 84.1% precision, 82.2% recall, 84.4% mAP@0.5, 60.3% mAP@0.75, 53.6% mAP@0.5:0.95, and 83.1% F1-score, evidencing its capability to accurately and efficiently recognize livestock of different species and ages against complex backgrounds.

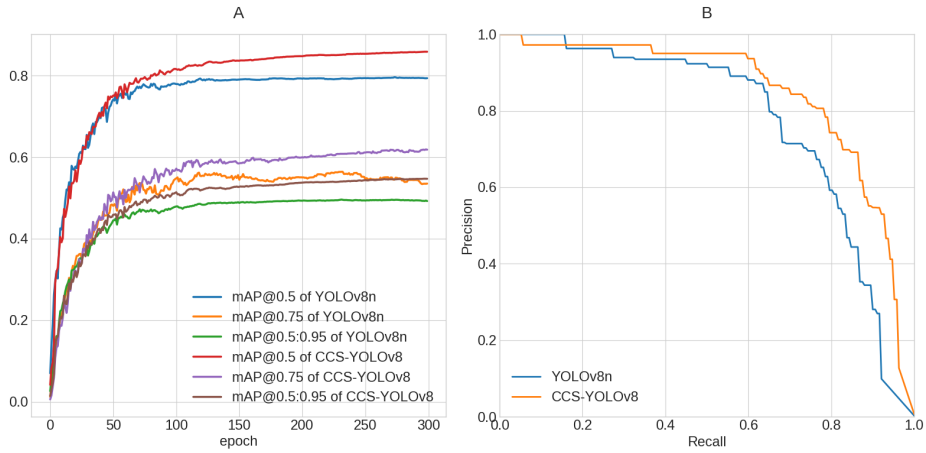

Figure 9. (**A**) The training mAP curves: the CCS-YOLOv8 model surpasses the YOLOv8n model on the curves of map0.5, map0.75, and map0.5:0.95. (**B**) The PR curve: The CCS-YOLOv8 model demonstrates superior performance for detecting and capturing most livestock instances with high recall and maintaining high accuracy.

3.3. In Contrast to Other Mainstream Models

By contrasting the performance of the CCS-YOLOv8 model with those of other popular one- and two-stage target detection models, we can further validate the enhanced model's efficacy. To assess the model's complexity in detail, these metrics include the number of model parameters and the quantity of computations. Faster R-CNN utilizes a two-stage computational inference design to effectively enhance detection accuracy. However, Faster R-CNN can encounter challenges with tasks that involve detecting small targets. This is primarily because the model heavily relies on high-dimensional feature mapping for prediction, which can lead to the overlooking of fine-grained feature information. Faster R-CNN's detection performance on the livestock dataset is noticeably worse than those of the one-stage algorithmic models, as Table 3 illustrates. Furthermore, the model has higher training and inference costs due to its large number of parameters and computing needs, which makes it challenging to satisfy the demands of livestock target monitoring. Among one-stage algorithms, the YOLOv8n model exhibits the lowest complexity. However, aside from SSD and YOLOv3tiny, it does not demonstrate a clear advantage in detection performance. By comparison, exceptional detection results are obtained by the CCS-YOLOv8 model. Meanwhile, the number of parameters and computational volume being low at 3.38 M and 13.7GFLOPs, respectively, indicate that the model is designed to be lightweight yet high-performing, making it more practical for real applications.

Table 3. Comparison with other models.

Model	Precision (%)	Recall (%)	mAP@0.5 (%)	mAP@0.75 (%)	mAP@0.5:0.95 (%)	F1-Score (%)	Parameters (M)	FLOPs (G)
Faster R-CNN	75.3	58.3	66.9	46.1	41	65.7	41.75	87.9
SSD [37]	75.6	65.2	75.7	45.6	43.5	70	25.12	88.2
YOLOv3tiny	80.1	68.5	71.1	48	44.5	73.8	12.14	19
YOLOv5s	84.1	80.1	81.1	55.8	50.6	82.1	7.04	16
YOLOXs [38]	80.5	75.8	81.3	53.5	49.6	78.1	8.97	13.4
YOLOv6n [39]	80.1	74.1	76.5	52.8	48.3	77	4.24	11.9
YOLOv7tiny [40]	84.1	79.6	81.7	53.2	49.3	81.8	6.03	13.2
YOLOv8n	83	74.3	78.6	53.7	48.8	78.4	3.01	8.2
YOLOv8s	83.9	80.2	82	59.3	52.8	82	11.1	28.7
CCS-YOLOv8	84.1	82.2	84.4	60.3	53.6	83.1	3.38	13.7

3.4. Robustness Test

The VisDrone2019 dataset [41] was collected by the AISKYEYE Machine Learning and Data Mining Lab team at Tianjin University. It consists of 288 video clips, totaling 261,908 frames, as well as 10,209 still images. These data samples were captured using a variety of UAV cameras in diverse scenarios. The dataset includes footage from 14 different cities in China, covering both urban and rural environments. The target objects in the dataset encompass pedestrians and vehicles, and the scenarios vary in density, ranging from sparse to congested [42,43].

By conducting experiments on this comprehensive dataset, the study aims to evaluate and compare the performance of the CCS-YOLOv8 model against the baseline algorithm. This assessment will provide insights into the model's robustness in recognizing and localizing objects in various challenging situations, contributing to the advancement of object detection techniques in aerial surveillance and monitoring applications.

Table 4 showcases the experimental results. On the VisDrone2019 dataset, the CCS-YOLOv8 model recorded 42.6% accuracy, 33.5% recall, 31.1% mAP@0.5, 16.8% mAP@0.75, 17.1% mAP@0.5:0.95, and 37.5% F1-score. Regarding accuracy, the CCS-YOLOv8 model shows an improvement of 2.2% over YOLOv8n, indicating its higher accuracy in detecting targets. Regarding recall, the CCS-YOLOv8 model demonstrates a 3.8% improvement over YOLOv8n, indicating its stronger recall ability with fewer missed detections. For mAP@0.5, mAP@0.75, and mAP@0.5:0.95, the CCS-YOLOv8 model registers improvements of 3.8%, 1.9%, and 2.1%, respectively, over YOLOv8n, indicating enhanced performance across different IoU thresholds. As a metric combining precision and recall, the F1-score of the CCS-YOLOv8 model sees a 3.3% improvement over YOLOv8n, highlighting its comprehensive performance superiority. To sum up, the CCS-YOLOv8 model's remarkable robustness has been validated by the experimental findings collected from the VisDrone2019 dataset.

Table 4. Results of the robustness test.

Model	Precision (%)	Recall (%)	mAP@0.5 (%)	mAP@0.75 (%)	mAP@0.5:0.95 (%)	F1-Score (%)
YOLOv8n	40.4	29.7	27.3	14.9	15.0	34.2
CCS-YOLOv8	42.6	33.5	31.1	16.8	17.1	37.5

4. Discussion

The CCS-YOLOv8 model is primarily utilized for the recognition of common livestock in the Qinghai region and holds potential applications in other animal and remote sensing target detection tasks. Despite significant improvements over the baseline model, we acknowledge that there are still instances of false alarms and missed detections. To address these issues, we plan to further optimize the model's training process by adjusting hyperparameters and exploring ensemble methods, such as combining multiple models to leverage their strengths and mitigate individual weaknesses. These steps are expected to enhance the overall detection accuracy and reliability of the model.

Our current dataset primarily includes three types of grasslands: alpine meadows, alpine steppes, and alpine desert grasslands. To improve our model's adaptability to diverse environments, we plan to collect additional datasets with varied backgrounds in the future. This will help ensure that our dataset encompasses all types of natural grasslands and improve the model's performance in various environmental contexts.

To further broaden the utility and applicability of our model, we intend to expand its capabilities to detect additional livestock species in future work. This expansion will involve including other commonly raised livestock species, thereby enhancing the model's performance and relevance in regions with diverse livestock populations.

As the model aims to strike a balance between accuracy and real-time processing capability, exploring lightweight optimization techniques is crucial. Techniques such as network pruning, quantization, and knowledge distillation can reduce model complexity without sacrificing performance, enabling more efficient inference on resource-constrained devices.

Going forward, the CCS-YOLOv8 model will be integrated into monitoring platforms such as drones and intelligent robots to refine the intelligent detection process and achieve real-time video detection of common livestock. This integration aims to enable proactive management strategies such as automated alerts for overgrazing, thereby enhancing the practical utility and impact of the model in real-world applications. By leveraging these advanced monitoring platforms, the model can provide timely and accurate livestock management support, ensuring sustainable and efficient livestock farming practices.

To ensure the continued effectiveness and relevance of the CCS-YOLOv8 model in evolving livestock management contexts, we plan to establish a framework for continuous evaluation and feedback. This framework will involve collecting feedback from end-users, monitoring the model's performance in different scenarios, and incorporating new data and insights into the model refinement processes. By implementing this continuous evaluation and feedback loop, we aim to iteratively improve the model over time, ensuring its adaptability and utility in various real-world applications. This approach will help us address emerging challenges and maintain the model's high performance in diverse livestock management contexts.

5. Conclusions

We have created a common livestock detection dataset in Qinghai by utilizing data collected by UAVs. As far as we are aware, this is the first dataset of its kind in this field. The introduction of the CBAM attention mechanism and the CARAFE operator into the model, coupled with the addition of a small object detection layer, aims to enhance the detection performance for livestock. Utilizing this dataset, the CCS-YOLOv8 model secured 84.1% accuracy, 82.2% recall, 84.4% mAP@0.5, 60.3% mAP@0.75, 53.6% mAP@0.5:0.95, and 83.1% F1-score, showing improvements of 1.1%, 7.9%, 5.8%, 6.6%, 4.8%, and 4.7%, respectively, over the baseline. Relative to other target detection models, the CCS-YOLOv8 model demonstrates excellent performance. By utilizing high-resolution images of livestock in Qinghai collected by UAVs and integrating the CCS-YOLOv8 model, this approach enables rapid detection of the ages and species of livestock. This method offers a new way to effectively manage the number of livestock and control overgrazing, aligning with the needs of modern precision livestock husbandry.

Author Contributions: Conceptualization, C.F.; methodology, C.F. and C.L.; software, C.F., J.N. and P.Y.; validation, Y.H., X.H. and J.N.; formal analysis, X.H. and Y.H.; investigation, C.F., C.L., S.K., P.Y. and X.H.; resources, C.F., S.K. and P.Y.; data curation, J.N. and Y.H.; writing—original draft preparation, C.F.; writing—review and editing, C.F., S.K. and C.L.; supervision, C.L.; project administration, C.L.; funding acquisition, C.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (grant No. 62166033) and the Basic Research Project of Qinghai Province, China (grant No. 2024-ZJ-788).

Institutional Review Board Statement: This study and all animal procedures therein were approved by the ethics committee for experimental animals of the Inner Qinghai University (No. qhdx-202306252).

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets generated and analyzed during the current study are available from the corresponding author upon reasonable request.

Acknowledgments: The authors thank the High-Performance Computing Center of Qinghai University for its support.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AP	Average Precision
box_loss	Bounding Box Loss
BCE	Binary Cross Entropy
CARAFE	Content-Aware ReAssembly of FEatures
CBAM	Convolutional Block Attention Module
CCS-YOLOv8	Comprehensive Contextual Sensing YOLOv8
cls_loss	Localization Loss
CSP	Cross Stage Partial DarkNet-53
C3	Cross Stage Partial Network with 3 Convolutions
dfl_loss	Distribution Focal Loss
DFL	Distribution Focal Loss
Faster R-CNN	Faster Region with CNN Feature
FN	False Negative
FP	False Positive
FPN	Feature Pyramid Network
GAP	Global Average Pooling
GMP	Global Maximum Pooling
IOU	Intersection over Union
mAP	Mean Average Precision
MLP	Multi-Layer Perception
PAN	Path Aggregation Network
R-CNN	Region with CNN Feature
SPPNet	Spatial Pyramid Pooling Network
TP	True Positive
UAV	Unmanned Aerial Vehicle
YOLO	You Only Look Once

References

1. Xue, Z.; Zhang, Z.; Lu, X.; Zou, Y.; Lu, Y.; Jiang, M.; Tong, S.; Zhang, K. Predicted areas of potential distributions of alpine wetlands under different scenarios in the Qinghai-Tibetan Plateau, China. *Glob. Planet. Chang.* **2014**, *123*, 77–85. [CrossRef]
2. Yang, H.; Gou, X.; Xue, B.; Xu, J.; Ma, W. How to effectively improve the ecosystem provisioning services in traditional agricultural and pastoral areas in China? *Ecol. Indic.* **2023**, *150*, 110244. [CrossRef]
3. Kou, Y.; Yuan, Q.; Dong, X.; Li, S.; Deng, W.; Ren, P. Dynamic Response and Adaptation of Grassland Ecosystems in the Three-River Headwaters Region under Changing Environment: A Review. *Int. J. Environ. Res. Public Health* **2023**, *20*, 4220. [CrossRef] [PubMed]
4. Zhou, W.; Gang, C.; Zhou, L.; Chen, Y.; Li, J.; Ju, W.; Odeh, I. Dynamic of grassland vegetation degradation and its quantitative assessment in the northwest China. *Acta Oecologica* **2014**, *55*, 86–96. [CrossRef]
5. Li, H.; Li, T.; Sun, W.; Zhang, W.; Zha, X. Degradation of wetlands on the Qinghai-Tibetan Plateau causing a loss in soil organic carbon in 1966–2016. *Plant Soil* **2021**, *467*, 253–265. [CrossRef]
6. Troiano, C.; Buglione, M.; Petrelli, S.; Belardinelli, S.; De Natale, A.; Svenning, J.C.; Fulgione, D. Traditional Free-Ranging Livestock Farming as a Management Strategy for Biological and Cultural Landscape Diversity: A Case from the Southern Apennines. *Land* **2021**, *10*, 957. [CrossRef]
7. Tan, K.; Ni, L.; Gong, B.; Jia, C.; Fang, Y.; Tang, L.; Huang, Z.; Ji, X.; Jia, K. Application of Yolo Algorithm in Livestock Counting and Identificatiom System. In Proceedings of the 2023 20th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP), Chengdu, China, 15–17 December 2003; pp. 1–6. [CrossRef]
8. Lallo, C.H.O.; Cohen, J.; Rankine, D.; Taylor, M.; Cambell, J.; Stephenson, T. Characterizing heat stress on livestock using the temperature humidity index (THI)-prospects for a warmer Caribbean. *Reg. Environ. Chang.* **2018**, *18*, 2329–2340. [CrossRef]
9. Bao, J.; Xie, Q. Artificial intelligence in animal farming: A systematic literature review. *J. Clean. Prod.* **2022**, *331*, 129956. [CrossRef]
10. Rančić, K.; Blagojević, B.; Bezdan, A.; Ivošević, B.; Tubić, B.; Vranešević, M.; Pejak, B.; Crnojević, V.; Marko, O. Animal Detection and Counting from UAV Images Using Convolutional Neural Networks. *Drones* **2023**, *7*, 179. [CrossRef]
11. Moradeyo, O.; Olaniyan, A.S.; Ojoawo, A.O.; Olawale, J.; Bello, R.W. YOLOv7 Applied to Livestock Image Detection and Segmentation Tasks in Cattle Grazing Behavior, Monitor and Intrusions. *J. Appl. Sci. Environ. Manag.* **2023**, *27*, 953–958. [CrossRef]
12. Ahmad, M.; Abbas, S.; Fatima, A.; Issa, G.F.; Ghazal, T.M.; Khan, M.A. Deep Transfer Learning-Based Animal Face Identification Model Empowered with Vision-Based Hybrid Approach. *Appl. Sci.* **2023**, *13*, 1178. [CrossRef]

13. Liang, F.; Zhou, Y.; Chen, X.; Liu, F.; Zhang, C.; Wu, X. Review of target detection technology based on deep learning. In Proceedings of the 5th International Conference on Control Engineering and Artificial Intelligence, Sanya, China, 14–16 January 2021; pp. 132–135. [CrossRef]
14. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587. [CrossRef]
15. Girshick, R. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015.
16. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef]
17. Guo, B.; Ling, S.; Tan, H.; Wang, S.; Wu, C.; Yang, D. Detection of the Grassland Weed Phlomoides umbrosa Using Multi-Source Imagery and an Improved YOLOv8 Network. *Agronomy* **2023**, *13*, 3001. [CrossRef]
18. Jiang, P.; Ergu, D.; Liu, F.; Cai, Y.; Ma, B. A Review of Yolo Algorithm Developments. *Procedia Comput. Sci.* **2022**, *199*, 1066–1073. [CrossRef]
19. Xiao, Z.; Wan, F.; Lei, G.; Xiong, Y.; Xu, L.; Ye, Z.; Liu, W.; Zhou, W.; Xu, C. FL-YOLOv7: A Lightweight Small Object Detection Algorithm in Forest Fire Detection. *Forests* **2023**, *14*, 1812. [CrossRef]
20. Du, X.; Qi, Y.; Zhu, J.; Li, Y.; Liu, L. Enhanced lightweight deep network for efficient livestock detection in grazing areas. *Int. J. Adv. Robot. Syst.* **2024**, *21*, 17298806231218865. [CrossRef]
21. Pu, J.; Yu, C.; Chen, X.; Zhang, Y.; Yang, X.; Li, J. Research on Chengdu Ma Goat Recognition Based on Computer Vison. *Animals* **2022**, *12*, 1746. [CrossRef] [PubMed]
22. Kurniadi, F.A.; Setianingsih, C.; Syaputra, R.E. Innovation in Livestock Surveillance: Applying the YOLO Algorithm to UAV Imagery and Videography. In Proceedings of the 2023 IEEE 9th International Conference on Smart Instrumentation, Measurement and Applications (ICSIMA), Kuala Lumpur, Malaysia, 17–18 October 2023; pp. 246–251. [CrossRef]
23. Zhang, X.; Xuan, C.; Ma, Y.; Liu, H.; Xue, J. Lightweight model-based sheep face recognition via face image recording channel. *J. Anim. Sci.* **2024**, *102* , skae066. [CrossRef] [PubMed]
24. Yao, H.; Qin, R.; Chen, X. Unmanned Aerial Vehicle for Remote Sensing Applications—A Review. *Remote Sens.* **2019**, *11*, 1443. [CrossRef]
25. Terven, J.; Córdova-Esparza, D.M.; Romero-González, J.A. A Comprehensive Review of YOLO Architectures in Computer Vision: From YOLOv1 to YOLOv8 and YOLO-NAS. *Mach. Learn. Knowl. Extr.* **2023**, *5*, 1680–1716. [CrossRef]
26. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
27. Wang, C.Y.; Liao, H.Y.M.; Wu, Y.H.; Chen, P.Y.; Hsieh, J.W.; Yeh, I.H. CSPNet: A new backbone that can enhance learning capability of CNN. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 390–391.
28. Li, Y.; Fan, Q.; Huang, H.; Han, Z.; Gu, Q. A Modified YOLOv8 Detection Network for UAV Aerial Image Recognition. *Drones* **2023**, *7*, 304. [CrossRef]
29. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. In Proceedings of the Computer Vision—ECCV 2014, Zurich, Switzerland, 6–12 September 2014; Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T., Eds.; Springer: Cham, Switzerland, 2014; pp. 346–361. [CrossRef]
30. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 936–944. [CrossRef]
31. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path Aggregation Network for Instance Segmentation. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8759–8768. [CrossRef]
32. Zhang, H.; Wang, Y.; Dayoub, F.; Sünderhauf, N. VarifocalNet: An IoU-aware Dense Object Detector. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 8510–8519. [CrossRef]
33. Li, X.; Wang, W.; Wu, L.; Chen, S.; Hu, X.; Li, J.; Tang, J.; Yang, J. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 21002–21012.
34. Zagoruyko, S.; Komodakis, N. Paying More Attention to Attention: Improving the Performance of Convolutional Neural Networks via Attention Transfer. *arXiv* **2017**, arXiv:1612.03928.
35. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In Proceedings of the Computer Vision—ECCV 2018, Munich, Germany, 8–14 September 2018; Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds.; Springer: Cham, Switzerland, 2018; pp. 3–19. [CrossRef]
36. Wang, J.; Chen, K.; Xu, R.; Liu, Z.; Loy, C.C.; Lin, D. CARAFE: Content-Aware ReAssembly of FEatures. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 3007–3016. [CrossRef]
37. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single Shot MultiBox Detector. In Proceedings of the Computer Vision—ECCV 2016, Amsterdam, The Netherlands, 11–14 October 2016; Leibe, B.; Matas, J.; Sebe, N.; Welling, M., Eds.; Springer: Cham, Switzerland, 2016; pp. 21–37. [CrossRef]

38. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. Yolox: Exceeding yolo series in 2021. *arXiv* **2021**, arXiv:2107.08430.
39. Li, C.; Li, L.; Jiang, H.; Weng, K.; Geng, Y.; Li, L.; Ke, Z.; Li, Q.; Cheng, M.; Nie, W.; et al. YOLOv6: A Single-Stage Object Detection Framework for Industrial Applications. *arXiv* **2022**, arXiv:2209.02976.
40. Wang, C.; Bochkovskiy, A.; Liao, H. YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–23 June 2023; pp. 7464–7475. [CrossRef]
41. Zhu, P.; Wen, L.; Du, D.; Bian, X.; Fan, H.; Hu, Q.; Ling, H. Detection and tracking meet drones challenge. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 7380–7399. [CrossRef]
42. Zhu, G.; Zhu, F.; Wang, Z.; Xiong, G.; Tian, B. Small Target Detection Algorithm Based On Multi-target Detection Head And Attention Mechanism. In Proceedings of the 2023 IEEE 3rd International Conference on Digital Twins and Parallel Intelligence (DTPI), Orlando, FL, USA, 7–9 November 2023; pp. 1–6. [CrossRef]
43. Youssef, Y.; Elshenawy, M. Automatic Vehicle Counting and Tracking in Aerial Video Feeds using Cascade Region-based Convolutional Neural Networks and Feature Pyramid Networks. *Transp. Res. Rec.* **2021**, *2675*, 304–317. [CrossRef]