drones

MDPI

Article

DLCAPL: Semi-Supervised Livestock Detection in the Ruoergai Alpine Grasslands Using Dual-View Localization Consistency and Reliability-Weighted Pseudo-Label Learning

Shengfang Zhu [1], Lianguo Wang [1,*] and Limin Hua [2,*]

[1] College of Information Science and Technology, Gansu Agricultural University, Lanzhou 730070, China; zhushengfang@gsau.edu.cn
[2] Engineering and Technology Research Center for Alpine Rodent Pest Control of National Forestry and Grassland Administration, Key Laboratory of Grassland Ecosystem of the Ministry of Education, College of Grassland Science, Gansu Agricultural University, Lanzhou 730070, China
* Correspondence: wanglg@gsau.edu.cn (L.W.); hualm@gsau.edu.cn (L.H.)

Highlights

What are the main findings?

- DLCAPL uses dual-view localization consistency to improve the use of unlabeled UAV imagery for livestock detection under limited-annotation conditions.
- With 30% of the training images labeled, DLCAPL approaches the detection performance of the fully supervised model trained with all training images labeled.

What are the implications of the main findings?

- Semi-supervised learning mitigates the heavy manual-annotation burden of high-resolution UAV-based livestock surveys.
- Reliability-aware pseudo-label learning built upon localization-stability evaluation facilitates scalable livestock monitoring within complex alpine-grassland scenarios.

Abstract

Unmanned aerial vehicle (UAV) remote sensing provides an efficient approach for livestock monitoring in alpine grassland environments, particularly in regions where traditional field surveys are constrained by complex terrain and large spatial coverage. However, livestock targets in nadir-view UAV imagery are typically small, densely distributed, partially occluded, and visually similar to the surrounding vegetation background, which makes accurate detection challenging. In addition, manually annotating large numbers of individual animals in high-resolution UAV images is time-consuming and labor-intensive. To reduce dependence on extensive manual annotation while improving the utilization of unlabeled UAV imagery, this study proposes DLCAPL (Dual-View Localization Consistency and Reliability-Weighted Pseudo-Label Learning), a semi-supervised livestock detection framework based on YOLOv13s. Within an exponential moving average (EMA)-based teacher–student framework, a dual-view localization consistency mechanism is introduced to evaluate pseudo-label reliability based on localization stability across augmented views. Teacher predictions are further hierarchically divided into reliable, weak, and ignored pseudo-labels using medium- and high-confidence thresholds. For weak pseudo-labels, classification confidence and localization consistency are jointly used to estimate reliability and assign reliability-aware supervision weights, thereby retaining informative difficult targets while reducing the influence of noisy pseudo-labels. Experiments were conducted on a self-constructed UAV livestock dataset collected in the

Academic Editor: Fei Liu

Received: 10 September 2026
Revised: 21 October 2026
Accepted: 23 October 2026
Published: 26 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

Ruoergai alpine grasslands containing 602 images and 57,390 livestock instances. With only 30% labeled data and 70% unlabeled data, DLCAPL achieved 73.28% mAP@0.5 and 50.68% mAP@0.5:0.95. Compared with the supervised baseline under the same annotation ratio, DLCAPL improved these two metrics by 17.96 and 16.43 percentage points, respectively, while approaching the performance of the fully supervised YOLOv13s model trained with 100% labeled data. These results demonstrate that DLCAPL can effectively exploit unlabeled UAV imagery, reduce annotation requirements, and maintain competitive detection performance for livestock monitoring in complex alpine grassland environments.

Keywords: UAV-based livestock monitoring; semi-supervised object detection; pseudo-label learning; teacher–student framework; dual-view localization consistency; Ruoergai alpine grasslands

1. Introduction

Livestock production is an important agricultural activity in alpine grassland regions. On the Qinghai–Tibetan Plateau, grasslands cover a substantial proportion of the region and support extensive livestock husbandry 1,2. Accordingly, accurate information on livestock abundance, species composition, and spatial distribution is essential for grazing management, grass–livestock balance assessment, animal health monitoring, and grassland resource conservation. Conventional livestock surveys mainly rely on manual field inspections, ground-based photography, or fixed monitoring devices, which are labor-intensive and time-consuming and are often constrained by survey coverage, terrain conditions, and livestock movement **Error! Reference source not found.**,4. With the increasing demand for efficient livestock monitoring in complex alpine environments, unmanned aerial vehicles (UAVs) have emerged as an effective remote sensing platform due to their operational flexibility, high data-acquisition efficiency, and high spatial resolution. UAV imagery can provide detailed spatial information over large grassland areas and serve as an important data source for automated livestock detection and monitoring 5.

In recent years, with the rapid development of deep learning-based object detection, convolutional neural network (CNN)-based detectors [6], particularly the You Only Look Once (YOLO) series 7, have been widely applied to UAV-based livestock detection and counting due to their favorable balance between detection accuracy and computational efficiency. Previous studies have shown that deep learning models can automatically extract discriminative visual features from complex grassland scenes, enabling the recognition and localization of different livestock species 8. However, livestock detection in UAV imagery collected from alpine grassland environments remains challenging due to complex backgrounds and target appearance variations. In nadir-view UAV images, livestock targets typically occupy only a small number of pixels and exhibit substantial scale variation. Moreover, animals are often densely distributed, partially occluded, and visually similar to surrounding vegetation and heterogeneous backgrounds. These characteristics weaken the texture and boundary information of small targets, making closely spaced individuals difficult to distinguish and increasing the likelihood of missed detections and inaccurate bounding-box localization.

Most existing UAV-based livestock detection methods rely on fully supervised learning and therefore require manual annotation of livestock targets in the training images 9. This annotation burden is particularly pronounced in high-resolution grassland UAV imagery, where a single image may contain numerous small and densely distributed livestock targets that must be individually localized and labeled. Consequently, annotating

such images is highly time- and labor-intensive 10. As UAV data acquisition becomes increasingly accessible, the availability of unlabeled imagery continues to grow, whereas manual annotation remains a major bottleneck for large-scale model development and deployment. Therefore, effectively exploiting unlabeled UAV imagery through semi-supervised learning provides a promising solution for reducing annotation requirements while maintaining competitive livestock detection performance.

Semi-supervised object detection jointly exploits a small amount of labeled data and a large amount of unlabeled data to extract useful supervisory information from unlabeled samples, thereby improving detection performance under limited annotation conditions 11. Most existing semi-supervised detectors follow a teacher–student paradigm, where the teacher network generates pseudo-labels from weakly augmented unlabeled images and the student network learns from strongly augmented images under pseudo-label supervision. The teacher parameters are typically updated using an exponential moving average (EMA) of the student parameters to provide stable pseudo-label predictions **Error! Reference source not found.**. Unbiased Teacher and Soft Teacher have demonstrated the effectiveness of teacher–student learning and online pseudo-label generation for semi-supervised object detection **Error! Reference source not found.**14. However, the quality of pseudo-labels remains a critical factor limiting the performance of teacher–student-based approaches, especially in complex scenes containing small-scale, densely distributed, or partially occluded targets. Subsequent studies have further investigated how to improve pseudo-label utilization for difficult targets and dense prediction scenarios. Dense Teacher showed that sparse pseudo-labels obtained through confidence thresholding and non-maximum suppression may discard useful information contained in teacher predictions, thereby limiting the exploitation of unlabeled data **Error! Reference source not found.**. Global Focal Learning mines difficult regions according to prediction consistency, improving the utilization of potential targets in unlabeled remote-sensing imagery 16. Denser Teacher further combines adaptive dense pseudo-label selection with multi-scale learning to achieve a better balance between informative predictions and pseudo-label noise in dense scenes 17. However, direct comparison among different semi-supervised object detection methods remains challenging because existing approaches are usually developed under different detection architectures, backbone networks, optimization strategies, and experimental settings. Therefore, in this study, all semi-supervised comparisons are conducted under the unified YOLOv13s framework to minimize the influence of detector architecture differences and more directly evaluate the contribution of the proposed pseudo-label optimization strategies.

However, directly applying existing semi-supervised object detection methods to UAV livestock imagery remains challenging due to the unique characteristics of livestock targets in complex alpine environments. First, distant, small-scale, or partially occluded livestock targets often receive relatively low prediction confidence. A fixed high-confidence threshold may therefore discard true targets that remain informative for learning, whereas lowering the threshold can introduce more background false positives and noisy pseudo-labels. MixTeacher has shown that objects at extreme scales are more likely to be ignored during pseudo-label selection because of their relatively low prediction confidence 18. Although lowering confidence thresholds can increase the number of selected pseudo-labels, it may also introduce unreliable predictions, making it difficult to balance pseudo-label completeness and reliability. Second, classification confidence does not necessarily reflect bounding-box localization quality. Even high-confidence predictions may exhibit inaccurate or unstable spatial localization, making it difficult to distinguish reliably localized difficult targets from noisy predictions with substantial localization errors 19. This limitation is particularly important for UAV livestock detection, where small objects occupy limited image regions and minor localization deviations may result in

considerable IoU variations. Therefore, confidence-based pseudo-label selection alone may be insufficient to characterize pseudo-label reliability. In addition, medium-confidence pseudo-label candidates may vary considerably in quality. If they are trained with the same supervision strength, unstable predictions caused by model updates and data augmentation may continuously introduce noise into student training and result in error accumulation 20. Consequently, semi-supervised UAV livestock detection requires a more comprehensive pseudo-label evaluation strategy that considers both prediction confidence and localization stability, as well as differentiated supervision according to pseudo-label reliability.

To address these challenges, we propose DLCAPL (Dual-View Localization Consistency and Reliability-Weighted Pseudo-Label Learning), a YOLOv13s-based semi-supervised livestock detector. Within an exponential moving average (EMA) teacher–student framework, dual-view localization consistency (DLC) assesses prediction stability across augmented views. Candidate pseudo-labels are divided into reliable, weak, and ignored groups using medium- and high-confidence thresholds. For weak pseudo-labels, classification confidence and localization consistency jointly determine continuous supervision weights. This design retains potentially informative difficult targets while reducing the influence of unreliable predictions.

Experiments were conducted on a self-constructed UAV livestock dataset collected in the Ruoergai alpine grasslands, containing 602 UAV images and 57,390 livestock instances. The experimental results demonstrate that DLCAPL can effectively exploit unlabeled UAV imagery under limited annotation conditions, substantially reduce annotation requirements, and achieve competitive detection performance for small, densely distributed, and partially occluded livestock targets.

The main contributions of this study are summarized as follows:

(1) A semi-supervised livestock detection framework, DLCAPL, is proposed by integrating dual-view localization consistency, hierarchical pseudo-label partitioning, and reliability-weighted supervision within a unified teacher–student architecture to effectively exploit unlabeled UAV imagery under limited annotation conditions.
(2) We introduce DLC to assess pseudo-label reliability from localization stability across augmented views, providing a spatial cue that complements classification confidence.
(3) A hierarchical pseudo-label partitioning and reliability-weighted supervision strategy is developed to retain informative medium-confidence pseudo-labels and assign continuous weights according to pseudo-label reliability, thereby reducing noisy supervision caused by low-quality pseudo-labels.

2. Materials

2.1. Study Area

The UAV imagery used in this study was acquired in Ruoergai County, Aba Tibetan and Qiang Autonomous Prefecture, Sichuan Province, China. The location of the study area and representative sampling sites are shown in Figure 1. Ruoergai County is located on the northeastern margin of the Qinghai–Tibet Plateau in northwestern Sichuan Province, with an average elevation of approximately 3500 m above sea level. It is a representative alpine pastoral region in northwestern Sichuan **Error! Reference source not found.**. The Ruoergai Plateau has a cold-temperate, humid to semi-humid continental monsoon climate, and its ecosystems are dominated by alpine meadows, alpine grasslands, and wetlands 22. Local livestock production relies primarily on grazing on natural grasslands. Yaks, Tibetan sheep, and horses are the major livestock species in the study area and were therefore selected as the target categories in this study.

Livestock in the study area occur both in dispersed groups and in localized clusters. Heterogeneous vegetation and surface textures, together with variation in animal size, posture, and occlusion, make small or closely spaced targets difficult to distinguish from the background. These conditions make Ruoergai a relevant setting for evaluating UAV-based livestock detection in alpine grasslands.

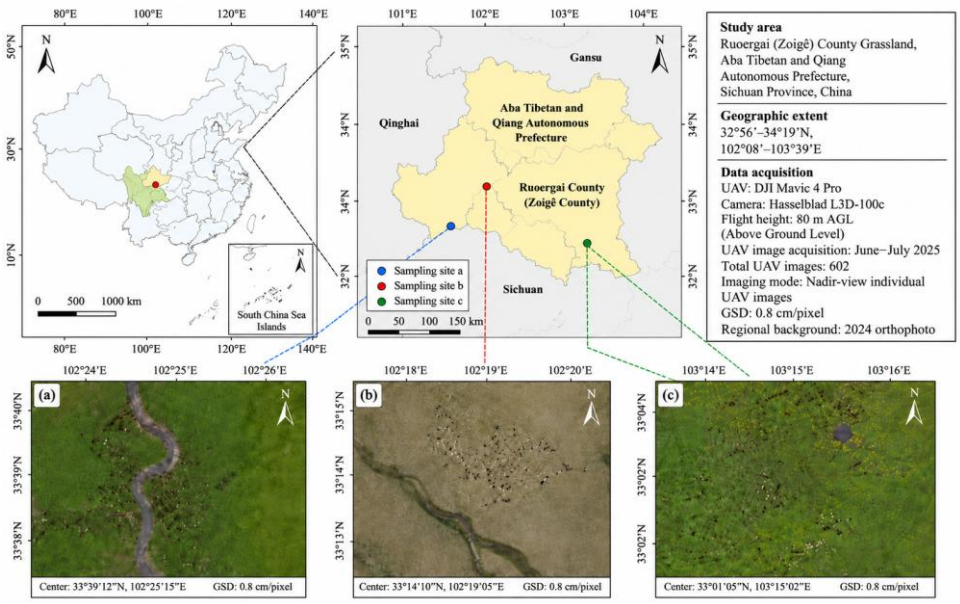


Figure 1. Study area and representative UAV images. The 2024 orthophoto provides geographic context for the study area and sampling sites. Panels (**a**–**c**) show individual UAV photographs acquired during field surveys in June and July 2025. The orthophoto was used for geographic reference only and was not included in the detection dataset.

2.2. UAV Image Acquisition

UAV images were acquired in June and July 2025 using a DJI Mavic 4 Pro (SZ DJI Technology Co., Ltd., Shenzhen, China) equipped with a Hasselblad L3D-100c camera (Hasselblad, Gothenburg, Sweden). The gimbal pitch was set to −90° to obtain nadir-view images. Surveys were conducted at multiple sites and on different dates under natural grazing conditions. The nominal flight altitude was approximately 80 m above ground level. Based on the camera parameters and original image dimensions, the theoretical ground sampling distance (GSD) was estimated at approximately 0.8 cm/pixel. The detection dataset comprises individual UAV photographs, not orthomosaics or manually generated cropped tiles. The original image dimensions were 12,288 × 8192 pixels. Image metadata recorded a focal length of 14 mm (28 mm in 35 mm-equivalent terms), a shutter speed of 1/30 s, an aperture of f/11, and ISO 100 [23].

The dataset contains 602 images collected across multiple grazing sites and dates. The scenes vary in vegetation cover, surface conditions, illumination, and livestock distribution, including meadows, wetlands, riverbanks, bare ground, and vegetation patches. Livestock range from dispersed individuals to dense clusters and vary in scale and occlusion, providing examples of the small targets, crowded scenes, and complex backgrounds encountered in alpine grasslands.

2.3. Data Processing

2.3.1. Data Preprocessing

To construct a dataset suitable for deep-learning-based object detection, the original UAV images were manually annotated using the open-source annotation tool LabelMe

[24]. According to the major livestock species present in the study area, the detection categories were defined as yak, Tibetan sheep, and horse, with the corresponding class labels set to "yak", "sheep", and "horse", respectively. During annotation, each visually identifiable livestock individual was manually delineated using polygon annotations in LabelMe to closely follow the visible target boundary. The resulting polygon annotations were stored in JSON format and subsequently converted into YOLO-format rectangular bounding boxes for object detection. The final dataset contained 57,390 livestock instances, including 43,080 yaks, 14,078 sheep, and 232 horses. The detailed distribution of livestock instances among the training, validation, and test sets is summarized in Table 1. Representative images of Tibetan sheep are shown in Figure 2, while representative images of yaks and horses are shown in Figure 3.

Table 1. Dataset composition and livestock instance distribution across different subsets.

Dataset Split	Images	Yak Instances	Sheep Instances	Horse Instances	Total Instances
Training	482	34,861	11,291	186	46,338
Validation	60	4351	1080	24	5455
Test	60	3868	1707	22	5597
Total	602	43,080	14,078	232	57,390

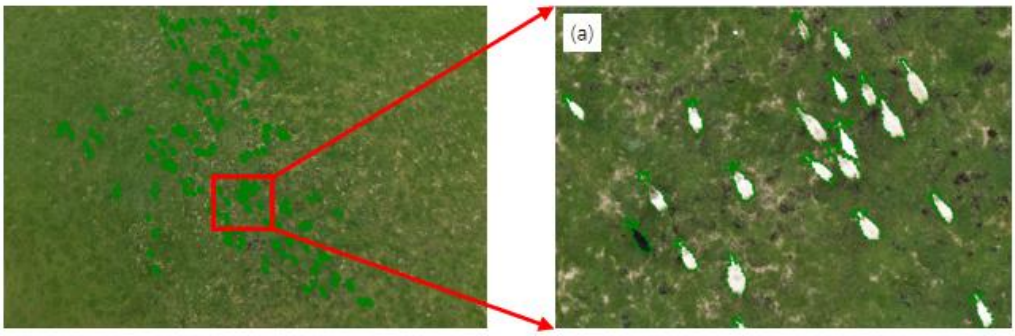


Figure 2. Representative UAV image of a Tibetan sheep flock with an enlarged local view. The red arrows indicate the correspondence between the selected region in the original image and the enlarged local view (a), which shows Tibetan sheep instances.

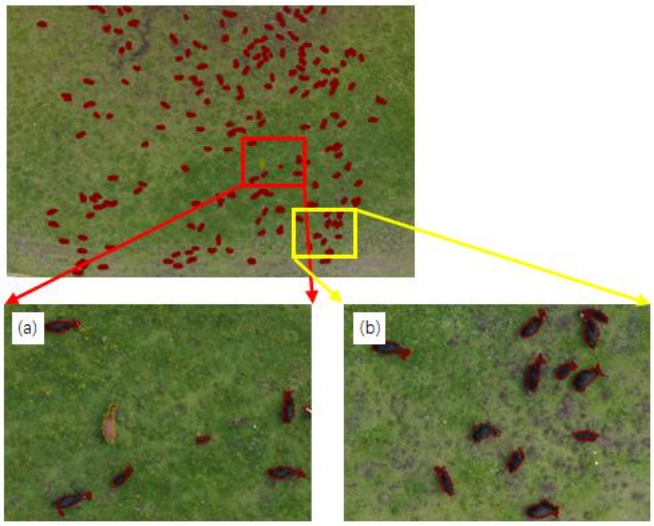


Figure 3. Representative UAV image containing yaks and horses with enlarged local views. The red and yellow rectangles indicate the regions selected for enlargement in panels (**a**) and (**b**),

respectively, and the corresponding colored arrows connect these regions to their enlarged views. (**a**) Horse instance; (**b**) yak instance.

2.3.2. Data Augmentation and Dataset Partitioning

To improve model robustness to complex grassland backgrounds and variations in UAV imaging conditions, data augmentation was applied only to the training set, whereas no stochastic augmentation was applied to the validation or test sets. The augmentation strategies included geometric transformations, photometric transformations, and image degradation to simulate variations in target position and scale, illumination, color, and image quality [25]. For geometric transformations involving changes in target position, annotation coordinates were synchronously updated and bounding boxes were recalculated. Targets extending beyond the image boundaries or retaining insufficient valid regions after augmentation were cropped or filtered to maintain consistency between images and annotations.

To prevent an original image and its augmented versions from being distributed across different dataset subsets, the 602 original images were first grouped according to flight segments and partitioned before any augmentation was applied, after which data augmentation was applied exclusively to the training set **Error! Reference source not found.**. Images acquired from the same or adjacent flight segments, particularly those containing overlapping scenes or the same livestock individuals, were assigned to the same subset to avoid spatial leakage across the training, validation, and test sets. The dataset was divided into training, validation, and test sets at an approximate ratio of 80%:10%:10%, corresponding to 482, 60, and 60 images, respectively. The validation and test sets retained their complete manual annotations and underwent only image resizing and normalization.

For the semi-supervised experiments, the 482 training images were further divided into labeled and unlabeled subsets. Using a fixed random seed of 42, 145 images were randomly selected as labeled data, accounting for 30% of the training set, while the remaining 337 images were treated as unlabeled data, accounting for 70%. A three-fold equivalent sampling strategy was adopted for the supervised branch, whereby each labeled image contributed the original sample and two online randomly augmented samples, resulting in an equivalent supervised sample size of 435 images per training epoch. This strategy did not introduce additional manually annotated images; instead, it increased the number of supervised training observations through repeated online augmentation of the same labeled images. The augmented samples were dynamically generated during training and were not stored as fixed image files. The supervised baseline with 30% labeled data (YOLOv13s-S30) also adopted the same three-fold equivalent sampling strategy to ensure a fair comparison with the semi-supervised experiments. The YOLOv13s-S30 baseline used the same data augmentation strategy and training configuration as the supervised branch of DLCAPL, except that no unlabeled images or pseudo-label learning process were involved.

For each unlabeled image, a weakly augmented view and a strongly augmented view were generated. Weak augmentation included random horizontal flipping (p = 0.5), aspect-ratio-preserving random resizing (scale factor 0.8–1.2), random brightness adjustment (±0.15), random contrast adjustment (±0.15), and image normalization. These mild transformations were designed to preserve target semantics and basic structural characteristics for stable pseudo-label generation. Strong augmentation included all weak augmentation operations and additionally applied random affine transformation (rotation within ±10° and translation), hue and saturation jitter (±0.2), Gaussian blur, random grayscale conversion (p = 0.1), and detection-oriented CutMix. Bounding-box coordinates were synchronously updated for all geometric transformations and CutMix operations. The

weakly augmented view was fed into the teacher model to generate pseudo-labels, whereas the strongly augmented view was fed into the student model for consistency learning. Accordingly, the 337 unlabeled images generated 674 augmented views during each complete sampling cycle, although the actual number of unlabeled source images remained 337. Their manual annotations were completely hidden during training and were used only for subsequent pseudo-label quality analysis. All comparison and ablation experiments employed the same dataset partition to ensure fair and reproducible comparisons.

3. Methods

3.1. Overall Architecture of DLCAPL

To address the challenges of missed detections of small livestock targets, unstable pseudo-label localization, and accumulated noisy supervision in UAV imagery under limited annotation conditions, this study proposes DLCAPL, a semi-supervised livestock detection method based on Dual-View Localization Consistency and Reliability-Weighted Pseudo-Label Learning, using YOLOv13s as the base detector. YOLOv13s adopts a one-stage detection architecture and performs object classification and bounding-box regression on multi-scale feature maps, providing a favorable balance between detection accuracy and computational efficiency [27]. DLCAPL consists of a training stage and an inference stage, and its overall architecture is illustrated in Figure 4.

During training, the framework consists of a supervised branch and an unsupervised branch, which share the same student network and model parameters. In the supervised branch, labeled images are augmented and fed into the student model, and the supervised loss is calculated based on the model predictions and ground-truth annotations. In the unsupervised branch, two views of the same unlabeled image are generated through weak and strong augmentations. The weakly augmented view is fed into the teacher model to generate candidate pseudo-labels, whereas the strongly augmented view is fed into the student model for consistency learning. The student model is optimized through backpropagation, while the teacher model is excluded from gradient computation. Teacher predictions are detached from the computational graph and used as fixed supervision targets for student training. Instead, its parameters are updated from those of the student model using an exponential moving average (EMA) strategy.

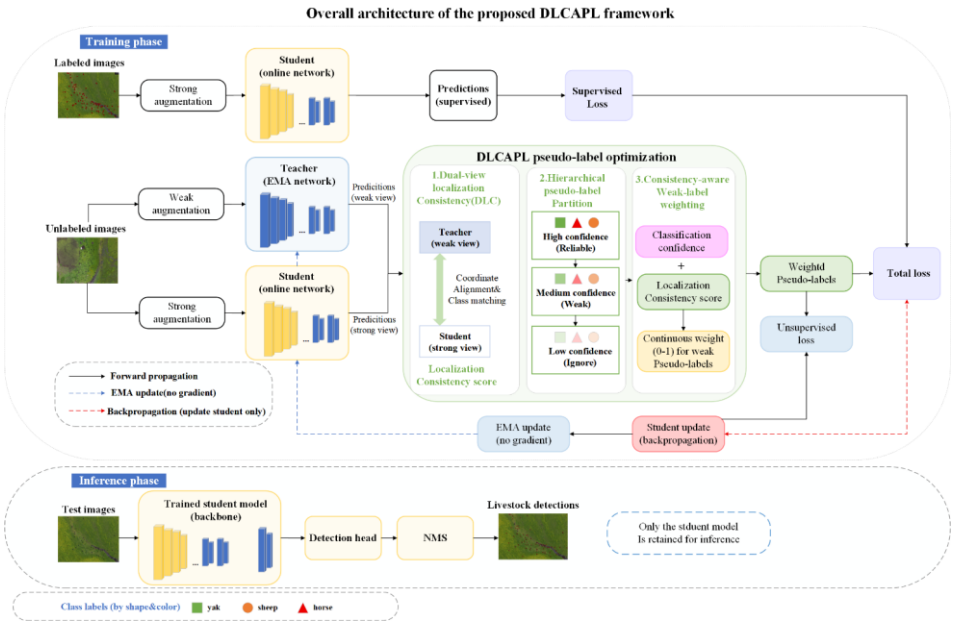


Figure 4. Overall architecture of the proposed DLCAPL semi-supervised livestock detection framework. Solid gray arrows indicate forward propagation, blue dashed arrows indicate exponential moving average (EMA) updates without gradients, and red dashed arrows indicate backpropagation for updating the student model only. In the inference phase, green circles, orange circles, and red triangles represent yaks, sheep, and horses, respectively.

To effectively exploit reliable supervisory information from unlabeled images, DLCAPL sequentially performs dual-view localization consistency evaluation, hierarchical pseudo-label partitioning, and reliability-weighted supervision. According to the geometric transformations recorded during weak and strong augmentation, predictions from the two views are mapped into a common image coordinate system before matching and consistency evaluation. Localization stability is then evaluated based on class correspondence and the spatial overlap between predicted bounding boxes. Meanwhile, according to the classification confidence produced by the teacher model, two confidence thresholds, namely a medium threshold ($\tau_m = 0.20$) and a high threshold ($\tau_h = 0.50$), are used to partition candidate predictions into reliable pseudo-labels, weak pseudo-labels, and ignored predictions. These threshold values were selected based on preliminary validation experiments, considering the trade-off between pseudo-label quantity and reliability. Reliable pseudo-labels above the high-confidence threshold are assigned full supervision weight, whereas predictions below the medium-confidence threshold are ignored during unsupervised training. For weak pseudo-labels with confidence scores between the medium and high thresholds, classification confidence and dual-view localization consistency are jointly integrated to estimate their reliability, based on which continuous supervision weights are assigned. In this manner, candidate predictions with different reliability levels are utilized through differentiated pseudo-label selection and adaptive supervision.

The student model is jointly optimized using the supervised loss and the weighted unsupervised loss. Considering that the teacher model has limited predictive capability and pseudo-label quality remains unstable during the early training stage, a progressive unsupervised-loss weighting strategy is adopted. The model therefore relies primarily on ground-truth annotations during the early stage to establish stable category discrimination and object localization, while the contribution of unlabeled data is gradually increased as training proceeds 28. After training, only the student model is retained for inference, whereas the teacher model and the DLCAPL pseudo-label optimization modules are not involved in deployment. Therefore, DLCAPL introduces no additional inference-time modules beyond the YOLOv13s student detector.

3.2. Teacher–Student Semi-Supervised Training Framework

To effectively exploit livestock information contained in unlabeled UAV images, a teacher–student semi-supervised training framework based on weak–strong augmentation and an exponential moving average (EMA) is constructed on top of YOLOv13s, upon which the DLCAPL pseudo-label quality-control mechanism is further introduced [29]. The proposed framework is illustrated in Figure 5.

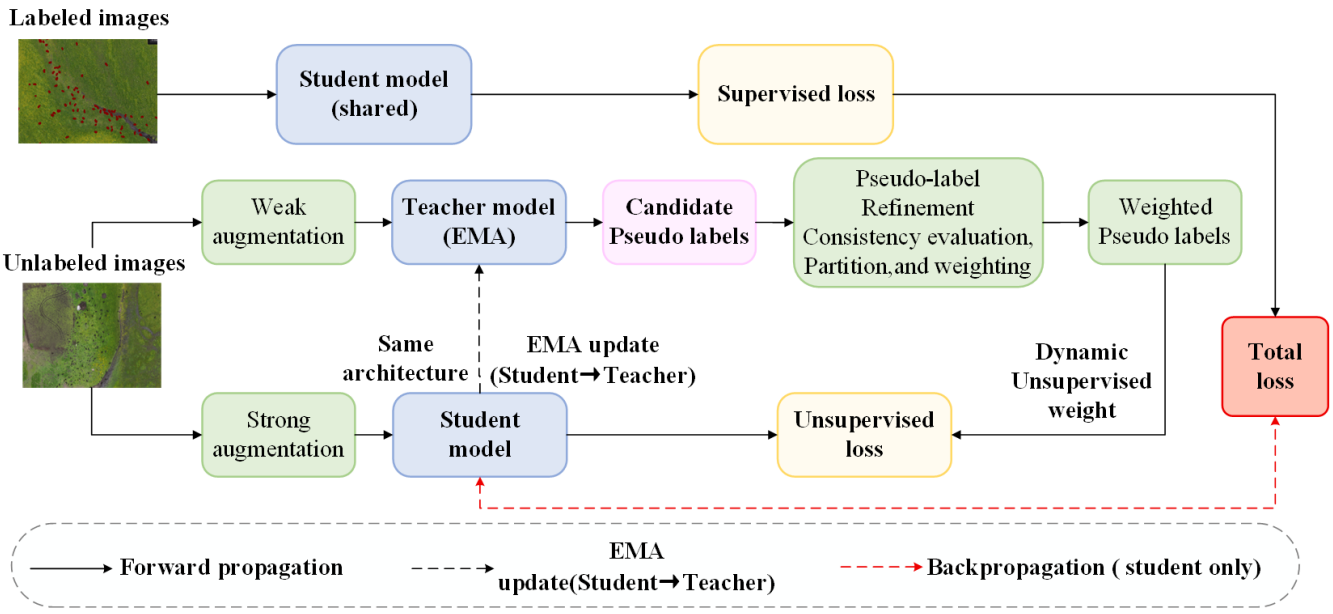


Figure 5. Teacher–student semi-supervised training framework. Different box colors visually distinguish the functional components of the framework. Solid black arrows indicate forward propagation, black dashed arrows indicate exponential moving average (EMA) updates from the student model to the teacher model, and red dashed arrows indicate backpropagation for updating the student model only.

The teacher and student models employ the same network architecture. Labeled images are augmented and directly fed into the student model, and the supervised loss is calculated based on the ground-truth categories and bounding boxes. For each unlabeled image, a weakly augmented view and a strongly augmented view are generated from the same original image. The weakly augmented view is fed into the teacher model to generate candidate pseudo-labels, whereas the strongly augmented view is fed into the student model for consistency learning. As described in Section 2.3.2, weak augmentation applies mild geometric and photometric perturbations to preserve target semantics and support stable pseudo-label generation, whereas strong augmentation introduces more intensive perturbations to improve the robustness of the student model.

The teacher model does not participate in gradient backpropagation, and its predictions are detached from the computational graph before being used as pseudo-label supervision. Instead, its parameters are updated using an exponential moving average:

$$\theta_T^{(t)} = \mu\theta_T^{(t-1)} + (1-\mu)\theta_S^{(t)} \quad (1)$$

where $\theta_T^{(t)}$ and $\theta_S^{(t)}$ denote the parameters of the teacher model and student model, respectively, at the t-th iteration, and μ denotes the EMA decay coefficient. Based on the sensitivity analysis presented in Section 4.5, μ is set to 0.999. By temporally averaging the student-model parameters across different training stages, EMA reduces prediction fluctuations induced by individual parameter updates and enables the teacher model to provide relatively stable classification confidence scores and bounding-box predictions for unlabeled samples.

Since the teacher and student models receive different augmented views, student predictions are first transformed back to the original image coordinate system according to the corresponding augmentation parameters. The localization consistency between teacher and student predictions is then evaluated through IoU-based matching. This operation eliminates the influence of geometric transformations and ensures that consistency is measured in the same spatial reference frame. After IoU-based localization consistency

evaluation, candidate predictions generated by the teacher model are further processed through hierarchical pseudo-label partitioning and reliability-weighted pseudo-label supervision. This process reduces the influence of localization-unstable predictions while retaining informative medium-confidence pseudo-labels with relatively stable localization, thereby improving the utilization of supervisory information from unlabeled data. A low IoU between teacher and student predictions is interpreted as localization instability caused by augmentation perturbations rather than simply classification uncertainty. Therefore, localization consistency provides complementary spatial reliability information beyond classification confidence. This consideration is particularly important for UAV livestock detection, where a considerable proportion of targets occupy relatively small pixel regions after image resizing. For small bounding boxes, a one-to-two pixel coordinate deviation may cause a substantial IoU variation, making localization consistency a more sensitive indicator of pseudo-label reliability than classification confidence alone. When strong augmentation causes severe cropping of a livestock target or removes the valid region required for reliable localization, the corresponding student prediction cannot be matched effectively with the teacher prediction. In such cases, the localization consistency score is assigned as zero, indicating that the prediction does not provide reliable spatial consistency information.

The student model is jointly optimized using the supervised loss and the weighted unsupervised loss:

$$\mathcal{L}_{total} = \mathcal{L}_{sup} + \lambda_u(t)\mathcal{L}_{unsup} \quad (2)$$

where $\mathcal{L}_{\text{sup}}$ denotes the supervised loss calculated from labeled data and $\mathcal{L}_{\text{unsup}}$ denotes the unsupervised loss calculated using weighted pseudo-labels, with individual pseudo-labels contributing according to their reliability weights. $\lambda_u(t)$ denotes the training-dependent unsupervised-loss coefficient used to regulate the overall contribution of unlabeled data at different training stages. Because the predictive capability of the teacher model and the quality of pseudo-labels are relatively unstable during the early stage of training, a progressive unsupervised-loss weighting strategy is adopted. The contribution of unlabeled supervision is therefore kept low during the early stage and gradually increased as training proceeds, reducing the influence of noisy early-stage pseudo-labels.

During each training iteration, the teacher model first generates candidate pseudo-labels from weakly augmented unlabeled images. The student model is then optimized using the supervised and weighted unsupervised losses, after which the teacher parameters are updated according to Equation (1). This procedure is repeated throughout semi-supervised training. After training, only the student model is retained for inference, whereas the teacher model and pseudo-label quality-control modules are used only during training. Therefore, DLCAPL introduces additional computation during training but does not add inference-time modules beyond the YOLOv13s student detector.

3.3. Dual-View Localization Consistency Module

Traditional semi-supervised object detection methods commonly select pseudo-labels based on classification confidence. However, classification confidence does not necessarily reflect the localization quality of predicted bounding boxes, and predictions with high classification scores may still exhibit substantial localization errors [30]. In UAV-based livestock imagery, targets are typically small and often appear in densely distributed scenes with partial occlusion, pose variations, and complex background interference, which can easily result in shifts in bounding-box centers or inaccurate boundary estimation. Due to the extremely limited pixel coverage of livestock targets, which occupy only approximately 6–17 pixels after resizing, even minor localization deviations may cause considerable IoU fluctuations and affect pseudo-label reliability evaluation. For extremely

small livestock targets, a displacement of only one or two pixels in the predicted bounding-box position may substantially change the IoU value, even when the predicted location remains visually close to the actual target region. Some predictions with considerable localization errors may still receive high classification confidence, whereas genuine small-scale or occluded targets may obtain only moderate confidence because of incomplete visual features. Therefore, classification confidence alone may be insufficient to characterize pseudo-label reliability, particularly for extremely small livestock targets whose IoU values are sensitive to minor spatial deviations.

We introduce the Dual-View Localization Consistency (DLC) module to provide a complementary reliability cue based on the spatial stability of predictions under different augmentations. DLC measures agreement between predictions across views; it does not directly estimate localization error against ground-truth boxes that are unavailable for unlabeled images. Figure 6 illustrates the module.

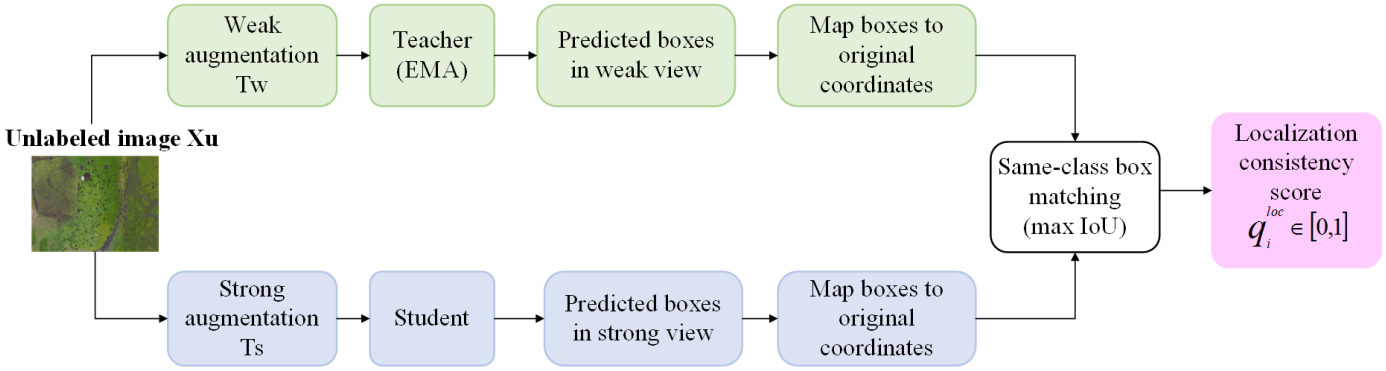


Figure 6. Dual-View Localization Consistency (DLC) module. Green boxes represent the weak-view teacher branch, blue boxes represent the strong-view student branch, and the pink box represents the localization consistency score. Black arrows indicate the data flow between the components.

For the same unlabeled image (x), a weak augmentation transformation (T_w) and a strong augmentation transformation (T_s) are applied to construct a weakly augmented view and a strongly augmented view, respectively. The weakly augmented view is fed into the teacher model to generate candidate pseudo-labels, whereas the strongly augmented view is fed into the student model to obtain the corresponding predictions for localization-consistency evaluation. To maintain a well-defined geometric correspondence between the two views, DLC matching is performed before applying CutMix augmentation. Therefore, only geometric transformations with recorded inverse mappings, including horizontal flipping, resizing, rotation, and translation, are considered in the coordinate mapping for localization-consistency evaluation. Appearance-only perturbations such as brightness adjustment, contrast adjustment, hue and saturation jitter, and Gaussian blur do not alter bounding-box coordinates and are therefore excluded from coordinate transformation.

Let the i-th predicted bounding box in the weakly augmented view be b_i^w and the j-th predicted bounding box in the strongly augmented view be b_j^s. After mapping them back to the original image coordinate system, the corresponding bounding boxes are expressed as

$$\tilde{b}_i^w = \mathcal{T}_w^{-1}(b_i^w), \qquad \tilde{b}_j^s = \mathcal{T}_s^{-1}(b_j^s) \tag{3}$$

For candidate predictions belonging to the same category, their pairwise localization consistency is defined as

$$q_{ij}^{\mathrm{loc}} = \mathrm{IoU}(\tilde{b}_i^w, \tilde{b}_j^s), \qquad q_{ij}^{\mathrm{loc}} \in [0,1] \tag{4}$$

where IoU(·) denotes the Intersection over Union between two predicted bounding boxes. A higher localization consistency score indicates that the same target maintains a stable center position and boundary extent under different augmentation perturbations, suggesting greater spatial reliability of the corresponding candidate pseudo-label. Conversely, a lower localization consistency score indicates that the prediction may be more susceptible to interference from background textures, target occlusion, or incomplete local features.

To avoid incorrect matching between different categories or adjacent targets in densely distributed livestock scenes, candidate boxes are first filtered according to their predicted categories, after which matching is performed only between predictions of the same category according to the highest-IoU matching criterion. For a candidate box b_i^w in the weakly augmented view, the index of its best-matching box in the strongly augmented view is defined as

$$j_i^* = \underset{j:c_i^w=c_j^s}{\operatorname{argmax}} \operatorname{IoU}\left(\tilde{b}_i^w, \tilde{b}_j^s\right) \tag{5}$$

where c_i^w and c_j^s denote the predicted categories of bounding boxes in the weakly and strongly augmented views, respectively. To ensure one-to-one correspondence, matching pairs are determined sequentially in descending order of IoU, and a predicted bounding box that has already been matched is excluded from subsequent matching. The final localization consistency score of the i-th candidate box in the weakly augmented view is expressed as

$$q_i^{\mathrm{loc}} = \operatorname{IoU}\left(\tilde{b}_i^w, \tilde{b}_{j_i^*}^s\right). \tag{6}$$

If no candidate box of the same category is available for matching, the prediction is regarded as unmatched, and its localization consistency score is set to 0.

Dual-view localization consistency and classification confidence provide complementary information. Classification confidence reflects the model's certainty regarding the predicted category, whereas localization consistency measures the spatial stability of a predicted bounding box under augmentation perturbations. By jointly considering these two types of information, predictions with high classification confidence but low cross-view localization consistency can be assigned lower reliability, whereas small-scale or occluded targets with moderate classification confidence but stable predictions across both views can receive higher reliability scores. Consequently, candidate predictions from unlabeled images are no longer evaluated solely according to a single classification score; instead, cross-view spatial verification is incorporated to provide an additional reliability criterion for subsequent pseudo-label partitioning and reliability-aware weak-label weighting. This design helps reduce the accumulation of noisy pseudo-labels and improves the utilization of difficult livestock targets.

3.4. Hierarchical Pseudo-Label Partitioning and Reliability-Aware Weak-Label Weighting

A single confidence threshold generally struggles to balance pseudo-label precision and recall. In UAV-based livestock imagery, distant, small-scale, and occluded targets often receive only moderate prediction confidence because of their weak visual features. A high confidence threshold may therefore discard potentially informative targets, whereas excessively lowering the threshold can introduce background false positives and noisy predictions with substantial localization uncertainty. To address this trade-off, this study introduces a fixed dual-threshold hierarchical pseudo-label strategy, where candidate predictions generated by the teacher model are first divided into different reliability levels according to classification confidence.

Specifically, a medium-confidence threshold and a high-confidence threshold are employed to partition candidate predictions into strong positive samples, weak positive

samples, and ignore regions. Strong positive samples provide reliable supervision directly, whereas weak positive samples are not discarded but further evaluated by jointly considering classification confidence and dual-view localization consistency. Based on this joint reliability estimation, continuous supervision weights are assigned to weak positive samples.

Therefore, the proposed strategy performs coarse-grained pseudo-label quality screening through confidence-based hierarchical partitioning, followed by fine-grained supervision adjustment through consistency-aware soft weighting. In this way, more difficult but informative targets can be retained while reducing the influence of unreliable pseudo-labels during training. The proposed hierarchical pseudo-label partitioning and consistency-aware weak-label weighting module is illustrated in Figure 7.

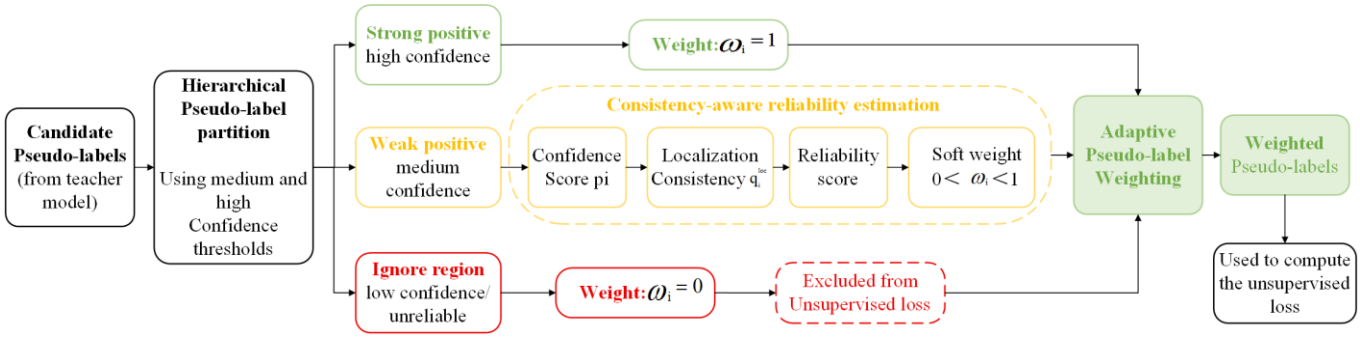


Figure 7. Hierarchical pseudo-label partitioning and reliability-aware weak-label weighting. Green, yellow, and red boxes denote strong-positive, weak-positive, and ignored regions, respectively. The dashed orange box represents reliability estimation based on classification confidence and localization consistency. Solid arrows indicate pseudo-label flow. Weights ($\omega_i = 1$), ($0 < \omega_i < 1$), and ($\omega_i = 0$) indicate full, partial, and zero contributions to the unsupervised loss, respectively.

To alleviate the difficulty of balancing pseudo-label precision and recall using a single confidence threshold, a medium-confidence threshold (τ_m) and a high-confidence threshold (τ_h) are introduced to distinguish different reliability levels of teacher-generated predictions, weak positive samples, and strong positive samples. Based on preliminary experiments conducted on the validation set, the medium- and high-confidence thresholds were determined as 0.20 and 0.50, respectively:

$$\tau_m = 0.20,\ \tau_h = 0.50 \tag{7}$$

The high-confidence threshold is used to select candidate targets with relatively reliable category predictions, which are treated as strong positive samples and provide full supervision. The medium-confidence threshold is used to retain predictions with relatively lower confidence but potentially useful learning information. Predictions whose confidence falls between the two thresholds are not directly discarded; instead, they are regarded as weak positive samples, and their reliability is further evaluated using dual-view localization consistency. Predictions below the medium-confidence threshold exhibit relatively high uncertainty and are assigned to ignore regions [31]. Accordingly, candidate predictions generated by the teacher model are partitioned into strong positive samples, weak positive samples, and ignore regions:

$$\tilde{y}_i = \begin{cases} y_{strong}, & p_i \geq \tau_h, \\ y_{weak}, & \tau_m \leq p_i < \tau_h, \\ y_{ignore}, & p_i < \tau_m, \end{cases} \tag{8}$$

where p_i denotes the classification confidence of the i-th candidate prediction. When p_i satisfies the high-confidence criterion, the prediction is classified as a strong positive sample with high category reliability and participates in training with the full supervision

weight. When p_i falls between the medium- and high-confidence thresholds, the prediction is retained as a weak positive sample. Such samples may contain a considerable proportion of distant, small-scale, or occluded targets and therefore require additional localization-stability information to determine their supervision strength. When p_i is below the medium-confidence threshold, the candidate prediction is considered highly uncertain and is assigned to the ignore region with a supervision weight of 0. Through differentiated supervision for pseudo-labels at different reliability levels, this hierarchical strategy improves pseudo-label utilization while reducing the influence of unreliable predictions during training.

For weak positive samples, classification confidence alone remains insufficient to fully characterize pseudo-label reliability 32. Therefore, classification confidence and dual-view localization consistency are further integrated to construct a comprehensive reliability score for each weak positive sample:

$$r_i = p_i^{\gamma}(q_i^{\mathrm{loc}})^{1-\gamma} \tag{9}$$

where p_i denotes the classification confidence of the i-th candidate prediction, q_i^{loc} represents the dual-view localization consistency score, and γ controls the relative contribution of classification confidence and spatial consistency. When a candidate box has relatively high classification confidence and remains spatially stable across both views, its comprehensive reliability score becomes higher. Conversely, when the predicted box exhibits low cross-view localization consistency under augmentation perturbations, its comprehensive reliability is reduced even if its classification confidence is relatively high. Since the weak-positive samples are initially selected according to the confidence interval defined by the medium and high thresholds, the reliability score is normalized within the same confidence range before being converted into supervision weights. This normalization provides a consistent mapping from the weak-positive reliability distribution to the supervision-weight interval, while preventing the confidence term from directly dominating the final weight assignment. Consequently, the contribution of dual-view localization consistency remains effective even for predictions with relatively high classification confidence.

Based on the comprehensive reliability score, supervision weights are further assigned to pseudo-labels at different hierarchical levels:

$$\omega_i = \begin{cases} 1, & \widetilde{y}_i \epsilon y_{strong} \\ [clip(\frac{r_i - \tau_m}{\tau_h - \tau_m + \varepsilon}, 0, 1)]^{\lambda}, & \widetilde{y}_i \epsilon y_{weak} \\ 0, & \widetilde{y}_i \epsilon y_{ignore} \end{cases} \tag{10}$$

where λ denotes the weak-label weight adjustment factor, ε is a small constant introduced to prevent numerical instability, and $\mathrm{clip}(\cdot, 0, 1)$ constrains the normalized reliability score to the interval of $[0, 1]$. The clipping operation is used only to bound the weight range and avoid numerical instability; it does not change the relative ranking among weak pseudo-labels before normalization. Strong positive samples are assigned the full supervision weight, whereas weak positive samples receive continuously varying soft weights according to their normalized comprehensive reliability scores derived from classification confidence and dual-view localization consistency.

Strong positive samples are assigned a fixed supervision weight of 1, whereas weak positive samples receive continuously varying reliability weights (ω_i) derived from classification confidence and dual-view localization consistency, as defined in Equation (10). The confidence thresholds are applied to candidate bounding boxes generated by the teacher model, rather than indiscriminately to all prediction locations of the dense detection head. After hierarchical selection, retained strong and weak pseudo-label boxes are assigned to the corresponding positive prediction locations of the student detector using

its positive-sample assignment strategy. Each matched positive prediction location inherits the reliability weight of its corresponding pseudo-label.

Ignore candidates are handled separately through an ignore mask and are excluded from the corresponding unsupervised pseudo-label supervision. Importantly, an unmatched prediction location is not automatically an ignored location. Unmatched locations outside the ignore mask remain eligible to contribute to the background classification loss, whereas locations excluded by the ignore mask do not contribute to the unsupervised classification loss. Neither ordinary background locations nor ignored locations receive pseudo-label-based bounding-box regression or Distribution Focal Loss (DFL) supervision. This distinction prevents uncertain teacher predictions from being incorrectly treated as background while preserving background supervision at valid unmatched locations.

The clipping operation bounds the normalized reliability term to the interval of [0, 1] before exponentiation by the weak-label adjustment factor (λ). Values below the lower bound are set to 0, whereas values above the upper bound are set to 1. The subsequent exponentiation preserves this bounded range while adjusting the contribution of intermediate reliability scores. In particular, when $\lambda > 1$, intermediate weights are reduced nonlinearly, whereas weights at the boundaries remain unchanged. Thus, the weighting mechanism limits the influence of unreliable weak pseudo-labels while retaining stronger supervision for predictions with higher estimated reliability.

The two-stage strategy first partitions candidates by confidence, retaining high-confidence predictions as strong positives, medium-confidence predictions as weak positives, and low-confidence or unreliable candidates as ignored. Reliability-aware weighting then differentiates supervision within the weak-positive group using both classification confidence and cross-view localization consistency. This separates candidate selection from supervision-strength assignment and is intended to preserve useful difficult-target information while limiting the effect of noisy predictions.

3.5. Joint Loss Function and Dynamic Unsupervised Loss Weight Scheduling

To enable collaborative learning from labeled and unlabeled data, the student model is jointly optimized using a supervised loss and a weighted unsupervised loss. Notably, DLC does not introduce an additional optimization objective; instead, it provides a localization reliability signal for pseudo-label weighting in the unsupervised branch. The supervised branch uses ground-truth annotations to provide stable supervision for category discrimination and bounding-box regression, whereas the unsupervised branch exploits pseudo-labels processed through hierarchical partitioning and reliability-aware weighting to further extract useful information from distant, small-scale, and occluded targets in unlabeled images. The reliability weights assigned to pseudo-labels are applied at the target level to regulate the contribution of individual candidate predictions during unsupervised optimization.

For labeled images, the supervised loss consists of a bounding-box regression loss, a classification loss, and a Distribution Focal Loss (DFL):

$$\mathcal{L}_{sup} = \lambda_{\text{box}}\mathcal{L}^{l}_{\text{box}} + \lambda_{\text{cls}}\mathcal{L}^{l}_{\text{cls}} + \lambda_{\text{dfl}}\mathcal{L}^{l}_{\text{dfl}} \quad (11)$$

where $\mathcal{L}^{l}_{\text{box}}$ constrains the positional and scale discrepancies between the predicted and ground-truth bounding boxes, $\mathcal{L}^{l}_{\text{cls}}$ is used for livestock category classification, and $\mathcal{L}^{l}_{\text{dfl}}$ denotes the Distribution Focal Loss for learning the discrete probability distribution of bounding-box locations 33. λ_{box}, λ_{cls}, and λ_{dfl} are balancing coefficients for the corresponding loss terms.

For unlabeled images, the unsupervised branch adopts the same detection loss formulation as the supervised branch, while the pseudo-label weights obtained in the

preceding section are used to modulate the supervision strength of individual candidate targets. Specifically, the reliability weight of each pseudo-label is multiplied with its corresponding detection loss components, including classification, bounding-box regression, and DFL losses. The weighted unsupervised loss is formulated as

$$\mathcal{L}_{pseudo} = \frac{\sum_{i=1}^{N_p} \omega_i (\lambda_{\text{box}} \ell_{\text{box}}^i + \lambda_{\text{cls}} \ell_{\text{cls}}^i + \lambda_{\text{dfl}} \ell_{\text{dfl}}^i)}{\sum_{i=1}^{N_p} \omega_i + \varepsilon} \tag{12}$$

where N_p denotes the number of candidate pseudo-labels retained after hierarchical processing in the current batch and ε is a small constant introduced to prevent division by zero. The loss terms ℓ_{box}^i, ℓ_{cls}^i, and ℓ_{dfl}^i denote the bounding-box regression loss, classification loss, and Distribution Focal Loss associated with the i-th candidate pseudo-label, respectively. Strong positive samples are assigned a weight of 1 and therefore provide full supervision. Weak positive samples receive continuous soft weights according to their classification confidence and localization consistency, whereas predictions in ignore regions are assigned a weight of 0 and thus contribute no effective gradient. They are not treated as background negatives in the unsupervised branch, preventing uncertain teacher predictions from introducing incorrect background constraints. Applying target-level reliability weights jointly to both classification and localization losses reduces the influence of unreliable pseudo-label supervision while retaining useful information from difficult targets.

Ordinary unmatched prediction locations outside the ignore mask remain eligible for background classification supervision. Let $\mathcal{N}_{bg}$ denote the set of these locations. The background classification loss is expressed as

$$\mathcal{L}_{bg} = \lambda_{\text{cls}} Reduce\left(\sum_{j \in \mathcal{N}_{bg}} \ell_{\text{cls}}^j\right) \tag{13}$$

where $Reduce$ denotes the loss reduction and normalization used in the implementation. The overall unsupervised loss is then expressed as

$$\mathcal{L}_{unsup} = \mathcal{L}_{pseudo} + \mathcal{L}_{bg} \tag{14}$$

Here, $\mathcal{N}_{bg}$ excludes positive prediction locations matched to selected pseudo-labels and locations masked as ignore regions. Ordinary unmatched locations contribute to background classification supervision but do not receive bounding-box regression or Distribution Focal Loss (DFL) supervision. Ignore locations are excluded from the corresponding unsupervised loss calculations and are not treated as background negatives.

During the early stage of training, the teacher model may generate relatively unstable pseudo-labels due to limited feature representation capability. Assigning a large weight to the unsupervised loss at this stage may cause erroneous pseudo-labels to be repeatedly learned by the student model, thereby leading to confirmation bias 34. Therefore, a progressive weight scheduling strategy is adopted:

$$\lambda_u(e) = \lambda_u^{\max} min(1, \frac{e}{E_{\text{ramp}}}) \tag{15}$$

where e denotes the relative training epoch within the semi-supervised joint-training stage, $\lambda_u^{\max}$ specifies the maximum coefficient of the unsupervised loss, and E_{ramp} determines the duration of the ramp-up period during which the unsupervised loss weight gradually increases.

In this study, these parameters are set as

$$\lambda_u^{\max} = 1.0, \; E_{\text{ramp}} = 20$$

The overall optimization objective of the student model is then defined as

$$\mathcal{L}_{total} = \mathcal{L}_{sup} + \lambda_u(e) \mathcal{L}_{unsup} \tag{16}$$

During the early training stage, the model relies primarily on ground-truth annotations to establish stable category discrimination and object localization capabilities. As training progresses, the teacher model gradually provides more reliable pseudo-label predictions through EMA-based parameter updating, and the weight of the unsupervised loss is progressively increased, allowing the supervisory information contained in unlabeled images to be effectively incorporated into model optimization. The supervised loss provides reliable constraints from ground-truth annotations, the pseudo-label reliability weights regulate the contribution of individual candidate targets according to their estimated reliability, and the dynamic weight scheduling strategy controls the overall contribution of the unsupervised branch at different training stages. The coordination of these three components reduces the adverse impact of unreliable pseudo-labels, improves the stability of semi-supervised optimization, and enhances the detection capability of the model for difficult livestock targets.

4. Experimental Results and Analysis

4.1. Experimental Setup and Evaluation Metrics

To ensure a fair and reproducible comparison among different methods and ablation experiments, all experiments were conducted using the same training/validation/test partition, training environment, input resolution, and basic optimization settings, except for method-specific modules and their corresponding hyperparameters. The dataset partition was fixed throughout all experiments, and no additional samples or annotations from the validation or test sets were used during model training. All experiments were performed on Windows 11 with an Intel® Core™ i9-14900K CPU, an NVIDIA GeForce RTX 4090 GPU, and 64 GB of system memory. Python 3.9 was used as the programming language, with PyTorch 2.0.0 as the deep learning framework and CUDA 11.8 for GPU acceleration.

Considering that livestock targets in UAV imagery are generally small and densely distributed, the model input resolution was uniformly set to 960 × 960 pixels to preserve sufficient spatial details while maintaining computational efficiency. The original UAV images had a resolution of 12,288 × 8192 pixels with a ground sampling distance (GSD) of approximately 0.8 cm/pixel. During training, the images were resized to 960 × 960 pixels, corresponding to a moderate downsampling ratio that balances the preservation of livestock appearance features and GPU memory consumption. After resizing, livestock targets remained represented by multiple pixels, allowing the model to retain essential localization information for small-object detection. After resizing to 960 × 960 pixels, livestock targets remained extremely small in the input images. Specifically, the average bounding-box dimensions of Tibetan sheep, yaks, and horses were approximately 6–8 pixels, 11–14 pixels, and 13–17 pixels, respectively. These limited spatial representations confirm that the dataset represents a typical small-object detection scenario, where accurate localization is highly sensitive to minor spatial deviations.

In addition to object size, livestock density was analyzed to characterize the challenge of dense detection scenarios. The number of livestock instances per image ranged from several individuals to more than 500, with an average of approximately 95 instances per image across the dataset. The maximum number of livestock instances in a single image reached 557, and 19 training images contained more than 300 livestock instances. A considerable proportion of images contained multiple adjacent livestock individuals, which increased the difficulty of individual localization and separation.

The model was first subjected to supervised pre-training for 30 epochs using the labeled data. For a fair comparison, the YOLOv13s-S30 supervised baseline was trained for the same total number of epochs using only labeled data, following the same optimization schedule and data augmentation strategy without the unlabeled branch. The resulting supervised pre-trained weights were then loaded for 120 epochs of semi-supervised joint

training. During this stage, the labeled and unlabeled branches were jointly optimized, while the teacher model was updated only through EMA-based parameter averaging without gradient backpropagation. During training, stochastic gradient descent (SGD) was employed to optimize the network parameters, with an initial learning rate of 0.005, a momentum coefficient of 0.937, and the weight decay coefficient set according to the original configuration. A cosine annealing strategy was adopted to dynamically adjust the learning rate, with a learning-rate warm-up applied during the first three epochs. During the semi-supervised training stage, the batch sizes for labeled and unlabeled data were both set to 2. All experiments used a fixed random seed of 42 unless otherwise specified. Hyperparameters, including pseudo-label thresholds and loss-weight coefficients, were selected based on preliminary validation experiments and then fixed for final evaluation on the test set.

During semi-supervised training, the EMA decay coefficient was set to 0.999. The teacher model was used only for pseudo-label generation during training and was not involved in final inference. For candidate pseudo-labels generated from unlabeled data, a hierarchical confidence-based filtering strategy was adopted to separate reliable pseudo-labels, weak positive samples, and ignored predictions according to their confidence levels. The medium- and high-confidence thresholds were set to 0.20 and 0.50, respectively, to partition the predictions into strong positive samples, weak positive samples, and ignore regions. For weak positive samples, the reliability balancing coefficient used in consistency-aware weighting was set to 0.5 to integrate classification confidence and dual-view localization consistency. The resulting reliability scores were used to assign target-level supervision weights during unsupervised optimization.

The main hyperparameters of DLCAPL were selected based on preliminary validation experiments and commonly adopted configurations in teacher–student semi-supervised detection frameworks. The EMA decay coefficient was set to 0.999 to provide stable teacher model updates. The pseudo-label thresholds (0.20 and 0.50) were selected by balancing pseudo-label quantity and reliability. The reliability balancing coefficient (0.5) and weak-label weight adjustment coefficient (2.0) were determined to achieve an appropriate contribution balance between classification confidence and localization consistency.

The weak-label weight adjustment coefficient was set to 2.0 to reduce the influence of low-reliability pseudo-labels on model optimization. During pseudo-label filtering, Non-Maximum Suppression (NMS) was applied to remove redundant bounding boxes, with the IoU threshold set to 0.60. The same NMS setting was maintained for all compared methods, pseudo-label generation, and final inference to ensure fair comparison. The maximum unsupervised loss weight was set to 1.0 and was linearly increased to this maximum value over the first 20 epochs after the start of semi-supervised joint training. This ramp-up strategy was applied only to the global contribution of the unsupervised branch and did not alter the target-level pseudo-label reliability weights. The detailed training parameters are listed in Table 2.

Table 2. Experimental training parameter settings.

Parameter	Setting
Input resolution	960 × 960
Optimizer	SGD
Initial learning rate	0.005
Momentum	0.937
Weight decay	5×10^{-4}
Learning rate scheduler	Cosine annealing
Warm-up epochs	3
Supervised pre-training epochs	30

Semi-supervised training epochs	120
Batch size (labeled/unlabeled)	2/2
EMA decay coefficient	0.999
Pseudo-label thresholds	0.20/0.50
NMS IoU threshold	0.60
Maximum unsupervised loss weight	1.0
Unsupervised loss warm-up epochs	20

To comprehensively evaluate the object recognition and bounding-box localization performance of the models, Precision (P), Recall (R), Average Precision (AP), and mean Average Precision (mAP) were adopted as evaluation metrics. Precision represents the proportion of correctly detected livestock targets among all predictions classified as livestock, whereas Recall represents the proportion of ground-truth livestock targets that are correctly detected by the model. They are calculated as follows:

$$P = \frac{TP}{TP + FP} \tag{17}$$

$$R = \frac{TP}{TP + FN} \tag{18}$$

where TP denotes correctly detected targets whose predicted category and bounding box satisfy the specified IoU matching criterion, FP denotes predicted targets that do not correspond to any ground-truth object, and FN denotes ground-truth targets that are missed by the detector.

Average Precision represents the area under the precision–recall curve for a given class and can theoretically be expressed as

$$AP_c = \int_0^1 P_c(r) d\, r \tag{19}$$

where AP_c denotes the Average Precision of the c-th object category and $P_c(r)$ represents the precision value at recall level (r). The mean Average Precision is calculated as the arithmetic mean of the AP values over all categories:

$$mAP = \frac{1}{C}\sum_{c=1}^{C} AP_c \tag{20}$$

where C denotes the number of detection categories. In this study, three livestock categories were considered: yak, Tibetan sheep, and horse.

The performance of the models was primarily evaluated using P, R, $mAP@0.5$, and mAP@0.5: 0.95 **Error! Reference source not found.**. Specifically, mAP@0.5 denotes the mean AP across all categories when the IoU threshold is fixed at 0.50 and primarily reflects the overall detection performance. In contrast, $mAP@0.5{:}\,0.95$ follows the COCO-style evaluation protocol, in which AP is calculated at IoU thresholds ranging from 0.50 to 0.95 with an interval of 0.05 and then averaged across both IoU thresholds and object categories. This metric imposes a more stringent requirement on bounding-box localization accuracy. Taken together, these metrics enable a comprehensive evaluation of different methods in terms of classification accuracy, missed detections, false detections, and bounding-box localization quality.

4.2. Comparative Experiments

To evaluate the effectiveness of DLCAPL under limited-label conditions, we compared it with a supervised baseline, conventional pseudo-label-based self-training, and representative teacher–student semi-supervised object detection methods. Except for YOLOv13s-Full, all methods used the same 30% labeled training subset. The semi-supervised methods additionally utilized the remaining 70% of the training images as unlabeled data without manual annotations. YOLOv13s-S30 was trained exclusively on the labeled

subset and served as the supervised baseline under limited-label conditions. All methods used YOLOv13s as the base detector and shared the same dataset partitions, input resolution, and common training settings wherever applicable. Method-specific learning mechanisms and associated hyperparameters were retained or adapted as necessary to integrate the methods into the common detection framework. The validation set was used for model selection and hyperparameter tuning, whereas the independent test set was reserved for final performance evaluation. Unless otherwise specified, all quantitative results reported in this section were obtained on the test set.

The compared methods included YOLOv13s-S30, Pseudo-Label 36, STAC-style 37, EMA Teacher–Student, Unbiased Teacher v2-style **Error! Reference source not found.**, Soft Teacher-style and DLCAPL, in addition to the YOLOv13s-S30 and YOLOv13s-Full reference models. Pseudo-Label adopts a self-training strategy in which predictions from a detector trained on labeled images are filtered according to a confidence threshold, and the retained predictions are used as pseudo-labels for subsequent training. STAC-style combines pseudo-label-based self-training with strong data augmentation, using pseudo-labels to supervise the detector on strongly augmented unlabeled images. The EMA Teacher–Student baseline updates the teacher parameters through the exponential moving averaging (EMA) of the student parameters, providing teacher predictions for unsupervised supervision.

Unbiased Teacher v2-style was adapted to the YOLOv13s framework to incorporate the original method's approach to controlling unreliable pseudo-label supervision in bounding-box regression. Soft Teacher-style uses teacher predictions to supervise unlabeled images, applies confidence-based weighting to classification loss, and employs box-jitter-based reliability estimation for regression pseudo-labels. These methods were implemented as adaptations to a common detection framework rather than exact reproductions of their original detector architectures and complete training pipelines. The suffix "-style" therefore indicates that the defining mechanisms were adapted to YOLOv13s, with architecture-related and implementation-specific differences acknowledged.

Building upon the teacher–student framework, DLCAPL integrates dual-view localization consistency (DLC), hierarchical pseudo-label partitioning, and reliability-aware weak-label weighting. Unlike approaches that rely primarily on classification confidence or perturbation-based pseudo-box reliability estimation, DLC evaluates the spatial consistency of teacher and student predictions across weakly and strongly augmented views after mapping the predictions to a common coordinate frame. The hierarchical strategy distinguishes high-confidence and medium-confidence pseudo-labels, while reliability-aware weighting adjusts the contribution of weak pseudo-labels according to their classification confidence and localization consistency. These components are designed to retain informative predictions for difficult livestock targets while reducing the influence of unreliable pseudo-label supervision. The main comparison results are presented in Table 3.

Table 3. Comparison of detection performance among different methods on the test set. Note: All results were obtained on the test set. Semi-supervised methods used 30% labeled data and 70% unlabeled data during training.

Model	Labeled (%)	Unlabeled	P (%)	R (%)	mAP@0.5 (%)	mAP@0.5:0.95 (%)
YOLOv13s-S30	30		58.23	52.17	55.32	34.25
Pseudo-Label	30	√	62.14	55.06	58.74	37.21
STAC-style	30	√	66.38	58.47	62.31	40.36
EMA Teacher–Student	30	√	70.26	61.35	65.48	42.91
Unbiased Teacher v2-style	30	√	73.84	64.22	68.16	45.27
Soft Teacher-style	30	√	77.56	67.31	70.92	47.68
DLCAPL (Ours)	30	√	82.13	70.46	73.28	50.68

YOLOv13s-Full	100	80.23	72.54	73.91	52.45

The experimental results show that all semi-supervised methods improved detection performance to varying degrees under limited-label conditions. YOLOv13s-S30, trained using only 30% labeled data, achieved an mAP@0.5 of 55.32% and an mAP@0.5: 0.95 of 34.25%. After incorporating pseudo-label learning, strong-augmentation consistency, and EMA-based teacher models, the semi-supervised methods achieved different levels of performance improvement. These results indicate that unlabeled images can provide additional supervisory signals, while effective pseudo-label generation and quality-control strategies are essential for extracting useful information from unlabeled data.

The proposed DLCAPL achieved the best performance among the semi-supervised methods, with a Precision of 82.13%, Recall of 70.46%, mAP@0.5 of 73.28%, and mAP@0.5: 0.95 of 50.68%. Compared with YOLOv13s-S30, these four metrics increased by 23.90, 18.29, 17.96, and 16.43 percentage points, respectively. This demonstrates that DLCAPL can more effectively exploit unlabeled data under limited-label conditions by improving pseudo-label reliability and reducing the adverse influence of inaccurate supervision.

Compared with Soft Teacher-style, the strongest competing semi-supervised method in this comparison, DLCAPL improved precision and recall by 4.57 and 3.15 percentage points, respectively, and increased mAP@0.5 and mAP@0.5:0.95 by 2.36 and 3.00 percentage points, respectively. These results indicate that DLCAPL achieved better overall detection performance under both the IoU = 0.50 evaluation criterion and the stricter evaluation averaged across IoU thresholds from 0.50 to 0.95. However, mAP@0.5:0.95 reflects detection performance across multiple IoU thresholds and does not independently isolate bounding-box localization accuracy. Therefore, the observed improvement should be interpreted as an overall detection-performance gain rather than direct evidence of improved localization accuracy. The proposed dual-view localization consistency module assesses prediction stability across augmented views, while the consistency-aware weighting strategy adjusts the contribution of weak pseudo-labels during optimization.

Compared with YOLOv13s-Full trained using 100% labeled data, DLCAPL achieved $mAP@0.5$ and $mAP@0.5{:}\,0.95$ values that were only 0.63 and 1.77 percentage points lower, respectively, while Recall was 2.08 percentage points lower and Precision was 1.90 percentage points higher. These results demonstrate that, by using only 30% manually labeled data, together with 70% unlabeled data, DLCAPL can achieve overall detection performance close to that of the fully supervised model, highlighting the potential of semi-supervised learning to reduce manual annotation requirements.

To further evaluate the robustness of DLCAPL under different annotation budgets, labeled-data ratios of 10%, 20%, and 30% were considered. For each setting, the selected labeled subset was used for supervised learning, while the remaining training images were treated as unlabeled data without additional annotations. For each setting, the corresponding proportion of training images was selected as labeled data, while the annotations of the remaining training images were hidden during training and these images were treated as unlabeled data for DLCAPL. Accordingly, the 10%, 20%, and 30% labeled-data settings corresponded to 90%, 80%, and 70% unlabeled training data, respectively. All experiments followed the same data partitioning principles and training configurations, and all training images not selected as labeled data were used as unlabeled data during semi-supervised training. The results are presented in Table 4.

Table 4. Detection performance under different labeled-data ratios on the test set.

Model	Labeled Ratio (%)	P (%)	R (%)	mAP@0.5 (%)	mAP@0.5:0.95 (%)
YOLOv13s	10	45.76	39.54	42.31	24.68

	20	52.14	46.23	49.47	29.56
	30	58.23	52.17	55.32	34.25
DLCAPL (Ours)	10	69.42	59.36	63.18	41.72
	20	76.28	65.36	68.62	46.35
	30	82.13	70.46	73.28	50.68
YOLOv13s-Full	100	80.23	72.54	73.91	52.45

As shown in Table 4, the detection performance of both the YOLOv13s supervised baseline and DLCAPL improved as the proportion of labeled data increased, indicating that additional labeled samples improve the model's ability to learn category-specific features and localization patterns. Under the same labeled-data ratio, DLCAPL consistently outperformed the corresponding supervised baseline across all evaluation metrics, demonstrating its ability to effectively utilize unlabeled imagery and mitigate the performance degradation caused by insufficient supervision.

Further analysis of the results under different labeled-data ratios shows that DLCAPL consistently outperformed the supervised baseline under the 10%, 20%, and 30% labeled-data settings, indicating that the proposed method can effectively exploit latent supervisory information contained in unlabeled UAV imagery. Under low-label conditions, the limited number of ground-truth annotations restricts the model's ability to learn discriminative target features and spatial distributions; therefore, pseudo-label supervision derived from unlabeled data provides more substantial performance gains. As the proportion of labeled data increases, the supervised baseline receives increasingly sufficient supervisory information, and the relative improvement provided by DLCAPL gradually decreases. Specifically, the improvement in $mAP@0.5$ decreased from 20.87 percentage points under the 10% labeled-data setting to 17.96 percentage points under the 30% setting, while the improvement in $mAP@0.5{:}\,0.95$, decreased from 17.04 to 16.43 percentage points.

This trend indicates that, when labeled information is limited, the proposed pseudo-label quality-control mechanism can effectively compensate for insufficient manual annotations by extracting complementary supervisory information from unlabeled images. As the amount of labeled data increases, supervision from ground-truth annotations gradually becomes dominant, resulting in a reduced additional gain from semi-supervised learning. Under the 30% labeled-data setting, DLCAPL achieved an $mAP@0.5$ of 73.28% and an $mAP@0.5{:}\,0.95$ of 50.68%, showing that semi-supervised learning can substantially reduce annotation requirements while maintaining comparable detection performance.

Since livestock categories differ in visual characteristics and sample availability, class-wise detection performance was evaluated on the independent test set, as reported in Table 5. The test set contains 3868 yak instances, 1707 Tibetan sheep instances, and only 22 horse instances, reflecting a substantial class imbalance. Class-wise precision and recall were evaluated using class-specific confidence thresholds selected on the validation set, whereas the overall precision and recall were evaluated at the global operating point selected on the validation set. These different operating points should be considered when interpreting the class-wise and overall precision and recall values. Average precision (AP) for each class was calculated from the test-set precision–recall curve, and the overall mAP was calculated as the mean of the class-wise AP values.

Table 5. Class-wise detection performance of DLCAPL on the test set under the 30% labeled-data setting. Note: Class-wise precision and recall are reported at class-specific confidence thresholds selected on the validation set, whereas overall precision and recall are reported at the global operating point selected on the validation set. AP values are calculated from the test-set precision–recall curves.

Class	Number of Instances	Precision (%)	Recall (%)	mAP@0.5 (%)	mAP@0.5:0.95 (%)
Yak	3868	83.46	72.15	75.12	51.84
Tibetan sheep	1707	80.72	67.34	72.46	49.26
Horse	22	74.58	62.05	72.26	50.94
Overall performance	5597	82.13	70.46	73.28	50.68

The class-wise results provide a more detailed view of DLCAPL's detection performance across livestock categories. However, the small number of horse instances in the test set limits the strength of conclusions that can be drawn about performance on this category. Therefore, the horse-specific metrics should be interpreted with caution and should not be regarded as definitive evidence of generalization to underrepresented livestock categories.

To further evaluate the practical applicability of DLCAPL for UAV-based livestock monitoring, the computational complexity and inference efficiency of the final detection model were analyzed. Since the teacher model, dual-view localization consistency evaluation, hierarchical pseudo-label partitioning, and reliability-aware weighting modules are only involved during the training stage, they are discarded after model optimization. Therefore, only the student detector is retained during inference, and no additional semi-supervised modules are introduced during deployment. The number of parameters, floating-point operations (FLOPs), inference latency, and frames per second (FPS) were measured under the same hardware environment and input resolution (960 × 960 pixels). The results are presented in Table 6.

Table 6. Computational efficiency comparison of inference models under the same hardware and input resolution.

Model	Parameters (M)	FLOPs (G)	Inference Latency (ms/Image)	FPS
YOLOv13s-S30	8.92	16.41	14.2	70.4
DLCAPL (student model)	8.92	16.41	14.4	69.4
YOLOv13s-Full	8.92	16.41	14.6	68.5

As shown in Table 6, YOLOv13s-S30, DLCAPL (student model), and YOLOv13s-Full have identical parameter counts (8.92 M) and computational complexity (16.41 G FLOPs) under the same input resolution of 960 × 960 pixels. Their measured inference latencies were 14.2, 14.4, and 14.6 ms per image, respectively, corresponding to 70.4, 69.4, and 68.5 FPS. The small differences in measured inference speed may be attributable to runtime fluctuations during GPU inference. Since the teacher model, dual-view localization consistency evaluation, hierarchical pseudo-label partitioning, and reliability-aware weighting modules are discarded after training, they introduce no additional computational overhead during deployment. These results indicate that DLCAPL improves detection performance under limited-label conditions without increasing the parameter count or theoretical computational complexity of the deployed student detector.

4.3. Ablation Experiments and Analysis

To evaluate the incremental contributions and complementary effects of the individual components of DLCAPL, YOLOv13s trained with 30% labeled data was adopted as

the baseline. The EMA teacher–student framework, dual-view localization consistency (DLC), hierarchical pseudo-label partitioning, and reliability-aware weak-label weighting were then progressively incorporated. Different combinations of these components were evaluated to analyze their effects on detection performance. Except for the component under investigation, all experiments used the same data split, training parameters, and evaluation criteria. Each ablation variant was defined by progressively activating different components of DLCAPL. Specifically, the experiments separately evaluated the contributions of EMA teacher updating, dual-view localization consistency, hierarchical pseudo-label partitioning, and reliability-aware weak-label weighting. In addition, a weighting-only variant without DLC was introduced to distinguish the contribution of localization consistency from confidence-based weighting. The results are presented in Table 7.

Table 7. Ablation study of individual components in DLCAPL on the test set. Note: YOLOv13s-S30 denotes the supervised baseline trained using 30% labeled data. EMA Teacher represents the teacher–student baseline without the DLCAPL-specific modules. The "+DLC" variant incorporates dual-view localization consistency into the EMA Teacher baseline. "+Hierarchical Pseudo-label Partition" introduces the hierarchical pseudo-label selection strategy. "Reliability-aware Weak-label Weighting" without DLC denotes confidence-based weak-label weighting without localization consistency. The additional control variants evaluate the effect of removing DLC while retaining the other components and replacing DLC with box-jitter variance as an alternative localization reliability measure. The complete DLCAPL model incorporates all proposed components. Boldface indicates the selected control and complete-model configurations.

Modules					P (%)	R (%)	mAP@0.5 (%)	mAP@0.5:0.95 (%)
YOLOv13s	EMA	DLC	Hierarchical Pseudo-Label Partition	Reliability-Aware Weak-Label Weighting				
√					58.23	52.17	55.32	34.25
√	√				70.26	61.35	65.48	42.91
√	√	√			74.32	64.18	68.27	45.36
√	√		√		73.45	65.06	67.81	44.92
√	√		√	√	75.62	66.83	69.65	46.91
√	√	√	√		78.64	68.29	71.35	48.26
√	√		√	√	77.02	67.11	69.77	47.03
√	√	box-jitter	√	√	79.45	68.30	71.64	48.71
√	√	√	√	√	82.13	70.46	73.28	50.68

As shown in Table 7, the progressive ablation experiments demonstrate that each component contributes to improved detection performance to varying degrees. The YOLOv13s baseline trained with only 30% labeled data achieved an $mAP@0.5$ of 55.32% and an $mAP@0.5{:}\,0.95$ of 34.25%. After introducing the EMA teacher–student framework, these two metrics increased to 65.48% and 42.91%, corresponding to improvements of 10.16 and 8.66 percentage points over the baseline, respectively. By temporally smoothing the student model parameters to construct the teacher model, EMA reduces short-term fluctuations in model parameters and predictions during training, thereby providing more stable pseudo-label supervision for unlabeled data. This is consistent with the general teacher–student paradigm, where temporal averaging of model parameters is used to obtain more stable teacher predictions.

Further incorporating DLC into the EMA framework increased precision and recall from 70.26% and 61.35% to 74.32% and 64.18%, respectively, while $mAP@0.5$ and $mAP@0.5{:}\,0.95$ increased to 68.27% and 45.36%, representing further improvements of 2.79 and 2.45 percentage points over the EMA framework. These results indicate that incorporating localization stability provides an additional criterion for assessing pseudo-label quality beyond classification confidence. Therefore, DLC complements confidence-

based pseudo-label selection by providing an additional localization reliability cue. Previous studies on semi-supervised object detection have also indicated that classification scores alone cannot adequately guarantee the localization accuracy of pseudo-boxes, whereas regression consistency can provide complementary information for evaluating pseudo-box localization quality.

When hierarchical pseudo-label partitioning was introduced independently into the EMA framework, recall increased from 61.35% to 65.06%, while $mAP@0.5$ and $mAP@0.5{:}\,0.95$ reached 67.81% and 44.92%, respectively. The relatively pronounced improvement in recall indicates that partitioning candidate predictions into strong positives, weak positives, and ignored regions using fixed medium and high confidence thresholds prevents moderately confident candidate targets from being directly discarded, thereby expanding the range of difficult targets available for training. To further distinguish the contribution of localization consistency from confidence-based weighting, a weighting-only variant without DLC was evaluated. In this configuration, weak pseudo-labels were assigned confidence-based weights without incorporating localization consistency information. This variant achieved an mAP@0.5 of 69.65% and an mAP@0.5:0.95 of 46.91%, which improved over the EMA teacher baseline. These results indicate that confidence-based weighting can reduce the influence of unreliable pseudo-labels by assigning different supervision strengths to candidate predictions. However, compared with DLC-based configurations, the performance remained lower, demonstrating that localization consistency provides additional spatial reliability information beyond classification confidence.

When DLC and hierarchical pseudo-label partitioning were incorporated simultaneously, precision and recall further increased to 78.64% and 68.29%, respectively, while $mAP@0.5$ and $mAP@0.5{:}\,0.95$ reached 71.35% and 48.26%. Compared with the configuration using DLC alone, the two mAP metrics were further improved by 3.08 and 2.90 percentage points, respectively; compared with the configuration using hierarchical pseudo-label partitioning alone, the corresponding improvements were 3.54 and 3.34 percentage points. These results suggest that hierarchical pseudo-label partitioning primarily expands the utilization of moderately confident candidate targets, whereas DLC provides information on spatial localization stability. The two mechanisms are therefore complementary in terms of candidate-target utilization and supervision reliability control.

Two control groups are designed in Table 7 to further verify the effectiveness of the DLC module. In the first group, hierarchical pseudo-label partitioning and reliability-aware weak-label weighting are fully retained while DLC is removed. The performance drops to 69.77% (mAP@0.5) and 47.03% (mAP@0.5:0.95), which confirms that performance gains are not merely derived from the confidence-based weighting strategy. In the second group, DLC is substituted with box-jitter variance, an existing localization reliability metric from Soft Teacher, with all other components unchanged. This setup yields mAP@0.5 of 71.64% and mAP@0.5:0.95 of 48.71%. Although box-jitter variance brings performance benefits, it still underperforms our DLC module, proving the superiority of dual-view localization consistency for this livestock-detection task.

On this basis, further introducing reliability-aware weak-label weighting yielded the complete DLCAPL model, achieving a precision of 82.13%, a recall of 70.46%, an $mAP@0.5$ of 73.28%, and an $mAP@0.5{:}\,0.95$ of 50.68%. Compared with the combination of DLC and hierarchical pseudo-label partitioning without weak-label weighting, the two mAP metrics were further improved by 1.93 and 2.42 percentage points, respectively, with a relatively larger improvement in $mAP@0.5{:}\,0.95$. This strategy calculates the overall reliability of weak positives by jointly considering classification confidence and dual-view localization consistency and assigns continuous supervision weights according to pseudo-label reliability, thereby reducing the contribution of uncertain predictions while

preserving useful supervision from reliable weak pseudo-labels. As a result, reliable weak pseudo-labels retain stronger supervision, whereas the gradient contributions of low-reliability weak pseudo-labels are reduced.

Overall, EMA provides stable teacher predictions for unlabeled data, DLC supplements pseudo-label evaluation with localization stability information, hierarchical pseudo-label partitioning expands the utilization of difficult targets with moderate confidence, and reliability-aware weak-label weighting further regulates the supervision contributions of different weak pseudo-labels.

To further validate the effectiveness of DLC as a localization reliability estimation strategy, a box-jitter variance-based reliability estimation method inspired by Soft Teacher was introduced for comparison. Unlike DLC, which evaluates prediction stability across weakly and strongly augmented views, box-jitter variance estimates localization reliability by measuring the variation in perturbed bounding boxes. Both strategies were incorporated into the same pseudo-label weighting framework for a controlled comparison, with only the localization reliability estimation strategy replaced and the other DLCAPL components kept unchanged. Note that the Soft Teacher-style baseline in Table 3 incorporates adapted Soft Teacher mechanisms within the common YOLOv13s framework, whereas the box-jitter variant in Table 8 replaces only the DLC module while retaining the other DLCAPL components. Therefore, the two results correspond to different experimental configurations and should not be interpreted as identical implementations. The results are shown in Table 8.

Table 8. Comparison of localization reliability estimation strategies on the test set.

Localization Reliability Strategy	mAP@0.5 (%)	mAP@0.5:0.95 (%)
Box-jitter variance-based estimation	71.64	48.71
Dual-view localization consistency (DLC)-based estimation	73.28	50.68

As shown in Table 8, dual-view localization consistency (DLC) achieved higher detection performance than box-jitter variance-based estimation under the same pseudo-label weighting framework. Replacing box-jitter variance with DLC improved mAP@0.5 from 71.64% to 73.28% and mAP@0.5:0.95 from 48.71% to 50.68%, corresponding to gains of 1.64 and 1.97 percentage points, respectively. These results suggest that cross-view localization consistency is a more effective alternative to box-jitter variance for improving overall detection performance within the evaluated DLCAPL framework. However, these results do not directly establish that DLC provides a more accurate estimate of absolute localization quality.

Since the primary motivation of semi-supervised learning is to reduce annotation requirements by exploiting readily available unlabeled UAV imagery, the influence of unlabeled data scale on DLCAPL was further investigated. In these experiments, the labeled subset was fixed to 145 images, while the number of unlabeled images was gradually increased from 145 to 337. All experiments were conducted using the same training configuration, dataset partition strategy, and random seed, with only the amount of unlabeled data changed. The results are presented in Table 9.

Table 9. Effect of unlabeled data scale on DLCAPL performance under fixed labeled-data conditions.

Labeled Images	Unlabeled Images	Unlabeled Ratio (%)	Precision (%)	Recall (%)	mAP@0.5 (%)	mAP@0.5:0.95 (%)
145	145	30	76.42	64.18	67.35	45.22
145	241	50	79.27	67.21	70.14	47.93
145	337	70	82.13	70.46	73.28	50.68

As shown in Table 9, DLCAPL consistently benefits from increasing the amount of unlabeled data. When the number of unlabeled images increased from 145 to 337, mAP@0.5 improved from 67.35% to 73.28%, while mAP@0.5:0.95 increased from 45.22% to 50.68%. This improvement indicates that additional unlabeled UAV images provide more diverse target appearances and background variations, allowing the teacher model to generate more informative pseudo-label supervision and enabling the student model to learn richer target representations. The results further demonstrate the scalability of DLCAPL for practical livestock monitoring scenarios, where UAV imagery can be collected efficiently while manual annotation remains expensive.

4.4. Visualization Analysis

To further compare the training convergence of different methods, the validation-set *mAP*@0.5 curves of YOLOv13s-S30, Pseudo-Label, STAC-style, EMA Teacher–Student, Unbiased Teacher v2-style, Soft Teacher-style, and DLCAPL were plotted, as shown in Figure 8. The validation set was only used for monitoring convergence behavior and did not participate in final performance evaluation. All methods employed the same dataset split and basic training configurations. The first 30 epochs corresponded to the supervised pre-training stage, followed by semi-supervised joint training from epochs 31 to 150.

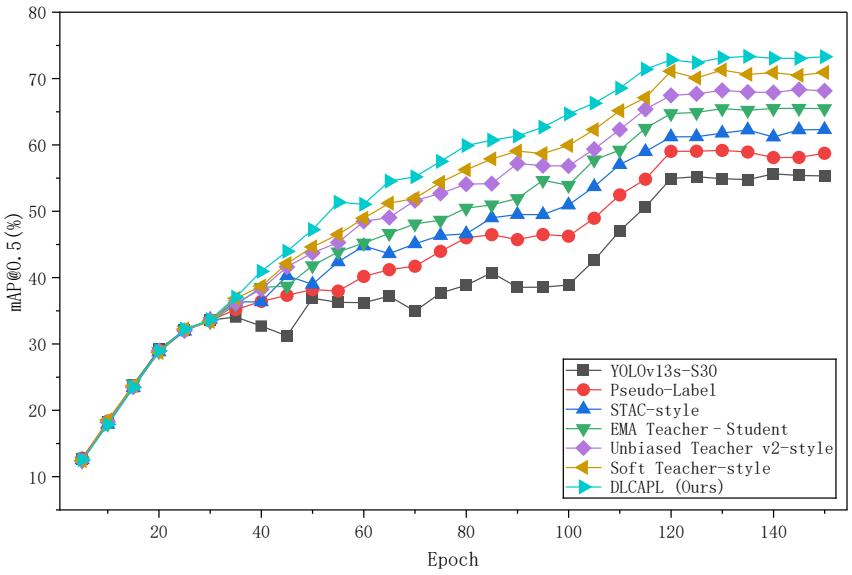


Figure 8. Convergence curves of different methods in terms of mAP@0.5 on the validation set.

As shown in Figure 8, during the supervised pretraining stage, all semi-supervised methods shared the same initialization model and therefore exhibited generally similar performance trends. After entering the semi-supervised joint training stage, the curves of different methods gradually diverged. Since YOLOv13s-S30 relies only on limited labeled data, its performance fluctuates more noticeably during training, and its final performance remains relatively low. By contrast, after introducing pseudo-label learning, strong-augmentation consistency, and the EMA teacher model, the convergence behavior of the model is progressively improved, indicating that unlabeled data and stable teacher predictions can provide effective supervision for model optimization.

Compared with other semi-supervised methods, DLCAPL maintains a consistently higher *mAP*@0.5 during the joint training stage and gradually stabilizes at a superior level in the later stage of training. This result demonstrates that the dual-view localization consistency strategy can reduce the influence of pseudo-labels with unstable localization,

while pseudo-label partitioning and reliability-aware weak-label weighting can further control noisy supervision, thereby improving the utilization efficiency of unlabeled data and enhancing training convergence stability.

To further provide an intuitive comparison of the detection performance of different methods in complex grassland scenarios, several representative scenes were selected from the test set for visualization analysis, including small-scale targets, densely distributed targets, partial occlusion, and complex backgrounds, as shown in Figure 9. Figure 9 presents, in order, the ground-truth annotations and the detection results of YOLOv13s-S30, Soft Teacher-style, and DLCAPL. All methods use the same confidence threshold and non-maximum suppression parameters to ensure a fair comparison.

As shown in Figure 9, notable differences in detection performance can be observed among the different methods under challenging conditions, including distant small-scale targets, densely distributed targets, partial occlusion, and complex backgrounds. Owing to the limited amount of labeled data and the weak visual features of livestock targets, YOLOv13s-S30 still exhibits missed detections, false detections, and bounding-box localization deviations to a certain extent. Soft Teacher-style improves the detection of some challenging targets; however, limitations remain in separating adjacent targets and localizing occluded targets.

In comparison, DLCAPL achieves more complete target detection and more accurate bounding-box localization across various complex scenarios. For distant small-scale targets, the proposed method successfully retains more valid targets. In densely distributed and partially occluded scenes, it more effectively distinguishes adjacent individuals, thereby reducing missed detections and localization deviations. In regions where the target appearance is similar to the grassland background, background-related false detections are also reduced. These results indicate that dual-view localization consistency can enhance the spatial reliability of pseudo-labels, while hierarchical pseudo-label partitioning and reliability-aware weak-label weighting help retain reliable challenging samples and suppress interference from noisy pseudo-labels.

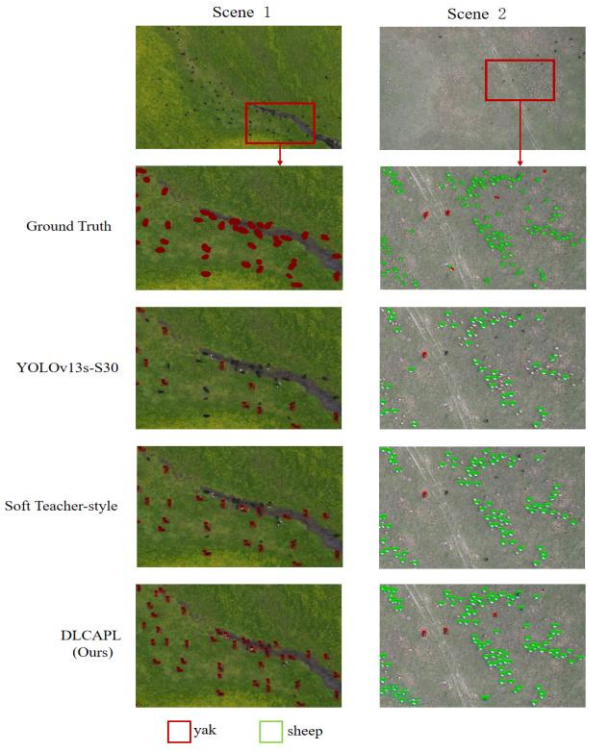


Figure 9. Comparison of detection results of different methods on the self-constructed UAV grassland livestock dataset.

https://doi.org/10.3390/drones10110804

Taken together, the results in Figures 8 and 9 demonstrate that DLCAPL not only exhibits favorable convergence stability during semi-supervised training but also achieves robust detection performance under challenging conditions involving small-scale targets, dense distributions, occlusion, and complex backgrounds. The visualization results are generally consistent with the findings of the preceding comparative and ablation experiments, further validating the effectiveness of the proposed method for livestock detection in complex UAV grassland scenarios.

4.5. Sensitivity Analysis

To investigate the influence of key hyperparameters on the performance of DLCAPL, sensitivity analyses were conducted on the validation set. Since these hyperparameters were determined before final evaluation, the validation set was used for parameter selection, while the test set was reserved exclusively for the final performance evaluation. The effects of the EMA decay coefficient and pseudo-label confidence thresholds were analyzed in detail, while the remaining hyperparameters were determined through preliminary validation experiments considering both detection performance and training stability.

Table 10 presents the model performance under different EMA decay coefficients. A smaller EMA decay coefficient causes the teacher model to rapidly follow the real-time updates of the student network, weakening the temporal smoothing effect and introducing unstable pseudo-label predictions. When the coefficient reaches 0.999, the model achieves the best performance on the validation set, with mAP@0.5 and mAP@0.5:0.95 of 72.14% and 49.36%, respectively. If the coefficient is further increased to 0.9995, the teacher network updates more slowly and may fail to adapt effectively to changes during training, resulting in a slight performance decrease. Therefore, an EMA decay coefficient of 0.999 was selected as a balance between prediction stability and model adaptability.

Table 10. Sensitivity analysis of the EMA decay coefficient on the validation set.

EMA Decay Coefficient	mAP@0.5 (%)	mAP@0.5:0.95 (%)
0.95	67.42	45.11
0.99	70.86	47.63
0.999	72.14	49.36
0.9995	71.57	48.82

Table 11 summarizes the sensitivity analysis of the medium- and high-confidence thresholds on the validation set. When the high-confidence threshold was fixed at (τ_h = 0.5), the model achieved its best validation performance at ($\tau_m = 0.2$), with mAP@0.5 and mAP@0.5:0.95 values of 72.14% and 49.36%, respectively. When the medium-confidence threshold was fixed at ($\tau_m = 0.2$), the model also achieved its highest values for both metrics at (τ_h = 0.5), yielding mAP@0.5 and mAP@0.5:0.95 values of 72.14% and 49.36%, respectively. A low medium-confidence threshold may retain more uncertain predictions as weak pseudo-labels, increasing the risk of noisy supervision, whereas an excessively high threshold may exclude potentially informative predictions. Similarly, the high-confidence threshold affects the balance between the quantity and reliability of strongly supervised pseudo-labels. These results support the use of ($\tau_m = 0.2$) and (τ_h = 0.5) as the default thresholds within the evaluated ranges.

Table 11. Sensitivity analysis of the medium- and high-confidence thresholds on the validation set. Note: When evaluating τ_m, the high-confidence threshold was fixed at ($\tau_h = 0.5$). When evaluating τ_h, the medium-confidence threshold was fixed at ($\tau_m = 0.2$). All results were obtained on the

validation set with the remaining experimental settings unchanged. The bold values indicate the selected default configuration for τ_m and the corresponding result at (τ_h = 0.50).

Varied Parameter	τ_m	τ_h	mAP@0.5 (%)	mAP@0.5:0.95 (%)
τ_m	0.1	0.5	70.71	47.94
τ_m	0.2	0.5	72.14	49.36
τ_m	0.3	0.5	69.92	47.28
τ_h	0.2	0.4	71.33	48.47
τ_h	0.2	0.5	72.14	49.36
τ_h	0.2	0.6	71.57	48.61

To further investigate the sensitivity of DLCAPL to the reliability-balancing coefficient, the weak-label adjustment coefficient, and the unsupervised loss ramp-up length, additional experiments were conducted on the validation set. Each parameter was varied independently while the remaining experimental settings were kept unchanged. The results are summarized in Table 12.

Table 12. Sensitivity analysis of the reliability-balancing coefficient, weak-label adjustment coefficient, and unsupervised loss ramp-up length on the validation set. Note: Each parameter was varied independently while the remaining hyperparameters were held at their default values. When evaluating γ, λ and the ramp-up length were fixed at 2.0 and 20 epochs, respectively. When evaluating λ, γ and the ramp-up length were fixed at 0.50 and 20 epochs, respectively. When evaluating the ramp-up length, γ and λ were fixed at 0.50 and 2.0, respectively. All results were obtained on the validation set under otherwise identical experimental settings.

Parameter	Value	mAP@0.5 (%)	mAP@0.5:0.95 (%)
γ	0.25	70.42	47.63
γ	0.50	72.14	49.36
γ	0.75	71.26	48.22
λ	1.0	71.03	48.51
λ	2.0	72.14	49.36
λ	4.0	70.58	47.79
Ramp-up length (epochs)	10	70.81	48.05
Ramp-up length (epochs)	20	72.14	49.36
Ramp-up length (epochs)	30	71.45	48.74

As shown in Table 12, the configuration with ($\gamma = 0.5$, $\lambda = 2.0$, and a ramp-up length of 20 epochs) achieved the highest validation performance, yielding mAP@0.5 and mAP@0.5:0.95 values of 72.14% and 49.36%, respectively. For the reliability-balancing coefficient, both metrics decreased when γ was set to 0.25 or 0.75, suggesting that the relative contributions of classification confidence and localization consistency affect pseudo-label reliability. Similarly, λ outperformed the tested alternatives, indicating that the strength of weak-label weighting influences the effectiveness of pseudo-label supervision. Regarding the ramp-up length, 20 epochs yielded better performance than either 10 or 30 epochs, suggesting that the timing of unsupervised loss introduction affects the balance between supervised learning and pseudo-label utilization. These results support the selected default configuration within the evaluated parameter ranges.

The NMS threshold was set to 0.60 to account for the dense distribution of livestock targets in alpine grassland UAV imagery, aiming to reduce redundant predictions while avoiding excessive suppression of adjacent individuals.

To evaluate the stability of model performance with respect to training randomness, YOLOv13s-S30, Soft Teacher-style, and DLCAPL were evaluated using three random seeds (42, 2025, and 2026) while keeping the dataset partitions, labeled subset, and

hyperparameter settings unchanged. The test-set performance across the three runs is summarized in Table 13.

Table 13. Robustness analysis of YOLOv13s-S30, Soft Teacher-style, and DLCAPL across three random seeds on the test set. Note: All methods were evaluated using the same training, validation, and test partitions. The labeled subset and hyperparameter settings were held constant across random seeds. Results are reported as mean ± standard deviation (SD) across three runs.

Model	Seed	mAP@0.5 (%)	mAP@0.5:0.95 (%)
YOLOv13s-S30	42	55.32	34.25
YOLOv13s-S30	2025	55.08	34.06
YOLOv13s-S30	2026	55.56	34.44
YOLOv13s-S30	Mean ± Std	55.32 ± 0.24	34.25 ± 0.19
Soft Teacher-style	42	70.92	47.68
Soft Teacher-style	2025	70.63	47.95
Soft Teacher-style	2026	69.82	47.31
Soft Teacher-style	Mean ± Std	70.46 ± 0.57	47.65 ± 0.32
DLCAPL	42	73.28	50.68
DLCAPL	2025	72.91	50.34
DLCAPL	2026	73.47	50.89
DLCAPL	Mean ± Std	73.22 ± 0.28	50.64 ± 0.28

As shown in Table 13, DLCAPL consistently outperformed YOLOv13s-S30 and Soft Teacher-style across all three random seeds. For DLCAPL, the standard deviations of mAP@0.5 and mAP@0.5:0.95 were both 0.28 percentage points, indicating limited performance variation across the three runs. These results suggest that the performance advantage of DLCAPL is robust to the evaluated random training seeds. However, this analysis keeps the labeled subset fixed and therefore does not assess sensitivity to the selection of labeled training images.

4.6. Robustness of DLCAPL to Labeled-Subset Selection

To investigate the sensitivity of DLCAPL to the selection of labeled training images, experiments were conducted using three independently sampled labeled subsets under the 30% labeled-data setting. The training, validation, and test partitions were kept unchanged, while the labeled subset was varied across runs. The training random seed was fixed at 42, and the model architecture, training schedule, and hyperparameter settings were kept unchanged. The same independent test set was used to evaluate all three runs. The results are summarized in Table 14.

Table 14. Robustness of DLCAPL to different labeled-subset selections under the 30% labeled-data setting. Note: The training, validation, and test partitions were held constant across runs. Three different labeled subsets were sampled from the training partition, with the labeled-data proportion fixed at 30%. The training random seed was fixed at 42, and all other experimental settings were kept unchanged. Performance was evaluated on the same independent test set. Results are reported as mean ± standard deviation (SD) across the three labeled-subset selections.

Labeled-Subset Selection	Training Seed	mAP@0.5 (%)	mAP@0.5:0.95 (%)
Draw 1	42	73.28	50.68
Draw 2	42	72.94	50.37
Draw 3	42	73.51	50.92
Mean ± Std		73.24 ± 0.29	50.66 ± 0.28

As shown in Table 14, DLCAPL achieved mean mAP@0.5 and mAP@0.5:0.95 values of 73.24% and 50.66%, respectively, across the three labeled-subset selections. The corresponding standard deviations were 0.29 and 0.28 percentage points. The mAP@0.5 values ranged from 72.94% to 73.51%, while the mAP@0.5:0.95 values ranged from 50.37% to 50.92%. The relatively small variation across these runs suggests that DLCAPL's detection performance was reasonably consistent across the three evaluated labeled subsets under the fixed training configuration. However, this analysis is limited to three subset selections and therefore provides preliminary evidence of robustness rather than establishing general robustness across all possible labeled subsets. Furthermore, because the baseline methods were not evaluated across the same alternative labeled-subset selections in this analysis, these results do not establish whether the performance advantage of DLCAPL over the baselines persists under different labeled-data selections.

5. Discussion

Livestock detection in UAV imagery of alpine grasslands remains challenging because animals are often small, densely distributed, partially occluded, and visually similar to the surrounding vegetation. These difficulties are amplified under limited-annotation conditions, when supervised detectors have fewer labeled examples from which to learn discriminative features. The quantitative results and qualitative visualizations suggest that DLCAPL improves the detection of small, occluded, and densely distributed livestock under the evaluated conditions. The visual examples also show fewer background-related false detections and more consistent bounding-box predictions in some challenging scenes. These findings support the potential of semi-supervised learning to improve livestock detection by exploiting unlabeled UAV imagery when manual annotations are limited.

Classification confidence alone does not fully characterize pseudo-label reliability. Small or partially occluded animals may have ambiguous boundaries, and their predicted boxes may vary across augmented views even when their classification confidence is relatively high. Conversely, some medium-confidence predictions may remain spatially stable across views. DLC addresses this issue by assessing the agreement between teacher and student predictions after mapping them to a common coordinate system. This cross-view consistency provides a spatial reliability cue that complements classification confidence. However, it measures prediction stability under augmentation rather than localization error relative to ground-truth boxes. Therefore, DLC should be interpreted as a proxy for spatial reliability, not as a direct estimate of absolute localization accuracy.

In the comparison reported in Table 3, DLCAPL improved mAP@0.5 and mAP@0.5:0.95 by 2.36 and 3.00 percentage points, respectively, relative to the Soft Teacher-style baseline. These results indicate improved overall detection performance under the evaluated test conditions. Nevertheless, mAP@0.5:0.95 summarizes detection performance across multiple IoU thresholds and does not isolate bounding-box localization accuracy. Its improvement therefore cannot, by itself, establish that DLC produces more accurate bounding boxes. The results should instead be interpreted as evidence that the proposed framework improves overall detection performance under the tested experimental configuration.

Soft Teacher, Consistent-Teacher, PseCo, and DLCAPL address pseudo-label reliability through different mechanisms. Soft Teacher uses box jittering to assess regression pseudo-label reliability by examining the variation in perturbed bounding boxes, whereas DLC measures agreement between teacher and student predictions across weakly and strongly augmented views. Consistent-Teacher addresses pseudo-target inconsistency and classification–regression misalignment through mechanisms including adaptive anchor assignment, feature alignment, and dynamic confidence-threshold adjustment.

PseCo combines prediction-guided label assignment, positive-proposal consistency voting, and multi-view scale-invariant learning to address noisy pseudo-boxes and scale-related consistency. DLCAPL combines cross-view localization consistency with hierarchical pseudo-label partitioning and reliability-aware weak-label weighting within a YOLOv13s-based framework. Because Consistent-Teacher and PseCo were not evaluated under the same experimental protocol in this study, these comparisons are methodological rather than quantitative and do not establish the superiority of DLCAPL over those methods. The "-style" baselines likewise represent adaptations to the common YOLOv13s framework rather than exact reproductions of the original methods and their complete training pipelines.

The ablation experiments suggest that the components of DLCAPL make complementary contributions. A single high-confidence threshold can exclude potentially informative predictions of small, distant, or partially occluded animals. Hierarchical pseudo-label partitioning retains medium-confidence candidates as weak pseudo-labels, while reliability-aware weighting adjusts their contributions according to classification confidence and cross-view localization consistency. The observed results support combining candidate selection with differentiated supervision rather than relying exclusively on a fixed confidence threshold. However, although the box-jitter comparison supports the utility of DLC within the evaluated DLCAPL configuration, the detection metrics do not directly demonstrate that DLC estimates absolute localization quality more accurately.

Experiments with different annotation ratios further indicate that DLCAPL can benefit from unlabeled imagery under limited-annotation conditions. DLCAPL outperformed the corresponding supervised baselines at the evaluated annotation ratios of 10%, 20%, and 30%. The magnitude of this benefit may depend on the amount and composition of labeled data, and the observed trend should be interpreted within the dataset and experimental settings used here. In addition, the teacher model and pseudo-label optimization modules are used only during training. The trained student detector is retained for inference, so these components do not introduce additional inference-time modules or theoretical model complexity.

Several limitations should be acknowledged. First, DLC measures cross-view prediction stability rather than absolute localization accuracy relative to ground-truth bounding boxes. Its correlation with ground-truth IoU has not been quantitatively evaluated using annotations for the unlabeled training images. Consequently, whether higher DLC scores reliably correspond to better localization against ground truth remains an open question. Second, the framework uses fixed medium- and high-confidence thresholds of 0.20 and 0.50. Although sensitivity analyses were conducted for these thresholds, the reliability-balancing coefficient, the weak-label adjustment coefficient, and the unsupervised-loss ramp-up length, the current framework does not explicitly adapt these parameters to class-specific confidence distributions, target scales, or training stages. Future work could investigate adaptive thresholding and parameter interactions to improve pseudo-label selection.

Third, the test set contains only 22 horse instances, making the corresponding class-wise performance estimates uncertain. The dataset was collected in a single study region in Ruoergai County, and the independent test set comprised 60 UAV images. Although images from the same or adjacent flight segments were assigned to the same subset to reduce spatial leakage, no independent site-held-out or date-held-out evaluation was conducted, and bootstrap confidence intervals were not calculated. The reported results should therefore be interpreted as point estimates for the available test set rather than definitive evidence of generalization across regions, acquisition dates, or environmental conditions. Fourth, the multi-seed experiments used a fixed labeled subset and therefore assessed variability associated with training randomness rather than sensitivity to

labeled-image selection. Although experiments with independently sampled labeled subsets were conducted, the number of selections was limited, and the baselines were not evaluated across the same alternative subsets. Further comparisons using multiple labeled-subset draws would provide stronger evidence of robustness. Finally, the teacher–student framework and dual-view prediction increase computational requirements during training, even though they add no inference-time modules. Future work should validate DLC against ground-truth localization quality, extend the evaluation across regions and acquisition dates, examine robustness to labeled-subset selection, and investigate adaptive pseudo-label selection and class-imbalance handling.

6. Conclusions

To address the annotation burden and limitations of confidence-based pseudo-labeling in UAV-based livestock detection, this study proposes DLCAPL, a semi-supervised object detection framework that integrates dual-view localization consistency (DLC), hierarchical pseudo-label partitioning, and reliability-aware weak-label weighting. Within a teacher–student framework, DLC assesses prediction stability across augmented views and complements classification confidence in pseudo-label reliability assessment. Hierarchical partitioning separates candidates into reliable, weak, and ignored groups, while reliability-aware weighting adjusts the contribution of weak pseudo-labels to the unsupervised loss. Together, these components provide a unified strategy for selecting pseudo-labels and adjusting their supervision strength under limited-annotation conditions.

On the UAV livestock dataset collected in the Ruoergai alpine grasslands, DLCAPL achieved a precision of 82.13%, a recall of 70.46%, an mAP@0.5 of 73.28%, and an mAP@0.5:0.95 of 50.68% when 30% of the training images were labeled. Compared with YOLOv13s-S30 trained using the same labeled-data ratio, these metrics improved by 23.90, 18.29, 17.96, and 16.43 percentage points, respectively. Compared with the Soft Teacher-style baseline, DLCAPL achieved gains of 2.36 percentage points in mAP@0.5 and 3.00 percentage points in mAP@0.5:0.95. These results indicate that DLCAPL can make effective use of unlabeled UAV imagery and improve overall detection performance under the evaluated conditions. However, the improvement in mAP@0.5:0.95 should not be interpreted as direct evidence of more accurate bounding-box localization because this metric summarizes detection performance across multiple IoU thresholds rather than isolating localization accuracy.

The ablation experiments suggest that DLC provides a complementary cross-view spatial-consistency cue, hierarchical pseudo-label partitioning helps retain potentially informative medium-confidence predictions, and reliability-aware weighting differentiates the supervision contributions of weak pseudo-labels. The contribution of DLCAPL therefore lies in combining confidence-based candidate selection with cross-view localization consistency to guide pseudo-label supervision, rather than relying on classification confidence alone. Nevertheless, the relationship between DLC scores and absolute localization accuracy has not yet been directly quantified using ground-truth IoU for unlabeled images. In addition, the evaluation was conducted in a single study region, and the test set contained only 22 horse instances. Further evaluation across regions and acquisition dates, together with direct validation of DLC against ground-truth localization quality, is needed to establish the generalizability of the proposed framework.

Author Contributions: Conceptualization, L.W.; methodology, S.Z. and L.W.; software, S.Z.; validation, S.Z.; formal analysis, S.Z.; investigation, S.Z.; resources, L.H.; data curation, S.Z.; writing—original draft preparation, S.Z.; writing—review and editing, L.W.; visualization, S.Z.; supervision, L.W.; project administration, L.W.; funding acquisition, L.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Disciplinary Team Construction Project (Grant No. GAU-XKTD-2026-077) and the National Key Research and Development Program of China (Grant No. 2024YFD1400500).

Data Availability Statement: The raw UAV images generated in this study are not publicly available due to confidentiality requirements associated with the National Key Research and Development Program of China (Grant No. 2024YFD1400500). The processed data supporting the findings of this study and the model source code are available from the corresponding authors upon reasonable request.

Acknowledgments: The authors thank the Key Laboratory of Grassland Ecosystem of the Ministry of Education at Gansu Agricultural University for providing technical support during UAV image collection and model experiments.

Conflicts of Interest: The authors declare no conflicts of interest. The funder had no role in the design of the study; in the collection, analysis, or interpretation of the data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

UAV	Unmanned Aerial Vehicle
YOLO	You Only Look Once
CNN	Convolutional Neural Network
mAP	mean Average Precision
FPS	Frames Per Second

References

1. Zhao, X.; Zhao, L.; Li, Q.; Chen, H.; Zhou, H.; Xu, S.; Dong, Q.; Wu, G.; He, Y. Using balance of seasonal herbage supply and demand to inform sustainable grassland management on the Qinghai–Tibetan Plateau. *Front. Agric. Sci. Eng.* **2018**, *5*, 1–8. https://doi.org/10.15302/J-FASE-2018203.
2. Yan, L.; Kong, L.; Wang, L.; Zhang, L.; Hu, J.; Ouyang, Z. Grass-livestock balance under the joint influences of climate change, human activities and ecological protection on Tibetan Plateau. *Ecol. Indic.* **2024**, *162*, 112040. https://doi.org/10.1016/j.ecolind.2024.112040.
3. Barbedo, J.G.A.; Koenigkan, L.V. Perspectives on the use of unmanned aerial systems to monitor cattle. *Outlook Agric.* **2018**, *47*, 214–222. https://doi.org/10.1177/0030727018781876.
4. Alanezi, M.A.; Shahriar, M.S.; Hasan, M.B.; Ahmed, S.; Sha'aban, Y.A.; Bouchekara, H.R.E.H. Livestock management with unmanned aerial vehicles: A review. *IEEE Access* **2022**, *10*, 45001–45028. https://doi.org/10.1109/ACCESS.2022.3168295.
5. Xiang, T.-Z.; Xia, G.-S.; Zhang, L. Mini-unmanned aerial vehicle-based remote sensing: Techniques, applications, and prospects. *IEEE Geosci. Remote Sens. Mag.* **2019**, *7*, 29–63. https://doi.org/10.1109/MGRS.2019.2918840.
6. Shao, W.; Kawakami, R.; Yoshihashi, R.; You, S.; Kawase, H.; Naemura, T. Cattle detection and counting in UAV images based on convolutional neural networks. *Int. J. Remote Sens.* **2020**, *41*, 31–52. https://doi.org/10.1080/01431161.2019.1624858.
7. 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.
8. Wang, Y.; Ma, L.; Wang, Q.; Wang, N.; Wang, D.; Wang, X.; Zheng, Q.; Hou, X.; Ouyang, G. A lightweight and high-accuracy deep learning method for grassland grazing livestock detection using UAV imagery. *Remote Sens.* **2023**, *15*, 1593. https://doi.org/10.3390/rs15061593.
9. Adam, M.; Song, J.; Yu, W.; Li, Q. Deep learning approaches for automatic livestock detection in UAV imagery: State-of-the-art and future directions. *Future Internet* **2025**, *17*, 431. https://doi.org/10.3390/fi17090431.
10. Zhao, T.; Zeng, Y.; Fang, Q.; Xu, X.; Xie, H. Semi-supervised object detection for remote sensing images using consistent dense pseudo-labels. *Remote Sens.* **2025**, *17*, 1474. https://doi.org/10.3390/rs17081474.
11. Jeong, J.; Lee, S.; Kim, J.; Kwak, N. Consistency-based semi-supervised learning for object detection. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 10758–10767.

12. Tang, Y.; Chen, W.; Luo, Y.; Zhang, Y. Humble teachers teach better students for semi-supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Virtual, 19–25 June 2021*; IEEE: Piscataway, NJ, USA, 2021; pp. 3132–3141. https://doi.org/10.1109/CVPR46437.2021.00315.
13. Xu, M.; Zhang, Z.; Hu, H.; Wang, J.; Wang, L.; Wei, F.; Bai, X.; Liu, Z. End-to-end semi-supervised object detection with Soft Teacher. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Virtual Event, 10–17 October 2021*; IEEE: Piscataway, NJ, USA, 2021; pp. 3060–3069. https://doi.org/10.1109/ICCV48922.2021.00305.
14. Liu, Y.-C.; Ma, C.-Y.; He, Z.; Kuo, C.-W.; Chen, K.; Zhang, P.; Wu, B.; Kira, Z.; Vajda, P. Unbiased teacher for semi-supervised object detection. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual Event, 3–7 May 2021.
15. Zhou, H.; Ge, Z.; Liu, S.; Mao, W.; Li, Z.; Yu, H.; Sun, J. Dense Teacher: Dense pseudo-labels for semi-supervised object detection. In *Computer Vision—ECCV 2022*; Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T., Eds.; Springer: Cham, Switzerland, 2022; Volume 13669, pp. 35–50. https://doi.org/10.1007/978-3-031-20077-9_3.
16. Wang, K.; Xiao, Z.; Wan, Q.; Xia, F.; Chen, P.; Li, D. Global focal learning for semi-supervised oriented object detection. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5636013. https://doi.org/10.1109/TGRS.2024.3438844.
17. Zhao, T.; Fang, Q.; Xu, X. Denser Teacher: Rethinking dense pseudo-label for semi-supervised oriented object detection. *IEEE Trans. Circuits Syst. Video Technol.* **2025**, *35*, 4549–4559. https://doi.org/10.1109/TCSVT.2024.3518452.
18. Liu, L.; Zhang, B.; Zhang, J.; Zhang, W.; Gan, Z.; Tian, G.; Zhu, W.; Wang, Y.; Wang, C. MixTeacher: Mining promising labels with mixed scale teacher for semi-supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023*; IEEE: Piscataway, NJ, USA, 2023; pp. 7370–7379. https://doi.org/10.1109/CVPR52729.2023.00712.
19. Li, H.; Wu, Z.; Shrivastava, A.; Davis, L.S. Rethinking pseudo labels for semi-supervised object detection. *Proc. AAAI Conf. Artif. Intell.* **2022**, *36*, 1314–1322. https://doi.org/10.1609/aaai.v36i2.20019.
20. Wang, X.; Yang, X.; Zhang, S.; Li, Y.; Feng, L.; Fang, S.; Lyu, C.; Chen, K.; Zhang, W. Consistent-Teacher: Towards reducing inconsistent pseudo-targets in semi-supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023*; IEEE: Piscataway, NJ, USA, 2023; pp. 3240–3249. https://doi.org/10.1109/CVPR52729.2023.00316.
21. Ruoergai County People's Government. Overview of Ruoergai County. Available online: https://www.ruoergai.gov.cn/regxrmzf/c100125/201908/2d4ee5b5d9bc4ec3a8a91b246ae9d02d.shtml (accessed on 23 July 2026).
22. Yang, J.; Su, P.; Zhou, Z.; Shi, R.; Ding, X. Environmental filtering rather than dispersal limitation dominated plant community assembly in the Zoige Plateau. *Ecol. Evol.* **2022**, *12*, e9117. https://doi.org/10.1002/ece3.9117.
23. *ISO 12232:2019*; Photography—Digital Still Cameras—Determination of Exposure Index, ISO Speed Ratings, Standard Output Sensitivity, and Recommended Exposure Index. International Organization for Standardization: Geneva, Switzerland, 2019.
24. Wada, K. *Labelme: Image Polygonal Annotation with Python*, version 4.6.0; Zenodo: Geneva, Switzerland, 2021. https://doi.org/10.5281/zenodo.5711226.
25. Buslaev, A.; Iglovikov, V.I.; Khvedchenya, E.; Parinov, A.; Druzhinin, M.; Kalinin, A.A. Albumentations: Fast and flexible image augmentations. *Information* **2020**, *11*, 125. https://doi.org/10.3390/info11020125.
26. Kaufman, S.; Rosset, S.; Perlich, C.; Stitelman, O. Leakage in data mining: Formulation, detection, and avoidance. *ACM Trans. Knowl. Discov. Data* **2012**, *6*, 1–21. https://doi.org/10.1145/2382577.2382579.
27. Lei, M.; Li, S.; Wu, Y.; Hu, H.; Zhou, Y.; Zheng, X.; Ding, G.; Du, S.; Wu, Z.; Gao, Y. YOLOv13: Real-time object detection with hypergraph-enhanced adaptive visual perception. *arXiv* **2025**, arXiv:2506.17733. https://doi.org/10.48550/arXiv.2506.17733.
28. Laine, S.; Aila, T. Temporal ensembling for semi-supervised learning. In Proceedings of the 5th International Conference on Learning Representations (ICLR), Toulon, France, 24–26 April 2017.
29. Tarvainen, A.; Valpola, H. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 1195–1204.
30. Liu, C.; Zhang, W.; Lin, X.; Zhang, W.; Tan, X.; Han, J.; Li, X.; Ding, E.; Wang, J. Ambiguity-resistant semi-supervised learning for dense object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023*; IEEE: Piscataway, NJ, USA, 2023; pp. 15579–15588. https://doi.org/10.1109/CVPR52729.2023.01495.
31. Chen, B.; Chen, W.; Yang, S.; Xuan, Y.; Song, J.; Xie, D.; Pu, S.; Song, M.; Zhuang, Y. Label matching semi-supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022*; IEEE: Piscataway, NJ, USA, 2022; pp. 14381–14390. https://doi.org/10.1109/CVPR52688.2022.01398.

32. Li, G.; Li, X.; Wang, Y.; Wu, Y.; Liang, D.; Zhang, S. PseCo: Pseudo labeling and consistency training for semi-supervised object detection. In *Computer Vision—ECCV 2022*; Springer: Cham, Switzerland, 2022; Volume 13669, pp. 457–472. https://doi.org/10.1007/978-3-031-20077-9_27.
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. Arazo, E.; Ortego, D.; Albert, P.; O'Connor, N.E.; McGuinness, K. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 19–24 July 2020; pp. 1–8. https://doi.org/10.1109/IJCNN48605.2020.9207304.
35. Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common objects in context. In *Computer Vision—ECCV 2014*; Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T., Eds.; Springer: Cham, Switzerland, 2014; Volume 8693, pp. 740–755. https://doi.org/10.1007/978-3-319-10602-1_48.
36. Lee, D.-H. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In Proceedings of the ICML 2013 Workshop on Challenges in Representation Learning, Atlanta, GA, USA, 21 June 2013; pp. 1–6.
37. Sohn, K.; Zhang, Z.; Li, C.-L.; Zhang, H.; Lee, C.-Y.; Pfister, T. A Simple Semi-Supervised Learning Framework for Object Detection. *arXiv* **2020**, arXiv:2005.04757.
38. Liu, Y.-C.; Ma, C.-Y.; Kira, Z. Unbiased Teacher v2: Semi-supervised object detection for anchor-free and anchor-based detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022*; IEEE: Piscataway, NJ, USA, 2022; pp. 9819–9828. https://doi.org/10.1109/CVPR52688.2022.00959.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.