Enhancing UAV Object Detection in Low-Light Conditions with ELS-YOLO: A Lightweight Model Based on Improved YOLOv11

Tianhang Weng and Xiaopeng Niu *

Posted Date: 27 June 2025

doi: 10.20944/preprints202506.2235.v1

Keywords: low-light conditions; YOLOv11; lightweight; model pruning

 doi:10.20944/preprints202506.2235.v1

Article

Enhancing UAV Object Detection in Low-Light Conditions with ELS-YOLO: A Lightweight Model Based on Improved YOLOv11

Tianhang Weng and Xiaopeng Niu *

School of Computer Science and Artificial Intelligence, Beijing Technology and Business University, Beijing 100048, China
* Correspondence: nxp_btbu@btbu.edu.cn

Abstract: Drone-view object detection models operating under low-light conditions face several challenges, such as object scale variations, high image noise, and limited computational resources. Existing models often struggle to balance accuracy and lightweight architecture. This paper introduces ELS-YOLO, a lightweight object detection model tailored for low-light environments, built upon the YOLOv11s framework. ELS-YOLO features a re-parameterized backbone (ER-HGNetV2) with integrated Re-parameterized Convolution and Efficient Channel Attention mechanisms, a Lightweight Feature Selection Pyramid Network (LFSPN) for multi-scale object detection, and a Shared Convolution Separate Batch Normalization Head (SCSHead) to reduce computational complexity. Layer-Adaptive Magnitude-Based Pruning (LAMP) is employed to compress the model size. Experiments on the ExDark and DroneVehicle datasets demonstrate that ELS-YOLO achieves high detection accuracy with a compact model. Here, we show that ELS-YOLO attains a mAP@0.5 of 74.3% and 68.7% on the ExDark and DroneVehicle datasets, respectively, while maintaining real-time inference capability.

Keywords: low-light conditions; YOLOv11; lightweight; model pruning

1. Introduction

Drone-view object detection (DVOD) aims to locate and classify objects in images or videos captured by unmanned aerial vehicles (UAVs) [1]. With the rapid advancements in computer vision and UAV technologies, DVOD has become prevalent in diverse applications, including security surveillance, intelligent transportation systems, and environmental monitoring, achieving significant results. For instance, Wu et al. [2] proposed CCR-Net, a multimodal feature fusion network that improves operational efficiency in disaster response and emergency relief missions. Huang et al. [3] developed UFPMP-Det, which accurately identifies crop diseases and pests from UAV imagery. Zhan et al. [4] introduced ARGNet, effectively detecting forest fire smoke. By integrating deep learning algorithms, UAV systems can monitor critical events such as traffic violations and accidents in real-time, providing essential support for informed decision-making.

Nighttime security patrols and search-and-rescue missions require real-time and precise monitoring over large regions. However, traditional manual inspections are inefficient and prone to missed detections. Conventional surveillance equipment struggles with capturing detailed imagery under low-light conditions and lacks automated data analysis capabilities. UAV systems equipped with object detection algorithms offer rapid deployment, mobility, and automated recognition, enabling efficient wide-area monitoring in a short time [5]. Therefore, developing robust and efficient object detection techniques for low-light conditions has become a critical research focus [6].

Currently, most object detection methods for low-light conditions primarily rely on image preprocessing techniques, such as brightness adjustment and noise suppression, to enhance image quality and thereby improve detection performance [7]. For example, Guo et al. [8] proposed an illumination map estimation method that initializes low-light images based on maximum RGB channel values. Hu et al. [9] mitigated color distortion in low-light images through saturation adjustment. Jeon et

 doi:10.20944/preprints202506.2235.v1

al. [10] combined atmospheric scattering models with pixel-adaptive gamma correction for image enhancement. However, these enhancement-based methods exhibit inherent limitations. First, the enhancement process can introduce artifacts that obscure essential image details. Second, reliance on fixed prior knowledge restricts adaptability to dynamically changing lighting conditions and limits the model's ability to learn deeper, high-level semantic features. Moreover, the computational overhead associated with image enhancement is significant for resource-constrained edge devices, severely limiting their real-time performance and practical deployment.

The YOLO (You Only Look Once) [11] series models represent single-stage object detection frameworks capable of performing localization and classification simultaneously in a single forward pass. These models, characterized by simple architectures, effectively balance detection accuracy and real-time performance, making them suitable for resource-constrained edge devices. Lightweight variants such as YOLOv8-nano, YOLOv9-tiny [12], and YOLOv10-nano [13] demonstrate robust performance on natural image datasets like Pascal VOC [14] and MS COCO [15], but they are not optimized specifically for low-light or UAV-captured imagery. Consequently, their performance significantly deteriorates under complex backgrounds and weak object features.

To address these issues, we propose a lightweight object detection model tailored specifically for low-light conditions called ELS-YOLO. This model builds upon the YOLOv11s framework and aims to balance detection accuracy with architectural efficiency. Specifically, we design a re-parameterized backbone network called ER-HGNetV2 that integrates Re-parameterized Convolution (RepConv) [16] and Efficient Channel Attention (ECA) [17] mechanisms to better capture critical features and suppress noise in complex environments. To address the challenge of detecting multi-scale objects, we develop the Lightweight Feature Selection Pyramid Network (LFSPN), which enables efficient cross-scale feature fusion and improves both model generalization and detection accuracy. To further reduce computational costs, we design the SCSHead that significantly reduces resource consumption during inference while maintaining detection accuracy. Given the limited computational resources in UAV applications, we introduce Layer-Adaptive Magnitude-Based Pruning (LAMP) [18] to precisely prune redundant parameters.

The main contributions of this work are summarized as follows:

1. We design the re-parameterized backbone ER-HGNetV2 for low-light environments, whcih effectively captures high-quality features, suppresses noise, and enhances feature representation.
2. We develop LFSPN, which enables efficient multi-scale feature fusion and enhances detection capability across diverse object scales.
3. We introduce SCSHead, a lightweight detection head leveraging shared convolutions and separate batch normalization layers to minimize computational complexity and enhance inference efficiency.
4. Extensive experiments conducted on the ExDark and DroneVehicle datasets demonstrate that ELS-YOLO achieves an optimal balance between detection accuracy and inference speed.

2. Related Work

2.1. DVOD: Drone-View Object Detection

As an emerging research direction in remote sensing, DVOD faces unique challenges compared to conventional ground-view detection. Images captured by drones often contain numerous targets with significant scale variations, complicating detection accuracy and robustness. Existing DVOD approaches can be broadly classified into three categories: super-resolution-based, context-based, and representation fusion-based methods.

Super-resolution-based methods [19,20] enhance small object detectability by reconstructing low-resolution regions into high-resolution representations, typically through a three-stage pipeline: candidate region proposal, super-resolution reconstruction, and detection. Although these methods significantly improve object perception, their multi-stage structures often introduce redundant com-

 doi:10.20944/preprints202506.2235.v1

putation and complicate the training process, limiting their practicality and end-to-end optimization capability.

Context-based methods [21–23] leverage local and global contextual information to build spatial relationships and semantic dependencies, enhancing semantic representation and scene understanding. However, drone imagery often presents complex backgrounds and ambiguous semantic boundaries, hindering effective context modeling and reducing overall detection accuracy.

Representation fusion-based methods [24,25] integrate fine-grained spatial details from shallow features with high-level semantic features from deeper layers, primarily using architectures such as the Feature Pyramid Network (FPN) and its variants. Nonetheless, under low-light conditions, the representational gap between scales is pronounced, and direct fusion may introduce noise, degrading the discriminative ability of the model.

2.2. LLOD: Low-Light Object Detection

Existing research on object detection in low-light environments mainly focuses on improving image quality through low-light image enhancement (LLIE) and enhancing detection performance through architectural optimization.

LLIE techniques aim to restore critical information in dark regions by enhancing image brightness, contrast, and overall visual quality. Early LLIE methods relied on pixel intensity mapping and local statistical modeling. Techniques such as exposure correction [26] adjust global brightness distributions to enhance visibility, whereas histogram equalization [27] redistributes pixel intensity histograms to increase contrast. The Retinex theory [28] offers a physically interpretable enhancement framework by decomposing an image into illumination and reflectance components, thereby modeling contributions from lighting and surface texture. In recent years, deep neural networks have achieved significant advancements in LLIE tasks. LLNet [29] was the first deep autoencoder-based network designed for simultaneous low-light enhancement and denoising. Wei et al. [30] combined Retinex theory with convolutional neural networks, incorporating Gaussian filtering and logarithmic transformation to perform adaptive brightness correction. Guo et al. proposed Zero-DCE [31,32], a method that achieves fast, reference-free image enhancement by learning pixel-wise luminance adjustment curves. Xu et al. [33] developed an SNR-aware network that adaptively enhances images through global attention mechanisms and local structure modeling.

Architectural optimization aims to improve feature extraction and object recognition under low-light conditions by refining network structures and incorporating attention mechanisms. Long et al. [34] proposed a multi-level illumination learning framework SCINet, which enhances feature extraction under complex backgrounds. Qiu et al. [22] introduced Efficient Attention Pyramid Transformer (EAPT), which integrates deformable attention and a global encoder-decoder structure to improve multi-scale feature modeling. Hu et al. [35] proposed an occlusion-aware attention module MPCM to alleviate detection difficulties caused by occlusion. Peng et al. [36] enhanced detection performance in low-light scenarios by optimizing attention mechanisms and the loss function. Wu et al. [37] developed the progressive enhancement network AENet, combining Yeo-Johnson transformation with a Transformer architecture to improve dynamic feature representation.

3. Baseline algorithm

YOLOv11 [38] is the latest version of the YOLO series, provides five model variants: n, s, m, l, and x, to support deployment across a spectrum of platforms, ranging from edge devices to high-performance servers. As illustrated in Figure 1, YOLOv11 adopts a classic three-stage architecture consisting of a backbone, neck, and head, which are responsible for feature extraction, feature fusion, and object detection, respectively.

 doi:10.20944/preprints202506.2235.v1

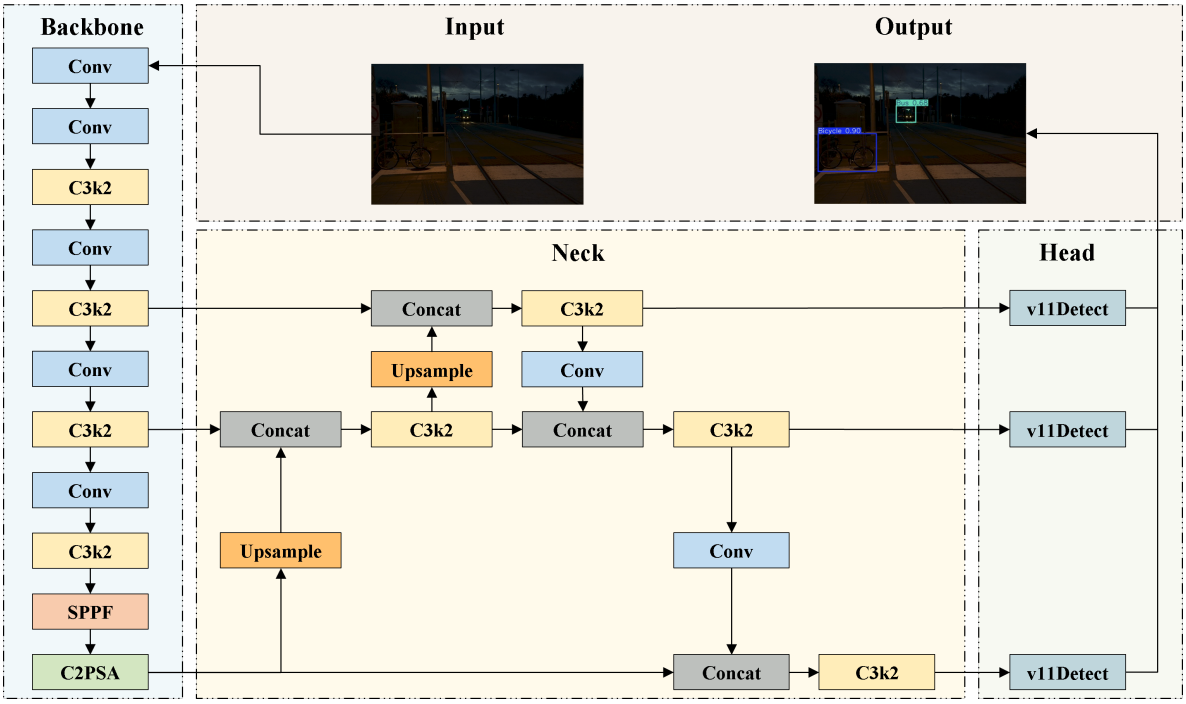

Figure 1. Overall architecture of the YOLOv11 model.

As a significant advancement in real-time object detection, YOLOv11 inherits the high efficiency and end-to-end detection capabilities of previous YOLO models, while incorporating multiple architectural innovations and optimizations. An overview of YOLOv11 primary modules is presented in Figure 2. The C3k2 module serves as the core structural unit of YOLOv11 and adjusts the kernel size by modifying the C3k parameter to meet the feature extraction requirements of different scenarios. The C2PSA module uses spatial attention to guide the model to focus on key regions and improve the accuracy of small or occluded objects. In the detection head, YOLOv11 employs depthwise separable convolutions (DWConv) to replace standard convolutions, thereby further reducing the parameter count and accelerating inference speed.

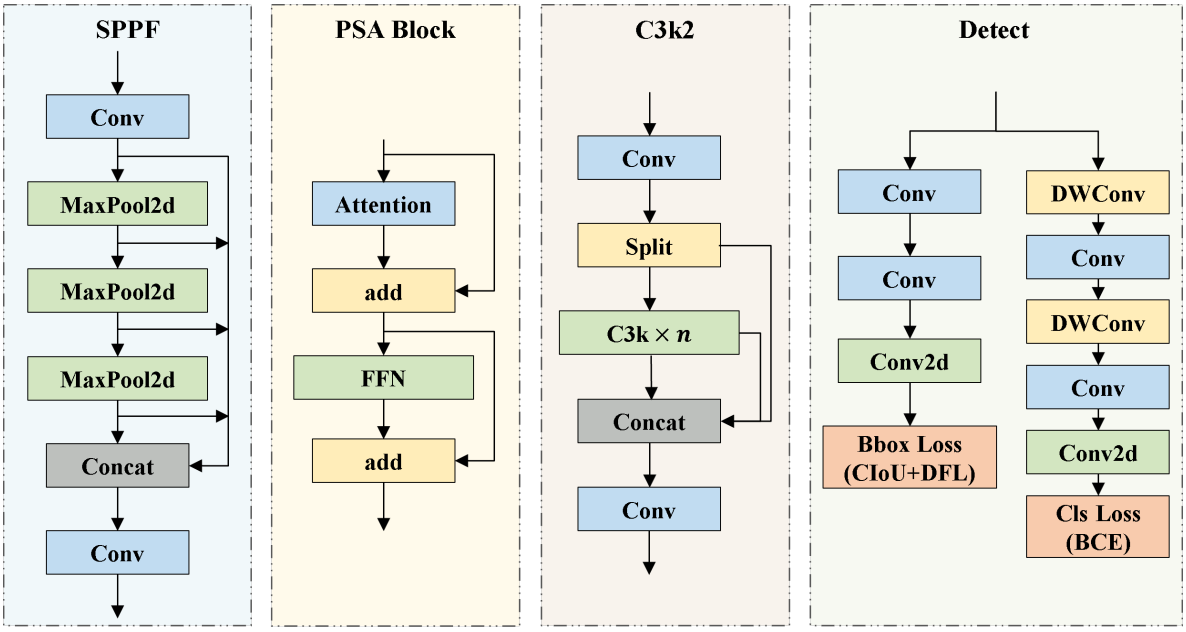

Figure 2. Detailed architectural design of core modules in YOLOv11.

 doi:10.20944/preprints202506.2235.v1

Although YOLOv11 performs well under normal lighting conditions, its ability to extract features in low-light environments is limited, making it difficult to identify critical objects. This limitation motivates us to redesign the network to better meet the detection needs of nighttime UAV imagery and improve performance under challenging illumination.

4. Methodology

4.1. ER-HGNetV2: Re-parameterized Backbone

The backbone network of YOLOv11 primarily consists of alternating stacks of standard convolutional layers and C3k2 modules. Although this design demonstrates strong feature extraction capabilities, its increasing depth and channel width lead to parameter redundancy and substantial computational overhead. Moreover, this structure struggles to effectively capture complex global semantic features in low-light UAV imagery. To address this issue, inspired by the HGNetV2 framework[39], we design a re-parameterized backbone ER-HGNetV2, which is illustrated in Figure 3.

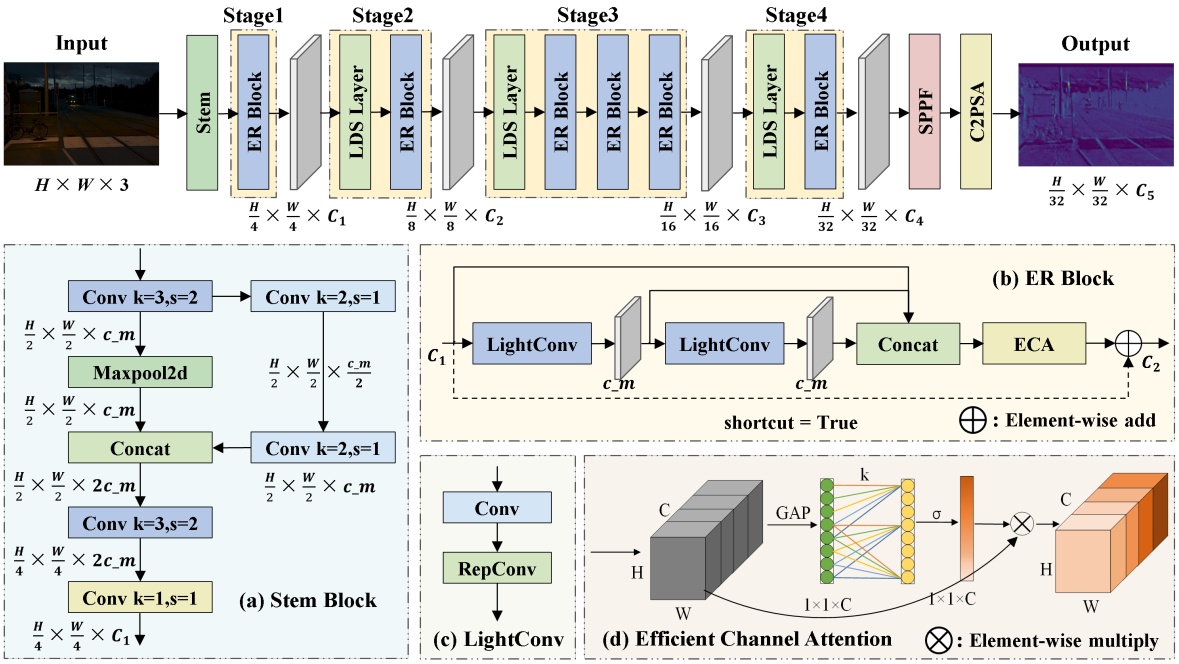

Figure 3. Network architecture of the proposed ER-HGNetV2 backbone.

ER-HGNetV2 begins with a Stem module that consists of standard convolution and max-pooling operations, performing initial spatial downsampling and extracting fundamental feature representations. ERBlock forms the core of the network and applies a multi-scale feature extraction strategy based on stacked convolution layers to refine features and enhance representational accuracy. LDS module leverages depthwise separable convolutions to efficiently reduce spatial resolution and expand feature channels, thereby enhancing global context awareness while minimizing model complexity.

We construct ERBlock using RepConv and ECA mechanism. As illustrated in Figure4, RepConv adopts a multi-branch training structure that incorporates 3×3 convolutions, 1×1 convolutions, and identity mappings to capture diverse feature patterns. During inference, RepConv applies re-parameterization to fuse convolutional layers and Batch Normalization into a single standard convolution, thereby achieves a balance between representational capacity and inference efficiency. As shown in Fig.3, ECA mechanism leverages a lightweight 1D convolution to adaptively compute channel-wise attention weights. This compact design effectively emphasizes critical feature channels with minimal computational cost, thereby improves the model's ability to distinguish targets under low-light conditions.

 doi:10.20944/preprints202506.2235.v1

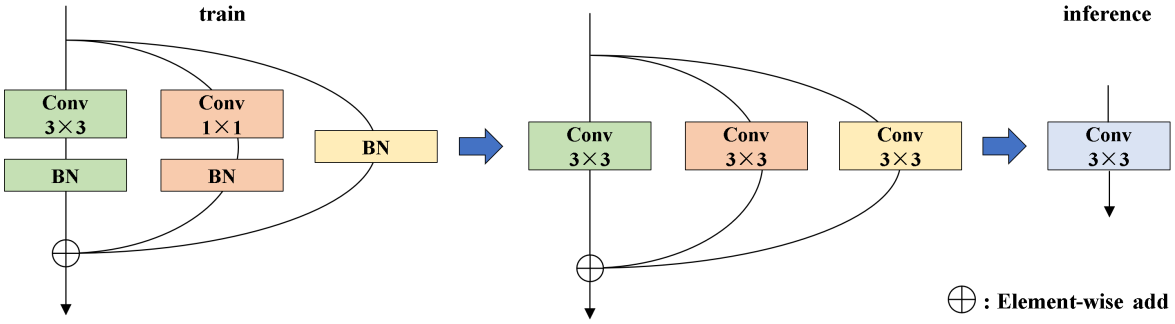

Figure 4. Re-parameterization process of the RepConv module.

4.2. LFSPN: Lightweight Feature Selection Pyramid Network

The Path Aggregation Feature Pyramid Network (PAFPN) used in YOLOv11 suffers from unselective aggregation and cross-level semantic inconsistency during feature fusion, making it difficult to obtain discriminative multi-scale representations. To address this issue, we design the LFSPN, which selectively strengthens semantically relevant features and suppresses redundant background information to significantly improve model robustness under challenging lighting conditions. As shown in Figure 5, LFSPN consists of two stages: an attention-weighted stage and a dynamic cross-scale feature fusion stage.

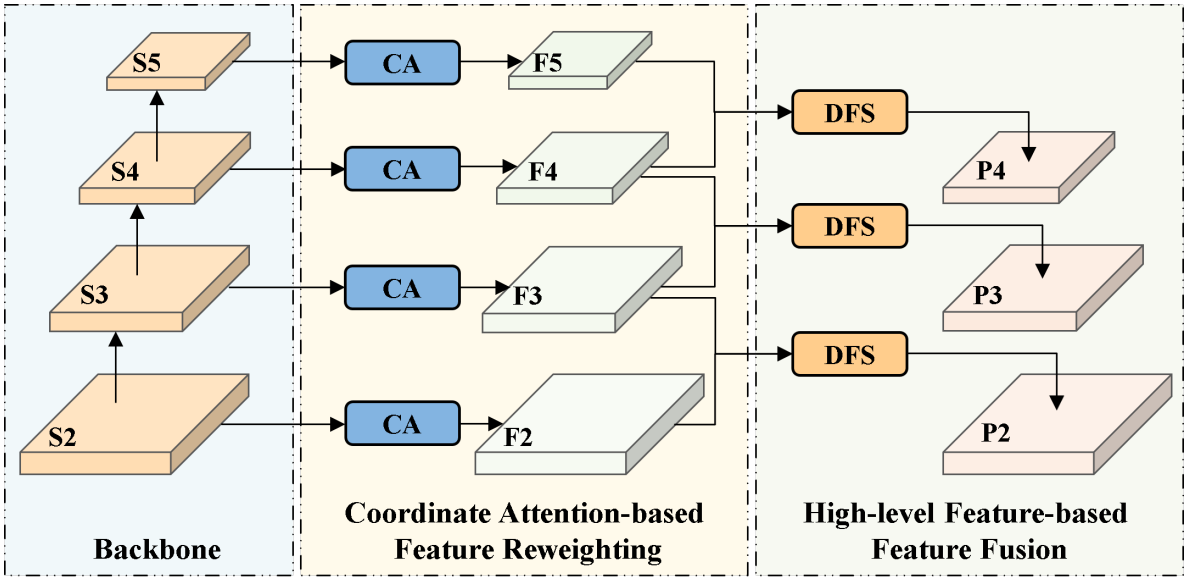

Figure 5. Architecture of the proposed Lightweight Feature Selection Pyramid Network.

The attention-weighted stage performs initial feature selection on the multi-scale feature maps extracted by the backbone and enhances spatial position awareness via the Coordinate Attention (CA) [40] mechanism. As shown in Figure 6, CA mechanism first applies global average pooling along the horizontal and vertical directions to capture spatial dependencies. It then concatenates the pooled features along the channel dimension and applies a 1×1 convolution to model cross-directional spatial relationships and generate intermediate attention weights. Finally, pixel-wise weighting is applied to the original feature map using the generated attention weights to emphasize informative regions and suppress redundant features, improving the quality of representations for the subsequent fusion stage.

 doi:10.20944/preprints202506.2235.v1

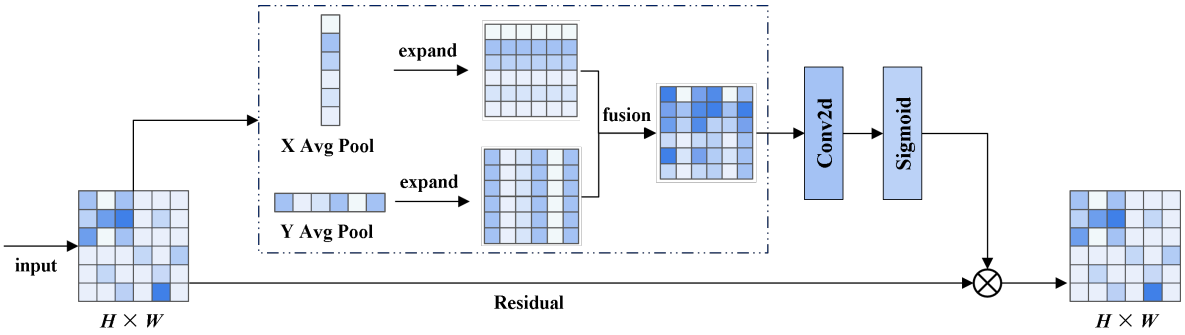

Figure 6. Illustration of the CA mechanism.

To further improve the effectiveness of feature fusion, we design a Dynamic Feature Selection (DFS) module, as shown in Figure 7. Taking the high-level feature map F_4 and the low-level feature map F_3 as an example, we first apply a 1×1 convolution to F_4 for nonlinear dimensionality reduction to lower computational cost. Then upsample F_4 using transposed convolution to match the spatial resolution of F_3. CA mechanism is applied to the upsampled feature map F_4 to generate attention weights for dynamic feature selection. These weights are element-wise multiplied with F_3 to selectively enhance spatially informative regions in the low-level feature map. The refined F_3 is then fused with F_4 through element-wise addition to produce a feature representation that preserves both fine-grained spatial details and high-level semantic information. The fused features are passed through a C3k2 module to further enhance feature representations, resulting in the final output feature map, as

$$f_{\text{tmp}} = \text{Trans}\Big(\text{Conv}_{1\times1}(f_{\text{high}})\Big) \tag{1}$$

$$f_{\text{out}} = \text{C3k2}\big(f_{\text{low}} \cdot \text{CA}(f_{\text{tmp}}) + f_{\text{tmp}}\big) \tag{2}$$

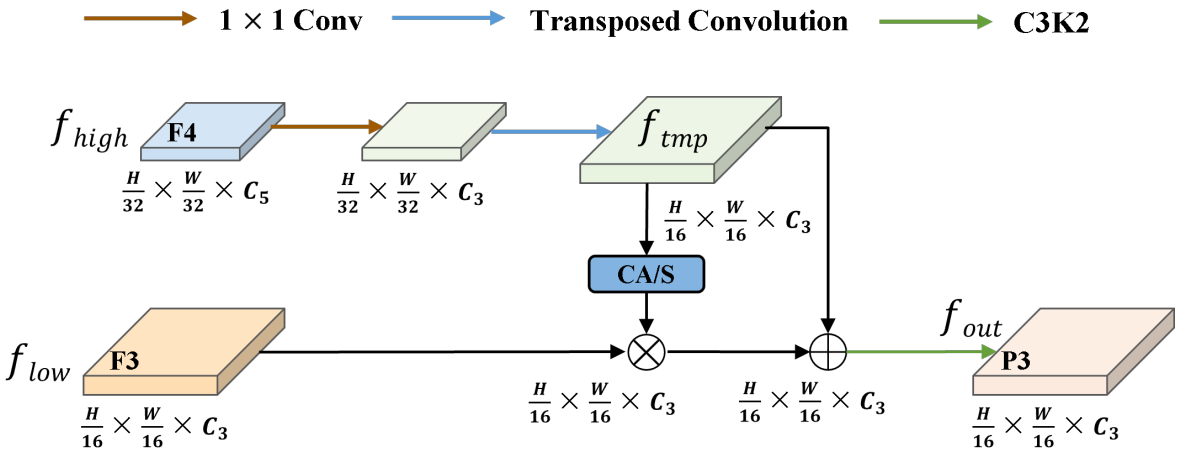

Figure 7. Framework of the Dynamic Feature Selection module.

4.3. SCSHead: Shared Convolution and Separate Batch Normalization Head

On UAV and other edge devices, detection head must balance high accuracy and low computational cost. However, conventional designs typically use separate convolutional branches for each prediction task, which leads to parameter redundancy. To address this issue, we propose the SCSHead, as shown in Figure 8.

 doi:10.20944/preprints202506.2235.v1

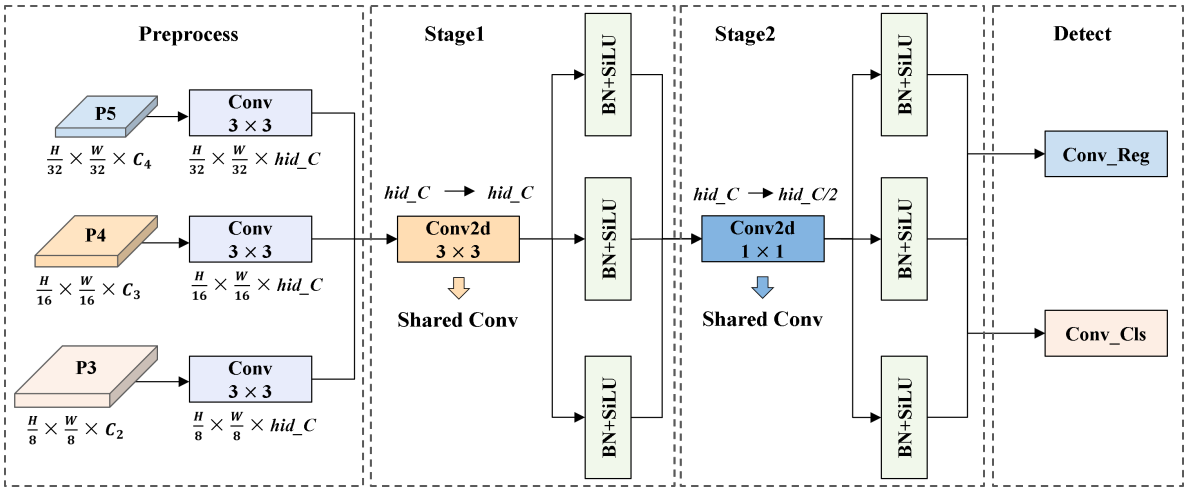

Figure 8. Architecture of the proposed SCSHead.

SCSHead adopts a shared-weight convolutional architecture. Feature maps from different scales are first processed by individual 3×3 convolutional layers to align their channel dimensions, thereby ensuring consistent cross-scale feature representation. The aligned feature maps are then passed through a shared-weight convolutional module consisting of two successive convolutional stages. First, a 3×3 convolutional layer is applied to extract cross-scale features and enhance representational capacity. Subsequently, a 1×1 convolutional layer performs channel-wise compression and reorganization, thereby reducing the number of parameters and computational cost.

To accommodate statistical discrepancies across feature maps of different scales, we introduce a scale-specific normalization strategy. Specifically, each input branch applies an independent BN layer. This design allows each scale to adaptively normalize its feature statistics, thereby preserving and emphasizing the discriminative properties of individual feature maps.

4.4. Network structure of ELS-YOLO

The overall architecture of the proposed ELS-YOLO is shown in Figure 9. The input image is first processed by the ER-HGNetV2 backbone to extract hierarchical features. These features are passed through LFSPN for dynamic cross-scale fusion to improve multi-scale representation. The fused features are then sent to SCSHead to produce the final detection results.

 doi:10.20944/preprints202506.2235.v1

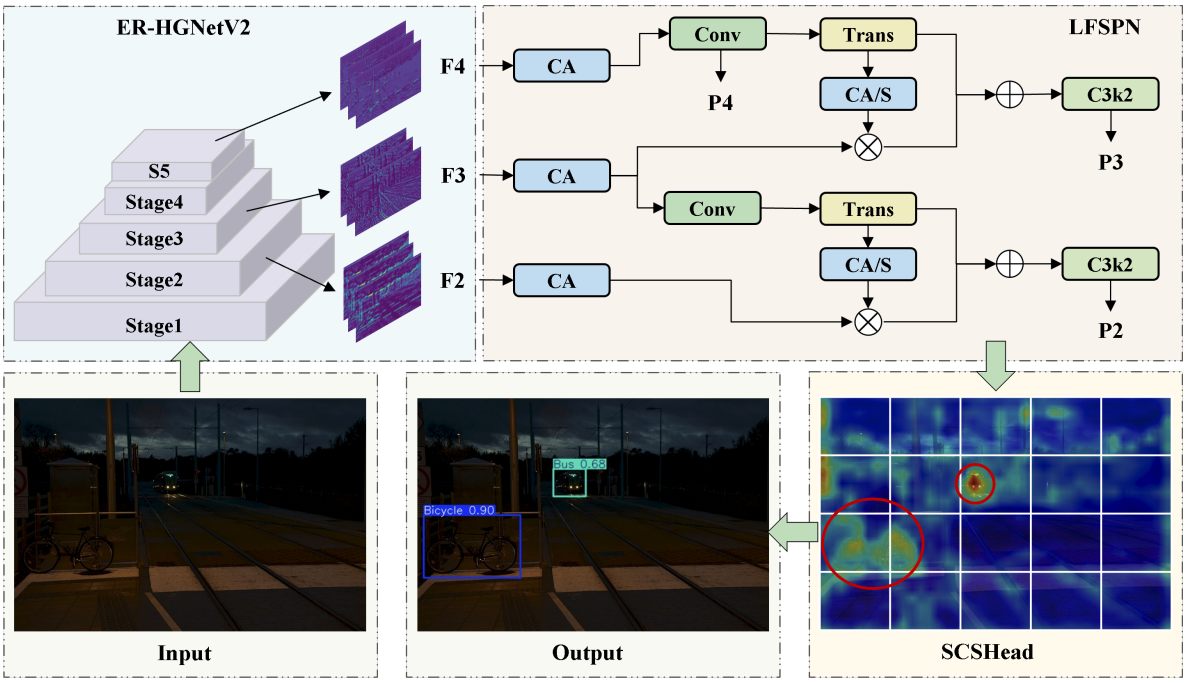

Figure 9. Overall architecture of the proposed ELS-YOLO model.

4.5. Channel Pruning

Deploying object detection systems on UAV platforms necessitates real-time processing of aerial imagery to accurately identify and localize traffic signs, vehicles, pedestrians, and other critical targets. However, UAVs are typically constrained by limited computational resources, memory, and battery capacity. Therefore, detection models must be carefully optimized to reduce both computational complexity and memory consumption. Model pruning has emerged as an effective structural optimization technique that significantly compresses model size while preserving near-original performance. To further improve the real-time inference capability of the proposed model on edge devices, we apply LAMP method to eliminate redundant parameters and reduce model complexity.

Traditional magnitude-based pruning methods apply a global threshold to prune all layers of the network uniformly. However, such approaches fail to account for the varying importance of parameters across layers, often resulting in excessive pruning of critical layers and subsequent degradation in model accuracy. To address this issue, the LAMP method introduces a layer-wise adaptive importance scoring mechanism. By reweighting the importance of weights within each layer, LAMP determines an appropriate pruning ratio for each layer. As shown in Figure 10, for a given layer's weight tensor W, its values are first flattened into a one-dimensional vector and sorted in ascending order based on magnitude. Based on this sorted vector, LAMP assigns each weight u a corresponding score that reflects its relative importance within the layer, as

$$\text{score}(u, W) = \frac{(W[u])^2}{\sum_{v \geq u} (W[v])^2} \tag{3}$$

where the numerator represents the squared value of the current weight, and the denominator denotes the sum of squares of all weights whose magnitudes are greater than or equal to that of the current weight.

 doi:10.20944/preprints202506.2235.v1

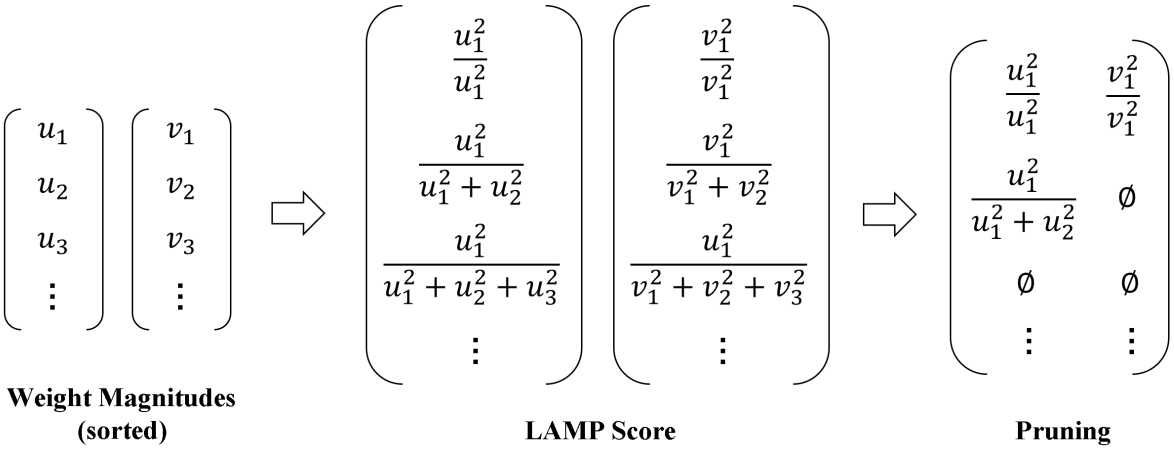

Figure 10. The LAMP pruning process.

The LAMP score reflects the relative contribution of each weight among the remaining connections. During pruning, the model removes connections with lower scores on a per-layer basis until the desired global sparsity level is reached. This strategy accounts for the varying contributions of different layers to the overall network, thereby enabling substantial model compression while preserving detection accuracy to the greatest extent possible.

5. Experimental results

5.1. Dataset

To evaluate the effectiveness and generalizability of the proposed ELS-YOLO model for object detection in low-light conditions, we conducted extensive experiments on two publicly available datasets: ExDark [46] and DroneVehicle [47]. We randomly split the dataset into training, validation, and test sets in an 8:1:1 ratio.

5.1.1. ExDark

The ExDark dataset is designed for object detection under low-light conditions. It contains 7,363 real-world images captured in various challenging lighting environments and provides annotations for 12 object categories. Figure 11 shows typical image examples and a statistical overview of the annotations.

 doi:10.20944/preprints202506.2235.v1

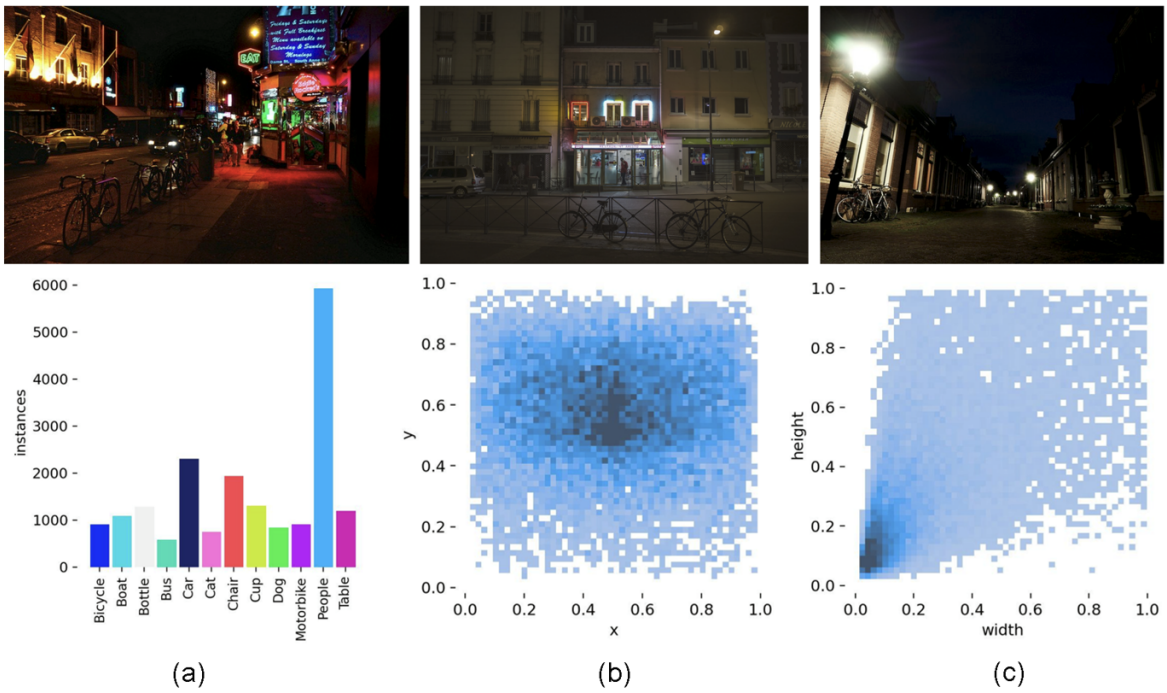

Figure 11. Illustration and statistical analysis of the ExDark dataset. (a) Class distribution of annotated objects; (b) Distribution of normalized bounding box sizes; (c) Spatial distribution of bounding box center points across the image plane.

5.1.2. DroneVehicle

The DroneVehicle dataset contains aerial images captured by UAVs across various urban scenes such as city roads, residential areas, parking lots, and highways. The images are categorized into four lighting conditions: Day, Hazy, Night, and Dark Night. Figure 12 shows typical examples under each condition. Since this study focuses on object detection in low-light environments, we retain only the images labeled as Night and Dark Night.

Figure 12. Images under varying illumination conditions.

5.2. Experimental environment

The experiment is conducted using Python 3.10.16 and PyTorch 2.3.1 on an NVIDIA GeForce RTX 4090 GPU. We use SGD as the optimizer with an initial learning rate of 0.01, momentum set to 0.937, and weight decay of 0.0005. Each model is trained for 200 epochs with a batch size of 16 and no pretrained weights. All comparison experiments use the same settings.

5.3. Evaluation indicators

The performance of the proposed model is evaluated using commonly adopted object detection metrics, including precision, recall, mean Average Precision (mAP), frames per second (FPS), and GFLOPs. Definitions of these metrics are presented below.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4}$$

 doi:10.20944/preprints202506.2235.v1

$$\text{Recall} = \frac{TP}{TP + FN} \tag{5}$$

$$AP = \int_0^1 \text{Precision}(\text{Recall})\, d(\text{Recall}) \tag{6}$$

$$\text{mAP} = \frac{1}{K} \sum_{i=1}^{K} AP_i \tag{7}$$

where TP denotes the number of correctly detected positive samples, FP denotes the number of falsely detected positive samples, FN denotes the number of ground truth objects that were missed by the detector, K denotes the total number of object categories in the dataset, and AP_i represents the average precision for the i-th category.

5.4. Experimental analysis on the ExDark dataset

5.4.1. ER-HGNetv2 experiment

To validate the effectiveness of the proposed ER-HGNetV2 backbone for low-light object detection, we conducted comparative experiments using YOLOv11s as the baseline and replacing its backbone with ER-HGNetV2 and other existing lightweight alternatives. Table 1 presents the experimental results on the ExDark dataset.

Table 1. Comparative experiments with different backbone networks.

Backbone	mAP@0.5/%	mAP/%	Params/M	GFLOPs/G
EfficientViT[41]	68.5	43.1	7.98	19.0
RepViT[42]	69.3	43.9	10.14	23.5
HGNetV2	<u>69.7</u>	<u>44.6</u>	<u>7.61</u>	18.9
MobileNetV4[43]	66.3	41.9	9.53	27.8
StarNet[44]	65.8	40.1	8.63	**17.6**
ER-HGNetV2	**72.6**	**46.5**	**7.17**	<u>18.3</u>

The best value for each metric is shown in bold and the second-best is underlined.

The proposed ER-HGNetV2 backbone demonstrates clear advantages across all evaluation metrics, achieving the highest scores in both mAP@0.5 and mAP. Compared to the original HGNetV2, ER-HGNetV2 yields better detection performance while incurring lower computational complexity.

5.4.2. Comparison with YOLOv11

To evaluate the performance advantages of ELS-YOLO over the YOLOv11 series models, we conducted a quantitative analysis on the ExDark dataset. The experimental results are presented in Table 2.

Table 2. Comparison with the YOLOv11 series on the ExDark dataset

Models	mAP@0.5/%	mAP/%	Params/M	GFLOPs/G
YOLO11n	67.6	42.2	**2.6**	**6.3**
YOLO11s	71.4	45.7	9.4	21.3
YOLO11m	73.2	47.7	20.0	67.7
YOLO11l	<u>74.6</u>	<u>48.9</u>	25.2	86.6
YOLO11x	**75.7**	**49.7**	56.8	194.5
ELS-YOLO	74.3	48.5	<u>4.6</u>	<u>15.0</u>

The best value for each metric is shown in bold and the second-best is underlined.

In terms of detection performance, ELS-YOLO achieves a mAP@0.5 of 74.3% and mAP of 48.5%. Compared to the baseline model YOLOv11s, ELS-YOLO improves mAP@0.5 by 2.9% and mAP by

 doi:10.20944/preprints202506.2235.v1

2.8%, demonstrating stronger robustness in low-light conditions. Regarding model complexity, ELS-YOLO contains only 48.9% of the parameters and 70.4% of the computational cost of YOLOv11s, demonstrating excellent lightweight characteristics. Figure 13 illustrates the training performance curves of four key metrics for both ELS-YOLO and YOLOv11s. The proposed ELS-YOLO outperforms YOLOv11s across all four metrics throughout the training process.

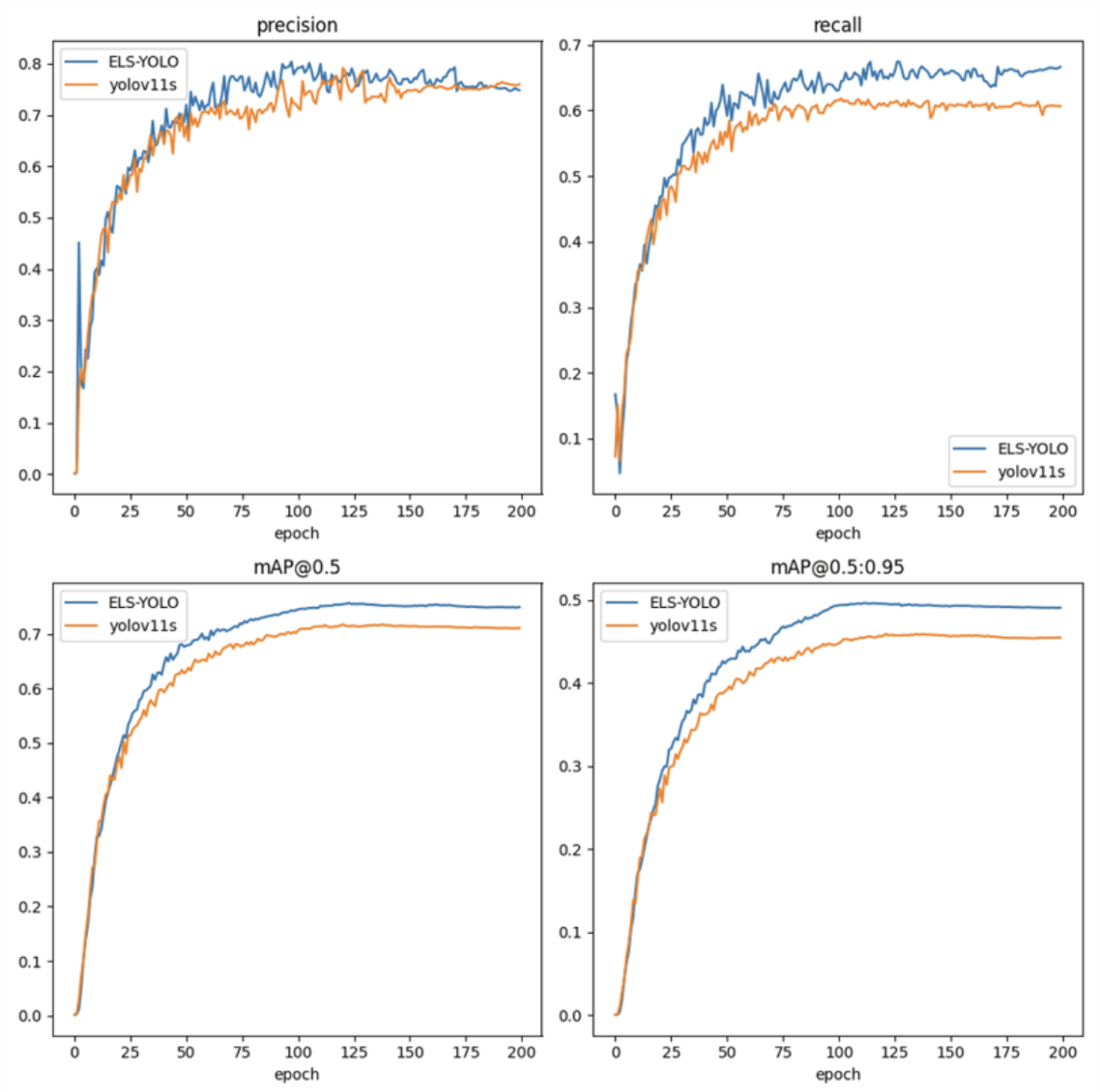

Figure 13. Training curves of ELS-YOLO and YOLOv11s.

5.4.3. LAMP experiment

We conducted a detailed analysis of the ELS-YOLO model's performance under different pruning ratios. The pruning ratio is defined as the ratio of computational cost before pruning to that after pruning. For example, a pruning ratio of 1.33 indicates a 25% reduction in computation, a ratio of 2 indicates a 50% reduction, and a ratio of 4 indicates a 75% reduction.

As shown in Table 3, when the pruning ratio is set to 1.33, the number of parameters is reduced from 4.6M to 2.4M, corresponding to a compression rate of 47.8%. In terms of detection performance, the mAP@0.5 remains at 74.3%, consistent with the unpruned model, while mAP decreases by only 0.1%. When the pruning ratio is increased to 2.0, the model achieves an optimal balance between compression and performance. Specifically, the parameter count further decreases to 1.3M, and the mAP@0.5 is maintained at 74.2%. However, when the pruning ratio increases to 4.0, although the

 doi:10.20944/preprints202506.2235.v1

number of parameters continue to decline, detection performance drops sharply, with mAP@0.5 decreasing to 62.4% and mAP falling to 37.5%.

Table 3. Model performance under different pruning rates

Models	mAP@0.5/%	mAP/%	Params/M	GFLOPs/G
ELS-YOLO	**74.3**	**48.5**	4.6	15.0
ELS-YOLO(ratio=1.33)	**74.3**	48.4	2.4	11.2
ELS-YOLO(ratio=2.0)	74.2	48.1	1.3	7.4
ELS-YOLO(ratio=4.0)	62.4	37.5	**0.5**	**3.7**

The best value for each metric is shown in bold and the second-best is underlined.

Based on the above experimental results, the pruning ratio is set to 2.0, which provides the best trade-off between model compactness and detection accuracy.

5.4.4. Ablation experiments

To comprehensively assess the contribution of each key component in the proposed ELS-YOLO, we conducted a series of ablation experiments. The corresponding experimental results are summarized in Table 4.

Table 4. Ablation experiments performed with the proposed ELS-YOLO.

Models	P/%	R/%	mAP@0.5/%	mAP/%	Params/M	GFLOPs/G
YOLOv11s	78.7	60.5	71.4	45.7	9.4	21.3
+A	**79.8**	63.1	72.6	46.5	7.6	18.3
+A+B	78.5	65.4	73.8	47.9	7.4	18.1
+A+B+C	79.2	**65.8**	**74.3**	**48.5**	**4.6**	**15.0**

A denotes ER-HGNetV2, B denotes SCSHead, and C denotes LFSPN.

First, we replace the original backbone with the proposed ER-HGNetV2 to enhance the model's ability to represent and generalize target features under low-light conditions. As shown in Table 4, this substitution resulted in a 1.2 increase in the mAP@0.5 metric, while the computational complexity was reduced by 3 GFLOPs. These results indicate that ER-HGNetV2 effectively eliminates redundant computation and improves the network's capacity to extract complex features in dark environments.

Next, we applied the proposed SCSHead, which further increased the mAP@0.5 to 73.8%. This result indicates that SCSHead, through its shared convolutional structure and scale-adaptive normalization, enhances the network's robustness and discriminative capacity for multi-scale object detection in low-light conditions.

Finally, we introduce LFSPN to enhance multi-scale feature fusion. The addition of LFSPN improved mAP@0.5 and mAP by 0.5% and 0.6%, respectively. Meanwhile, the model's parameter and computational complexity were significantly reduced. These results confirm that LFSPN effectively integrates multi-scale information while suppressing redundant features, thereby improving both the network's generalization capability and computational efficiency.

5.4.5. Comparison experiments with other baseline methods

To further evaluate the effectiveness and performance advantages of the proposed ELS-YOLO, we conduct comparison experiments under the same settings with several mainstream detection models.

As presented in Table 5, ELS-YOLO demonstrates clear advantages across all key performance metrics, outperforming lightweight YOLO variants and DETR-based models. In terms of model complexity, ELS-YOLO maintains a compact architecture with only 4.6M parameters and 15.0 GFLOPs, which is significantly lower than that of larger models such as Faster R-CNN and DETR. These results further support the feasibility of deploying ELS-YOLO for real-time inference on resource-constrained edge devices.

 doi:10.20944/preprints202506.2235.v1

Table 5. Comparison with other models on the ExDark dataset.

Models	P/%	R/%	mAP@0.5/%	mAP/%	Params/M	GFLOPs/G
YOLOv8n	70.2	59.6	65.7	41.1	3.0	8.1
YOLOv8s	73.9	62.7	70.4	44.3	11.1	28.5
YOLOv9t	74.0	56.7	65.2	40.8	**2.0**	**7.6**
YOLOv9s	74.1	62.1	69.8	44.8	7.2	26.8
YOLOv10n	71.8	58.1	65.0	40.5	2.7	8.2
YOLOv10s	77.2	60.2	69.0	43.8	8.1	24.5
Faster R-CNN	67.4	52.6	58.9	35.2	68.9	80.2
RetinaNet	66.3	50.7	57.6	33.9	41.2	78.2
DETR	71.9	57.3	63.8	39.7	40.8	186.2
RT-DETR-r50	75.4	61.5	67.1	42.2	41.9	125.7
RT-DETR-L	73.1	58.1	64.6	39.9	32.0	103.5
ELS-YOLO	**79.2**	**65.8**	**74.3**	**48.5**	4.6	15.0

The best value for each metric is shown in bold and the second-best is underlined.

5.4.6. Visualization analysis

To intuitively evaluate the detection performance of the proposed ELS-YOLO under low-light conditions, four representative scenes from the ExDark dataset were selected. These scenes encompass various challenges, including small object detection, multi-scale targets, severe occlusion, and complex lighting environments.

From the visualization results in Figure 14, it is evident that ELS-YOLO demonstrates superior object detection and localization capabilities compared to YOLOv11s. Specifically, in the first row, under extremely poor illumination, ELS-YOLO accurately detects partially occluded targets in dim lighting conditions. In the second row, which depicts a distant and weakly illuminated scene, ELS-YOLO effectively detects small-scale pedestrians and vehicles, whereas YOLOv11s exhibits clear omissions. These results suggest that the ER-HGNetV2 backbone and the LFSPN structure enhance ELS-YOLO's ability to capture fine-grained features and small object information in low-light environments. In the third row, representing a multi-scale detection scenario, YOLOv11s shows noticeable false detections. In contrast, ELS-YOLO successfully identifies multiple object instances with higher precision and better localization, highlighting the benefits of the proposed SCSHead in cross-scale feature sharing and adaptive normalization. Additionally, in the fourth row, an urban night scene, ELS-YOLO provides clear and accurate detections across objects of various scales. In comparison, YOLOv11s performs poorly on distant targets, leading to both false negatives and false positives.

We use Gradient-weighted Class Activation Mapping (Grad-CAM)[45] to visualize the focus regions of both ELS-YOLO and the baseline YOLOv11s during prediction. As shown in Figure 15, ELS-YOLO exhibits more accurate attention to critical object regions compared to YOLOv11s. In the first row, under a dimly lit scene, YOLOv11s displays weak activation for distant pedestrians and bicycles. In contrast, ELS-YOLO significantly enhances attention to these small objects. In the second row, which depicts a complex nighttime environment, YOLOv11s either overlooks or diffusely attends to vehicles and distant pedestrians. However, ELS-YOLO produces more concentrated attention centered on the key regions of the targets. The third row further confirms the robustness and superiority of ELS-YOLO. Under partial occlusion, YOLOv11s generates vague and dispersed attention, whereas ELS-YOLO accurately highlights the partially occluded small objects.

 doi:10.20944/preprints202506.2235.v1

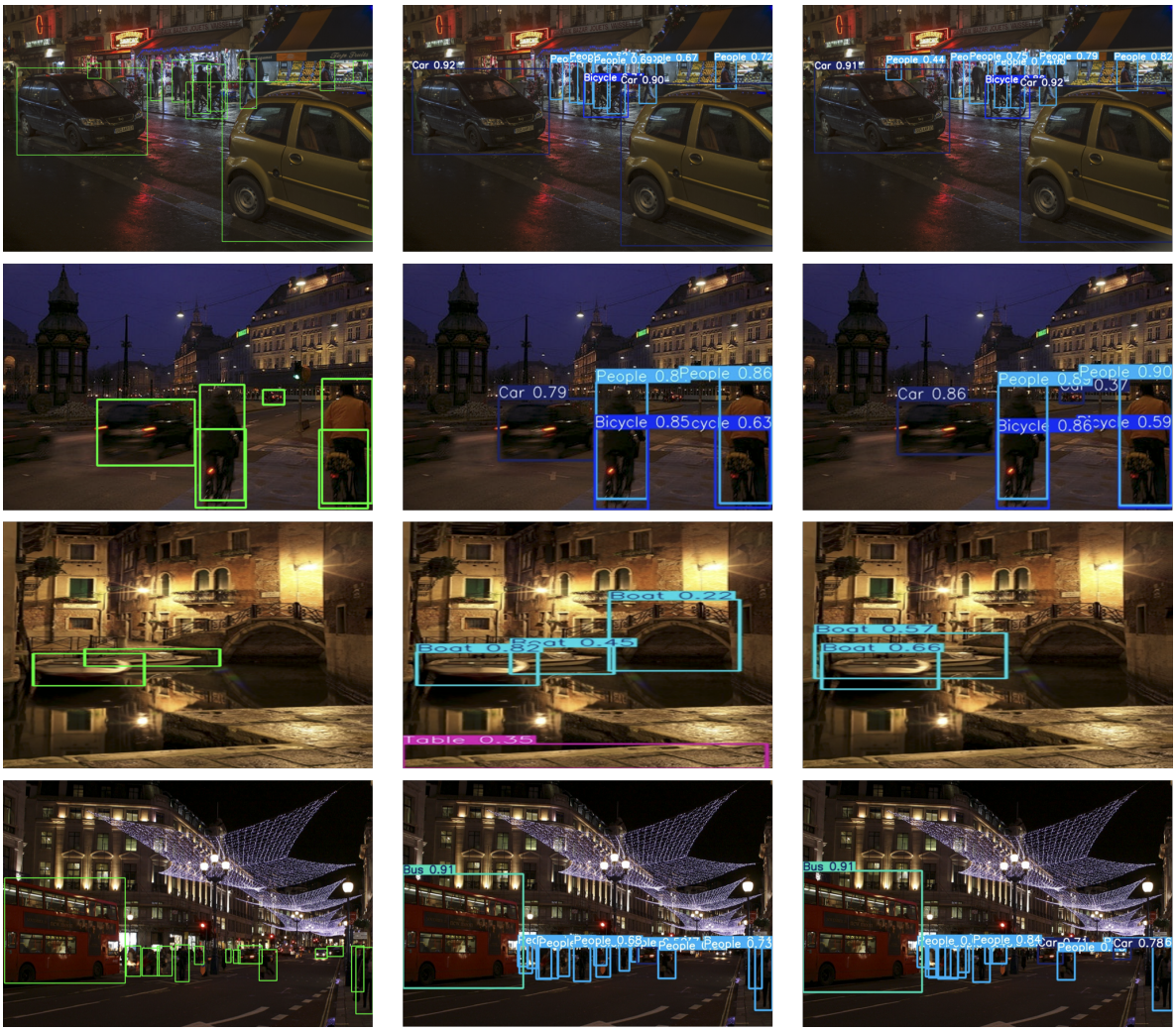

Figure 14. Detection results of different models on the ExDark dataset. From left to right: Ground Truth,YOLOv11s,ELS-YOLO.

 doi:10.20944/preprints202506.2235.v1

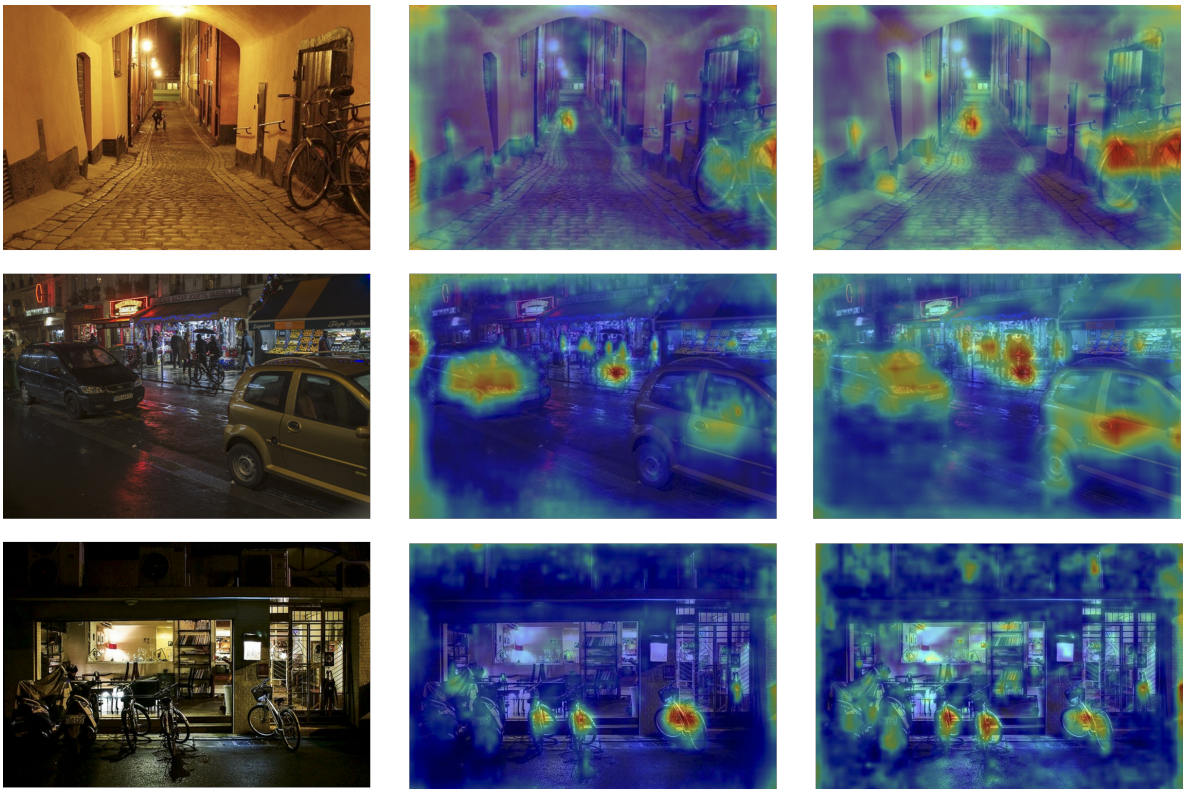

Figure 15. Heatmap results. From left to right: Original images,YOLOv11s,ELS-YOLO.

5.5. Experimental analysis on the DroneVehicle dataset

To further assess the performance of the proposed ELS-YOLO model in DVOD tasks, we conducted experiments on the DroneVehicle dataset. This dataset comprises aerial images captured by UAVs in various urban scenarios, including city roads, residential areas, parking lots, and highways. Representative inference results are presented in Figure 16, and the corresponding performance comparison is summarized in Table6.

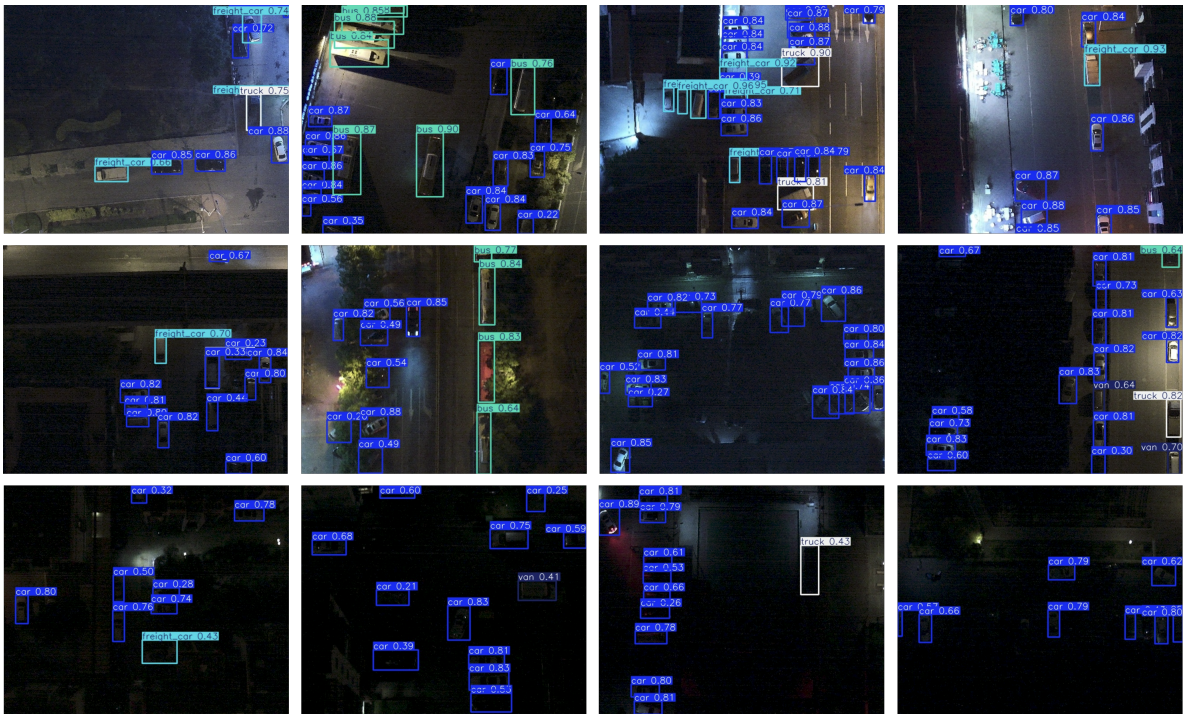

Figure 16. Detection results of ELS-YOLO on the DroneVehicle dataset.

 doi:10.20944/preprints202506.2235.v1

Table 6. Comparative experiments of different object detection algorithms on the DroneVehicle dataset.

Models	P/%	R/%	mAP@0.5/%	mAP/%
YOLO11n	58.4	58.9	61.7	38.6
YOLO11s	**68.2**	63.7	67.2	42.9
RT-DETR-r50	65.7	63.2	66.7	41.2
RT-DETR-L	67.9	66.4	68.1	43.3
ELS-YOLO	68.3	**67.5**	**68.7**	**44.5**
ELS-YOLO (ratio=2.0)	68.2	67.3	68.5	44.2

The best value for each metric is shown in bold and the second-best is underlined.

As shown in Table 6, ELS-YOLO achieves a precision of 68.3%, recall of 67.5%, mAP@0.5 of 68.7%, and mAP of 44.5%. Compared to the baseline model YOLOv11s, ELS-YOLO improves mAP@0.5 by 1.5% and mAP by 1.6%, demonstrating superior target localization capabilities.

6. Discussion

We conducted extensive experiments on the representative ExDark and DroneVehicle datasets to evaluate the object detection performance of ELS-YOLO under complex low-light environments. These datasets include diverse low-light scenarios and drone perspectives characterized by substantial scale variations and complex backgrounds. The experimental results indicate that ELS-YOLO achieves mAP@0.5 of 74.3% and 68.7% on ExDark and DroneVehicle, respectively, significantly surpassing YOLOv11s and other mainstream detection models. This performance enhancement is primarily attributed to structural innovations, resulting in superior capabilities in feature extraction and fusion strategies. Compared to other lightweight models, ELS-YOLO exhibits advantages in parameter scale, computational complexity, and inference efficiency, highlighting its potential for deployment in edge computing environments.

Although ELS-YOLO demonstrates strong overall performance, there remains room for improvement under certain conditions. For instance, its detection accuracy still needs enhancement in scenarios involving extreme low-light conditions or severe occlusion. In addition, although this study has significantly reduced the model's parameter size, further improving inference speed without sacrificing accuracy remains an important direction for future research.

7. Conclusions

In this paper, we propose the lightweight object detection model ELS-YOLO to address the challenges of limited detection performance in complex low-light conditions and constrained computational resources on edge devices. We design the re-parameterized backbone ER-HGNetV2 to enhance the ability to extract and represent critical features under low-light environments. To overcome the limitations of conventional fusion strategies, we introduce LFSPN, which enables efficient multi-scale feature integration. We also develop the lightweight detection head SCSHead to reduce computational cost and parameter count. Furthermore, we apply the LAMP pruning strategy to compress model size without sacrificing accuracy. Extensive experiments on the ExDark and DroneVehicle datasets demonstrate that ELS-YOLO achieves superior detection accuracy and real-time inference efficiency compared to existing lightweight models.

In future work, we will further explore optimization strategies for the model under extreme conditions, such as incorporating advanced attention mechanisms or dynamic feature fusion methods, to enhance the model's adaptability in complex environments. Additionally, we will investigate integrating techniques like knowledge distillation and quantization training to further improve the real-time inference capability of the model, thereby facilitating the deployment and application of object detection technologies in a broader range of practical scenarios.

Author Contributions: Conceptualization, T.W. and X.N.; methodology, T.W.; software, T.W.; validation, T.W.; formal analysis, T.W.; investigation, T.W.; resources, T.W.; data curation, T.W.; writing—original draft preparation,

 doi:10.20944/preprints202506.2235.v1

T.W.; writing—review and editing, X.N.; visualization, T.W.; supervision, X.N.; project administration, X.N.; funding acquisition, X.N. All authors have read and agreed to the published version of the manuscript.

Funding: "This research was funded by National Natural Science Foundation of China grant number 62472010."

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available from the corresponding author upon reasonable request.

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

References

1. P. Zhu, L. Wen, D. Du, X. Bian, H. Fan, Q. Hu, and H. Ling, "Detection and Tracking Meet Drones Challenge," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 11, pp. 7380–7399, 2022. doi: 10.1109/TPAMI.2021.3119563.
2. X. Wu, D. Hong, and J. Chanussot, "Convolutional Neural Networks for Multimodal Remote Sensing Data Classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–10, 2022. doi: 10.1109/TGRS.2021.3124913.
3. Y. Huang, J. Chen, and D. Huang, "UFPMP-Det: Toward Accurate and Efficient Object Detection on Drone Imagery," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022.
4. J. Zhan, Y. Hu, G. Zhou, Y. Wang, W. Cai, and L. Li, "A high-precision forest fire smoke detection approach based on ARGNet," *Computers and Electronics in Agriculture*, vol. 196, p. 106874, 2022. doi: 10.1016/j.compag.2022.106874.
5. L. Zhou, Y. Dong, B. Ma, et al., "Object detection in low-light conditions based on DBS-YOLOv8," *Cluster Computing*, vol. 28, no. 55, 2025. doi: 10.1007/s10586-024-04829-1.
6. R. Kaur and S. Singh, "A comprehensive review of object detection with deep learning," *Digital Signal Processing*, vol. 132, pp. 103812, 2023. doi: 10.1016/j.dsp.2022.103812.
7. X. Liu, Z. Wu, A. Li, et al., "NTIRE 2024 Challenge on Low Light Image Enhancement: Methods and Results," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 6571–6594, 2024. doi: 10.1109/CVPRW63382.2024.00655.
8. X. Guo, Y. Li, and H. Ling, "LIME: Low-Light Image Enhancement via Illumination Map Estimation," *IEEE Transactions on Image Processing*, vol. 26, no. 2, pp. 982–993, 2017. doi: 10.1109/TIP.2016.2639450.
9. C. Hu, W. Yi, K. Hu, Y. Guo, X. Jing, and P. Liu, "FHSI and QRCPE-Based Low-Light Enhancement With Application to Night Traffic Monitoring Images," *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 7, pp. 6978–6993, 2024. doi: 10.1109/TITS.2023.3342799.
10. J. J. Jeon, J. Y. Park, and I. K. Eom, "Low-light image enhancement using gamma correction prior in mixed color spaces," *Pattern Recognition*, vol. 146, p. 110001, 2024. doi: 10.1016/j.patcog.2023.110001.
11. J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 779–788, 2016. doi: 10.1109/CVPR.2016.91.
12. C.-Y. Wang, I.-H. Yeh, and H.-Y. M. Liao, "YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information," in *Proceedings of the European Conference on Computer Vision (ECCV)*, Milan, Italy, pp. 1–21, 2024. doi: 10.1007/978-3-031-72751-1_1.
13. A. Wang, H. Chen, L. Liu, K. Chen, Z. Lin, J. Han, and G. Ding, "YOLOv10: Real-Time End-to-End Object Detection," in *Advances in Neural Information Processing Systems*, vol. 37, pp. 107984–108011, 2024.
14. M. Everingham, L. Van Gool, C. K. I. Williams, et al., "The PASCAL Visual Object Classes (VOC) Challenge," *International Journal of Computer Vision*, vol. 88, pp. 303–338, 2010. doi: 10.1007/s11263-009-0275-4.
15. T.-Y. Lin, M. Maire, S. Belongie, et al., "Microsoft COCO: Common Objects in Context," in *Computer Vision – ECCV 2014*, Lecture Notes in Computer Science, vol. 8693, Springer, Cham, 2014. doi: 10.1007/978-3-319-10602-1_48.
16. X. Ding, X. Zhang, N. Ma, J. Han, G. Ding, and J. Sun, "RepVGG: Making VGG-style ConvNets Great Again," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13728–13737, 2021. doi: 10.1109/CVPR46437.2021.01352.

 doi:10.20944/preprints202506.2235.v1

17. Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11531–11539, 2020. doi: 10.1109/CVPR42600.2020.01155.
18. J. Lee, S. Park, S. Mo, S. Ahn, and J. Shin, "Layer-Adaptive Sparsity for the Magnitude-Based Pruning," in *Proceedings of the 9th International Conference on Learning Representations (ICLR)*, 2021.
19. S. Deng, S. Li, K. Xie, W. Song, X. Liao, A. Hao, and H. Qin, "A Global-Local Self-Adaptive Network for Drone-View Object Detection," *IEEE Transactions on Image Processing*, vol. 30, pp. 1556–1569, 2021. doi: 10.1109/TIP.2020.3045636.
20. Y. Bai, Y. Zhang, M. Ding, and B. Ghanem, "SOD-MTGAN: Small Object Detection via Multi-Task Generative Adversarial Network," in *Computer Vision – ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Lecture Notes in Computer Science, vol. 11217, Springer, Cham, 2018. doi: 10.1007/978-3-030-01261-8_13.
21. Y. Xi, J. Zheng, X. He, W. Jia, H. Li, Y. Xie, M. Feng, and X. Li, "Beyond context: Exploring semantic similarity for small object detection in crowded scenes," *Pattern Recognition Letters*, vol. 137, pp. 53–60, 2020. doi: 10.1016/j.patrec.2019.03.009.
22. H. Qiu, H. Li, Q. Wu, F. Meng, L. Xu, K. N. Ngan, and H. Shi, "Hierarchical context features embedding for object detection," *IEEE Transactions on Multimedia*, vol. 22, no. 12, pp. 3039–3050, 2020. doi: 10.1109/TMM.2020.2971175.
23. G. Li, Z. Liu, D. Zeng, W. Lin, and H. Ling, "Adjacent context coordination network for salient object detection in optical remote sensing images," *IEEE Transactions on Cybernetics*, vol. 53, no. 1, pp. 526–538, 2023. doi: 10.1109/TCYB.2022.3162945.
24. W. Zhao, Y. Kang, H. Chen, Z. Zhao, Z. Zhao, and Y. Zhai, "Adaptively attentional feature fusion oriented to multiscale object detection in remote sensing images," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–11, 2023. doi: 10.1109/TIM.2023.3246536.
25. Y. Sun, B. Cao, P. Zhu, and Q. Hu, "Drone-based RGB-Infrared cross-modality vehicle detection via uncertainty-aware learning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 10, pp. 6700–6713, 2022. doi: 10.1109/TCSVT.2022.3168279.
26. H. Farid, "Blind inverse gamma correction," *IEEE Transactions on Image Processing*, vol. 10, no. 10, pp. 1428–1433, 2001. doi: 10.1109/83.951529.
27. K. Zuiderveld, "Contrast limited adaptive histogram equalization," in *Graphics Gems IV*, USA: Academic Press Professional, Inc., 1994, pp. 474–485.
28. M. Li, J. Liu, W. Yang, X. Sun, and Z. Guo, "Structure-revealing low-light image enhancement via robust Retinex model," *IEEE Transactions on Image Processing*, vol. 27, no. 6, pp. 2828–2841, 2018. doi: 10.1109/TIP.2018.2810539.
29. K. G. Lore, A. Akintayo, and S. Sarkar, "LLNet: A deep autoencoder approach to natural low-light image enhancement," *Pattern Recognition*, vol. 61, pp. 650–662, 2017. doi: 10.1016/j.patcog.2016.06.008.
30. X. Li, W. Wang, X. Feng, and M. Li, "Deep parametric Retinex decomposition model for low-light image enhancement," *Computer Vision and Image Understanding*, vol. 241, p. 103948, 2024. doi: 10.1016/j.cviu.2024.103948.
31. C. Guo, C. Li, J. Guo, C. C. Loy, J. Hou, S. Kwong, and R. Cong, "Zero-reference deep curve estimation for low-light image enhancement," in *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1777–1786, 2020. doi: 10.1109/CVPR42600.2020.00185.
32. C. Li, C. Guo, and C. C. Loy, "Learning to enhance low-light image via zero-reference deep curve estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 8, pp. 4225–4238, 2022. doi: 10.1109/TPAMI.2021.3063604.
33. X. Xu, R. Wang, C.-W. Fu, and J. Jia, "SNR-aware low-light image enhancement," in *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17693–17703, 2022. doi: 10.1109/CVPR52688.2022.01719.
34. L. Ma, T. Ma, R. Liu, X. Fan, and Z. Luo, "Toward fast, flexible, and robust low-light image enhancement," in *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5627–5636, 2022. doi: 10.1109/CVPR52688.2022.00555.
35. J. Hu and Z. Cui, "YOLO-Owl: An occlusion aware detector for low illuminance environment," in *Proceedings of the 2023 3rd International Conference on Neural Networks, Information and Communication Engineering (NNICE)*, pp. 167–170, 2023. doi: 10.1109/NNICE58320.2023.10105800.

 doi:10.20944/preprints202506.2235.v1

36. Y. Zhang, C. Wu, T. Zhang, Y. Liu, and Y. Zheng, "Self-attention guidance and multiscale feature fusion-based UAV image object detection," *IEEE Geoscience and Remote Sensing Letters*, vol. 20, pp. 1–5, 2023. doi: 10.1109/LGRS.2023.3265995.
37. R. Wu, W. Huang, and X. Xu, "AE-YOLO: Asymptotic enhancement for low-light object detection," in *Proceedings of the 2024 17th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pp. 1–6, 2024. doi: 10.1109/CISP-BMEI64163.2024.10906253.
38. R. Khanam and M. Hussain, "YOLOv11: An Overview of the Key Architectural Enhancements," *arXiv preprint*, arXiv:2410.17725, 2024. doi: 10.48550/arXiv.2410.17725.
39. Y. Zhao, W. Lv, S. Xu, J. Wei, G. Wang, Q. Dang, Y. Liu, and J. Chen, "DETRs Beat YOLOs on Real-time Object Detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16965–16974, 2024. doi: 10.1109/CVPR52733.2024.01605.
40. Q. Hou, D. Zhou, and J. Feng, "Coordinate Attention for Efficient Mobile Network Design," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13708–13717, 2021. doi: 10.1109/CVPR46437.2021.01350.
41. H. Cai, J. Li, M. Hu, C. Gan, and S. Han, "EfficientViT: Lightweight Multi-Scale Attention for High-Resolution Dense Prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 17256–17267. doi: 10.1109/ICCV51070.2023.01587.
42. A. Wang, H. Chen, Z. Lin, J. Han, and G. Ding, "Rep ViT: Revisiting Mobile CNN From ViT Perspective," in *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 15909–15920. doi: 10.1109/CVPR52733.2024.01506.
43. D. Qin *et al.*, "MobileNetV4: Universal Models for the Mobile Ecosystem," in *Computer Vision – ECCV 2024*, A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, and G. Varol, Eds. Cham: Springer, 2025, Lecture Notes in Computer Science, vol. 15098. doi: 10.1007/978-3-031-73661-2_5.
44. X. Ma, X. Dai, Y. Bai, Y. Wang, and Y. Fu, "Rewrite the Stars," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024, pp. 5694–5703. doi: 10.1109/CVPR52733.2024.00544.
45. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626, 2017. doi: 10.1109/ICCV.2017.74.
46. Y. P. Loh and C. S. Chan, "Getting to Know Low-light Images with The Exclusively Dark Dataset," *Computer Vision and Image Understanding*, vol. 178, pp. 30–42, 2019. doi: 10.1016/j.cviu.2018.10.010.
47. Y. Sun, B. Cao, P. Zhu, and Q. Hu, "Drone-based RGB-Infrared Cross-Modality Vehicle Detection via Uncertainty-Aware Learning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, pp. 1–1, 2022. doi: 10.1109/TCSVT.2022.3168279.