Article

Lightweight Zero-DCE-Driven YOLO for Nighttime Low-Light Object Detection

Tianxiang Wang*, Abudhahir Buhari

Faculty of Engineering, Science and Technology, Kuala Lumpur University of Science and Technology, Kajang, 43000, Selangor, Malaysia

* Email: 242924823@s.klust.edu.my

Published: 7 November 2026 | https://doi.org/10.64376/dgnynx73

Abstract: To address the difficulty of balancing accuracy improvement and computational overhead in the joint deployment of low-light enhancement and target detection, this paper constructs a lightweight Zero-DCE-driven YOLO framework for nighttime low-light object detection. The framework integrates depthwise separable convolution, spatial attention, and SE attention in a task-oriented manner: based on the brightness distribution, local gradient, and channel response characteristics of low-light images, the applicable conditions of these common components are analyzed, and the enhancement and detection modules are jointly optimized so that they collaborate effectively for target detection. ExDark experiments show that the complete model improves mAP@0.5 from 74.6% to 87.2%, an improvement of 12.6 percentage points. The end-to-end inference speed reaches 41.3 FPS, compared with 45.2 FPS for the original YOLOv8s. The proposed method thus achieves a substantial accuracy improvement at the cost of a limited 3.9 FPS reduction while maintaining real-time inference capability.

Keywords: Low-light object detection; Zero-DCE; YOLO; Lightweight network; End-to-end training

1. Introduction

Nighttime images captured by visible-light cameras often contain compressed intensity ranges, weak contrast in dark regions, and reduced texture visibility. These degradations directly affect object detection because boundaries and local appearance cues become less reliable. Pedestrians, vehicles, and traffic signs are particularly difficult to distinguish when their contours are weak or their color information is unstable. The problem is relevant to applications such as autonomous driving, surveillance, industrial inspection, and other vision systems that must operate after dark [1]. Detection models trained mainly on well-lit imagery can therefore lose accuracy when the image statistics shift substantially under low illumination [2].

The practical effect of this degradation is application dependent. In urban surveillance, missed pedestrians or vehicles can reduce the reliability of automatic alarms. For autonomous driving, poor nighttime localization of road users and obstacles directly affects the perception pipeline [3]. Similar requirements arise in UAV inspection, industrial parks, and parking facilities, where illumination is often uneven and cannot be controlled. These use cases motivate detection methods that remain efficient while extracting useful object cues from dark images.

YOLO is a representative one-stage detection framework that performs object classification and localization within a unified network [4]. Its relatively simple inference path makes the YOLO family attractive for real-time applications. Existing work on low-light object detection can be grouped into two broad directions. One direction modifies the detector itself to improve feature extraction from dark inputs. This approach keeps the inference pipeline compact, but it cannot recover information that has already been heavily attenuated by poor illumination or noise. LLEYOLO, for example, adapts the YOLOv5 architecture to strengthen feature representation in low-light environments [5].

A second direction enhances the image before detection. Pre-enhancement can reveal structures that are difficult to distinguish in the original input, but the enhancement objective is usually defined independently of the detector. As a result, visually improved images do not necessarily preserve the edges, textures, and scale cues that are most useful for recognition; artifacts, local overexposure, or smoothing may instead be introduced [6]. Zero-DCE is attractive in this setting because it learns pixel-wise curve parameters without paired training images [7,8]. Nevertheless, even a compact enhancer creates an additional processing stage, and the original Zero-DCE losses were designed for image

enhancement rather than for downstream object detection [7].

Recently, several studies have moved toward coordinated enhancement–detection design. For example, adaptive enhancement with dynamic feature fusion has been shown to benefit low-light object detection [9], and lightweight detection frameworks that integrate image enhancement with adaptive feature fusion have also been explored [10]. Similar to these studies, the present work focuses on task-level coordination between enhancement and detection; in contrast, it adopts a Zero-DCE-driven lightweight enhancement module together with an end-to-end joint optimization strategy, and places particular emphasis on the analysis of attention placement and on a unified experimental comparison under a fixed data split.

The method proposed here addresses these issues by coupling a lightweight Zero-DCE variant with a modified YOLOv8s detector. Depthwise separable convolution is used in the enhancement branch to reduce convolutional cost, while spatial attention adjusts the curve-parameter responses in regions that are more informative for detection. The detector is simplified with lightweight backbone modules and supplemented with channel recalibration. The main contribution of this study lies in a task-oriented integration of existing components rather than in proposing new convolution or attention operators: we analyze the applicable conditions of attention mechanisms under low-light conditions, examine where spatial and channel attention are most effective within the enhancement–detection pipeline, compare a frozen two-stage enhancement-detection pipeline with joint optimization under the same data split, and evaluate the resulting accuracy, speed, and parameter trade-off. The reported 41.3 FPS is treated as end-to-end throughput of the complete system rather than as an acceleration over the original YOLOv8s.

2. Methods

2.1 Design of the Lightweight Zero-DCE Image Enhancement Module

Zero-DCE formulates low-light enhancement as a learnable pixel-wise curve transformation. Instead of applying a fixed global threshold or manually selected correction parameter, the network predicts curve parameters from the input image and applies them directly to pixel intensities [7]. This formulation allows the adjustment to vary across the image and is therefore suitable for scenes in which illumination is spatially non-uniform. The goal in this study is to retain that adaptive mapping while reducing the cost of using the enhancer inside a detection pipeline.

The original Zero-DCE predicts pixel-wise curve parameters with a seven-layer convolutional network and contains about 79K (0.079M) trainable parameters. Although the parameter count is small, placing the enhancer in front of a detector still adds a complete image-processing stage and increases end-to-end latency. In the redesigned branch, standard convolutions are replaced by depthwise separable convolutions [11]. Each depthwise operation processes channels independently, and the following 1×1 pointwise convolution mixes information across channels. This factorization reduces convolutional computation while preserving cross-channel interaction. Spatial attention is then added to the enhancer so that the predicted curve parameters can respond more strongly to locally informative regions in dark scenes.

The redesigned lightweight Zero-DCE module contains five depthwise separable convolutional layers, each followed by a LeakyReLU activation function. The network takes a three-channel low-light image as input and outputs a 24-channel curve-parameter map, corresponding to eight sets of curve parameters for the three RGB channels. Brightness adjustment is progressively achieved through eight iterative curve mappings. The curve transformation is defined in Eq. (1). The depth of five layers was chosen as a balance between estimation capacity and computational cost: it retains sufficient capacity for pixel-wise curve-parameter estimation while substantially reducing the computational overhead of the original seven-layer stack, and its overall effectiveness is reflected by the pipeline-level results in Section 3.2.

$$I^{(\mathrm{n})}(x) = I^{(\mathrm{n}-1)}(x) + \alpha_n(x) I^{(\mathrm{n}-1)}(x)[1 - I^{(\mathrm{n}-1)}(x)] \quad (1)$$

where $I^{(0)}(x)$ denotes the original low-light pixel value at position x, $I^{(\mathrm{n})}(x)$ denotes the pixel value after the n-th curve mapping, and $\alpha_n(x)$ denotes the curve parameter corresponding to the n-th iteration, where $n = 1, \ldots, 8$. After eight recursive curve mappings, the final enhanced result $I_{\mathrm{enh}}(x) = I^{(8)}(x)$ is obtained.

For RGB images, each iteration corresponds to a set of curve parameters for the three color channels; therefore, the network ultimately predicts a 24-channel parameter map.

Repeated curve mapping changes the brightness progressively rather than through a single global correction. This helps retain the ordering of pixel intensities and limits abrupt color changes. Spatial attention is applied to the predicted curve-parameter map rather than used as a hard region selector. In this way, local responses around objects such as pedestrians and vehicles can influence the strength of enhancement without discarding the remaining image content. The modified Zero-DCE contains approximately 0.8M trainable parameters. Although this is larger than the original Zero-DCE, the design is intended to reduce practical computation through depthwise separable convolution and to improve task relevance through spatial weighting. Its efficiency is therefore assessed at the level of the complete enhancement-detection pipeline (Table 1). This placement is motivated by the characteristics of low-light images: illumination varies spatially and the local gradients that define object boundaries are concentrated in specific regions, so directly modulating the per-pixel curve parameters with spatial weights is a natural way to focus enhancement on

detection-relevant areas.

2.2 Lightweight Improvement of the YOLOv8s Backbone Network

Lightweight object detectors have been explored as a way to reduce computation in low-light deployment settings [12,13]. In this study, YOLOv8s is used as the starting detector because it offers a practical compromise between accuracy and throughput [14]. Dark inputs, however, weaken boundaries and textures and can make small pedestrians or distant vehicles difficult to separate from background responses. Related lightweight YOLOv8 designs have also shown that attention can improve small-object representation without necessarily requiring a large model [15]. The backbone is therefore modified with two aims: removing redundant computation and retaining features that remain informative under low illumination. Several C2f blocks are replaced by C2f-Lite modules [16]. These blocks preserve the feature-splitting path of C2f but use depthwise separable convolution and fewer bottleneck units, reducing repeated transformations while maintaining multi-level feature propagation. Blocks located in the deeper backbone stages, where repeated transformations contribute the most redundant computation, were prioritized for replacement, whereas the shallow stages that carry edge and texture responses critical under low illumination were retained.

Channel recalibration is inserted at selected backbone stages. The SE block summarizes each channel by global average pooling, models the channel dependencies with two fully connected layers, and produces multiplicative weights through a Sigmoid activation [17]. These weights rescale feature channels rather than replacing the original feature map. In the neck, selected upsampling operations are replaced with lightweight transposed convolution, and skip connections introduce shallow spatial information into deeper features. The resulting detector has approximately 6.2M parameters, compared with 11.2M for the original YOLOv8s, a reduction of about 44.6%. Its standalone inference speed increases from 45.2 to 49.1 FPS in the test environment. These measurements indicate that the detector-side changes reduce model size while preserving efficient feature extraction. SE recalibration is placed in the detector rather than in the enhancement branch because it operates on the global channel statistics of the feature maps produced by the backbone; the enhancement branch instead deals with spatially varying pixel statistics, which is the role assigned to spatial attention (Section 2.1).

2.3 Applicability and Limitations of Attention Mechanisms under Low-Light Conditions

Spatial attention is most useful when an object still differs from its surroundings through local contrast, edge continuity, or a recognizable semantic response. That condition does not always hold in very dark scenes. If both foreground and background are close to the noise floor, or if reflections create strong bright structures, the attention map can emphasize the wrong region. SE recalibration has a different limitation: it relies on global channel statistics, which can be unrepresentative when useful objects occupy only a small part of the image or when illumination varies sharply across the scene. For this reason, neither attention block is used as an explicit object prior. Spatial weights only modulate the curve parameters, while channel weights only rescale feature amplitudes; residual and skip paths retain the underlying features. This soft use of attention reduces the risk of hard information suppression, although local over-enhancement and over-suppression can still occur. These considerations are analytical rather than independently isolated by dedicated experiments; they are used to motivate the architectural choices, whose combined effectiveness is validated at the pipeline level in Section 3.2. Isolating the contribution of each attention placement through dedicated ablations is left for future work.

2.4 End-to-End Joint Training of Enhancement and Detection

In a conventional two-stage workflow, the enhancer is trained first and then frozen, or enhanced images are generated offline before detector training. This separation is convenient, but the enhancement losses are optimized for image quality rather than specifically for the boundaries, textures, and local contrast used by a detector [18]. To make the comparison explicit, this study uses "Lightweight Zero-DCE + Original YOLOv8s" as the two-stage baseline: the enhancer is pretrained and frozen, and YOLOv8s is trained on the resulting enhanced images. The complete model instead combines the improved YOLOv8s with the enhancer in one computational graph, allowing the detection loss to propagate into the enhancement branch. Because the detector architecture also changes, the difference between the complete model and the two-stage baseline represents the combined effect of detector modification and joint optimization; it should not be attributed to joint training alone.

During joint training, each low-light image first passes through the lightweight Zero-DCE branch and the enhanced output is then processed by the modified YOLOv8s. The detector produces class and localization predictions, while its loss is propagated backward through the enhancement stage. At the same time, the enhancer remains constrained by its image-quality objectives. This coupling allows the curve estimator to respond not only to brightness-related criteria but also to the features that affect recognition, while the detector is trained on the distribution produced by the enhancer [19].

The overall loss function consists of an image enhancement loss and an object detection loss. The image enhancement loss comprises a spatial consistency loss, an exposure control loss, a color constancy loss, and an illumination smoothness loss, and is formulated as follows:

$$L_{\text{enh}} = L_{\text{spa}} + L_{\text{exp}} + L_{\text{col}} + L_{\text{tv}} \tag{2}$$

where L_{spa}, L_{exp}, L_{col}, and L_{tv} denote the spatial consistency loss, exposure control loss, color constancy loss, and illumination smoothness loss, respectively.

The object detection loss adopts the standard YOLOv8 detection loss, which consists of classification loss, bounding-box regression loss, and Distribution Focal Loss (DFL) [14], and is formulated as follows:

$$L_{\text{det}} = L_{\text{cls}} + L_{\text{box}} + L_{\text{DFL}} \tag{3}$$

where L_{cls}, L_{box}, and L_{DFL} denote the classification loss, bounding-box regression loss, and Distribution Focal Loss, respectively. During end-to-end joint training, the image enhancement loss and object detection loss jointly constitute the overall optimization objective of the network:

$$L_{\text{total}} = L_{\text{enh}} + L_{\text{det}} \tag{4}$$

where L_{total} denotes the total loss during the joint training stage.

The detection loss is backpropagated through the complete computational graph to the enhancement module, enabling the enhancement network to preserve basic image-quality constraints while placing greater emphasis on structural and feature information that is beneficial for subsequent object recognition. Formally, the total loss couples the enhancement and detection objectives within a single computational graph: gradients of the detection loss propagate into the enhancement branch, while the enhancement-loss terms continue to constrain brightness, color, and smoothness, thereby steering the curve-parameter estimator toward features that benefit recognition. A quantitative analysis of the gradient interactions and possible weight conflicts among the enhancement loss components under detection backpropagation is beyond the scope of this study and is left for future work.

A staged training strategy is adopted with strictly controlled data boundaries. First, the lightweight Zero-DCE module is pretrained using only the 5,890 images in the ExDark training set, without accessing any test images or test-set bounding-box annotations. The improved YOLOv8s is then integrated and jointly fine-tuned on the same training set. The 1,473 test images remain completely unseen during training, hyperparameter selection, and enhancement-module pretraining. The two-stage baseline also uses only the same 5,890 training images: the enhancer is first frozen to generate enhanced training inputs, after which the detector is trained independently. This setting prevents ExDark test images from entering the training process of either the enhancement or detection module and ensures that all experiments are compared under identical data boundaries.

3. Results

3.1 Experimental Dataset and Evaluation Metrics

Experiments use the publicly available ExDark (Exclusively Dark) dataset, which contains 7,363 low-light images from 12 object categories [20]. A fixed split of 5,890 training images and 1,473 test images was created before training. All model fitting, enhancement pretraining, joint fine-tuning, and hyperparameter selection were restricted to the training subset. The test subset was reserved for final evaluation and was not used to tune either the enhancement or detection branch. The ExDark dataset does not prescribe a single official train/test partition, and different studies adopt different fixed partitions; the partition used here was fixed before any training, is fully reproducible, and is shared by every configuration compared in this study.

The primary evaluation metrics are mAP@0.5, end-to-end inference speed, and parameter count. mAP@0.5 measures the mean detection accuracy across all categories at an IoU threshold of 0.5 [21]. Inference speed was measured with a batch size of 1, with all models using the same input resolution and numerical precision. Before testing, a fixed number of GPU warm-up iterations was performed, after which the average inference time over the complete test process was calculated. End-to-end FPS includes both image enhancement and object detection computation but excludes disk I/O time, and was measured on an NVIDIA RTX 3090 under Ubuntu 20.04 and PyTorch 2.0. The standalone inference speed of the detection subnetwork is reported only for analyzing module-level computational overhead and is not mixed with the end-to-end inference speed of the complete enhancement–detection pipeline. mAP@0.5 was adopted as the primary accuracy metric because it is the most widely reported criterion in low-light detection studies on ExDark and directly supports the accuracy–speed–parameter trade-off analysis of this study.

All models were trained with an input resolution of 640 × 640, a batch size of 8, and the AdamW optimizer with an initial learning rate of 0.01, a final learning-rate factor of 0.01, cosine annealing scheduling, three warm-up epochs, and gradient accumulation over two steps, for 120 epochs. The same split and training protocol were applied to every configuration compared in this study.

3.2 Ablation Experiments and Analysis of Module Effectiveness

Table 1 compares five configurations under the same ExDark split and test environment. The original YOLOv8s obtains 74.6% mAP@0.5 at 45.2 FPS. Adding the original Zero-DCE raises mAP@0.5 to 81.3%, but throughput falls to 23.7 FPS. Replacing it with the redesigned enhancer increases mAP@0.5 to 84.5% while restoring end-to-end speed to 42.1 FPS, giving a better accuracy-efficiency trade-off than the original enhancement front end in this setting. The modified YOLOv8s alone reaches 78.9% mAP@0.5 and 49.1 FPS. With both branches combined and jointly optimized, the complete model achieves 87.2% mAP@0.5 at 41.3 FPS. Relative to the original YOLOv8s, this corresponds to a gain of 12.6 percentage points in mAP@0.5 with a 3.9 FPS reduction in end-to-end throughput. Because the complete model differs from the two-stage baseline in both the detector

architecture and the training strategy, the difference between them reflects the combined effect of these two factors and is not attributed to joint training alone; the independent contributions of the enhancement module and the modified detector are separately shown by the first four configurations of Table 1.

Table 1. Ablation Study Results

Method	mAP@0.5 (%)	Inference Speed (FPS)
Original YOLOv8s	74.6	45.2
YOLOv8s + Original Zero-DCE	81.3	23.7
YOLOv8s + Lightweight Zero-DCE	84.5	42.1
Improved YOLOv8s	78.9	49.1
Proposed Method	87.2	41.3

Figure 1 shows the enhancement outputs for one representative low-light test image. Because ExDark is a real low-light dataset without paired reference images, full-reference image-quality metrics such as PSNR and SSIM are not directly applicable to the enhancement output; no-reference metrics such as LOE and PIQE were not computed in the present study and are left for future evaluation.

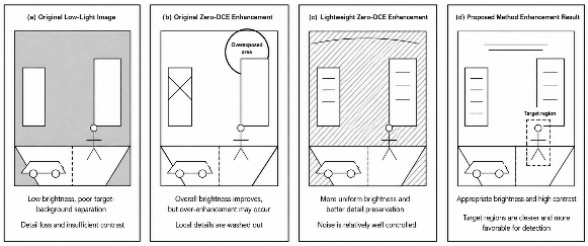


Figure 1. Schematic illustration of the enhancement results of different methods on a representative low-light test image. The redesigned enhancer yields a more even brightness distribution and clearer object boundaries than the original Zero-DCE, while the complete method preserves detection-relevant contrast.

For this example, the original Zero-DCE brightens the scene but also produces local overexposure. The redesigned enhancer gives a more even brightness distribution, and the complete method retains clearer object boundaries and stronger local contrast in the regions containing objects. No obvious increase in background noise is visible in this particular image.

3.3 Comparative Experiments and Performance Analysis

Table 2 places the proposed method alongside several representative enhancement-detection combinations: RetinexNet+YOLOv5s, KinD+YOLOv5s, LLNet+YOLOv8s, and MBLLEN+YOLOv8s [18]. Their published settings are not fully identical, so the table is used as a reference comparison rather than as a controlled ranking. The reported values are 76.8% mAP@0.5, 32.5 FPS, and 15.7M parameters for RetinexNet+YOLOv5s; 79.2%, 28.3 FPS, and 18.9M for KinD+YOLOv5s; 82.1%, 35.7 FPS, and 12.4M for LLNet+YOLOv8s; and 83.5%, 31.2 FPS, and 14.6M for MBLLEN+YOLOv8s. The proposed model reports 87.2% mAP@0.5, 41.3 FPS, and 7.0M parameters. Numerically, these values indicate a favorable accuracy-efficiency-size trade-off, but the differences should be interpreted with caution because the underlying protocols are not identical.

Table 2. Performance Comparison of Different Methods

Method	mAP@0.5 (%)	Inference Speed (FPS)	Parameters (M)
RetinexNet + YOLOv5s	76.8	32.5	15.7
KinD + YOLOv5s	79.2	28.3	18.9
LLNet + YOLOv8s	82.1	35.7	12.4
MBLLEN + YOLOv8s	83.5	31.2	14.6
Proposed Method	87.2	41.3	7.0

Note. The numerical results of the compared methods are quoted from their original publications and were obtained under their respective experimental settings. They are presented here as a reference comparison to indicate the general performance level of the field rather than as a strictly controlled evaluation, and direct numerical comparisons should be made with caution.

3.4 Visualization and Single-Image Quantitative Analysis

To provide an intuitive comparison of detection performance, test images from street and parking-lot scenes were selected for visualization. The street-scene example contains one vehicle and one distant pedestrian. RetinexNet+YOLOv5s correctly detects the vehicle but misses the pedestrian, yielding TP/FP/FN values of 1/0/1, with a single-image precision of 100% and a recall of 50%. KinD+YOLOv5s detects both ground-truth objects but falsely identifies a streetlight as a vehicle, resulting in TP/FP/FN values of 2/1/0, with a precision of 66.7% and a recall of 100%. The proposed method achieves TP/FP/FN values of 2/0/0, corresponding to both precision and recall of 100%. Therefore, this example supports the conclusion that the proposed method reduces missed detections and false positives in this specific case, rather than establishing a generalized advantage across all street scenes, as shown in Fig. 2.

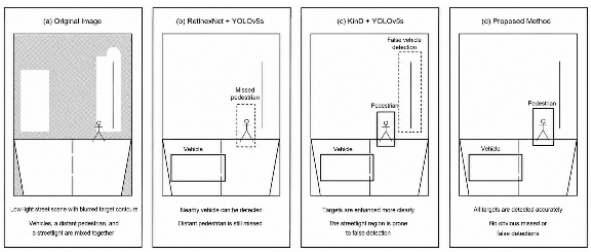


Figure 2. Schematic illustration of the detection results of different methods in a street scene. The example contains one vehicle and one distant pedestrian; the proposed method detects both objects without false

positives (TP/FP/FN = 2/0/0).

The street scene contains distant pedestrians, nearby vehicles, streetlights, and other objects. In this example, RetinexNet+YOLOv5s successfully detects the nearby vehicles but fails to detect the distant pedestrians. As shown in Fig. 3, the parking-lot scene contains five vehicles, three of which are distant small objects. LLNet+YOLOv8s detects two vehicles and misses three, yielding TP/FP/FN values of 2/0/3, an overall recall of 40%, and a distant small-object recall of 0%. MBLLEN+YOLOv8s stably detects three vehicles and misses two, with TP/FP/FN values of 3/0/2, an overall recall of 60%, and a distant small-object recall of 33.3%. In contrast, the proposed method detects all five vehicles, achieving TP/FP/FN values of 5/0/0, with both the overall recall and small-object recall reaching 100%.

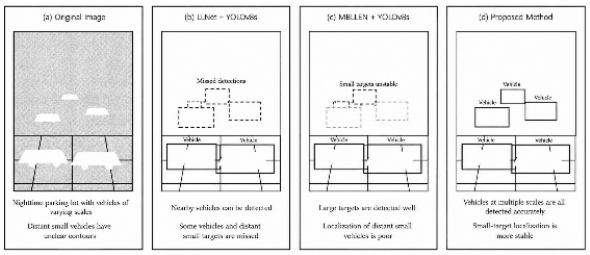


Figure 3. Schematic illustration of the detection results of different methods in a parking-lot scene. The scene contains five vehicles, including three distant small objects; the proposed method detects all five (TP/FP/FN = 5/0/0).

The parking-lot scene contains multiple vehicles with substantial scale variations. LLNet+YOLOv8s fails to detect some vehicles, while MBLLEN+YOLOv8s exhibits insufficient localization stability for small vehicles in the distance. In contrast, the proposed method maintains robust detection performance for both nearby large objects and distant small objects. This example further demonstrates the potential advantage of the proposed method for nighttime small-object detection. The examples presented above are representative cases selected to illustrate the behaviour of the compared methods; they are not a random sample, and the failure boundaries of the models, discussed qualitatively in Sections 2.3 and 4, are not fully reflected here. A systematic failure analysis is left for future work.

4. Discussion

Recent studies on low-light object detection have gradually shifted from independently improving either the enhancement module or the detector toward coordinated enhancement–detection design, such as adaptive enhancement with dynamic feature fusion [9] and lightweight detection frameworks integrating image enhancement with adaptive feature fusion [10]. Similar to these studies, this work focuses on task-level coordination between enhancement and detection, but specifically adopts a Zero-DCE-driven lightweight enhancement module together with an end-to-end joint optimization strategy. Depthwise separable convolution and attention mechanisms are both general-purpose network components; therefore, the contribution of this study lies primarily in their task-specific integration, constraint-aware design, and evaluation within the low-light detection pipeline, rather than in proposing new convolutional or attention operators. Under low-light conditions, the slowly varying illumination component and the relatively localized gradients around object boundaries provide exploitable cues for spatial weighting; however, strong light sources, reflections, and noise may also be amplified. SE can enhance channel responses with stable global statistics, but it may introduce excessive suppression under non-uniform illumination or when small objects are sparsely distributed.

The efficiency results should also be interpreted hierarchically. Although the improved enhancement module contains more parameters than the original Zero-DCE, its computationally efficient design substantially improves the practical inference efficiency of the complete enhancement–detection pipeline. Under the same experimental environment, the end-to-end inference speed of Original Zero-DCE+YOLOv8s is 23.7 FPS, whereas the combination of the improved enhancement module and YOLOv8s achieves 42.1 FPS. Therefore, the efficiency of the enhancement module discussed in this study is evaluated primarily according to its actual end-to-end inference performance rather than solely on the basis of the parameter count of the enhancement network.

Several limitations remain. Transformer-based methods have also combined low-light enhancement with object detection [22], and diffusion-based restoration approaches such as LightDiffu-DCE have been proposed for difficult, spatially varying illumination [23]. These methods were not reproduced in the present codebase, so their published results are treated only as technical context rather than as directly controlled baselines. The current experiments also do not provide class-wise AP or full-test scale-specific recall. Future evaluation should therefore reproduce stronger enhancement baselines with the same detector and hardware and report class- and scale-level results on the fixed test split. In addition, although the enhancement-module pretraining and all fine-tuning use only the training subset, the enhancement branch is nevertheless adapted to the overall luminance distribution of the ExDark dataset; evaluating the pipeline on other low-light benchmarks would provide stronger evidence of cross-dataset generalization. Finally, several ablations were not included in the current study, such as a comparison of end-to-end joint training with frozen two-stage training under exactly the same architecture, the effect of swapping the attention placements, the independent contribution of each enhancement loss component, and a per-luminance-interval performance breakdown; these are important directions for future work.

5. Conclusion

This paper develops a Zero-DCE-driven YOLO framework for nighttime low-light object detection. Although the original Zero-DCE contains only

approximately 79K parameters, using it as an independent enhancement front end still introduces an additional inference stage, and its optimization objective is not fully aligned with the object detection task. To address this issue, the enhancement module is redesigned in a task-oriented and computationally efficient manner, while the parameter count of YOLOv8s is reduced from 11.2M to 6.2M, representing a reduction of approximately 44.6%. The complete enhancement–detection model contains approximately 7.0M parameters and achieves an mAP@0.5 of 87.2% and an end-to-end inference speed of 41.3 FPS on the ExDark dataset. Compared with the original YOLOv8s, the mAP@0.5 increases from 74.6% to 87.2%, corresponding to an improvement of 12.6 percentage points. Compared with the two-stage combination of the lightweight Zero-DCE and the original YOLOv8s, the complete model further improves mAP@0.5 by 2.7 percentage points. This gain reflects the combined effects of the structural improvements to the detection network and end-to-end joint optimization.

Funding: This research received no external funding.

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

Ethics Statement and Informed Consent Statement: Not applicable. This study used only the publicly available ExDark dataset and did not involve human participants or animals.

Data Availability Statement: The ExDark dataset used in this study is publicly available at https://github.com/cs-chan/Exclusively-Dark-Dataset. Additional implementation details are available from the corresponding author upon reasonable request.

Author Contributions: Conceptualization, T.W.; 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, T.W.; Writing – review and editing, T.W. and A.B.; Visualization, T.W.; Supervision, A.B.; Project administration, T.W. All authors have read and agreed to the published version of the manuscript. (T.W. = Tianxiang Wang; A.B. = Abudhahir Buhari)

References

[1] Peng D, Ding W, Zhen T. A novel low light object detection method based on the YOLOv5 fusion feature enhancement. *Sci Rep*. 2024;14(1):4486. https://doi.org/10.1038/s41598-024-54428-8

[2] Liu Y, Li S, Zhou L, Liu H, Li Z. Dark-YOLO: a low-light object detection algorithm integrating multiple attention mechanisms. *Appl Sci*. 2025;15(9):5170. https://doi.org/10.3390/app15095170

[3] Deng K, Xing L, Wu H, Ma H, Ling Y, Gao J. Advances in object detection for autonomous driving using mmWave radar and camera: a comprehensive survey. *J King Saud Univ Comput Inf Sci*. 2025;37(10):328. https://doi.org/10.1007/s44443-025-00302-0

[4] Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: unified, real-time object detection. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas (NV): IEEE; 2016. p. 779–788. https://doi.org/10.1109/CVPR.2016.91

[5] Jiang L, Qian L. LLEYOLO: a target detection algorithm based on improved YOLOv5 for low-light environments. *J King Saud Univ Comput Inf Sci*. 2025;37(9):284. https://doi.org/10.1007/s44443-025-00252-7

[6] Lim WT, Ang K, Loh YP. Deep enhancement-object features fusion for low-light object detection. In: Proceedings of the 4th ACM International Conference on Multimedia in Asia. New York (NY): ACM; 2022. p. 1–6. https://doi.org/10.1145/3551626.3564947

[7] Guo C, Li C, Guo J, Loy CC, Hou J, Kwong S, et al. 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). Seattle (WA): IEEE; 2020. p. 1777–1786. https://doi.org/10.1109/CVPR42600.2020.00185

[8] Li C, Guo C, Loy CC. Learning to enhance low-light image via zero-reference deep curve estimation. *IEEE Trans Pattern Anal Mach Intell*. 2022;44(8):4225–4238. https://doi.org/10.1109/TPAMI.2021.3063604

[9] Fu S, Zhao Q, Liu H, Tao Q, Liu D. Low-light object detection via adaptive enhancement and dynamic feature fusion. *Alex Eng J*. 2025;126:60–69. https://doi.org/10.1016/j.aej.2025.04.047

[10] Ci WY, Hou HY, Lu S, Deng H, Ma JH. Multi-module enhanced low-light object detection: from image enhancement to adaptive feature fusion. *Adv Eng Inform*. 2026;69:104097. https://doi.org/10.1016/j.aei.2025.104097

[11] Sandler M, Howard A, Zhu M, Zhmoginov A, Chen LC. MobileNetV2: inverted residuals and linear bottlenecks. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City (UT): IEEE; 2018. p. 4510–4520. https://doi.org/10.1109/CVPR.2018.00474

[12] Wan G, Ming T. A lightweight object detection algorithm based on improved YOLOv8. *Comput Artif Intell*. 2025;2(2):37–43. https://doi.org/10.70267/cai.25v2n2.3743

[13] Vinoth K, Sasikumar P. Lightweight object detection in low light: pixel-wise depth refinement and TensorRT optimization. *Results Eng*. 2024;23:102510. https://doi.org/10.1016/j.rineng.2024.102510

[14] Jocher G, Chaurasia A, Qiu J. Ultralytics YOLOv8 [software]. Version 8.0.0. Ultralytics; 2023 [cited 2026 Nov 7]. Available from: https://github.com/ultralytics/ultralytics

[15] Zhao Y, Chen G. LPAE-YOLOv8: lightweight aerial small object detection via LSE-Head and adaptive attention. *Sci Rep*. 2025;15(1):44780. https://doi.org/10.1038/s41598-025-28741-9

[16] Ding J, Hu J, Lin J, Zhang X. Lightweight enhanced YOLOv8n underwater object detection network for low light environments. *Sci Rep*. 2024;14(1):27922. https://doi.org/10.1038/s41598-024-79211-7

[17] Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City (UT): IEEE; 2018. p. 7132–7141. https://doi.org/10.1109/CVPR.2018.00745

[18] Al Sobbahi R, Tekli J. Comparing deep learning models for low-light natural scene image enhancement and their impact on object detection and classification: overview, empirical evaluation, and challenges. *Signal Process Image Commun*. 2022;109:116848. https://doi.org/10.1016/j.image.2022.116848

[19] Qiao C, Jin J, Zhang J, Zhu Z, Cao X. Low-light object detection via joint image enhancement and feature adaptation. *J Electron Inf Technol*. 2025;47(10):3929–3940. https://doi.org/10.11999/JEIT250302. Chinese.

[20] Loh YP, Chan CS. Getting to know low-light images with the Exclusively Dark dataset. *Comput Vis Image Underst*. 2019;178:30–42. https://doi.org/10.1016/j.cviu.2018.10.010

[21] Padilla R, Passos WL, Dias TLB, Netto SL, da Silva EAB. A comparative analysis of object detection metrics with a companion open-source toolkit. *Electronics*. 2021;10(3):279. https://doi.org/10.3390/electronics10030279

[22] Kou K, Yin X, Gao X, Nie F, Liu J, Zhang G. Lightweight two-stage transformer for low-light image enhancement and object detection. *Digit Signal Process*. 2024;150:104521. https://doi.org/10.1016/j.dsp.2024.104521

[23] Yan G, Wu B, Ma L. LightDiffu-DCE: low light image enhancement based on light intensity diffusion. *Opt Precis Eng*. 2025;33(7):1114–1129. https://doi.org/10.37188/OPE.20253307.1114. Chinese.