symmetry

MDPI

Article

DCM-YOLO: Robust Electric Bicycle Detection in Confined Indoor Environments Under Occlusion and Image Degradation

Guanfang Zuo [1,2], Yuxuan Wang [3], Yanyou Sha [3], Yuchen Xia [3], Mohan Tang [2], Hengkuo Jia [2] and Ronghua Chi [2,*]

[1] Jiangsu Province Engineering Research Center of Photonic Devices and System Integration for Communication Sensing Convergence, Wuxi 214105, China; zgf@cwxu.edu.cn
[2] School of Electronic and Information Engineering, Wuxi University, Wuxi 214105, China; 23400225@stu.cwxu.edu.cn (M.T.); 23321313@stu.cwxu.edu.cn (H.J.)
[3] School of Electronics and Information Engineering, Nanjing University of Information Science and Technology, Nanjing 210044, China; 202512491104@nuist.edu.cn (Y.W.); 202312490329@nuist.edu.cn (Y.S.); 202412492557@nuist.edu.cn (Y.X.)
* Correspondence: ronghchi@cwxu.edu.cn

Abstract

To address electric bicycle detection in confined indoor environments affected by occlusion and image degradation, this study proposes DCM-YOLO, a robustness-oriented detection framework designed to improve detection accuracy under complex indoor visual conditions. First, the Dual-branch Adaptive Fusion (DAF) module combines lightweight feature generation with adaptive modulation to preserve local structures and channel diversity when target appearances are incomplete. Second, the Spatial–Channel Synergistic Attention (SCSA) mechanism sequentially refines informative regions and semantic channels, allowing the detector to suppress background interference more effectively. Third, the Multi-Scale Group-Aware Head (MSGA-Head) introduces multi-branch receptive-field modeling and grouped refinement to improve scale-sensitive classification and localization. These components form a coordinated backbone–attention–head design, reducing detection ambiguity caused by partial visibility and degraded image quality, including underexposure, overexposure, low contrast, and blur. Experimental results on a public dataset collected from representative indoor environments indicate that DCM-YOLO achieves 87.6% Precision, 83.7% Recall, 86.2% mAP50, and 65.1% mAP50-95, exceeding the baseline model by 2.5, 2.9, 2.8, and 1.7 percentage points, respectively. Additional evaluations on public benchmark datasets further verify the effectiveness and robustness of DCM-YOLO.

Keywords: occluded detection; adaptive fusion; attention mechanism; detection head; YOLO

Check for updates

Academic Editor: Zhixun Su

Received: 23 April 2026
Revised: 1 June 2026
Accepted: 11 June 2026
Published: 16 June 2026

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

1. Introduction

In recent years, fire incidents caused by electric bicycles have occurred with increasing frequency. This issue is particularly severe in high-rise residential communities, where electric bicycles are often brought into enclosed elevators, thereby posing serious safety hazards. At present, most residential communities rely on manual real-time monitoring to prevent such safety hazards. However, this approach requires substantial human resources, suffers from low efficiency, and is prone to missed detections. As a result, accurate and reliable detection of electric bicycles in confined indoor spaces has become an urgent requirement. Deep learning-based object detection methods have emerged as a promising solution.

Deep learning-based object detection algorithms can be broadly divided into two-stage and single-stage methods. Two-stage approaches, such as the Region-based Convolutional Neural Network proposed by Girshick et al. [1] and its extensions, Fast R-CNN [2] and Faster R-CNN [3], rely on region proposals followed by classification and regression. These methods achieve high detection accuracy but involve heavy computation and limited inference speed. Single-stage methods, such as the Single-Shot MultiBox Detector [4] and the You Only Look Once series [5–13], perform object localization and classification in an end-to-end manner on feature maps. By predicting bounding boxes and class probabilities in a single forward pass, YOLO-based detectors achieve favorable speed–accuracy trade-offs. They are widely adopted in real-time industrial applications.

YOLOv8 further increases detection performance with anchor-free prediction, dynamic label assignment, and efficient feature fusion strategies, achieving competitive accuracy while maintaining real-time inference capability [14]. Despite these advances, object detection in challenging indoor surveillance scenarios remains difficult in practice. Such environments are typically confined indoor spaces with fixed camera viewpoints and limited spatial margins. During normal operation, image quality may degrade abruptly due to environmental interactions such as opening and closing structures [15,16]. This results in underexposure, overexposure, low contrast, or blur within a short time, introducing biased feature activation and unstable visual responses [17].

In addition to image degradation, there are other challenges that occur frequently. Targets are often partially obstructed by pedestrians, carried objects, or surrounding structures. Under such conditions, target features become fragmented and spatially incomplete, which weakens semantic continuity and degrades detection robustness. Moreover, the confined geometry of indoor spaces causes targets to appear at different spatial positions and scales [18], further increasing the difficulty of stable feature extraction and localization.

Meanwhile, practical deployment in residential indoor environments still imposes certain constraints on model complexity. Surveillance systems are often deployed on resource-limited platforms, where excessively heavy models remain difficult to apply in practice [19,20]. However, improving robustness does not solely depend on blindly increasing network depth or introducing overly complex attention structures. In this work, we adopt a more effective design strategy that moderately increases model complexity while focusing on feature representation quality, thereby achieving improved detection performance under occlusion and image degradation [21]. At the same time, lightweight designs are also insufficient, as excessive parameter reduction may weaken feature representation when target visibility is incomplete or image quality is degraded.

Importantly, the aforementioned challenges are not isolated but closely coupled in confined indoor vision scenarios. Image degradation, occlusion, and model design constraints interact across different stages of feature processing. Optimizing only one factor often leads to unstable responses in other stages rather than genuine robustness improvement. For example, strong attention mechanisms can become sensitive to noise under abrupt image degradation, while overly simplified convolutions may weaken semantic continuity [22]. This coupling indicates that partial architectural modification is inadequate. Robust detection requires a coordinated view of feature extraction, feature selection, and prediction stability throughout the network.

The motivation of the proposed DCM-YOLO is to address feature instability in challenging indoor surveillance scenarios caused by image degradation and dense occlusion [23]. In confined indoor spaces, sudden changes in environmental conditions are more likely to disturb feature responses at early stages. At the same time, continuous occlusion weakens semantic completeness and makes stable target representation more difficult. To address these intertwined factors, DCM-YOLO is designed as a robust electric

bicycle detection framework for confined indoor environments. The framework focuses on improving feature extraction, feature selection, and prediction stability through coordinated module design. These stages are jointly optimized to improve robustness under image degradation and dense occlusion. As a result, DCM-YOLO improves detection accuracy under challenging visual conditions and provides a more reliable solution for indoor monitoring tasks:

1. We design a Dual-branch Adaptive Fusion (DAF) module to improve feature extraction in early network stages. The module combines GhostConv and Channel Shuffle to reduce channel redundancy while preserving spatial diversity. Through dual-branch interaction, DAF enhances the representation of degraded and partially occluded targets.
2. We introduce a Spatial–Channel Synergistic Attention (SCSA) mechanism to improve feature selection robustness in complex scenes. The spatial branch emphasizes informative regions affected by occlusion and image degradation, while the channel branch recalibrates semantic responses. Through coordinated spatial and channel refinement, irrelevant activations are suppressed and discriminative feature representation is improved.
3. We design a Multi-Scale Group-Aware Head (MSGA-Head) to improve prediction stability under image degradation. The head combines multi-scale perception and grouped feature refinement to enhance classification and localization reliability when targets are affected by underexposure, overexposure, low contrast, and blur.

The paper is organized as follows. Section 2 reviews related works on detecting electric bicycles in elevators, object detection under occlusion, attention mechanisms, and detection head design under image degradation. In Section 3, the DCM-YOLO framework and the design of all core modules are introduced. Section 4 shows the experimental results, ablation studies, and performance tests on public datasets. Section 5 summarizes the main content and discusses possible directions for future work.

2. Related Work

2.1. Object Detection in Confined Indoor Environments

Recent research on electric bicycle detection in confined indoor environments has mainly focused on two directions: improving detection accuracy and reducing model complexity. Elevator cabins are often used as a representative scenario in these studies because the space is limited and the visual conditions are usually challenging. Zhang et al. [24] proposed a lightweight electric bicycle detection method for confined indoor environments based on YOLOv5s. By replacing the Backbone network with MobileNetV2, a 58.4% reduction in parameters and a 50.6% reduction in computational cost were achieved. By introducing a Bidirectional Feature Pyramid Network (BiFPN) and Convolutional Block Attention (CBAM) into the neck, multi-scale feature fusion was improved and important regions were emphasized more effectively. Although this method reduced the computational burden of traditional YOLOv5 in confined indoor scenarios, optimization for severe occlusion and degraded image quality was not sufficiently considered. As a result, the detection performance remained limited when targets became partially invisible or visual cues became weak.

Wan et al. [25] developed an electric bicycle detection algorithm for confined indoor surveillance scenarios based on YOLOv8n-FS. By replacing the standard convolution modules with FasterNet blocks, the computational cost was reduced by 37%, while the inference speed changed by 7.6%. The introduction of the Squeeze-and-Excitation (SE) attention mechanism helped suppress noise interference and improved robustness under

complex lighting and small-object conditions. However, in safety-critical scenarios like elevators, improving detection accuracy is more important as long as the basic operational requirements are satisfied.

Lai et al. [26] proposed EEIE-YOLOv8 for electric bicycle detection in confined indoor environments. To reduce the accuracy drop caused by target occlusion and overlap, Efficient Multiscale Attention (EMA) and Involution blocks were introduced to strengthen edge information extraction. Experiments on a custom dataset collected from confined indoor environments showed moderate improvements in average precision and precision. However, the increase in model complexity restricted practical deployment on resource-limited devices, and the prediction stability under degraded imaging conditions was still not fully addressed.

Overall, existing methods for confined indoor detection usually emphasize either accuracy or efficiency. Accuracy-oriented models often increase structural complexity, whereas lightweight designs may lose discriminative cues when targets are partially visible or image quality is degraded. These limitations motivate a detector that balances robustness and moderate complexity in confined indoor scenes.

2.2. Occluded Object Detection

The aim of occluded object detection is to maintain reliable target perception when objects are partially or heavily occluded by other instances or scene structures. Existing approaches are generally grouped into three paradigms: occlusion-aware feature modeling, context reasoning, and relational or interaction modeling. Occlusion-aware feature modeling attempts to separate target features from occluding interference so that fragmented feature representations can be handled more effectively. He et al. [27] proposed a grouped orthogonal representation to reduce interference from occluding targets and foreground responses in traffic scenes. Better discrimination under occlusion was achieved, but the overall interaction among features became harder to model.

Localization-oriented approaches to occlusion adapt bounding box regression for partial visibility. Guo et al. [28] presented an adaptive occlusion-aware localization strategy with OL-IOU, in which bounding boxes were adjusted according to visible target regions. Localization accuracy was improved in occluded cases, but additional geometric constraints were introduced during optimization.

Occlusion robustness has also been investigated in tracking scenarios. Zheng et al. [29] indicated that motion-guided and occlusion-aware hierarchical matching can maintain stable target association during interruptions by preserving semantic consistency. This work also highlights the importance of keeping target representation stable when appearance cues are missing [30]. Additional support can be found in recent representation learning studies. Jiang et al. [31] showed that cross-modal complementary learning with memory prompting can produce stable target representations when appearance regions are absent. Zhao et al. [32] proposed a mixed local channel attention mechanism that improves feature discrimination under background interference while maintaining low computational cost.

Overall, these studies show that occlusion robustness depends strongly on the stability of early feature extraction and the effectiveness of subsequent feature selection, rather than only on explicit occlusion-aware modeling. Motivated by this observation, DCM-YOLO improves the representation of partially visible targets through DAF and further enhances discriminative responses through SCSA, so that semantic disruption caused by occlusion can be reduced more effectively.

2.3. Attention Mechanism

Attention mechanisms are used to improve feature selection by dynamically reweighting spatial or channel responses [33]. In this way, detection networks can focus more on informative target regions and reduce the influence of irrelevant background information. In object detection, attention mechanisms are widely adopted to deal with cluttered backgrounds, image degradation, and partial occlusion by reallocating representational emphasis toward more discriminative features. Recent attention designs have increasingly emphasized the balance between discriminative feature enhancement, stable information aggregation, and structural complexity. Chen and Li [34] showed that expanding the effective receptive field inside lightweight attention is beneficial for preserving detailed responses while maintaining efficient dependency modeling. Yu et al. [35] further demonstrated that multi-scale spatial modeling and long-range channel interaction can be jointly exploited to produce more informative and structured attention responses. These studies indicate that effective attention design can improve detection reliability in challenging environments.

Inspired by transformers, structured attention routing strategies have also been explored to stabilize feature responses under constrained computation. Zhu et al. [36] showed that query-adaptive grouped routing can improve attention effectiveness by filtering irrelevant regions before fine-grained interaction. This suggests that the effectiveness of attention depends not only on response intensity, but also on coordinated regulation across different feature dimensions. However, many existing attention mechanisms are still applied in a relatively independent manner. Under sudden image degradation or strong interference, unstable responses may still be introduced. Evidence from recent studies therefore supports the view that more organized spatial–channel coordination is needed to form stable and discriminative feature responses. Zhao et al. [37] showed that suppressing unreliable spectral responses enables difference-aware models to perceive targets more stably under strong interference. SiamSTU state-aware template updating was proposed to handle attention drift under appearance variation, which further highlights the importance of maintaining semantically stable responses [38].

These studies suggest that independent spatial or channel attention is often not sufficient in challenging indoor detection scenarios. More organized coordination between spatial emphasis and channel refinement is needed to form stable and discriminative responses under occlusion and image degradation. Motivated by this, DCM-YOLO introduces the SCSA mechanism to improve discriminative feature selection by jointly refining informative regions and semantic channels.

2.4. Detection Head

Recent studies on object detection under degraded visual conditions have shown that detection performance is influenced not only by feature extraction, but also by the stability of the prediction stage. In low-resolution and low-quality image detection, Li et al. [39] proposed a two-stage enhancement framework in which image-level and feature-level enhancement were jointly considered to improve detection accuracy under different degradation types. Xue et al. [40] further showed that practical architectural augmentation can effectively improve low-quality object detection, indicating that detector robustness cannot be ensured by Backbone design alone. These studies suggest that when image quality is degraded, the final prediction results become highly sensitive to the way useful features are organized and delivered to the detector output.

Recent journal studies have also indicated that stable prediction under degraded imaging conditions often requires more task-oriented feature interaction. Xiong et al. [41] improved unsupervised domain adaptive low-light object detection by fine-tuning feature interaction, which shows that low-light degradation can disturb the consistency of detector

responses if cross-level information is not properly adjusted. In adverse-weather detection, Nguyen et al. [42] proposed a model-agnostic framework that improves robustness across detectors under fog, rain, and snow conditions. Zhang et al. [43] further reduced the spurious influence of weather-specific information through feature decorrelation, showing that stable prediction depends not only on cleaner inputs, but also on more reliable use of detection features during inference.

Although these methods improve robustness under degraded conditions, most of them mainly focus on enhancement, adaptation, or feature decoupling. The prediction head itself is still usually treated as a shared output component, and explicit redesign for scale-sensitive and degradation-sensitive inference remains limited. However, under underexposure, overexposure, low contrast, and blur, the response quality of different-scale features often becomes inconsistent. If a unified prediction strategy is still adopted, classification confidence and localization accuracy can easily become unstable. Motivated by this observation, DCM-YOLO employs an MSGA-Head to improve prediction stability through multi-scale perception and grouped feature refinement, so that more reliable classification and localization can be achieved under degraded imaging conditions.

3. Proposed Method

To address electric bicycle detection in confined indoor environments under occlusion and image degradation, this study proposes a robustness-oriented detection framework termed DCM-YOLO. The overall architecture is illustrated in Figure 1. YOLOv8n is selected as the baseline because it is a representative and mature lightweight detector. It has been widely used in practical object detection projects and is friendly to embedded edge deployment. Therefore, YOLOv8n provides a suitable and reproducible baseline for evaluating the proposed method. Built upon YOLOv8n, the proposed method includes three key components, namely a DAF-enhanced Backbone, an introduced SCSA attention module, and a redesigned MSGA-Head. These components are jointly organized to improve feature extraction, discriminative feature selection, and prediction stability under complex visual conditions.

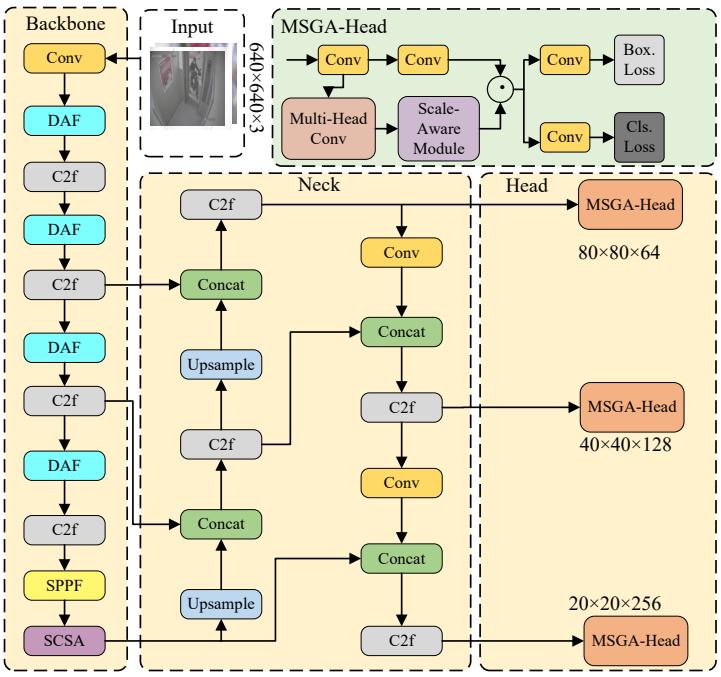


Figure 1. DCM-YOLO network structure.

Different from directly stacking existing lightweight convolutions, attention modules, or multi-scale prediction structures, the proposed framework assigns each component to a specific failure stage in confined indoor electric bicycle detection. DAF is used to alleviate early-stage feature instability, SCSA is used to improve discriminative feature selection, and MSGA-Head is used to enhance prediction stability under degraded multi-scale feature responses.

Firstly, the Backbone is enhanced with DAF as the core feature extraction unit. By combining GhostConv and Channel Shuffle, DAF reduces channel redundancy while preserving spatial diversity and local structural information. Through repeated insertion at different Backbone stages, more stable multi-scale representations can be generated for degraded and partially occluded targets. This design provides a stronger feature foundation for subsequent refinement and prediction.

Secondly, one SCSA module is introduced after the Backbone output for selective refinement before features are sent to the neck. SCSA jointly models spatial and channel dependencies, so informative target regions can be emphasized and semantically useful channel responses can be strengthened. In this way, background interference can be suppressed more effectively, and discriminative feature selection can be improved under occlusion and image degradation.

Thirdly, the prediction stage is redesigned with MSGA-Head. As shown in Figure 1, the neck preserves a multi-scale feature aggregation structure and outputs three feature maps, which are then delivered to the MSGA-Head branches at resolutions of 80×80, 40×40, and 20×20. Within MSGA-Head, multi-head convolution and scale-aware modulation are used to improve scale-sensitive prediction, so that classification and localization can remain more stable when targets are affected by underexposure, overexposure, low contrast, and blur.

Through this progressive design, DCM-YOLO links backbone representation, attention refinement, and detection-head prediction into a unified framework for robust indoor object detection.

3.1. Dual-Branch Adaptive Fusion

The motivation for designing DAF is to improve feature extraction quality at early network stages under occlusion and image degradation in challenging object detection scenarios. Conventional convolution modules often produce insufficiently discriminative feature responses when target structures are partially visible or visual quality is degraded, which limits the representation of fine-grained local details and broader semantic cues. To address this issue, a Dual-branch Adaptive Fusion module is designed as the basic feature extraction unit in the Backbone. By combining GhostConv [44] with Channel Shuffle [45] operations, DAF enhances feature diversity while preserving informative spatial structures. Through dual-branch interaction, the module strengthens the extraction of fine-grained local patterns and improves structural continuity in early-stage feature representation. The novelty of DAF does not lie in the isolated use of GhostConv or Channel Shuffle, but in the dual-branch adaptive fusion process that couples lightweight feature extraction with adaptive modulation. Through this design, local structural preservation, channel interaction, and adaptive response enhancement are jointly performed at early network stages before degraded features are propagated to deeper layers.

The overall architecture of DAF is illustrated in Figure 2.

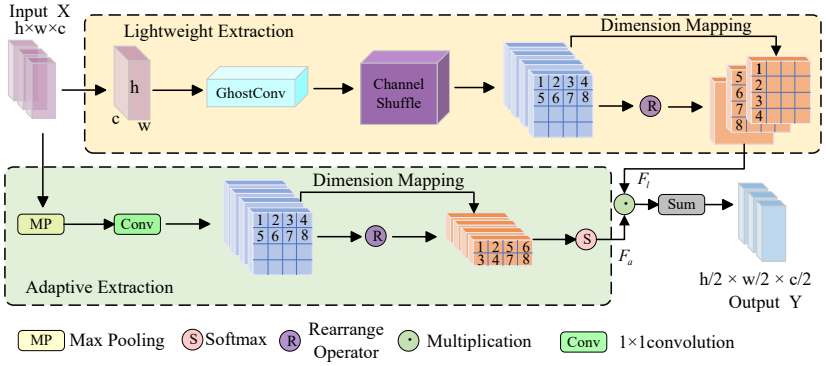


Figure 2. DAF structure diagram.

As shown in Figure 2, the input feature map $X \in \mathbb{R}^{h \times w \times c}$ is fed into two parallel branches, namely a feature extraction branch and an adaptive modulation branch. In the first branch, feature generation is performed through Ghost convolution and Channel Shuffle, so that local structural details and channel interaction can be better preserved [46]. Then, a mapping operation is applied to align the spatial resolution and channel width of the branch output with the target stage dimensions of the current Backbone stage, and the extracted feature F_l is generated [47].

In parallel, the adaptive modulation branch is designed to generate adaptive weighting information. The input feature is first processed by Max Pooling to aggregate contextual information, followed by a 1×1 convolution to reduce channel dimensions and reorganize feature responses. The compressed feature is denoted as X_c. It is then rearranged by the operator $R(\cdot)$ and normalized by a Softmax function to produce an adaptive weight map W. In this way, more informative spatial–channel locations can be emphasized, while unreliable or interference-sensitive responses can be suppressed [48,49]. The adaptive weight generation process is defined in Equation (1).

$$W = \mathrm{softmax}(R(X_c)) \tag{1}$$

where X_c denotes the compressed feature after channel reduction, $R(\cdot)$ denotes the rearrangement operation, and W denotes the adaptive weight map. The adaptive feature F_a is obtained through element-wise modulation between the extracted feature F_l and the adaptive weight map W. This operation strengthens reliable feature responses while preserving the structural information generated by the extraction branch. Then, the output feature Y is obtained through the residual summation of F_l and F_a. The fusion process is defined in Equation (2).

$$Y = F_l + F_a \tag{2}$$

where F_l denotes the extracted feature from the feature extraction branch, F_a denotes the adaptively modulated feature, and Y denotes the final output feature of DAF. Through this dual-branch adaptive fusion mechanism, DAF improves early-stage feature representation for degraded and partially occluded targets. More stable local structures and more informative spatial details can therefore be preserved before subsequent feature refinement and prediction. As a result, DAF provides a stronger feature foundation for robust object detection under occlusion and image degradation.

3.2. Spatial–Channel Synergistic Attention

The motivation for introducing the SCSA mechanism is that occlusion and image degradation in challenging object detection scenarios destabilize attention responses and reduce feature reliability. Under underexposure, overexposure, low contrast, and blur,

spatial localization becomes less stable, and channel-wise semantic consistency is more easily disturbed. Existing attention mechanisms, such as SE and CBAM model spatial or channel dimensions independently, and they are often insufficient for complex scenes where target visibility is incomplete and visual quality is degraded [50]. To address this limitation, SCSA is introduced after the Backbone to jointly regulate spatial emphasis and channel reliability. By coordinating spatial and channel attention, SCSA suppresses background interference and strengthens informative target responses [51]. Discriminative features are therefore refined more effectively under occlusion and image degradation, as illustrated in Figure 3.

Unlike attention modules that enhance spatial or channel responses in an independent or weakly coupled manner, SCSA adopts a sequential spatial–channel refinement strategy. SMSA first stabilizes spatial emphasis through multi-receptive-field context, and PCSA then recalibrates channel semantics, so the module is designed for discriminative feature selection rather than general attention weighting.

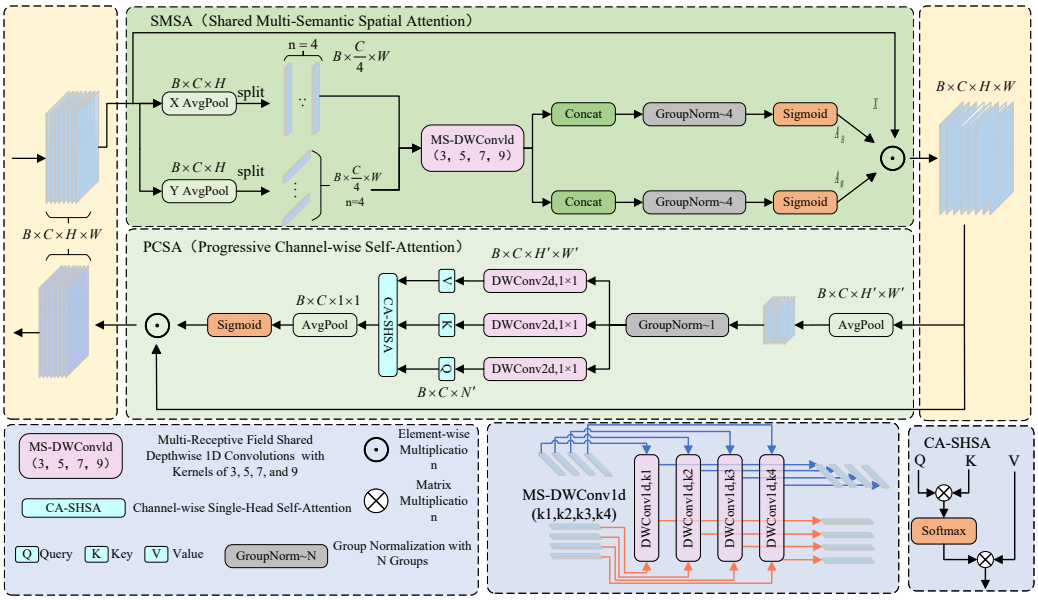


Figure 3. SCSA structure diagram.

This module is composed of two complementary branches:

(1) Multi-Semantic Spatial Attention (SMSA) improves spatial feature selection by aggregating multi-receptive-field context. This design enables the model to maintain more reliable spatial emphasis when target boundaries and local details are weakened by occlusion or image degradation. Concretely, the input feature is split into four low-dimensional branches. Each branch is processed by shared depthwise one-dimensional convolutions with different kernel sizes. This operation captures both local edge information and broader semantic context. The resulting features are then normalized and gated to generate spatial attention maps.

(2) Progressive Channel Self-Attention (PCSA) improves channel discrimination through progressive refinement of channel dependencies. This process is driven by globally pooled channel descriptors. Irrelevant channels are suppressed during refinement, while semantically useful responses are selectively emphasized. In this way, more discriminative channel representations can be obtained for degraded and partially occluded targets.

As illustrated in Figure 3, SCSA refines features through a sequential spatial–channel attention pipeline. The SMSA branch is applied first, followed by the PCSA branch.

Given an input feature map X, SMSA is used to refine spatial responses before channel-wise recalibration.

In the SMSA branch, directional average pooling is performed along the height and width dimensions to extract complementary spatial context cues. The pooled features are split into $n = 4$ low-dimensional channel groups to improve spatial sensitivity. Each group is processed by shared multi-scale depthwise one-dimensional convolutions with kernel sizes $\{3, 5, 7, 9\}$, enabling multi-receptive-field perception under occlusion and image degradation. The outputs from different scales are concatenated and normalized using Group Normalization. Sigmoid activation is then applied to generate spatial attention maps along the height and width directions, denoted as A_H and A_W, respectively. The spatially refined feature X_s is obtained by element-wise modulation of the input feature, as defined in Equation (3), where $\odot$ denotes element-wise multiplication.

$$X_s = X \odot A_H \odot A_W \tag{3}$$

Based on the spatially refined feature X_s, the PCSA branch performs progressive channel-wise refinement to enhance semantic discrimination. Global average pooling is first applied to X_s to compress spatial information and generate channel descriptors. These descriptors are then projected into query, key, and value embeddings, denoted as Q, K, and V, respectively, using depthwise 1×1 convolutions. A channel-wise self-attention mechanism is subsequently employed to capture inter-channel dependencies. Specifically, the channel attention response A_C is computed as shown in Equation (4), where $\sigma(\cdot)$ denotes the Sigmoid activation function.

$$A_C = \sigma\left(\text{Softmax}\left(QK^T\right)V\right) \tag{4}$$

The final output of the SCSA module is obtained by applying channel-wise modulation to the spatially refined feature, as defined in Equation (5).

$$Y = X_s \odot A_C \tag{5}$$

Through the sequential integration of SMSA and PCSA, SCSA first improves spatial emphasis under target degradation and partial occlusion, followed by selective recalibration of channel-wise semantics. This coordinated refinement enhances discriminative feature selection and suppresses background interference, thereby providing more reliable feature representations for subsequent prediction.

3.3. Multi-Scale Group-Aware Head

The motivation for designing the MSGA-Head is that, under image degradation, reliable detection depends not only on whether discriminative features are preserved, but also on whether prediction responses remain stable across different scales. Under underexposure, overexposure, low contrast, and blur conditions, target boundaries become ambiguous, local textures are weakened, and semantic responses tend to be scattered [52]. Even when the preceding feature extraction and feature selection stages have retained useful information, the detection head still has to perform stable classification and localization under multi-scale feature distributions that are easily disturbed by image degradation. If features at different scales are still processed with an overly uniform prediction strategy, the detector may suffer from confidence fluctuation, unstable classification, and localization deviation [53]. To address this issue, an MSGA-Head is designed at the prediction stage to improve prediction stability for degraded and partially visible targets through scale-sensitive multi-branch perception and grouped subspace refinement. The novelty of MSGA-

Head lies in moving degradation-aware refinement to the prediction stage. Instead of using a uniform head for all feature scales, MSGA-Head combines scale-sensitive receptive-field perception with grouped subspace refinement, allowing classification and localization to remain more stable when feature quality differs across scales.

The overall structure of the MSGA-Head is illustrated in Figure 4.

Conventional YOLO detection heads usually employ a shared prediction structure for features of different scales. Although such a design is effective in common scenarios, it is less suitable for degraded imaging conditions, where the representation quality of features at different scales often differs significantly [54]. Small-scale features rely more on local edge and detailed information, whereas large-scale features are more responsible for global semantic representation. A unified prediction strategy cannot sufficiently accommodate these differences, which reduces prediction stability in degraded scenes. Therefore, the proposed MSGA-Head is designed to incorporate multi-scale convolutional perception and group-wise feature refinement to improve classification and localization reliability under image degradation.

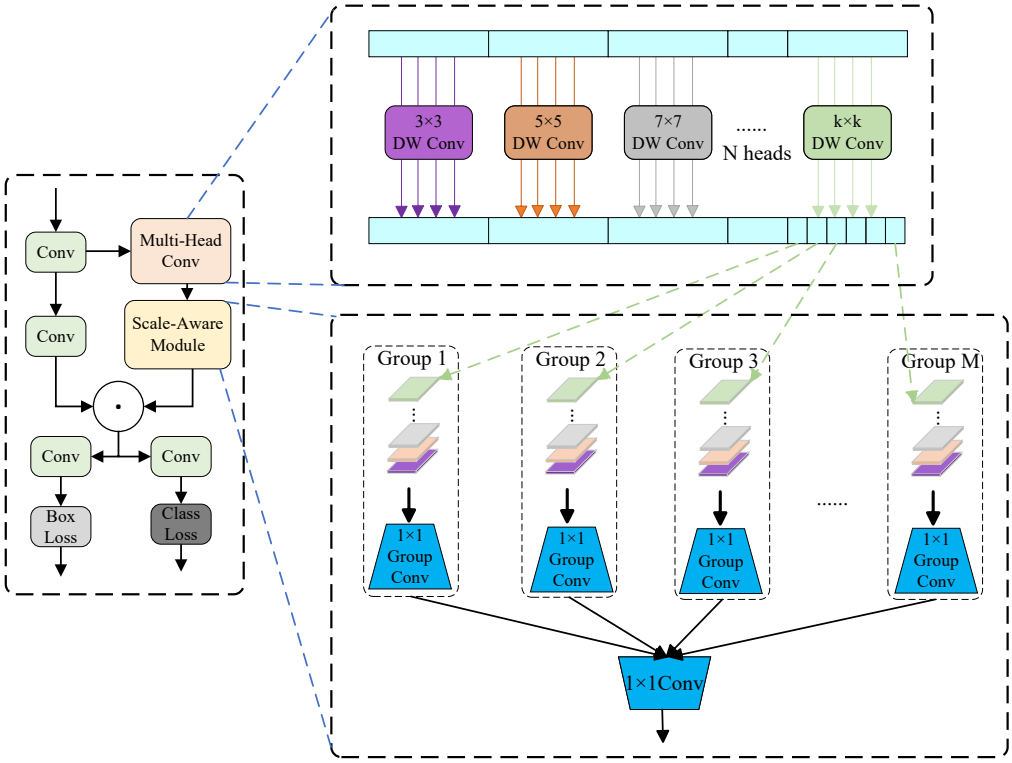


Figure 4. Architecture of the proposed Multi-Scale Group-Aware Head.

As shown in Figure 4, the MSGA-Head first applies multi-head depthwise separable convolution to the input feature map. Different branches adopt convolution kernels of different sizes, so that feature responses with diverse receptive fields can be generated in parallel. In this way, contextual information at multiple spatial scales can be modeled simultaneously, which improves the perception of target contours, edges, and local structures under degraded imaging conditions. This design is particularly beneficial when visual cues become weak or ambiguous due to adverse illumination and blur effects.

Let the input feature from the neck be denoted as $F \in \mathbb{R}^{H \times W \times C}$, where H and W are the spatial dimensions of the feature map, and C is the number of channels. After multi-head depthwise separable convolution, the output feature of the h-th branch is defined as shown in Equation (6):

$$F^{(h)} = D_{k_h}(F), \quad h = 1, 2, \ldots, N \tag{6}$$

where $D_{k_h}(\cdot)$ denotes the depthwise separable convolution operation with kernel size $k_h \times k_h$, and N is the number of parallel branches. Based on the multi-scale responses produced by different branches, the feature channels are further partitioned into M groups along the channel dimension. Each group is independently transformed by a 1×1 convolution to refine semantic distributions within each feature subspace. Unlike a conventional full-channel 1×1 convolution, this grouped transformation focuses on local semantic adjustment within each group, which helps reduce interference among heterogeneous feature channels. Such a design enables different subspaces to concentrate on target representations at different scales or semantic levels. The group-wise mapping process is expressed as shown in Equation (7):

$$G_m = C_{1\times1}\left(F_m^{(h)}\right), \quad m = 1, 2, \ldots, M \tag{7}$$

where $F_m^{(h)}$ denotes the m-th grouped feature, and $C_{1\times1}(\cdot)$ represents the 1×1 convolution operation. Finally, all grouped features are concatenated along the channel dimension and fused through a final 1×1 convolution to generate the output feature for target prediction, as shown in Equation (8):

$$F_{out} = C_{1\times1}(\text{Concat}(G_1, G_2, \ldots, G_M)) \tag{8}$$

Through this design, the MSGA-Head explicitly models both scale-level variation and subspace-level semantic differences during prediction. The multi-head depthwise separable convolution branches provide diverse receptive-field perception for degraded targets, while the grouped channel transformation refines feature responses in a more discriminative manner. Their combination improves the stability of classification and localization under underexposure, overexposure, low contrast, and blur conditions. As a result, the proposed MSGA-Head enhances prediction robustness in challenging monitoring environments and provides more reliable support for degraded-target detection.

3.4. Optimization-Related Discussion

Although the proposed modules are mainly designed from an architectural perspective, their effectiveness can also be interpreted from the viewpoints of feature modulation, gradient propagation, and response regularization. Let X denote the input feature and $\mathcal{F}(\cdot)$ denote a general nonlinear transformation. In degraded or occluded scenes, unreliable responses may be introduced by weak illumination, blur, or incomplete target appearance. The proposed modules aim to reduce such unreliable responses by adaptively modulating feature activations and preserving informative target-related cues.

For DAF, the residual adaptive fusion process defined in Equation (2) provides an identity-preserving transformation. The residual path allows the extracted feature F_l to be directly propagated to the output, which helps maintain stable feature transmission and alleviates optimization difficulty. Meanwhile, the adaptive branch generates a weight map to emphasize reliable spatial–channel responses and suppress interference-prone activations. Therefore, DAF not only enriches feature representation but also provides a more stable optimization path for early backbone learning. The residual path allows the extracted feature F_l to be directly propagated to the output, which helps maintain stable feature transmission and alleviates optimization difficulty. Meanwhile, the adaptive branch generates a weight map to emphasize reliable spatial–channel responses and suppress interference-prone activations. Therefore, DAF not only enriches feature representation but also provides a more stable optimization path for early backbone learning.

For SCSA, by combining Equations (3) and (5), the sequential spatial–channel refinement can be interpreted as a two-stage gating process:

$$Y = X \odot A_H \odot A_W \odot A_C.$$

The spatial gates A_H and A_W first regulate the response distribution over spatial locations, while the channel gate A_C further recalibrates semantic responses. This sequential modulation reduces the influence of irrelevant background regions and unreliable channels. From an optimization perspective, such gated refinement constrains the feature response space and encourages the network to focus on target-related activations, which is beneficial for learning more discriminative representations under occlusion and image degradation.

For MSGA-Head, the multi-branch and grouped prediction design reduces the interference caused by heterogeneous feature responses across different scales and channel subspaces. Different receptive-field branches capture complementary local and contextual information, while grouped refinement allows each channel subspace to be adjusted more independently. Compared with a uniform prediction head, this design provides a more flexible prediction function for scale-dependent feature distributions. As a result, the classification and localization branches can obtain more stable inputs, improving prediction reliability under underexposure, overexposure, low contrast, and blur.

Overall, DAF improves the stability of early feature transmission, SCSA regularizes spatial–channel responses through sequential gating, and MSGA-Head reduces cross-scale and cross-channel interference during prediction. These optimization-related properties provide a theoretical explanation for the improved robustness of DCM-YOLO beyond the computational flow of the proposed modules.

4. Experiments and Results

4.1. Experimental Setup

4.1.1. Experimental Environment and Parameters

All experiments were performed on an Ubuntu 20.04 system with Python 3.8, PyTorch 2.0.0, and CUDA 11.8. The hardware configuration consisted of an NVIDIA RTX 3090 GPU (24GB VRAM) and a 14-core Intel Xeon Platinum 8362 CPU @ 2.80 GHz. Training hyperparameters included a batch size of 32, momentum of 0.937, weight decay of 0.0005, initial learning rate of 0.01, and the SGD optimizer. To ensure fair comparison, all compared models were trained and evaluated under the same experimental protocol. Specifically, the same dataset split, input resolution of 640×640, data augmentation strategy, batch size of 32, SGD optimizer, initial learning rate of 0.01, momentum of 0.937, weight decay of 0.0005, and default YOLOv8 training configuration were used. The model underwent training for 200 epochs to ensure convergence. To evaluate the stability of the experimental results, the main models were trained and evaluated three times using different random seeds, and the results are reported as mean $\pm$ standard deviation where applicable.

4.1.2. Dataset

We selected a publicly available dataset collected from representative confined indoor environments [55], which consists of 5461 high-resolution images captured under real-world indoor surveillance conditions. This dataset is designed to reflect the challenges commonly encountered in confined indoor vision scenarios. These challenges include frequent occlusions caused by human activity, tight spatial constraints, and pronounced illumination variation within limited indoor spaces. Such factors pose significant difficulties for conventional object detection models. Each image is annotated with bounding boxes for three categories: electric bicycle, bicycle, and person. This annotation strategy

helps mitigate common misclassification issues, particularly between bicycles and electric bicycles. The dataset is systematically divided into training, validation, and test sets with proportions of 70%, 20%, and 10%, respectively. The training subset contains 4780 person instances, 4302 electric bicycle instances, and 3823 bicycle instances. The validation subset contains 1365 person instances, 1229 electric bicycle instances, and 1092 bicycle instances. The test subset contains 683 person instances, 614 electric bicycle instances, and 547 bicycle instances.

4.1.3. Evaluation Metrics

Model performance was assessed using standard detection metrics, including Precision (P), Recall (R), Average Precision (AP), and Mean Average Precision (mAP). Computational efficiency was measured via FLOPs and model size (Params). The mathematical formulations for these metrics are provided in Equations (9)–(12).

$$P = \frac{TP}{TP + FP} \times 100\% \tag{9}$$

$$R = \frac{TP}{TP + FN} \times 100\% \tag{10}$$

$$AP = \int_0^1 P(R)\, dR \times 100\% \tag{11}$$

$$mAP = \frac{\sum_{i=1}^{k} AP_i}{k} \tag{12}$$

Definitions of Evaluation Metrics: True Positive (TP) denotes the number of samples correctly predicted as positive instances; False Positive (FP) represents the number of samples incorrectly predicted as positive instances; False Negative (FN) indicates the number of samples incorrectly predicted as negative instances; and True Negative (TN) corresponds to the number of samples correctly predicted as negative instances. Average Precision (AP) is used to evaluate the detection performance for a single object class by summarizing the shape of the precision–recall (P-R) curve. It is computed by integrating the precision values over all recall levels from 0 to 1, where precision is treated as a function of recall. In this context, precision reflects the proportion of correctly predicted positive samples among all predicted positives at each recall level. The resulting AP value, typically expressed as a percentage, captures the ability of the model to maintain high precision across varying degrees of recall. mAP extends this concept to multi-class object detection by averaging the AP values across all object classes. Specifically, mAP is obtained by taking the mean of the AP scores for each individual class. Here, k denotes the total number of classes, and AP_i represents the Average Precision for the i-th class. As such, mAP provides a single, comprehensive metric that reflects the overall model detection accuracy across all categories.

4.2. Ablation Study

4.2.1. Component-Wise Ablation Results

To evaluate the contribution of each component, ablation experiments were conducted under identical settings, and the results are summarized in Table 1. The baseline YOLOv8n achieves 85.1% Precision, 80.8% Recall, 83.4% mAP50, and 63.4% mAP50-95 with 3.0 M parameters and 8.1 GFLOPs.

Adding DAF improves Precision, Recall, mAP50, and mAP50-95 from 85.1%, 80.8%, 83.4%, and 63.4% to 85.5%, 81.1%, 83.8%, and 63.7%, respectively, with only a small increase in parameters and FLOPs. This improvement suggests that the DAF-enhanced backbone

retains more useful target-related features when local structures are weakened by occlusion or image degradation.

The single-module results also show that SCSA and MSGA-Head contribute from different stages of the network. With only SCSA, the model achieves 85.7% Precision, 81.4% Recall, 84.2% mAP50, and 63.9% mAP50-95, showing the benefit of spatial–channel refinement for feature selection. With only MSGA-Head, the model obtains 85.9% Precision, 81.6% Recall, 84.2% mAP50, and 64.0% mAP50-95, indicating that prediction-stage refinement improves classification and localization stability.

Table 1. Results of the ablation experiment.

DAF	SCSA	MSGA-Head	P%	R%	mAP50/%	mAP50-95/%	Params/10^6	FLOPs/G
–	–	–	85.1	80.8	83.4	63.4	3.0	8.1
√	–	–	85.5	81.1	83.8	63.7	3.2	8.2
–	√	–	85.7	81.4	84.2	63.9	3.1	8.3
–	–	√	85.9	81.6	84.2	64.0	3.2	8.4
√	√	–	86.3	82.0	84.5	64.1	3.3	8.2
√	–	√	86.8	82.6	85.1	64.5	3.4	8.5
–	√	√	87.1	83.0	85.7	64.8	3.5	8.6
√	√	√	87.6	83.7	86.2	65.1	3.7	8.8

The two-module combinations further support the complementarity of the proposed components. When DAF and SCSA are used together, Precision and Recall increase to 86.3% and 82.0%, while mAP50 and mAP50-95 reach 84.5% and 64.1%. This result suggests that SCSA can further refine the spatial and channel responses after the backbone representation is strengthened by DAF. When MSGA-Head is introduced on top of DAF, the model achieves 86.8% Precision, 82.6% Recall, 85.1% mAP50, and 64.5% mAP50-95, with 3.4 M parameters and 8.5 GFLOPs. This indicates that multi-scale perception and grouped feature refinement help stabilize prediction for degraded and partially visible targets. When SCSA and MSGA-Head are jointly adopted, the model reaches 87.1% Precision, 83.0% Recall, 85.7% mAP50, and 64.8% mAP50-95, confirming that feature selection and prediction refinement work better when used together.

The complete model achieves the best overall performance, with 87.6% Precision, 83.7% Recall, 86.2% mAP50, and 65.1% mAP50-95. Compared with YOLOv8n, DCM-YOLO improves Precision, Recall, mAP50, and mAP50-95 by 2.5, 2.9, 2.8, and 1.7 percentage points, respectively. This improvement is accompanied by a moderate increase in parameters from 3.0 M to 3.7 M and FLOPs from 8.1 G to 8.8 G. Therefore, the proposed model does not simply aim to minimize computational cost but seeks a better trade-off between model complexity and detection reliability. For the considered safety-monitoring task, the improvement in Recall is particularly important because missed detections may lead to a higher practical risk.

In the repeated runs, the standard deviations of the main metrics are small. For YOLOv8n, the Precision, Recall, mAP50, and mAP50-95 values are 85.1 $\pm$ 0.2%, 80.8 $\pm$ 0.2%, 83.4 $\pm$ 0.2%, and 63.4 $\pm$ 0.1%, respectively. For DCM-YOLO, the corresponding values are 87.6 $\pm$ 0.2%, 83.7 $\pm$ 0.2%, 86.2 $\pm$ 0.2%, and 65.1 $\pm$ 0.1%, respectively. These results show that the performance gain remains consistent under different random seeds.

4.2.2. Branch-Level Ablation of SCSA

To analyze the structural necessity of SCSA and clarify the contribution of each attention branch, a branch-level ablation study is conducted by separately evaluating SMSA and PCSA. All experiments are performed under identical training settings with the same Backbone, neck, and detection head. The results are summarized in Table 2.

Table 2. Ablation Study of SCSA.

Model	P%	R%	mAP50/%	mAP50-95/%	Params/10^6	FLOPs/G
YOLOv8n	85.1	80.8	83.4	63.4	3.0	8.1
Only SMSA	85.3	81.0	83.6	63.5	3.0	8.2
Only PCSA	85.4	81.1	83.7	63.6	3.0	8.2
YOLOv8n+SCSA	85.7	81.4	84.2	63.9	3.1	8.2

When only the spatial attention branch is retained, denoted as Only-SMSA, the model achieves 85.3% Precision, 81.0% Recall, 83.6% mAP50, and 63.5% mAP50-95. The number of parameters remains unchanged at 3.0 M, while the computational cost slightly increases to 8.2 GFLOPs. This result indicates that SMSA can enhance spatial response stability and improve localization reliability under occlusion and image degradation. The increases in Precision and Recall also suggest that spatial refinement contributes to reducing false detections and missed detections to some extent. However, since it does not explicitly model channel-wise semantic dependency, its ability to suppress semantically irrelevant responses remains limited. When only the channel attention branch is adopted, denoted as Only-PCSA, the model reaches 85.4% Precision, 81.1% Recall, 83.7% mAP50, and 63.6% mAP50-95, with 3.0 M parameters and 8.2 GFLOPs. Compared with Only-SMSA, Only-PCSA achieves slightly better performance in all four accuracy-related metrics, suggesting that channel-wise feature reweighting is effective in highlighting informative semantic responses and reducing background interference. The slightly higher Precision and Recall further indicate that PCSA improves both false-detection suppression and missed-detection reduction. Nevertheless, without explicit spatial refinement, its ability to stabilize target localization is still constrained.

A further improvement is achieved when both branches are jointly integrated in the full SCSA module. In this case, the model attains 85.7% Precision, 81.4% Recall, 84.2% mAP50, and 63.9% mAP50-95, with 3.1 M parameters and 8.2 GFLOPs. Compared with the baseline YOLOv8n, the complete SCSA module improves Precision and Recall by 0.6 percentage points, while mAP50 and mAP50-95 increase by 0.8 and 0.5 percentage points, respectively. Compared with Only-SMSA and Only-PCSA, the complete SCSA module also achieves higher Precision, Recall, mAP50, and mAP50-95. These results demonstrate that spatial refinement and channel reweighting play complementary roles in discriminative feature selection. Their coordinated integration enables more effective feature refinement and more reliable target detection under occlusion and image degradation.

4.3. Comparative Experiments

4.3.1. Comparison with Different Attention Mechanisms

To evaluate the effectiveness of the proposed attention mechanism, comparative experiments were conducted with six representative attention modules, including Squeeze-and-Excitation (SE) [56], Global Group Attention (GGA) [57], Large Separable Kernel Attention (LSKA) [58], Multi-Scale Dilated Attention (MSDA) [59], Deformable Attention Transformer (DAT) [60], and Mixed Local Channel Attention (MLCA) [61,62]. All methods were evaluated under identical settings, and the results are summarized in Table 3.

As shown in Table 3, the proposed SCSA module obtains the best overall results among the compared attention mechanisms, with 85.7% Precision, 81.4% Recall, 84.2% mAP50, and 63.9% mAP50-95. Compared with YOLOv8n, the gains are 0.6, 0.6, 0.8, and 0.5 percentage points in Precision, Recall, mAP50, and mAP50-95, respectively. Although the numerical improvement is moderate, the consistent increase across all four metrics shows

that spatial–channel coordination is beneficial for feature refinement under occlusion and image degradation.

Table 3. Comparison of different attention mechanisms.

Model	P%	R%	mAP50/%	mAP50-95/%	Params/10^6	FLOPs/G
YOLOv8n	85.1	80.8	83.4	63.4	3.0	8.1
+SE	84.9	80.5	83.2	63.1	3.1	8.2
+GGA	85.2	80.9	83.5	63.4	3.1	8.2
+LSKA	85.3	81.0	83.6	63.5	3.2	8.3
+MSDA	85.6	81.3	84.1	63.7	3.6	8.9
+DAT	85.4	81.1	83.8	63.5	3.3	8.4
+MLCA	85.5	81.2	83.9	63.6	3.2	8.5
+Ours	85.7	81.4	84.2	63.9	3.1	8.2

SE performs slightly worse than the baseline, with 84.9% Precision and 80.5% Recall. This result suggests that channel recalibration alone may be insufficient when target regions are partially occluded or degraded by poor illumination. In such cases, spatially informative regions need to be localized before channel-wise semantic responses are recalibrated. Compared with GGA, LSKA, DAT, and MLCA, SCSA also achieves higher Precision and Recall with a relatively small model size. This indicates that isolated response enhancement or increased attention complexity does not necessarily lead to more stable detection in confined indoor scenes.

MSDA provides the closest performance to SCSA, reaching 85.6% Precision, 81.3% Recall, 84.1% mAP50, and 63.7% mAP50-95. However, MSDA requires 3.6 M parameters and 8.9 GFLOPs, whereas SCSA reaches slightly higher accuracy with 3.1 M parameters and 8.2 GFLOPs. This comparison shows that the advantage of SCSA comes from its sequential spatial–channel refinement rather than from heavier attention computation. By first refining spatial emphasis and then recalibrating channel responses, SCSA provides a more suitable attention design for degraded and partially occluded indoor detection.

4.3.2. Comparative Experiment on Backbone Structure

To evaluate the effectiveness of the designed Backbone structure, the proposed Backbone was compared with several representative Backbone networks, including MobileNetV3 [63], SENetV2 [64], ShuffleNetV2 [45], PP-HGNetV2 [65], and GhostNetV2 [66]. All Backbones were integrated into the same YOLOv8 detector framework to ensure a fair comparison [67]. The quantitative results are summarized in Table 4.

Table 4. Comparison results of different backbone networks.

Model	P%	R%	mAP50/%	mAP50-95/%	Params/10^6	FLOPs/G
YOLOv8n	85.1	80.8	83.4	63.4	3.0	8.1
MobileNetV3	84.4	79.8	82.6	62.8	5.6	9.5
SENetV2	84.8	80.4	83.1	63.2	3.0	8.2
ShuffleNetV2	85.4	81.2	83.7	63.6	3.2	8.3
PP-HGNetV2	85.8	81.6	84.1	63.8	4.5	8.7
GhostNetV2	85.0	80.6	83.3	63.2	6.3	8.7
Ours	86.3	82.0	84.5	64.1	3.3	8.2

As shown in Table 4, the proposed Backbone obtains the best performance among the compared backbone structures, reaching 86.3% Precision, 82.0% Recall, 84.5% mAP50,

and 64.1% mAP50-95. Compared with YOLOv8n, it improves Precision, Recall, mAP50, and mAP50-95 by 1.2, 1.2, 1.1, and 0.7 percentage points, respectively, while increasing the parameters and FLOPs only slightly. This result indicates that the DAF-based backbone improves feature representation without introducing excessive computational overhead.

MobileNetV3 and GhostNetV2 show lower accuracy than the proposed Backbone. MobileNetV3 obtains 84.4% Precision and 79.8% Recall, while GhostNetV2 obtains 85.0% Precision and 80.6% Recall. Although these lightweight structures reduce redundancy, their compressed representations may weaken fine-grained local structures when targets are partially visible or image quality is degraded. By contrast, DAF preserves spatial diversity and local structural cues in the early backbone stages, which provides a more stable feature basis for subsequent refinement.

SENetV2 and ShuffleNetV2 also improve efficiency through channel recalibration or channel shuffling, but their performance remains lower than that of the proposed Backbone. This suggests that channel-level adjustment alone is insufficient when spatial ambiguity and semantic inconsistency occur simultaneously. PP-HGNetV2 achieves relatively strong results, with 85.8% Precision, 81.6% Recall, 84.1% mAP50, and 63.8% mAP50-95. However, it uses 4.5 M parameters and 8.7 GFLOPs, whereas the proposed Backbone achieves higher accuracy with 3.3 M parameters and 8.2 GFLOPs. These comparisons show that the DAF-based design provides a better accuracy–complexity trade-off for degraded and partially occluded indoor scenes.

4.3.3. Comparison with Different Detection Algorithms

As shown in Table 5, DCM-YOLO achieves the best accuracy among all compared detectors, with 87.6% Precision, 83.7% Recall, 86.2% mAP50, and 65.1% mAP50-95. The FPS values in Table 5 were measured on the RK3588 embedded development board under the same input resolution. Compared with YOLOv8n, DCM-YOLO improves Precision, Recall, mAP50, and mAP50-95 by 2.5, 2.9, 2.8, and 1.7 percentage points, respectively. This result shows that the coordinated use of DAF, SCSA, and MSGA-Head improves feature extraction, feature selection, and prediction stability beyond the original baseline.

Table 5. Comparison results of different detection algorithms.

Model	P%	R%	mAP50/%	mAP50-95/%	Params/10^6	FLOPs/G	FPS
Faster R-CNN	67.4	58.9	61.2	42.7	41.37	158	22.8
YOLOv5n	79.8	74.6	78.7	59.8	2.5	7.2	52.4
YOLOv6	80.2	75.3	79.2	60.9	4.2	11.9	39.7
YOLOv7-tiny	80.7	75.6	79.4	61.4	6.3	13.2	36.8
YOLOv8n	85.1	80.8	83.4	63.4	3.0	8.1	47.5
YOLOv8-FS [25]	85.6	81.6	84.3	63.9	2.8	7.6	50.2
EEIE-YOLOv8 [26]	86.2	82.4	85.0	64.5	3.4	9.2	44.9
YOLOv8s	85.0	80.5	83.5	63.1	11.38	28.8	24.6
YOLOv9s	85.8	81.5	84.0	63.8	7.2	12.3	38.9
YOLOv10n	85.5	81.0	83.8	63.3	2.69	8.2	48.3
YOLOv11n	86.0	82.2	84.5	64.1	2.58	6.3	57.6
YOLOv12n	86.1	82.3	84.8	64.2	2.57	6.5	53.8
YOLOv13n	86.4	82.8	85.3	64.7	2.4	6.3	55.1
RT-DETR	76.3	70.1	72.6	56.3	42.79	134.5	28.3
Ours	87.6	83.7	86.2	65.1	3.7	8.8	46.2

Among recent YOLO-series detectors, YOLOv11n, YOLOv12n, and YOLOv13n obtain higher FPS or lower computational cost than DCM-YOLO. However, their Precision, Recall, mAP50, and mAP50-95 are still lower than those of DCM-YOLO. For example, YOLOv13n achieves 86.4% Precision, 82.8% Recall, 85.3% mAP50, and 64.7% mAP50-95, while DCM-YOLO improves these values to 87.6%, 83.7%, 86.2%, and 65.1%, respectively. This comparison suggests that, for confined indoor electric bicycle detection, reducing

model size alone does not fully address the accuracy loss caused by occlusion and image degradation.

Compared with heavier generic detectors, DCM-YOLO also shows a more practical complexity-performance relationship. Faster R-CNN and RT-DETR require much larger computational costs, with 158 G and 134.5 G FLOPs, respectively, but their detection accuracy is clearly lower than that of DCM-YOLO. YOLOv8s also has a larger parameter count and higher FLOPs than YOLOv8n, but its accuracy gain is limited in this task. These results indicate that directly increasing model scale or using a more complex generic detector is not an efficient solution for the target scenario.

Compared with YOLOv8-FS and EEIE-YOLOv8, which are both designed for confined indoor electric bicycle detection, DCM-YOLO achieves higher Precision, Recall, mAP50, and mAP50-95 while maintaining a moderate computational cost. Specifically, compared with YOLOv8-FS, DCM-YOLO improves Precision, Recall, mAP50, and mAP50-95 by 2.0, 2.1, 1.9, and 1.2 percentage points, respectively. Compared with EEIE-YOLOv8, the corresponding improvements are 1.4, 1.3, 1.2, and 0.6 percentage points. Although DCM-YOLO has slightly more parameters than these two task-specific methods, it reduces the FLOPs compared with EEIE-YOLOv8 and maintains competitive FPS. These comparisons show that the proposed model improves detection reliability without relying on excessive model expansion.

4.4. Qualitative Analysis and Visualization

4.4.1. Visualization of Detection Results

In Figures 5–8, the upper row shows the detection results of YOLOv8n, while the lower row shows the results of DCM-YOLO. The degraded images were manually generated from the original test images to provide a controllable robustness comparison. In Figures 5–7, each row contains one original image followed by four progressively degraded images. Specifically, the brightness is reduced by 20%, 40%, 60%, and 80% for underexposure, increased by 20%, 40%, 60%, and 80% for overexposure, and the contrast is reduced by 20%, 40%, 60%, and 80% for low-contrast degradation. In Figure 8, each row contains one original image followed by three blurred images, with blur strengths of 25%, 50%, and 75%.

The visual results show that YOLOv8n becomes less stable as the degradation level increases. In underexposure and low-contrast cases, it produces false detections when target boundaries and texture cues become weak. Under overexposure and blur, missed detections occur when brightness saturation or detail loss makes the electric bicycle less distinguishable from the background. In contrast, DCM-YOLO maintains more stable localization and classification results across the four degradation types. This improvement is mainly related to the coordinated design of the three proposed modules: DAF preserves more useful early-stage features, SCSA suppresses interference from degraded regions, and MSGA-Head improves prediction stability when local details are weakened. Therefore, the visual comparison provides additional evidence that DCM-YOLO is more robust than YOLOv8n under representative image degradation conditions.

4.4.2. Visualization Comparison of Detection Responses

To further analyze detection differences, heatmap visualizations and response distribution statistics were employed under the same challenging conditions. A qualitative comparison was conducted between YOLOv8n and DCM-YOLO under identical input settings, as shown in Figure 9. Row (a) presents the original input images. Row (b) shows the detection responses produced by YOLOv8n, and row (c) presents the responses of DCM-YOLO under the same conditions. From left to right, the four columns correspond to partial occlusion, incomplete surveillance view, occlusion, and severe occlusion, respectively.



Figure 5. Underexposure detection results: (**a**) YOLOv8n; (**b**) DCM-YOLO.



Figure 6. Overexposure detection results: (**a**) YOLOv8n; (**b**) DCM-YOLO.



Figure 7. Low-contrast detection results: (**a**) YOLOv8n; (**b**) DCM-YOLO.

As shown in row (b), YOLOv8n produces relatively diffuse and fragmented heatmap responses in all four cases. In the first and third columns, where the targets are partially occluded, the activations are distributed across both visible target regions and surrounding interference, which weakens localization focus. In the second column, where the surveillance image is incomplete, the response is less concentrated on the valid target area, indicating insufficient robustness to incomplete visual information. In the fourth column, under severe occlusion, the response becomes more scattered and less continuous, suggesting that the baseline model is more easily disturbed when visible target cues are heavily reduced.

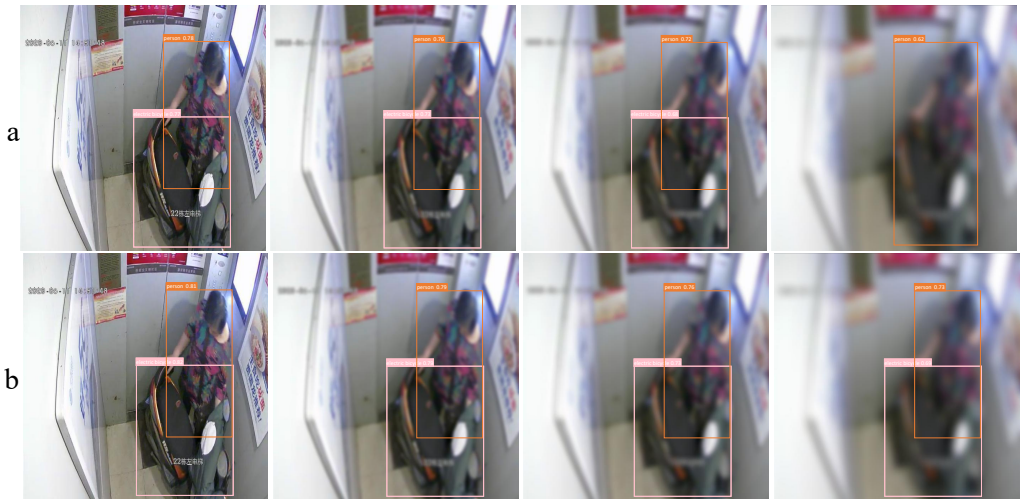


Figure 8. Blur detection results: (**a**) YOLOv8n; (**b**) DCM-YOLO.

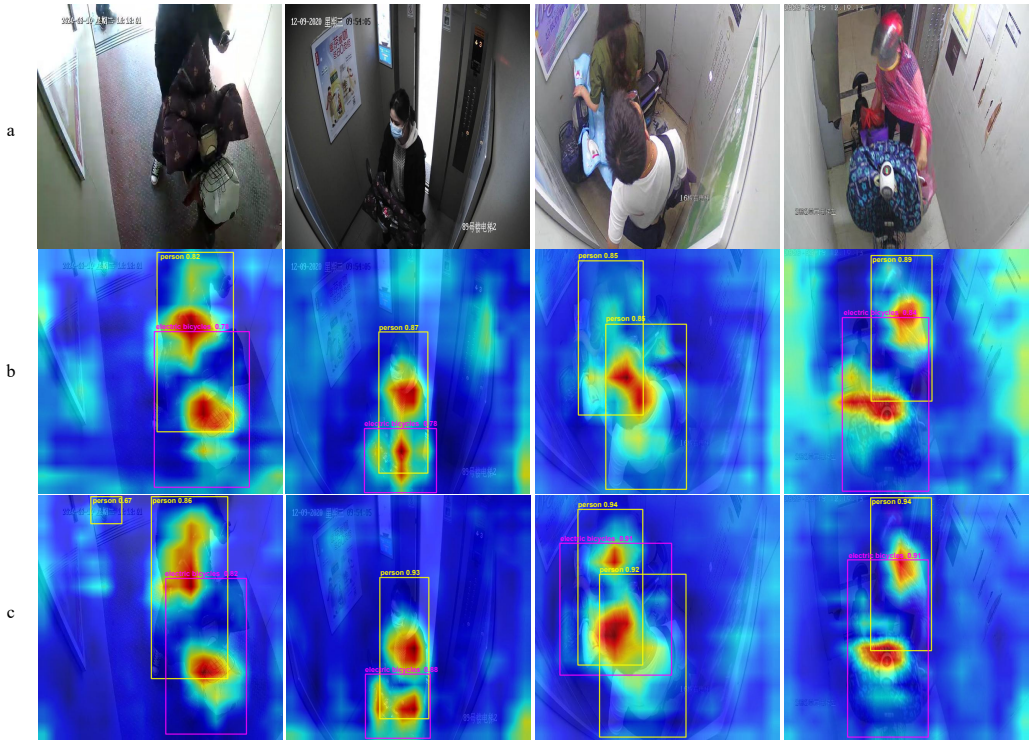


Figure 9. Heatmap Comparison. (**a**) Original; (**b**) YOLOv8n; (**c**) DCM-YOLO.

By contrast, row (c) shows that DCM-YOLO produces more concentrated and continuous responses on the visible target regions. In the partial occlusion and general occlusion cases, stronger activations are maintained around the target body and electric bicycle, while less attention is allocated to irrelevant background regions. In the incomplete-view and severe-occlusion cases, the response remains more stable and target-oriented, indicating better robustness when structural information is missing or only partially preserved.

The response distribution in Figure 9 is further summarized in Table 6. From left to right, the four columns correspond to partial occlusion, incomplete surveillance view, occlusion, and severe occlusion, respectively. Target Activation Ratio (TAR) denotes the proportion of heatmap response located inside the annotated target boxes, while Background Activation Ratio (BAR) denotes the proportion of responses distributed outside the target boxes.

Across the four occlusion-related conditions, DCM-YOLO consistently produces higher TAR and lower BAR than YOLOv8n. Specifically, TAR increases from 60.4%, 58.7%, 56.9%, and 55.8% to 70.8%, 69.5%, 67.8%, and 66.9%, respectively. Meanwhile, BAR decreases from 39.6%, 41.3%, 43.1%, and 44.2% to 29.3%, 30.5%, 32.2%, and 33.1%, respectively. These changes indicate that DCM-YOLO shifts more activation responses from background regions to visible target regions, which is consistent with the more concentrated heatmap responses shown in row (c) of Figure 9.

Table 6. Heatmap response distribution under different occlusion conditions.

Condition	Model	TAR/%	BAR/%
Partial occlusion	YOLOv8n	60.4	39.6
	DCM-YOLO	70.8	29.3
Incomplete surveillance view	YOLOv8n	58.7	41.3
	DCM-YOLO	69.5	30.5
Occlusion	YOLOv8n	56.9	43.1
	DCM-YOLO	67.8	32.2
Severe occlusion	YOLOv8n	55.8	44.2
	DCM-YOLO	66.9	33.1

This response redistribution is related to the coordinated design of DCM-YOLO. DAF helps preserve local structural cues when the target is only partially visible, SCSA suppresses interference from irrelevant spatial and channel responses, and MSGA-Head stabilizes prediction using multi-scale perception and grouped refinement. Together, these modules make the detector less dependent on background regions and produce more focused responses under different occlusion conditions.

4.4.3. Stage-Wise Analysis of Feature Responses

The response concentration differences indicate progressive feature refinement inside DCM-YOLO. To clarify how such refinement is formed, feature response visualizations are analyzed at different network stages. Feature responses are extracted before SCSA, after SCSA, and before the prediction head. The results are illustrated in Figure 10 using four representative scenarios, namely partial occlusion, incomplete surveillance view, occlusion, and severe occlusion.

4.5. Generalization Validation of DCM-YOLO

To evaluate the generalization performance of the proposed DCM-YOLO, experiments were conducted on three challenging public datasets. The first dataset is the Mobility Devices dataset, which contains 15,546 outdoor images covering 13 types of vehicles under diverse conditions, including low-light and fog. The second dataset is the Motorcar dataset, which contains 26,452 images focusing exclusively on electric bicycles. The third dataset is derived from COCO2017, where only images containing person and vehicle categories are retained. All experiments follow the same settings described in Section 4.1 to ensure fair and comparable evaluation across different datasets.

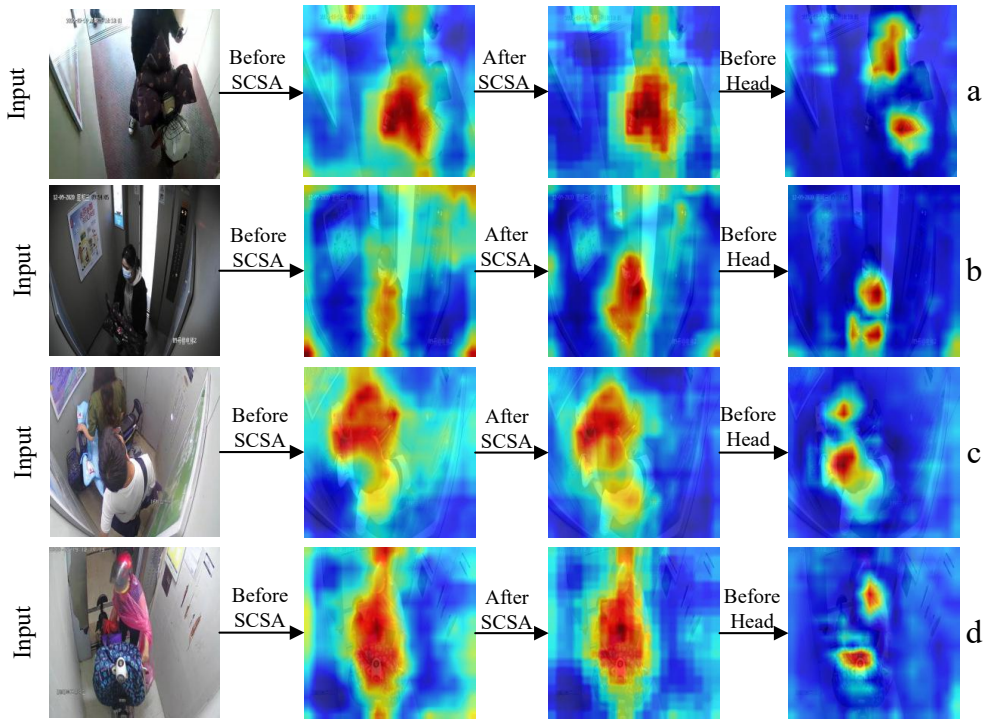


Figure 10. Feature response visualization at different network stages. (**a**) Partial occlusion; (**b**) incomplete surveillance view; (**c**) occlusion; and (**d**) severe occlusion. The columns show the input image, feature responses before SCSA, feature responses after SCSA, and feature responses before the prediction head, respectively.

As shown in Table 7, DCM-YOLO consistently outperforms the baseline YOLOv8n across all three datasets. On the Mobility Devices dataset, which contains more diverse outdoor conditions and multiple vehicle categories, DCM-YOLO still achieves stable improvements, indicating that the proposed design is not limited to the target scenario of the main dataset. On the Motorcar dataset, which is more closely related to electric bicycle detection, the gains are more pronounced, suggesting that the coordinated design of DAF, SCSA, and MSGA-Head is especially effective when target appearance and partial occlusion are closely related to the original task. On the filtered COCO2017 subset, DCM-YOLO also maintains consistent performance gains, which further verifies that the proposed method remains effective under broader category distributions and more complex visual conditions.

Overall, the experimental results on the three public datasets demonstrate that DCM-YOLO achieves consistent performance improvements under different data distributions. These results indicate that the proposed method has good generalization ability across different detection scenarios. The coordinated enhancement of feature extraction, discriminative feature selection, and prediction stability enables DCM-YOLO to remain effective not only in the original task setting, but also in broader indoor and outdoor visual environments.

Table 7. Performance of DCM-YOLO on Other Datasets.

Dataset	Algorithm	P%	R%	mAP50/%	mAP50-95/%
Mobility-Devices	YOLOv8n	66.8	59.4	63.1	46.8
	DCM-YOLO	69.2	61.8	65.2	47.9
Motorcar	YOLOv8n	88.1	82.6	86.3	65.2
	DCM-YOLO	90.3	85.1	88.9	66.5
COCO2017	YOLOv8n	69.7	62.5	66.4	48.5
	DCM-YOLO	72.3	65.0	68.8	49.8

5. Conclusions

This study proposes DCM-YOLO for robust electric bicycle detection in confined indoor environments affected by occlusion and image degradation. Unlike methods that rely mainly on general architectural scaling or isolated module replacement, DCM-YOLO introduces a coordinated backbone–attention–head design to address weakened visual cues and incomplete target appearances. Specifically, DAF provides adaptive backbone representation by combining lightweight feature generation with response modulation, SCSA performs sequential spatial–channel refinement to suppress interference-prone responses, and MSGA-Head introduces scale-aware prediction for more reliable classification and localization.

Comprehensive experiments verify the effectiveness of the proposed design. Ablation studies show that the three components are complementary and jointly contribute to the final performance gain. Comparative experiments further demonstrate that DCM-YOLO achieves the best overall detection performance among the compared methods. On the main dataset, DCM-YOLO reaches 86.2% $mAP50$ and 65.1% mAP50-95, improving upon the baseline YOLOv8n by 2.8% and 1.7%, respectively. Additional evaluations on public datasets, including Mobility Devices, Motorcar, and COCO2017, also show consistent improvements, which confirms the generalization ability of the proposed method under different data distributions. Visual comparison and heatmap analysis further indicate that DCM-YOLO produces more reliable and concentrated responses under underexposure, overexposure, low contrast, blur, and different occlusion-related conditions.

Although DCM-YOLO achieves strong overall performance, several limitations still remain. First, the current model is mainly evaluated under fixed-camera settings, and its robustness under moving-camera conditions or more dynamic viewpoints still requires further investigation. Second, although the proposed design is effective for occlusion and image degradation, more severe real-world domain shifts may still affect detection stability. Third, the current study mainly focuses on detection accuracy and robustness, while further optimization for practical deployment can still be explored in future work.

In future work, the proposed framework will be extended to more dynamic and diverse visual scenarios, including mobile-camera and cross-domain detection settings. More robust adaptation strategies for unseen degradations and complex structural interference will also be investigated. In addition, further refinement of prediction-stage design and more efficient deployment-oriented optimization may help broaden the applicability of DCM-YOLO in real-world vision tasks.

Author Contributions: Conceptualization, G.Z.; Methodology, G.Z.; Software, Y.S.; Validation, Y.X.; Formal analysis, Y.S.; Investigation, G.Z.; Resources, R.C.; Data curation, Y.S.; Writing—original draft, Y.X.; Writing—review & editing, Y.W. and H.J.; Visualization, M.T.; Funding acquisition, R.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The dataset used in this study is publicly available at: https://aistudio.baidu.com/datasetdetail/350592 (accessed on 19 July 2025). The source code, trained weights, and experimental configuration files of DCM-YOLO will be made publicly available at https://github.com/syy-zt/DCM-YOLO (accessed on 21 May 2026). The released materials will include the implementations of DAF, SCSA, and MSGA-Head, the model configuration files, training scripts, inference scripts, and the hyperparameter settings used in the experiments.

Acknowledgments: During the preparation of this manuscript, the authors used ChatGPT (GPT-5.5 Thinking, OpenAI), for language polishing and improving the readability of selected analysis paragraphs. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The author Guanfang Zuo was employed by the Jiangsu Province Engineering Research Center of Photonic Devices and System Integration for Communication Sensing Convergence. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2014; pp. 580–587.
2. Girshick, R. Fast R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*; IEEE: Piscataway, NJ, USA, 2015; pp. 1440–1448.
3. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Nice, France, 2015; Volume 28, pp. 91–99.
4. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. SSD: Single Shot MultiBox Detector. In *Computer Vision–ECCV 2016*; Leibe, B., Matas, J., Sebe, N., Welling, M., Eds.; Springer: Cham, Switzerland, 2016; pp. 21–37.
5. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2016; pp. 779–788.
6. Redmon, J.; Farhadi, A. YOLO9000: Better, Faster, Stronger. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2017; pp. 7263–7271.
7. Bochkovskiy, A.; Wang, C.-Y.; Liao, H.-Y.M. YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv* **2020**, arXiv:2004.10934.
8. Li, C.; Li, L.; Jiang, H.; Weng, K.; Geng, Y.; Li, L.; Ke, Z.; Li, Q.; Cheng, M.; Nie, W.; et al. YOLOv6: A Single-Stage Object Detection Framework for Industrial Applications. *arXiv* **2022**, arXiv:2209.02976.
9. Wang, C.-Y.; Bochkovskiy, A.; Liao, H.-Y.M. YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2023; pp. 7464–7475.
10. Wang, C.-Y.; Yeh, I.-H.; Liao, H.-Y.M. YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information. In *Computer Vision–ECCV 2024*; Springer: Cham, Switzerland, 2024.
11. Wang, A.; Chen, H.; Liu, L.; Chen, K.; Lin, Z.; Han, J.; Ding, G. YOLOv10: Real-Time End-to-End Object Detection. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Nice, France, 2024; Volume 37.
12. Tian, Y.; Ye, Q.; Doermann, D. YOLOv12: Attention-Centric Real-Time Object Detectors. *arXiv* **2025**, arXiv:2502.12524.
13. Lei, M.; Li, S.; Wu, Y.; Hu, H.; Zhou, Y.; Zheng, X.; Ding, G.; Du, S.; Wu, Z.; Gao, Y. YOLOv13: Real-Time Object Detection with Hypergraph-Enhanced Adaptive Visual Perception. *arXiv* **2025**, arXiv:2506.17733.
14. Li, F.; Lv, X.; Zhao, M.; Wu, W. IFD-YOLO: A Lightweight Infrared Sensor-Based Detector for Small UAV Targets. *Sensors* **2025**, *25*, 7449. [PubMed]
15. Lu, Y.-F.; Gao, J.-W.; Yu, Q.; Li, Y.; Lv, Y.-S.; Qiao, H. A Cross-Scale and Illumination Invariance-Based Model for Robust Object Detection in Traffic Surveillance Scenarios. *IEEE Trans. Intell. Transp. Syst.* **2023**, *24*, 6989–6999.
16. Zhao, D.; Zhou, L.; Li, Y.; Wu, Y.; Hu, J.; Zhou, H. Visibility Estimation via Near-Infrared Bispectral Real-Time Imaging in Bad Weather. *Infrared Phys. Technol.* **2024**, *136*, 105008.
17. Cui, X.; Ma, L.; Ma, T.; Liu, J.; Tang, Y.; Zhang, Y. Trash to Treasure: Low-Light Object Detection via Decomposition-and-Aggregation. In *Proceedings of the AAAI Conference on Artificial Intelligence*; AAAI Press: Washington, DC, USA, 2024; Volume 38, pp. 1417–1425.
18. Tran, D.Q.; Aboah, A.; Jeon, Y.; Shoman, M.; Park, M.; Park, S. Low-Light Image Enhancement Framework for Improved Object Detection in Fisheye Lens Datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*; IEEE: Piscataway, NJ, USA, 2024; pp. 7056–7065.
19. Gupta, H.; Kotlyar, O.; Andreasson, H.; Lilienthal, A.J. Robust Object Detection in Challenging Weather Conditions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*; IEEE: Piscataway, NJ, USA, 2024; pp. 7523–7532.
20. Chen, M.; Liu, Y.; Zhang, Z.; Guo, W. RCRFNet: Enhancing Object Detection with Self-Supervised Radar–Camera Fusion and Open-Set Recognition. *Sensors* **2024**, *24*, 4803. [PubMed]
21. Li, G.; Wang, Y.; He, B.; Pang, T.; Gao, M. Low-Light Multimodal Object Detection: A Survey. *Comput. Sci. Rev.* **2025**, *58*, 100804. [CrossRef]
22. Zhao, D.; Wang, M.; Huang, K.; Zhong, W.; Arun, P.V.; Li, Y.; Asano, Y.; Wu, L.; Zhou, H. OCSCNet-Tracker: Hyperspectral Video Tracker Based on Octave Convolution and Spatial–Spectral Capsule Network. *Remote Sens.* **2025**, *17*, 693.
23. Zhao, D.; Yan, W.; You, M.; Zhang, J.; Arun, P.V.; Jiao, C.; Wang, Q.; Zhou, H. Hyperspectral Anomaly Detection Based on Empirical Mode Decomposition and Local Weighted Contrast. *IEEE Sens. J.* **2024**, *24*, 33847–33861. [CrossRef]

24. Zhang, Z.; Yang, X.; Wu, C. An Improved Lightweight YOLOv5s-Based Method for Detecting Electric Bicycles in Elevators. *Electronics* **2024**, *13*, 2660.
25. Wan, Y.; Hu, Y.; Li, X.; Song, Z.; Hou, T. Improved Algorithm of Elevator Blocking System Based on YOLOv8. In *Proceedings of the 12th International Conference on Information Systems and Computing Technology (ISCTech)*; IEEE: Piscataway, NJ, USA, 2024.
26. Lai, L.; Xiong, A.; Yan, A.; Zhang, G. EEIE-YOLO: Enhanced Edge Information Extraction YOLO for Electric Bicycle Detection in Elevator. In *Proceedings of the International Conference on Image Processing, Computer Vision and Machine Learning*; SPIE: Philadelphia, PA, USA, 2024; pp. 617–621.
27. He, J.; Chen, H.; Liu, B.; Luo, S.; Liu, J. Enhancing YOLO for Occluded Vehicle Detection with Grouped Orthogonal Attention and Dense Object Repulsion. *Sci. Rep.* **2024**, *14*, 19650. [CrossRef] [PubMed]
28. Guo, B.; Zhang, H.; Wang, H.; Li, X.; Jin, L. Adaptive Occlusion Object Detection Algorithm Based on OL-IoU. *Sci. Rep.* **2024**, *14*, 27644. [CrossRef] [PubMed]
29. Zheng, Y.; Qi, H.; Li, L.; Li, S.; Huang, Y.; He, C.; Wang, D. Motion-Guided and Occlusion-Aware Multi-Object Tracking with Hierarchical Matching. *Pattern Recognit.* **2024**, *151*, 110369.
30. Teng, X.; Zhao, D.; Xiang, P.; Yu, X.; Li, Z.; Hsu, C.-C.; Yang, T.; Zhou, H.; Ren, J. UAV-Based Multispectral Object Tracking with Positive-Negative Prompt Mining Network. *IEEE Trans. Geosci. Remote Sens.* **2026**, *64*, 5515418. [CrossRef]
31. Jiang, W.; Zhao, D.; Wang, C.; Yu, X.; Arun, P.V.; Asano, Y.; Xiang, P.; Zhou, H. Hyperspectral Video Object Tracking with Cross-Modal Spectral Complementary and Memory Prompt Network. *Knowl.-Based Syst.* **2025**, *330*, 114595. [CrossRef]
32. Zhao, D.; Zhong, W.; Ge, M.; Arun, P.V.; Li, Y.; Asano, Y.; Wu, L.; Zhou, H. SiamBSI: Hyperspectral Video Tracker Based on Band Correlation Grouping and Spatial–Spectral Information Interaction. *Infrared Phys. Technol.* **2025**, *151*, 106063.
33. Yao, Y.; Wang, Q.; Zhao, D.; You, M.; Xiang, P.; Asano, Y.; Yu, X.; Wang, C.; Zhou, H.; Ren, J. DFBSNet: Dual Frequency-domain Branch Fusion and Selection Network for Hyperspectral Anomaly Detection. *Pattern Recognit.* **2026**, *180*, 113967. [CrossRef]
34. Chen, C.; Li, B. A Transform Module to Enhance Lightweight Attention by Expanding Receptive Field. *Expert Syst. Appl.* **2024**, *248*, 123359. [CrossRef]
35. Yu, Y.; Zhang, Y.; Cheng, Z.; Xu, Y. Multi-Scale Spatial Pyramid Attention Mechanism for Image Recognition: An Effective Approach. *Eng. Appl. Artif. Intell.* **2024**, *133*, 108261. [CrossRef]
36. Zhu, L.; Wang, X.; Ke, Z.; Zhang, W.; Lau, R.W.H. BiFormer: Vision Transformer with Bi-Level Routing Attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2023; pp. 2493–2502.
37. Zhao, D.; Hu, B.; Jiang, W.; Zhong, W.; Arun, P.V.; Cheng, K.; Zhao, Z.; Zhou, H. Hyperspectral Video Tracker Based on Spectral Difference Matching Reduction and Deep Spectral Target Perception Features. *Opt. Lasers Eng.* **2025**, *194*, 109124. [CrossRef]
38. Zhao, D.; Zhang, H.; Arun, P.V.; Jiao, C.; Zhou, H.; Xiang, P.; Cheng, K. SiamSTU: Hyperspectral Video Tracker Based on Spectral Spatial Angle Mapping Enhancement and State-Aware Template Update. *Infrared Phys. Technol.* **2025**, *150*, 105919.
39. Li, M.; Zhao, Y.; Gui, G.; Zhang, F.; Luo, B.; Yang, C.; Gui, W.; Chang, K.; Wang, H. Object Detection on Low-Resolution Images with Two-Stage Enhancement. *Knowl.-Based Syst.* **2024**, *299*, 111985.
40. Xue, X.; Zheng, H.; Gao, Y.; Ma, T.; Jia, Q.; Wang, H. Crossing the Chasm: A Practical Architecture Augmentation for Low-Quality Object Detection. *Neurocomputing* **2025**, *625*, 129553. [CrossRef]
41. Xiong, M.; Zhang, Q.; Li, D.; Wang, W.; Zhang, Z.; Zhang, K.; Liu, C.; Chen, D.; Zhang, J. Fine-Tuning Feature Interaction for Unsupervised Domain Adaptive Low-Light Object Detection. *Neurocomputing* **2025**, *658*, 131717.
42. Nguyen, T.-D.; Le, D.-T. MODE: A Model-Agnostic Framework for Object Detection under Adverse Weather Conditions. *Pattern Recognit.* **2026**, *173*, 112868.
43. Zhang, Y.; Xuan, S.; Li, Z. Robust Object Detection in Adverse Weather with Feature Decorrelation via Independence Learning. *Pattern Recognit.* **2026**, *169*, 111790.
44. Han, K.; Wang, Y.; Tian, Q.; Guo, J.; Xu, C.; Xu, C. GhostNet: More Features from Cheap Operations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020*; IEEE: Piscataway, NJ, USA, 2020; pp. 1580–1589.
45. Ma, N.; Zhang, X.; Zheng, H.-T.; Sun, J. ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design. In *Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 116–131.
46. Dikici, B.; Bekciogullari, M.F.; Acikgoz, H.; Ozbay, S. A Lightweight and Improved You Only Look Once Model Using GhostWise Convolution and Attention Mechanism for Accurate Plant Disease Detection. *Eng. Appl. Artif. Intell.* **2025**, *161*, 112163. [CrossRef]
47. Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017*; IEEE: Piscataway, NJ, USA, 2017; pp. 2117–2125.
48. Zhao, L.; Zhang, Z. An Improved Pooling Method for Convolutional Neural Networks. *Sci. Rep.* **2024**, *14*, 1589. [CrossRef] [PubMed]

49. Zhao, Z.; Song, J.; Zhao, D.; Zhang, J.; Zhou, H.; Zhou, J. Multi-Scale Frequency-Guided Two-Stream Network for Hyperspectral Anomaly Detection. *Int. J. Appl. Earth Obs. Geoinf.* **2025**, *136*, 104355.
50. Chen, J.; Kao, S.-H.; He, H.; Zhuo, W.; Wen, S.; Lee, C.-H.; Chan, S.-H.G. Run, Don't Walk: Chasing Higher FLOPS for Faster Neural Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2023; pp. 12021–12031.
51. Si, Y.; Xu, H.; Zhu, X.; Zhang, W.; Dong, Y.; Chen, Y.; Li, H. SCSA: Exploring the Synergistic Effects Between Spatial and Channel Attention. *Neurocomputing* **2025**, *634*, 129866. [CrossRef]
52. Lin, W.; Wu, Z.; Chen, J.; Huang, J.; Jin, L. Scale-Aware Modulation Meet Transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: Piscataway, NJ, USA, 2023; pp. 5992–6003.
53. Zuo, F.; Liu, J.; Chen, Z.; Fu, M.; Wang, L. TGADHead: An Efficient and Accurate Task-Guided Attention-Decoupled Head for Single-Stage Object Detection. *Knowl.-Based Syst.* **2024**, *302*, 112349.
54. Hua, J.; Wang, Z.; Chang, J.; Zou, H.; Wu, G.; Ma, J. MSTDNet: Multi-Scale Traffic Object Detection Network with Smooth Information Perception. *Pattern Recognit.* **2026**, *172*, 112579.
55. Baidu AI Studio. Elevator Electric Bicycle Detection Dataset. Available online: https://aistudio.baidu.com/datasetdetail/350592 (accessed on 21 May 2026).
56. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2018; pp. 7132–7141.
57. Liu, X.; Peng, H.; Zheng, N.; Yang, Y.; Hu, H.; Yuan, Y. EfficientViT: Memory Efficient Vision Transformer with Cascaded Group Attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2023; pp. 14420–14430.
58. Lau, K.W.; Po, L.-M.; Rehman, Y.A.U. Large Separable Kernel Attention: Rethinking the Large Kernel Attention Design in CNN. *Expert Syst. Appl.* **2024**, *236*, 121352.
59. Jiao, J.; Tang, Y.-M.; Lin, K.-Y.; Gao, X.; Sang, N.; Zheng, W.-S. DilateFormer: Multi-Scale Dilated Transformer for Visual Recognition. *IEEE Trans. Multimed.* **2023**, *25*, 8906–8919.
60. Xia, Z.; Pan, X.; Song, S.; Li, L.E.; Huang, G. Vision Transformer with Deformable Attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2022; pp. 4794–4803.
61. Wan, D.; Lu, R.; Shen, S.; Xu, T.; Lang, X.; Ren, Z. Mixed Local Channel Attention for Object Detection. *Eng. Appl. Artif. Intell.* **2023**, *123*, 106442. [CrossRef]
62. Zhao, D.; Zhang, H.; Huang, K.; Zhu, X.; Arun, P.V.; Jiang, W.; Li, S.; Pei, X.; Zhou, H. SASU-Net: Hyperspectral Video Tracker Based on Spectral Adaptive Aggregation Weighting and Scale Updating. *Expert Syst. Appl.* **2025**, *272*, 126721. [CrossRef]
63. Howard, A.; Sandler, M.; Chu, G.; Chen, L.-C.; Chen, B.; Tan, M.; Wang, W.; Zhu, Y.; Pang, R.; Vasudevan, V.; Le, Q.V.; Adam, H. Searching for MobileNetV3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: Piscataway, NJ, USA, 2019; pp. 1314–1324.
64. Narayanan, M. SENetV2: Aggregated Dense Layer for Channelwise and Global Representations. *arXiv* **2023**, arXiv:2311.10807.
65. Xu, Z.; Ma, G.; Gao, R. An Optimized PP-HGNetV2-Enhanced Lightweight YOLOv11 for Robust Object Detection in VisDrone UAV Scenes. In *Proceedings of the 2025 5th International Conference on Electronic Information Engineering and Computer Communication (EIECC)*; IEEE: Piscataway, NJ, USA, 2025; pp. 868–872.
66. Tang, Y.; Han, K.; Guo, J.; Xu, C.; Xu, C.; Wang, Y. GhostNetV2: Enhance Cheap Operation with Long-Range Attention. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 9969–9982.
67. Zhao, D.; Xu, X.; You, M.; Arun, P.V.; Zhao, Z.; Ren, J.; Wu, L.; Zhou, H. Local Sub-Block Contrast and Spatial–Spectral Gradient Features Fusion for Hyperspectral Anomaly Detection. *Remote Sens.* **2025**, *17*, 695.

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