https://doi.org/10.31449/inf.v50i14.12887 Informatica **50** (2026) 123–136 **123**

YOLO-SCEMA: Efficient Multi-scale Attention Fusion Module Combining Spatial and Channel Reconstruction Convolution

Zhuang Yang[1], Haiying Wang[1], Xiaolong Fu[1], Ming Hui[1,2], Xiaolong Gao[1], Kun Ning[1], Zufeng Fu[1,2,*]
[1]Nanyang Normal University, Department of Artificial Intelligence and Software Engineering, Nanyang, Henan, 473061, China
[2]Collaborative Innovation Center of Intelligent Explosion-proof Equipment, Nanyang, Henan, 473061, China
E-mail:2242450789@qq.com, why8206@163.com, xl_f@foxmail.com, 20131130@nynu.edu.cn, xl_g@foxmail.com, nk122@foxmail.com, fuzufeng@outlook.com
*Corresponding author

Keywords: Complex scene object detection, lightweight, attention mechanisms, YOLO, multiscale fusion

Received: January 5, 2026

Convolutional Neural Networks (CNNs) have achieved significant performance in various computer vision tasks, but at the cost of enormous computing resources. To alleviate this dilemma, this paper proposes a lightweight and efficient multiscale attention fusion module (SCEMA) by combining spatial, and channel reconstruction convolution with an efficient attention mechanism and adds the module to the YOLOV8 network structure named YOLO-SCEMA. SCEMA adopts a parallel processing strategy, with the left branch performing feature refinement operations through spatial and channel reconstruction units to reduce redundant calculations, and the right branch effectively integrating features of different scales through feature grouping and cross-spatial learning, enhancing the model's understanding of multiscale image content and improving its performance in handling complex image structures. The experimental results on the open source datasets ExDark,VisDrone2019 and FYP show that YOLO-SCEMA has increased the mAP(50) score by 7.37%,3.24% and 1.5% , and compared to the YOLOv8 benchmark while reducing the parameter and computational complexity by 36.9% and 8.6%, respectively. Compared with the latest YOLO series, YOLO-SCEMA performs better in detection accuracy and parameter quantity.

Povzetek: Članek predstavi YOLO-SCEMA, lahek nadgradni model YOLOv8 z večmerilnim modulom pozornostne fuzije, ki z rekonstrukcijo prostorskih/kanalnih značilk in učinkovito integracijo različnih meril izboljša detekcijo ob manj parametrih in nižji računski zahtevnosti.

1 Introduction

Convolutional Neural Networks (CNNs) can automatically extract local features from images and combine low-level features into high-level feature representations through layer-by-layer transmission. Due to their strong robustness and adaptability, CNNs have been widely used in computer vision tasks [16]. However, this success is largely based on intensive computing and storage resources, which poses a serious challenge to its effective deployment in resource-constrained environments. As CNN layers deepen, progressively increasing the number of feature channels from layer to layer increases the expressive capacity of the features [1]. However, this leads to redundancy in space and channels and requires a significant amount of memory and computational resources, which is the main drawback of building deep CNNs [17]. As an alternative method of increasing the number of feature channels layer by layer, by introducing attention mechanisms, neural networks can automatically learn and selectively focus on important information in the input, improving the performance and generalization ability of the model. Thanks to its flexible structure, the attention mechanism not only enhances the learning of discriminative features but also integrates easily into CNN backbones. Therefore, attention mechanisms have aroused widespread interest in the field of computer vision research. At present, there are mainly two types of attention mechanisms, such as channel attention and spatial attention. As a representative channel attention, Squeeze and Excitation (SE) simulates cross dimensional interactions to extract channel attention [4]. The Convolutional Block Attention Module (CBAM) establishes cross-channel and cross-spatial information through semantic interdependence between spatial and channel dimensions in feature maps [23]. Therefore, CBAM shows great potential in integrating cross-dimensional attention weights into input features. However, CBAM requires attention to channel and space calculations and introduces additional parameters, which increases computational complexity and leads to increased training and inference time for the model. To overcome the limitation of computational cost, Alex et al. proposed a more effective method of using feature grouping to divide features into multiple groups in different spatial dimensions [6], which can make each group of fea-

tures evenly distributed in space. According to this setting, Spatial Group-wise Enhance (SGE) [10] pays attention to generating attention factors through the similarity between global and local feature descriptors, enhancing the spatial distribution of semantic features and suppressing noise.

One of the most effective ways to simplify model complexity is by utilizing Depthwise Separable Convolution in conjunction with an attention mechanism to enhance the original model. Separable depth convolution initially conducts depth convolution on each input channel before consolidating information across channels through point convolution, leading to a significant reduction in model parameters and making the model lighter [13]. The SBC-FormerBlock [12] improves model performance by refining the handling of local and global features, merging them to produce feature maps that encompass both local and global details.

The Multidimensional Collaborative Attention Module (MCA) [26] comprises a three-branch architecture capable of simultaneously capturing attention in three dimensions: channel, height, and width, thereby enhancing feature representation and boosting network performance with minimal computational overhead. The Efficient Channel Attention Module (ECA) [21] dynamically captures inter-channel dependencies through a straightforward and effective one-dimensional convolution, eliminating the necessity for dimensionality reduction or enhancement. This approach sidesteps the intricate multi-layer perceptron (MLP) structure in conventional attention mechanisms, reducing model complexity and computational load. Ouyang et al. modified the sequential processing approach of Coordinate Attention (CA) [3] through a grouping structure and introduced an Efficient Multi-scale Attention Module (EMA) [14] without dimensionality reduction. The EMA module adeptly establishes short-range and long-range dependencies by grouping features and processing multi-scale structures in various spatial dimensions to achieve superior performance.

Inspired by the above convolution module and attention mechanism, we adopt a multi-branch structure. By replacing the convolution module in the original network structure and combining the attention mechanism, it helps to improve the feature expression ability of the model and reduce the complexity. We propose a lightweight and efficient multi-scale attention fusion module (YOLO-SCEMA) by incorporating SCConv [9] convolution with EMA and CBAM Attention into the YOLOV8 network architecture. This module reconstructs the original spatial and channel units to reduce redundant calculations and integrate features of different scales, enhancing the model's understanding of multi-scale image content and improving its performance in handling complex image structures.

The primary objective of this study is to validate the universality and efficiency of the proposed YOLO-SCEMA module across diverse and challenging detection scenarios. The research questions addressed are:

Can YOLO-SCEMA improve detection accuracy in low-light, dense small-object, and complex image environments compared not only to the benchmark YOLOv8n but also to other state-of-the-art (SOTA) models?

Can YOLO-SCEMA achieve these improvements while simultaneously reducing parameter count and computational complexity, thereby enhancing deployment efficiency?

In comparison with existing attention mechanisms and lightweight designs, what specific innovations does SCEMA introduce to balance accuracy and efficiency?

To address these questions, Section 4 presents a series of comparative experiments conducted on three representative datasets: ExDark (low-light detection), VisDrone2019 (dense small-object detection), and FYP (complex image structures). The experimental results demonstrate that YOLO-SCEMA achieves consistent improvements in mAP(50) of 7.37%, 3.24%, and 1.5% across the three datasets, while reducing parameters by 36.9% and FLOPs by 8.6%. These findings confirm that YOLO-SCEMA not only enhances accuracy but also maintains efficiency. Compared with existing methods, SCEMA's innovation lies in its efficient multi-scale fusion of spatial and channel reconstruction, which enables robust feature representation with reduced computational overhead. This design supports its suitability for real-world deployment in resource-constrained environments.

2 Related work

In this section, we will present some prominent works on the structure of target detection networks and attention modules.

2.1 YOLO object detection

Compared with the two-stage R-CNN series algorithms [2], the one-stage YOLO series algorithms [5] skip the candidate region generation process and can directly conduct forward propagation on the image to achieve detection results. YOLO exhibits superior overall performance in terms of both detection speed and accuracy. It is the pioneering object detection algorithm to unify object detection and classification as a regression issue, utilizing a single convolutional neural network design to identify and categorize objects within the entire image. The object detection and recognition process of YOLO algorithms primarily involves three key steps: resizing the image, predicting with the neural network, and processing detection information. Initially, YOLO resizes the image to a suitable dimension for neural network prediction. Subsequently, the neural network extracts and processes features from the input image through the backbone, neck, and head structures. Following this, the neural network provides the detection information of the target. Finally, redundant detection information undergoes non-maximum suppression to yield the ultimate object detection data. Building upon

its predecessor, YOLOv8 [18] introduces novel functionalities and enhancements, positioning it as a preferred solution for diverse object detection assignments across various domains. YOLOv8 implements cutting-edge backbone and neck architectures to enhance feature extraction and object detection capabilities. Moreover, YOLOv8 integrates an anchor-free split head, which enhances the precision and efficiency of the detection process in comparison to anchor-based techniques.

2.2 Attention mechanism

Attention mechanisms have become an important research direction in recent years, especially in natural language processing and computer vision. Attention mechanism can simulate human's ability to focus attention and dynamically allocate computing resources, thus improving the performance of the model. SENet improves the representativeness of the convolutional neural network through the introduction of a novel structural unit known as the "Squeeze and Excitation" (SE) block. This innovation explicitly captures the relationships between convolutional feature channels, leading to notable enhancements in network performance without a rise in computational expenses. The Bottleneck Attention Module (BAM) [15] enhances the network's ability to represent information by incorporating channel and spatial attention. This module generates attention maps using two distinct pathways (channel and spatial), offering compatibility with various feedforward convolutional neural networks. The Shuffle Attention module, as introduced by Zhang et al. (2021) [27], organizes channel dimensions into several sub-features, processes these sub-features simultaneously, and employs the "channel shuffle" technique to facilitate information exchange among them. This approach enhances the efficiency and efficacy of the SA module. The Global Attention Mechanism (GAM) proposed by Liu et al. (2021) [11] aims to mitigate information dispersion and boost global feature interaction by incorporating channel and spatial dual attention. This approach ultimately enhances the performance of various visual tasks. CBAM combines Channel Attention and Spatial Attention for the first time to extract input features in two stages. This design prioritizes identifying 'which channels are important' initially, followed by determining 'which positions in space are important'. This sequential approach enables the model to capture crucial information within the features more comprehensively. The core mechanism of ECA module is to adaptively capture the dependencies between channels through a simple and efficient one-dimensional convolution, without the process of dimension reduction and dimension increase. This design avoids the complex multilayer perceptron (MLP) structure in the traditional attention mechanism, and reduces the model complexity and computational burden.

2.3 Multi-attention fusion module

Multi-attention fusion module has demonstrated exceptional performance across a range of computer vision tasks. Typical fusion techniques involve multi-scale feature fusion and parallel multi-branch structures. The former integrates features at different levels to capture multi-scale information, while the latter addresses multi-scale structures in various dimensions using parallel multi-branch structures. The High-Similarity-Pass Attention Module (HSPAM) [19] combines High-Similarity-Pass Attention (HSPA) and Locality Exploration Block (LEB) to extract both non local and local features. HSPA is responsible for extracting useful nonlocal information, while LEB enhances local feature expression through convolution. The combination of the two improves the feature extraction ability and super-resolution reconstruction effect of the model. The Lightweight Cross Attention Module (LCA) [24] uses the cross attention mechanism to let the brightness branch and color branch provide information to each other, so as to avoid brightness deviation or color distortion caused by separate processing. By calculating the attention distribution of brightness features to color features and the attention of color features to brightness features, this module can more accurately adjust the light intensity, optimize the color consistency, and make the enhanced image more natural. The Pinwheel-Shaped Convolutional Module (PSConv) [25] proposes a scale based dynamic (SD) Loss, which dynamically adjusts the impact of scale and location loss according to the size of the target, thus improving the network's ability to detect targets of different scales. Traditional global attention mechanisms tend to focus on the overall semantic structure of images, and may ignore local non semantic features when conducting global interaction. The Multi-Scale Linear Attention U-Net (MSLAU-Net)[7] is a hybrid CNN-Transformer encoder-decoder architecture, which core introduces the multi-scale linear attention (MSLA) module. It integrates the advantages of multi-scale local feature extraction of CNN and global dependency modeling of Transformer, restores spatial details with low computational complexity. The Mamba-YOLO[22] innovatively integrates the state space model (SSM) into the classic YOLO detection paradigm, takes ODMamba as the backbone network, and combines RG Block to enhance local detail capture and occluded target localization. While ensuring detection accuracy superior to peer YOLO models, it effectively reduces computational overhead and improves inference speed. The Shunted Self-Attention Module (SSA) [20] suppresses the expression of semantic information by restricting attention calculation to specific tensor blocks, allowing the model to focus more on capturing these non-semantic local features. The Contrast-Driven Feature Aggregation Module (CDFA) [8] realizes multi-level feature fusion and enhancement by fusing multi-level feature maps from the encoder and combining foreground and background feature maps. By comparing foreground and background features, CDFA module can highlight key

Table 1: Comparison of representative attention mechanisms and lightweight models

Method	Core Idea	Params	GFLOPs	mAP@50	Limitations
EMA	Multi-scale feature grouping	3.02	8.2	FYP: 68.94	Complex structure
ECA	1D convolution for channel dependencies	2.65	8.4	VisDrone2019: 32.27	Ignores spatial features
CBAM	Channel + spatial attention	3.07	8.2	ExDark: 75.46 VisDrone2019: 32.62 FYP: 69.12	Higher latency
MSLAU-Net	Integrates multi-scale linear attention into U-Net	21.9	N/A	ExDark: 76.86	Large parameters
Mamba-YOLO-T	Combines state-space modeling with a lightweight YOLO backbone	5.8	13.2	VisDrone2019: 35.47	Heavy computation

features and suppress non key features. This makes the model pay more attention to the important foreground area in the segmentation process, reduce background interference,and improve the segmentation accuracy.

The inclusion of Table 1 provides a clear overview of the strengths and limitations of existing approaches. For example, while MSLAU-Net and Mamba-YOLO-T achieve slightly higher accuracy than SCEMA, they require substantially more parameters and computation, which limits their suitability for resource-constrained environments. By contrast, SCEMA achieves competitive accuracy with significantly fewer parameters and lower computational cost, thereby addressing the identified gaps in prior works. Unlike traditional multi-attention approaches that rely on sequential or single-path mechanisms and risk information loss and redundancy, SCEMA introduces a parallel multi-branch fusion strategy integrating SRU, CRU, CSL, and CBAM. SRU and CRU reconstruct spatial and channel features in a lightweight manner to reduce redundancy and improve efficiency, while CSL and CBAM enhance accuracy through multi-scale integration and selective attention. By combining these modules in parallel, SCEMA achieves dual optimization in efficiency and accuracy, avoids single-path bottlenecks, and ensures richer feature representation. Ablation studies further validate that this parallel fusion consistently outperforms partial or sequential configurations, underscoring the novelty and practicality of SCEMA for lightweight object detection in complex scenarios.

3 Methodology

In this section, we introduce the proposed SCEMA module and the YOLO-SCEMA model constructed upon it. The architecture of SCEMA is illustrated in Figure 1. SCEMA adopts a parallel substructure processing strategy. In the left branch, the Spatial-Refined Feature X^w is extracted through the SRU and spatial attention operations, followed by the derivation of the Channel-Refined Feature Y via the CRU and channel attention operations. The right branch emphasizes learning an effective channel representation by grouping features without reducing the channel dimension during convolution. This design enhances pixel-level attention in advanced feature maps, thereby improving the ability to handle multi-scale images. The feature grouping process is indicated by the blue dotted line. The structure of the YOLO-SCEMA model is shown in Figure 2. We integrate SCEMA into the YOLOv8 backbone. In the Convolutional Block (CBS), BatchNorm is replaced by the Feature Refinement Module and the Cross-Spatial Learning Module, yielding two variants: CBS-FR and CBS-CSL. In the Bottleneck structure, the second convolutional layer is substituted with the Feature Refinement Module, resulting in Bottleneck-FR. Moreover, the Convolutional Blocks (CBS) in the Spatial Pyramid Pooling Fast (SPPF) are replaced with CBS-FR and CBS-CSL.

3.1 Feature refinement module

The module consists of the Spatial Reconstruction Unit (SRU), the Spatial Attention Module, the Channel Reconstruction Unit (CRU), and the Channel Attention Module. Figure 3 illustrates that the SRU employs separation and reconstruction operations. These operations divide the feature map into two parts: one with a substantial amount of information and the other with a minimal amount of information. The information content of different feature maps is evaluated using a scaling factor in the group normalization (GN) layer. Suppose that we have an intermediate feature map $X \in \mathbb{R}^{N\times C\times H\times W}$, where N represents the batch size, C represents the number of channels, and H and W represent the spatial height and width, respectively. Initially, we normalize the input feature X by subtracting the mean μ and dividing it by the standard deviation σ. This normalization process can be represented by Formula 1:

$$X_{out} = GN(X) = \gamma\frac{X-\mu}{\sqrt{\sigma^2+\varepsilon}} + \beta \tag{1}$$

where μ and σ represent the mean and standard deviation of X, ϵ is a small constant added for division stability, and γ and β denote trainable affine transformations.

Subsequently, the weight value of the reweighted feature map undergoes mapping through the sigmoid function and is controlled by a threshold. We assign a weight of 1 to values above the threshold to derive the information weight W_1, and a weight of 0 to those below to derive the non-information weight W_2. The entire process of determining W can be mathematically described as Formula 2:

$$W = Gate(Sigmoid(GN(X))) \tag{2}$$

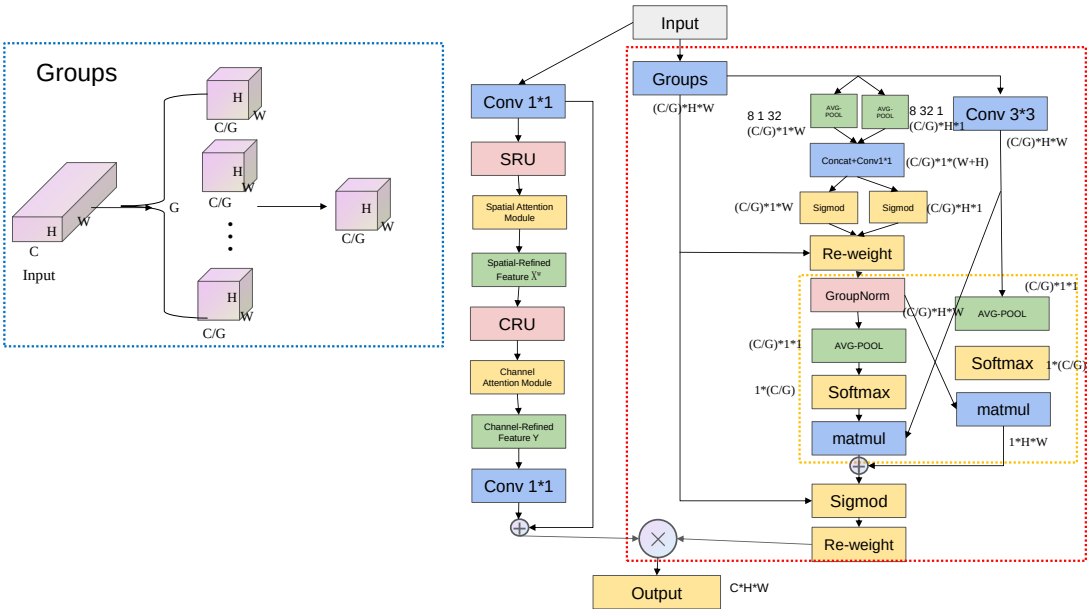

Figure 1: In the SCEMA model diagram, the blue dotted line represents feature grouping, the left branch of SCEMA is the feature refinement module, the right branch is the cross space learning module, and the yellow dotted line represents the aggregation output of matrix dot product operation

Finally, the input feature X is multiplied by W_1 and W_2 to obtain two weighted features: X_1^W containing more information and X_2^W containing less information. Features with less information are typically deemed redundant. To mitigate spatial redundancy, we introduce a cross-reconstruction operation. This operation combines the feature X_{11}^W containing rich information with the feature X_{12}^W containing less information, resulting in more informative features while conserving space. Subsequently, the two cross-reconstructed features are integrated to yield the spatially-thinned feature map X^W. The entire reconstruction process can be mathematically represented as shown in Formula 3:

$$\begin{cases} X_1^W = W_1 \otimes X \\ X_2^W = W_2 \otimes X \\ X_{11}^W \oplus X_{22}^W = X^{W1} \\ X_{21}^W \oplus X_{12}^W = X^{W2} \\ X^{W1} \cup X^{W2} = X^W \end{cases} \tag{3}$$

where $\otimes$ denotes element-wise multiplication, $\oplus$signifies element-wise summation, and $\cup$ indicates cascading.

As illustrated in Figure 4, the CRU employs split, transformation, and fusion strategies to enhance the spatial refinement feature map X^W by minimizing redundancy in the channel dimension.

Split: For a given spatial thinning feature $X^W \in \mathbb{R}^{c\times h\times w}$, the channels of X^W are initially divided into two parts: αC channels and $(1-\alpha)$ C channels, where $0 \leq \alpha \leq 1$ represents the segmentation ratio. Subsequently, a 1×1 convolution is applied to compress the feature map channels, enhancing computational efficiency. To regulate the feature channel and balance the computational expense of the CRU, a compression ratio r is introduced. Following segmentation and compression, the spatial thinning feature X^W is split into upper X_{up} and lower X_{low}.

Transform: X_{up} serves as the input to the upper transformation stage, functioning as a 'rich feature extractor'. employing efficient convolution operations, namely Depth-wise Separable Convolution (GWC) and Pointwise Convolution (PWC), we aim to substitute the costly standard k × k convolution with the objective of extracting high-level representative data while simultaneously minimizing computational expenses. The transformation process of the upper layer can be mathematically represented as Formula 4:

$$Y_1 = M^G X_{up} + M^{P1} X_{up} \tag{4}$$

where $M^G \in \mathbb{R}^{\frac{\alpha c}{gr}\times k\times k\times c}$ and $M^{P_1} \in \mathbb{R}^{\frac{\alpha c}{r}\times 1\times 1\times c}$ represent the learning weight matrices for GWC and PWC, respectively. In addition, $X_{up} \in \mathbb{R}^{\frac{\alpha c}{r}\times h\times w}$ and $Y_1 \in \mathbb{R}^{c\times h\times w}$ denote the input and output characteristic graphs of the upper layer. X_{low} serves as the input to the lower conversion stage. In this context, we employ a cost-effective 1×1 PWC operation to produce a feature map containing superficial hidden details, complementing the rich feature extractor. By reusing features from X_{low}, we obtain additional feature maps without incurring extra costs. Subsequently, we combine the newly generated features with the reused ones to construct the output of the lower level, denoted as Y_2, as illustrated in Formula 5:

$$Y_2 = M^{P_2} X_{low} \cup X_{low} \tag{5}$$

where $M^{P_2} \in \mathbb{R}^{\frac{(1-\alpha)c}{r}\times 1\times 1\times (1-\frac{1-\alpha}{r})c}$ is the learnable weight matrix of PWC, $\cup$ is the join operation, $M_{low} \in \mathbb{R}^{\frac{(1-\alpha)c}{r}\times h\times w}$ and $Y_2 \in \mathbb{R}^{c\times h\times w}$ are the lower layer input and output characteristic graphs respectively.

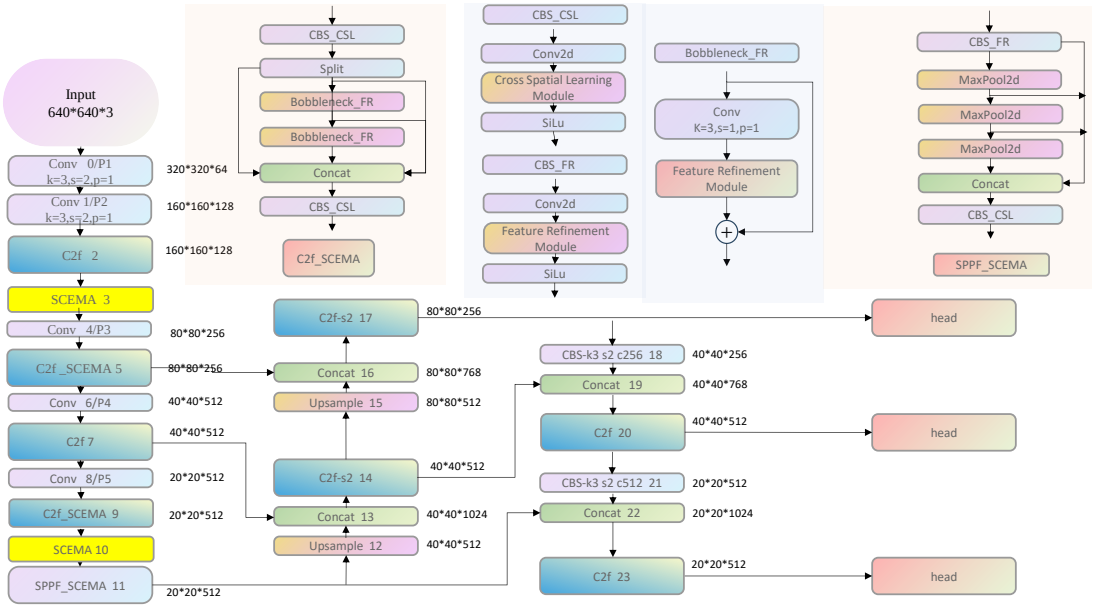

Figure 2: YOLO-SCEMA model structure diagram.Modify the original YOLOV8Backbone structure. The Feature Refinement Module and Cross Spatial Learning Module in SCEMA have been modified for C2f and SPPF

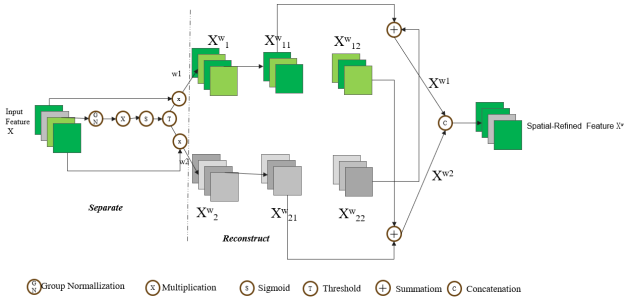

Figure 3: Structure diagram of spatial reconstruction unit (SRU)

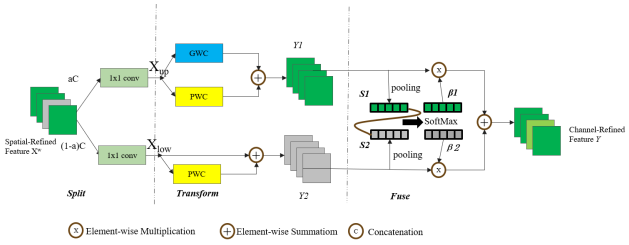

Figure 4: Channel reconstruction unit (CRU) structure diagram

Fuse:After the conversion is completed, the two types of features are no longer directly connected or added, but the output features Y_1 and Y_2 in the up and down conversion phase are adaptively combined, as shown in the Fuse section of Figure 4. We first apply GAPooling to collect global spatial information with channel statistics $S_m \in \mathbb{R}^{c\times1\times1}$. Next, we stack the upper and lower global channel descriptors $S_1 and S_2$ together, and use channel soft attention operations to generate feature importance vector $\beta_1, \beta_2 \in \mathbb{R}^c$. Finally, under the guidance of feature importance vectors$\beta_1 and \beta_2$, channel refinement feature Y can be obtained by merging upper layer feature Y_1 and lower layer feature Y_2 channel by channel,as shown in Formula 6:

$$Y = \beta_1 Y_1 + \beta_2 Y_2 \tag{6}$$

3.2 Cross spatial learning module

As depicted in the right branch of Figure 1, the CSL module employs a processing approach based on parallel sub-structures. Initially, it utilizes a 1x1 convolution as a common element, designating it as the 1x1 convolution branch. Subsequently, to rapidly integrate spatial information across multiple scales, it simultaneously positions the 3x3 convolution core alongside the 1x1 convolution branch. This configuration is referred to as the 3x3 convolution branch.

Feature Grouping: For instance, as illustrated in Figure 5, the red dotted line represents a division in a given input feature graph $X \in \mathbb{R}^{C\times H\times W}$. The CSL module segments X into G distinct sub feature groups, with each group focusing on learning distinct semantics. This approach of grouping features allows the model to effectively distribute and

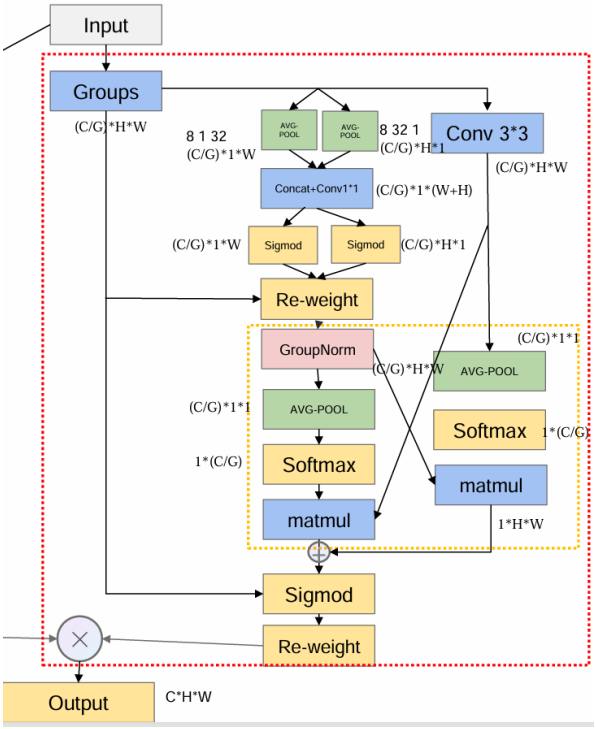

Figure 5: Structure diagram of Cross Space Learning module(CSL).Aggregating 1x1 and 3x3 branch outputs through matrix dot product operation through cross spatial information aggregation

handle GPU resources. Moreover, this grouping strategy enhances the learning of semantic regions while reducing noise interference.

Parallel Subnetwork: The CSL module uses three parallel paths to extract the attention weight descriptor of the grouping feature graph. As shown in Figure 5, the two paths on the left are 1x1 branches, and the third path on the right is 3x3 branches. One dimensional global average pooling operation is adopted in 1x1 branch to encode channel information in two spatial directions respectively.The two 1×1 branches on the left adopt a processing strategy analogous to Coordinate Attention (CA), whereby the encoded features are concatenated along the vertical dimension of the image. This design enables both branches to share a common 1×1 convolution without incurring dimensionality reduction. The output of the convolution is subsequently decomposed into two vectors, which are modeled using nonlinear Sigmoid functions to approximate a two-dimensional binomial distribution over the linear convolution. Finally, the attention maps from the two channel directions are aggregated within each group through element-wise multiplication, thereby facilitating diverse intersection points and channel interaction patterns between the parallel paths in the 1×1 branches. On the right, the 3×3 branch employs a 3×3 convolution to effectively capture multi-scale feature representations, thereby enriching the model's capacity for hierarchical feature extraction. In this way, CSL can not only encode cross channel information to adjust the importance of different channels, but also reserve accurate spatial structure information into channels.

Cross Space learning: The Cross-Spatial Learning (CSL) module is designed to enhance feature aggregation by introducing cross-spatial information fusion across multiple spatial dimensions. Specifically, the 1×1 branch leverages two-dimensional global average pooling to encode global spatial information, thereby effectively capturing holistic contextual cues and strengthening the modeling of long-range dependencies. In parallel, the 3×3 branch directly maps features into the corresponding dimensional shape, preserving fine-grained local structural patterns and complementing the global representation. During the feature interaction stage, the outputs of these two branches are integrated to generate the initial spatial attention map via a matrix dot-product operation, as illustrated by the yellow dotted line in Figure 5. This attention mechanism establishes connections across different spatial dimensions, enabling multi-scale feature fusion. To further refine the representation, two sigmoid functions are applied to the spatial attention weights, which aggregate the output feature maps within each group. This operation captures pixel-level pairing relationships and emphasizes global contextual dependencies among all pixels, thereby producing more robust and discriminative feature representations.

3.3 Model complexity analysis

The SCEMA we propose can be easily embedded into various existing well-designed CNNs architectures to reduce computing and storage costs. In the SCEMA module, all parameters are concentrated in the conversion phase. The parameter of standard convolution $Y = M^k X$ can be calculated as formula 7:

$$P_s = k^2 C_1 C_2 \tag{7}$$

where k is the size of the convolution kernel, C_1 and C_2 are the number of input and output feature channels. The parameters of the proposed SCEMA module can be expressed as Formula 8:

$$\begin{aligned} P_{SCEMA} &= M^G X_{up} + M^{P1} X_{up} + M^{P_2} X_{low} \cup X_{low} \\ &= 1 \times 1 \times \alpha C_1 \times \frac{\alpha C_1}{r} + k \times k \times \frac{\alpha C_1}{gr} \times \frac{C_2}{g} \times g \\ &+ 1 \times 1 \times \frac{\alpha C_1}{r} \times C_2 + (1-\alpha) C_1 \times \frac{(1-\alpha) C_1}{r} \\ &+ 1 \times 1 \times \frac{(1-\alpha) C_1}{r} \times (C_2 - \frac{(1-\alpha)}{r} C_1) \end{aligned} \tag{8}$$

where α represents the division ratio, r represents the compression ratio, g is the group size of GWC operation, C_1 and C_2 are the input and output characteristic channels respectively. In the experiment, the general parameter set is $\alpha = \frac{1}{2}$, r = 2,g = 2,k = 3,$C_1 = C_2 = C$, The number of parameters can be reduced by 5 times, of which $\frac{P_S}{P_{SCEMA}} \approx 5$. The improvement of model performance after improvement

Table 2: Deployment efficiency comparison of lightweight detectors on RTX 4070 Ti with different input sizes.

Models	Input Size	Latency (ms)	FPS	Allocated (MB)	Reserved (MB)
Mamba-YOLO	320×320	2.53	395.90	49.12	70.00
	640×640	2.52	396.47	54.92	90.00
	1280×1280	4.31	231.79	79.09	190.00
MSLAU-Net	320×320	2.38	421.01	11.67	34.00
	640×640	2.46	406.26	17.95	56.00
	1280×1280	3.34	299.28	41.46	136.00
YOLOv12	320×320	2.16	463.97	13.59	36.00
	640×640	2.14	467.21	19.88	58.00
	1280×1280	3.26	307.02	43.03	138.00
YOLOv8n	320×320	3.13	319.61	41.33	66.00
	640×640	3.07	325.31	47.61	86.00
	1280×1280	4.83	207.09	70.32	218.00
YOLO-SCEMA	320×320	1.99	501.46	13.54	38.00
	640×640	2.02	496.12	19.83	58.00
	1280×1280	3.12	320.02	44.30	138.00

will be described in detail in the subsequent comparative experiment in Chapter 4.

We have added deployment efficiency experiments on an RTX 4070 Ti GPU, including inference latency, FPS, and memory footprint. As shown in Table 2, YOLO-SCEMA demonstrates superior deployment efficiency by achieving the lowest latency (2.02 ms at 640×640), the highest FPS (496), and significantly reduced memory usage (19.83 MB) compared with YOLOv8n, while maintaining comparable efficiency to YOLOv12 and Mamba-YOLO. These results confirm that YOLO-SCEMA not only preserves competitive accuracy but also offers lower latency, higher throughput, and reduced memory footprint, making it well-suited for deployment in resource-constrained environments such as embedded systems, UAV monitoring, and edge computing.

4 Experimental

In this section, we will provide details of experiments and results to prove the performance of YOLO-SCEMA, including complex image structure experiments on the open-source dataset FYP, small object detection experiments in dense scenes on the VisDrone2019, and low light environment object detection experiments on the ExDark dataset. All experiments were conducted on hardware equipped with an RTX4070Ti GPU and an Intel (R) i5-9300H CPU @ 2.40GHz. The software environment: Python 3.10.8, PyTorch 1.8.1 and CUDA 11.8. The key training hyperparameters configured as follows: the input image size (imgsz) was set to 640×640 pixels; the total training epochs were 300 (ExDark), 150 (VisDrone2019), 300 (FYP); the number of worker threads (workers) for data loading was 8 (ExDark), 4 (VisDrone2019), 8 (FYP); the image scaling factor (scale) was 0.5 ; Mosaic data augmentation was fully enabled with a value of 1.0; and the probability of Copy-Paste data augmentation was set to 0.1.

4.1 Object detection in weak-light scenes

In order to verify the detection performance of the proposed SCEMA module in low-light scenes, we chose to verify it on the ExDark dataset. This dataset was specifically created for low-light object detection, containing images under 12 different lighting conditions ranging from extremely low light to twilight; it includes 5,891 training images, 1,472 test images, and 12 object categories. From the results in Figure 6 and Table 3, we can see that MCA, GAM and CBAM can improve the baseline performance of object detection. The proposed SCEMA module consistently outperforms networks based on MCA, GAM, and CBAM in terms of mAP (50) and mAP (95). It should be noted that GAM and CBAM improve YOLOV8N performance by 4.32% and 6.51%, respectively, and are higher than MCA, but at the cost of more parameters and calculations. Specifically, SCEMA reduces parameters by 36.9% and computational cost by 8.6% compared with the baseline method, achieves a 7.37% improvement in mAP (50) compared with YOLOv8n, and is 4.37%, 3.05%, and 0.86% higher than MCA, GAM, and CBAM, respectively. Compared with the latest YOLOV12N, SCEMA has increased mAP (50) by 0.5% while reducing parameters by 0.66M. On ExDark, MSLAU-Net achieves 76.86% mAP@50, Mamba-YOLO-T achieves 76.21%, and SCEMA achieves 76.32%. SCEMA slightly outperforms Mamba-YOLO-T (+0.11%) but is marginally lower than MSLAU-Net (-0.54%).

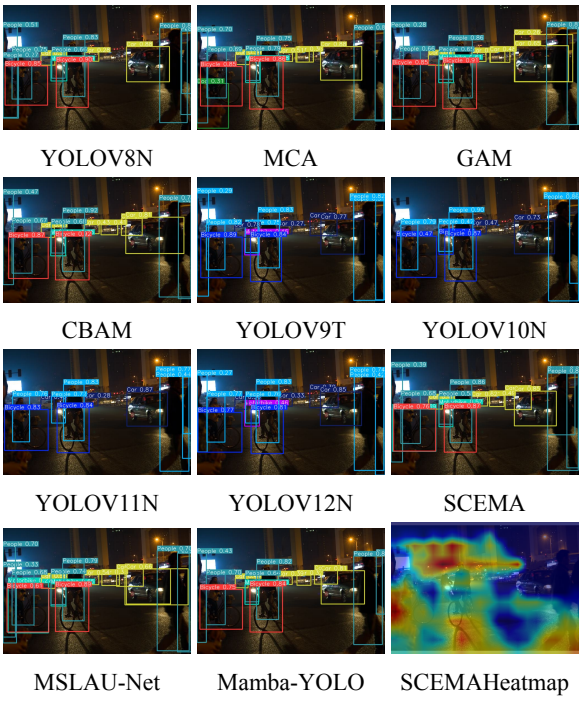

Figure 6: The visualization results of the ExDark dataset.In the dark environment, the baseline YOLOV8N, YOLOV10N and YOLOV11N have the problem of car type missing detection. The SCEMA we proposed can accurately identify the cars blocked in the distance.

4.2 Object detection in dense scenes

In order to verify the generalization capability of the proposed SCEMA module, we evaluated it on the VisDrone2019 dataset. The dataset includes varying perspectives, occlusions, vehicle sizes, and lighting conditions, which pose significant challenges to object detection in dense scenes, and put forward higher requirements for the stability and adaptability of the algorithm. The VisDrone2019 dataset contains 10,209 still images: 6,471 for training, 548 for validation, and 3,190 for testing. From the results in Figure 7 and Table 4, The proposed SCEMA module consistently outperforms networks based on PConv, ECA, and CBAM in terms of mAP (50) and mAP (95). Compared with the latest YOLOv12n, SCEMA improves mAP (50) by 0.33%, while reducing parameter count by 25.8% and computational complexity by 14.9%. CBAM improves YOLOv8n performance by 0.10%, surpassing PConv and ECA, but at the cost of additional parameters and computation. For the PConv module, it achieves almost the same performance as baseline, and is 0.03% higher than YOLOV8N in terms of mAP (50), while ECA requires more parameters and computation than PConv (2.65M vs. 1.86M and 8.4G vs. 7.1G). Specifically, SCEMA has 1.50M fewer parameters than the baseline method, 3.24% more than YOLOV8N in terms of mAP (50), and 3.21% more than PConv in terms of mAP (50) with slightly higher parameters and computation. On

Table 3: On the ExDark dataset, SCEMA reduces 36.9% of the parameters and 8.6% of the computation compared with the baseline YOLOv8n, and increases mAP (50) by 7.37%. Compared with the latest YOLOv12n, SCEMA improves mAP (50) by 0.5% while reducing parameters by 0.66M. For MSLAU-Net, only parameter counts (21.9M) were reported in the original paper, and FLOPs were not provided, so computation is marked as "N/A" in the comparison.

Models	mAP@50	mAP@95	Parameters (M)	GFLOPs
YOLOv8n	68.95	43.03	3.01	8.1
MCA	71.95	44.24	1.87	7.1
GAM	73.27	46.68	3.94	9.5
CBAM	75.46	48.53	3.07	8.2
YOLOv9t	74.62	47.96	1.97	7.6
YOLOv10n	75.35	48.26	2.69	8.2
YOLOv11n	75.52	48.61	2.58	6.3
YOLOv12n	75.82	48.94	2.56	6.3
MSLAU-Net	76.86	49.52	21.90	N/A
Mamba-YOLO-T	76.21	49.01	5.80	13.2
SCEMA	76.32	49.26	1.90	7.4

VisDrone2019, MSLAU-Net achieves 34.42% mAP@50, Mamba-YOLO-T achieves 35.47%, and SCEMA achieves 34.86%. SCEMA is higher than MSLAU-Net (+0.44%) but lower than Mamba-YOLO-T (-0.61%). These results demonstrate that SCEMA is an effective object detection module, further validating the proposed method.

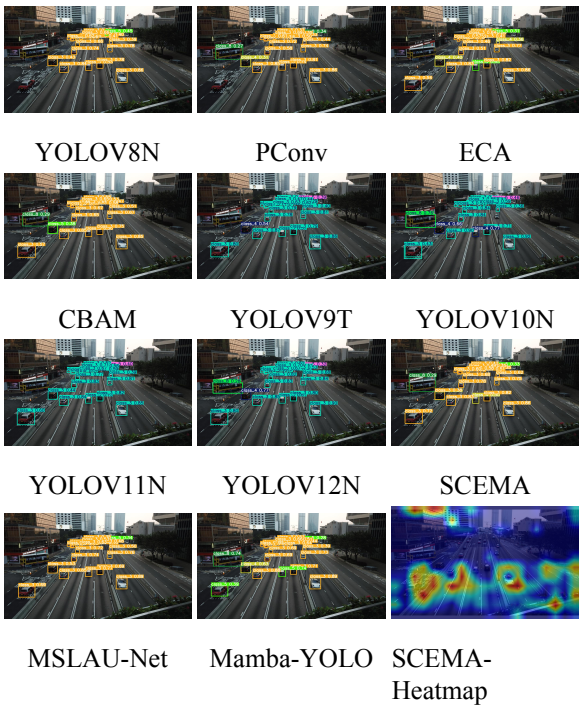

Figure 7: The visualization results of the Visdrone2019 dataset,The YOLOV8N, PConv, ECA, YOLOV9T, YOLOV11N and MSLAU-Net models missed detecting the bus or car on the left side of the image.

Table 4: On the VisDrone2019 dataset, Although SCEMA is slightly higher than PConv in terms of parameters and computation, mAP(50) has increased by 3.21%. Compared with the latest YOLOV12, the mAP (50) has increased by 0.33%.

Models	mAP@50	mAP@95	Parameters (M)	GFLOPs
YOLOV8N	31.62	18.16	3.01	8.1
PConv	31.65	18.20	1.86	7.1
ECA	32.27	18.41	2.65	8.4
CBAM	32.62	18.63	3.07	8.2
YOLOV9T	34.46	20.03	1.97	7.6
YOLOV10N	34.18	19.89	2.69	8.2
YOLOV11N	34.42	20.12	2.58	6.3
YOLOV12N	34.53	20.24	2.56	6.3
MSLAU-Net	34.42	20.16	21.90	N/A
Mamba-YOLO-T	35.47	20.83	5.80	13.2
SCEMA	34.86	20.46	1.90	7.4

Table 5: On the FYP dataset,Performance comparison of different models based on mAP, parameters, and GFLOPs.Compared with other latest YOLO series, SCEMA has improved mAP (50) by 0.93%, 0.84%, 0.67%, and 0.56% respectively, while having fewer parameters and computational complexity.

Models	mAP@50	mAP@95	Parameters (M)	GFLOPs
YOLOV8N	67.89	51.53	3.01	8.1
EMA	68.94	53.04	3.02	8.2
MCA	68.41	53.36	1.87	7.1
CBAM	69.12	53.12	3.07	8.2
YOLOV9T	68.46	53.41	1.97	7.6
YOLOV10N	68.55	53.49	2.69	8.2
YOLOV11N	68.72	53.62	2.58	6.3
YOLOV12N	68.83	53.71	2.56	6.3
MSLAU-Net	68.52	53.45	21.90	N/A
Mamba-YOLO-T	68.67	53.57	5.80	13.2
SCEMA	69.39	53.89	1.90	7.4

4.3 Complex image structure object detection

To demonstrate the advantages of the proposed method for complex image structure detection, we conducted experiments on the open-source FYP dataset. The dataset contains complex scenes such as vehicle occlusion, dense traffic, and nighttime conditions, with target objects that are often overlapping, occluded, or truncated. The experimental results are shown in Figure 8 and Table 5. On FYP, MSLAU-Net achieves 68.52% mAP@50, Mamba-YOLO-T achieves 68.67%, and SCEMA achieves 69.39%. SCEMA outperforms both MSLAU-Net (+0.87%) and Mamba-YOLO-T (+0.72%). Compared with recent YOLO series models, SCEMA improves mAP (50) by 0.93%, 0.84%, 0.67%, and 0.56%, while requiring fewer parameters and lower computational complexity. Compared with the benchmark YOLOv8n and MCA models, EMA and CBAM achieve higher accuracy, but SCEMA remains competitive. SCEMA is slightly better than CBAM in mAP (50) and 0.45% higher than EMA. In addition, the SCEMA model has 1.90M parameters, only 0.03M more than the MCA model. Although the FLOPs of the SCEMA are 7.4G, which is only 0.3G larger than the MCA model, the SCEMA has achieved 69.39% mAP (50) and 53.89% mAP (95) on all five categories, which is significantly higher than YOLOV8N benchmark and other methods.

4.4 Ablation studies

In this section, ablation studies were conducted to evaluate the relative contributions of the individual components within the proposed SCEMA module. The results are summarized in Table 6. Compared with the YOLOv8n baseline, all modified variants achieved performance improvements across the three datasets, yet none surpassed the complete SCEMA configuration. Specifically, decomposing SCConv into its two sub-units, SRU and CRU, revealed that each contributes positively to detection accuracy, but their joint integration yields superior results. Removing SRU led to performance drops of 1.05%, 2.79%, and 3.67% on FYP, VisDrone2019, and ExDark, respectively, while eliminating CRU resulted in decreases of 1.18%, 2.91%, and 4.21%. Similarly, removing the cross-spatial learning (CSL) module reduced detection performance by 0.97%, 2.7%, and 2.81% across the same datasets. These findings demonstrate that SRU and CRU are complementary, and their combination within SCConv is essential for maximizing detection accuracy.

5 Discussion

5.1 Comparison with related works

In this subsection, we explicitly compare SCEMA with representative lightweight detectors and multi-attention frameworks, including the YOLO family, CBAM, ECA, EMA, MSLAU-Net, and Mamba-YOLO-T. While these approaches have demonstrated improvements in feature representation, they often introduce substantial parameter overhead and computational complexity, limiting their applicability in resource-constrained environments. Our experimental results highlight that YOLO-SCEMA achieves superior performance compared with the YOLOv8n baseline, with mAP@50 improvements of 7.37% on ExDark, 3.24% on VisDrone2019, and 1.5% on FYP, while simultaneously reducing parameters by 36.9% and computational load by 8.6%. These results indicate that SCEMA effectively balances accuracy and efficiency, addressing the limitations observed in prior methods. To further strengthen the comparison with state-of-the-art lightweight detectors beyond the YOLO family, we incorporated MSLAU-Net and Mamba-YOLO-T into our evaluation. The results demonstrate that, on the ExDark dataset, MSLAU-Net achieves an mAP@50 that is 0.54% higher than SCEMA, but requires 21.9M parameters compared to SCEMA's

Table 6: Results of ablation experiments on ExDark, VisDrone2019, and FYP datasets. Models A–E represent variant models with different modules. The data on the right side of the table represents mAP (50) scores.

Models	SRU	CRU	CSL	CBAM	FYP-mAP(50)	VisDrone2019-mAP(50)	ExDark-mAP(50)
YOLOv8n					67.89	31.62	68.95
A	✓		✓	✓	68.21	31.95	72.11
B		✓	✓	✓	68.34	32.07	72.65
C	✓	✓		✓	68.42	32.16	73.51
D	✓	✓	✓		68.83	32.43	74.62
E	✓	✓	✓	✓	69.39	34.86	76.32

1.9M. On the VisDrone2019 dataset, Mamba-YOLO-T achieves an mAP@50 that is 0.61% higher than SCEMA, but requires 5.8M parameters and 13.2 GFLOPs, compared with SCEMA's 7.4 GFLOPs. Thus, although these models achieve marginally higher accuracy, their substantially larger parameter counts and computational demands highlight SCEMA's advantage in delivering competitive accuracy with markedly superior efficiency.

Ovrerall, SCEMA distinguishes itself from existing multi-attention frameworks and lightweight detectors by achieving a favorable trade-off between detection accuracy and computational efficiency. Its parallel fusion strategy, which simultaneously reduces redundancy and enhances multi-scale feature integration, provides a novel and practical solution for the detection of complex objects in the scene.

5.2 Design choices and trade-offs

To evaluate the rationale behind SCEMA's design, we conducted ablation experiments on ExDark, VisDrone2019, and FYP datasets, as shown in Table 6. Models A–E represent different combinations of SRU, CRU, CSL, and CBAM, enabling analysis of the individual contributions of each module. The results demonstrate that each component plays a distinct role in either efficiency optimization or accuracy enhancement, while the final parallel fusion strategy achieves the best overall balance. The Spatial Reconstruction Unit (SRU) and Channel Reconstruction Unit (CRU) primarily contribute to efficiency.

For instance, Model C(SRU+CRU+CBAM) achieves 73.51% mAP@50 on ExDark, compared with 68.95% for YOLOv8n, while maintaining relatively low computational overhead. SRU reduces spatial redundancy by separating and reconstructing features, whereas CRU leverages grouped and pointwise convolutions to compress channels and improve representational capacity. Together, they enable lightweight processing without sacrificing accuracy. The Cross Spatial Learning (CSL) module plays a critical role in accuracy improvement. Model D (SRU+CRU+CSL) achieves 74.62% mAP@50 on ExDark and 32.43% on VisDrone2019, outperforming Model C, which lacks CSL. This indicates that CSL's feature grouping and parallel substructures enhance multi-scale integration, particularly in complex image scenarios.The CBAM module further strengthens feature selectivity. Model B (CRU+CSL+CBAM) achieves 72.65% mAP@50 on ExDark and 32.07% on VisDrone2019, showing that CBAM improves attention to key regions, especially for small-object detection. Although CBAM introduces additional parameters, its contribution to accuracy is evident.

Finally, the full SCEMA configuration (Model E) integrates all four modules and achieves the highest performance across datasets: 76.32% mAP@50 on ExDark, 34.86% on VisDrone2019, and 69.39% on FYP. These results confirm that the parallel multi-branch fusion strategy—combining SRU/CRU's redundancy reduction with CSL/CBAM's accuracy gains provides dual optimization in efficiency and accuracy. Unlike traditional single-path or sequential attention mechanisms, SCEMA's parallel design avoids information loss and improves multi-scale representation, which explains its superior performance in ablation studies and highlights its novelty in lightweight object detection.

5.3 Significance of results, applications, and future work

The experimental results demonstrate that SCEMA achieves a favorable balance between accuracy and efficiency, making it highly suitable for deployment in real-world scenarios. On the ExDark, VisDrone2019, and FYP datasets, SCEMA consistently outperforms the YOLOv8n baseline, with mAP@50 improvements of 7.37%, 3.24%, and 1.5%, respectively, while reducing parameters by 36.9% and computational cost by 8.6%. These gains are particularly meaningful in practical applications: ExDark highlights SCEMA's robustness in low-light environments, VisDrone2019 validates its effectiveness in dense UAV scenes with small objects, and FYP demonstrates its adaptability to complex image structures. The ability to maintain lightweight design while achieving accuracy comparable to or exceeding state-of-the-art models underscores SCEMA's potential for resource-constrained deployments such as embedded systems, UAV monitoring, and edge computing.

Despite these promising results, the current study is focused on YOLO-based object detection tasks. Future work will extend SCEMA to broader tasks, including object tracking and image classification, to further validate its generalization capability. Additionally, optimizing cross-scale fusion strategies and exploring more efficient attention mechanisms will be pursued to enhance both accu-

Figure 8: The visualization results of the FYP dataset. In dense scenes, our proposed SCEMA can accurately identify motorcycles obstructed by buses and cars.

racy and efficiency. These directions aim to strengthen SCEMA's applicability across diverse tasks and environments, ensuring its continued relevance in lightweight computer vision research.

6 Conclusion

This article proposes a lightweight and efficient multi-scale attention fusion module (SCEMA), which adopts a multi-branch parallel processing strategy. The left branch reduces the common spatial and channel redundancy in standard convolution through a feature refinement module. This reduction in redundancy not only lowers computational costs and model storage requirements, but also improves model performance. The right branch cleverly combines features of different scales through feature grouping and cross spatial learning, thereby enhancing the model's ability to understand multi-scale image content and improving performance in handling complex image structures. Extensive experiments conducted on three different datasets using this module and various methods have shown that YOLO-SCEMA exhibits excellent detection accuracy while reducing parameters and computational load. In future work, we will combine SCEMA with other models for tasks such as object tracking and object classification.

Acknowledgments

This research was funded by the Ph.D. Scientific Research Project of Nanyang Normal University (No.2022ZX022, 2026PY018) and the Natural Science Foundation Project of Henan Province(262300421805).

Code, data, and materials availability

Part of the code and data related to the paper have been uploaded to https://github.com/yz563/YOLO-SCEMA. The complete code and data will be made public after the publication of the paper.

Conflicts of interest statement

The authors declare that there are no financial interests, commercial affiliations, or other potential conflicts of interest that could have influenced the objectivity of this research or the writing of this paper.

References

[1] Chen, L., Zhang, H., Xiao, J., Nie, L., Shao, J., Liu, W., & Chua, T.-S. (2017). Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 6298–6306). doi:https://doi.org/10.31449/inf.v50i14.12887https://doi.org/10.1109/CVPR.2017.667.

[2] He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask r-cnn. In *2017 IEEE International Conference on Computer Vision (ICCV)* (pp. 2980–2988). doi:https://doi.org/10.31449/inf.v50i14.12887https://doi.org/10.1109/ICCV.2017.322.

[3] Hou, Q., Zhou, D., & Feng, J. (2021). Coordinate attention for efficient mobile network design. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 13708–13717). doi:https://doi.org/10.31449/inf.v50i14.12887https://doi.org/10.1109/CVPR46437.2021.01350.

[4] Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 7132–7141). doi:https://doi.org/10.31449/inf.v50i14.12887https://doi.org/10.1109/CVPR.2018.00745.

[5] Jiang, P., Ergu, D., Liu, F., Cai, Y., & Ma, B. (2022). A review of yolo algorithm developments. *Procedia Computer Science*, *199*, 1066–1073. doi:https://doi.org/10.31449/inf.v50i14.12887https://doi.org/10.1016/j.procs.2022.01.135. Publisher: Elsevier.

[6] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, *60*, 84–90. doi:https://doi.org/10.31449/inf.v50i14.12887https://doi.org/10.1145/3065386.

[7] Lan, L., Li, Y., Liu, X., Zhou, J., Zhang, J., Huang, N., & Zhang, Y. (2025). Mslau-net: A hybird cnn-transformer network for medical image segmentation. *arXiv preprint arXiv:2505.18823*, . doi:https://doi.org/10.31449/inf.v50i14.12887https://arxiv.org/abs/2505.18823.

[8] Lei, M., Wu, H., Lv, X., & Wang, X. (2025). Condseg: A general medical image segmentation framework via contrast-driven feature enhancement. In *Proceedings of the AAAI Conference on Artificial Intelligence* (pp. 4571–4579). volume 39. doi:https://doi.org/10.31449/inf.v50i14.12887https://doi.org/10.1609/aaai.v39i5.32482.

[9] Li, J., Wen, Y., & He, L. (2023). Scconv: Spatial and channel reconstruction convolution for feature redundancy. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 6153–6162). doi:https://doi.org/10.31449/inf.v50i14.12887https://doi.org/10.1109/CVPR52729.2023.00596.

[10] Li, X., Hu, X., & Yang, J. (2019). Spatial group-wise enhance: Improving semantic feature learning in convolutional networks. *arXiv preprint arXiv:1905.09646*, . doi:https://doi.org/10.31449/inf.v50i14.12887https://doi.org/10.48550/arXiv.1905.09646.

[11] Liu, Y., Shao, Z., & Hoffmann, N. (2021). Global attention mechanism: Retain information to enhance channel-spatial interactions. *arXiv preprint arXiv:2112.05561*, . doi:https://doi.org/10.31449/inf.v50i14.12887https://doi.org/10.48550/arXiv.2112.05561.

[12] Lu, X., Suganuma, M., & Okatani, T. (2024). Sbcformer: lightweight network capable of full-size imagenet classification at 1 fps on single board computers. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* (pp. 1112–1122). doi:https://doi.org/10.31449/inf.v50i14.12887https://doi.org/10.1109/WACV57701.2024.00116.

[13] Ma, N., Zhang, X., Zheng, H.-T., & Sun, J. (2018). Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Computer Vision – ECCV 2018* (pp. 122–138). Springer International Publishing. doi:https://doi.org/10.31449/inf.v50i14.12887https://doi.org/10.1007/978-3-030-01264-9_8.

[14] Ouyang, D., He, S., Zhang, G., Luo, M., Guo, H., Zhan, J., & Huang, Z. (2023). Efficient multi-scale attention module with cross-spatial learning. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1–5). IEEE. doi:https://doi.org/10.31449/inf.v50i14.12887https://doi.org/10.1109/ICASSP49357.2023.10096516.

[15] Park, J., Woo, S., Lee, J.-Y., & Kweon, I. S. (2018). Bam: Bottleneck attention module. *arXiv preprint arXiv:1807.06514*, . doi:https://doi.org/10.31449/inf.v50i14.12887https://doi.org/10.48550/arXiv.1807.06514.

[16] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., & others (2015). Imagenet large scale visual recognition challenge. *International journal of computer vision*, *115*, 211–252. doi:https://doi.org/10.31449/inf.v50i14.12887https://doi.org/10.1007/s11263-015-0816-y.

[17] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 4510–4520). doi:https://doi.org/10.31449/inf.v50i14.12887https://doi.org/10.1109/CVPR.2018.00474.

[18] Sohan, M., Sai Ram, T., & Rami Reddy, C. V. (2024). A review on yolov8 and its advancements. In *Data Intelligence and Cognitive Informatics* (pp. 529–545). Springer. doi:https://doi.org/10.31449/inf.v50i14.12887https://doi.org/10.1007/978-981-99-7962-2_39.

[19] Su, J.-N., Gan, M., Chen, G.-Y., Guo, W., & Chen, C. P. (2024). High-similarity-pass attention for single image super-resolution. *IEEE Transactions on Image Processing*, *33*, 610–624. doi:https://doi.org/10.31449/inf.v50i14.12887https://doi.org/10.1109/TIP.2023.3348293. Publisher: IEEE.

[20] Su, L., Ma, X., Zhu, X., Niu, C., Lei, Z., & Zhou, J.-Z. (2025). Can we get rid of handcrafted feature extractors? sparsevit: Nonsemantics-centered, parameter-efficient image manipulation localization through spare-coding transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence* (pp. 7024–7032). volume 39.

doi:https://doi.org/10.31449/inf.v50i14.12887https://doi.org/10.1609/aaai.v39i7.32754 issue: 7.

[21] Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., & Hu, Q. (2020). ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 11531–11539). doi:https://doi.org/10.31449/inf.v50i14.12887https://doi.org/10.1109/CVPR42600.2020.01155.

[22] Wang, Z., Li, C., Xu, H., Zhu, X., & Li, H. (2024). Mamba yolo: A simple baseline for object detection with state space model. In *Proceedings of the AAAI Conference on Artificial Intelligence* (pp. 8205–8213). volume 39. doi:https://doi.org/10.31449/inf.v50i14.12887https://doi.org/10.1609/aaai.v39i8.32885.

[23] Woo, S., Park, J., Lee, J.-Y., & Kweon, I. S. (2018). CBAM: Convolutional Block Attention Module. In *Computer Vision – ECCV 2018* (pp. 3–19). Springer International Publishing. doi:https://doi.org/10.31449/inf.v50i14.12887https://doi.org/10.1007/978-3-030-01234-2_1.

[24] Yan, Q., Feng, Y., Zhang, C., Pang, G., Shi, K., Wu, P., Dong, W., Sun, J., & Zhang, Y. (2025). Hvi: A new color space for low-light image enhancement. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (pp. 5678–5687). doi:https://doi.org/10.31449/inf.v50i14.12887https://doi.org/10.1109/CVPR52734.2025.00533.

[25] Yang, J., Liu, S., Wu, J., Su, X., Hai, N., & Huang, X. (2025). Pinwheel-shaped convolution and scale-based dynamic loss for infrared small target detection. In *Proceedings of the AAAI Conference on Artificial Intelligence* (pp. 9202–9210). volume 39. doi:https://doi.org/10.31449/inf.v50i14.12887https://doi.org/10.1609/aaai.v39i9.32996 issue: 9.

[26] Yu, Y., Zhang, Y., Cheng, Z., Song, Z., & Tang, C. (2023). MCA: Multidimensional collaborative attention in deep convolutional neural networks for image recognition. *Engineering Applications of Artificial Intelligence*, *126*, 107079. doi:https://doi.org/10.31449/inf.v50i14.12887https://doi.org/10.1016/j.engappai.2023.107079. Publisher: Elsevier.

[27] Zhang, Q.-L., & Yang, Y.-B. (2021). Sa-net: Shuffle attention for deep convolutional neural networks. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 2235–2239). IEEE. doi:https://doi.org/10.31449/inf.v50i14.12887https://doi.org/10.1109/ICASSP39728.2021.9414568.