Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

RFSC-net: Re-parameterization forward semantic compensation network in low-light environments

Wenhao Zhang [a], Huiying Xu [a,*], Xinzhong Zhu [a,b,c], Yunzhong Si [a], Yao Dong [a], Xiao Huang [d], Hongbo Li [c]

[a] School of Computer Science and Technology, Zhejiang Normal University, Jinhua, Zhejiang 321004, China
[b] Research Institute of Hangzhou Artificial Intelligence, Zhejiang Normal University, Hangzhou, Zhejiang 311231, China
[c] Beijing Geekplus Technology Co., Ltd, Beijing 100101, China
[d] College of Education, Zhejiang Normal University, Jinhua 321004, China

ARTICLE INFO

Keywords:
Low-light detection
Feature fusion
Re-parameterization
Forward semantic compensation

ABSTRACT

Although detectors currently perform well in well-light conditions, their accuracy decreases due to insufficient object information. In addressing this issue, we propose the Re-parameterization Forward Semantic Compensation Network (RFSC-Net). We propose the Reparameterization Residual Efficient Layer Aggregation Networks (RSELAN) for feature extraction, which integrates the concepts of re-parameterization and the Efficient Layer Aggregation Networks (ELAN). While focusing on the fusion of feature maps of the same dimension, it also incorporates upward fusion of lower-level feature maps, enhancing the detailed texture information in higher-level features. Our proposed Forward Semantic Compensation Feature Fusion (FSCFF) network reduces interference from high-level to low-level semantic information, retaining finer details to improve detection accuracy in low-light conditions. Experiments on the low-light ExDark and DarkFace datasets show that RFSC-Net improves mAP by 2% on ExDark and 0.5% on DarkFace over the YOLOv8n baseline, without an increase in parameter counts. Additionally, AP50 is enhanced by 2.1% on ExDark and 1.1% on DarkFace, with a mere 3.7 ms detection latency on ExDark.

1. Introduction

In recent years, object detection has made remarkable progress and has been widely applied in various fields. However, while most current detectors [1–5] achieve satisfactory detection performance under normal lighting conditions with abundant feature information, they exhibit poor detection accuracy in low-light or dark environments where feature information is insufficient. Low-light scenarios are an inevitable occurrence in practical situations, such as at night, on overcast or rainy days, and under shadow conditions. Therefore, it is particularly important to enhance research on object detection in low-light environments.

In low-light conditions, the noise of the camera sensor increases significantly, resulting in a substantial decrease in image contrast. The gray level difference becomes smaller between the object and the background and affects the detection accuracy. In the process of forward propagation of images, multiple down sampling operations are performed to obtain high-level semantic information. However, some information of image is lost during feature extraction. Due to noise and insufficient feature information, the detected objects are often mistaken for the background and considered invalid so that not learned. However, this information is crucial for the precise identification and localization of the objects. Therefore, we consider preserving feature information: if we minimize the discard of shallow feature information during forward propagation, will it significantly improve the accuracy of the detector? Following this guiding philosophy, we design the RFSC-Net forward semantic compensation network. The primary design concept is to continuously compensate for the lost shallow information in the higher-level feature maps during the forward propagation process of the feature extraction network. To mitigate the influence of high-level semantic information on low level semantic information, we reduce the fusion of high-level feature maps into low-level feature maps in the feature network. Additionally, to enhance the representational capability of feature maps while maintaining inference efficiency, we adopt multi-branch structure and re-parameterization techniques.

[*] Corresponding author.
E-mail addresses: zwh2012918201@zjnu.edu.cn (W. Zhang), xhy@zjnu.edu.cn (H. Xu), zxz@zjnu.edu.cn (X. Zhu), siyuzhong@zjnu.edu.cn (Y. Si), dongyao@zjnu.edu.cn (Y. Dong), Huangxiao@zjnu.edu.cn (X. Huang), Jason.li@geekplus.com (H. Li).

https://doi.org/10.1016/j.imavis.2024.105271
Received 31 May 2024; Received in revised form 1 September 2024; Accepted 6 September 2024
Available online 7 September 2024

Currently, the main stream real-time detection algorithms primarily adopt single-stage detectors, such as the YOLO [1–4,6–9] series, which are based on Convolutional Neural Networks (CNN). Simultaneously, detectors that integrate Transformers [5] and Mamba [10] are gradually being applied to real-time detection. These detection algorithms typically start by extracting features through a lightweight backbone [11–18], and then the resulting multi-scale feature maps are fed into a feature fusion network [19–22] for multi-scale feature fusion. A substantial amount of research indicates that the combined use of backbones and feature fusion networks can significantly enhance detection accuracy. The backbone initially extracts features from the original image, obtaining feature maps of different scales at various stages. These multi-scale feature maps are then fed into the feature fusion network for integrating features across different scales. The use of the feature fusion network markedly increases the detection precision, which is important for accurately obtaining the final object semantic information.

In the backbones of the aforementioned detectors, downsampling methods that increase the number of feature map channels are used to obtain features at different scales. After feature extraction is completed at each stage, the process moves vertically upwards to the next stage for further feature extraction. As the depth of the network increases, the network is able to capture higher-level semantic information. Under normal circumstances, semantic information extracted from low-level feature maps aids in the detection of small objects, information extracted from high-level feature maps is focused on the detection of large objects. Although this feature extraction method can capture the vast majority of valid information, it may lose some detailed information. Because, in the process of downsampling, higher-level semantic information is obtained, and the most important information from the previous stage relative to the current stage is retained. To address this issue, we propose RSELAN to extract image features, which utilizes re-parameterization convolution to enhance the feature extraction capability of the network. It performs weighted fusion between adjacent scale feature maps to increase the effect of integration among them. Thus, more beneficial shallow semantic information is retained during the process of feature extraction.

In the existing feature extraction networks, the Feature Pyramid Network (FPN) [19] is the most widely used. FPN integrates adjacent feature maps through top-down upsampling and lateral connections, combining the characteristics of high-level feature maps with rich semantic information and low-level feature maps with detailed spatial information, resulting in new feature maps that are enriched with more comprehensive semantic information. In order to give high-level feature maps more precise positional information, the Path Aggregation Network (PANet) [21] is developed based on the FPN. PANet has shown excellent performance across various datasets and is widely used in most detectors today. Adaptive Spatial Feature Fusion (ASFF) [22] filters the information from different feature maps, alleviating semantic conflicts caused by scale inconsistencies. Although ASFF can address the problem of conflicting information at the same location, when aligning high-level feature maps with low-level feature maps, the richer semantic information in the high-level feature maps and the larger size of the objects make them easier to detect, leading to cases where the low-level feature maps consider the objects in the high-level maps as primary information. Moreover, because the objects in the high-level feature maps are large, there is also a situation where the objects in the high-level feature maps are considered as background by the low-level feature maps. In response to these issues, we propose a Forward Space Adaptive Compensation Feature Fusion Network (FSCFF), based on the integration of PA-FPN and ASFF. This method fully fuses high-level feature maps with low-level ones, conducting forward feature fusion for each layer. It only allows low-level feature maps to adaptively spatially merge with high-level feature maps, preventing high-level features from dominating in the low-level feature maps.

To balance accuracy and real-time performance, we select YOLOv8n as the baseline model. We make modifications to its backbone and feature extraction networks to validate the effectiveness of the forward semantic compensation concept. For this purpose, we conduct extensive experiments. The experiments in this paper were conducted on the low-light dataset ExDark [23], where we used the detection results of YOLOv8n [4] as the baseline, and conducted extensive comparative and ablation studies on the backbone and feature fusion network of RFSC-Net. RFSC-Net has a 2% higher mAP than the baseline model on the ExDark dataset, and its AP50 is 2.1% higher than that of the baseline. Additionally, the inference latency per image is only 3.7 ms, while the number of parameters remains unchanged compared to the baseline model. To verify the generalizability of our model, we also conduct experiments on the COCO [24] and VOC [25] datasets, where the final mAP surpasses the baseline model by 1.9% and 2.5% respectively. On the small object dataset VisDrone [26], the mAP and AP50 exceed the baseline by 1.6% and 2.3% respectively. On the small object detection dataset DarkFace [27,28] under dark conditions, the mAP and AP50 are 0.5% and 1.1% higher than the baseline model, respectively.

We summarize the work of this paper as follows:

- To address the issue of losing detailed information from shallow feature maps when backbone extracts features upward, we utilize re-parameterization convolution to enrich feature extraction. Additionally, we adopt a multi-branch structure to enhance the representation ability of the network, designing RSELAN for image feature extraction.
- In order to solve the problem that the importance of correct object semantic information weakens due to the fusion of high-level feature maps and low-level feature maps, we propose FSCFF, which not only solves the aforementioned problem but also takes into account the semantic loss and full integration of high-level and low-level feature maps during forward propagation.
- Extensive experimental results on the ExDark dataset have shown that RFSC-Net surpasses many of the most advanced real-time detection networks under almost same parameter counts and computation conditions. Additionally, it also performs well on the COCO, VOC, VisDrone, and DarkFace datasets.

2. Related work

2.1. Multi-branch feature extraction

The concept of residuals proposed by ResNet [29] addresses the degradation phenomenon that occurs as deep neural networks deepen. By incorporating shortcut connections, a linear transformation is added to the nonlinear neural network, allowing the network to possess both linear and nonlinear transformation capabilities. This ensures that the network can more easily map back to the original data during backpropagation. Under the guidance of residual thinking, a large number of efficient classification networks [11–15] have emerged. DenseNet [30] connects every feature map with a backward shortcut, achieving feature reuse. MobileNetV2 [11] and MobileNetV3 [12] introduced the inverted residual structure and the inverted residual structure augmented with Squeeze- and- Excitation (SE) [31]. ShuffleNetV1 [13] uses group convolution (GConv) and depthwise separable convolution (DWConv) to replace the standard 1 × 1 and 3 × 3 convolutions, respectively, and introduces a channel shuffle operation to facilitate intergroup infomation exchange. ShuffleNetV2 [14], building on the foundation of ShuffleNetV1, proposes four efficient design principles, providing effective guidance for the design of neural network architectures. VoVNet [32] proposes a One-Shot Aggregation module (OSA) to remove a large amount of redundant feature information, retaining connections with distinct feature information, thereby reducing memory access costs and enhancing the detection performance of the model. CSPNet [33] splits the gradient flow, allowing it to propagate through different network paths, achieving a rich combination of gradients with low computational cost, and enhancing the learning capability of Convolutional Neural

Networks (CNN). ELAN [34] reduces the use of transition layers and increases the minimum gradient path length, thus allowing the network to maintain good accuracy even at greater depths. RepVGG [35] employs reparameterization for multi-branch training and single-branch inference, resulting in a high-performance, fast network. Building on the ideas of re-parameterization and ELAN, by incorporating a forward information compensation branch, we propose RSELAN, which combines multi-branch with re-parameterization, featuring both lightweight and high-performance advantages.

2.2. *Multi-scale feature fusion*

FPN [19] has emerged as the predominant approach for multi-scale feature fusion. The top-down connections in FPN greatly enhance the representational capacity of feature maps. Following FPN, a large number of multi-scale feature fusion methods have emerged. PA-FPN [21] adds a topdown fusion path on the basis of FPN, integrating detailed information from lower-level features into higher-level features. NAS-FPN [36] uses neural architecture search (NAS) to obtain more efficient feature fusion connection methods. RFB [37] increases the spatial distribution and eccentricity of receptive fields by simulating receptive fields of human vision. BiFPN [20] proposes the idea of efficient bidirectional cross-scale connections and weighted feature fusion, merging high-level features with low-level features effectively. ASFF [22] proposes an adaptive feature fusion method to address the problem of inconsistent feature fusion scales. FPT [38] enables feature maps of different scale to capture non-local information of objects at various scales through self-attention mechanisms applied to different feature maps. DetectoRS [39] adopts the Recursive Feature Pyramid (RFP) to feed back fused features into the feature maps extracted by the backbone, achieving looking and thinking twice and utilizing a feedback mechanism to capture hidden information. Our proposed FSCFF integrates shallow features with deep features in the forward propagation process, which alleviates the problem of losing high-level semantic information in deep feature maps.

2.3. *Real-time object detector*

YOLO series detectors [1–4,6–9], the most widely used single-stage real-time detectors at present, which are mainly divided into two types: anchored and unanchored, and have always dominated the real-time detection mission. They generally use CSPNet [33] or ELAN [34] as the backbone for feature extraction of images. The images are processed through a feature extraction network to obtain multi-scale feature maps, which are then fused across different scales using FPN or an enhanced version of FPN. Finally, the fused multi-scale feature maps are sent to the detection head for prediction. Anchor-based detectors [1–5] use a coupled detection head for prediction, while anchor-free detectors [5,40] generally use a decoupled detection head. Recently, many end-to-end detectors based Transformers, such as DETR [40] and Deformable DETR [5], have emerged. They employ a bipartite graph matching algorithm to match each predicted object with a real label on a one-to-one basis, directly obtaining the category and location information of the object. This eliminates the time-consuming Non-Maximum Suppression (NMS) process in traditional object detection. However, these types of detectors converge slowly, and they require a large amount of computation and parameters, making the training very costly. Without corresponding pre-trained models, it is difficult to apply them in relevant fields. In the field of real-time detection, the DETR series can not yet replace the YOLO series. Therefore, we choose the framework of YOLOv8 [4] as the baseline to improve, use RSELAN as the backbone for feature extraction and FSCFF as the feature fusion network for feature fusion, resulting in a new real-time detector RFSC-Net.

3. Method

In this section, we provide a detailed introduction to RFSC-Net. The overall structure of RFSC-Net is shown in Fig. 1, consisting of a feature extraction network RSELAN, a multi-scale feature fusion network FSCFF, and a prediction head. Based on the YOLOv8n structure, we make improvements to the backbone and the multi-scale feature fusion network. Now, we will focus on RSELAN and FSCFF, describing their internal structures.

3.1. *The design of the RSELAN structure*

3.1.1. *The design of RepELAN*

RepELAN adopts a multi-branch structure design similar to ELAN, as depicted in Fig. 1. The multi-branch design allows gradients to flow through multiple paths, enhancing gradient propagation, reducing gradient vanishing, and strengthening the network's ability to process and receive gradients, enabling it to become deeper. What's more, different branches learn different features during feature learning, avoiding the model falling into local optimum. Using a multi-branch structure for feature extraction can capture features of different receptive field sizes, increasing the diversity of feature learning.Typically, during model training, multi-branch convolution is used to enhance the non-linear capabilities of network and provide diverse connections, thereby improving the representational power of the network. However, directly using multi-branch convolution to enhance the representational capacity of the model also increases inference costs. In contrast, RepConv [35] is multi-branch during training but single-branch during inference, which increases only the training cost without adding to the inference burden. Considering both model accuracy and speed, we choose RepConv to replace conventional convolution.

RepELAN inspired by the multi-branch design, drawing mainly from the C2f design in YOLOv8 [4], removes the 1 ×1 convolution from the firstbranch compared to ELAN and performs multi-branch operations post-convolution. We replace the regular 3 × 3 convolution with a reparameterized convolution, as shown in Fig. 2 (b), where normal Bottleneck uses two regular 3 × 3 convolutions, while DRepBottleneck uses two re-parameterization 3 × 3 convolutions followed by a shortcut connection.

3.1.2. *Feature compensation branch*

Due to the insufficient feature information of objects under low light conditions, directly using pooling operations for downsampling can lead to a reduction in image information and fail to adequately preserve detailed information. We use SPD-Conv [41] (Space-to-depth Convolution) for downsampling, as detailed in Fig. 3, which achieves downsampling by splitting and reorganizing the pixel values of the input feature map. Since the pixel values of the feature map do not change, this method preserves as much detailed information as possible in the lower-level feature maps. After downsampling, dimension alignment is achieved through 1 × 1 convolution. The use of SPD-Conv is illustrated in RSELAN, as shown in Fig. 1. Before each downsampling convolution operation, the SPDConv downsampling operation is performed to retain more shallow layer detailed information, which compensates for the higher-level feature maps, thereby adding more detailed information to the high-level semantic feature maps. Such as edge detailed information, texture information of object features, these details which are already scarce in low-light environments, are significantly lost after feature extraction, and as the network deepens, detailed information is severely missing, which will reduce the final detection accuracy. We minimize the loss of details during forward propagation by progressively compensating low-level semantic information into high-level feature maps.

After obtaining the feature maps of the compensation branch and the main branch, it is necessary to fuse the features. General feature fusion methods usually perform direct addition operations, which overlooks

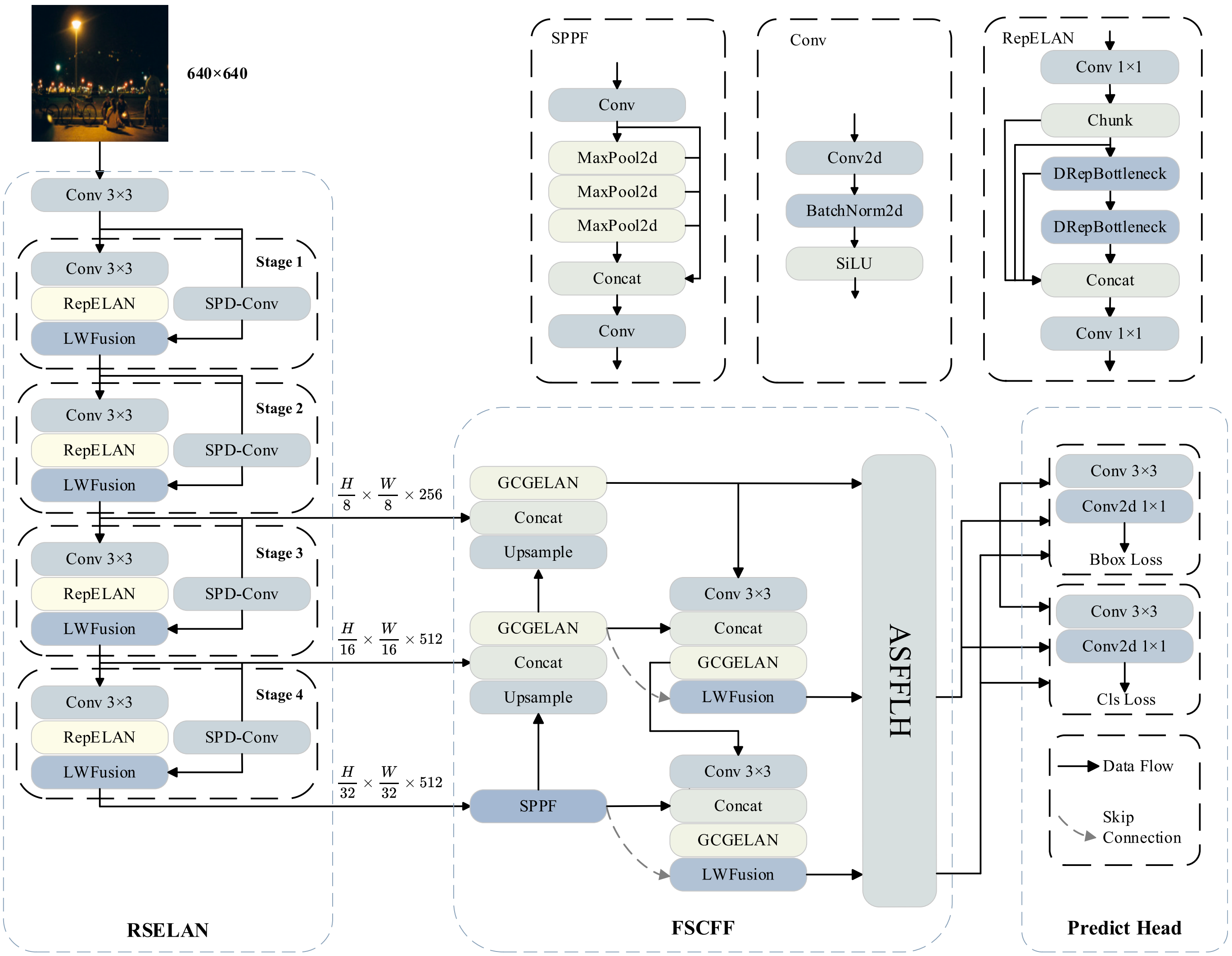
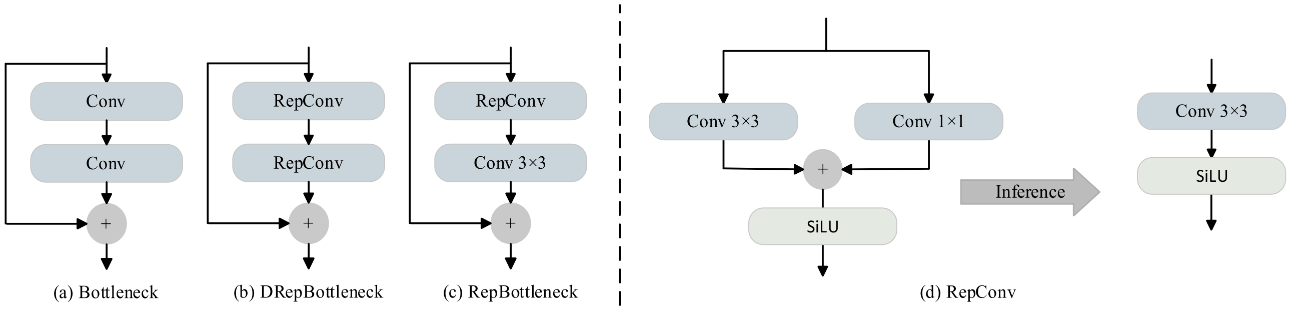


Fig. 1. The RFSC-Net consists of three components: RSELAN, FSCFF, and the Predict Head.



(a) Bottleneck (b) DRepBottleneck (c) RepBottleneck (d) RepConv

Fig. 2. The architecture of DRBottleneck and RepConv.

the varying importance of different feature maps. It is unreasonable to assign a higher weight to the feature maps that contribute less. This can result in containing less valid information in the fused feature maps, thus degrading the performance of object detection. Therefore, we use learnable weight feature fusion method (LWFusion), with the specific calculation method as follows:

$$p_i = \sum_{i=0}^{n} \frac{w_i}{\epsilon + \Sigma_j w_j} \cdot I_i \#(1)$$

the parameter w_j is derived from w_i through the *SiLU* activation function. Here, w_i is a learnable parameter. The i^{th} feature map involved in the fusion is represented as I_i. A very small number ϵ, set to 0.0001 in this paper, is used to prevent numerical instability. The value n epresents the number of feature maps in the fusion, and p_i is the resulting feature map output after fusion. This normalization method restricts the weight values within the range [0–1], avoiding the occurrence of outliers and making the training process more stable.

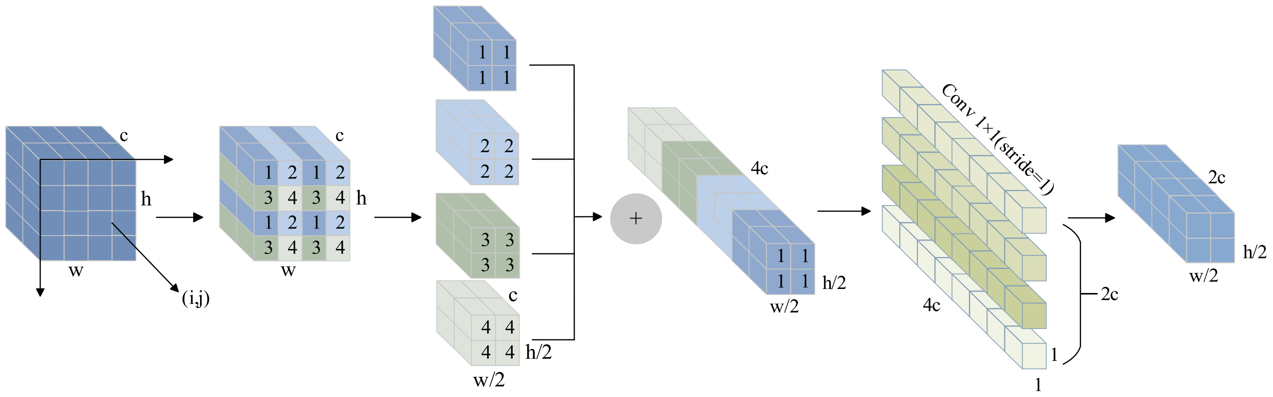


Fig. 3. The structure of SPD-Conv.

3.2. *Design of the FSCFF architecture*

3.2.1. *The design of GCGEALN*

This section describes the GCGELAN structure, as showcased in Fig. 4. The GELAN [42] structure is a generalized efficient aggregation network. It balances lightweight design, inference speed, and detection accuracy by combining ideas from CSPNet [33] and ELAN [34]. Internally, GELAN uses RepConv to replace regular convolution, enhancing the richness of feature extraction. Within GELAN, a RepNCSP structure is utilized. The detailed architecture of RepNCSP is depicted in Fig. 5. RepNCSP uses reparameterization convolution within CSPDarkNet. The RepBottleneck structure is illustrated in Fig. 2 (c). It replaces the first regular convolution in the Bottleneck with a RepConv, thus adding only a minimal training cost without increasing inference expenses.

The design of GELAN [42] allows the branches inside RepNCSP to access multi-branch feature information and extract richer semantic information. However, for the two branches that have not undergone the RepNCSP structure, they are connected by identity, without any feature transformation operations. For these two branches, the retained information tends more towards the original input information. Since the original feature map contains a lot of information, during the feature fusion stage, it is not necessary to extract as much information as in the feature extraction network. Instead, it should focus more on which features to fuse and how to perform the fusion. The GELAN network has considered the question of how to do the fusion, but has not chosen which features to integrate. Therefore, we conduct a screening of the input feature maps to allow the model to focus on important information and reduce focus on secondary information. Based on this idea, we designed three structures using a Global Context (GC) [43] attention mechanism to filter features. The specific locations where GC is added are shown in Fig. 4. A structural diagram of GC can be found in Fig. 6.

3.2.2. *Integration of features at the same scale*

The top-down information flow pathway in the FPN [19] network allows the lower-level feature maps to contain the high-level semantic information found in upper-level feature maps, enriching the semantic information in the lower-level feature maps. Building on the FPN network, PANet [21] adds a bottom-up pathway that integrates the precise location information contained in lower-level feature maps into the upper-level feature maps, compensating for the lack of positional information in the higher-level feature maps. After feature fusion through PA-FPN, both low-level and high-level feature maps contain rich semantic and positional information. In the process of bidirectional feature fusion by PA-FPN, due to the increase in propagation paths, there is also an increase in feature loss. Although it fuses with the feature maps of the same scale in the FPN path each time, the PANet path is relatively deeper compared to the initially ex- tracted feature maps. Therefore, using only a fusion may not be sufficient to compensate for the feature loss. Therefore, in the three feature maps of the bottom-up path output in PA-FPN, we use identity connections on the two higher-level feature

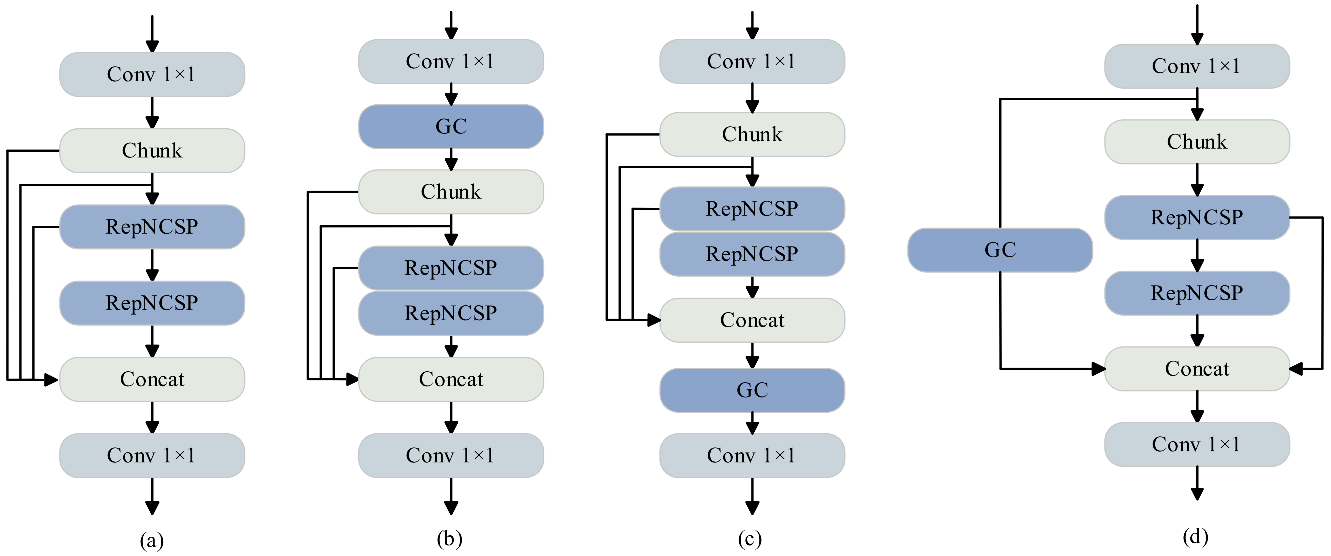


Fig. 4. The architecture of GCGELAN: (a) GELAN, (b) integrate a Global Context (GC) block antecedent to the Chunk processing stage, (c) applies a GC block subsequent to the Concatenation operation and (d) implements a GC block in a parallel fashion with respect to the primary computational branch.

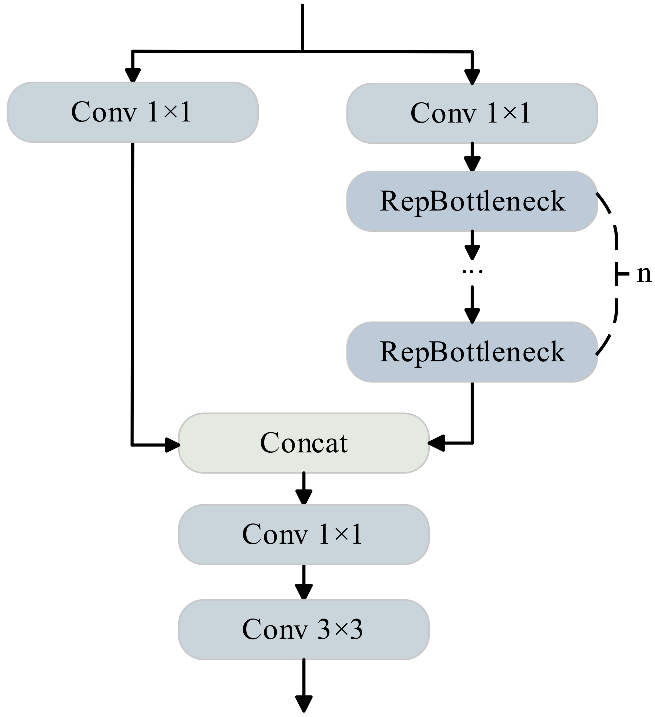


Fig. 5. The structure of RepNCSP.

maps output by PA-FPN to increase the degree of fusion of the output feature maps. The type of connection is shown in Fig. 1 as a Skip Connection. The feature fusion operation used is the same as the method of the feature extraction network, adopting the LWFusion method, which achieves a greater degree of feature fusion at a minimal cost.

3.2.3. The design of ASFFLH

The output results of PA-FPN are used as the input for ASFFLH, whose process flowchart is shown in Fig. 7. ASFFLH first scales the feature maps from different layers to the same size, and then performs adaptive fusion on the scaled feature maps. The number of channels and the size of the feature maps after adaptive fusion are the same as the input of this layer. Since the data flow in feature fusion is only from bottom to top, which means, from larger feature maps to smaller scale feature maps for information compensation, there is only downsampling during feature map scaling, with no upsampling involved. For feature maps that are downsampled by a factor of two, we use a 3 × 3 convolution kernel with a stride of 2. For feature maps that are downsampled by a factor of four, maximum pooling is first used for one scaling, followed by downsampling using a convolution kernel with a stride of 2 and a size of 3 × 3. After scaling the feature map size for channel alignment, all feature maps are subjected to channel compression. A general practice is used here, employing a convolution kernel of size 1 × 1 with a stride of 1, compressing the number of channels in all feature maps to the same amount. The number of channels we set is the same as the implementation in ASFF [22], using eight channels. After scaling, the low-level feature maps are adaptively fused with the high-level feature maps, as shown in the ASFFLH2 and ASFFLH3 modules in Fig. 7. The fusion formula is as follows:

$$y_{ij}^{l} = \sum_{k=0}^{l} \alpha_{ij}^{k\to l} \cdot x_{ij}^{k\to l} \tag{2}$$

here $l \geq 1$, where l is the layer number, k represents the feature map from the k^{th} layer, $x_{ij}^{k\to l}$ is the feature map resized from the k^{th} layer to the l^{th} layer, with ij as spatial coordinates, $\alpha_{ij}^{k\to l}$ is the weight of each position in the feature map, and y_{ij}^{l} is the output feature map after fusing the feature maps of l^{th} layer. The calculation formula for $\alpha_{ij}^{k\to l}$ is as follows:

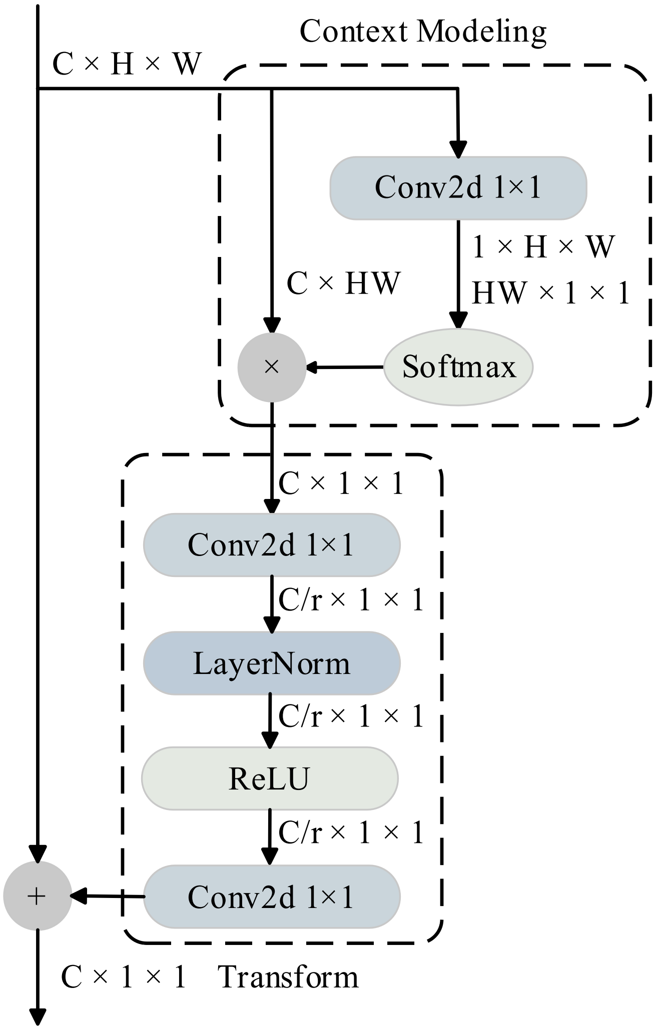


Fig. 6. The Global Context (GC) block structure consists of two parts: the Context and the Transform structures.

$$\alpha_{ij}^{k\to l} = \frac{e^{\alpha_{ij}^{k\to l}}}{\sum_{k=0}^{l} e^{\alpha_{ij}^{k\to l}}} \tag{3}$$

the calculation of is completed through the Softmax activation function. After normalization, the range of the weight is restricted to the interval [0–1], making the process more stable.

At $l = 0$, the shallowest feature map is obtained, which does not undergo feature fusion with other layers. Therefore, the aforementioned adaptive fusion operations for feature compensation are not conducted. Given the characteristics of the feature map at layer 0, lower-level feature maps have a lesser degree of feature fusion. In the PA-FPN process, only one GCGELAN fusion process is conducted, and the feature map of this layer might have insufficient feature fusion. In the layer 0 of ASFFLH, fusion operation is performed again to enhance the degree of feature information fusion. Considering that the objects in the lower-level feature maps are smaller, retaining more detailed information helps improve detection accuracy. Therefore, no attention mechanism is used during fusion to maintain uniformity with the structure in PA-FPN. For instance, if GCGELAN is used in PA-FPN, then the fusion module used in the layer 0 of ASFFLH would be GELAN without GC attention.

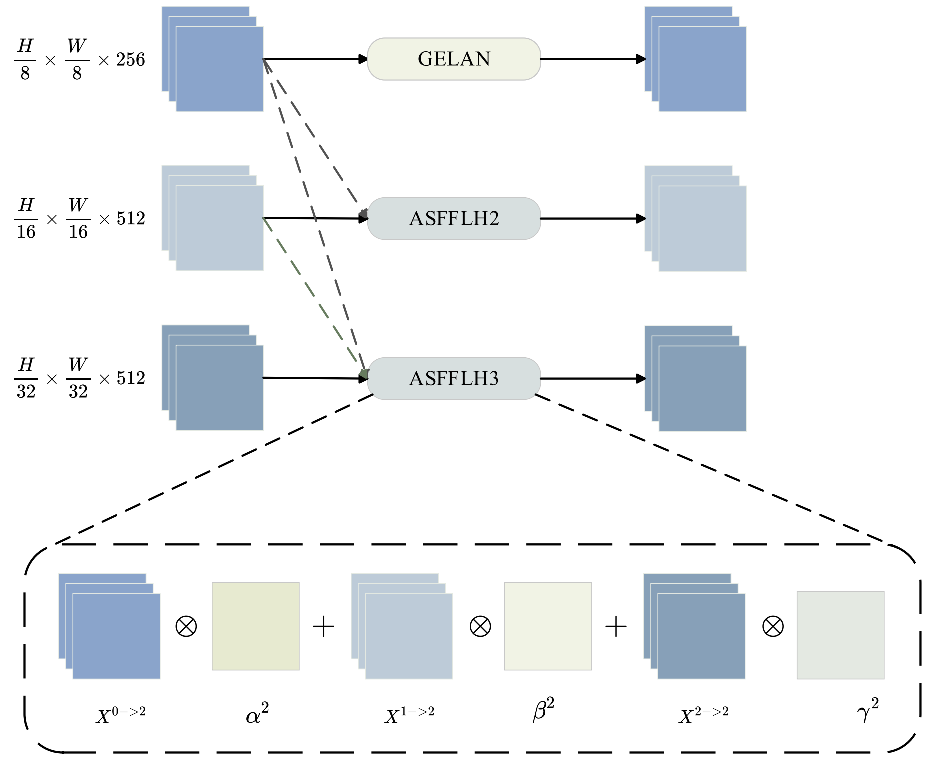

$\frac{H}{8} \times \frac{W}{8} \times 256$
$\frac{H}{16} \times \frac{W}{16} \times 512$
$\frac{H}{32} \times \frac{W}{32} \times 512$
GELAN
ASFFLH2
ASFFLH3
$X^{0->2}$
α^2
$X^{1->2}$
β^2
$X^{2->2}$
γ^2

Fig. 7. Adaptive Spatial Feature Fusion Mechanism from Low-level Semantics to High-level Semantics (ASFFLH) requires scaling the features of other layers entering each layer to the same size, and then fusing them based on learnable weight.

After the feature maps pass through the ASFFLH, they are once again merged with the shallow feature maps, thereby enriching the feature information in the high-level feature maps. For the lowest layer feature maps, after passing through GELAN, the features in the feature maps are more fully integrated.

4. Experiments

4.1. Experimental setup

Implementation details: This paper evaluates experiments conducted on ExDark, and generalization experiments on VisDrone, DarkFace, COCO 2017, and PASCAL VOC 2007 + 2012. Table 1 shows the number of categories as well as the training and validation sets for each dataset. We train all the models from scratch, standardizing the training for 300 epochs with a batch set to 32. The optimizer uses Stochastic Gradient Descent (SGD) with weight decay, setting the weight decay rate at 0.005 and the initial learning rate at 0.01. Warmup is used for the learning rate during the first three epochs, followed by a linear change in the learning rate thereafter. In the preprocessing, data augmentation techniques such as Mosaic, MixUp, Random HSV, Random Affine, and Horizontal Flip are used, with the Mosaic data augmentation off in the last 10 epochs. Our experiments are conducted on GeForce RTX 2080 Ti.

Table 1
Sizes and Number of Categories of Training and Validation Sets in Different Datasets.

Dataset	Training Set	Validation Set	Number of Classes
Exdark [23]	5890	1473	12
VisDrone [26]	6471	548	10
DarkFace [27]	4800	1200	1
COCO 2017 [24]	118,287	5000	80
PASCAL VOC 2007 + 2012 [25]	16,551	4952	20

Evaluation Metrics: We use mean Average Precision as the evaluation metric for experiment, which includes mAP and AP50. Additionally, the parameters (Params), Giga Floatingpoint Operations Per Second (GFLOPs), and Latency of the model are assessed.

4.2. Comparison experiment

Comparison of lightweight feature extraction networks: Table 2 shows a comparison of RSELAN with different lightweight backbones in RFSC-Net, achieving the highest accuracies of 44.0% and 69.3% compared to other backbones trained from scratch. In comparative experiments, GhostNetV2 1.0 ×, based on the CNN method, has the highest detection accuracy among the backbones. RSELAN exceeds GhostNetV2 1.0× by 0.4% in mAP and 0.6% in AP50, while the overall model has less than half the parameters of GhostNetV2 1.0 × and requires 0.2 G less in computational resources. Compared to MobileNetV2 0.5× [11], which has the lowest number of parameters, RSELAN achieves a 1.3%

Table 2
Comparison of Different Lightweight Backbones in RFSC-Net.

Backbone	Params (M)	FLOPs (G)	mAP	AP50
CSPDarknet [33] +FSCFF	4.4	10.8	42.8%	68.2%
MobileNetV2 0.5× [11] + FSCFF	2.6	8.1	42.7%	67.6%
MobileNetV2 0.75× [11] + FSCFF	3.3	9.6	42.6%	67.6%
MobileNetV3-Small 1.0 [12] + FSCFF	3.0	8.0	40.7%	64.6%
ShuffleNetV1 0.5× [13] + FSCFF	2.9	8.2	36.1%	59.9%
ShuffleNetV1 1.0× [13] + FSCFF	3.6	10.2	38.2%	62.6%
ShuffleNetV2 1.5× [14]+ FSCFF	4.4	11.7	40.6%	65.1%
ShuffleNetV2 1× [14] + FSCFF	3.2	9.2	39.6%	63.9%
FasterNet-T0 [17] + FSCFF	4.5	12.4	39.0%	63.6%
GhostNetV2 1.0× [16] + FSCFF	6.5	10.3	43.6%	68.7%
MobileViT-XXS [15] + FSCFF	3.6	11.7	43.7%	68.9%
RepViT-M0.6 [18] + FSCFF	4.5	13.6	40.3%	65.4%
RSELAN + FSCFF (Ours)	3.0	10.1	44.0%	69.3%

and 1.7% increase in mAP and AP50 respectively, with only an additional 0.4 M parameters. Among lightweight Vision Transforms, RepViT [18] and MobileViT [15] are two of the better-performing backbones. Additionaly, RSELAN exceeds Mobile ViT-XXS by 0.3% in mAP and 0.4% in AP50, while also reducing the computational load by 1.6 G. RSELAN achieves the highest accuracy with parameter counts and computational load comparable to other backbones.

Comparison of multi-scale feature fusion: Table 3 shows a comparison between FSCFF in RFSC-Net and the current generally used multi-scale feature fusion methods. Using the C2f structure from the baseline in the feature fusion network for comparison. Compared to other feature fusion methods, it achieves the highest mAP of 44.0% and AP50 of 69.3% on the low-light dataset ExDark. Compared to PA-FPN [19], the mAP and AP50 of FSCFF increased by 2.0% and 1.5%, respectively, with only a 0.3 M increase in the number of parameters. The feature fusion network used in ASFF [22] is a combination of FPN and ASFF. Compared to using FPN for feature fusion, the mAP and AP50 using FSCFF improved by 3.9% and 3%, respectively.

Comparison of attention mechanisms: The effects of using different attention mechanisms in GCGELAN are compared in Table 4: channel attention, coordinate attention, serial channel-space attention, and parallel channel-space attention. The results indicate that the attention mechanism plays a crucial role in final detection accuracy, and using the appropriate attention method will significantly improve recognition accuracy. Among the four types of attention mechanisms, channel attention, coordinate attention, and serial channel-space attention perform poorly, showing a significant drop in accuracy compared to not using any attention mechanism. The parallel BAM [44] perform better than the other three, but there is still a decrease of 0.5% in mAP and 0.7% in AP50. GC is the only attention mechanism in all tested that surpassed the RFSC-Net without using attention. This suggests that GC can capture more global, which helps improve detection ac- curacy under low-light conditions.

Comparison with various SOTA methods: Table 5 shows a comparison between RFSC-Net and other real-time detectors trained from scratch. With the same number of parameters as YOLOv8n, the mAP and AP50 of RFSC- Net increase by 2.0% and 2.1% respectively compared to the baseline, with only a 0.4 ms increase in inference speed. Gold-YOLO-n achieved an AP50 of 43.7% with a latency of only 2.9 ms. However, compared to RSFC-Net, its mAP is 0.3% lower, and its parameters count is four times that of RSFC-Net. Compared to YOLOv9t, the mAP of RFSCNet improve by 1.5%, and the AP50 is higher 2.3%. Among all the models compared, YOLOv9s exhibits the highest accuracy; however, it also has significantly more parameters and a higher computational cost compared to RFSC-Net. In summary, RFSC-Net strikes a favorable balance between model size, accuracy, and speed compared to the other models.

4.3. Ablation study

Ablation among the components of RFSC-Net: We use the architecture of YOLOv8n as the baseline model to verify the effects of different structures relative to the baseline model. As shown in Table 6, each structure is used in replacement of the base YOLOv8n. Notably, RSELAN employs 512 channels in stage 4, which is half the number of channels used by the YOLOv8 model. Correspondingly, in the feature fusion network, the number of output channels for the High Prediction Layer part is also 512. Experimental results indicate that while the mAP of RSELAN did not in- crease, the number of channels used in the fourth stage of RSELAN is half of that in YOLOv8n. As a result, the parameters count of the model is reduced by 0.3 M, and AP50 is increased by 0.6%. The use of GCGELAN and RSELAN improves both mAP and AP50. It can be seen from the results that the combined use of FSCFF* and GCGELAN more effective than using them independently. Overall, different combinations contribute to the improvement of accuracy, but the combination of the three is the most effective, achieving 44.0% mAP and 69.3% AP50.

Table 3
Comparison of different multi-scale feature fusion methods in RFSC-Net.

Method	Params (M)	FLOPs (G)	mAP	AP50
RSELAN + FPN [21]	1.6	9.3	40.1%	66.0%
RSELAN + PA-FPN [19]	2.7	8.4	42.0%	67.8%
RSELAN + ASFF [22]	2.3	8.7	41.1%	67.5%
RSELAN + BiFPN [20]	2.4	8.3	41.9%	67.6%
RSELAN + AFPN [45]	1.9	7.4	39.3%	63.8%
RSELAN + ResGFPN [46]	2.9	8.6	42.2%	68.4%
RSELAN + FSCFF(Ours)	3.0	10.1	44.0%	69.3%

Table 4
Comparison of different attentions in GCGELAN.

Attention Module	Params (M)	FLOPs (G)	mAP	AP50
GELAN [42]	3.0	10.1	43.7%	68.9%
GEALN+BAM [44]	3.0	10.1	43.2%	68.2%
GEALN+CA [47]	3.0	10.1	42.3%	67.1%
GEALN+ECA [48]	3.0	10.1	42.4%	67.6%
GEALN+ELA [49]	3.0	10.1	42.9%	67.7%
GEALN+EMA [50]	3.0	10.2	42.3%	67.3%
GEALN+GAM [51]	4.3	14.4	41.8%	66.6%
GEALN+SE [52]	3.0	10.1	42.5%	67.5%
GEALN+CBAM [53]	3.0	10.1	42.4%	67.4%
GEALN+ShuffleAttention [54]	3.0	10.1	42.5%	66.9%
GEALN+GC [43]	3.0	10.1	44.0%	69.3%

Ablation among the components of RSELAN: In order to verify the effect of Feature Compensation Branch, we conduct ablation experiments on Feature Compensation Branch in RSELAN, and the results of these explorations are presented in Table 7. We conduct experiments with different Feature Compensation Branches. From the results of the experiments, we can find that preserving shallow semantic information has different degrees of benefits for the accuracy of RFSC-Net by using Feature Compensation Branches. Among them, the combination of SPD-Conv and LWFusion has the best performance, with mAP and AP50 increasing by 1.1% and 1.4% respectively compared to the case without using the feature compensation branch.

Ablation of GC at different locations in GCGELAN: We conduct ablation experiments on different positions of GC in GCGELAN, Table 8. Compared with not using GC, when GC is placed before and after Res NCSP in GCGLEAN, the accuracy is lower than that without using GC. We believe that applying the attention mechanism to the features processed by the RepNCSP branch will lead to overconcentration and information loss. In order to reduce information loss, we only apply the attention mechanism to branches that do not pass through RepNCSP. The final experimental results show that both mAP and AP50 have increased.

Ablation of ASFFLH module: As shown in Table 9, the results of using ASFFLH and ASFF in RFSC-Net indicate that the accuracy of using ASFF decreases by 0.1% compared to direct prediction. This validates our previous hypothesis that the fusion of higher-level semantic information with lower-level semantics can cause semantic conflicts, thereby affecting the predictions at that level. Due to the use of adaptive feature fusion, the influence of semantic conflicts is minimal, but the number of parameters and computations significantly increases. The fusion of ASFFLH improves by 1.1% compared to not using it, and AP50 improves by 0.8%. The number of parameters and computations is the same as using ASFF, validating that the semantic supplementation from lower-level semantics to higher-level feature maps contributes greatly to the improvement of final detection accuracy.

Ablation of different blocks in ASFFLH 0*th* layer: As illustrated in Table 10, in the **0*th*** layer of ASFFLH, we experimented with some generally used blocks in current lightweight networks. Obviously, from the experimental results, GELAN has the highest accuracy. Using blocks different from those in feature fusion networks is not conducive to

Table 5
Comparison of RFSC-Net with the Latest Real-Time Detectors. Latency is the result of averaging the inference across all validation sets in ExDark. Latency is measured on a a GeForce RTX 2080 Ti.

Method	Input Size	Params(M)	FLOPs(G)	mAP	AP50	Latency$_{bs=32}$
YOLOv8n (baseline) [4]	640	3.0	8.1	42.0%	67.2%	3.3 ms
YOLOv3tiny [6]	640	12.1	18.9	33.8%	60.2%	7.6 ms
YOLOv5n	640	1.8	4.2	33.3%	61.3%	2.3 ms
YOLOv5s	640	7.0	15.9	38.0%	67.2%	8.6 ms
YOLOv6n [1]	640	4.6	11.4	42.7%	68.8%	4.3 ms
YOLOXtiny [3]	416	5.0	7.6	33.7%	63.4%	5.4 ms
YOLOv7tiny [2]	416	6.0	13.1	39.6%	67.9%	12.2 ms
YOLOv8s [4]	640	11.1	28.5	43.7%	69.7%	4.0 ms
YOLOv9t [9]	640	2.0	7.6	42.5%	67.0%	2.7 ms
YOLOv9s [9]	640	7.2	26.8	45.7%	71.6%	4.1 ms
YOLOv10n [8]	640	2.7	8.3	41.0%	65.9%	2.4 ms
YOLOv10s [8]	640	8.0	24.5	43.8%	69.8%	3.7 ms
Gold-YOLO-n [55]	640	12.0	5.6	43.7%	70.1%	2.9 ms
RFSC-Net (Ours)	640	3.0	10.1	44.0%	69.3%	3.7 ms

Table 6
Ablation of different structures in RFSC-Net. FSCFF* refers to the FSCFF that uses C2f instead of GCGELAN.

RSELAN	GCGELAN	FSCFF*	Params (M)	FLOPs (G)	mAP	AP50
×	×	×	3.0	8.1	42.0%	67.2%
√	×	×	2.7	8.4	42.0%	67.8%
×	√	×	2.8	8.4	42.7%	68.0%
×	×	√	4.3	10.0	42.3%	67.7%
√	√	×	2.4	8.6	42.1%	67.2%
√	×	√	4.0	10.3	42.7%	67.8%
×	√	√	4.4	10.8	42.8%	68.2%
√	√	√	3.0	10.1	44.0%	69.3%

Table 7
Ablation study of different Feature Compensation Branches in RSELAN. No use means that the feature compensation branch is not used.

Method	Params (M)	FLOPs (G)	mAP	AP50
No use	2.9	9.8	42.9%	67.9%
AvgPool + LWFusion	2.9	9.9	43.0%	67.6%
MaxPool + LWFusion	2.9	9.9	43.6%	68.9%
SPD-Conv + LWFusion	3.0	10.1	44.0%	69.3%

Table 8
Ablation of GC at different locations in GCGELAN.

Location	Params (M)	FLOPs (G)	mAP	AP50
None	3.0	10.1	43.8%	68.9%
Before	3.0	10.1	42.0%	66.9%
After	3.2	10.3	43.3%	68.8%
Middle (Ours)	3.0	10.1	44.0%	69.3%

Table 9
Ablation of multi-scale adaptive fusion in RFSC-Net divides the fusion into three cases: No use, ASFF, and ASFFLH. No use makes predictions directly after PA-FPN.

Method	Params (M)	FLOPs (G)	mAP	AP50
No use	2.3	8.6	42.9%	68.5%
ASFF [22]	3.0	10.1	42.8%	68.5%
ASFFLH	3.0	10.1	44.0%	69.3%

Table 10
Ablation of different blocks in ASFFLH 0th layer. Identity is the act of returning the input directly without applying any operations.

Block	Params (M)	FLOPs (G)	mAP	AP50
Identity	3.0	9.6	43.2%	68.6%
GCGELAN	3.0	10.1	43.5%	68.9%
C2f [4]	3.0	9.9	43.2%	68.3%
C3 [33]	3.0	9.8	43.7%	69.0%
Residual Block [29]	3.0	9.6	43.4%	68.3%
FasterNet Block [17]	3.0	9.6	42.4%	67.2%
StartNet Block [56]	3.0	10.1	42.3%	67.3%
ELAN [34]	3.0	10.2	43.5%	68.2%
GELAN [42]	3.0	10.1	44.0%	69.3%

improving detection accuracy, verifying what we mentioned in 3.2.3 that using blocks with consistent overall structure can maximize the effectiveness of ASFFLH.

Result visualization: Fig. 9 shows the detection results of RFSC-Net with YOLOv8n YOLOv9t and YOLOv10n on ExDark, which significantly improves the detection of underrepresented objects compared to other models. Additionally, we use GradCAM [57] to visualize the final three layers of prediction feature maps, as shown in Fig. 8. Under low light conditions, the four detectors have comparable detection performance for large objects, but RFSC-Net demonstrates superior performance in learning about and small objects.

4.4. Generalization experiment

To validate the generalizability of RFSC-Net, we conduct experiments on COCO 2017, PASCAL VOC 2007 + 2012, VisDrone, and DarkFace. The experimental results are shown in Table 11. For the small object dataset VisDrone, RFSC-Net improves the mAP by 1.6% compared to the baseline. On the low-light small object dataset DarkFace, RFSCNet also achieves a 0.5% improvement in mAP compared to the baseline.

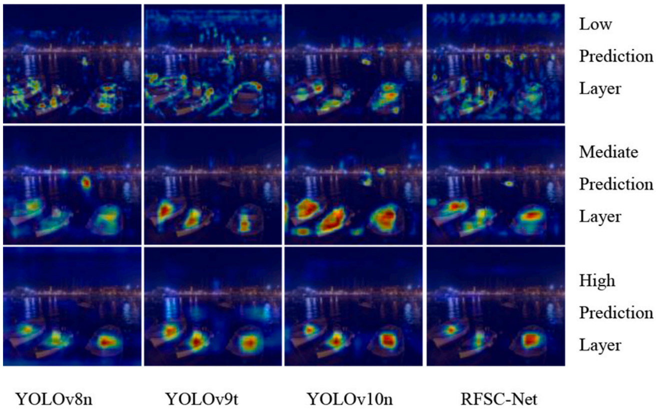


Fig. 8. Comparison of RFSC-Net with YOLOv8n, YOLOv9t, and YOLOv10n for heatmap prediction.

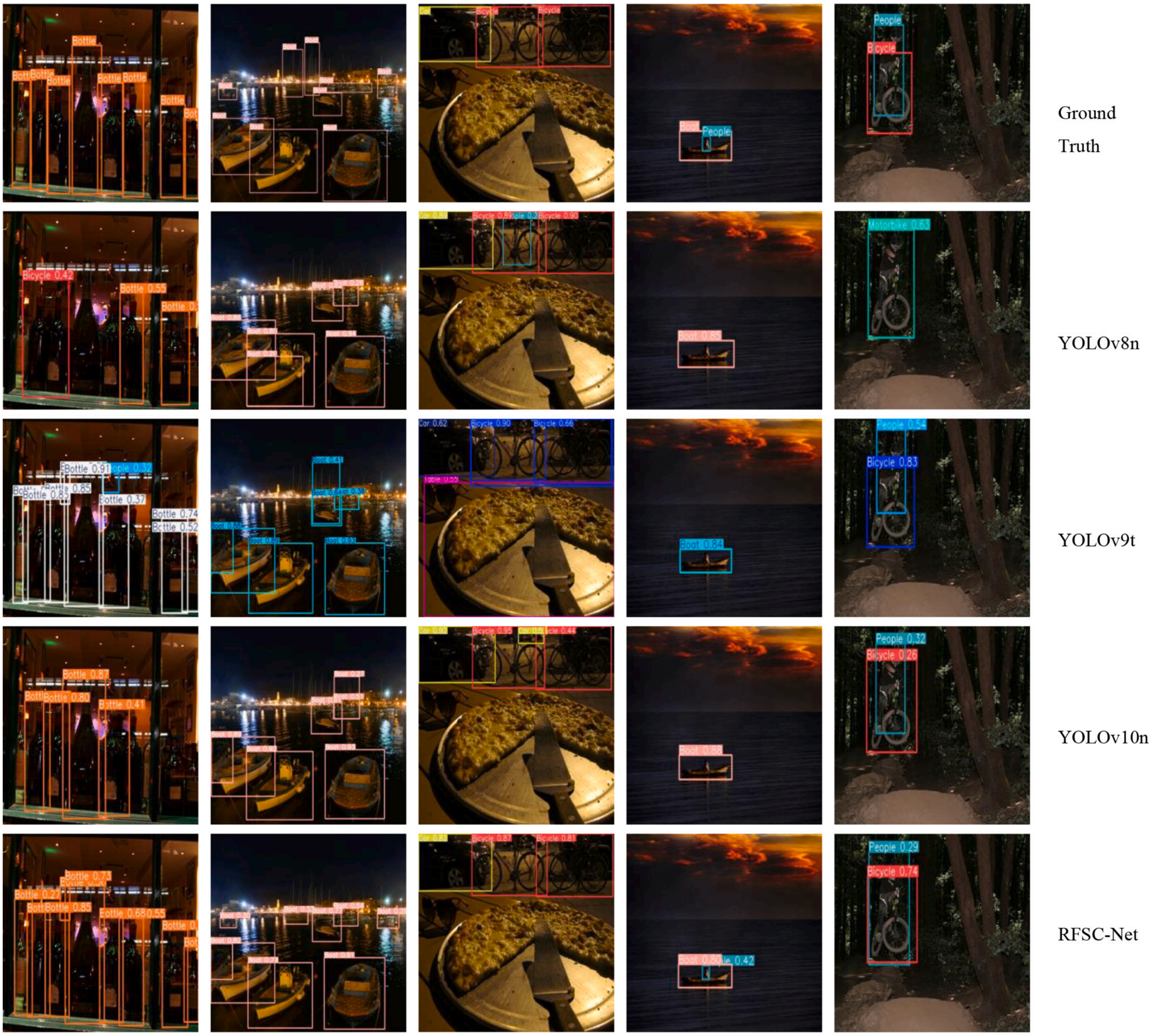


Fig. 9. Comparison of detection results of RFSC-Net with YOLOv8n, YOLOv9t, and YOLOv10n on the ExDark validation set.

Table 11
The performance of RFSC-Net on different datasets.

Dataset	mAP		AP50	
	YOLOv8n	RFSC-Net	YOLOv8n	RFSC-Net
VisDrone	19.4%	21.0%	33.3%	35.6%
DarkFace	19.4%	19.9%	45.8%	46.9%
PASCAL VOC 2007 + 2012	59.1%	61.6%	79.6%	81.0%
COCO2017	36.1%	38.0%	50.8%	52.9%

Furthermore, on COCO 2017 and PASCAL VOC 2007 + 2012, RFSC-Net outperforms the baseline by 1.9% and 2.5% in mAP, respectively. It is evident that RFSC-Net is not only effective in low-light condition detection but also shows significant improvement on the more challenging DarkFace dataset.

5. Conclusion

This paper introduces RFSC-Net to tackle the problem of detail information loss in images during the forward propagation process. We propose a forward information compensation concept to counteract the loss of information during the forward propagation process, thus preserving sufficient details for the final detection and improving detection accuracy. RFSC-Net achieved a 2.0% increase in mAP compared to YOLOv8n, without adding more parameters, and the inference speed increased by only 0.4 ms.

However, there are still some issues in the current work. We find that RFSC-Net reduces the loss of detailed information during the forward propagation process, retaining more shallow information. However, due to the limited effective information in low-light images, the retained detailed features may be discontinuous. Consequently, the detector may interpret these discontinuous features as separate objects, leading to a single object being misidentified as multiple objects. In addition, although retaining more shallow semantic information can provide more useful information, it also retains more noise, which can affect the judgment of detector and reduce detection accuracy. Directly using image enhancement techniques results in a high computational cost. We are exploring the possibility of filtering out noise from the retained information in real-time conditions, aiming to enhance the quality of this information.

CRediT authorship contribution statement

Wenhao Zhang: Writing – original draft, Methodology, Investigation, Formal analysis, Conceptualization. **Huiying Xu:** Writing – review & editing, Funding acquisition. **Xinzhong Zhu:** Supervision, Funding acquisition. **Yunzhong Si:** Validation, Investigation. **Yao Dong:** Data curation. **Xiao Huang:** Project administration. **Hongbo Li:** Supervision, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (62376252, 61976196, U22A20102), the Key Project of Natural Science Foundation of Zhejiang Province (LZ22F030003).

References

[1] Chuyi Li, Lulu Li, Hongliang Jiang, Kaiheng Weng, Yifei Geng, Liang Li, Zaidan Ke, Qingyuan Li, Meng Cheng, Weiqiang Nie, et al., Yolov6: A single-stage object detection framework for industrial applications. arXiv preprint arXiv:2209.02976, *2022*.

[2] Chien-Yao Wang, Alexey Bochkovskiy, Hong-Yuan Mark Liao, Yolov7: Trainable bag-of-freebies sets new state-of-the-art for realtime object detectors, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 7464–7475.

[3] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, Jian Sun, Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021.

[4] Dillon Reis, Jordan Kupec, Jacqueline Hong, Ahmad Daoudi, Real-time flying object detection with yolov8. *arXiv preprint arXiv:2305.09972*, 2023.

[5] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, Jifeng Dai, Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv: 2010.04159*, 2020.

[6] Joseph Redmon, Ali Farhadi, Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[7] Alexey Bochkovskiy, Chien-Yao Wang, Hong-Yuan Mark Liao, Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020.

[8] Ao Wang, Hui Chen, Lihao Liu, Kai Chen, Zijia Lin, Jungong Han, Guiguang Ding, Yolov10: Real-time end-to-end object detection, 2024.

[9] Chien-Yao Wang, I-Hau Yeh, Hong-Yuan Mark Liao, Yolov9: Learning what you want to learn using programmable gradient information. *arXiv preprint arXiv: 2402.13616*, 2024.

[10] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, Xinggang Wang, Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417*, 2024.

[11] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, Liang-Chieh Chen, Mobilenetv2: Inverted residuals and linear bottlenecks, in: In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4510–4520.

[12] Brett Koonce, Brett Koonce, Mobilenetv3, in: Convolutional Neural Networks with Swift for Tensorflow: Image Recognition and Dataset Categorization, 2021, pp. 125–144.

[13] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, Jian Sun, Shufflenet: An extremely efficient convolutional neural network for mobile devices, in: In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6848–6856.

[14] Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, Jian Sun, Shufflenet v2: Practical guidelines for efficient cnn architecture design, in: In Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 116–131.

[15] Sachin Mehta, Mohammad Rastegari, Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer. *arXiv preprint arXiv:2110.02178*, 2021.

[16] Yehui Tang, Kai Han, Jianyuan Guo, Xu Chang, Xu Chao, Yunhe Wang, Ghostnetv2: enhance cheap operation with longrange attention, Adv. Neural Inf. Proces. Syst. 35 (2022) 9969–9982.

[17] Jierun Chen, Shiu-hong Kao, Hao He, Weipeng Zhuo, Song Wen, Chul-Ho Lee, S.-H. Gary Chan, Run, don't walk: Chasing higher flops for faster neural networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 12021–12031.

[18] Ao Wang, Hui Chen, Zijia Lin, Pu Hengjun, Guiguang Ding, Repvit: Revisiting mobile cnn from vit perspective. *arXiv preprint arXiv: 2307. 09283*, 2023.

[19] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, Serge Belongie, Feature pyramid networks for object detection, in: In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2117–2125.

[20] Jun Chen, HongSheng Mai, Linbo Luo, Xiaoqiang Chen, Wu. Kangle, Effective feature fusion network in bifpn for small object detection, in: In 2021 IEEE International Conference on Image Processing (ICIP), IEEE, 2021, pp. 699–703.

[21] Shu Liu, Qi Lu, Haifang Qin, Jianping Shi, Jiaya Jia, Path aggregation network for instance segmentation, in: In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8759–8768.

[22] Songtao Liu, Di Huang, Yunhong Wang, Learning spatial fusion for single-shot object detection. *arXiv preprint arXiv:1911.09516*, 2019.

[23] Yuen Peng Loh, Chee Seng Chan, Getting to know low-light images with the exclusively dark dataset, Comput. Vis. Image Underst. 178 (2019) 30–42.

[24] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, C. Lawrence Zitnick, Microsoft coco: Common objects in context, in: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 612, 2014, Proceedings, Part V 13, Springer, 2014, pp. 740–755.

[25] S.M. Mark Everingham, Ali Eslami, Luc Van Gool, Christopher K.I. Williams, John Winn, Andrew Zisserman, The pascal visual object classes challenge: a retrospective, Int. J. Comput. Vis. 111 (2015) 98–136.

[26] Pengfei Zhu, Longyin Wen, Dawei Du, Xiao Bian, Heng Fan, Qinghua Hu, Haibin Ling, Detection and tracking meet drones challenge, IEEE Trans. Pattern Anal. Mach. Intell. 44 (11) (2021) 7380–7399.

[27] Wenhan Yang Jiaying Liu Chen Wei, Wenjing Wang, Deep retinex decomposition for low-light enhancement, in: In British Machine Vision Conference, 2018.

[28] Wenhan Yang, Ye Yuan, Wenqi Ren, Jiaying Liu, Walter J. Scheirer, Zhangyang Wang, Zhang, et al., Advancing image understanding in poor visibility environments: a collective benchmark study, IEEE Trans. Image Process. 29 (2020) 5737–5752.

[29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, Deep residual learning for image recognition, in: In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[30] Huang Gao, Zhuang Liu, Laurens Van Der Maaten, Kilian Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4700–4708.

[31] Jie Hu, Li Shen, Gang Sun, Squeeze-and-excitation networks, in: In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141.

[32] Youngwan Lee, Joong-won Hwang, Sangrok Lee, Yuseok Bae, Jongyoul Park, An energy and gpu-computation efficient backbone network for real-time object detection, in: In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2019 pages 0–0.

[33] Chien-Yao Wang, Hong-Yuan Mark Liao, Yueh-Hua Wu, Ping-Yang Chen, Jun-Wei Hsieh, I-Hau Yeh, Cspnet: A new backbone that can enhance learning capability of cnn, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 390–391.

[34] Chien-Yao Wang, Hong-Yuan Mark Liao, I-Hau Yeh, Designing network design strategies through gradient path analysis. *arXiv preprint arXiv:2211.04800*, 2022.

[35] Xiaohan Ding, Xiangyu Zhang, Ningning Ma, Jungong Han, Guiguang Ding, Jian Sun, Repvgg: Making vgg-style convnets great again, in: In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 13733–13742.

[36] Golnaz Ghiasi, Tsung-Yi Lin, Quoc V. Le, Nas-fpn: Learning scalable feature pyramid architecture for object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 7036–7045.

[37] Songtao Liu, Di Huang, et al., Receptive field block net for accurate and fast object detection, in: In Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 385–400.

[38] Zhang Dong, Hanwang Zhang, Jinhui Tang, Meng Wang, Xiansheng Hua, Qianru Sun, Feature pyramid transformer, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII 16, Springer, 2020, pp. 323–339.

[39] Siyuan Qiao, Liang-Chieh Chen, Alan Yuille, Detectors: Detecting objects with recursive feature pyramid and switchable atrous convolution, in: In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 10213–10224.

[40] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, Sergey Zagoruyko, End-to-end object detection with transformers, in: European Conference on Computer Vision, Springer, 2020, pp. 213–229.

[41] Raja Sunkara, Tie Luo, No more strided convolutions or pooling: Anewcnnbuildingblockforlow-resolutionimagesandsmallobjects, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2022, pp. 443–459.

[42] Chien-Yao Wang, I-Hau Yeh, Hong-Yuan Mark Liao, Yolov9: Learning what you want to learn using programmable gradient information. *arXiv preprint arXiv: 2402.13616*, 2024.

[43] Yue Cao, Jiarui Xu, Stephen Lin, Fangyun Wei, Han Hu, Gcnet: Non-local networks meet squeeze-excitation networks and beyond, in: In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, pages 0–0, 2019.

[44] Jongchan Park, Sanghyun Woo, Joon-Young Lee, In So Kweon, Bam: Bottleneck attention module. *arXiv preprint arXiv:1807.06514*, 2018.

[45] Guoyu Yang, Jie Lei, Zhikuan Zhu, Siyu Cheng, Zunlei Feng, Ronghua Liang, Afpn: Asymptotic feature pyramid network for object detection, in: In 2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC), IEEE, 2023, pp. 2184–2189.

[46] Xianzhe Xu, Yiqi Jiang, Weihua Chen, Yilun Huang, Yuan Zhang, Xiuyu Sun, Damo-yolo: A report on real-time object detection design. *arXiv preprint arXiv: 2211.15444*, 2022.

[47] Qibin Hou, Daquan Zhou, Jiashi Feng, Coordinate attention for efficient mobile network design, in: In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 13713–13722.

[48] Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, Qinghua Hu, Eca-net: Efficient channel attention for deep convolutional neural networks, in: In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 11534–11542.

[49] Wei Xu, Yi Wan, Ela: Efficient local attention for deep convolutional neural networks. *arXiv preprint arXiv:2403.01123*, 2024.

[50] Su Daliang Ouyang, Guozhong Zhang He, Mingzhu Luo, Huaiyong Guo, Jian Zhan, Zhijie Huang, Efficient multi-scale attention module with cross-spatial learning, in: ICASSP 2023–2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2023, pp. 1–5.

[51] Yichao Liu, Zongru Shao, Nico Hoffmann, Global attention mechanism: Retain information to enhance channel-spatial interactions. *arXiv preprint arXiv: 2112.05561*, 2021.

[52] Jie Hu, Li Shen, Gang Sun, Squeeze-and-excitation networks, in: In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141.

[53] Sanghyun Woo, Jongchan Park, Joon-Young Lee, In So Kweon, Cbam: Convolutional block attention module, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 3–19.

[54] Qing-Long Zhang, Yu-Bin Yang, Sa-net: Shuffle attention for deep convolutional neural networks, in: ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2021, pp. 2235–2239.

[55] Chengcheng Wang, Wei He, Ying Nie, Jianyuan Guo, Chuanjian Liu, Yunhe Wang, Kai Han, Gold-yolo: efficient object detector via gather-and-distribute mechanism, Adv. Neural Inf. Proces. Syst. 36 (2024).

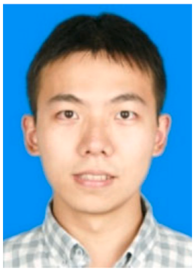
[56] Xu Ma, Xiyang Dai, Yue Bai, Yizhou Wang, Yun Fu, Rewrite the stars. *arXiv preprint arXiv:2403.19967*, 2024.

[57] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 618–626.

Wenhao Zhang received the B.S. degree from Anhui Agricultural University, Hefei, China, in 2023. He is currently pursuing a M.E. degree in Computer Science at Zhejiang Normal University, focusing on deep learning and computer vision, specifically in object detection.

Huiying Xu received the M.S. degree from National University of Defense Technology (NUDT), China. She is an associate professor with the School of Computer Science and Technology, Zhejiang Normal University, and also the researcher of Research Institute of Ningbo Cixing Co. Ltd., PR, China. Her research interests include Kernel learning and feature selection, Object Detection, Vision SLAM, Computer vision, Image processing, Pattern recognition, Computer simulation, Deep clustering, Generative Adversarial Network, Diffusion Model, Clustering Ensemble, Multiple Kernel Learning, Learning with incomplete data and their applications. She is a member of the China Computer Federation. She has published papers, including those in highly regarded journals such International Journal of Intelligent Systems, IEEE Transactions on Cybernetics, IEEE Transactions on Multimedia, etc.

Xinzhong Zhu received the Ph.D. degree from Xidian University and M.S. degree from National University of Defense Technology (NUDT), China. He is a professor with the School of Computer Science and Technology, Zhejiang Normal University, and also the chief scientist of Beijing Geekplus Technology Co., Ltd. and president of Research Institute of Ningbo Cixing Co., Ltd., China. His research interests include Machine learning, Deep clustering, Computer vision, Object detection, Segmentation, Recognition and Tracking, Diffusion Model, Manufacturing informatiza-tion, Manufacturing informatization, Robotics and System integration, Laser SLAM, Vision SLAM, Low Quality Data Learning, Multiple Kernel Learning, and Intelligent manufacturing. He is a member of the ACM and certified as CCF distinguished member. Dr. Zhu has published more than 30 peer-reviewed papers, including those in highly regarded journals and conferences such as the IEEE Transactionson Pattern Analysis and Machine Intelligence, the IEEE Transactionson Image Processing, the IEEE Transactions on Multimedia, the IEEE Transactions on Knowledge and Data Engineering, CVPR, NeurIPS, AAAI, IJCAI, etc. He served on the Technical Program Committees of IJCAI 2020 and AAAI 2020.



Yunzhong Si received B.E. degrees from Hunan University of Technology, China. He is currently pursuing an M.E. degree in electronic information from Zhejiang Normal University, China. His current research focuses are centered around deep learning, computer vision, and object detection.

Yao Dong received the B.S. degree from Ningbo Tech University, Ningbo, China, in 2023. He is currently pursuing a M.E. degree in Computer Science at Zhejiang Normal University, focusing on deep learning and computer vision, specifically in object detection.

Xiao Huang has received a PhD degree from East China Normal University. She is the Dean of the College of Education, the Joint Education Institute of Zhejiang Normal University and University of Kansas. She has worked as a professor in 2016, and also served as a PhD Supervisor and the Director of Science Education Research Center. She is the chief expert of Research Institute of Education Reform and Development in Zhejiang Philosophy and Social Sciences Key Cultivation Research Base, expert of international ISO standard TC/286/WG 4 for school-enterprise cooperation. Her research fields include STEM education, Nature of science and Scientific inquiry. She is the member of NARST(National Association for Research in Science Teaching), ESERA(European Science Education Research Association) and AAPT(American Association of Physics Teachers).

Hongbo Li received his Ph.D. degree in computer science from Tsinghua University in 2009. Currently, he holds the position of the Chief Technology Officer and Co-founder of Beijing Geek+ Technology Co., Ltd. China. In addition, he also serves as the secretary-general of Chinese Intelligent service Society and is an Editorial Board Member of several high-profile journals. His research interests include the design and application of intelligent robots, intelligent information process, and intelligent logistic systems. He has published more than 70 papers in prestigious journals and conference, and has been awarded more than 120 patents, including 46 international invention patents.