Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.082531

ARTICLE

VPCW-YOLO: An Improved YOLOv8 Algorithm for Vulnerable Pedestrian Detection under Complex Weather Conditions

Jian Su[1,2,*] **and Jiaqi Wang**[2]

[1]School of Information Science and Engineering (School of Cyber Science and Technology), Zhejiang Sci-Tech University, Hangzhou, China

[2]School of Computer Science, Nanjing University of Information Science and Technology, Nanjing, China

*Corresponding Author: Jian Su. Email: sj890718@gmail.com

Received: 17 March 2026; Accepted: 27 April 2026; Published: 15 June 2026

ABSTRACT: Despite significant advances in object detection technology, vulnerable pedestrian detection in intelligent transportation systems remains highly challenging under complex weather conditions. Environmental factors such as fog, rain, and snow often lead to occlusion, motion blur, and low-contrast images, making small-scale or weak-featured vulnerable pedestrians difficult to accurately identify. Therefore, improving the detection accuracy and robustness of vulnerable pedestrians in complex weather scenarios has become an urgent research problem. To address this issue, this paper proposes an improved YOLOv8-based vulnerable pedestrian complex weather detection algorithm, termed VPCW-YOLO. The proposed method enhances detection performance through multiple structural optimizations. First, a C2f_ST module is designed by integrating Spatial and Channel Reconstruction Convolution (SCConv) with a triplet attention mechanism to strengthen feature representation and improve the model's focus on critical regions. Second, a residual self-attention based (RSAB) module based on a self-attention mechanism is introduced to enhance global feature modeling capability under complex weather conditions. In addition, a Space-to-Depth operation is embedded into the backbone network to preserve more fine-grained information. Finally, a P2 small-object detection layer is added to improve the detection performance for distant and tiny pedestrians. Experimental results on the augmented BGVP dataset demonstrate that VPCW-YOLO achieves 73.0% mean Average Precision (mAP)@0.5 and 49.2% mAP@0.5:0.95, representing improvements of 5.4% and 5.7%, respectively, compared with the original YOLOv8 model. Generalization experiments on the Real-world Traffic Sign Detection in the Wild dataset (RTTS) pedestrian subset show that VPCW-YOLO achieves a 6.3% improvement in mAP@0.5. The results indicate that the proposed method effectively improves the detection accuracy and robustness of vulnerable pedestrians in complex weather scenarios while maintaining certain generalization capability, providing a promising solution for pedestrian safety perception in intelligent transportation systems.

KEYWORDS: Intelligent transportation systems; object detection; pedestrian detection; attention mechanism

1 Introduction

As global urbanization accelerates, an increasing number of people are migrating to cities, leading to a sharp rise in the number of vehicles. Consequently, urban transportation problems—such as traffic congestion and accidents—have become increasingly complex and severe. To address these challenges, intelligent transportation systems have emerged, aiming to enhance traffic management efficiency, reduce accidents, and improve traffic flow through advanced technological solutions.

Object detection, as one of the core technologies in intelligent transportation systems, enables the accurate identification of various traffic participants, including pedestrians, vehicles, and traffic signals. This significantly enhances traffic management and safety measures. Recent breakthroughs in artificial intelligence and deep learning has provided robust technical support for object detection. Deep learning-based object detection algorithms, such as YOLO [1] and R-CNN [2], have demonstrated outstanding performance and speed across various application scenarios, particularly in intelligent transportation systems. These technologies can process large volumes of traffic data in real-time, enabling fast and accurate detection of road users. As a result, they significantly reduce the risk of traffic accidents and improve overall transportation efficiency.

Pedestrians are the most vulnerable participants in transportation systems, making pedestrian detection a crucial technology in intelligent transportation systems. It has rapidly become an important area of research in traffic safety. Notably, in recent years, the proportion of vulnerable pedestrians injured in traffic accidents has been steadily increasing, bringing their safety into sharp focus. Vulnerable pedestrians include the elderly, children, and individuals with disabilities, who are more susceptible to accidents due to physical limitations and slower reaction speeds. Their behavioral patterns and safety requirements on urban roads differ significantly from those of ordinary pedestrians, making traditional pedestrian detection systems potentially ineffective in ensuring their safety. Furthermore, complex weather conditions such as heavy rain, haze, and snowfall exacerbate the challenges of detecting vulnerable pedestrians. For instance, heavy rain can cause camera images to become blurred, reducing image clarity, while haze and snowfall decrease visibility, making it difficult to accurately recognize pedestrian contours and features. These factors increase the complexity of pedestrian detection, reducing detection accuracy and reliability, leading to false positives and missed detections, and ultimately contributing to traffic accidents. Therefore, this paper presents a vulnerable pedestrian detection system tailored for autonomous driving in complex weather conditions. Given that object detection models in autonomous driving require both high accuracy and fast detection speeds, this study adopts YOLOv8 [3] as the baseline model for improvement.

To address the above issues, this study proposes an improved YOLOv8-based framework for detecting vulnerable pedestrians in complex weather conditions. Unlike existing studies that mainly focus on general pedestrian detection or directly apply existing modules, this work enhances YOLOv8 through task-specific improvements, optimizing feature representation, attention modeling, and small-object detection for vulnerable pedestrians under complex weather scenarios. The main contributions are summarized as follows:

1. A YOLOv8-based framework for detecting vulnerable pedestrians in complex weather scenarios, termed VPCW-YOLO, is developed to improve detection accuracy and robustness under adverse weather conditions.

2. To address the insufficient feature representation of pedestrian targets under complex weather conditions—often caused by occlusion, blur, and low contrast—an enhanced feature extraction module, C2f_ST, is designed by improving the original C2f module. This allows the network to more effectively capture key features of vulnerable pedestrian targets, significantly enhancing their feature representation capability.

3. To improve the model's perception of targets under complex weather noise, the RSAB module is introduced to model global contextual relationships, thereby enhancing the model's focus on vulnerable pedestrian regions.

4. To enhance the detection performance for small vulnerable pedestrians under complex weather conditions, a Space-to-Depth feature reorganization strategy is integrated into the YOLOv8 architecture,

along with an additional P2-level detection head, which preserves more shallow semantic information and improves small-object detection capability.

Experimental results demonstrate that the VPCW-YOLO model improved mAP@0.5 and mAP@0.5:0.95 by 5.4% and 5.7%, reaching 73.0% and 49.2%, respectively.

The remainder of this paper is organized as follows: Section 2 describes related work on object detection; Section 3 provides a detailed introduction to the VPCW-YOLO model; Section 4 presents the dataset, configuration, and evaluation metrics, followed by comparative experiments and ablation studies, along with an analysis of the experimental results; Finally, Section 5 concludes the paper.

2 Related Work

In the field of autonomous driving, pedestrian detection is a critical task that is primarily handled using either single-stage detectors or two-stage detectors. The core objective of this task is to accurately identify and localize pedestrians to enhance traffic safety. Two-stage detectors typically employ candidate region generation and classification-regression steps to ensure detection accuracy, whereas single-stage detectors extract pedestrian features directly from images to improve computational efficiency.

Two-stage object detectors follow a stepwise processing strategy, first generating candidate regions and then performing classification and location regression, leading to high detection accuracy. Early R-CNN [2] used the Selective Search method to generate candidate bounding boxes and conducted independent feature extraction and classification to achieve precise object detection. However, its high computational overhead made it unsuitable for real-time applications. Fast R-CNN [4] optimized computational efficiency by sharing the feature extraction process and employing RoI pooling to reduce redundant computations. Subsequently, Faster R-CNN [5] introduced the Region Proposal Network (RPN) to replace the Selective Search method, significantly improving the efficiency of candidate region generation and becoming one of the most widely used two-stage methods in deep learning-based object detection. Mask R-CNN [6] further extended Faster R-CNN by incorporating instance segmentation capabilities, enabling both object recognition and shape extraction, proving highly effective in tasks such as medical image analysis. Cascade R-CNN [7] employed a multi-stage classifier cascade strategy to progressively refine detection results, improving bounding box precision, making it particularly suitable for high-quality object detection. Additionally, Sparse R-CNN [8] leveraged sparse sampling of candidate regions to reduce computational costs while maintaining high detection accuracy, offering better adaptability in resource-constrained environments. However, despite their high detection accuracy and stability in complex scenarios, two-stage detectors suffer from high computational costs and slow inference speeds, making them unsuitable for real-time processing. In applications that require efficient object detection, such as autonomous driving and intelligent surveillance, single-stage detectors present a more viable choice. Therefore, this study opts for a single-stage detector to balance detection speed and accuracy while further optimizing the object detection task.

Single-stage detectors complete object detection tasks through a single forward pass. Compared to two-stage detectors, they offer higher computational efficiency and are well-suited for applications requiring real-time performance. SSD [9] predicts using multi-scale feature maps, balancing detection speed and accuracy to some extent; however, it struggles with small-object detection and low-resolution images. RetinaNet [10] leverages Focal Loss to address the challenge of disproportionate positive and negative sample distribution, improving detection performance for small and hard-to-detect objects, but its inference speed is relatively slow, compromising its applicability in scenarios demanding rigorous real-time performance. EfficientDet [11] adopts EfficientNet as the backbone network and integrates bi-directional feature pyramids for multi-scale feature fusion, achieving strong detection performance at the cost of increased computational

complexity. In contrast, the You Only Look Once (YOLO) series, a representative single-stage detection algorithm, employs an end-to-end detection approach alongside efficient network optimizations, continuously improving detection speed and accuracy. YOLOv3 [12] relies on Darknet-53 and a multi-scale prediction strategy, effectively enhancing small-object detection capability. YOLOv4 [13] introduces the CSPDarknet53 backbone network and the Spatial Pyramid Pooling (SPP) module, significantly improving detection accuracy. YOLOv5 [14] incorporates a more flexible network architecture and automated anchor optimization, simplifying the training process while enhancing detection efficiency. YOLOv7 [15] further advances feature extraction and object detection accuracy by integrating the Efficient Layer Aggregation Network (ELAN) module. YOLOv8 builds upon the strengths of its predecessors by optimizing network structure and loss function design, achieving high-speed inference while greatly improving detection robustness and result stability, meeting the demands for both real-time processing and high accuracy. Therefore, considering detection precision, computational efficiency, and real-time performance requirements, this study selects YOLOv8 as the baseline model for further improvements.

Current research on vulnerable pedestrian detection primarily relies on behavioral features to identify vulnerable individuals. For instance, Song et al. [16] proposed a video stream–driven motion analysis module to protect vulnerable pedestrians, while Liu et al. [17] introduced a pedestrian-oriented forewarning system (POFS) aimed at protecting individuals distracted by smartphone use, classifying them as vulnerable pedestrians. Ni et al. [18] constructed a thermal imaging dataset for individuals with mobility impairments who use assistive devices or carry mobility burdens. While these methods have made progress in identifying pedestrians who are vulnerable due to behavioral factors, they often fail to adequately account for groups such as the elderly, children, and individuals with disabilities—people who are at a disadvantage in traffic environments due to physiological or cognitive limitations. This limitation is common in current studies. Additionally, most existing studies focus on pedestrian detection under normal weather conditions. For instance, Xie et al. [19] proposed an improved algorithm, YOLO-ACE, based on the YOLOv10 framework. By introducing the additive-convolutional gated linear unit (Add-CGLU) structure to replace the C2f module, designing the feature pyramid shared conv (FPSC) module to optimize the SPPF structure, and constructing the Efficient multibranch scale (EMBS) multi-branch scale feature pyramid network, the model is further optimized with a dual distillation strategy. These improvements effectively reduce the number of parameters and computational cost while enhancing the accuracy and speed of vehicle and pedestrian detection. Chen et al. [20] addressed the limited detection accuracy of vehicles and pedestrians in infrared images by proposing an Adaptive Feature Manipulation Network. This method introduces a Fine Spatial Pooling module, a Shuffle Feature Manipulation module, and a multi-scale adaptive connection mechanism to achieve effective fusion of features across different scales and channels, thereby improving the detection accuracy of vehicles and pedestrians while maintaining a lightweight structure and high detection speed. Ge et al. [21] proposed a Lightweight and Efficient Pedestrian Detection Network. By designing a PoolFormer-based lightweight detection head to reduce computational cost and incorporating a Tri-branch Joint Attention Module to enhance global semantic representation and spatial dependency modeling, the method enables accurate detection of small and occluded pedestrians, providing an efficient solution for pedestrian detection in resource-constrained environments. Wang and Xie [22] proposed the FCD-Net framework, which enhances feature diversity through feature decorrelation loss and achieves adaptive feature integration via confidence-driven dynamic fusion. In addition, a score–accuracy mapping and a distribution calibration module are introduced to ensure the reliability of confidence estimation. These designs improve detection performance while maintaining real-time capability and effectively reduce missed detections and false positives in real driving scenarios.

Notably, some studies have also explored methods specifically designed for adverse weather conditions, such as improving the quality of blurred or low-contrast images through image restoration or enhancing model generalization under different weather conditions via domain adaptation. An et al. [23] proposed a Mixture-of-Experts based unified image restoration framework, MUIRF, which achieves unified restoration of degraded images under various adverse weather conditions through a channel-level parameter sharing strategy and a meta-vector guided gradient homogenization algorithm. Xu et al. [24] analyzed the differential impact of adverse weather conditions on different scene regions and proposed an image restoration method that integrates weather priors with a unified imaging model, further enhancing scene recovery under diverse weather conditions through a weather-aware cross-attention module. Li et al. [25] proposed a domain-adaptive object detection framework for foggy and rainy scenarios, which reduces cross-domain discrepancies under different weather conditions through image-level and object-level feature alignment along with an adversarial gradient reversal mechanism, thereby improving object detection performance.

However, existing studies still face many challenges in pedestrian detection under complex weather conditions. Therefore, this paper focuses on the problem of vulnerable pedestrian detection in complex weather environments, aiming to improve the detection performance of models for vulnerable pedestrians under adverse weather conditions.

3 Methods

3.1 Improved Overall Structure

VPCW-YOLO is composed of three main components: a backbone for extracting features, a neck for fusing those features, and four detection heads dedicated to producing the final outputs. As shown in Fig. 1, the input 640 × 640 × 3 RGB image is processed through convolutional layers, the Space-to-Depth module, the C2f module, and the Spatial Pyramid Pooling Fast (SPPF) layer in the backbone network, gradually extracting four feature maps with diverse scales: P2, P3, P4, and P5. Among them, the Space-to-Depth operation helps preserve fine-grained detail information in the image. In the neck network, the RSAB module is employed to capture global contextual information and long-range dependencies, enhancing the distinguishability of target features. Additionally, the C2f_ST feature fusion module effectively captures the key features of the target. Finally, the four detection heads output the detection results corresponding to different scales.

3.2 C2f_ST

Under complex weather conditions, images often suffer from occlusion, blur, and low contrast, which makes the features of vulnerable pedestrians less prominent and poses significant challenges for detection. To enhance the model's feature extraction and perception capabilities in such extreme environments, this paper proposes an improved feature extraction module, C2f_ST, to replace the original C2f module in YOLOv8. As illustrated in the Fig. 2, the C2f_ST module includes two core improvements:

First, the standard Bottleneck structure in the original C2f module is replaced with a Bottleneck_SCConv module. Each Bottleneck_SCConv unit consists of convolution, SCConv [26], BatchNorm, and Gaussian Error Linear Unit (GELU). This module reconstructs both spatial and channel information in the input feature maps, reducing redundant interference and enhancing feature representation capability. The SCConv module is composed of a Spatial Reconstruction Unit (SRU) and a Channel Reconstruction Unit (CRU), which are designed to optimize spatial and channel features, respectively. The SRU employs a separate-and-reconstruct strategy, using scaling factors from Group Normalization to evaluate the information content of the feature maps. It separates high-information maps from low-information ones and

then fuses them via cross-reconstruction to generate higher-information features while saving computational resources. The CRU adopts a split-transform-fuse strategy. Firstly, the input spatially refined features are divided into two parts and compressed using 1 × 1 convolutions to obtain Xup and Xlow. Xup is then used for rich feature extraction, where it passes through Group-wise Convolution (GWC) and Point-wise Convolution (PWC), and then sum the results to obtain the output Y1. Xlow serves as a supplementary input for rich feature extraction and is processed by PWC; the result is merged with the original input to form Y2. At last, a simplified Selective Kernel Network (SKNet)-based method is used to fuse Y1 and Y2, enhancing the channel-wise feature representation. This improvement not only reduces interference from redundant features but also enhances the model's representational capacity and computational efficiency, enabling the network to more accurately capture key features and improve overall performance.

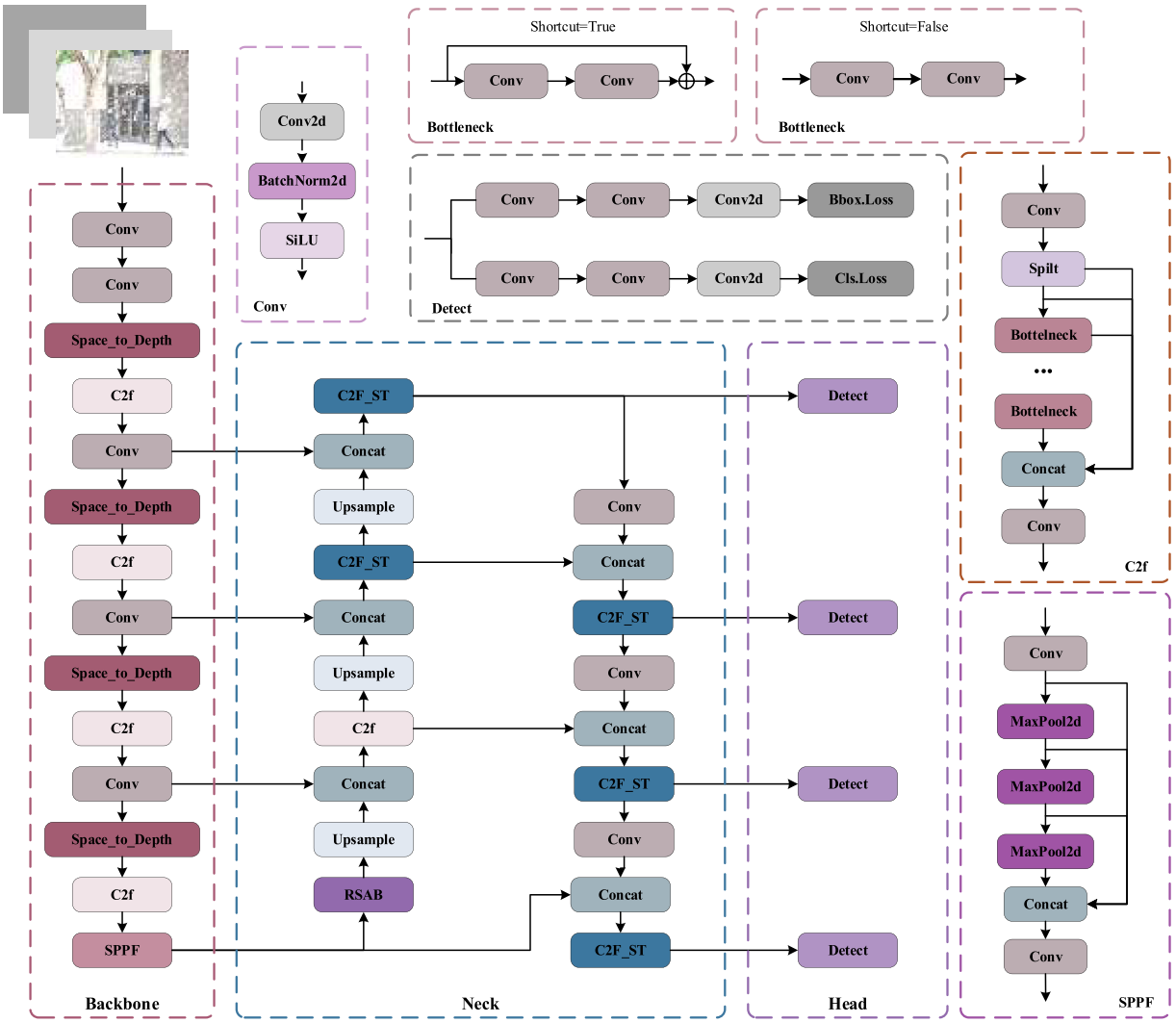

Figure 1: Network architecture diagram of VPCW-YOLO.

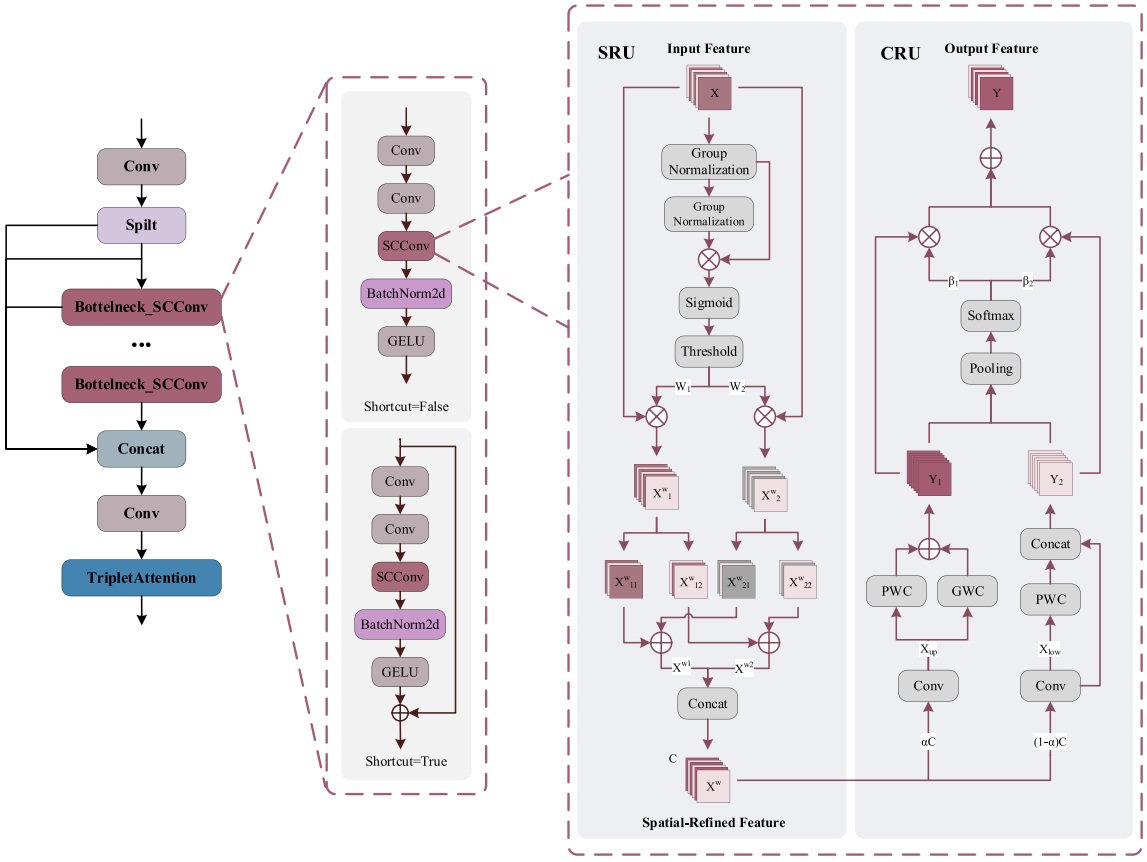

Figure 2: C2f_ST module structure diagram.

Secondly, a Triplet Attention mechanism [27] is introduced at the end of this module to further emphasize target regions and empower the model to learn richer features. This mechanism adopts a three-branch structure that computes attention weights through interactions among the channel (C), height (H), and width (W) dimensions, thereby significantly enhancing feature representation while maintaining computational efficiency. Compared to traditional single-channel or dual-channel attention mechanisms, Triplet Attention not only focuses on independent features across each dimension but also fully exploits their interdependencies, allowing the model to understand target regions more comprehensively and accurately. Specifically, the module first performs a rotation operation on the input tensor, enabling interactive mapping of information across different dimensions, thereby enriching the spatial features. Then, a residual transformation is applied to further optimize the information flow, allowing the model to efficiently learn dependencies among different dimensions. This approach ensures that the attention mechanism is not limited to local regions but can effectively integrate information across H-C, C-W, and H-W dimensions, improving the effectiveness of feature fusion. As illustrated in Fig. 3, the first branch captures interactions between H and C dimensions, the second between C and W dimensions, and the third between H and W dimensions. After generating attention weights, each branch performs a permutation operation on the input, and the outputs of the three branches are averaged to obtain the final triplet attention output. This approach further suppresses redundant background interference, highlights key target regions, and significantly enhances the model's ability to detect subtle targets in complex backgrounds.

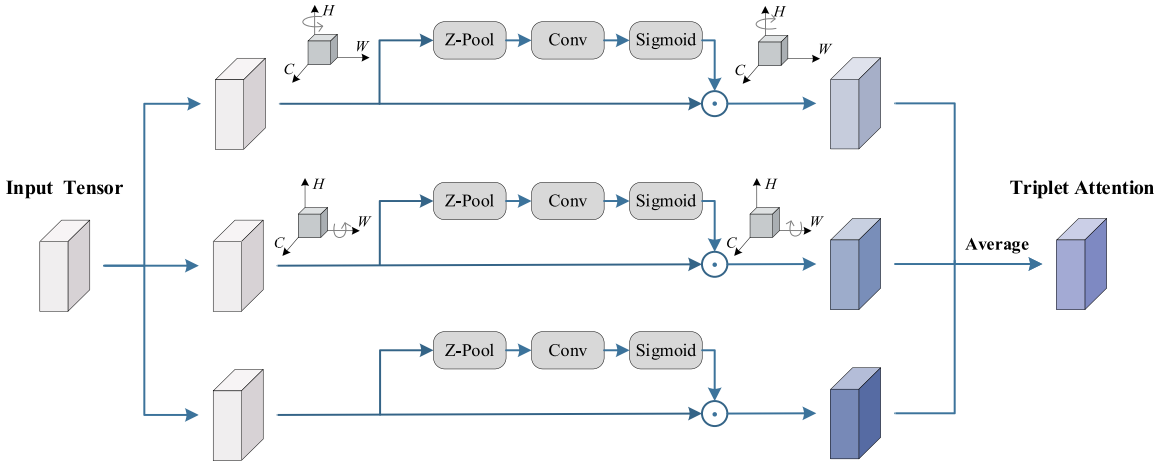

Figure 3: Triplet attention mechanism structure diagram.

In summary, the C2f_ST module enhances feature representation and reduces redundant information through SCConv, while incorporating a triplet attention mechanism to strengthen the perception of target regions. This significantly improves the model's ability to represent low-quality vulnerable pedestrian targets. Under complex weather conditions, the module effectively mitigates recognition challenges caused by feature blurring and background interference.

3.3 RSAB

Pedestrian detection under complex weather conditions poses significant challenges. Due to environmental factors, images often contain severe interference from mixed targets and weather noise, as well as issues such as target blurring and occlusion. These problems are particularly detrimental to the detection of vulnerable pedestrians who are distant, small in size, or structurally incomplete. In such scenarios, traditional object detection models often struggle to accurately localize and identify these critical targets, severely limiting both detection accuracy and robustness.

In order to deal with this problem, we embed an improved attention module, RSAB, into YOLOv8. Although the original Residual Attention Block (RAB) [28] enhances feature representation to some extent, its spatial attention module (SAM) still suffers from clear shortcomings. Conventional spatial attention relies on local, convolution-based weighting within a fixed receptive field and cannot effectively capture long-range dependencies between distant pixels. When pedestrian targets lie far from the main region, are occluded, or are heavily blurred, this spatial attention often fails to deliver sufficient global semantic support, resulting in weak key target responses and reduced feature extraction capability.

Accordingly, we replace the SAM in RAB with a self-attention mechanism [29]. As shown in Fig. 4, the enhanced module retains RAB's multi-branch residual structure, stacking multiple convolution and Rectified Linear Unit (ReLU) units to harvest abundant local features while employing self-attention to model global spatial relationships. By constructing Query–Key–Value triplets, self-attention enables information exchange among arbitrary positions, allowing the model to capture semantic correlations between distant targets and highlight the importance of target regions under complex background interference. As illustrated in Fig. 5, the self-attention mechanism is implemented through the following steps:

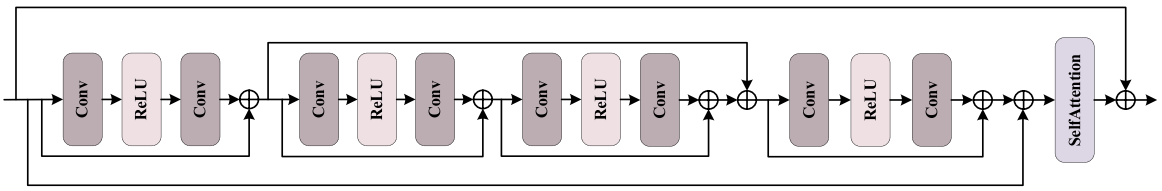

Figure 4: RSAB module structure diagram.

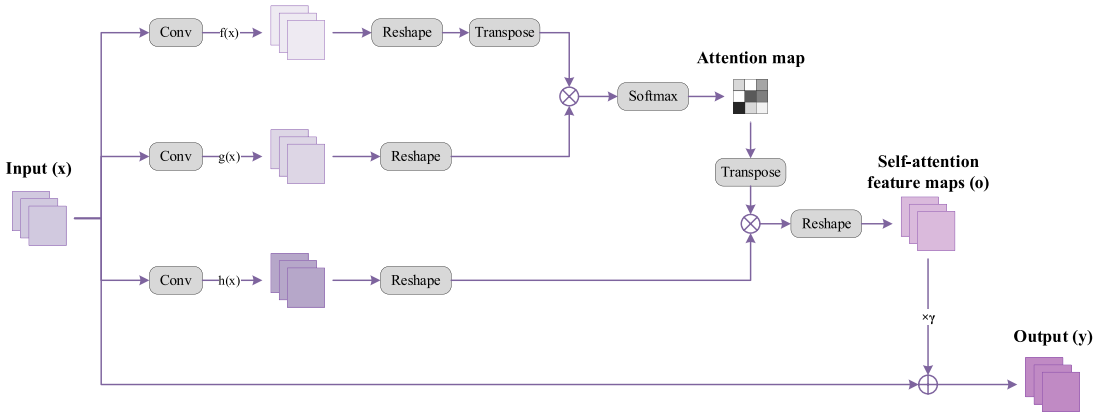

Figure 5: Self-attention mechanism structure diagram.

Assume the input feature is $x \in R^{C\times N}$, where C is the number of channels and N is the number of feature positions. The input is first projected into two feature spaces:

$$\mathrm{f}(x) = W_f x \tag{1}$$

$$\mathrm{g}(x) = W_g x \tag{2}$$

which are used to calculate correlations between positions and further construct the attention weight matrix:

$$\beta_{j,i} = \frac{\exp\left(s_{ij}\right)}{\sum_{i=1}^{N}\exp\left(s_{ij}\right)} \tag{3}$$

$$S_{\mathrm{ij}} = f\left(x_i\right)^T g\left(x_j\right) \tag{4}$$

where $\beta_{j,i}$ represents the degree to which the model focuses on input position i when synthesizing output position j. Then, the corresponding self-attention output features can be expressed as:

$$o_{\mathrm{j}} = \sum_{i=1}^{N}\beta_{j,i}h\left(x_i\right) \tag{5}$$

$$\mathrm{h}(x) = W_h x \tag{6}$$

where the weight matrixs W_f, W_g and W_h are implemented using a 1×1 convolution.

To stabilize training and enhance representational capacity, a learnable scaling factor γ is introduced and combined with the input via residual connection. The final output is given by:

$$\mathrm{y}_i = \gamma o_i + x_i \tag{7}$$

where γ is initialized to 0, allowing the model to initially rely more on the original features, and gradually learn optimized attention allocation strategy during training.

This improvement not only significantly enhances the perceptibility of vulnerable pedestrians but also strengthens the model's robustness against blurred and occluded targets. Meanwhile, the self-attention mechanism, with its larger receptive field and stronger feature representation capabilities, empowers the model to capture semantic connections between distant pixels, improving the recognizability and detection accuracy of vulnerable pedestrians under complex weather conditions. It is particularly stable in low-resolution or partially occluded scenarios, thereby effectively boosting the overall detection performance of the model in complex weather environments.

3.4 Space-to-Depth

In vulnerable pedestrian detection under complex weather conditions, low resolution and small object size are key factors affecting detection performance. The original YOLOv8 mainly uses stride convolutions (stride ≥ 2) and pooling operations in the backbone for downsampling. While this helps accelerate the downsampling process, it also significantly compromises the integrity of shallow features, leading to the loss of fine-grained information. This issue is especially critical for small objects, whose key structural details may be weakened or lost early in the network, negatively impacting detection performance.

To tackle the above issue, this study adopts a Space-to-Depth-based downsampling strategy in the YOLOv8 backbone. The stride of the original convolutions in the backbone is adjusted from 2 to 1 to reduce information loss during the feature compression process. At the same time, the traditional downsampling approach is replaced with the Space-to-Depth technique [30]. This method transforms spatial information into the channel dimension, allowing details in the original image to be preserved and further exploited by deeper network layers, thereby effectively enhancing the perception of small objects.

To illustrate the working mechanism of the Space-to-Depth, let the input feature map be denoted as $X \in R^{S\times S\times C_1}$, where S represents the spatial dimensions and C denotes the channel count. The Space-to-Depth slices the input feature map into multiple sub-regions along the spatial dimensions and rearranges these sub-regions into the channel dimension. Given a partition factor of scale, the operation produces $scale^2$ sub-feature mapps

$$f_{i,j} = X\left[i{:}\,S{:}\,scale, j{:}\,S{:}\,scale\right], 0 \leq i, j < scale \tag{8}$$

each sub-feature map $f_{i,j} \in R^{\frac{S}{scale}\times\frac{S}{scale}\times C}$ represents the sub-regions obtained by interval sampling from the original feature map. These sub-regions are then concatenated along the channel dimension to generate the output feature map:

$$X' = Concat_{channel}\left(f_{0,0}, f_{0,1}, \cdots, f_{scale-1,scale-1}\right) \in R^{\frac{S}{scale}\times\frac{S}{scale}\times scale^2\cdot C} \tag{9}$$

When $scale = 2$, the Space-to-Depth operation divides the original feature map into four sub-maps: $f_{0,0}$, $f_{0,1}$, $f_{1,0}$, $f_{1,1}$, as illustrated in Fig. 6. The output feature map is compressed to 50% of its original spatial dimensions, while the number of channels is expanded to four times the original.

This strategy enhances the model's ability to perceive small objects while preserving critical detail information, thereby reducing false detections caused by shallow feature compression and significantly improving detection performance in complex weather conditions.

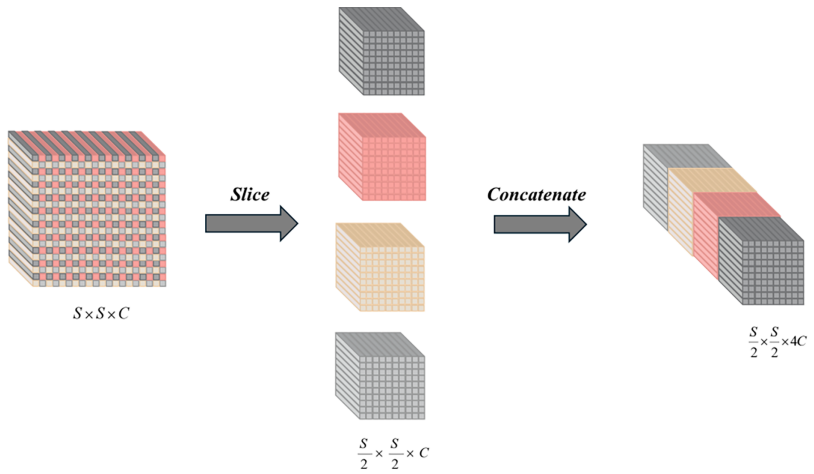

Figure 6: Workflow diagram of space-to-depth when scale = 2.

3.5 Small Object Detection Layer

Vulnerable pedestrian detection under complex weather conditions faces numerous challenges, among which the limited ability to detect distant and small-scale targets is a key issue. These targets often occupy only a tiny portion of the image and are highly susceptible to low contrast, uneven lighting, and occlusion, leading to insufficient feature representation and ultimately reduced detection accuracy.

To enhance the model's perception of small vulnerable pedestrians, this paper introduces an additional P2 detection layer into the YOLOv8 architecture, as shown in the Fig. 7. This layer is designed to process higher-resolution features with a smaller receptive field. Compared to the shallowest detection layer P3 in the original YOLOv8, the P2 detection layer further preserves more spatial detail information, enabling more effective extraction of structural features of weak and small targets and enhancing the model's ability to distinguish low-scale targets. By incorporating the P2 layer, the model can exploit high-resolution feature maps to extract richer detail, thus mitigating the challenges of difficult target detection and weak feature representation to a certain extent.

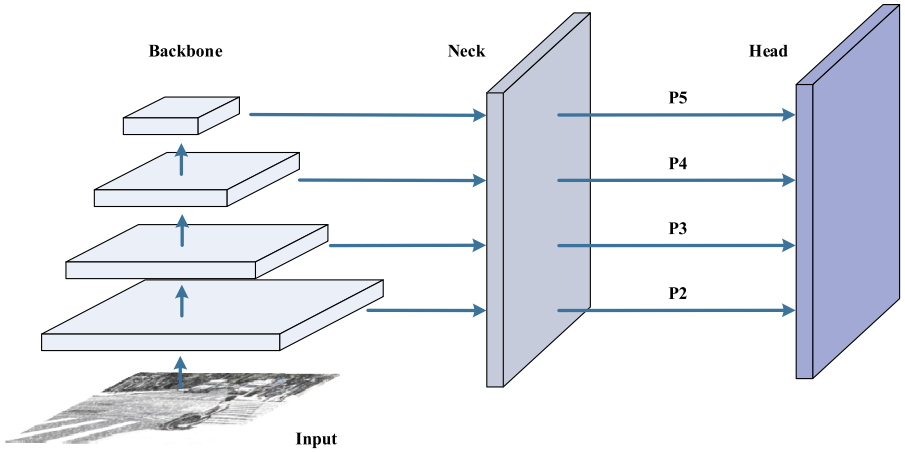

Figure 7: Detection network architecture diagram using the P2 detection layer.

4 Experiments

4.1 Data Set and Configuration

Due to the extremely limited availability of real-world data resources for vulnerable pedestrian detection in complex weather scenarios, we constructed an augmented dataset suitable for adverse weather detection tasks based on the Bowling Green Vulnerable Pedestrian (BGVP) dataset [31] released in 2024. The BGVP dataset focuses on vulnerable pedestrian detection and covers four categories of people: (1) Children without disability: aged 1–16, considered vulnerable pedestrians due to unpredictable behavior and limited understanding of traffic rules. (2) Elderly without disability: aged 50 and above, without disabilities but may experience mobility challenges due to age-related factors. (3) With disability: pedestrians of all ages with physical disabilities who may require wheelchairs, walkers, or other assistive devices. (4) Non-vulnerable: those not belonging to the above categories and less likely to be affected by traffic conditions compared to other groups.

To better reflect the visual perception challenges under adverse weather conditions in practical applications, we introduced multiple adverse weather effects into the original BGVP dataset, creating a dataset with greater weather diversity. Specifically, the rain effect is generated using the imgaug library by setting the raindrop size to 0.4–0.5 and the rain speed to 0.1–0.2. The snow effect is also produced using imgaug with a moderate intensity (severity = 2) to simulate snowflake distribution. The fog effect is simulated using an exponential attenuation function based on the pixel distance to the image center, expressed as $t_d = \exp(\beta \times d)$ ($\beta = 0.08$), which is linearly blended with a brightness coefficient B = 0.8. The sandstorm effect is generated by combining the cloud-like fog effect provided by imgaug with a yellow color overlay, where the fog intensity is set to 1, the yellow tint strength is set to 0.6, and the global brightness shift is set to 0.1. This dataset not only includes images from the original dataset but also simulates four typical adverse weather scenarios—rain, fog, snow, and sandstorms—through image augmentation techniques, as shown in Fig. 8, thereby effectively enhancing the diversity and challenge of the data.

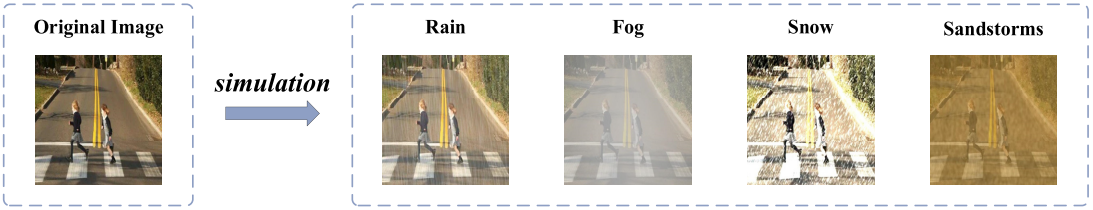

Figure 8: Weather condition simulation including rain, fog, snow, and sandstorms.

The dataset consists of 9929 images and includes 29,409 bounding box annotations. Specifically, there are 8178 bounding box annotations for the "children without disability" category, 4046 instances for the "elderly without disability" category, 4677 instances for the "with disability" category, and 12,508 instances for the "non-vulnerable" category. The dataset is divided into three subsets: a training set with 7017 images, a validation set with 1932 images, and a test set with 980 images. This dataset combines original clear-weather images with synthesized adverse-weather images, covering a wide range of scenes and lighting conditions, significantly enhancing the representativeness of the samples. During the model training phase, we used this dataset to optimize performance across different climatic scenarios; in the testing and evaluation phase, assessments were conducted on the test subset of this dataset to comprehensively evaluate the model's ability to detect vulnerable pedestrians under complex weather conditions.

The experiments were conducted on a computer equipped with an NVIDIA RTX 4090 GPU and an Intel(R) Xeon(R) Platinum 8352V CPU. The programming language used was Python 3.8.19, and the CUDA

version was 11.1. The input image size for training and testing was 640 × 640. The entire network was optimized using the Stochastic Gradient Descent (SGD) optimizer with an initial learning rate of 0.01. The batch size was set to 64, and the number of training epochs was 200.

4.2 Evaluation Metrics of the Algorithm

To comprehensively evaluate the performance of VPCW-YOLO in detecting vulnerable pedestrians under complex weather conditions, this paper analyzes the model using several widely adopted evaluation metrics, including precision, recall, mAP and parameters.

Among them, precision measures how many of the samples predicted as positive are truly positive. Its calculation formula is:

$$precision = \frac{TP}{TP + FP} \times 100\% \tag{10}$$

where, TP denotes correctly detected positive samples, whereas FP represents negative samples that the model mistakenly identified as positive.

Recall quantifies the model's ability to correctly retrieve true positive cases from all actual positive samples. Its calculation formula is:

$$recall = \frac{TP}{TP + FN} \times 100\% \tag{11}$$

where, FN represents the count of actual targets that are missed by the model.

To more intuitively reflect the overall detection precision of the model, this paper uses mAP for evaluation. The calculation formula is as follows:

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

where AP represents the average precision for a single category.

In addition, parameters denote the overall quantity of trainable weights in the model—fewer parameters results in a more lightweight model.

4.3 Experimental Results

To verify the effectiveness of VPCW-YOLO for vulnerable pedestrian detection under complex weather conditions, we selected multiple representative object detection models for comparison on the augmented BGVP dataset. These models include Faster R-CNN, Cascade R-CNN, YOLOv6n, YOLOv7-tiny, YOLOv8n, YOLOv8s, YOLOv9s, YOLOv10s, and YOLOv11s. All models were trained and evaluated under identical training environments and parameter settings to ensure the fairness and reproducibility of the comparison.

As shown in Table 1, the proposed VPCW-YOLO achieves a detection accuracy of 73.0% in terms of mAP@0.5, which is the highest among all compared models. It significantly outperforms YOLOv6n (67.4%), YOLOv7-tiny (67.1%), YOLOv8n (67.6%), YOLOv8s (69.1%), YOLOv9s (69.9%), YOLOv10s (66.5%), and YOLOv11s (71.1%), with improvements of 5.6%, 5.9%, 5.4%, 3.9%, 3.1%, 6.5%, and 1.9%, respectively. In addition, compared with the two-stage detectors Faster R-CNN and Cascade R-CNN, the proposed method improves the detection accuracy by 3.1% and 5.5%, respectively. In terms of computational complexity, VPCW-YOLO has 19.5 Floating Point Operations (FLOPs) and only 4.33M parameters, maintaining a relatively low model complexity. Compared with high-complexity models such as Faster R-CNN and Cascade R-CNN, the proposed method requires significantly fewer floating-point operations. Although the number

of parameters slightly increases compared with YOLOv8n, the proposed model improves the mAP@0.5 by 5.4 percentage points, reaching 73.0%. Meanwhile, its mAP@0.5:0.95 reaches 49.2%, which is significantly higher than that of YOLOv8s and other models with higher computational complexity, achieving a better balance between detection accuracy and computational efficiency. In terms of inference speed, under the same experimental settings, VPCW-YOLO achieves 126 Frames Per Second (FPS), compared with 178 FPS for YOLOv8n and 120 FPS for YOLOv8s. Although the inference speed is lower than that of YOLOv8n, it remains comparable to YOLOv8s and still satisfies real-time detection requirements. This indicates that the proposed method maintains competitive real-time performance while significantly improving detection accuracy over baseline models. Furthermore, the complexity of the proposed model is significantly lower than that of other models with comparable accuracy. For example, although YOLOv11s achieves relatively high detection accuracy, it requires higher computational cost, while the number of parameters of Cascade R-CNN is as high as 69.16M. In contrast, VPCW-YOLO not only achieves the highest detection accuracy but also effectively controls model size and computational cost, demonstrating stronger practicality and deployment potential.

Table 1: Performance comparison of object detection models on the augmented BGVP dataset.

Model	Precision (%)	Recall (%)	mAP@0.5 (%)	mAP@0.5:0.95 (%)	Params (M)	FLOPs (G)	FPS (Frame/s)
Faster R-CNN	55.3	54.0	69.9	46.6	41.36	41.4	21
Casacade R-CNN	63.7	55.7	67.5	46.9	69.16	69.2	13
YOLOv6n [32]	66.8	66.8	67.4	44.3	4.23	11.8	141
YOLOv7-tiny	65.5	64.5	67.1	41.8	6.23	6.9	169
YOLOv8n	68.8	64.7	67.6	43.5	3.01	8.2	178
YOLOv8s	65.5	69.8	69.1	45.6	11.13	28.4	120
YOLOv9s [33]	70.9	66.0	69.9	46.8	7.17	26.7	123
YOLOv10s [34]	71.0	60.9	66.5	44.6	8.04	24.5	128
YOLOv11s [35]	67.7	70.1	71.1	47.3	9.41	21.3	132
VPCW-YOLO(Ours)	69.7	69.6	73.0	49.2	4.33	19.5	126

We selected three representative image samples from the enhanced BGVP dataset, covering typical scenarios such as complex backgrounds, object occlusion, small targets, and weather-induced noise interference. We compared the detection performance of the pre-improvement model (a) and its improved version (b), as shown in Fig. 9.

In the first group of images, the targets exhibit significant mutual occlusion, and the background is relatively complex. The unimproved model was only able to detect three “children without disability” who were either unoccluded or only slightly occluded, failing to identify other heavily occluded targets, resulting in a high rate of missed detections. In contrast, the improved model accurately detected all foreground targets, indicating stronger robustness in handling occlusion and background interference. The second group mainly consists of small, distant targets, posing challenges such as small scale and weak information. Under heavy snow noise interference, the original model identified only a few pedestrian targets, while the improved model successfully detected more small-scale pedestrian targets, showcasing enhanced perceptual ability in handling distant small objects. The third group simulates a sandstorm scenario with significant noise interference and overall low image contrast. The original model misclassified two “non-vulnerable” targets

as “elderly without disability” and “children without disability,” while also missing several targets. After the improvement, the model not only recognized all targets in the images but also maintained extremely high detection accuracy under low-quality image conditions, demonstrating superior anti-interference capability.

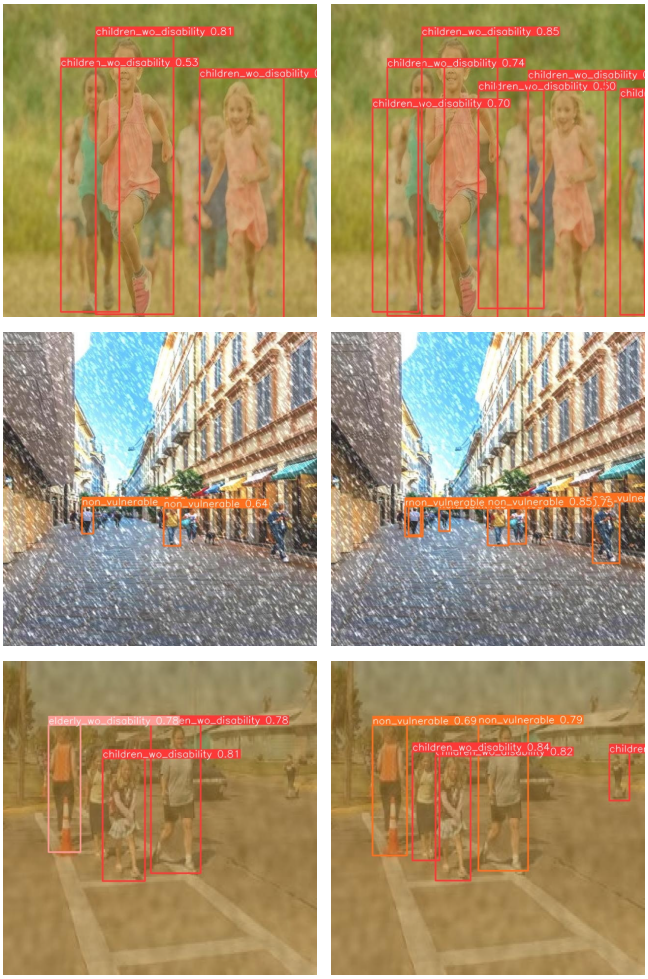

Figure 9: Detection of vulnerable pedestrians under complex weather conditions. (**a**) Before enhancement; (**b**) After enhancement.

To more intuitively demonstrate the effectiveness of the model under complex weather conditions, we introduced attention heatmaps for visual analysis. As shown in Fig. 10, the corresponding attention heatmaps for the three sample groups are provided, where a denotes the response of the model before improvement, and b corresponds to the improved version. From the visualized attention maps, it can be observed that in the first group, the attention distribution in (a) is relatively scattered, with some heat regions focusing on background areas, leading to insufficient response to occluded targets. In contrast, (b) shows more concentrated attention on human regions, significantly enhancing the model’s ability to focus on foreground pedestrians. In the second group, the baseline model shows weak responses to distant small-scale objects, with under-activated attention regions, while the improved model exhibits clearly concentrated heat on critical regions. In the third group, the heatmap from the baseline model produces a large number of irrelevant responses in noisy

areas of the image, indicating a high risk of false detection; conversely, the improved model demonstrates more focused attention on true target regions, with non-target responses effectively suppressed.

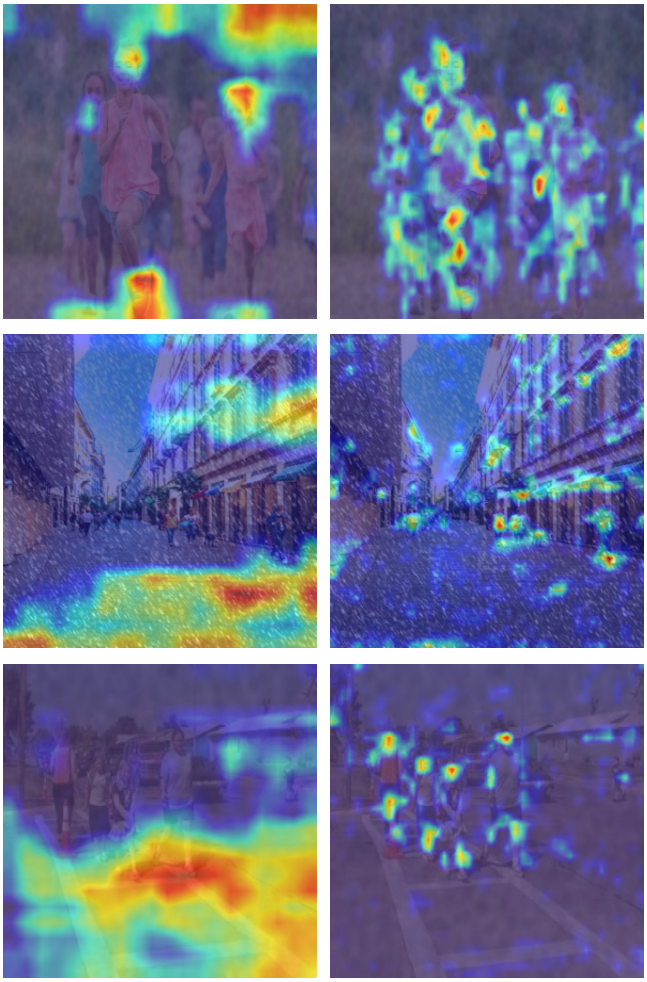

Figure 10: Visualization of attention heatmaps for detecting vulnerable pedestrians under complex weather scenarios. (**a**) Before enhancement; (**b**) After enhancement.

To further verify the effectiveness of VPCW-YOLO under various complex weather conditions, experiments were conducted to evaluate the detection performance of the model in four types of weather scenarios: fog, snow, rain, and sandstorm. As shown in Table 2, VPCW-YOLO outperforms YOLOv8n in all four complex weather conditions. Specifically, the mAP@0.5 increases by 4.5%, 5.9%, 7.0%, and 6.7% in foggy, snowy, rainy, and sandstorm scenarios, respectively. These results indicate that the proposed method has stronger anti-interference capability and feature representation ability under complex weather conditions, achieving stable performance improvements.

To verify the cross-dataset generalization ability of the proposed model, this study constructs a pedestrian subset on the RTTS dataset, which is designed for real-world complex weather scenarios. Specifically, images containing pedestrian targets are selected from the original dataset, and only the pedestrian category annotations are retained while other category information is ignored. In this way, a single-class pedestrian detection dataset is formed to evaluate the model's generalization performance in new scenarios. The

experimental results are shown in Table 3. It can be observed that the proposed method still maintains good detection performance on this dataset. Compared with the original YOLOv8n model, the proposed VPCW-YOLO achieves significant improvements across all evaluation metrics. Although the RTTS dataset differs significantly from the training data in terms of weather conditions, imaging quality, and target characteristics, the proposed method still outperforms the baseline model, which demonstrates that the proposed model has strong cross-dataset generalization capability in complex weather scenarios.

Table 2: Performance comparison of different object detection models under various weather conditions on the enhanced BGVP dataset.

Model	Foggy mAP@0.5 (%)	Snowy mAP@0.5 (%)	Rainy mAP@0.5 (%)	Sandstorm mAP@0.5 (%)
Yolov8n	70.3	63.1	68.5	67.6
VPCW-YOLO (Ours)	74.8	69.0	75.5	74.3

Table 3: Generalization experiment on the pedestrian subset of the RTTS dataset.

Model	Precision (%)	Recall (%)	mAP@0.5 (%)	mAP@0.5:0.95 (%)
Yolov8n	76.4	51.6	57.2	30.0
VPCW-YOLO(Ours)	78.6	56.3	63.5	35.6

4.4 Ablation Experiments

To validate the performance gains brought by the proposed modules in detecting vulnerable pedestrians under complex weather conditions, ablation experiments were conducted by progressively integrating each component into the YOLOv8 framework. The experiments evaluated AP across different pedestrian categories, including "children without disability," "elderly without disability," "with disability," and "non-vulnerable," using mAP@0.5 as the overall performance metric. The results are summarized in Table 4.

Table 4: Ablation experiments on the enhanced BGVP dataset.

Method	AP (Children without Disability) (%)	AP (Elderly without Disability) (%)	AP (With Disability) (%)	AP (Non-Vulnerable) (%)	mAP (%)
YOLOv8	74.5	51.8	87.6	56.5	67.6
YOLOv8 + C2f_ST	73.3	54.6	88.0	60.8	69.2
YOLOv8 + C2f_ST + Space-to-Depth	76.8	54.6	90.1	60.5	70.5
YOLOv8 + C2f_ST + Space-to-Depth+P2	78.8	55.3	89.5	63.8	71.9
YOLOv8 + C2f_ST + Space-to-Depth + P2 + RSAB	78.4	60.1	91.9	61.9	73.0

Firstly, replacing the original C2f module in YOLOv8 with the C2f_ST module increased mAP@0.5 to 69.2, indicating enhanced perception of vulnerable pedestrians. Subsequently, incorporating the Space-to-Depth strategy further raised mAP@0.5 to 70.5, with notable improvements of 2.1 and 3.5 percentage points in AP for "with disabilities" and "children without disability," respectively, validating the effectiveness of the Space-to-Depth design. Building on this, the addition of a P2 small-object detection layer led to a further increase in mAP@0.5 to 71.9, with the most significant AP gain of +3.3 observed for the "non-vulnerable" group. However, the detection accuracy of the "with disability" category shows a slight decrease. This phenomenon can be attributed to the introduction of the P2 layer, which enhances small-object detection by incorporating more shallow features. However, the additional shallow features may also introduce certain background noise, slightly affecting the feature representation of some targets. Finally, upon integrating the RSAB mechanism, detection performance across all categories achieved significant improvements, with improvements of 3.9, 8.3, 4.3, and 5.4 percentage points compared to the baseline. The RSAB module can capture global contextual relationships and suppress irrelevant information, thereby mitigating the interference caused by shallow features and enhancing feature discrimination capability. As a result, the overall mAP@0.5 reached 73.0, representing a 5.4-point improvement over the original YOLOv8. These results clearly demonstrate that the proposed enhancements significantly boost the model's ability to detect vulnerable pedestrians under adverse weather conditions, and each module contributes positively to detection accuracy.

It should be noted that the AP for the elderly category in Table 2 is relatively low, which is due to the class imbalance in the dataset, where the number of samples for "elderly without disabilities" is significantly smaller than that of the "non-vulnerable group." Potential mitigation strategies, such as re-weighting or re-sampling, could be explored in future work to alleviate this imbalance and further improve detection performance.

In addition, we further validated the effectiveness of incorporating SCConv and Triplet Attention into the C2f_ST module in enhancing performance. As shown in Table 5, compared to the original C2f module in YOLOv8, integrating SCConv alone led to a 0.7% increase in mAP@0.5, while introducing the Triplet Attention mechanism alone resulted in a 0.9% improvement. When the two components were combined in the proposed C2f_ST module, mAP@0.5 increased by 1.6%, achieving the best performance. These results demonstrate that the two components are structurally complementary and both contribute positively to enhancing overall detection performance.

Table 5: Ablation study on the internal structure of the C2f_ST module.

Method	AP (Children without Disability) (%)	AP (Elderly without Disability) (%)	AP (With Disability) (%)	AP (Non-Vulnerable) (%)	mAP (%)
YOLOv8-C2f	74.5	51.8	87.6	56.5	67.6
YOLOv8-C2f_Scconv	76.2	50.6	89.7	56.7	68.3
YOLOv8-C2f_Triplet Attention	74.6	52.4	88.1	58.8	68.5
YOLOv8-C2f_ST	73.3	54.6	88.0	60.8	69.2

To further validate the effectiveness of the proposed RSAB mechanism in detecting vulnerable pedestrians under complex weather conditions, this study introduces several mainstream attention mechanisms into the YOLOv8 architecture for comparative experiments, including Convolutional Block Attention Module

(CBAM), Efficient Channel Attention (ECA), hybrid dilated residual attention block (HDRED), and RAB. Each attention module was embedded in the same designated position without modifying the remaining parts of the model. The detection performance was evaluated using the mAP@0.5 metric.

As shown in the Table 6, compared to the original YOLOv8, introducing the CBAM module improved the mAP by 1.6%. With the ECA module, detection accuracy for various categories of vulnerable pedestrians improved to varying degrees, leading to a 1.4% increase in mAP. Although the HDRED module caused a slight drop in detection accuracy for some categories, it maintained overall performance stability with a 1.9% mAP improvement. Incorporating the RAB module resulted in an mAP of 69.3%, an increase of 1.7%. The RSAB module achieved the best performance, with the mAP reaching 70.4%, representing a 2.8% improvement. It should be noted that, due to different experimental settings, the performance gains of RSAB differ between Tables 4 and 6. In Table 4, RSAB operates alongside other modules within the complete proposed framework, reflecting its collaborative effect; in Table 6, RSAB is evaluated as a standalone module on YOLOv8 to compare its individual effectiveness with other attention mechanisms. Based on the experimental results, the RSAB module achieves the most significant improvement in detection accuracy for vulnerable pedestrians under complex weather conditions.

Table 6: Performance comparison of different attention mechanisms on the augmented BGVP.

Method	AP (Children without Disability) (%)	AP (Elderly without Disability) (%)	AP (With Disability) (%)	AP (Non-Vulnerable) (%)	mAP (%)
YOLOv8	74.5	51.8	87.6	56.5	67.6
YOLOv8 + CBAM [36]	74.1	55.0	88.4	59.3	69.2
YOLOv8 + ECA [37]	76.6	52.6	90.2	56.6	69.0
YOLOv8 + HDRED [28]	80.7	48.0	90.2	59.0	69.5
YOLOv8 + RAB [28]	74.6	52.7	89.9	60.1	69.3
YOLOv8 + RSAB	75.3	54.7	89.3	62.5	70.4

5 Conclusion

This study proposes an improved YOLOv8 model, VPCW-YOLO, for detecting vulnerable pedestrians under complex weather conditions. To address challenges such as occlusion, blur, and low contrast caused by adverse weather, VPCW-YOLO enhances detection performance through multiple optimizations: the C2f_ST feature extraction module integrates SCConv with triplet attention, effectively improving key feature representation and region perception; the RSAB module strengthens global feature modeling to enhance target perception under complex weather; meanwhile, the Space-to-Depth feature reorganization and P2 small-object detection layer further improve detection of distant and small vulnerable pedestrians. Experimental results on the enhanced BGVP dataset show that VPCW-YOLO achieves an mAP@0.5 of 73.0% and an mAP@0.5:0.95 of 49.2%, representing improvements of 5.4% and 5.7% over the original YOLOv8, respectively. Furthermore, generalization experiments on the RTTS pedestrian subset demonstrate a 6.3% improvement in mAP@0.5, further validating the method's generalization capability.

These results indicate that the proposed method maintains high detection accuracy while offering strong generalization potential, providing reliable technical support for vulnerable pedestrian perception in intelligent transportation systems, autonomous driving, and surveillance scenarios.

Acknowledgement: None.

Funding Statement: This supported by "the Fundamental Research Funds of Zhejiang Sci-Tech University" (No. 26222178-Y) and Research Startup Fund of Zhejiang Sci-Tech University.

Author Contributions: The authors confirm contribution to the paper as follows: study conception and design: Jiaqi Wang; draft manuscript preparation: Jiaqi Wang; review: Jian Su. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The authors confirm that the datasets supporting the findings of this study, including the BGVP dataset [31] and the RTTS dataset [38], are available within the article.

Ethics Approval: Not applicable.

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

References

1. Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: unified, real-time object detection. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016 Jun 27–30; Las Vegas, NV, USA. p. 779–88. doi:10.1109/cvpr.2016.91.
2. Girshick R, Donahue J, Darrell T, Malik J. Region-based convolutional networks for accurate object detection and segmentation. IEEE Trans Pattern Anal Mach Intell. 2016;38(1):142–58. doi:10.1109/tpami.2015.2437384.
3. Yaseen M. What is YOLOv8: an in-depth exploration of the internal features of the next-generation object detector. arXiv:2408.15857. 2024. doi:10.48550/arXiv.2408.15857.
4. Girshick R. Fast R-CNN. In: Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV); 2015 Dec 7–13; Santiago, Chile. p. 1440–8. doi:10.1109/iccv.2015.169.
5. Ren S, He K, Girshick R, Sun J. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell. 2017;39(6):1137–49. doi:10.1109/tpami.2016.2577031.
6. He K, Gkioxari G, Dollar P, Girshick R. Mask R-CNN. In: Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV); 2017 Oct 22–29; Venice, Italy. p. 2980–8. doi:10.1109/iccv.2017.322.
7. Cai Z, Vasconcelos N. Cascade R-CNN: high quality object detection and instance segmentation. IEEE Trans Pattern Anal Mach Intell. 2021;43(5):1483–98. doi:10.1109/tpami.2019.2956516.
8. Sun P, Zhang R, Jiang Y, Kong T, Xu C, Zhan W, et al. Sparse R-CNN: end-to-end object detection with learnable proposals. In: Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2021 Jun 20–25; Nashville, TN, USA. p. 14449–58. doi:10.1109/cvpr46437.2021.01422.
9. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY, et al. SSD: single shot MultiBox detector. In: Leibe B, Matas J, Sebe N, Welling M, editors. Computer Vision—ECCV 2016. Cham, Switzerland: Springer International Publishing; 2016. p. 21–37. doi:10.1007/978-3-319-46448-0_2.
10. Lin TY, Goyal P, Girshick R, He K, Dollar P. Focal loss for dense object detection. In: Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV); 2017 Oct 22–29; Venice, Italy. p. 2999–3007. doi:10.1109/iccv.2017.324.
11. Tan M, Pang R, Le QV. EfficientDet: scalable and efficient object detection. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun 13–19; Seattle, WA, USA. p. 10778–87. doi:10.1109/cvpr42600.2020.01079.
12. Redmon J, Farhadi A. YOLOv3: an incremental improvement. arXiv:1804.02767. 2018. doi:10.48550/arXiv.1804.02767.
13. Bochkovskiy A, Wang CY, Liao HYM. YOLOv4: optimal speed and accuracy of object detection. arXiv:2004.10934. 2020. doi:10.48550/arXiv.2004.10934.
14. Khanam R, Hussain M. What is YOLOv5: a deep look into the internal features of the popular object detector. arXiv:2407.20892. 2024. doi:10.48550/arXiv.2407.20892.

15. Wang CY, Bochkovskiy A, Liao HM. YOLOv7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In: Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023 Jun 17–24; Vancouver, BC, Canada. p. 7464–75. doi:10.1109/cvpr52729.2023.00721.
16. Song H, Choi IK, Ko MS, Bae J, Kwak S, Yoo J. Vulnerable pedestrian detection and tracking using deep learning. In: Proceedings of the 2018 International Conference on Electronics, Information, and Communication (ICEIC); 2018 Jan 24–27; Honolulu, HI, USA. p. 1–2. doi:10.23919/elinfocom.2018.8330547.
17. Liu Z, Pu L, Meng Z, Yang X, Zhu K, Zhang L. POFS: a novel pedestrian-oriented forewarning system for vulnerable pedestrian safety. In: Proceedings of the 2015 International Conference on Connected Vehicles and Expo (ICCVE); 2015 Oct 19–23; Shenzhen, China. p. 100–5. doi:10.1109/iccve.2015.63.
18. Ni X, Kühnel C, Jiang X. Thermal detection of people with mobility restrictions for barrier reduction at traffic lights controlled intersections. IEEE Open J Intell Transp Syst. 2026;7(11):908–22. doi:10.1109/ojits.2026.3672067.
19. Xie Y, Du D, Bi M. YOLO-ACE: a vehicle and pedestrian detection algorithm for autonomous driving scenarios based on knowledge distillation of YOLOv10. IEEE Internet Things J. 2025;12(15):30086–97. doi:10.1109/jiot.2025.3569735.
20. Chen G, Zhang P, Zhang Y, He Z, Shi B. Adaptive feature-manipulated vehicle and pedestrian detection in infrared images. IEEE Trans Intell Transp Syst. 2025;26(4):4579–91. doi:10.1109/tits.2025.3545844.
21. Ge W, Huang S, Li M, Jiao Y. LEPD-net: a lightweight and efficient network for pedestrian detection. IEEE Trans Neural Netw Learn Syst. 2026;37(4):1717–25. doi:10.1109/tnnls.2025.3624356.
22. Wang C, Xie H. FCD-net: feature decorrelation and confidence-driven dynamic fusion for robust pedestrian recognition in autonomous driving. IEEE Trans Intell Transp Syst. 2025;26(11):19468–80. doi:10.1109/tits.2025.3596166.
23. An T, Gao H, Liu R, Dai K, Xie T, Li R, et al. An MoE-driven unified image restoration framework for adverse weather conditions. IEEE Trans Circuits Syst Video Technol. 2026;36(3):3101–16. doi:10.1109/tcsvt.2025.3625191.
24. Xu J, Hu X, Zhu L, Heng PA. Unifying physically-informed weather priors in a single model for image restoration across multiple adverse weather conditions. IEEE Trans Circuits Syst Video Technol. 2025;35(10):9575–91. doi:10.1109/tcsvt.2025.3561470.
25. Li J, Xu R, Liu X, Ma J, Li B, Zou Q, et al. Domain adaptation based object detection for autonomous driving in foggy and rainy weather. IEEE Trans Intell Veh. 2025;10(2):900–11. doi:10.1109/tiv.2024.3419689.
26. Li J, Wen Y, He L. SCConv: spatial and channel reconstruction convolution for feature redundancy. In: Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023 Jun 17–24; Vancouver, BC, Canada. p. 6153–62. doi:10.1109/cvpr52729.2023.00596.
27. Misra D, Nalamada T, Arasanipalai AU, Hou Q. Rotate to attend: convolutional triplet attention module. In: Proceedings of the 2021 IEEE Winter Conference on Applications of Computer Vision (WACV); 2021 Jan 3–8; Waikoloa, HI, USA. p. 3138–47. doi:10.1109/wacv48630.2021.00318.
28. Wu W, Liu S, Xia Y, Zhang Y. Dual residual attention network for image denoising. Pattern Recognit. 2024;149(8):110291. doi:10.1016/j.patcog.2024.110291.
29. Zhang H, Goodfellow I, Metaxas D, Odena A. Self-attention generative adversarial networks. arXiv:1805.08318. 2019. doi:10.48550/arXiv.1805.08318.
30. Sunkara R, Luo T. No more strided convolutions or pooling: a new CNN building block for low-resolution images and small objects. In: Amini MR, Canu S, Fischer A, Guns T, Kralj Novak P, Tsoumakas G, editors. Machine learning and knowledge discovery in databases. Cham, Switzerland: Springer Nature; 2023. p. 443–59. doi:10.1007/978-3-031-26409-2_27.
31. Sharma D, Hade T, Tian Q. Comparison of deep object detectors on a new vulnerable pedestrian dataset. In: Proceedings of the 2024 IEEE 27th International Conference on Intelligent Transportation Systems (ITSC); 2024 Sep 24–27; Edmonton, AB, Canada. p. 278–83. doi:10.1109/itsc58415.2024.10920135.
32. Li C, Li L, Jiang H, Weng K, Geng Y, Li L, et al. YOLOv6: a single-stage object detection framework for industrial applications. arXiv:2209.02976. 2022. doi:10.48550/arXiv.2209.02976.

33. Wang CY, Yeh IH, Mark Liao HY. YOLOv9: learning what you want to learn using programmable gradient information. In: Leonardis A, Ricci E, Roth S, Russakovsky O, Sattler T, Varol G, editors. Computer Vision—ECCV 2024. Cham, Switzerland: Springer Nature; 2024. p. 1–21. doi:10.1007/978-3-031-72751-1_1.
34. Wang A, Chen H, Liu L, Chen K, Lin Z, Han J, et al. YOLOv10: real-time end-to-end object detection. arXiv:2405.14458. 2024. doi:10.48550/arXiv.2405.14458.
35. Khanam R, Hussain M. YOLOv11: an overview of the key architectural enhancements. arXiv:2410.17725. 2024. doi:10.48550/arXiv.2410.17725.
36. Woo S, Park J, Lee JY, Kweon IS. CBAM: convolutional block attention module. In: Ferrari V, Hebert M, Sminchisescu C, Weiss Y, editors. Computer Vision—ECCV 2018. Cham, Switzerland: Springer International Publishing; 2018. p. 3–19. doi:10.1007/978-3-030-01234-2_1.
37. Wang Q, Wu B, Zhu P, Li P, Zuo W, Hu Q. ECA-net: efficient channel attention for deep convolutional neural networks. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun 13–19; Seattle, WA, USA. p. 11531–9. doi:10.1109/cvpr42600.2020.01155.
38. Li B, Ren W, Fu D, Tao D, Feng D, Zeng W, et al. Benchmarking single-image dehazing and beyond. IEEE Trans Image Process. 2019;28(1):492–505. doi:10.1109/TIP.2018.2867951.