Lightweight Real-Time Object Detection via Enhanced Global Perception and Intra-Layer Interaction for Complex Traffic Scenarios

Ben Liang, Jia Su, Kangkang Feng, Yongqiang Zhang, and Weimin Hou

***Abstract*—Due to unfavorable factors such as cluttered spatial and temporal distribution of multiple types of targets, occlusion of background objects of different shapes, and blurring of feature information by inclement weather, the low detection accuracy in complex traffic scenarios has been a troubling issue. Regarding the above-mentioned issues, the paper proposes a lightweight real-time detection network to augment multi-scale object perception capabilities in traffic scenarios while ensuring real-time detection speed. First, we construct a novel global feature extraction (GFE) structure by cascading orthogonal band convolution kernels that capture the global dependencies between pixels to improve feature discrimination. Then, an intra-layer multi-scale feature interaction (IMFI) module is proposed to reinforce the effective reuse and multi-level transfer of salient features. In addition, we build a multi-branch scale-aware aggregation (MSA) module that captures abundant context-associated features to improve the target decision-making capability and the self-adaptive capability of the model when dealing with diverse object scales. Experimental results demonstrate that the proposed approach attains a significant improvement of 5.6 percentage points in AP50 with fewer parameters and computational power compared to the baseline model, with an improved FPS of 73. Furthermore, our approach strikes the optimal speed-accuracy balance when compared against other excellent object detection algorithms of the same magnitude.**

Link to graphical and video abstracts, and to code: https://latamt.ieeer9.org/index.php/transactions/article/view/8420

***Index Terms*—Object detection, Complex autonomous driving, Real-time, Global dependencies, Multi-branch scale-aware**

I. Introduction

Object detection plays a fundamental role in addressing computer vision problems explored, enabling accurate identification of object classes and recognition of object contour sizes from specified scenes. Due to tremendous leaps in innovative technologies and artificial intelligence, this task has been broadly implemented across various fields, such as autonomous driving [1], intelligent transport systems [2], [3], and smart city [4]. In particular, as autonomous driving perception technology has increasingly matured within the realm of intelligent transportation, it has attracted considerable attention from scholars and society. Moreover, an important core technology of autonomous driving is the accurate detection and classification of diverse road traffic objects. However, the problem of achieving real-time and accurate object detection in complex traffic scenarios has been a persistent and perplexing issue.

Deep learning-based object detection methods have come out ahead of traditional detection methods developed in the past few years and have demonstrated better performance regarding accuracy and speed. In general, object detection methods can be categorized into two-stage pipelines and single-stage pipelines. Two-stage detection approaches typically employ selective search techniques to generate potential region proposals prior to determining the object class and position, such as Fast R-CNN [5] and Faster R-CNN [6]. In contrast, single-stage detection algorithms transform the detection as an issue of category labeling and placement estimation using regression, such as SSD [7] and YOLO [8], [9], [10], [11]. Both types of detectors achieve excellent detection performance in simple general-purpose scenarios, but when applied in more intricate circumstances, their detection accuracy and robustness are greatly diminished. In complex traffic environments, issues like similar targets or different classes of targets with different scales, different degrees of occlusion overlap between various targets, and interference from complex buildings of different shapes can make accurate object detection more challenging. Furthermore, when in foggy weather conditions, the brightness and contrast of the detection environment are decreased, the image background is blurred, and the feature details are severely lost, which also further increases the difficulty of the detection algorithm to make decisions on object categories.

For the above-mentioned issues, many scholars and researchers have utilized techniques like Feature Pyramid Network (FPN) [12], Path Aggregation Network (PAN) [13], or Bi-directional Feature Pyramid Network (BiFPN) [14] to improve feature utilization and cross-scale information flow. By leveraging multi-level feature maps and top-down and lateral connections between pyramid levels, these methods allow better aggregation of semantic and spatial information across different scales. However, feature pyramid-based architectures mainly focus on feature propagation across layers, while overlooking the inter-dependencies between features within each layer. Additionally, since the Vision Transformer (ViT) [15] model was first applied to visual tasks and achieved impressive performance, some researchers have started to leverage the powerful global context modeling capabilities of the transformer to improve object detection accuracy, e.g.,

B. Liang, J. Su, K. Feng, W. Hou and Y. Zhang are with the School of Information Science and Engineering, Hebei University of Science and Technology, Shijiazhuang, China (e-mails: benhebust@163.com, sujia@hebust.edu.cn, fengkhbkj@163.com, hwm@hebust.edu.cn and zyq@hebust.edu.cn.).

PVT [16], Swin Transformer [17], and DETR [18]. However, self-attention layers in Transformers exhibit quadratic computational complexity with respect to sequence length, these models typically possess a substantial number of parameters and require massive computational resources. Especially in detection tasks with high input image resolutions, this can result in an enormous computational burden and severely slow down real-time detection speeds.

In the following, a lightweight real-time object detection algorithm is proposed to reinforce the model's robustness and multi-object detection capability in complicated traffic environments while sustaining swift detection speeds. Firstly, we design an enhanced feature extraction structure, called Global Feature Extraction (GFE), which can capture the global position of the object features through the interaction of horizontal and vertical information. The GFE module utilizes fully convolution neural networks to achieve transformer-like global modeling capability and requires only a small amount of computation. Secondly, to ensure the network can obtain richer feature presentation and learn an implicit relation between pixels/objects. An intra-layer multi-scale feature interaction (IMFI) module is proposed to enrich the diversity and completeness of image features. Thirdly, larger receptive fields are beneficial for capturing rich contextual features and neighboring pixel information, so we propose a multi-branch scale-aware aggregation (MSA) structure to improve long-range dependencies between objects and suppress redundant background clutter. In addition, considering that collecting and annotating real foggy images is extremely difficult, and existing foggy image datasets are few in number and lacking in diversity. Therefore, to ensure the applicability and robustness of the model to more complex environments, we use data enhancement techniques to simulate complex foggy traffic scenarios by leveraging the depth information contained within the images to increase the diversity of the dataset [19].

II. Methods

We propose a lightweight real-time detection framework in this section that can markedly improve object detection accuracy under difficult traffic conditions while maintaining high detection efficiency in real-time. To accomplish our goal, we are focusing on three main aspects. Firstly, to enhance the quality and precision of extracted features under complex conditions, we construct the GFE structure to strengthen the representation of global position contexts within the image features. Secondly, we embed the proposed IMFI module in the feature pyramid network to ensure that more effective and significant features are retained during feature transfer. Finally, the different scale perceptual fields of the output feature maps are aggregated using the MSA module to improve the adaptability to objects at different scales. The complete structure is represented in Fig. 1.

A. Global Feature Extraction (GFE)

The quality of the effective feature representations extracted by the backbone network plays a crucial role in the subsequent effective fusion and interaction of different feature information. In the backbone, a series of cascaded convolutional layers with varying characteristics are used to progressively extract higher-level features from the input image in a hierarchical manner. While the convolutional operation captures pixel-level information within its local receptive field through a sliding window approach, it lacks sensitivity to associations between distant or non-adjacent regions within the feature maps and does not explicitly encode the global positional information of each pixel or region relative to the full image context. However, in traffic scenes with objects appearing at varying scales and locations, recognizing the global context and precise position of the object is crucial for accurate detection and localization tasks. To address this, we propose the GFE module to compensate for the inherent limitations of convolutional layers in modeling long-range relationships and accessing global positional awareness, aiding in understanding the entire scene and locating the target from a more macro view.

Specifically, we employ two large convolutional kernels oriented orthogonal to each other in the horizontal and vertical axes to aggregate the relationships between pixels across different directions. By interactively stacking the horizontal and vertical pixel tensors, generates a global receptive field to obtain the global spatial dependencies of each pixel in the feature map. In other words, it encodes each pixel's dependencies not just within its local neighborhood, but also concerning distant pixels in both the horizontal and vertical directions across the entire feature map. Essentially, it expands the local neighborhood conventionally processed by convolutional filters into a holistic global scope. The use of orthogonal convolutional kernels that operate along different axes effectively expands the traditional local receptive fields of convolutional layers into a true global receptive field, enabling the modeling of long-range dependencies between pixels throughout the feature map.

As shown in the top part of Fig. 1, the input feature maps are each passed through two parallel 1×1 regular convolutions to halve the number of channels. Then, a cheap operation, PDConv, is performed on one of the branches, using a smaller number of parameters to achieve the effect of a normal convolution. Specifically, given input feature $F^{H\times W\times C}$ with height H, width W, and channel's number C. Firstly, a 1×1 point-wise convolution (PWConv) is applied to extract the important features and compress the number of channels, as shown in the following Eq. (1)

$$Y_1 = F * PWConv_{1\times 1} \tag{1}$$

where $Y_1 \in H \times W \times C1$ is the output feature map, and its output channel number $C1$ is half of C. * denotes the convolution operation. Then, a 3×3 depth-wise convolution (DWConv) is utilized to independently extract distinct features for each channel, as shown in Eq. (2)

$$Y_2 = Y_1 * DWConv_{3\times 3} \tag{2}$$

where the number of output channels of $Y_2 \in H \times W \times C1$ remains the same. Finally, the two convolution outputs are

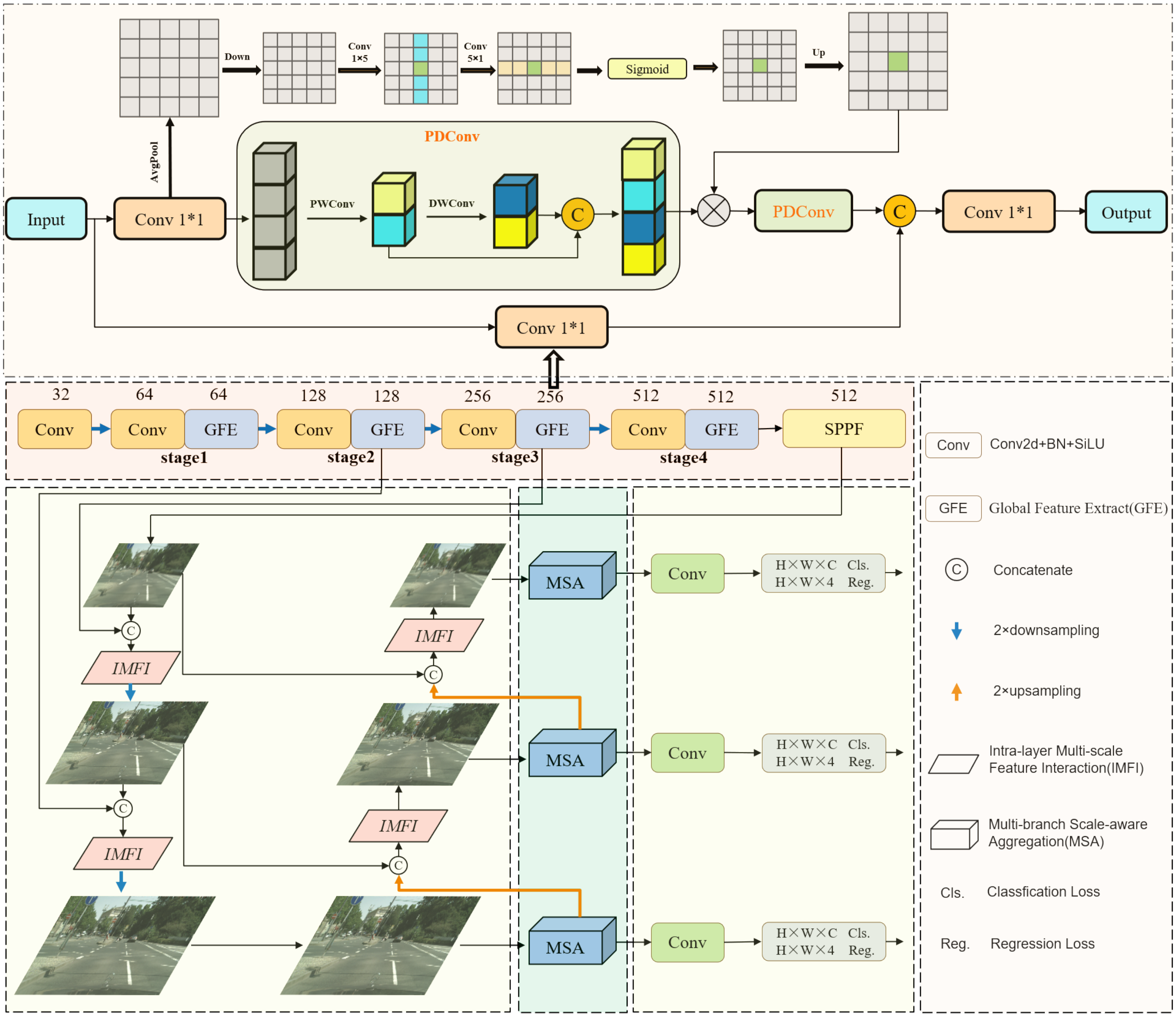


Fig. 1. The overall architecture of the proposed method.

combined by concatenating them along the channel dimension, as demonstrated in Eq. (3)

$$F_{out} = Concat[Y_1, Y_2] \tag{3}$$

where $F_{out} \in H \times W \times C_{out}$ is the output feature. Next, we proceed by using orthogonal convolutions to aggregate distinct locations of each pixel in the horizontal and vertical directions. The concatenated outputs undergo a Sigmoid normalization and ranking operation to generate a global perceptual attention feature map and then are upsampled to restore the original feature map size. This attention map is then element-wise multiplied with the input feature F_{out}. This multiplication operation effectively strengthens the representation by incorporating long-range spatial dependencies captured distinctly by the orthogonal convolutions. Furthermore, unlike self-attention [15], our design does not incur the high memory overhead or increased inference latency associated with computing pairwise interactions between all spatial positions explicitly. As such, it enables capturing of global contextual cues to augment feature learning, while maintaining the computational efficiency of standard convolutional networks.

B. Intra-layer Multi-scale Feature Interaction (IMFI)

A single input image often contains objects of different scales with varying sizes and contour shapes. To enhance better scale-invariant object detection ability of the network, FPN [12] enables the propagation of high-level semantic features to lower feature maps to facilitate inter-layer feature interactions. PAN [13] incorporates a bottom-up path into the original FPN design, transferring lower-level location cues to higher pyramid levels. This inter-layer feature interaction mechanism combines shallow positional features and deep semantic features, fusing multi-scale representations from different network layers to enhance object detection performance. However, such approaches are still focused on layer-to-layer

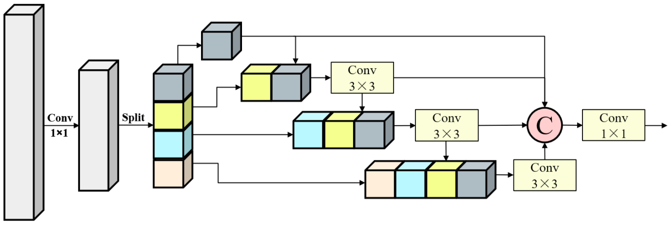


Fig. 2. The architecture of IMFI.

information transmission. In this paper, we propose the IMFI module to perform multi-scale feature interaction at a finer granularity level and better integrate information within a single block. It constructs hierarchical feature cascading within a block via feature splitting and grouped convolutions, progressively increasing the receptive field ranges between groups. By modeling intra-layer multi-scale interactions and aggregating outputs with parallel group convolutional branches and varied receptive fields, the IMFI module combines positional and semantic cues at a more granular level and realizes a more optimal combination of local and global scales through its granular intra-block design compared to solely relying on inter-layer propagation.

Specifically, as demonstrated in Fig. 2, the input features are equally divided into four groups along the channel dimensions after 1×1 convolution to reduce the dimensionality, as shown in Eq. (4)

$$F \rightarrow Split(X_1, X_2, X_3, X_4) \tag{4}$$

where F denotes the output after 1×1 convolution and X_i represents four different parts. Next, the previous group of features is concatenated with the neighboring next group of features and then convolved by a regular convolution of 3×3 for significant feature extraction. It can be summarized as Eq. (5

$$Y_i = Conv_{3\times3} * (X_{i-1} \oplus X_i), 2 \leq i \leq 4 \tag{5}$$

where * denotes the convolution operations and $\oplus$ represents the summation operation. Finally, to better integrate information of different scales, outputs from different groups are concatenated and fused through a 1×1 convolution layer, as shown in Eq. (6)

$$F_{out} = Conv_{1\times1} * (Concat[X_1, X_2, X_3, X_4]) \tag{6}$$

As mentioned above, multiple groups of convolutional residual linking operations within the intra-layer allow for progressive adjustments in receptive field scales, capturing both local and global information at a more fine-grained level. The combined effect of different inter-group features and multiple 3×3 convolutional overlays allows the output of IMFI to contain different combinations of different amounts and scales of receptive field information, facilitating the extraction of more equivalent feature scales and contextual information. Consequently, the proposed IMFI module achieves a finer network multi-scale modeling by transferring and fusing different intra-layer features, which further enhances the feature pyramid's ability to detect various objects.

C. Multi-branch Scale-aware Aggregation (MSA)

In CNNs, larger receptive fields allow the network to capture more global context and consider a wider area of the image. However, blindly pursuing expansive receptive fields without justification is inadvisable, as the size of the field of perception required for different scale targets is different. While a small receptive field is better suited for detecting small objects, a field close to or smaller than the object scale does not leverage enough contextual information. With a limited field of view that only includes the target, important surrounding context is missed. This makes it difficult to discriminate the target from its surroundings and hinders object detection performance.

We introduce a novel MSA module to address the fixed receptive field limitations of CNNs for multi-scale target detection. The MSA module comprises parallel convolution branches with identically configured filters except for the dilation rates. This design allows the network to simultaneously incorporate diversified receptive field sizes. As shown in Fig. 3, the module first splits the input feature F equally into three branches and applies 3×3 convolution with unique dilation rates ranging from 1 to 3. This results in effective receptive fields across branches spanning from local to larger contextual scales. The convolved feature maps from all branches are then concatenated to form a stacked multi-scale feature $H = \{F1, F2, F3\}$. Subsequently, 1×1 convolutions encode channel dependencies between hierarchical scales, facilitating the fusion of fine-grained and global representations.

The MSA module captures multi-scale features through its branches with varied dilation rates. Smaller dilations extract fine-grained details, while larger ones encompass a broader context. Features captured from different dilation branches can provide complementary representations, capturing fine details as well as global context. By aggregating representations from multiple branches, the MSA combines complementary scale-specific details. The aggregation of multi-branch receptive fields allows the network to perceive more differences between objects at different scales, richer contextual feature information, and correlation of feature information, thus enhancing the model's perceptual adaptability to scale variations and improving the target detection accuracy. Its multi-branch scale-aware design overcomes the limitations of fixed receptive fields, providing a robust and scalable solution for multi-scale visual tasks.

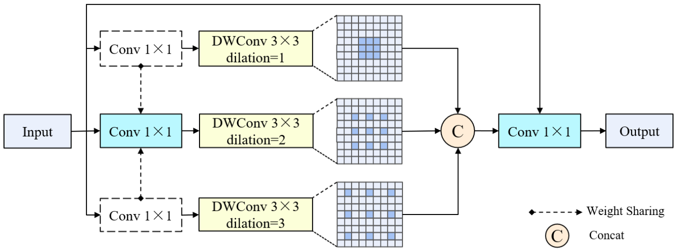


Fig. 3. The architecture of MSA.

III. Experiments and Discussion

A. Datasets

This paper is based on extensive experiments with the publicly available complex cityscape dataset Cityscapes [20] and the PASCAL VOC [21].

Cityscapes comprises street view images captured by in-car cameras across 50 diverse urban streets, including 2975 training images and 500 test images. It includes eight categories such as people, riders, cars, trucks, buses, motorbikes, and bicycles. Based on this dataset, We use the data enhancement method in the literature [19] to perform fogging operations on the images in the training and test sets, respectively, so as to achieve the simulation of complex foggy environments. The final hybrid dataset utilized for this paper comprised 5950 images for training and 1000 for testing, with a 1:1 ratio between foggy and clear images.

PASCAL VOC is a prevalent benchmark dataset for object detection comprising VOC2007 and VOC2012, which contain images across 20 common object classes. In this work, the training and validation subsets from VOC2007 and VOC2012, totaling 16,551 images, are used for model training. The VOC2007 test set, containing 4,952 images, is employed for model testing.

B. Evaluation Metrics and Implementation Details

For a fair comparison, our experiments mainly adopt primary evaluation metrics used in MS COCO [22] to gauge detection performance. These established measures are Average Precision (AP) and Average Recall (AR). Besides, it can be subdivided into AP_{50} (IoU=0.50), AP_{75} (IoU=0.75), and the overall AP (IoU=0.50:0.95) depending on different IoUs. In addition, the AP_S, AP_M, and AP_L represent small, medium, and large objects respectively. To provide a holistic assessment of the proposed methods, we also use Parameters (Params), GFLOPs, and Frames Per Second (FPS) as our comparison criteria.

To demonstrate the effectiveness of our proposed method, we use YOLOv5-6.1 as the baseline model without pre-trained weights. All experiments are conducted on a PC with an Intel i5-12600KF CPU at 3.70GHz and an NVIDIA GeForce GTX3060 GPU (12GB) based on the PyTorch 1.10 framework and CUDA 11.3. During the training phase, the optimizer utilizes SGD with a momentum of 0.937, a weight decay of 1e-5, and a warmup momentum of 0.8. The initial learning rate (lr) is set to 1e-2, and the cosine function is employed to dynamically reduce the lr. In addition, both training and test image sizes are uniformly resized to 640×640 size. Only basic common image enhancement methods such as 50% random horizontal flip, 0.0-0.5 times random scale, and 10% image translation are applied to the training images. These parameters are kept consistent for all experiments, training for 100 epochs with a batch size of 16.

C. Main Results

Considering the autonomous driving traffic scenarios often require fast and real-time detection of objects, models with a smaller number of parameters and lower computational requirements are better positioned for pragmatic engineering practices and tangible deployment scenarios. We compare the proposed algorithm with several popular real-time detectors, including YOLOX [23], YOLOv6 [24], and YOLOv7 [25]. At the same time, we also conduct comparison experiments using the newly released lightweight backbone FasterNet [26] replacement in the benchmark model. To fully compare the effectiveness of the algorithms, we also perform multiple data comparisons of algorithms with larger model volumes, such as QARepNext [27] and Swin Transformer [17].

As can be seen from the results in Table I, our method achieves the best performance among lightweight models in terms of AP_{50}, AP_{75}, and AP. Notably, all the results in Table I are for the mixed test set containing 500 original images and 500 synthetic foggy images. Comparing the first and last rows, our model decreases parameters and GFLOPs compared to the YOLOv5s baseline, while improving the $AP50$ metric by 5.6 percentage points. Furthermore, as can be seen from the data in the penultimate row of the table, our application of the improvements proposed in this paper to the YOLOv7-Tiny model also yielded stunning results, with AP, $AP50$, and $AP75$ improving by 2.0%, 4.6%, and 1.3%, respectively. Against state-of-the-art single-stage lightweight models YOLOX, YOLOv6, and YOLOv7-Tiny, the proposed model also achieves the highest $AP50$, achieving the best trade-off between speed and accuracy. When replacing the YOLOv5s backbone with the lighter FasterNet, parameters, and computation are further reduced but at the cost of decreased AP. In addition, compared with algorithms with larger model sizes, while our lightweight model is lower than theirs in AP, the huge computational parameters can severely reduce the model inference speed and increase the deployment cost. As can be seen from the comparison of FPS in the table, the FPS of the large model is much lower than that of our 73 FPS. In addition, to evaluate the generalizability and effectiveness of the proposed method, we conduct comparative experiments on the commonly used generalized object detection dataset PASCAL VOC [21]. As shown in the results in Table II, the proposed method obtains the highest mAP of 81.5% on the VOC2007 test set.

D. Ablation Study

To assess the effectiveness of the proposed methods, we conduct extensive ablation experiments. As shown in Tables III and IV, where M1, M2, and M3 represent the GFE module, IMFI structure, and MSA module, respectively. From the data comparison between the first two rows in Table III and Table IV, the GFE module enhances the global position of different targets in the entire feature map by overlaying between the pixels in the horizontal and vertical directions, which effectively improves the object detection rate in complex environments. The GFE module improves the AP50 of the baseline model from 32.6% to 35.1% on the original test set containing only clear images, and the AP50 improves by 2.2% on the mixed test set containing foggy images. Then, when we insert the proposed IMFI module into the feature pyramid

TABLE I
COMPARISON WITH STATE-OF-THE-ARTS ON MIXED TEST SET

Methods	Params	GFLOPs	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	FPS_{ba1}
YOLOv5s	7.2M	16.5G	16.7	31.6	15.2	1.5	10.0	30.4	71
YOLOX	5.1M	7.6G	17.1	32.6	15.6	0.6	8.2	32.4	72
YOLOv6	4.7M	11.4G	16.9	31.2	16.0	0.6	9.6	32.4	76
YOLOv7-Tiny	6.2M	13.7G	18.9	35.8	17.0	1.3	10.4	33.8	76
FasterNet	6.1M	13.3G	16.0	30.2	14.7	1.2	9.1	29.3	74
QARepNext	10.06M	25.3G	17.5	33.7	14.8	1.0	9.9	31.9	67
YOLOv5m	20.89M	48.3G	22.3	39.7	21.5	1.2	12.5	39.3	53
YOLOv7	37.23M	105.2G	24.3	43.9	22.3	2.8	15.3	39.6	50
Swin Transformer	33.72M	88.1G	23.4	42.6	22.0	1.1	14.0	41.2	31
Ours-v7	6.3M	14.1G	20.9	40.4	18.3	2.1	12.0	36.0	77
Ours-v5	6.8M	14.8G	19.8	37.2	18.3	1.3	10.5	35.7	73

TABLE II
COMPARATIVE DETECTION RESULTS ON VOC2007 TEST SET

Methods	Backbone	Train	mAP(%)
Faster RCNN [6]	VGGNet	VOC07+12	73.2
Faster RCNN [6]	ResNet-101	VOC07+12	76.4
R-FCN [28]	ResNet-101	VOC07+12	80.5
FCOS [29]	ResNet-50	VOC07+12	77.8
ATSS [30]	ResNet-50	VOC07+12	78.2
FAENet [31]	VGGNet	VOC07+12	80.1
RetinaNet [32]	VGGNet	VOC07+12	80.1
SSD300 [7]	VGGNet	VOC07+12	77.2
YOLOv3 [10]	DarkNet-53	VOC07+12	79.4
YOLOv5	DarkNet-53	VOC07+12	80.1
Ours	DarkNet-53	VOC07+12	81.5

TABLE III
ABLATION RESULTS FOR THE PROPOSED METHODS ON THE ORIGINAL TEST TET OF CITYSCAPES

Methods	Params	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Baseline	7.2M	17.2	32.6	15.4	2.2	11.3	30.5
M1	5.8M	18.7	35.1	17.6	1.4	11.5	33.8
M2	7.3M	20.4	37.9	19.4	1.4	12.6	36.4
M3	7.7M	20.3	37.5	18.9	1.7	12.3	35.9
M1+M2	6.1M	20.0	38.2	17.4	1.5	12.4	35.5
M1+M2+M3	6.8M	20.8	39.3	19.3	1.9	11.7	36.3

TABLE IV
ABLATION RESULTS FOR THE PROPOSED METHODS ON THE MIXED TEST SET OF CITYSCAPES

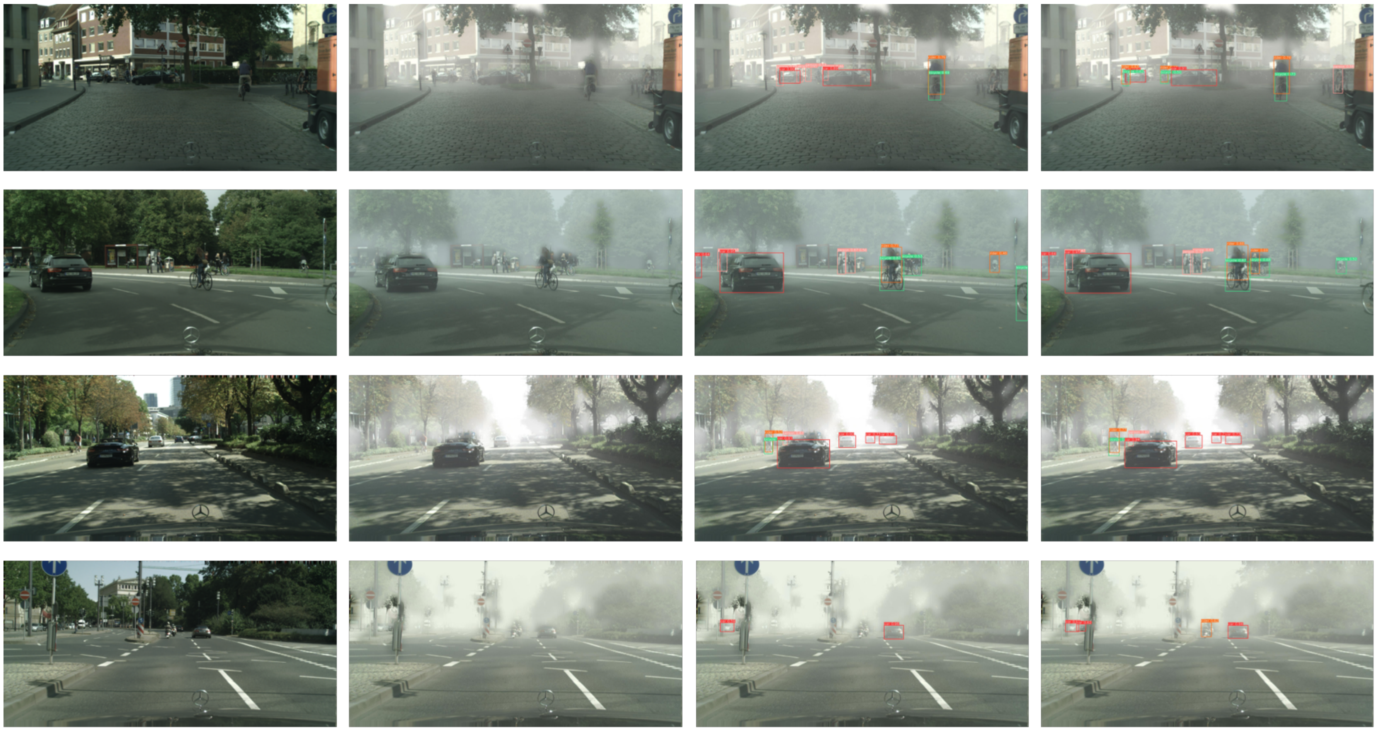
Methods	Params	AP	AP_{50}	AP_{75}	AR_1	AR_{10}	AR_{100}
Baseline	7.2M	16.7	31.6	15.2	14.5	26.4	31.7
M1	5.8M	18.0	33.8	16.5	15.3	27.4	32.3
M2	7.3M	19.7	36.2	18.3	15.9	28.3	32.3
M3	7.7M	19.5	35.8	18.2	15.9	28.6	32.5
M1+M2	6.1M	19.3	36.6	17.0	16.4	28.0	32.3
M1+M2+M3	6.8M	19.8	37.2	18.3	15.9	29.0	33.8

network, the larger and multi-scale receptive field interactions generated by feature grouping and convolutional superposition result in a huge improvement in the model's detection of large-scale targets. As indicated by the results in the third row of Table IV, the module improves by 3.0, 4.6, and 3.1 percentage points in AP, AP_{50}, and AP_{75}, respectively. The MSA module aggregates contextual information corresponding to different scales of targets using null convolutional branches with different expansion rates, allowing for better assignment of more appropriate feature representations based on the scale of the detected targets. We integrated this module separately into the detection process before the final decision-making phase to improve the model's ability to detect multiple targets and enhance its robustness. Subsequently, we incorporated the proposed method jointly into the model, and the proposed method was significantly improved in terms of detection accuracy and recall compared to the baseline model. The AP, AP_{50}, and AP_{75} reached a maxi-mum of 19.8%, 37.2%, and 18.3% respectively. The AR (maxdet=1, 10, 100) improved by 1.4%, 2.6%, and 2.1% respectively.

For qualitative assessment, detection results from the baseline and proposed methods are displayed in Fig. 4 to illustrate and compare their effects. The detected images are obtained from the Cityscapes dataset. To make the contrast more visible, we set the confidence threshold to 0.45, which filters out some of the lower-quality detection boxes. It can be seen from the results in Fig. 4 that the results in the first line that YOLOv5 does not detect the pedestrian object in the right corner. By carefully comparing the third and fourth columns of the figure, we can see that the proposed method outperforms the baseline model in both detection accuracy and false alarm rate. In summary, the proposed model effectively improves the performance of object detection in complex traffic scenarios.

E. Discussion

As we know, road traffic object detection is crucial for autonomous driving perception, where real-time and accurate detection is equally important. Even delays of 1 second or milliseconds can cause accidents. By considering the trade-off between speed and accuracy, we propose a lightweight detection model that improves multi-object detection precision while maintaining computational efficiency. Meanwhile, a smaller model size also helps reduce system development costs for autonomous systems. The experimental results demonstrate that our lightweight algorithm effectively improves the detection performance for objects in complex traffic scenarios while maintaining a relatively high FPS.



(a) Original Image (b) Foggy Image (a) YOLOv5 (a) Proposed Method

Fig. 4. Detection results of YOLOv5 and the proposed model.

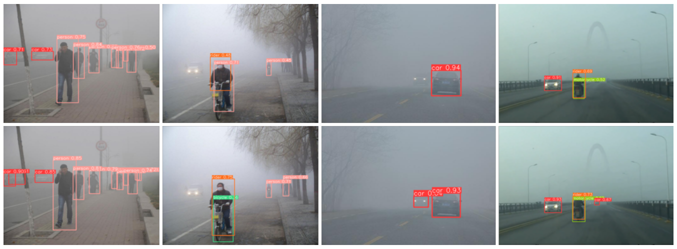


Fig. 5. Detection results of real foggy traffic image, the first and second rows are the results of the baseline YOLOv5 and the proposed model in this paper, respectively.

In addition, fog is one of the common adverse weather conditions encountered in autonomous driving scenarios, which can easily cause blurred target features and reduce detection accuracy. In order to ensure the robustness of the detection model for foggy detection environments, we leverage depth information to synthesize transmittance-based fog images. As shown in Fig. 4, we test our model on a mixed dataset containing clear and foggy images. Our model was able to detect more blurred and distant targets compared to the baseline model. Furthermore, we conduct real-world testing of the model using images captured from actual foggy traffic scenes. As shown in Fig. 5, our model is effective at reducing false detections when operating in foggy conditions. However, as fog density varies naturally in density, synthetic single-density fog cannot fully represent complex real scenarios, risking generalization. To better simulate diverse fogs, future work will introduce multi-density variations and robustness training. When deployed, other adverse conditions like rain and snow may cause even more severe interference. In summary, a lighter model design with lower computational costs is more conducive to subsequent real-world system deployment. The inclusion of synthesized fog images has enabled the model to better adapt to real object environments. The lightweight algorithm proposed in this paper achieves a better balance between detection accuracy and speed compared to other methods, bringing us one step closer to reliable perception systems for autonomous vehicles operating in diverse weather conditions.

IV. Conclusion

In this work, we propose a lightweight real-time detection algorithm for complex traffic scenarios. A novel global feature extraction (GFE) structure is proposed to capture dependencies between long-range pixels and enhance the model's ability for feature recognition and localization of objects. Then, intra-layer multi-scale feature interaction (IMFI) is designed to create multi-level feature transfer between layers and increase the flow of information. Besides, we use the multi-branch scale-aware aggregation (MSA) module to improve the adaptability of the network to different scales of targets by aggregating different sizes of receptive field information. Experiments show that the proposed method leads to substantial improvements in both average precision and average recall. Compared to other excellent lightweight algorithms, the model proposed in this paper achieves the best speed-accuracy trade-off, striking a perfect balance between rapid processing and accurate results.

ACKNOWLEDGMENT

The authors extend their sincere appreciation to the reviewers for their insightful and constructive comments, which have helped improve the overall presentation of this paper. Sincere thanks also go to the Hebei University of Science and Technology for their support throughout the research process.

REFERENCES

[1] X. Dai, X. Yuan, and X. Wei, "Tirnet: Object detection in thermal infrared images for autonomous driving," *Applied Intelligence*, vol. 51, pp. 1244–1261, 2021.

[2] D. C. Santos, F. A. da Silva, D. R. Pereira, L. L. de Almeida, A. O. Artero, M. A. Piteri, and V. H. Albuquerque, "Real-time traffic sign detection and recognition using cnn," *IEEE Latin America Transactions*, vol. 18, no. 03, pp. 522–529, 2020.

[3] V. Kshirsagar, R. H. Bhalerao, and M. Chaturvedi, "Modified yolo module for efficient object tracking in a video," *IEEE Latin America Transactions*, vol. 21, no. 3, pp. 389–398, 2023.

[4] P. Zhu, L. Wen, D. Du, X. Bian, H. Fan, Q. Hu, and H. Ling, "Detection and tracking meet drones challenge," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 11, pp. 7380–7399, 2021.

[5] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, pp. 1440–1448, 2015.

[6] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.

[7] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pp. 21–37, Springer, 2016.

[8] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779–788, 2016.

[9] J. Redmon and A. Farhadi, "Yolo9000: better, faster, stronger," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7263–7271, 2017.

[10] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[11] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.

[12] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2117–2125, 2017.

[13] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8759–8768, 2018.

[14] M. Tan, R. Pang, and Q. V. Le, "Efficientdet: Scalable and efficient object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10781–10790, 2020.

[15] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[16] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 568–578, 2021.

[17] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

[18] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*, pp. 213–229, Springer, 2020.

[19] C. Sakaridis, D. Dai, and L. Van Gool, "Semantic foggy scene understanding with synthetic data," *International Journal of Computer Vision*, vol. 126, pp. 973–992, 2018.

[20] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3213–3223, 2016.

[21] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *International journal of computer vision*, vol. 111, pp. 98–136, 2015.

[22] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755, Springer, 2014.

[23] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "Yolox: Exceeding yolo series in 2021," *arXiv preprint arXiv:2107.08430*, 2021.

[24] C. Li, L. Li, Y. Geng, H. Jiang, M. Cheng, B. Zhang, Z. Ke, X. Xu, and X. Chu, "Yolov6 v3. 0: A full-scale reloading," *arXiv preprint arXiv:2301.05586*, 2023.

[25] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7464–7475, 2023.

[26] J. Chen, S.-h. Kao, H. He, W. Zhuo, S. Wen, C.-H. Lee, and S.-H. G. 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*, pp. 12021–12031, 2023.

[27] X. Chu, L. Li, and B. Zhang, "Make repvgg greater again: A quantization-aware approach," *arXiv preprint arXiv:2212.01593*, 2022.

[28] J. Dai, Y. Li, K. He, and J. Sun, "R-fcn: Object detection via region-based fully convolutional networks," *Advances in neural information processing systems*, vol. 29, 2016.

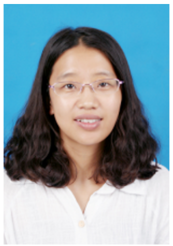
[29] Z. Tian, C. Shen, H. Chen, and T. He, "Fcos: Fully convolutional one-stage object detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9627–9636, 2019.

[30] S. Zhang, C. Chi, Y. Yao, Z. Lei, and S. Z. Li, "Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9759–9768, 2020.

[31] W. Li and G. Liu, "A single-shot object detector with feature aggregation and enhancement," in *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 3910–3914, IEEE, 2019.

[32] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li, "Single-shot refinement neural network for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4203–4212, 2018.

Ben Liang received the B.S. degree in communication engineering from West Anhui University, Anhui, China. He is currently pursuing the M.E. degree with the Hebei University of Science and Technology. His current research interests include computer vision, image processing, and object detection.



Jia Su received the Ph.D. degree in communication and information systems from Harbin Engineering University, Heilongjiang, China, in 2010. She is currently a Professor with the School of Information Science and Engineering, Hebei University of Science and Technology. Her research interests include multi-antenna array, image processing, computer vision, and object detection.

Kangkang Feng received the B.S. degree in communication engineering from the Polytechnic College of Hebei University of Science and Technology, Shijiazhuang, China. He is currently pursuing the M.E. degree with the Hebei University of Science and Technology. His current research interests include image processing and object detection.

Yongqaing Zhang received the M.E. degree in Anhui University of Technology, majoring in computer application technology. He is currently pursuing a doctor's degree in Army Engineering University of PLA. He is a director of the Hebei Industrial Internet Industry Alliance and deputy director of the Hebei Intelligent Internet of Things Technology Innovation Center. His research interests include artificial intelligence technology and Internet of things technology.

Weimin Hou received the Ph.D. degree in signal and information processing from the Institute of Acoustics, Chinese Academy of Sciences, Beijing, China, in 2007. He is currently a Professor with the School of Information Science and Engineering, Hebei University of Science and Technology. His research interests include array signal processing, wireless communication, remote sensing image processing, and artificial intelligence.