Received: 26 April 2023 | Revised: 20 July 2023 | Accepted: 11 August 2023

IET Image Processing

DOI: 10.1049/ipr2.12912

ORIGINAL RESEARCH

Small object detection based on hierarchical attention mechanism and multi-scale separable detection

Yafeng Zhang[1] | Junyang Yu[1] | Yuanyuan Wang[2] | Shuang Tang[3] | Han Li[3] | Zhiyi Xin[1] | Chaoyi Wang[4] | Ziming Zhao[1]

[1]School of Software, Henan University, Kaifeng, Henan, China

[2]Economic and Technical Research Institute, State Grid Henan Province Electric Power Company, Zhengzhou, Henan, China

[3]School of Economics, Henan University, Kaifeng, Henan, China

[4]Electrical and Computer Engineering, Shanghai Institute of Microsystem and Information Technology, Shanghai, China

Correspondence
Shuang Tang, School of Economics, Henan University, Kaifeng, Henan, China.
Email: 2683826561@qq.com

Funding information
Henan Province Science and Technology Research Project, Grant/Award Numbers: 222102210229, 232102210031

Abstract

The ability of modern detectors to detect small targets is still an unresolved topic compared to their capability of detecting medium and large targets in the field of object detection. Accurately detecting and identifying small objects in the real-world scenario suffer from sub-optimal performance due to various factors such as small target size, complex background, variability in illumination, occlusions, and target distortion. Here, a small object detection method for complex traffic scenarios named deformable local and global attention (DLGADet) is proposed, which seamlessly merges the ability of hierarchical attention mechanisms (HAMs) with the versatility of deformable multi-scale feature fusion, effectively improving recognition and detection performance. First, DLGADet introduces the combination of multi-scale separable detection and multi-scale feature fusion mechanism to obtain richer contextual information for feature fusion while solving the misalignment problem of classification and localisation tasks. Second, a deformation feature extraction module (DFEM) is designed to address the deformation of objects. Finally, a HAM combining global and local attention mechanisms is designed to obtain discriminative features from complex backgrounds. Extensive experiments on three datasets demonstrate the effectiveness of the proposed methods. Code is available at https://github.com/ACAMPUS/DLGADet

1 | INTRODUCTION

Small object detection refers to the task of detecting and localising small objects in images or videos. Small objects are characterized by limited spatial extent and low pixel density compared to larger objects, making their accurate detection challenging [1]. This field has gained significant attention due to its numerous real-world applications in areas such as surveillance, autonomous driving, medical imaging, and robotics. The notion of a ‘small object’ refers to the size of the object being considered. Typically, two distinct approaches are used to characterize small objects. The first approach involves absolute size, as exemplified in the COCO [2] dataset, where any target occupying an area less than 32×32 pixels is considered a small target. Conversely, the second approach relies on relative size, classifying a target that occupies less than 0.12% of the original image dimensions as a small object.

Despite significant advancements [3–9] in computer vision, recognizing small objects in real-world scenarios remains challenging. Figure 1 illustrates some challenging scenarios for detection tasks. The detection results of modern detectors on the COCO dataset show that the detection of small targets is still unsatisfactory compared to that of medium and large objects. This can be attributed to several reasons:

1. Size problem: Small targets are typically small in size, and the deep feature map may have only a few pixels left after the multi-layer convolutional downsampling process, making it difficult to extract effective features.
2. Target deformation problem: Due to different camera shooting angles, the same target exhibits different characteristics at different angles.
3. Receptive field problem: There exists a gap [10] between the theoretical receptive field and the real receptive field of convolutional neural networks (CNNs).

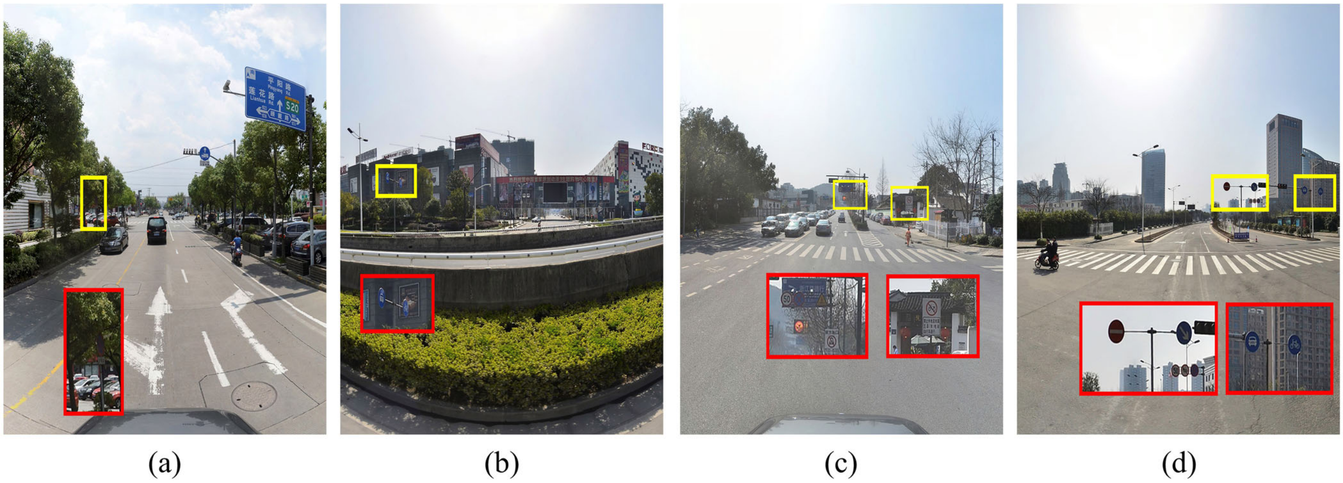

FIGURE 1 Examples of the challenging scenarios for the detection tasks in the dataset. The yellow rectangular boxes in the figures contain the small targets to be identified, and the red rectangular boxes are the corresponding zoom area for easy observation. (a) Poorly illuminated environments. (b) Severe target deformation. (c) Tiny target and complex background. (d) Complex illumination and tiny target.

4. Background complexity and occlusion problem: The backgrounds in which small targets are located in real scenes often have low contrast and high complexity. The colour or brightness of the small targets may be similar to the surrounding environment under different lighting or weather conditions, increasing the difficulty of detection algorithms.

CNN-based computer vision algorithms have shown promising results in detecting small targets. Researchers have proposed various methods [11, 12] to enhance localisation and detection accuracy, such as multi-scale feature fusion network structures [13]. Shallow feature maps contain valuable information like edges and textures for small target localisation, while deep feature maps provide semantic information beneficial for classification. These methods effectively address the scale problem by transferring shallow features to deep layers. However, the transfer process involving multiple convolutional and down-sampling operations often leads to feature loss and limited effectiveness [1]. Furthermore, although multi-scale detection is used, the scale of tiny targets is not adequately considered, as accurate localisation requires higher resolution. Thus, it is crucial to integrate deep semantic information while preserving high resolution for accurate detection of small objects. Small objects often overlap or get occluded by the surrounding background or other objects, making it challenging for detection algorithms to locate the targets accurately. Therefore, the algorithms need to handle the occlusion problem and incorporate an attention mechanism. Some computer vision algorithms [14–17] employ CNNs to extract features from images, utilize data augmentation methods to enrich the features of small targets, and address the resolution issue. However, these algorithms struggle to effectively leverage local contextual features in the presence of significant target deformations. The traditional convolutional approach with fixed kernels sampling fixed locations may struggle to adequately address target deformations, even with multiple convolutional and down-sampling layers. There remains a gap between the theoretical receptive field and the actual receptive field, posing difficulties in handling target deformations [18, 19]. Some algorithms [20–22] employ an attention mechanism to tackle occlusion, suppressing extraneous noise interference and significantly improving detection accuracy. However, these algorithms encounter difficulties in effectively capturing global image feature dependencies while simultaneously emphasizing local features.

To address the aforementioned challenges, we propose an algorithm based on a hierarchical attention mechanism (HAM) with deformable multi-scale detection and fusion, inspired by the concept of a one-stage detection algorithm. Our approach includes the following key components and contributions: First, we introduce Multi-scale separable detection (MSD) and multi-scale contextual feature fusion block (MCFF) to increase the number of detection layers while preserving high-resolution feature maps for detailed information retention. Additionally, we address the misalignment issue between object classification and localisation by decoupling them through the separation of detection heads. Second, we design deformable feature extraction module (DFEM) to dynamically adapt to different target shapes and sizes, effectively addressing target deformation and the problem of insufficient actual receptive field. Finally, the multi-head self-attention (MHSA)[23] mechanism is introduced to capture the long-range dependencies between features, followed by the use of shuffle attention [24] to bias the allocation of the most informative feature expressions. We also use an effective data augmentation strategy to further improve the training accuracy. In summary, our main contributions can be outlined as follows:

1. Proposal of the MSD and multi-scale context feature fusion (MCFF) block for accurate detection of extremely small

targets. Introduction of the decoupling idea to address the misalignment problem between classification and localisation in small object detection.
2. Design of the DFEM module within the feature extraction network to handle target deformation and address the issue of insufficient actual receptive field.
3. Introduction of a HAM to mitigate interference from irrelevant backgrounds and capture global dependencies.
4. Performance improvements on three challenging public datasets outperform current state-of-the-art models while maintaining real-time detection capabilities.

2 | RELATED WORK

2.1 | Small object detection

The challenge of detecting small targets has been addressed from four main perspectives in current research. First, researchers have focused on constructing multi-scale feature representations to extract comprehensive semantic and detailed information. Liu et al. [25] add a deconvolution layer and normalization layer to the output of the convolution layer. They concatenate the features of different layers into a fused feature map and propose a two-stage adaptive classification loss function to improve training effectiveness. Second, contextual information can be utilized to establish the relationship between the target and its surroundings. Yuan et al. [26] use a multi-resolution feature fusion approach and a vertical spatial sequence attention module. Their network consists of two stages: the first stage extracts multi-resolution feature maps using MobileNet and deconvolution layers, while the second stage constructs a vertical spatial sequence attention module to fully exploit context information. Third, super-resolution images can be constructed to localise more details. Ren et al. [27] utilize a region context network (RCN) as the backbone for efficient feature extraction. They incorporate a generative adversarial network (GAN) with distribution transformation and super-resolution enhancement modules to improve target visibility and resolution. The GAN-based approach effectively enhances the detailed information of images, especially for super-resolution applications. It can be applied to any type of generator network, without the need for a specific architecture. However, the training process of GANs is challenging. Last, data augmentation strategies can be utilized. Wang et al. [28] employ an image segmentation strategy to augment the data and increase the number of small objects, thereby facilitating the full training of the algorithm.

2.2 | Multi-scale features representations

The feature pyramid network (FPN) [29] serves as the foundation for constructing multi-scale feature representations. Utilizing multi-scale features effectively mitigates the challenges posed by target scale variation, as it integrates semantic information with high-resolution details to enhance object detection. Numerous studies have explored approaches to enhance multi-scale feature representation. Liu et al. [30] improve information flow by augmenting the entire feature hierarchy with precise localisation signals in lower layers through bottom-up path augmentation, thereby shortening the information path between lower layers and the topmost feature. Qiao et al. [13] reformulate feature pyramid construction as a feature reconfiguration process and introduce a new reconfiguration architecture that efficiently combines low-level representations with high-level semantic features in a highly non-linear manner. Tan et al. [31] propose a weighted bi-directional FPN (BiFPN) for rapid multi-scale feature fusion. Wang et al. [32] present a novel multi-scale context-aware FPN that enhances object detection performance by addressing the context information gap across different levels, yielding substantial improvements compared to existing FPN-based methods on the MS-COCO dataset.

2.3 | Attention mechanism

Attention mechanisms have gained widespread popularity in the field of deep learning due to their ability to enhance the performance of deep learning models in various computer vision tasks, such as image recognition, object detection, and semantic segmentation. These mechanisms enable models to focus on the most significant features and allocate processing resources effectively, leading to a better understanding of feature relationships and more accurate predictions. The squeeze-and-excitation block [33] dynamically recalibrates channel-wise feature responses in a CNN by computing a global average pooling (GAP) of the feature maps, which summarizes spatial information. It then models the channel relationships using two fully connected layers. The resulting channel weights are used to rescale the feature maps, allowing the network to prioritize the most informative feature channels. Guo et al. [34] introduce a novel linear attention mechanism called large kernel attention (LKA), which enables self-adaptive and long-range correlations, addressing the limitations of traditional self-attention. They also present a visual attention network (VAN) that utilizes LKA and outperforms similarly sized vision transformers (ViTs) and CNNs across multiple tasks. Zhang et al. [35] present the ensemble transformer with attention modules (ETAM) encoder, a powerful approach for detecting small objects by extracting subtle features. This method leads to significant improvements in small object detection performance across multiple datasets. The ViT has garnered considerable attention in the field of computer vision for its remarkable performance in various image tasks. Instead of the conventional convolutional and pooling layers used in computer vision, ViT employs self-attention mechanisms. This substitution enables the model to selectively attend to different regions of an image at various scales, capturing complex long-range dependencies between pixels.

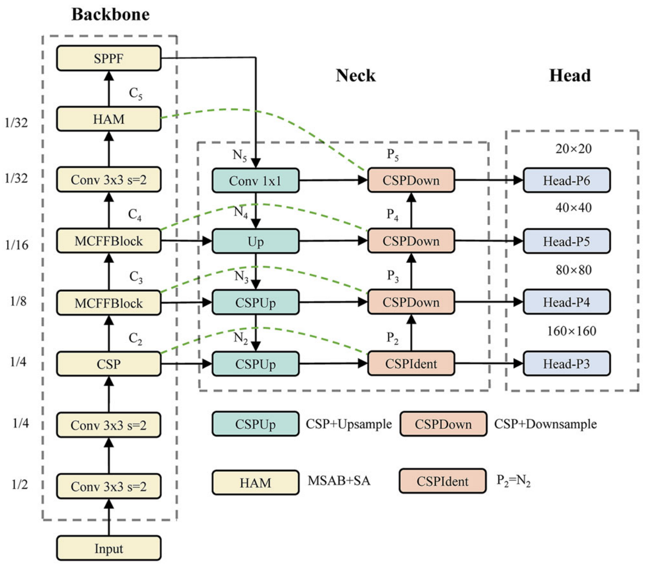

FIGURE 2 Architecture of our proposed deformable local and global attention. Initially, the input image is fed into the backbone network, which applies the MCFFBlock (multi-scale context feature fusion) module proposed in Section 3.1 and the hierarchical attention mechanism. The backbone network extracts essential features from the input image. Next, the features are passed to the Neck section for further processing. The Neck section includes CSPUp and CSPDown modules, which are variations of the cross stage partial (CSP) modules [36] with upsampling and downsampling operations based on CSPDarknet. These modules utilize convolution, batch normalization, and Silu activation functions. The numbers S 3 × 3 and 1 × 1) denote the sizes of the convolution kernels, and 's' represents the stride. The Neck section refines the feature map by aggregating features from different backbone layers into different detector levels. Finally, the proposed Multi-scale separable detection (MAD) detector head is employed to predict boxes at four different scales. For more detailed information on the structure of the MSD, please refer to Figure 3d.

3 | PROPOSED METHODS

This section presents the deformable local and global attention (DLGADet) model, which is depicted in Figure 2. The section is organized as follows: In Section 3.1, we introduce the general structure of the model, including the (MSD module and the MCFF module. Next, in Section 3.2, we introduce the proposed deformation feature extraction module (DFEM). Last, in Section 3.3, we provide a detailed description of the proposed HAM.

3.1 | Multi-scale context feature fusion block and multi-scale separable detection

The design of the method follows the following principle: enhancing the ability to capture small targets by leveraging shallow layer features such as textures and edges for localisation, while propagating features from lower to higher layers to improve the overall localisation ability of the feature layer. This is based on the observation that shallow textures and edges are more conducive to localisation. Two key improvements have been made. The first improvement is the introduction of the MCFFBlock for the backbone network. Additionally, a skip connection to the backbone network has been added to the head part, similar to the feature fusion operation used in FPN. This is depicted by the green dashed section in Figure 2. The second improvement involves the inclusion of a downsampling quadruple detection layer with a decoupling header design.

3.1.1 | Multi-scale context feature fusion block

Given the input feature map X_{c_i} of layer i in backbone, the specific flow from MCFFBlock to the detection head can be expressed as the following process:

$$F_{f_M} = F_{\text{conv}}(F_{\text{conv}}(X_{c_i})) + F_{\text{csp\_n}}(F_{\text{conv}}(X_{c_i})) \tag{1}$$

$$F_{f_P} = \Phi(F_{\text{conv}}(X_{P_{i-1}})) \tag{2}$$

$$F_{f_N} = \hbar(F_{\text{conv}}(X_{N_{i+1}})) \tag{3}$$

$$F_{fm_i} = \psi(F_{\text{conv}}(\psi(F_{f_M}, F_{f_N})), F_{f_P}) + F_{f_M} \tag{4}$$

Here, the convolution operation is denoted as $F\text{conv}(\cdot)$, $\hbar(\cdot)$ represents the CSP module and upsampling operation, Φ represents the CSP module and downsampling operation, $F_{\text{csp\_n}}$ represents n CSP modules, and $\psi(\cdot)$ represents a channel direction concatenation operation. Each convolution operation is followed by batch normalization and Silu activation functions, which are not shown for simplicity. The top-down feature map is denoted as $X_{N_{i+1}}$, and the bottom-up feature map is denoted as $X_{P_{i-1}}$. F_{f_M} represents the feature map obtained through MCFF processing, F_{f_N} represents the feature map obtained through convolution, CSP module, and upsampling, and F_{f_P} represents the feature map obtained through convolution, CSP module, and downsampling. As shown in Figure 3a, the input feature map needs to undergo several convolutional operations and $N \times CSP$ modules in the backbone network. However, this process inevitably leads to the loss of shallow textures, edges, and other features that contribute to localisation during network deepening. In contrast, the lateral connection in MCFF consists of only two convolutional layers, and the fused feature maps are directly subjected to an elementwise summation operation with the feature maps of the Neck part. This allows shallow features that aid in localisation to be merged into the deeper network, thereby enhancing the model's ability to detect small targets. Additionally, this design provides a shortcut for the gradient backpropagation process of the higher-level network, speeding up the convergence of the network.

3.1.2 | Multi-scale separable detection

We introduce an additional feature layer specifically designed for small target detection, which is added on top of the original detection feature layer. This expands the total number of detection layers from three to four, known as MSD as

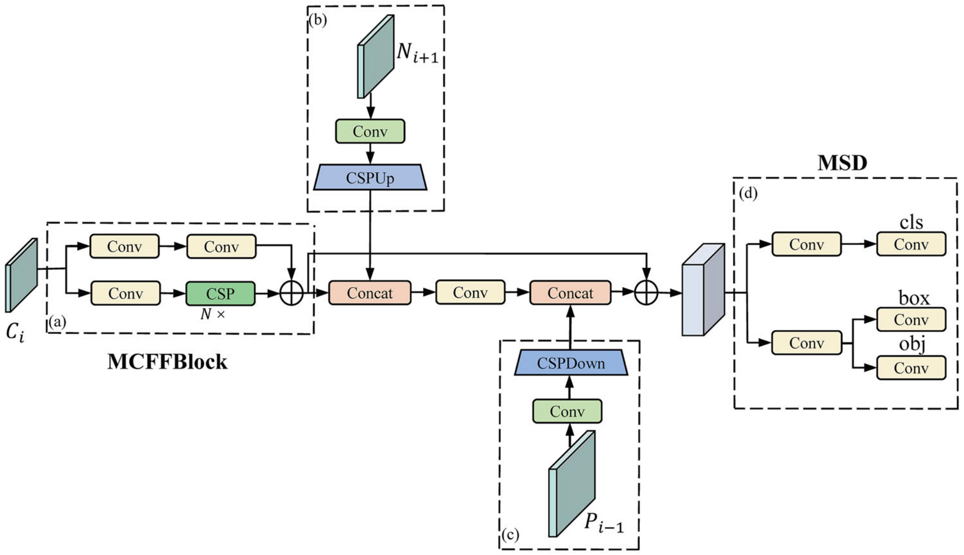

FIGURE 3 A detailed explanation of the partial process in Figure 2 is as follows: Each letter 'i' indicates a feature layer that has been downsampled 2^i times. For example, C_3 represents downsampling 8 times. Let us consider the feature map processing of the C_3 layer as an example. The C_3 layer undergoes processing through the MCFFBlock and is concatenated with the upsampled feature map of the N_4 layer, resulting in the feature map N_3. N_3 is then processed by the cross stage partial (CSP) module and upsampled to obtain N_2 (P_2). Additionally, N_3 undergoes a convolution operation and is concatenated with the downsampled feature map of P_2, resulting in the feature map P_3. To enhance the shallow features in P_3, a skip connection is established between the feature map obtained from the MCFFBlock and P_3. The obtained P_3 is further processed using the proposed MSD to obtain the classification and position coordinate detection results.

shown in Figure 3d. To address the misalignment between classification and localisation in object detection, we use an approach that separates the classification and regression tasks, which is achieved by replacing the 1 × 1 convolution with two parallel detection layers, each stacked with convolutional layers [37-39]. Through this approach, our network is able to effectively capture semantic features while maintaining high resolution for accurate localisation, thus solving the problem of classification and localisation misalignment simultaneously. Notably, when comparing the training curves of our decoupled head method with the YOLO [17] head method on the TT100k dataset, as shown in Figure 4, our approach demonstrates superior performance by converging faster and achieving better results.

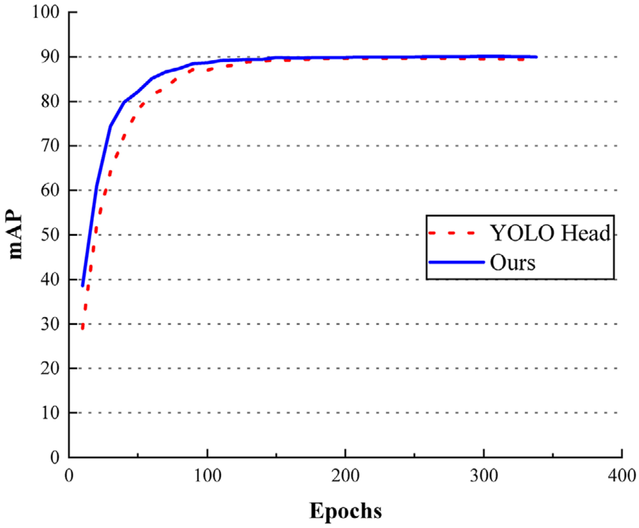

FIGURE 4 The mean average precision training curve of our head and YOLO head.

3.2 | Deformation feature extraction module

3.2.1 | Starting with deformable convolution

We extract complex geometric transformation visual features using deformable convolution. Deformable convolution introduces a 2D offset to the regular grid sampling locations of standard convolution, allowing adaptable geometric structures for the convolution kernels. In Figure 5, the basic block of the DFEM module resembles the widely used bottlenecks in traditional CNNs. However, instead of using ordinary convolution, we employ deformable convolution (denoted as DCNv2). Deformable convolution is a variant of standard convolution that enables the modulation of spatial sampling locations within the convolutional kernel. Let us begin with the standard convolutional operation. Given an output feature map y, the operation can be formulated as:

$$y(p_0) = \sum_{p_n \in V} w_k(p_n) \cdot x(p_0 + p_n) \tag{5}$$

Here, p_0 represents the current location of the output feature map, w_k denotes the weight, p_n represents each location in set V, and V is the set of sampling locations on the predefined grid ($V = (-1, -1), (-1, 0), \ldots, (0, +1), \ldots, (+1, +1)$) over the input feature map x. Deformable convolution introduces the use of learnable offset values, which replace the fixed

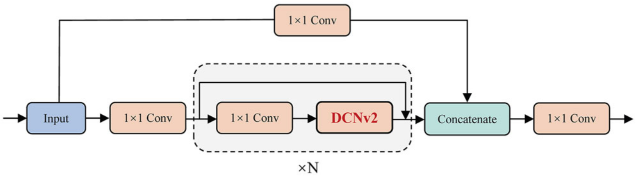

FIGURE 5 The architecture of the deformation feature extractionmodule.

kernel shifts. These offsets are integrated into the kernel positions before performing the convolution. Mathematically, this operation can be represented as:

$$y(p_0) = \sum_{p_n \in V} w_k(p_n) \cdot x(p_0 + p_n + \Delta p_n) \tag{6}$$

Here, Δp_n represents the learnable offset, and each new sampling location is adjusted with an offset. Bilinear interpolation is commonly used to compute the equation above.

3.2.2 | Extending DCNv2 for deformation feature extraction

However, this operation suffers from the issue of extending beyond the region of interest and influencing features with irrelevant image content. To address this, we introduce a modified version denoted as follows:

$$y(p_0) = \sum_{p_n \in V} w_k(p_n) \cdot x(p_0 + p_n + \Delta p_n) \cdot \Delta m_k \tag{7}$$

Here, Δp_n and Δm_k are learnable parameters. They are obtained from 3K channels, which are the result of another convolutional layer having the same input and spatial resolution as the current convolutional branch. The first 2K channels correspond to Δp_n, while the remaining K channels are passed through a sigmoid activation function to obtain modulation scalars. It is worth noting that Δp_n is an unrestricted value obtained after training, whereas Δm_k is constrained between 0 and 1. This constraint ensures that the aforementioned problem is resolved. To achieve adaptive extraction of deformation targets, we simply replace the 3×3 convolution in the CSP module of the MCFFBlock with DCNv2. Unlike standard convolution, which applies the same operation to all inputs with fixed parameters, deformable convolution can dynamically adjust to fit different target shapes and sizes based on the input content. This design allows for the adaptation to various small deformed traffic signs, dynamically expanding the model's perceptual field, and effectively reducing gradient reuse and computational complexity [36]. The effectiveness of our method is demonstrated in the experiment in Section 4.

3.3 | Hierarchical attention mechanism

3.3.1 | MSAB module

We propose a HAM called HAM, which combines a multi-head self-attention block (MSAB) and shuffle attention (SA). The design of the MSAB block is as follows: First, we embed a multi-headed self-attention (MHSA) layer into the CSP-DarkNet backbone, replacing the 3×3 convolution operation in the top CSP module of the bottleneck. Next, we apply SA operations to the obtained feature maps. Given the output feature map X from the backbone network, we flatten X to $X \in \mathbb{R}^{C \times (H*W)}$, where H and W represent the original resolution of the feature map. The processing flow of the MHSA module with M heads can be described as follows:

$$X = \text{position\_embedding}(X) + X \tag{8}$$

$$Q = XW_q, \quad K = XW_k, \quad V = XW_v \tag{9}$$

$$f^{(m)} = \psi\left(\frac{Q^{(m)} \times (K^{(m)})^T}{\sqrt{d}}\right) V^{(m)}, \quad m \in [1, \dots, M] \tag{10}$$

$$f = \text{Concat}(f^{(1)}, \dots, f^{(m)}) W_p \tag{11}$$

Here, position_embedding(·) represents the position embedding vector. We use a learnable one-dimensional position embedding and a linear layer to obtain the position information. The function $\psi(\cdot)$ denotes the softmax function, and d=C/M represents the dimension of each head. f^m represents the mth attention head, and $Q^{(m)}$, $K^{(m)}$, and $V^{(m)}$ are obtained by linearly transforming X. W_q, W_k, W_v, and W_p denote the corresponding weight matrices. Finally, the obtained matrix X is resized back to the original dimensions $C \times H \times W$ and used as input to the SA module.

3.3.2 | SA module

The SA module divides the feature map into g groups along the channel direction, resulting in matrices $[X_1, \dots, X_g]$. Each X_k in the g feature maps has a shape size of $H \times W \times C/g$. X_k is further split into X_{k1} and X_{k2} along the channel direction, with a shape size of $H \times W \times C/2g$. Channel attention and spatial attention operations are performed on each X_{k1} and X_{k2}, respectively. The attention processing can be described as follows:

$$s = \text{GAP}(X_{k1}) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} X_{k1}(i, j) \tag{12}$$

Here, GAP denotes global average pooling, which generates a compact feature to guide precise and adaptive selection using a gating mechanism incorporating a sigmoid activation function.

The final result is obtained by scaling X_{k1} as follows:

$$X'_{k1} = \sigma(W_1 s + b_1) \cdot X_{k1} \tag{13}$$

where W_1 represents the scaling parameter and b_1 represents the bias. Spatial attention is complementary to channel attention and is performed after group norm (GN) [40]. The spatial attention operation is similar to the channel attention operation and can be expressed as:

$$X'_{k2} = \sigma(W_2 \cdot GN(X_{k2}) + b_2) \cdot X_{k2} \tag{14}$$

Here, σ represents the sigmoid activation function, W_2 is the scaling parameter, b_2 is the bias, and GN denotes group norm. Finally, a 'channel shuffle' operator is applied to transfer information across groups along the channel dimension. The output of the SA module maintains the same dimensions as the input, making it convenient to integrate into contemporary architectural designs.

3.3.3 | Designing rules

To prevent the premature enforcement of regression boundaries and preserve meaningful context information, we limit the usage of the transformer layer. Additionally, considering the low resolution of the high-level feature map, this approach helps reduce computational complexity and memory space occupation. Consequently, in our HAM module, the MSAB module is only employed in the penultimate third layer of the backbone section. It is then followed by the shuffle attention (SA) module, which combines channel attention and spatial attention.

4 | EXPERIMENT

Here, we provide an overview of the benchmark datasets in Section 4.1. Following that, we describe the implementation details and evaluation metrics used in the experiments in Section 4.2. Next, we conduct a comprehensive performance evaluation of the proposed DLGADet model, comparing it with the current state-of-the-art convolution-based neural network model on the publicly available TT100K dataset in Section 4.3. Furthermore, we present a thorough ablation experiment of the proposed method in Section 4.4. Finally, we test the proposed method on VisDrone2019 [41] and SODA-D [1] in Section 4.5.

4.1 | Benchmark datasets

Experiments were conducted using three datasets. The primary experiments were carried out on the challenging TT100K dataset, which serves as a comprehensive benchmark with 100,000 Tencent Street View panoramas. Additionally, the proposed method was tested on the VisDrone2019 and SODA-D datasets. It is worth noting that all of the selected datasets consist of a significant number of small objects, each measuring less than 32 × 32 pixels.

4.2 | Implementation details and evaluation criterion

The original TT100K dataset consists of approximately 150 classes of traffic signs. We followed a similar approach as in reference [42] and excluded classes with fewer than 100 instances, resulting in a final training dataset comprising 45 classes of traffic signs. For classes with 100 to 1000 instances, we utilized data augmentation techniques to expand their quantity to 1000 instances. Drawing inspiration from the cut–mix [43] and copy–paste [44] data augmentation strategies, we employed a similar technique of pasting the targets into different training images. However, we did not introduce new background images that were not present in the original training set. The training phase involved approximately 6103 images with various resolutions, while 3067 images were reserved for testing. All experiments were conducted using the PyTorch deep learning framework on a cloud server equipped with an NVIDIA GeForce RTX A5000 (24GB) GPU, a 15 vCPU Intel(R) Xeon(R) Platinum 8358P CPU, and 80GB of RAM. The models were pretrained using weights based on the COCO dataset and fine-tuned on the TT100K dataset using the DLGADet model to expedite the training process. The experimental hyperparameter settings followed those outlined in YOLOv5. The initial learning rate was set to 0.01 and gradually decayed to 0.0001 using cosine annealing. We employed a stochastic gradient descent optimizer with a batch size of 16 and trained the models for 400 epochs. Here, various evaluation metrics were employed for comparisons, including mean average precision (mAP), average precision for small objects (APS), average precision for medium objects (APM), average precision for large objects (APL), precision, recall, frames per second (FPS), and F1 measure.

4.3 | Comparison with state-of-the-arts

The detection results were analyzed in detail on the TT100K dataset using images of different resolutions. Table 1 shows the results.

We compared our method with other state-of-the-art detectors on the TT100K benchmark. As shown in Table 1, the proposed DLGADet model achieves an mAP of 92.0 at a resolution of 640, which is 1.5 AP higher than the state-of-the-art competitors. Notably, in terms of APS, which is an important metric for small target detection, DLGADet shows significant improvements. It outperforms YOLOv5-L by 1.1 APS and 0.4 APS at resolutions of 640 and 1024, respectively. Compared to general-purpose state-of-the-art target detectors, DLGA demonstrates considerable advancements in addressing the challenges of small target detection. This improvement can be attributed to DLGADet's effective handling of scale processing, feature preservation during information transmission in small object detection, object deformation, occlusion, and the

TABLE 1 Comparison of latest state-of-art detectors on TT100K.

Resol-ution	Method	mAP	APS	APM	APL	Prec-ision	Recall	F1 measure
640 × 640	YOLOv3	84.3	37.3	67.3	68.3	-	-	-
	YOLOv4-CSP	86.2	43.6	72.6	81.3	85.0	79.7	82.3
	YOLOv5-L	88.3	55.6	74.4	83.1	86.9	85.5	86.2
	YOLOv6-L	88.2	46.4	77.9	85.5	83.8	81.0	82.4
	YOLOR-CSP-X	89.0	47.8	74.3	83.3	84.6	84.6	84.6
	YOLOv7	90.1	50.3	76.2	82.5	87.7	84.8	86.2
	YOLOv7-X	90.5	50.4	76.0	84.2	87.2	84.7	85.9
	Ours	92.0	57.7	77.4	83.7	90.6	89.2	89.9
1024 × 1024	YOLOv3	92.7	55.1	74.2	80.5	-	-	-
	YOLOv4-CSP	91.8	54.4	77.8	85.9	87.2	89.0	88.1
	YOLOv5-L	94.5	62.3	79.9	87.1	91.7	93.0	92.3
	YOLOv6-L	88.2	46.4	77.9	85.5	83.8	81.0	82.4
	YOLOR-CSP-X	92.3	53.7	77.5	86.3	90.0	89.6	89.8
	YOLOv7	93.8	57.6	79.0	87.9	89.0	90.6	89.8
	YOLOv7-X	95.1	61.1	79.8	88.8	91.0	91.9	91.4
	Ours	94.8	62.7	79.9	87.1	92.0	93.7	92.8

recognition of tiny targets as critical issues. It should be noted that DLGADet may not perform optimally in terms of APM and APL metrics at a resolution of 640. However, when compared to the primary focus of this paper, which is small object detection, the difference in accuracy for large object detection is minimal compared to other state-of-the-art detectors. This discrepancy could be attributed to DLGADet's passing of more shallow features to the upper feature map, potentially affecting the model's ability to discern semantic information in the upper feature map. Furthermore, although the mAP metric is lower than that of YOLOv7-X at a resolution of 1024, it is important to note that YOLOv7-X is a larger and more complex model with a higher number of parameters. Additionally, DLGADet still outperforms YOLOv7-X by 1.6 APS. Finally, our proposed method achieves the highest F1-measure metric at both resolutions, indicating that DLGADet attains an optimal balance between precision and recall.

Figure 6 illustrates the temporal variation of AP values, precision, and recall across epochs for our proposed model, as well as YOLOv4, YOLOv5-L, YOLOv7, YOLOv7-X, and YOLOR-CSP-X. Upon careful analysis of the figures, it is clear that our proposed model consistently outperforms the other models in terms of AP metrics and exhibits a faster convergence rate.

Table 2 shows the experiment that evaluated the performance of several methods for traffic sign recognition on the TT100K dataset. The evaluation metrics used in this experiment were precision, recall, and F1 measure. Precision measures the percentage of correctly recognized traffic signs out of all the detected traffic signs. Recall measures the percentage of correctly recognized traffic signs out of all the actual traffic signs. F1 measure is the harmonic mean of precision and recall. Based on the experimental results, it is evident that the proposed method outperforms the other evaluated methods in the majority of cases, exhibiting the highest F1 measure across various traffic sign symbols such as i2, il100, il60, and so on. It is worth noting that we also found that DLGADet does not perform as an optimal solution on all categorical entries. This may be attributed to the fact that we bring shallow features to the deeper layers resulting in ambiguity and loss of semantic information, which in turn is closely related to classification accuracy. Furthermore, in the instances where our method does not achieve the top performance, the difference between its F1 measure and the best result is negligible. These findings underscore the accuracy and robustness of the proposed method, which demonstrates its potential as a reliable and promising approach for traffic sign recognition tasks.

Figure 7 showcases the robustness of the proposed model in detecting small traffic signs under challenging conditions, including deformation, occlusion, illumination variations, small object size, and combinations thereof. To facilitate visualization and comprehensive analysis of the model's detection performance, the recognition results are magnified and displayed at the bottom of the figure. Remarkably, our model demonstrates precise recognition of small objects that make up less than 1% of the overall scene, even in extremely unfavourable conditions. This highlights the model's ability to handle objects with low visibility, thereby elevating its potential for practical applications.

4.4 | Ablation studies

Table 3 presents the results of an ablation experiment conducted on the TT100K dataset to evaluate the effectiveness of different components in a method. The components investigated include the MCFF block, MSD, shuffle attention (SA), data augmentation (*) described in Section 4.2, (DFEM, and MSAB. The evaluation metric used in this study is mAP, expressed as a percentage.

The results of the ablation experiment demonstrate that the inclusion of each component improves the overall performance, as measured by mAP. The initial baseline, without any components, achieves an mAP of 88.3%. The addition of the MCFF block increases the mAP to 89.5%, while the incorporation of MSD further improves it to 90.1%. The introduction of DFEM has a more substantial impact, raising the mAP to 91.7%. The highest mAP value of 92.0% is achieved when all components are used together. Through analysis, it has been determined that the synergistic utilization of all constituent elements shows potential for further advancement in the model's performance. This observation suggests that the functions of the aforementioned modules do not conflict with each other.

We conducted ablation experiments on the MCFFBlock component. In the Neck section, the elementwise summation from the backbone to the Neck feature fusion strategy is denoted as 'Fusion' For the ablation experiments, we excluded the MCFFBlock component ('-MCFFBlock'), the fusion strategy ('-Fusion'), and both components ('-All'). The results of these experiments, as shown in Table 4, demonstrate the

TABLE 2 Comparison of each category on TT100K.

Methods	Metrics	i2	i4	i5	il100	il60	il80	io	ip	p10	p11	p12	p19	p23	p26	p27
YOLOv4-CSP	Precision	87.3	86.5	90.6	94.9	93.4	84.0	74.6	85.6	79.9	85.7	96.1	96.7	94.4	85.3	100.0
	recall	77.8	89.6	95.0	95.2	98.6	96.9	89.1	92.9	77.8	90.1	75.6	88.5	90.3	89.2	89.0
	F1-measure	82.3	88.0	92.7	95.0	95.9	90.0	81.2	89.1	78.8	87.8	84.6	92.4	92.3	87.2	94.2
YOLOv5-L	precision	87.6	92.1	93.7	92.4	95.2	87.6	90.2	88.4	86.8	89.8	86.8	87.7	92.3	86.2	94.1
	recall	85.2	95.2	94.6	97.4	98.3	88.2	89.5	89.1	82.9	81.9	75.8	87.9	93.1	90.5	89.4
	F1-measure	86.4	93.6	94.1	94.8	96.7	87.9	89.8	88.7	84.8	85.7	80.9	87.8	92.7	88.3	91.7
YOLOv6-L	precision	86.7	95.4	95.2	95.0	98.5	92.1	88.0	90.1	82.9	90.8	84.3	96.7	97.0	85.4	95.5
	recall	82.0	81.0	82.0	97.0	94.0	96.0	80.0	92.0	78.0	81.0	89.0	87.0	93.0	87.0	89.0
	F1-measure	84.3	87.6	88.1	96.0	96.2	94.0	83.8	91.0	80.4	85.6	86.6	91.6	94.9	86.2	92.1
YOLOR-CSP-X	precision	83.1	88.6	93.3	84.9	96.3	88.6	77.7	85.9	82.3	86.5	78.4	95.2	94.9	84.6	98.3
	recall	84.5	94.8	95.6	92.3	97.1	97.1	90.4	94.8	80.4	91.3	81.8	93.9	90.3	88.9	91.5
	F1-measure	83.8	91.6	94.4	88.4	96.7	92.7	83.6	90.1	81.3	88.8	80.1	94.5	92.5	86.7	94.8
YOLOv7	precision	85.6	90.8	92.3	95.6	95.8	91.1	77.6	89.9	83.7	89.7	90.5	96.9	94.0	89.6	97.7
	recall	83.7	94.5	95.4	100.0	98.6	99.0	88.3	94.5	78.2	91.1	92.4	93.3	93.2	87.6	91.9
	F1-measure	84.6	92.6	93.8	97.8	97.2	94.9	82.6	92.1	80.9	90.4	91.4	95.1	93.6	88.6	94.7
YOLOv7-X	precision	86.7	86.4	91.6	92.5	97.6	86.2	80.9	86.8	83.7	88.2	86.5	92.7	94.1	86.9	93.9
	recall	85.2	95.2	94.9	95.1	97.9	97.9	88.7	95.3	77.0	90.1	93.9	93.9	92.3	90.8	95.7
	F1-measure	85.9	90.6	93.2	93.8	97.7	91.7	84.6	90.9	80.2	89.1	90.0	93.3	93.2	88.8	94.8
Ours	precision	**89.2**	91.3	93.9	**97.3**	**99.3**	**92.1**	87.0	**91.4**	**87.0**	**92.6**	91.1	93.6	95.9	89.0	**100.0**
	recall	**86.8**	94.8	95.3	**100.0**	**97.9**	**99.0**	88.7	89.8	**84.8**	90.1	93.5	**93.9**	**94.2**	**90.9**	92.9
	F1-measure	**88.0**	93.0	**94.6**	**98.6**	**98.6**	**95.4**	87.8	90.6	**85.9**	**91.3**	**92.3**	93.7	**95.0**	**89.9**	**96.3**
Methods	**Metrics**	**p3**	**p5**	**p6**	**pg**	**ph4**	**ph4.5**	**ph5**	**pl100**	**pl120**	**pl20**	**pl30**	**pl40**	**pl5**	**pl50**	**pl60**
YOLOv4-CSP	Precision	83.2	87.9	72.2	88.1	95.2	74.9	71.2	93.3	97.4	82.6	95.2	90.2	87.8	89.7	95.9
	recall	81.0	92.1	56.4	95.3	67.6	85.0	42.5	96.3	86.6	67.7	68.3	76.1	79.8	69.4	69.1
	F1-measure	82.1	90.0	63.3	91.6	79.1	79.6	53.2	94.8	91.7	74.4	79.5	82.6	83.6	78.3	80.3
YOLOv5-L	precision	90.2	88.8	78.2	88.4	79.4	84.6	87.1	92.4	90.8	80.6	87.1	88.5	85.4	88.3	85.7
	recall	79.7	93.9	83.0	95.3	70.3	83.3	67.4	92.6	90.8	82.1	82.5	87.0	89.4	85.3	83.2
	F1-measure	84.6	91.3	80.5	91.7	74.6	83.9	76.0	92.5	90.8	81.3	84.7	87.7	87.4	86.8	84.4
YOLOv6-L	precision	98.0	91.4	86.5	91.1	85.7	88.3	82.8	95.4	96.5	94.9	92.1	90.7	92.3	80.1	93.8
	recall	84.0	89.0	82.0	95.0	81.0	88.0	60.0	95.0	95.0	66.0	74.0	79.0	83.0	77.0	77.0
	F1-measure	90.5	90.2	84.2	93.0	83.3	88.2	69.6	95.2	95.7	77.8	82.1	84.5	87.4	78.5	84.6
YOLOR-CSP-X	precision	79.2	92.3	64.7	84.5	97.9	78.6	70.3	95.4	97.6	85.7	96.2	87.4	86.9	91.5	96.4
	recall	87.9	94.9	69.2	95.3	62.2	88.3	60.0	96.8	94.7	74.9	73.3	84.0	87.7	75.9	72.6
	F1-measure	83.3	93.6	66.9	89.6	76.1	83.2	64.7	96.1	96.1	79.9	83.2	85.7	87.3	83.0	82.8
YOLOv7	precision	87.9	92.5	81.1	77.6	100.0	73.8	73.6	96.4	97.7	88.6	94.3	92.0	88.0	90.7	93.2
	recall	91.4	94.7	76.8	96.7	68.9	85.0	70.0	97.2	96.1	55.7	72.7	79.6	85.7	77.1	69.0
	F1-measure	89.6	93.6	78.9	86.1	81.6	79.0	71.8	96.8	96.9	68.4	82.1	85.4	86.8	83.3	79.3
YOLOv7-X	precision	82.0	90.6	80.9	82.7	96.4	79.9	75.5	96.8	97.6	89.6	98.0	90.5	88.9	92.4	97.4
	recall	91.4	92.4	74.4	97.7	72.6	85.9	65.0	97.5	94.3	64.3	72.1	85.2	85.2	75.4	71.5
	F1-measure	86.4	91.5	77.5	89.6	82.8	82.8	69.9	97.1	95.9	74.9	83.1	87.8	87.0	83.0	82.5
Ours	precision	96.4	90.8	**89.0**	85.2	96.0	84.7	**90.5**	96.6	**97.7**	88.6	89.1	**94.5**	89.5	**92.9**	95.4
	recall	**93.3**	**96.6**	**84.6**	**97.7**	73.0	**88.3**	**71.9**	96.8	**96.3**	**82.9**	**86.8**	**88.1**	**92.7**	84.5	**87.6**
	F1-measure	**94.8**	**93.6**	**86.7**	91.0	82.9	86.5	**80.1**	96.7	**97.0**	**85.7**	**87.9**	**91.2**	**91.1**	**88.5**	**91.3**
Methods	**Metrics**	**pl70**	**pl80**	**pm20**	**pm30**	**pm55**	**pn**	**pne**	**po**	**pr40**	**w13**	**w32**	**w55**	**w57**	**w59**	**wo**
YOLOv4-CSP	Precision	100.0	95.5	84.3	92.3	78.8	86.4	88.3	80.0	80.5	48.7	62.2	77.6	78.6	73.1	61.0
	recall	54.6	69.9	69.4	74.7	76.3	94.6	96.4	65.6	98.1	73.4	70.6	75.0	90.2	84.5	23.7
	F1-measure	70.6	80.7	76.1	82.6	77.5	90.3	92.2	72.1	88.4	58.6	66.1	76.3	84.0	78.4	34.1

(Continues)

TABLE 2 (Continued)

Methods	Metrics	pl70	pl80	pm20	pm30	pm55	pn	pne	po	pr40	w13	w32	w55	w57	w59	wo
YOLOv5-L	precision	95.0	83.6	83.5	70.4	92.7	91.9	96.5	81.8	89.2	77.6	84.9	80.0	88.0	75.5	72.1
	recall	87.0	89.2	73.5	71.9	86.8	90.9	96.2	73.7	96.8	83.9	70.6	80.0	92.6	89.7	61.1
	F1-measure	90.8	86.3	78.2	71.1	89.7	91.4	96.3	77.5	92.8	80.6	77.1	80.0	90.2	82.0	66.1
YOLOv6-L	precision	100.0	92.8	86.4	92.3	88.1	86.8	92.9	78.8	98.4	88.9	80.0	86.4	92.9	76.8	74.1
	recall	70.0	84.0	77.0	75.0	97.0	85.0	85.0	71.0	95.0	77.0	82.0	85.0	86.0	91.0	52.0
	F1-measure	82.4	88.2	81.4	82.8	92.3	85.9	88.8	74.7	96.7	82.5	81.0	85.7	89.3	83.3	61.1
YOLOR-CSP-X	precision	100.0	85.6	83.9	89.9	69.1	87.3	92.7	76.7	78.7	51.1	73.4	82.2	80.1	67.1	66.4
	recall	67.6	86.9	75.5	83.6	81.6	94.7	96.7	67.8	95.2	80.6	82.4	76.8	92.4	89.7	52.6
	F1-measure	80.7	86.2	79.5	86.6	74.8	90.8	94.7	72.0	86.2	62.5	77.6	79.4	85.8	76.8	58.7
YOLOv7	precision	92.8	90.0	88.0	86.5	79.5	89.2	92.0	84.4	90.4	59.2	73.3	91.5	83.5	72.4	84.9
	recall	68.2	85.4	74.9	80.4	86.8	93.8	96.4	70.7	96.8	77.4	82.4	78.3	93.4	90.3	44.7
	F1-measure	78.6	87.6	80.9	83.3	83.0	91.4	94.1	76.9	93.5	67.1	77.6	84.4	88.2	80.4	58.6
YOLOv7-X	precision	97.0	91.0	89.2	86.4	78.0	88.3	91.4	82.2	88.0	64.1	76.1	86.2	87.4	71.6	72.0
	recall	72.5	85.0	75.5	79.6	83.7	93.8	97.1	66.6	98.4	87.1	75.0	73.1	91.0	91.1	34.2
	F1-measure	83.0	87.9	81.8	82.9	80.7	91.0	94.2	73.6	92.9	73.9	75.5	79.1	89.2	80.2	46.4
Ours	precision	97.2	91.4	**96.0**	**94.5**	79.6	**92.1**	94.9	**89.2**	93.1	64.7	**89.8**	90.3	89.9	75.2	69.5
	recall	79.5	**93.0**	**85.7**	**84.4**	94.7	92.4	96.4	73.0	96.8	**90.3**	**91.2**	83.3	92.6	**94.8**	52.6
	F1-measure	**87.5**	**92.2**	**90.6**	**89.2**	86.5	**92.2**	95.6	**80.3**	94.9	75.4	**90.5**	**86.7**	**91.2**	**83.9**	59.9

TABLE 3 Ablation studies of components on TT100K.

Baseline	MCFF	MSD	SA	DFEM	MSAB	mAP(%)
YOLOv5						88.3
YOLOv5	✓					89.5
YOLOv5	✓	✓				90.1
YOLOv5	✓	✓	✓			90.2
YOLOv5*	✓	✓	✓	✓		91.7
YOLOv5*	✓	✓	✓	✓	✓	92.0

TABLE 4 Effect of the MCFFBlock and feature fusion strategy.

Component	AP50:95	AP50	Precision	Recall	F1 measure
Base	70.0	89.5	89.5	87.9	87.5
-MCFFBlock	69.2	88.6	89.2	85.7	87.4
-Fusion	69.0	88.9	88.1	85.5	86.8
-All	68.9	88.3	86.9	85.5	86.2

significant impact of removing the various components. These findings confirm the effectiveness of the proposed module. Furthermore, the results indicate that omitting either the MCFFBlock or the feature fusion strategy leads to a modest decline in system performance, suggesting that both components contribute significantly to performance improvement. On the other hand, excluding both components results in a substantial reduction in performance, highlighting their critical role in the module.

TABLE 5 Experiments using different attention mechanisms.

Component	AP50:95	AP50	Precision	Recall	F1 measure
SE-Attention	71.3	91.9	90.6	88.3	89.4
GAM-Attention	71.7	91.9	90.1	89.4	89.8
Shuffle-Attention	72.1	92.0	90.6	89.2	89.9

We also conducted ablation experiments using different attention mechanisms. The three attention mechanisms considered were SE-Attention, GAM-Attention [45], and Shuffle-Attention. The results of these experiments are presented in Table 5. The findings indicate that all three attention mechanisms contribute to the improvement of the object detection system's performance. Among them, Shuffle-Attention achieves the most significant improvements in terms of AP50:95, AP50, and F1 scores. As a result, the Shuffle-Attention mechanism is selected and utilized in this study to construct the HAM module.

4.5 | Experiments on more datasets

To assess the robustness and generalization capability of the proposed method, we conducted experiments on additional large-scale datasets containing images of varying resolutions. These datasets, such as SODA-D and VisDrone2019, encompass both large-scale images and numerous small objects. The experimental results are presented in Tables 6 and 7. The results demonstrate that our method achieves significantly higher mAP and APS metrics compared to the baseline on both datasets,

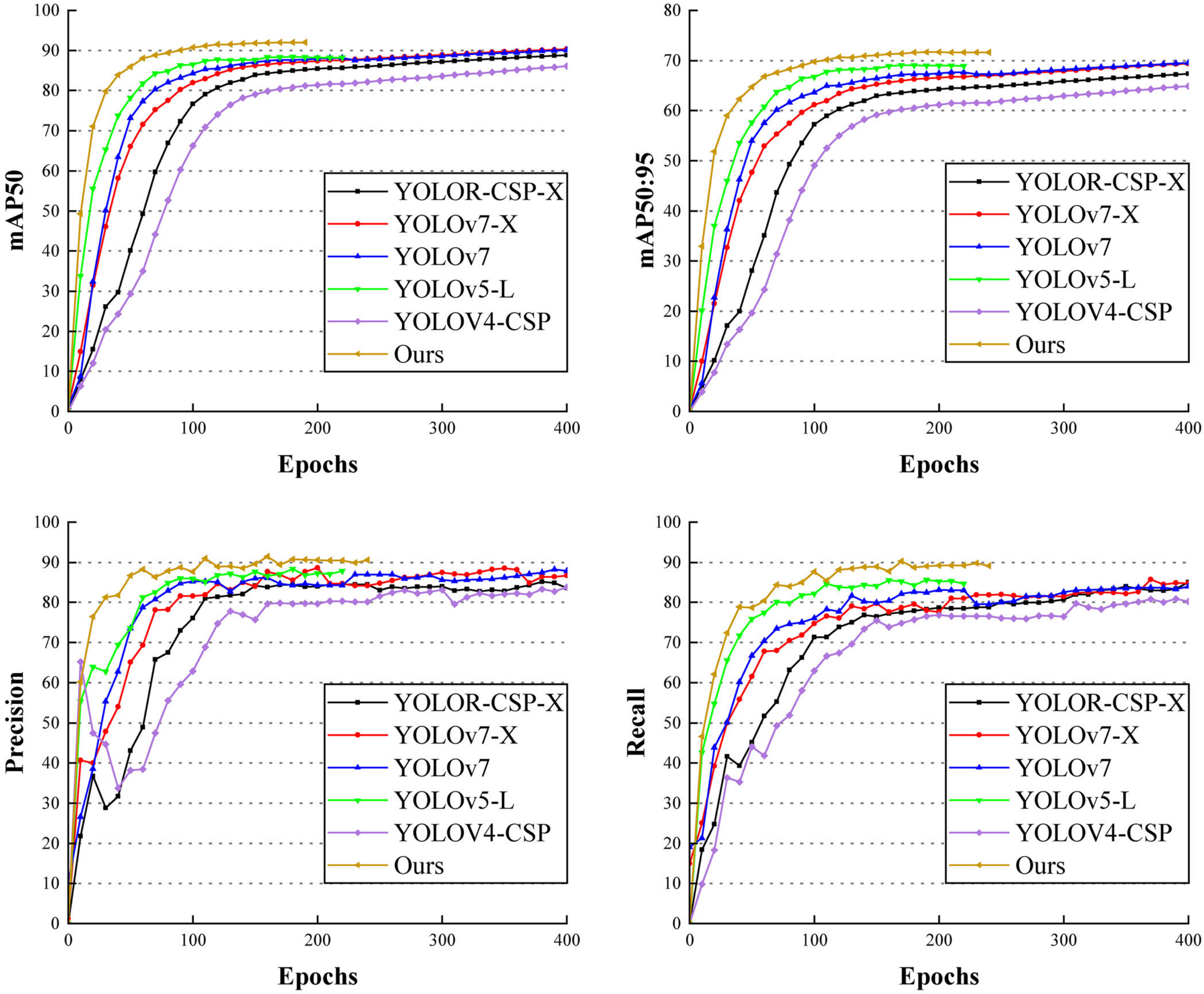

FIGURE 6 Training curve of mAP50, mAP50:95, precision, and recall metrics.

regardless of the resolution. These findings indicate that our method exhibits a relatively strong generalization ability, effectively detecting and recognizing small objects in diverse settings and varying image resolutions.

5 | CONCLUSION

The focus of this paper is to address the challenges in small object detection using c). We find that although existing methods have improved in terms of localisation and detection accuracy, they still have limitations in terms of scale handling, feature loss during transmission, object deformation problems, and insufficient consideration of tiny targets. In addition, occlusion and global feature dependency pose further challenges. To address these issues, we propose an algorithm that combines a HAM with deformable multi-scale detection and fusion. The scale problem is solved by introducing MSD and MCFF blocks, which increase the number of detection layers while preserving the high-resolution feature maps needed to detect small objects. By adding jump connections to the detection layer, the problem of feature loss during feature transmission with multiple convolutional downsampling is reduced. During training, we observed slow convergence of simultaneous classification and localisation using coupled heads, and therefore improved it by decoupling the classification and regression heads. To cope with the problem of object deformation due to filming, and the problem of insufficient effective receptive field, we use the property of deformable convolutional dynamic sampling to integrate into the feature extraction module DFEM, which effectively improves the above problems. In order to cope with the complex environments housing small objects and the occlusion issues that small objects encounter, the property of capturing global dependence using transformer is integrated

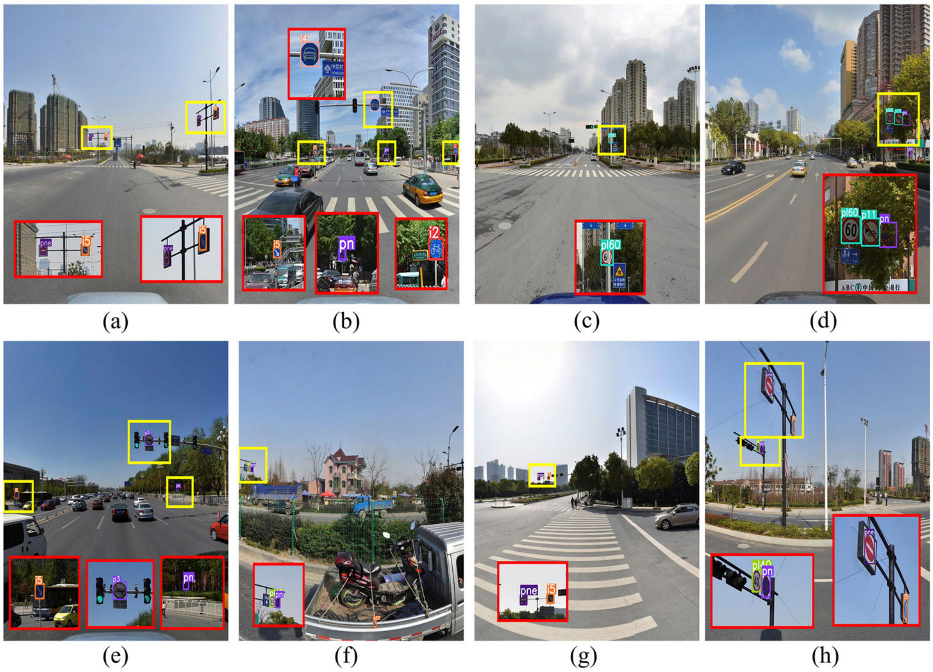

FIGURE 7 Examples of detection on the TT100K testing set. The yellow rectangular boxes in the figures contain the small targets to be identified, and the red rectangular boxes are the corresponding zoom area for easy observation. It is worth noting that the objects to be detected in all the figures occupy only a very small percentage of the overall image. In addition, figures (a), (f), and (h) with sharp target deformation, figures (b), (e), and (g) with low contrast of the target under illumination, and figures (c) and (d) with severe occlusion of the target.

TABLE 6 Experimental results on the VisDrone2019 dataset.

Resolution	Method	mAP	APS
576 × 576	YOLOv5-L	36.2	11.4
	Ours	41.4$^{+5.2}$	14.8$^{+3.4}$
640 × 640	YOLOv5-L	39.7	13.2
	Ours	44.2$^{+4.5}$	17.0$^{+3.8}$
768 × 768	YOLOv5-L	43.8	16.0
	Ours	47.7$^{+3.9}$	18.9$^{+2.9}$

TABLE 7 Experimental results on the SODA-D dataset.

Resolution	Method	mAP	APS
576 × 576	YOLOv5-L	9.5	2.2
	Ours	10.8$^{+1.3}$	2.6$^{+0.4}$
640 × 640	YOLOv5-L	11.8	2.9
	Ours	15.4$^{+3.6}$	4.0$^{+1.1}$
768 × 768	YOLOv5-L	17.6	5.2
	Ours	23.1$^{+5.5}$	7.0$^{+1.8}$

into the proposed MSAB module, which effectively solves the problems such as occlusion in detection. Experiments at different resolutions on multiple datasets demonstrate the superiority and robustness of the proposed method. At the same time, we also observe that our method has sub-optimal FPS metrics. This may be due to the increased model complexity. In addition, our understanding of how to improve the detection of large objects based on how to ensure high-precision detection of small targets is preliminary. These two issues await future work.

AUTHOR CONTRIBUTIONS

All authors contributed to the study's conception and design. Material preparation, data collection, and analysis were performed by Junyang Yu, Yuanyuan Wang, Shuang Tang, and Han Li. The first draft of the manuscript was written by Yafeng Zhang and all authors commented on previous versions of the manuscript. All authors approved the final manuscript.

ACKNOWLEDGEMENTS

This work was supported by the Major Science and Technology Project of Henan Province (Grant No. 222102210229) and the Henan Province Science and Technology Research Project (Grant No. 232102210031).

CONFLICT OF INTEREST STATEMENT

All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

DATA AVAILABILITY STATEMENT

The data used to support the findings of this study is available from the corresponding author upon reasonable request.

ORCID

Shuang Tang https://orcid.org/0009-0007-7807-5474

REFERENCES

1. Cheng, G., Yuan, X., Yao, X., Yan, K., Zeng, Q., Xie, X., Han, J.: Towards large-scale small object detection: survey and benchmarks. IEEE Trans. Pattern Anal. Mach. Intell. 1–20 (2023)
2. Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: common objects in context. In: Proceedings of the Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, Part V 13, pp. 740–755. Springer, Berlin (2014)
3. Yu, L., Zhang, H., Yu, J., Qiao, B.: Online-adaptive classification and regression network with sample-efficient meta learning for long-term tracking. Image Vision Comput. 112, 104181 (2021)
4. Yu, J., Zuo, M., Dong, L., Zhang, H., He, X.: The multi-level classification and regression network for visual tracking via residual channel attention. Digital Signal Process. 120, 103269 (2022)
5. Pan, W., Zhao, Z., Huang, W., Zhang, Z., Fu, L., Pan, Z., Yu, J., Wu, F.: Video moment retrieval with noisy labels. IEEE Trans. Neural Networks Learn. Syst. 1–13 (2022)
6. Xu, M., Yoon, S., Fuentes, A., Park, D.S.: A comprehensive survey of image augmentation techniques for deep learning. Pattern Recognit. 137, 109347 (2023)
7. Ma, L., Zheng, Y., Zhang, Z., Yao, Y., Fan, X., Ye, Q.: Motion stimulation for compositional action recognition. IEEE Trans. Circuits Syst. Video Technol. 33, 2061-2074 (2022)
8. Fu, L., Zhang, D., Ye, Q.: Recurrent thrifty attention network for remote sensing scene recognition. IEEE Trans. Geosci. Remote Sens. 59(10), 8257–8268 (2020)
9. Qi, W., Su, H.: A cybertwin based multimodal network for ECG patterns monitoring using deep learning. IEEE Trans. Ind. Inf. 18(10), 6663–6670 (2022)
10. Luo, W., Li, Y., Urtasun, R., Zemel, R.: Understanding the effective receptive field in deep convolutional neural networks. Adv. Neural Inf. Process. Syst. 29, 4905-4913 (2016)
11. Bo, W., Liu, J., Fan, X., Tjahjadi, T., Ye, Q., Fu, L.: Basnet: burned area segmentation network for real-time detection of damage maps in remote sensing images. IEEE Trans. Geosci. Remote Sens. 60, 1–13 (2022)
12. Shen, Y., Zhang, D., Song, Z., Jiang, X., Ye, Q.: Learning to reduce information bottleneck for object detection in aerial images. IEEE Geosci. Remote Sens. Lett. 20, 1-5 (2023)
13. Qiao, S., Chen, L.-C., Yuille, A.: Detectors: detecting objects with recursive feature pyramid and switchable atrous convolution. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 10213–10224 (2021)
14. Rehman, Y., Amanullah, H., Shirazi, M.A., Kim, M.Y.: Small traffic sign detection in big images: searching needle in a hay. IEEE Access 10, 18667–18680 (2022)
15. Liu, Z., Du, J., Tian, F., Wen, J.: MR-CNN: a multi-scale region-based convolutional neural network for small traffic sign recognition. IEEE Access 7, 57120–57128 (2019)
16. Chen, J., Jia, K., Chen, W., Lv, Z., Zhang, R.: A real-time and high-precision method for small traffic-signs recognition. Neural Comput. Appl. 34, 2233-2245 (2022)
17. Jocher, G., Chaurasia, A., Stoken, A., Borovec, J., NanoCode012, Kwon, Y., Xie, T., Fang, J., imyhxy, Michael, K., Lorna, V.A., Montes, D., Nadar, J., Laughing, tkianai, yxNONG, Skalski, P., Wang, Z., Hogan, A., Fati, C., Mammana, L., AlexWang1900, Patel, D., Yiwei, D., You, F., Hajek, J., Diaconu, L., Minh, M.T.: ultralytics/yolov5: v6.1 - TensorRT, TensorFlow Edge TPU and OpenVINO Export and Inference. (2022). https://doi.org/10.5281/zenodo.6222936
18. Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., Wei, Y.: Deformable convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 764–773 (2017)
19. Zhu, X., Hu, H., Lin, S., Dai, J.: Deformable convnets v2: more deformable, better results. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9308–9316 (2019)
20. Xi, X., Wang, J., Li, F., Li, D.: Irsdet: infrared small-object detection network based on sparse-skip connection and guide maps. Electronics 11(14), 2154 (2022)
21. Ji, S.-J., Ling, Q.-H., Han, F.: An improved algorithm for small object detection based on yolo v4 and multi-scale contextual information. Comput. Electr. Eng. 105, 108490 (2023)
22. Zhang, T.-Y., Li, J., Chai, J., Zhao, Z.-Q., Tian, W.-D.: Improved yolov5 network with attention and context for small object detection. In: International Conference on Intelligent Computing, pp. 341–352. Springer, Berlin (2022)
23. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: transformers for image recognition at scale. arXiv:2010.11929 (2020)
24. Zhang, X., Zhou, X., Lin, M., Sun, J.: Shufflenet: an extremely efficient convolutional neural network for mobile devices. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6848–6856 (2018)
25. Liu, Z., Li, D., Ge, S.S., Tian, F.: small traffic sign detection from large image. Appl. Intell. 50, 1–13 (2020)
26. Yuan, Y., Xiong, Z., Wang, Q.: Vssa-net: vertical spatial sequence attention network for traffic sign detection. IEEE Trans. Image Process. 28(7), 3423–3434 (2019)
27. Ren, K., Gao, Y., Wan, M., Gu, G., Chen, Q.: Infrared small target detection via region super resolution generative adversarial network. Appl. Intell. 52(10), 11725–11737 (2022)
28. Wang, X., Zhu, D., Yan, Y.: Towards efficient detection for small objects via attention-guided detection network and data augmentation. Sensors 22(19), 7663 (2022)
29. Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2117–2125 (2017)
30. Liu, S., Qi, L., Qin, H., Shi, J., Jia, J.: Path aggregation network for instance segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8759–8768 (2018)
31. Tan, M., Pang, R., Le, Q.V.: Efficientdet: scalable and efficient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Cision and Pattern Recognition, pp. 10781–10790 (2020)
32. Wang, B., Ji, R., Zhang, L., Wu, Y.: Bridging multi-scale context-aware representation for object detection. IEEE Trans. Circuits Syst. Video Technol. 33, 2317-2329 (2022)
33. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2018)
34. Guo, M.-H., Lu, C.-Z., Liu, Z.-N., Cheng, M.-M., Hu, S.-M.: Visual attention network. arXiv:2202.09741 (2022)
35. Zhang, J., Xia, K., Huang, Z., Wang, S., Akindele, R.G.: Etam: ensemble transformer with attention modules for detection of small objects. Expert Syst. Appl. 224, 119997 (2023)
36. Wang, C.-Y., Liao, H.-Y.M., Wu, Y.-H., Chen, P.-Y., Hsieh, J.-W., Yeh, I.-H.: Cspnet: a new backbone that can enhance learning capability of cnn. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 390–391 (2020)
37. Wu, Y., Chen, Y., Yuan, L., Liu, Z., Wang, L., Li, H., Fu, Y.: Rethinking classification and localization for object detection. In: Proceedings of the IEEE/CVF Conference on Computer Cision and Pattern Recognition, pp. 10186–10195 (2020)
38. Song, G., Liu, Y., Wang, X.: Revisiting the sibling head in object detector. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11563–11572 (2020)
39. Ge, Z., Liu, S., Wang, F., Li, Z., Sun, J.: Yolox: Exceeding yolo series in 2021. arXiv:2107.08430. (2021)
40. Li, P., Zhao, H., Liu, P., Cao, F.: Rtm3d: real-time monocular 3d detection from object keypoints for autonomous driving. In: Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Part III 16, pp. 644–660. Springer, Berlin (2020)
41. Du, D., Zhu, P., Wen, L., Bian, X., Lin, H., Hu, Q., Peng, T., Zheng, J., Wang, X., Zhang, Y., et al.: Visdrone-det2019: The vision meets drone object detection in image challenge results. In: Proceedings of the

IEEE/CVF International Conference on Computer Vision Workshops (2019)

42. Liu, Y., Peng, J., Xue, J.-H., Chen, Y., Fu, Z.-H.: TSingNet: scale-aware and context-rich feature learning for traffic sign detection and recognition in the wild. Neurocomputing 447, 10–22 (2021).
43. Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y.: Cutmix: regularization strategy to train strong classifiers with localizable features. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6023–6032 (2019)
44. Ghiasi, G., Cui, Y., Srinivas, A., Qian, R., Lin, T.-Y., Cubuk, E.D., Le, Q.V., Zoph, B.: Simple copy-paste is a strong data augmentation method for instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2918–2928 (2021)
45. Liu, Y., Shao, Z., Hoffmann, N.: Global attention mechanism: retain information to enhance channel-spatial interactions. arXiv:2112.05561 (2021)

How to cite this article: Zhang, Y., Yu, J., Wang, Y., Tang, S., Li, H., Xin, Z., Wang, C., Zhao, Z.: Small object detection based on hierarchical attention mechanism and multi-scale separable detection. IET Image Process. 17, 3986–3999 (2023). https://doi.org/10.1049/ipr2.12912