Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

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

Full length article

FE-YOLOv5: Feature enhancement network based on YOLOv5 for small object detection[☆]

Min Wang[a,b], Wenzhong Yang[a,b,*], Liejun Wang[a], Danny Chen[a,b], Fuyuan Wei[a,b], HaiLaTi KeZiErBieKe[a], Yuanyuan Liao[a]

[a] *School of Information Science and Engineering, Xinjiang University, Urumqi, Xinjiang, China*
[b] *Xinjiang Key Laboratory of Multilingual Information Technology, Xinjiang University, Urumqi, Xinjiang, China*

ARTICLE INFO

Keywords:
Small object detection
Feature enhancement
Spatial-aware

ABSTRACT

Due to their inherent characteristics, small objects have weaker feature representation after multiple down-sampling and are even annihilated in the background. FPN's simple feature concatenation does not fully utilize multi-scale information and introduces irrelevant context into the information transfer, further reducing the detection performance of the small object. To address the above issues, we propose the simple but effective FE-YOLOv5. (1) We designed the feature enhancement module (FEM) to capture more discriminative features of the small object. Global attention and high-level global contextual information are used to guide shallow, high-resolution features. Global attention interacts with cross-dimensional feature interaction and reduces information loss. High-level context complements more detailed semantic information by modeling global relationships through non-local networks. (2) We design the spatially aware module (SAM) to filter spatial information and enhance the robustness of features. Deformable convolution performs sparse sampling and adaptive spatial learning to better focus on foreground objects. According to the experimental results, our proposed FE-YOLOv5 outperforms the other architectures in the VisDrone2019 dataset and Tsinghua-Tencent100K dataset. Compared to YOLOv5, the AP_S was improved by 2.8% and 2.9%, respectively.

1. Introduction

CNN can learn the richer appearance and spatial representation from large-scale image datasets, significantly improving detection performance, opening up a new path for object detection research, and promoting significant progress. Object detection algorithms fall into two main categories. The two-stage network represented by RCNN [1–3] first generates candidate regions and then classifies and regresses them. The one-stage networks represented by YOLO [4–6] and SSD [7] directly classify and regress objects. In contrast, the former has higher accuracy and the latter has faster speed.Large-scale objects are generally easier to detect due to their large area and rich features, while small objects fail to be detected satisfactorily due to their small area and susceptibility to noise. In recent years, small object detection has played a very important role in practical scenarios such as autonomous driving and UAV aerial photography analysis. It is a hot topic in the field of object detection research.

The main reasons for the poor performance of small object detection are as follows. (1) Small objects have limited expression of the features extracted by the network due to their characteristics. After multiple down-sampling, the features are even weaker and are even lost in the background. (2) Small objects lack common training datasets, and the data in the public datasets are unevenly distributed so that the network is more inclined to label a larger number of large objects during the training process. (3) Localization is difficult because small objects are more sensitive than large objects. The existing small object detection methods are all improvements on the general deep learning detection methods to improve the detection effect. FPN [8] fuses shallow detail information and deep high-level semantic information to get more discriminative features, which greatly improves the expression of small objects.

FPN is highly structured. Objects of different scales are detected on different feature layers. Small objects are generally detected on the feature map P_3 with 8x downsampling. P_2 (4x downsampled feature map) near the original image contains more detailed information about the small object, but does not participate in the transmission of pyramid information. To obtain more discriminative features, we propose

[☆] This paper has been recommended for acceptance by Zicheng Liu.

[*] Corresponding author at: School of Information Science and Engineering, Xinjiang University, Urumqi, Xinjiang, China.

E-mail addresses: wangmin@stu.xju.edu.cn (M. Wang), yangwenzhong@xju.edu.cn (W. Yang), wljxju@xju.edu.cn (L. Wang), kabuoxygen@163.com (D. Chen), 641868693@qq.com (F. Wei), hayrat805@163.com (H. KeZiErBieKe), liaoyuan@xju.edu.cn (Y. Liao).

https://doi.org/10.1016/j.jvcir.2023.103752
Received 22 July 2022; Received in revised form 15 November 2022; Accepted 2 January 2023
Available online 5 January 2023

the feature enhancement module (FEM). The shallow high-resolution feature P_2 is guided by global attention and high-level global contextual information. Global attention carries out cross-dimensional feature interaction to reduce the loss of information; the latter complements more detailed semantic information by modeling global relationships through non-local networks. RSOD [9] utilizes more fine-grained P_2 and then adaptively performs feature fusion by assigning weights to the importance of each layer calculated by ASFF. QueryDet [10] also incorporates P_2 into the pyramid, and to make the model fully learn, the authors include balance coefficients in the loss function of each layer. Unlike this, we do not weigh the features of each layer. The focus is on enhancing the feature of P_2, reducing background interference and information loss, and making the network more sensitive to small objects.

The simple feature concatenation of FPN cannot fully use multi-scale information. To enhance the robustness of features, many FPN deformations have been proposed. For example, adding bottom-up paths, adding weights, etc. [11] proposed a fusion factor to control the information transfer from deep to shallow layers of FPNs. It aims to eliminate the negative effects of top-down connections as small object bands. We found that adding deformable convolution [12] for sparse learning before deep and shallow feature fusion can improve the performance of small object detection, so we propose the spatial perception module (SAM). Deformable convolution is invariant in channel dimension and adaptively learns according to the shape of the object in 2D space.

Based on the above, this paper proposes FE-YOLOv5 to improve the detection performance of the small object. The model adds FEM and SAM to YOLOv5 and changes the original detection layers. The VisDrone2019 dataset [13] and the Tsinghua-Tencent100K dataset [14] contain a larger number of small objects. Since the Tsinghua-Tencent-100K dataset has a serious category imbalance problem, we filtered and reclassified this dataset. The experimental results show that our proposed method has better detection results.

To sum up, the main contributions of this paper are summarized as follows:

(1) Uses global attention and high-level contextual information of global relationship modeling to guide the fine-grained shallow high-resolution feature, and enhances the feature representation of small objects.

(2) Applies deformable convolution in the neck part of YOLOv5 to enhance the robustness of FPN features, filter the spatial information, and focus on the foreground objects.

(3) Effective experiments on two small object detection benchmark datasets: the VisDrone2019 and Tsinghua-Tencent100K to evaluate the performance of the model.

2. Related work

Traditional object detection algorithms rely on manually designed features that have poor generalization and are not robust to changes in diversity. Meanwhile, the region selection strategy based on sliding windows is time-consuming and unreliable in real time. With the development of deep learning and computing resources, research scholars have found that better features can be extracted to accurately classify and localize objects using convolution neural networks, which has greatly advanced the development of object detection. However, studies have shown that the detection accuracy of small objects differs significantly from that of regular-sized objects, which is only half that of large objects, indicating that there is still much room for improvement. The characteristics and data distribution of small objects have made detection difficult. Data augmentation, new training methods, and multi-scale feature fusion detection have emerged as the most important means of improving small object detection performance.

2.1. Data augmentation

To address the problems of insufficient small object detection dataset and unbalanced samples, kisantal et al. [15] increased the diversity of the number and location of small objects by oversampling and copy-pasting, but it is necessary to ensure that these objects appear in the correct context. Yolov4 [16] employs the Mosaic data augmentation method, in which four randomly selected images were scaled and stitched and then fed into the neural network for training.

2.2. New training methods

Singh and Davis proposed a new training strategy SNIP [17] to address the domain shift during pre-trained model migration by selectively back-propagating the gradients of object instances of different sizes according to the variation of image scales. The core of deformable DETR [18] is the multi-scale variability attention module, which can naturally be extended to aggregate multi-scale features by focusing on only a small group of key sampling points. The algorithm achieves better performance than DERT [19], especially on small objects. QueryDet accelerates small object detection through cascading sparse queries and uses the rough positions of small objects predicted on the low-resolution feature map to guide the calculation of high-resolution features.

2.3. Multi-scale feature fusion detection

Lin et al. proposed the FPN to reduce the information loss of small objects during pooling by combining low-resolution feature maps and high-resolution feature maps so that each level of the feature pyramid has rich semantic information and detailed information, which is currently the most representative method for multi-scale fusion. AugFPN [20] employs the consistency supervision approach to ensure the laterally connected feature maps contain similar semantic information to reduce information loss in FPN's highest-level features. To improve the overall feature and shorten the information path between the bottom and top features, PANet [21] adds a bottom-up path to transmit positioning information based on FPN. Bi-FPN [22] introduces weights based on PANet to better balance the information of different scales. NAS-FPN [23] provides the optimal search method to combine and update the extracted feature maps to obtain better detection results. Trident [24] constructs a parallel multi-branch structure with different receptive fields by using dilated convolution to deal with the scale change problem of object detection by scale-aware training.

To some extent, multi-scale feature fusion methods improved the detection performance of small objects, but some methods ignore the semantic spacing and noise between different scales of information, and others introduce significant overheads. We discovered that using deformable convolution in the feature fusion process can effectively solve these problems. Simultaneously, YOLOv5 employs Mosaic data augmentation to improve the network's sensitivity to small objects. As a result, our proposed model has a higher detection performance.

3. Method

YOLOv5 uses CSPNet as the backbone to extract features, achieving a better balance of inference speed and accuracy. YOLOv5 can dynamically adjust the depth and width of the network to meet the needs of different scenes. We chose the smallest scale YOLOv5-S (the network width parameter is 0.50 and the depth parameter is 0.33). Based on this, we made improvements and proposed the FE-YOLOv5 model (Fig. 1). We added the FEM and the SAM to YOLOv5-S and changed the previous detection layers to improve small object detection while maintaining other object detection performance. $\{C_2, C_3, C_4, C_5\}$ is the 2^i downsampled feature maps, where $i = (2, 3, 4, 5)$. C_2 contains more detailed information and P_3 learns richer semantic information

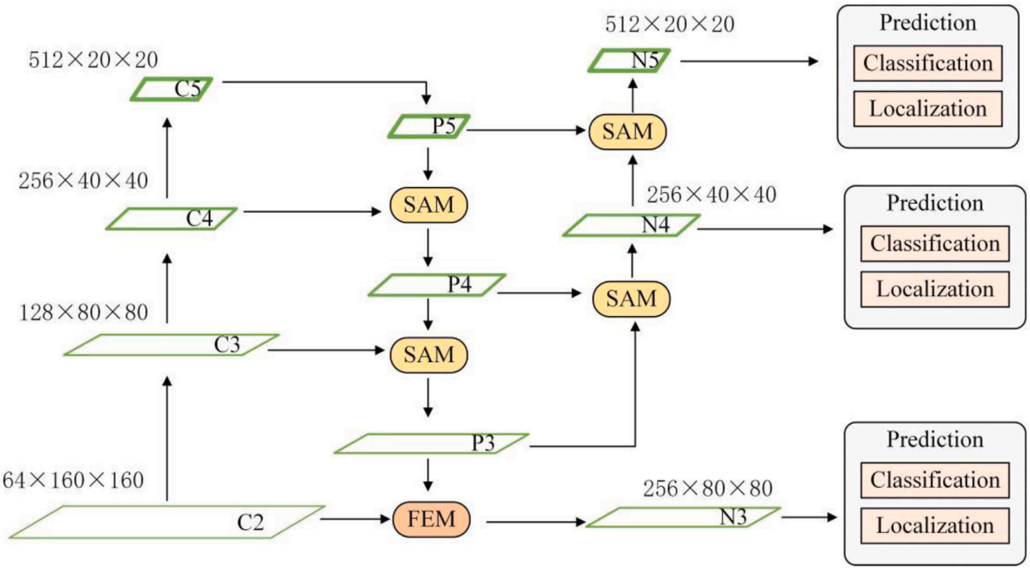


Fig. 1. The framework of the FE-YOLOv5 model. The size of the quadrilateral represents the size of the feature maps, and the thickness of the lines represents the number of channels.

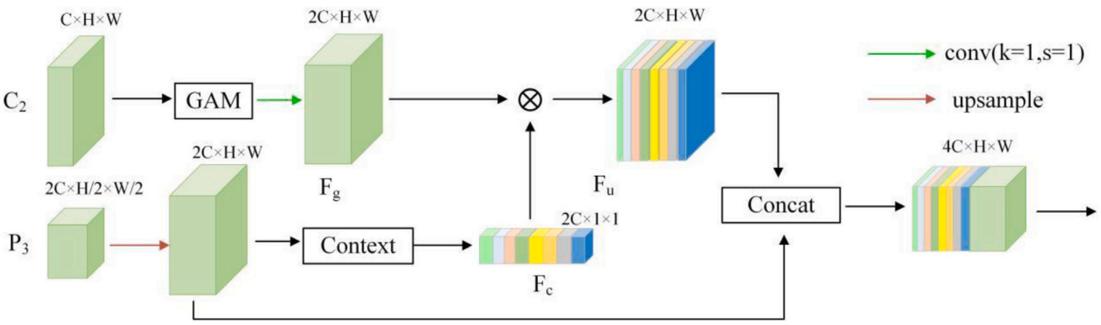


Fig. 2. The structure of FEM. ⊗ represents elementwise product. The green line represents 1 × 1 convolution, and the brown line represents an upsampling operation.

via top-down paths and lateral connections. FEM combines the advantages of these two layers to obtain more discriminative features. Small objects, due to their characteristics, become smaller or even lose all information in the deep network, while YOLOv5 discards the shallower layer to detect small objects on P_3. The recall has significantly improved since the addition of FEM. SAM mainly performs sparse sampling by deformable convolution and spatially filters conflicting information to improve feature quality, resulting in improved detection of deformed and occluded objects. $\{P_3, P_4, P_5\}$ is the intermediate feature generated by FPN, and FE-YOLOv5 modifies the original detection layers, detecting objects at different scales on $\{N_2, N_3, N_4, N_5\}$.

3.1. Feature enhancement module

As shown in Fig. 2, C_2 and P_3 as the input to the FEM. The shallow layer C_2 is closer to the original image and contains some contour, edge, and location information. P_3 has a larger receptive field and contains rich semantic information about the object. The FEM draws on the experience of GAU [25] (Global Attention Upsample) using high-level contextual information to guide the shallow information, as shown in Fig. 3(a), and we improve it.

C_2 through the global attention mechanism (GAM) [26], followed by 1 × 1 convolution to increase the number of channels, where GAM includes both channel attention and spatial attention. SENet [27] modeled the interdependence between channels. CBAM [28] used a sequence of focusing on important features from channel to space. Self-attention mechanism effectively captured full context information. These attention mechanisms ignored the interaction between channels and spatial dimensions. GAM has improved CBAM and performs cross-dimensional feature interaction to retain more information and capture more important features. Assuming that input feature mapping $F_1 \in R^{C\times H\times W}$, F_2 represents intermediate features and F_3 is output feature, GAM can be expressed as:

$$F_2 = M_C(F_1) \otimes F_1 \tag{1}$$

$$F_3 = M_S(F_2) \otimes F_2 \tag{2}$$

where ⊗ denotes the multiplication operation by the element and the descriptions of channel attention M_C and spatial attention M_S are shown below:

$$M_C(F_1) = \sigma(f(MLP(permutation(F_1)))) \tag{3}$$

$$M_S(F_2) = \sigma(Conv_{7\times 7}(Conv_{7\times 7}(F_2))) \tag{4}$$

MLP is a two-layer encoder–decoder perceptron. Permutation operation will convert $C \times W \times H$ to $W \times H \times C$, retaining the information across dimensions, the f operation reduces the dimension of the permutation, σ denoting the sigmoid function.

For small objects, the similarity with the background makes their classification difficult. GAM can effectively enhance useful features while suppressing useless ones, being more sensitive to small objects.

P_3 obtains attention weights by upsampling and then using 1 × 1 convolution and softmax function to perform attention mapping to

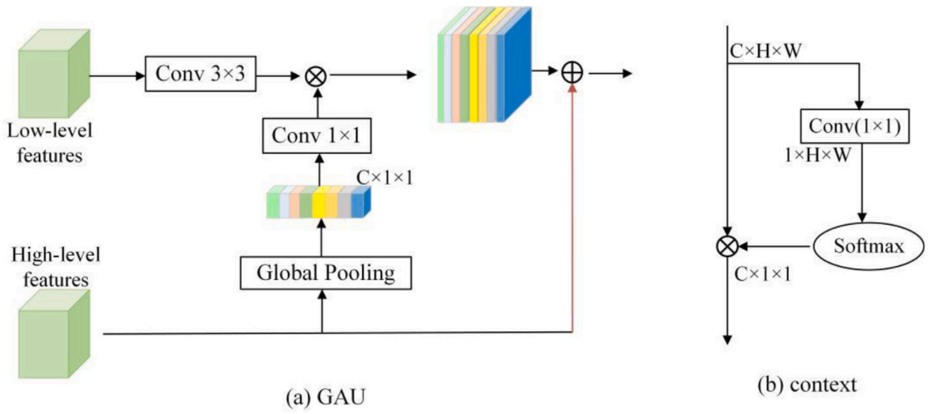


Fig. 3. The structure of GAU and context module. ⊗ represents the elementwise product, ⊕ represents elementwise add, and the brown line represents upsampling. (b) modeling global relationships through non-local networks.

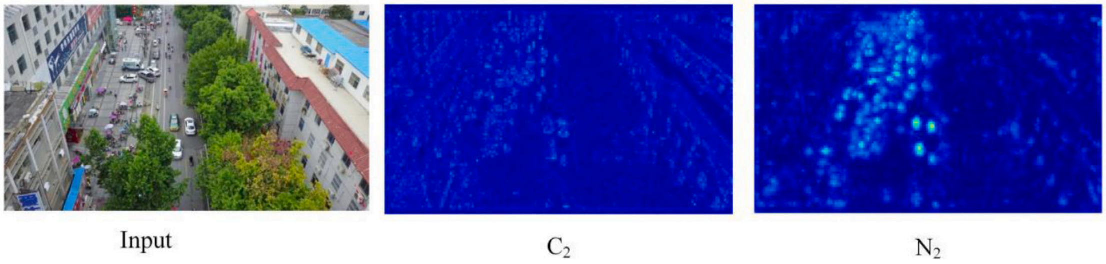


Fig. 4. Results of feature map visualization.

get the global contextual feature F_3 (as shown in Fig. 3(b)), which guides for C_2. In contrast, GAU uses global pooling to compress the information of the spatial dimension on the channel dimension, which leads to information loss and cannot model the global relationships, so we use a simplified version of the Non-local network instead of global pooling. Each pixel is processed in an adjacent region because the CNN convolution kernel is a fixed square region. CNN expands the receptive field by stacking convolution blocks to establish long-range dependencies, the context module [29] combines features at all positions by weighted averaging, reducing the number of parametric calculations and conveying effective long-range dependency information.

Finally, the feature F_u multiplied by F_c and F_g is concatenated with the feature sampled on P_3. Then the feature is convoluted to obtain N_2 (256 × 160×160) and detected. Fig. 4 shows some visualization results. It can be seen that the features obtained after the FEM of C_2 and P_3 are more obvious and beneficial for object detection.

3.2. Spatial-aware module

Different spatial positions correspond to different scales of objects, and the importance of pixels in different positions for detecting features is also different. In the FPN-based structure, the deep features are simply concatenated with the shallow features after upsampling, which cannot fully express the multi-scale features and introduce a lot of irrelevant context information. CNN uses a square convolutional kernel structure for sampling, resulting in the same size of receptive fields of the same layer of activation units. For small objects, this receptive field may contain many regions unrelated to the object, causing background interference, thus reducing the quality of features. We propose the spatial-aware module (Fig. 5) that employs deformable convolution to adaptively learn the optimal convolution kernel structure and sampling points. So that the effective receptive field can better match the shape of the object. Therefore, SAM can better focus on the foreground object according to the discrimination ability of various spatial locations.

The deep feature X_2 after upsampling and the shallow feature X_1 have the same dimension. They perform a 3 × 3 convolution to learn the bias offset (offset on x, offset on y) and the weight mask of each sampled point. Then feed these two tensors together with the original feature map into the deformable convolution to capture the features that are more suitable for the object. The output of this step is:

$$y(p) = \sum_{k=0}^{k} w_k \cdot x(p + p_k + \triangle p_k) \cdot \triangle m_k \tag{5}$$

where k is the size of the convolution kernel, w_k and p_k denote the weights and predefined offsets at the kth convolution kernel position, $y(p)$ denotes the feature at position p in the output feature map y. $\triangle p_k$ and $\triangle m_k$ denote the offset and weight coefficients at the kth position, $\triangle m_k \in [0,1]$.

4. Experiments

4.1. Datasets and evaluation metrics

We experimented on the VisDrone2019 dataset and the Tsinghua-Tencent100K dataset. Fig. 6 describes the object instance distribution of the two datasets. The horizontal axis indicates the percentage of objects of different scales in different sets. Small objects have an area of fewer than 32 × 32 pixels, medium-sized objects have an area between 32 × 32 pixels and 96 × 96 pixels, and large objects have an area greater than 96 × 96 pixels.

The VisDrone2019 dataset using UAVs captures various scenes from our daily lives, with a total of 10 classes. It covers different altitudes, weather, and lighting conditions, containing many small objects with varying degrees of dense object deformation and occlusion. The dataset contains 6471 images for training, 548 for validation, and 3190 for testing, with 1580 challenging images in the test set.

The Tsinghua-Tencent100K dataset is the largest public dataset of traffic signs in China, covering a large area, including warning

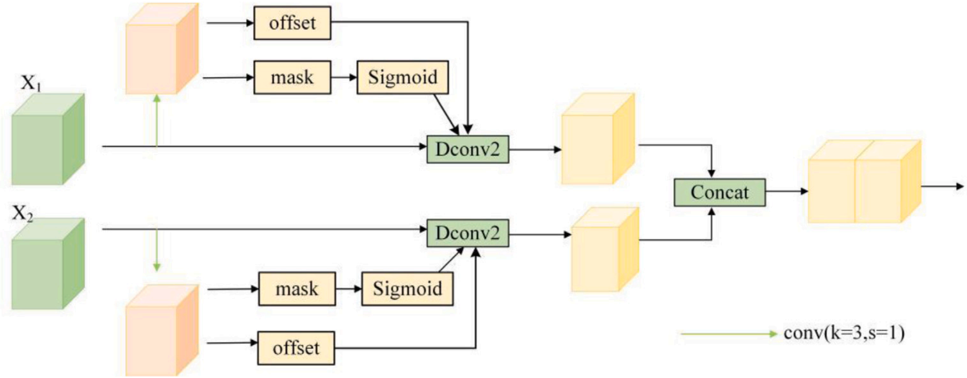


Fig. 5. The structure of SAM. The green line represents 3 × 3 convolution. X_1 and X_2 are the feature layers for feature fusion in the neck part.

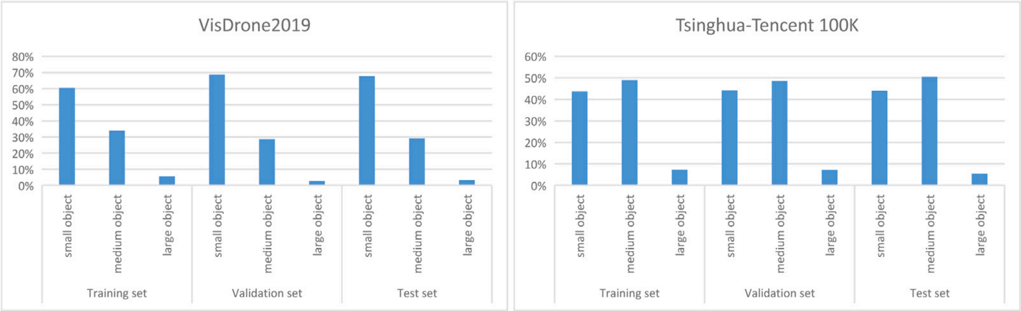


Fig. 6. (a) shows the data distribution on the VisDrone2019 dataset, and (b) shows the data distribution on the Tsinghua-Tencent100K dataset. It can be seen from the figure that the proportion of small objects is relatively high, which can be used as the benchmark dataset for studying small object detection.

signs, prohibition signs, and indication signs installed on urban centers and suburban roads. Only 10,000 images contain traffic sign symbols after removing the background images from 100,000 Tencent street scenes. A total of 30,000 traffic sign symbols and 128 categories were extracted, and each traffic symbol was manually labeled in the shape of a polygon or ellipse. Compared with the previous traffic sign detection dataset, this dataset contains more images and has a higher resolution (2048 × 2048). The majority of the traffic signs in the images are very small and have significant variations in illumination and weather. There is a serious category imbalance in this dataset, and some traffic signs, such as beware of falling rocks, under construction, etc., rarely appear in urban scenes. Statistically, some categories do not even appear in the training set, so we reserved 45 categories with more than 100 traffic sign instances. The image categories are the most common traffic symbol categories in this figure. We counted these categories and re-divided the dataset according to the ratio of 7:2:1, with 6657 images in the training set, 1909 images in the validation set, and 978 images in the test set.

We used the evaluation criteria of COCO to evaluate the effectiveness of the model. Average Precision(AP) is divided into 10 different intervals according to the threshold of IoU, and AP is calculated every 0.05 in [0.5,0.95], and then go to calculate the average of these AP as the final AP, generally, AP, AP_{50}, and AP_{75} are used. AP Across Scales is the average accuracy of objects at various scales, primarily AP_S, AP_M, and AP_L.

4.2. Implementation details

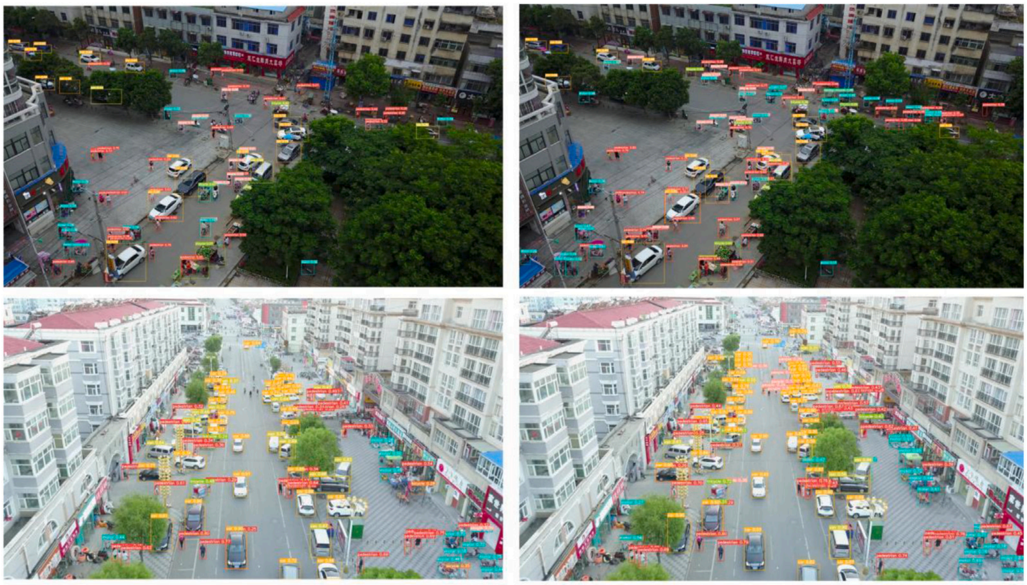
We used the weights pre-trained by YOLOv5-S on the COCO dataset to initialize the network, which facilitates faster convergence of the model, and the hyperparameters were set using the YOLOv5-S configuration. We use an NVIDIA GeForce RTX 2080 Ti graphics card for model training and testing, and the input images are uniformly resized to 640 × 640 and optimized using the SGD optimizer. The initial learning rate was 0.01, the weight decay coefficient was 0.0005, and a total of 300 epochs were iterated.

Table 1
Experimental results on VisDrone2019 dataset.

Model	AP	AP_{50}	AP_{70}	AP_S	AP_M	AP_L
RetinaNet [30]	15.0	26.4	15.3	6.3	25.6	34.4
Libra R-CNN [31]	14.9	25.2	15.2	5.9	25.6	31.4
Fcos [32]	17.9	30.4	18.3	9.2	27.6	35.4
ATSS [33]	20.4	33.8	**20.9**	11.6	**31.7**	36.7
Tridentnet	19.8	35.0	19.5	11.4	29.6	36.6
YOLOv5	18.2	32.9	17.4	10.4	27.0	35.3
FE-YOLOv5(ours)	**21.0**	**37.0**	20.7	**13.2**	29.5	**39.1**

4.3. Experiments on VisDrone2019

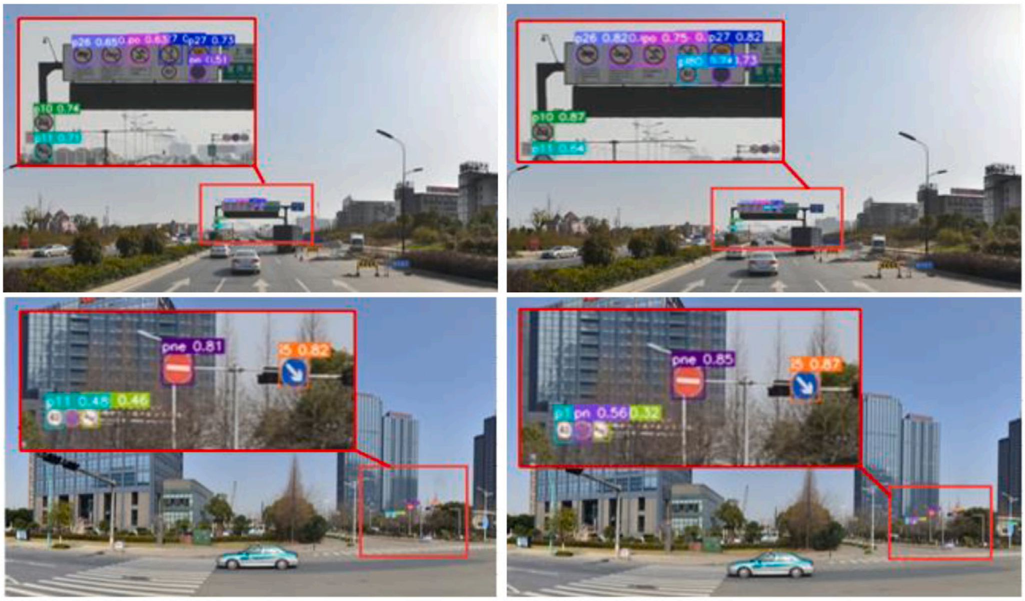
To evaluate the effectiveness of our proposed method, we compared it with some other models on the VisDrone2019 dataset, and the experimental results are shown in Table 1. These methods are representative models based on two-stage anchor_based, one-stage anchor_based, and one-stage anchor free, respectively. As can be seen in Table 1, our proposed model improves on all evaluation metrics when compared to YOLOv5. There is an increase of 4.1% on AP50, 2.8%, 2.5%, and 3.8% on small, medium, and large objects, respectively. FE-YOLOv5 performed better than the other models in the table on AP, AP_{50}, AP_S, and AP_L, but slightly lower than ATSS on AP_{75}, and behind ATSS and Tridentnet on AP_M. ATSS has a more flexible strategy for assigning positive and negative samples using statistical methods, so AP_{75} performs better. While Tridentnet utilizes parallel multiple branches with different dilated convolutions to detect, the receptive field is better matched with medium objects. Fig. 7 shows some detection results of FE-YOLOv5, and it can be seen that FE-YOLOv5 detects more small objects.



(a)

(b)

Fig. 7. Visualization of the detection results on the VisDrone2019 dataset. (a) indicates the YOLOv5 detection results and (b) is our proposed FE-YOLOv5 detection results.



(a)

(b)

Fig. 8. Visualization of detection results on Tsinghua-Tencent100K dataset. (a) indicates the YOLOv5 detection results, and (b) is the detection result of our proposed FE-YOLOv5.

4.4. *Experiments on Tsinghua-Tencent100K*

We also conducted experiments on the Tsinghua-Tencent100K dataset, and the results are shown in Table 2. From Table 2, we can see that our proposed model outperforms the YOLOv5 model before improvement by 2.2%, 2.9%, 0.7%, and 3.4% in AP_{50}, AP_S, AP_M, and AP_L, respectively. It is slightly lower than the Libra R-CNN model by 0.2% on AP_M. Furthermore, we can find that our proposed model has the best performance on AP_S in Table 2.

Fig. 8 shows some detection results on this dataset. Because the image resolution of this dataset is large and the proportion of objects in the original image is very small, we have enlarged the region to show it. It can be seen from the figure that FE-YOLOv5 has higher detection accuracy and also detects the objects missed by YOLOv5.

4.5. *Ablation studies*

To demonstrate the effectiveness of FEM and SAM, we performed ablation experiments on the VisDrone2019 dataset, and the results are shown in Table 3.

FEM enhances the shallow features using the global attention mechanism and then supplements the shallow information using the higher-level contextual information to make the model more focused on small objects and increase the detection performance of small objects. As can be seen from the data distribution in Fig. 5, the proportion of large objects is relatively small, so AP_L has decreased. Rows 3 and 4 of Table 3 are the ablation experiments of FEM. The experimental results in row 3 illustrate that P_3 is necessary for C_2 guidance, while the experimental results in row 4 illustrate that GAM is effective for the

Table 2
Experimental results on the Tsinghua-Tencent100K dataset.

Model	AP	AP_{50}	AP_{70}	AP_S	AP_M	AP_L
RetinaNet	49.1	63.0	58.2	33.4	58.7	68.2
Libra R-CNN	60.4	83.5	75.8	44.4	**70.2**	71.7
Fcos	54.9	69.9	65.7	36.8	62.8	72.4
ATSS	55.4	69.9	66.1	37.2	65.4	72.7
Tridentnet	58.2	81.9	70.5	37.7	68.8	74.0
YOLOv5	62.2	80.8	74.0	49.9	69.3	74.4
FE-YOLOv5(ours)	**63.6**	**82.6**	**76.9**	**52.8**	70.0	**77.7**

Table 3
Ablation experiment of FE-YOLOv5 and FEM. ✓-GAM means to remove GAM from FEM;✓-content means to remove content from FEM, and is directly spliced with after upsampling by GAM and 1× 1 convolution operation.

YOLOv5	FEM	SAM	AP	AP_{50}	AP_{70}	AP_S	AP_M	AP_L
✓			18.2	32.9	17.4	10.4	27.0	35.3
✓	✓		19.5	34.5	19.2	12.2	28.0	33.3
✓	✓-GAM		19.0	33.9	18.5	11.5	27.5	33.9
✓	✓-content		19.5	34.8	19.0	12.4	27.9	33.7
✓	✓-content	✓	20.1	35.3	19.5	12.6	28.4	36.3
✓		✓	19.7	34.9	18.9	12.0	28.0	35.4
✓	✓	✓	**21.0**	**37.0**	**20.7**	**13.2**	**29.5**	**39.1**

Table 4
Other comparative experiments. We compared FE-YOLOv5, YOLO-S, and YOLO-M models. At the same time, the effectiveness of FEM and GAU is verified.

Model	AP	AP_{50}	AP_{70}	AP_S	AP_M	AP_L
YOLOv5-S	18.2	32.9	17.4	10.4	27.0	35.3
YOLOv5-S+GAU+SAM	19.6	34.7	19.1	12.5	27.7	37.5
YOLOv5-S+FEM+SAM	21.0	37.0	20.7	13.2	29.5	39.1
YOLOv5-M	21.0	36.6	20.6	12.6	30.9	39.4

Table 5
Comparison of parameter amount and floating point operations under the same input size of YOLO-S, YOLO-M, and FE-YOLOv5.

Model	Input size (pixels)	params (M)	FLOPs (G)
YOLOv5-S	640× 640	7.03	15.9
FE-YOLOv5	640× 640	9.14	31.0
YOLOv5-M	640× 640	20.88	48.1

detection of small objects, some of the evaluation metrics are higher than FEM. Therefore, we guess whether the performance of the whole model will be improved if the content module is removed for FEM. The corresponding experiments are in row 5 of Table 3. The results show a decrease in performance, so we still used the FEM as shown in Fig. 2. SAM improved in all metrics, proving the effectiveness of the module.

Table 4 shows the comparison results of FEM and GAU (mentioned in 3.1), FE-YOLOv5, and YOLOv5-M. In contrast, FE-YOLOv5(YOLOv5-S+FEM+SAM) outperforms the model composed of YOLOv5-S+GAU+SAM in all indicators. As can be seen from the last two rows of the table, our proposed model is only slightly lower than YOLOv5-M in AP_M and AP_L, has better detection results for all other metrics, and has fewer model parameters and computational effort, as shown in Table 5.

5. Conclusion

In this paper, we propose the FE-YOLOv5 model with structural modifications to the YOLOv5 neck part to improve the detection performance of small objects. FEM incorporates shallow high-resolution features into pyramid information transmission and uses GAM and high-level features to guide shallow features to obtain features based on discrimination. From the visual feature map, this module can reduce the interference of background information and make the effective features more obvious. Secondly, we add SAM to the neck part and use the deformable convolution adaptive learning optimal convolution kernel structure to increase the effective receptive field and filter the interference of irrelevant information. FE-YOLOv5 has shown good detection results on the VisDrone2019 dataset and the Tsinghua-Tenent100K dataset, superior to the original YOLOv5-S model, and even beyond YOLOv5-M. FE-YOLOv5 applies to natural scenes or scenes taken by UAVs, providing a new scheme for small target detection. In the future, we will continue to improve and explore its application and effectiveness in actual industrial scenarios.

CRediT authorship contribution statement

Min Wang: Software, Writing – original draft, Writing – review & editing, Visualization. **Wenzhong Yang:** Conceptualization, Funding acquisition, Supervision. **Liejun Wang:** Resources. **Danny Chen:** Formal analysis, Data curation. **Fuyuan Wei:** Formal analysis, Data curation. **HaiLaTi KeZiErBieKe:** Investigation. **Yuanyuan Liao:** Project administration.

Declaration of competing interest

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

Data availability

Data will be made available on request.

Acknowledgments

This research was funded by [the National Natural Science Foundation of China] grant number [No. U1603115], [Science and Technology Project of Autonomous Region] grant number [No. 2020A02001-1], [Research on short-term and impending precipitation prediction model and accuracy evaluation in Northern Xinjiang Based on deep learning] grant number [2021D01C080] and [Jiangxi Natural Science Foundation Project] grant number [No. 20202BAB202023].

References

[1] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 580–587.

[2] R. Girshick, Fast r-cnn, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1440–1448.

[3] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, Adv. Neural Inf. Process. Syst. 28 (2015).

[4] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 779–788.

[5] J. Redmon, A. Farhadi, YOLO9000: better, faster, stronger, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 7263–7271.

[6] J. Redmon, A. Farhadi, Yolov3: An incremental improvement, 2018, arXiv preprint arXiv:1804.02767.

[7] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A.C. Berg, Ssd: Single shot multibox detector, in: European Conference on Computer Vision, 2016, pp. 21–37.

[8] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2117–2125.

[9] W. Sun, L. Dai, X. Zhang, P. Chang, X. He, RSOD: Real-time small object detection algorithm in UAV-based traffic monitoring, Appl. Intell. (2021) 1–16.

[10] C. Yang, Z. Huang, N. Wang, QueryDet: Cascaded sparse query for accelerating high-resolution small object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 13668–13677.

[11] Y. Gong, X. Yu, Y. Ding, X. Peng, J. Zhao, Z. Han, Effective fusion factor in FPN for tiny object detection, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, pp. 1160–1168.

[12] X. Zhu, H. Hu, S. Lin, J. Dai, Deformable convnets v2: More deformable, better results, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 9308–9316.

[13] P. Zhu, L. Wen, D. Du, X. Bian, H. Fan, Q. Hu, H. Ling, Detection and tracking meet drones challenge, IEEE Trans. Pattern Anal. Mach. Intell. 44 (2021) 7380–7399.

[14] Z. Zhu, D. Liang, S. Zhang, X. Huang, B. Li, S. Hu, Traffic-sign detection and classification in the wild, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2110–2118.

[15] M. Kisantal, Z. Wojna, J. Murawski, J. Naruniec, K. Cho, Augmentation for small object detection, 2019, arXiv preprint arXiv:1902.07296.

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

[17] B. Singh, L.S. Davis, An analysis of scale invariance in object detection snip, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3578–3587.

[18] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, J. Dai, Deformable detr: Deformable transformers for end-to-end object detection, 2020, arXiv preprint arXiv:2010.04159.

[19] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, End-to-end object detection with transformers, in: European Conference on Computer Vision, 2020, pp. 213–229.

[20] C. Guo, B. Fan, Q. Zhang, S. Xiang, C. Pan, Augfpn: Improving multi-scale feature learning for object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 12595–12604.

[21] S. Liu, L. Qi, H. Qin, J. Shi, J. Jia, Path aggregation network for instance segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8759–8768.

[22] M. Tan, R. Pang, Q.V. Le, Efficientdet: Scalable and efficient object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 10781–10790.

[23] G. Ghiasi, T.-Y. Lin, Q.V. Le, Nas-fpn: Learning scalable feature pyramid architecture for object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 7036–7045.

[24] Y. Li, Y. Chen, N. Wang, Z. Zhang, Scale-aware trident networks for object detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 6054–6063.

[25] H. Li, P. Xiong, J. An, L. Wang, Pyramid attention network for semantic segmentation, 2018, arXiv preprint arXiv:1805.10180.

[26] Y. Liu, Z. Shao, N. Hoffmann, Global attention mechanism: Retain information to enhance channel-spatial interactions, 2021, arXiv preprint arXiv:2112.05561.

[27] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141.

[28] S. Woo, J. Park, J.-Y. Lee, I.S. Kweon, Cbam: Convolutional block attention module, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 3–19.

[29] Y. Cao, J. Xu, S. Lin, F. Wei, H. Hu, Gcnet: Non-local networks meet squeeze-excitation networks and beyond, in: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, 2019.

[30] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2980–2988.

[31] J. Pang, K. Chen, J. Shi, H. Feng, W. Ouyang, D. Lin, Libra r-cnn: Towards balanced learning for object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 821–830.

[32] Z. Tian, C. Shen, H. Chen, T. He, Fcos: Fully convolutional one-stage object detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 9627–9636.

[33] S. Zhang, C. Chi, Y. Yao, Z. Lei, 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, 2020, pp. 9759–9768.