scientific reports

Check for updates

OPEN

R-AFPN: a residual asymptotic feature pyramid network for UAV aerial photography of small targets

Zuowen Chen, Yahong Ma✉, Zi'an Gong, Minghao Cao, Yuyao Yang, Zhiyuan Wang, Tengjie Wang, Jing Li & Yuxi Liu

This study proposes an improved Residual Asymptotic Feature Pyramid Network (R-AFPN) to address challenges in small target detection from the Unmanned Aerial Vehicle (UAV) perspectives, such as scale imbalance, feature extraction difficulty, occlusion, and computational constraints. The R-AFPN integrates three key modules: Residual Asymptotic Feature Fusion (RAFF) for adaptive spatial fusion and cross-scale linking, Shallow Information Extraction (SIE) for capturing detailed shallow features, and Hierarchical Feature Fusion (HFF) for bottom-up incremental fusion to enhance deep feature details. Experimental results demonstrate that R-AFPN-L achieves 50.7% AP_{50} on the TinyPerson dataset and 48.9% mAP_{50} on the VisDrone2019 dataset, outperforming the baseline by 3% and 1.2%, respectively, while reducing parameters by 15.1%. This approach offers a lightweight, efficient solution for small target detection in UAV applications.

Keywords UAV, Small target detection, Residual connectivity, Asymptotic feature fusion, Context learning

In recent years, with the rapid progress of UAV technology and the vigorous development of target detection algorithms, UAVs play an increasingly important role in many practical application scenarios, such as agricultural production[1], traffic monitoring[2], emergency rescue[3], etc., by virtue of their lightweight and convenient advantages, high flexibility, and strong endurance. However, with the increase of UAV flight altitude and the change of shooting viewpoints, the object scales in the images show remarkable diversity, especially containing numerous small-scale objects. These small-scale targets occupy very few pixels in the image and often lose some of the detail information due to occlusion, making detection and recognition extremely challenging and difficult to deploy in edge devices such as camera sensors in UAVs. Therefore, how to effectively improve the detection accuracy of small targets in UAV-captured scenes has become a key challenge to be overcome in the current technological development.

With the rapid popularity of deep learning-based target detection techniques in many fields[4–11], the problem of small target detection from the UAV perspective has taken a new turn. Deep learning-based target detection methods are usually divided into two categories: one-stage detection and two-stage detection. One-stage detection methods[12–18] directly classify and locate the prediction of the input image, pursuing high efficiency, while two-stage detection methods[19,20] first generate a series of candidate regions, and then meticulously classify and locate the prediction of these regions, which is known for its accuracy. However, directly applying existing deep learning-based object detection techniques to small object detection from UAV perspectives achieves suboptimal performance. The main challenge of the small target detection task is how to effectively avoid feature redundancy while retaining key details and semantic information in useful features. Although most deep learning detectors adopt the feature pyramid network (FPN)[12] as the basic structure to cope with the problem of excessive changes in object scales, this structure tends to lose detail information during feature transfer, leading to information degradation. To compensate for this deficiency, PANet[13] designs bottom-up and top-down dual-path structures on the basis of FPN, which alleviates the problem of detail loss between features of different layers to a certain extent. However, PANet still does not fully consider the detail loss or degradation when features of the same level are passed, as well as the semantic gap between non-adjacent hierarchical features. In addition, recent studies have shown that although some advanced networks[14,15] pursue higher detection scores through a large amount of computational resources, this is contrary to the high requirement of real-time target detection in the UAV perspective.

In the traditional design of Feature Pyramid Network (FPN), a common practice is to use up-sampling techniques to convert high-level features generated by the backbone network into low-level features. However,

School of Electronic Information and Xi'an Key Laboratory of High Precision Industrial Intelligent Vision Measurement Technology, Xijing University, Xi'an 710123, China. ✉email: yahongma@sina.com

this approach commonly suffers from the problem of information loss and degradation of the features during transmission and interaction. Notably, Bidirectional Feature Pyramid Network (BiFPN)[14] effectively reduces the information loss and degradation of same-level features in the transfer process by introducing Same-Scale Residual Connections. This improvement enables features to flow more smoothly between different levels of the pyramid and enhances the consistency and robustness of features. On the other hand, in the feature extraction stage, Asymptotic Feature Pyramid Network (AFPN)[15] gradually and repeatedly fuses low-level features with high-level features, and this process not only facilitates the in-depth interaction of cross-level features, but also generates richer and more hierarchical feature information. This progressive fusion helps to capture finer target features, especially when dealing with small-scale targets, and can significantly improve the detection accuracy. In order to solve the above problems, this research work focuses on a computationally network model structure that effectively reduces computational redundancy while improving the efficiency of small target detection, so as to meet the urgent needs of small target detection from the UAV perspective.

The main innovations and specific contributions of this paper are as follows:

(1) A residual asymptotic feature pyramid network (R-AFPN) is proposed to reduce the feature redundancy and degradation of the FPN by adaptive feature fusion and residual connectivity on the same scale while ensuring the probability of small-scale target detection.
(2) In order to utilise the shallower features, this paper introduces a context extractor-Acmix to capture the location information of the shallower features to enhance the recognition of small targets.
(3) A bottom-up progressive fusion path is proposed to make up for the lack of some detail information caused by multi-target information conflict in the feature fusion process, so as to reduce the leakage detection rate and false detection rate of small targets.
(4) Extensive experiments are conducted on TinyPerson and VisDrone 2019 datasets, and the comparison results with other feature pyramid networks and state-of-the-art detectors all demonstrate that the proposed method is more competitive in small target detection.

Related work

Currently, most of the small target detectors are improved from general purpose detectors. Among these methods, feature pyramid based on feature extraction and feature fusion, and context learning are the most popular methods. Next, related work in recent years is highlighted.

Feature pyramid

At the beginning of the development of deep learning, deep detectors based on convolutional neural networks (CNNs), such as Fast R-CNN and Faster R-CNN, mainly rely on the last layer of features of the backbone network for object detection. However, this approach appears to be incompetent when dealing with scenes that span a wide range of object scales. To solve this problem, Lin et al.[12] proposed the feature pyramid network (FPN), which exhibits robustness to multi-scale objects and has gradually become the basic structure of modern depth detectors. The FPN achieves the detection of objects of different scales by up-sampling the high-level features in the last layer of the backbone network and converting them to low-level features in a top-down manner. However, high-level features are not directly fused with low-level features in the up-sampling process of FPN. In view of this, Liu et al.[13] argued that this top-down structure of FPN lacks the rich localisation information of high-level features, and proposed PANet. PANet adds an extra bottom-up branching path on top of FPN, aiming to pass the low-level localisation information to high-level features so as to improve the accuracy of object localisation. Subsequently, Tan et al.[14] found in their study that there is a problem of missing and degraded information in the feature transfer process. In order to solve this problem, they proposed an additional end-to-end connection of input and output features of the same level in BiFPN. This approach fuses more features while only adding less computational cost, thus effectively addressing the problem of information loss and degradation in same-level transfer. In recent years, Yang et al.[15] also proposed an asymptotic feature pyramid network (AFPN) structure. They applied the adaptive spatial feature fusion module to the feature pyramid network, which not only avoids the loss of the highest level feature information in the FPN, but more importantly reduces the semantic information gap between the non-adjacent features by progressively fusing the neighbouring layer features and the non-adjacent layer features. In essence, these network structures are based on multi-scale learning and endeavour to reduce the loss and degradation of information in order to improve the accuracy and robustness of object detection.

Contextual learning

Contextual information plays a key role in helping networks to better understand images. It has been demonstrated that appropriately associating contextual contexts can provide additional information about an object, which is crucial for improving the accuracy of object recognition[21]. In particular, combining contextual information is especially important in small object scenes when external information is severely scarce. Zhang et al.[22] proposed a single-shot scale-invariant method. The method uses equidistantly distributed small anchor points and large filters to model the contextual information of an image. Subsequently, Tang et al.[23] further enhanced the accuracy of object recognition by simultaneously detecting the object and its surrounding background features. In addition, there are also studies that utilise attentional mechanisms to expand the receptive field to capture more contextual information. Zhang et al.[24] learned contextual information by establishing the correlation between the features obtained from the convolutional layer and those extracted by the Convolutional Block Attention Module (CBAM). However, some methods, such as Vision Transformer proposed by Xia et al.[25], while having a huge receptive field and being able to capture contextual information easily, use only self-attention modules, resulting in easy loss of positional information. To compensate for

this deficiency, Pan et al.[26] proposed ACmix, which combines self-attention and convolution and successfully solves the problem of location information loss.

Our approach

In this paper, we propose a residual asymptotic feature pyramid network structure based on same-scale connectivity, named R-AFPN, and its general architecture is shown in Fig. 1. The network takes CSP-Darknet53 as the backbone network and incorporates the residual asymptotic feature fusion module (RAFF), shallow feature fusion module (SIE), and progressive feature fusion module on this basis. In the feature extraction stage, the R-AFPN firstly utilises the adaptive spatial feature fusion (ASFF) technique to effectively fuse the P_3, P_4 and P_5 layers of features in the backbone network. Subsequently, in order to deliver the multi-layer semantic information to the output as completely and semantically similar as possible, RAFF adds a residual connection from the backbone input features to the output feature map at the same scale. In terms of shallow information extraction, R-AFPN introduces a context extractor called ACmix at the P_2 layer, which is a module that combines self-attention and convolutional operations, aiming to capture shallower positional information for more accurate positional sensing capability. Further, in order to compensate for some of the missing detail information that may be caused by multi-target information conflict during the feature fusion process, the R-AFPN designs a bottom-up progressive fusion path (HFF). This path passes the positional and detail information of shallow objects in the P_2 layer layer by layer to the higher-level adaptively fused features, in order to calibrate the global features and further enhance the multiscale features after residual linking. Finally, after a series of fine feature processing and fusion steps, the R-AFPN passes the output feature maps of the multiscale features to the detection head for classification and localisation, thus achieving accurate target detection. This network structure design not only effectively reduces the semantic gap between non-neighbouring layers, but also significantly reduces the information loss and degradation problems in the feature transfer process.

RAFF

As mentioned above, FPNs may retain some features that do not contribute much to the detection during feature processing, which not only reduces the detection speed of the model, but also causes a waste of computational resources. In order to optimise the network model and achieve a lightweight design, this paper proposes a residual asymptotic feature fusion module (RAFF), as shown in Fig. 2. This module combines the advantages of adaptive spatial feature fusion technique and residual connectivity, and can replace the two up-sampling process in FPN. In RAFF, adaptive spatial feature fusion is first performed on the P_3, P_4, and P_5 layer features of the backbone network to reduce the difference in semantic information between the features of non-adjacent layers, making the fused features more representative. Meanwhile, in order to further enhance the representativeness of the output features, an additional connecting line is added between the input features and the output features to form a residual connection. This design not only makes the output features retain enough similar semantic information, but also effectively avoids unnecessary feature redundancy. Therefore, by combining the advantages of adaptive spatial feature fusion and residual connectivity, the RAFF module achieves the improvement and optimisation of the FPN, which provides a new idea for the lightweight design of network models.

Considering that the shallower features contain rich detail information, as in Fig. 1, in this paper, P_2 is used as the lowest level feature, and $P_n \in R^{H_n \times W_n \times C_n}$ is used to denote the features output from the last four convolutional blocks in the backbone. Where $n \in \{2, 3, 4, 5\}$, H_n, W_n, and C_n represent the height, width, and number of channels of the features in the nth layer, respectively, and the computational formulas are shown in Eqs. (1)–(3).

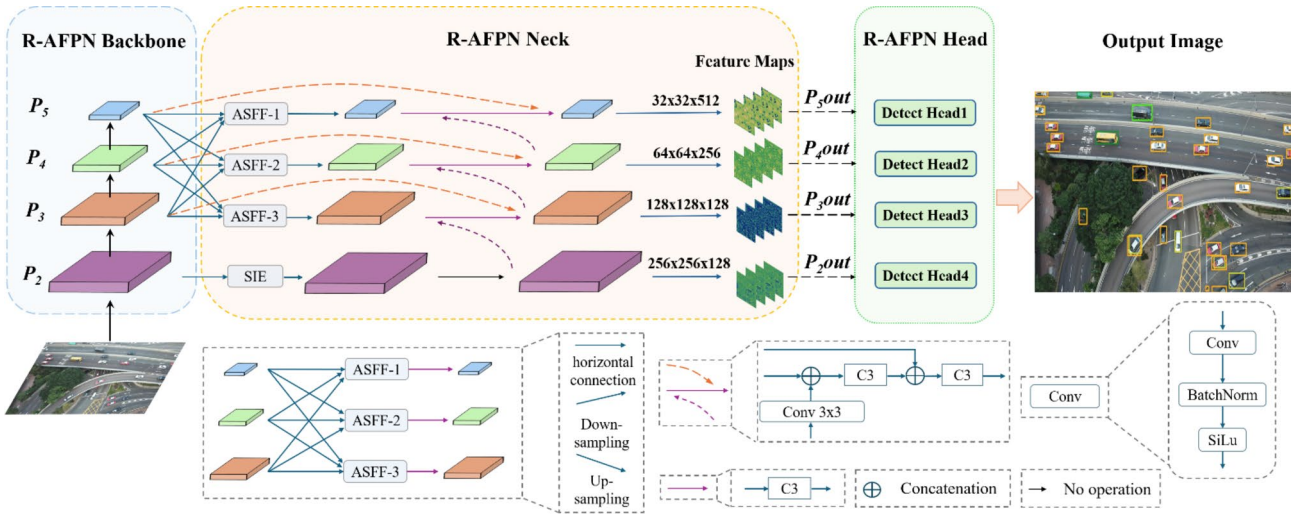

Fig. 1. Overall framework of the R-AFPN.

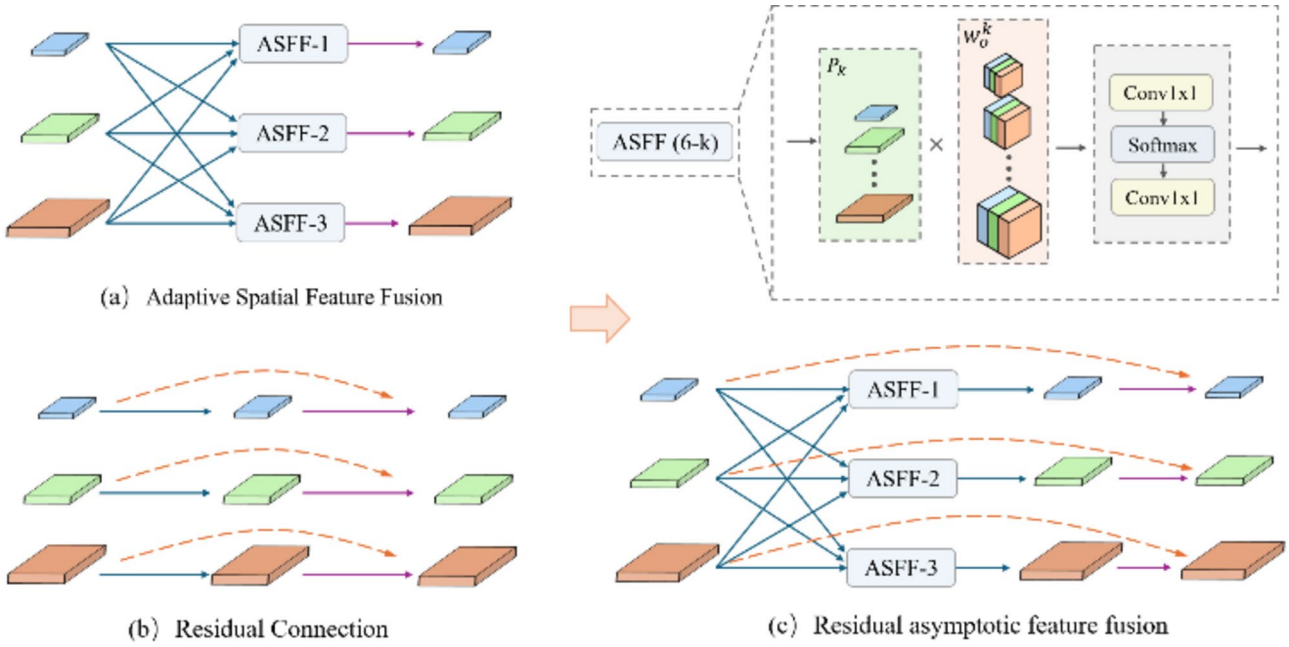

Fig. 2. RAFF module.

$$H_n = \frac{H_i}{2^n} \tag{1}$$

$$W_n = \frac{W_i}{2^n} \tag{2}$$

$$C_n = 2^{n+5} \tag{3}$$

where Hi, Wi, and Ci represent the height, width, and number of channels of the input image, respectively. As shown in Figs. 1 and 2a, the process of up-sampling, down-sampling and horizontal join is included in ASFF[27], which is to splice the weights of the three layers of features with the same spatial dimensions after up-/down-sampling, and then the weights of the spliced weights are assigned to the input features according to the different layers of spatial dimensions respectively and are summed up, so that the three layers of features reconstructed in this way can reduce the semantic gaps between P_3, P_4, and P_5. In this case, the up-sampling mode is bilinear interpolation and the down-sampling mode is 1×1 convolution. The global convolution is actually of CBS type, consisting of Conv, Batch Normalisation and Silu, Batch Normalisation is used to speed up the computation, and Silu is an activation function that combines linear and nonlinear properties, which can be expressed as Eq. (4).

$$Silu\,(x) = \frac{x}{1 + e^{-x}} \tag{4}$$

P_k denotes the feature map before spatial dimension splicing, W_o denotes the initial weight, and $k \in \{3, 4, 5\}$ denotes the kth layer feature map, then the reconstructed weight W_k and feature $P_k\prime$ in Fig. 2a can be expressed as Eq. (5) and Eq. (6), respectively.

$$W_k = soft\max\left[Conv_{1\times1}\left(\sum\nolimits_{k=3}^{5} w_o^k\right)\right] \tag{5}$$

$$P_k' = Conv_{1\times1}\left(\sum\nolimits_{k=3}^{5} P_k \times W_K\right) \tag{6}$$

The splicing operation of the feature map is denoted by $\|$ in Fig. 2b, then the output features after residual connection can be expressed as Eq. (7).

$$P_n'' = C3\left[P_n \parallel Conv_{1\times1}\left(P_n\right)\right] \tag{7}$$

Eventually, driven by residual linking horizontally and feature fusion vertically together, RAFF generates multiscale features $\left\{P_3^{(3)}, P_4^{(3)}, P_5^{(3)}\right\}$ with some detail conflicts but similar semantics and suppression of degradation as in Fig. 2c. The RAFF module is obtained after fusion and its output features can be expressed as Eq. (8).

$$P_n^{(3)} = C3\left(P_n \parallel P_k'\right) \tag{8}$$

SIE

Given that shallow low-level features contain richer detail information, a shallow information extraction (SIE) module is designed in this paper. In SIE, a module that mixes convolution and self-attention, called ACmix, is introduced to enhance the location information of shallow features, as shown in Fig. 3.

This paper uses eight-head self-attention to align the P_2 layer network, and use P_{conv} and P_{att} to denote the features after 3-kernel convolution and feature linking, self-attention and feature linking, respectively, and $\{P_{conv}, P_{att}\} \in R^{H_2 \times W_2 \times C_2}$. Thus, the output of ACmix can be easily expressed as Eq. (9).

$$P_{AC} = (\alpha \times P_{conv} + \beta \times P_{att}) \tag{9}$$

where P_{AC} represents the output features after ACmix, α and β denote different scaling factors respectively, and the sum of the two factors is 1.

Considering that when the spatial dimension is constant, the reduction of the number of channels will lose part of the detail information, which is not favourable for the detection of small-scale targets. In the SIE module, the number of channels is consciously kept consistent throughout the process. Finally, with the help of the contextual context, the backbone shallow input features are subjected to a point convolution and a C3 module to obtain the location- and detail-rich output feature $P_2' \in R^{H_2 \times W_2 \times C_2}$, which can be expressed as Eq. (10).

$$P_2' = C3\left[Conv_{1\times1}\left(P_{AC}\right)\right] \tag{10}$$

HFF

Considering that some detail information conflicts existing in ASFF will be passed to the output features, which will reduce the detection probability of small targets. Therefore, this paper designs a hierarchical feature fusion path (HFF), which passes the low-level output features with richer detail information to the high-level input features for calibration layer by layer, as shown in Fig. 4.

In this paper, the final output feature map obtained by adding the residual join is denoted as $P_n out \in R^{H_n \times W_n \times C_n}$, the input features of HFF are $\{P_2', P_3', P_4', P_5'\}$, and the output features without residual connection are expressed as $P_n^{(4)} \in R^{H_n \times W_n \times C_n}$. The HFF passes diagonally upwards through a 3×3 convolutional kernel for each pass to complete the downsampling and splice the downsampled low-level features with the high-level input features, and horizontally through a C3 module for each pass to keep the number of channels consistent. The output features of the HFF can be expressed as Eq. (11).

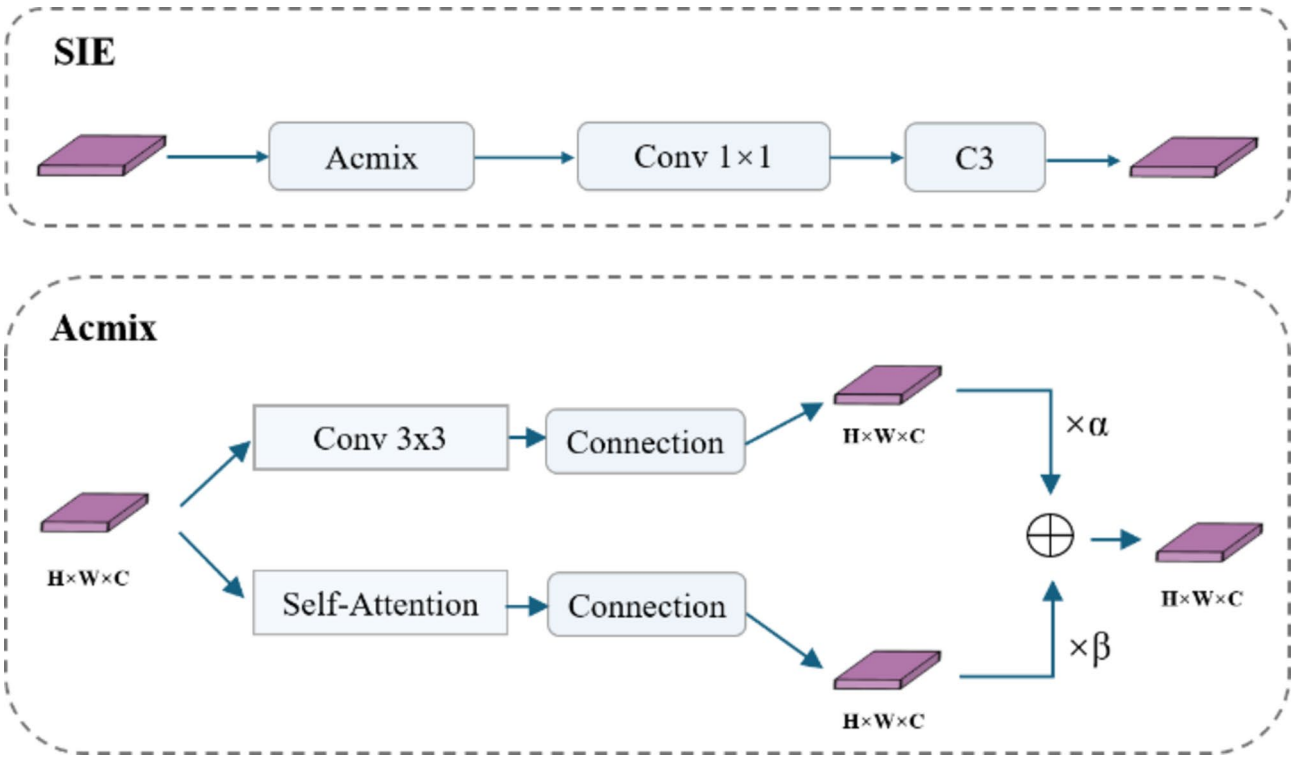

Fig. 3. SIE module.

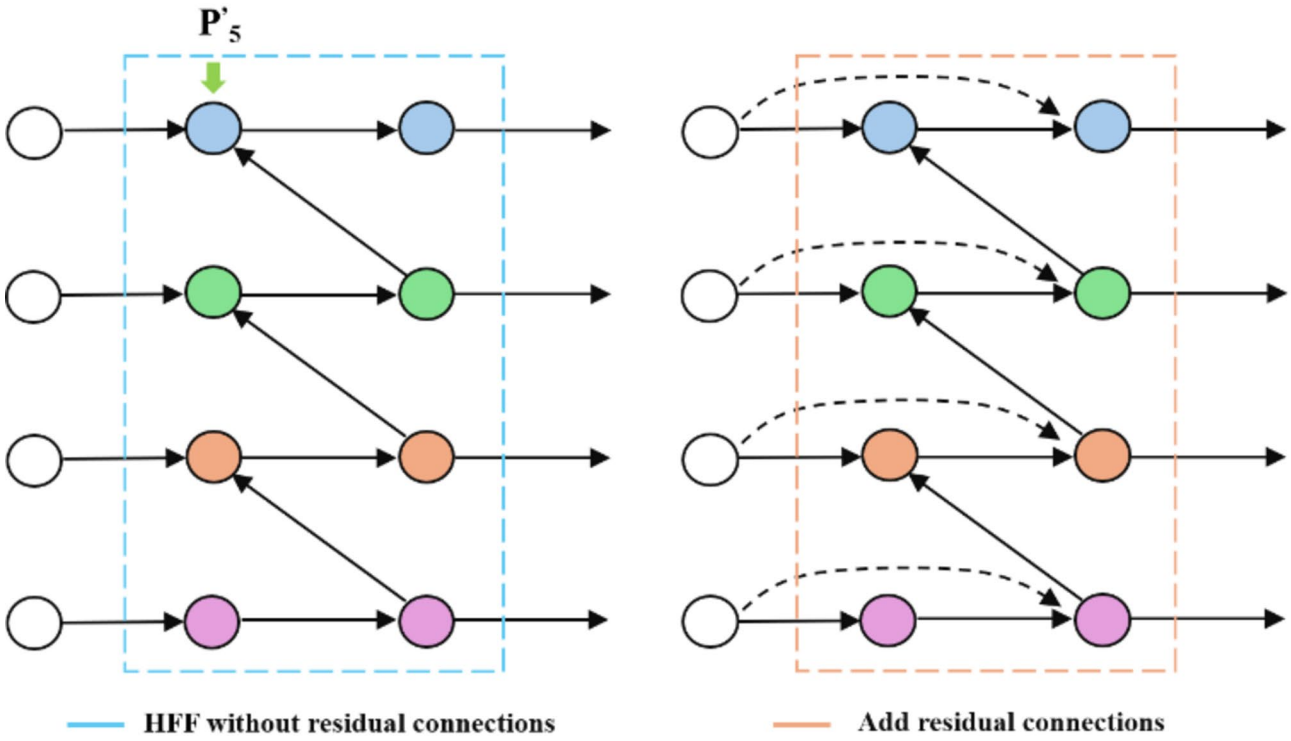

Fig. 4. HFF module.

Environment	Parameter
CPU	13th Gen Intel(R) Core(TM) i9-13900 K@3.00 GHz
GPU	NVIDIA GeForce RTX 4090
VRAM	24.0 GB
RAM	64.0 GB
Operating system	Windows 11 Professional Edition
Language	Python 3.8.19
Frame	Pytorch 2.2.2
CUDA version	11.8

Table 1. Configuration of training and testing experiment environments.

$$P_n^{(4)} = \begin{cases} P_2', & n = 2 \\ C3\left[Conv_{1\times1}\left(P_{n-1}^{(4)}\right)\right], & n > 2 \end{cases} \tag{11}$$

Ultimately, the multiscale feature map output from the R-AFPN network can be expressed as Eq. (12).

$$P_n = \begin{cases} P_2', & n = 2 \\ C3\left(P_n \parallel P_n^{(4)}\right), & n > 2 \end{cases} \tag{12}$$

Experimental configuration and evaluation indicators

Experimental configurations

All experiments in this study were conducted on a computer equipped with NVIDIA GeForce RTX 4090 GPU, 64 GB RAM, Python 3.8.19, and CUDA 11.8; see Table 1 for detailed hardware configuration.

During the training process, the input image resolution was uniformly set to 1024×1024, batch size to 4, Epoch to 300 times, and IoU threshold to 0.2, and the SGD optimiser was selected to cope with possible overfitting. A cosine decay strategy is also used to help generalise the model, and the learning rate is updated from 0.01. In addition, in order to balance the localisation loss, classification loss and confidence prediction loss, λ_b, λ_c and λ_o were set to 0.05, 0.5 and 1.0, respectively. The detailed hyperparameter configurations are shown in Table 2.

Epochs	Batch size	Initial learning rate	Final Learning rate	Optimizer	Patience	λ_b	λ_c	λ_O	IoU threshold	Weight-decay
300	4	1e−02	1e−03	SGD	100	0.05	0.5	1.0	0.20	5*1e−4

Table 2. Training Parameters Setting.

Datasets

Compared to YOLOv8L, YOLOv5L, which has faster inference speed, is chosen as the benchmark model in this study, with version 6.0. Considering the large difference between the R-AFPN and YOLOv5L in this paper, no pre-trained model is used in the experiments. In order to verify the effectiveness of this method for small target detection, the TinyPerson and VisDrone-DET 2019 datasets were used for training, respectively. All YOLO-based models were trained with random cropping and random splicing for data enhancement.

TinyPerson is a benchmark dataset designed for detection of tiny pedestrians at long range and in large backgrounds. It provides a new research direction in the field of tiny target detection. The dataset contains 736 training set images, 796 test set images, and a total of 72,651 manually labelled bounding box annotations. These annotations are classified into five categories based on the actual pixel size and scene features: 'sea', 'ground', 'ignore', 'uncertain' and 'dense'. However, in this study, in order to simplify the processing, all people labelled as tiny were treated as a single category 'people' and were evaluated uniformly on the test set.

VisDrone-DET 2019 is a dataset dedicated to small-target aerial image detection, originally designed for the ICCV 2019 International Computer Vision Competition, and later collected and manually bounding-box annotated by Tianjin University to serve as a benchmark dataset. The dataset covers 10,209 still images from 14 different cities in China, under a variety of scenes, weather and lighting conditions, and over 2.6 million annotations. Among them, the training set contains 6,471 images, the validation set contains 548 images, and the test set actually contains 1,610 images. To further validate the effectiveness of the network proposed in this paper, experiments were additionally conducted on this dataset and the network performance was evaluated on the test set.

Evaluation metrics

In this paper, average precision (AP), mean average precision (mAP), parameters, GFLOPS and FPS are chosen as evaluation metrics, and F1-score, Recall are included. The main consideration here is that small targets have less information, more false detections will seriously affect the final detection results, and focusing only on AP and mAP does not represent a better increase in model performance. We hope that the data alone can intuitively and accurately see the advantages and disadvantages of the detection methods. In addition, since the TinyPerson dataset contains only one category, the model is evaluated by AP. In contrast, the evaluation was performed by mAP on the VisDrone-DET 2019 dataset, which contains ten categories.

Second, the contribution of each module is demonstrated separately through ablation experiments. Next, the present network is trained on the TinyPerson and VisDrone-DET 2019 datasets in the UAV perspective, respectively, and compared with other advanced network models. Finally, the performance of the present network is comprehensively evaluated in terms of weighting parameters, average accuracy, and computational cost.

Experimental results and analysis

Different FPN performance comparison experiment

In order to evaluate the performance of the proposed algorithm in this paper, YOLOv5 V6.0 is used as the framework and compared with other feature pyramid networks on the TinyPerson dataset. Specifically, in this paper, CSP-Darknet53 is used as the backbone and the neck is compared using different feature pyramid networks as shown in Fig. 5. As can be seen from Fig. 5, the proposed R-AFPN is not only more competitive than the state-of-the-art AFPN, but also has the lowest weight parameters (Parameters). In particular, compared to PANet, the AP of R-AFPN is improved by 3% and the number of total parameters is reduced by 15.1%.

Ablation experiments and results

YOLOv5L V6.0 is used as the Baseline to do the ablation study and all the experiments are done on TinyPerson dataset. The impact of different modules, feature maps, input image size and up-sampling method on the model performance and the results are analysed below.

Impact of different modules on model performance

1) *RAFF:* To ensure consistency of the variables, here only the PANet of the original YOLOv5L neck needs to be replaced with the BAFF module, aligning the layers and connecting the multiscale output features to the head correctly. As shown in row 2 of Table 3, this module gains 22.6% and 21.6% improvement in Parameter and GFLOPS, respectively, over baseline, with only 0.4% reduction in AP_{50}. The original YOLOv5L dual path uses more convolution during upsampling and downsampling, which inevitably leads to the accumulation of weight parameters. In contrast, RAFF's adaptive fusion process and dimensional matching after residual concatenation contain fewer small convolution kernels, which is the key to the module's ability to achieve a lighter weight. Experimental results also confirm that RAFF effectively reduces redundant features while only slightly sacrificing object recognition accuracy.

2) *SIE:* Considering that SIE single outputs the feature information of P_2 layer, we chose to add SIE on top of BAFF to verify the role of SIE in the experiment. In Table 3, RAFF + SIE improves AP_{50} by 3.7%, Recall by

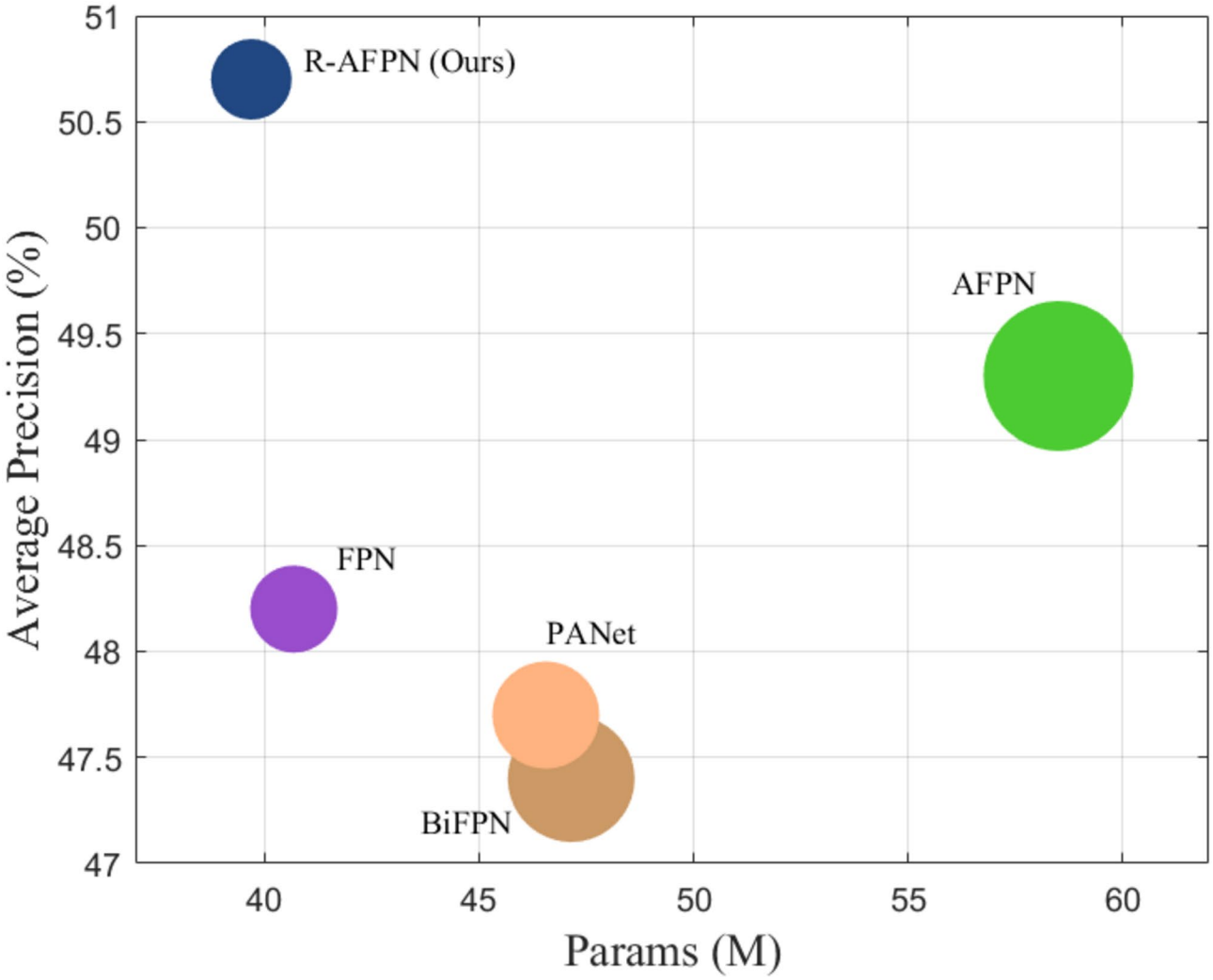

Fig. 5. HFF module.

Method	RAFF	SIE	HFF	F1-score↑	Recall↑	AP_{50}↑	Params↓	GFLOPS↓
Baseline	–	–	–	38	26.5	47.7	46.5 M	280.5
	✓	–	–	38	26.5	47.3	**36.0 M**	**219.9**
	✓	✓	–	44	32.6	**51.0**	36.2 M	246.7
	✓	✓	✓	**47**	**36.2**	50.7	39.5 M	259.6

Table 3. Results of Ablation Experiments. Significant values are in (bold).

6.1%, and F1-score by 6 points over BAFF alone. Correspondingly, the number of references increases by only 0.2 MB, and the GFLOPS increase by 26.8. Overall, RAFF + SIE obtains an all-round performance improvement over baseline. The experimental data shows that using SIE contributes significantly to the network proposed in this paper. In order to further analyse the contribution of shallower features and context learning, an additional comparison experiment is added: one group focuses on the P_2 layer features while the other group does not focus on the P_2 layer features without using the SIE module, and one group uses SIE while the other group does not use SIE while focusing on the commonality of the shallower P_2 features, where the P_2 is the shallower features and SIE is the shallow information extraction. module, indicating that Acmix was used to capture contextual information. As shown in Table 4, P2 + without-SIE that focuses on shallower features improves AP_{50} and Recall by 2.1% and 2.9% respectively compared to R-AFPN without-P_2, which proves that the shallow features have richer information about small target details. In Table 4, the complete R-AFPN with the SIE module improves AP_{50} and Recall by 0.6% and 5.3% over without-SIE with only 0.2 MB of parameter counts and one point of computational cost, confirming that adopting contextual learning to focus on shallower features can increase the probability of object discovery, and is not caused by focusing on shallower features alone.

Method	Recall↑	AP_{50}↑	$AP_{50\text{-}95}$↑	Params↓	GFLOPS↓
without-P_2	28	48	29.3	**39.1 M**	**227.5**
P_2 + without-SIE	30.9	50.1	**31.4**	39.3 M	236.5
Full R-AFPN	**36.2**	**50.7**	31.2	39.5 M	259.6

Table 4. Results of ablation experiments. Significant values are in (bold).

3) *HFF:* In Table 3, RAFF + SIE + HFF corresponds to the highest F1-score and Recall. Compared to RAFF + SIE, F1-score increases by 3% and Recall improves by 3.6%, but the number of parameters increases by 8.8%, AP_{50} decreases by 0.3%, and GFLOPS also improves slightly. As expected, the bottom-up progressive fusion path is able to trade off higher scores and further reduce the risk of false and missed detections by giving up only a portion of the lightweighting. It is believed that this approach is well worth the effort. In addition, overall compared with baseline, Recall improves by 9.7 percentage points, the score increases from 38 to 47, AP_{50} improves by 3%, the weight parameter decreases by 15.1%, and the GFLOPS decreases by 20.9, which significantly improves the object recognition effect while possessing a certain degree of lightness, and also shows that the three modules are compatible with each other.

Comparative analysis of output feature maps
The input images and output feature maps are compared in order to visualize the impact of each module in the ablation experiment on the model performance. In this paper, P3 layer output feature maps with consistent resolution size are captured in the ablation experiments and heat maps are generated to visualize the regions that the model pays more attention to. As shown in Fig. 6, the models in Baseline pay less and varying attention to the region of the person to be detected. When using only the RAFF module, the model successfully shifts its attention to the object to be detected and starts to prioritize the approximate region of the person to be detected. After adding the SIE module, the model will prioritize the background region to determine the region of the person to be detected, so that it can pay better attention to the location point of the person, but it is difficult to focus on all the people. Eventually, after adding the HFF module to the complete R-AFPN, the model is able to accurately focus on the location point of each person to be detected.

Comparative analysis of input image size
This experiment evaluates the effect of input image size on model performance by training and testing network models using input images of different resolutions. Specifically, three additional 768×768, 1024×1024 and 1280×1280 were selected for comparison on top of the default 640×640 of the one-stage detector, and evaluated by the area under the PR curve (AUC-PR). As shown in Fig. 7, both AUC-PRs increased significantly from 640×640 to 1024×1024. This indicates that the model performance of both the baseline and the present method has been significantly improved. While from 1024×1024 to 1280×1280, the AUC-ROC is not improved much. Considering the fact that increasing the input size in a single step will increase the computational complexity of the model and the risk of overfitting. Therefore, in this paper, the input size of the network is set to 1024×1024 to help the model acquire finer-grained features in small target detection and improve the detection precision and recall.

Comparative analysis of up-sampling modes
This paper further explores the impact of the up-sampling modes used for adaptive feature fusion in the RAFF module on the R-AFPN, including nearest neighbour interpolation, cubic interpolation, bilinear interpolation and regional interpolation. The structure and settings of all models are kept the same, except for the different sampling modes, and twofold and fourfold upsampling is performed according to ASFF. As shown in Table 5, R-AFPN using Area is the highest in AP_{50} and R-AFPN using Nearest is optimal in FPS, but neither is the best choice. On the contrary, BiAFPN using Bilinear is optimal in recall, the difference in AP_{50} is very small compared to Area, and the FPS can meet the real-time requirement. Taking all the considerations into account, this paper adopts Bilinear as the upsampling mode in RAFF.

Comparisons with other networks
In this section, the present method is compared with other methods on two datasets. All models were trained in the same configuration as described earlier.

Qualitative comparison
The qualitative comparison of the output feature maps on the TinyPerson dataset and the VisDrone-DET 2019 dataset is shown in Figs. 8 and 9, the methodology is the same as that in Chapter 5, Section B, Part b. The qualitative comparison of the image detection is shown in Figs. 10 and 11. Where the baseline is YOLOv5L + PANet and FPN is a one-stage detector based on YOLOv5L.

In Fig. 8, the feature map of this paper is the most effective among similar models, which can focus on the background and the people to be tested in an orderly manner, and then focus on the specific areas of all the people to be tested, which well solves the problem of distraction in similar models. Especially compared to YOLOv10 and YOLOv11, this phenomenon is more obvious. It also shows that the present method rides on the small target detection in a single category. In contrast, in Fig. 9, this approach of the present model cannot focus on all the objects to be detected at the same time, but focuses on them successively according to the category they

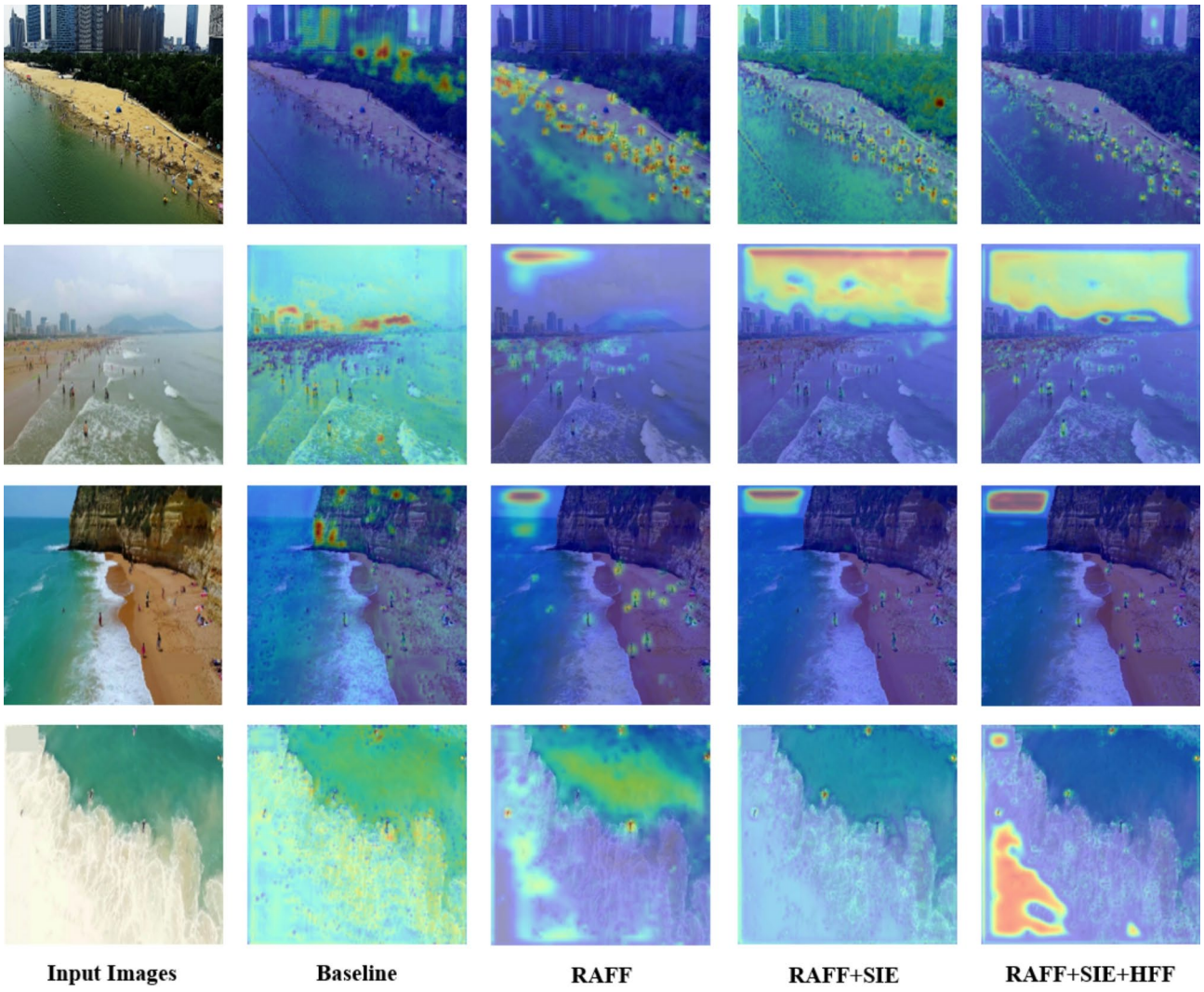

Fig. 6. Visualisation of output feature maps in ablation experiments.

belong to. While YOLOv10 and YOLOv11 are improved on the basis of YOLOv5, the model can better focus on all the objects to be detected. Next, this paper will evaluate the advantages and disadvantages of these two approaches through empirical results.

In Fig. 10, it is clear that the R-AFPN designed in this paper is able to accurately detect more tiny people, while the baseline model and the FPN barely detect this part of the objects, which solves the problem that the previous models have a high number of detection failures on the TinyPerson dataset with a proportion of small-scale objects. For example, the baseline and FPN missed detecting a higher number of TinyPersons in the distant dense regions #1 and #3 in Fig. 10. In contrast, R-AFPN performs very well for dense region targets. Meanwhile, in region #2, where half of the body remains on the surface of the seawater, the baseline and FPN detected fewer persons than the R-AFPN. The advantages of R-AFPN are also evident in region #4 where the brightness is dark under the masking umbrella. These results show that the network model designed according to the method in this paper is highly adaptable to the scene under the UAV viewpoint.

In Fig. 11, the R-AFPN in this paper is better at detecting small and distant objects in UAV scenes with large scale differences, such as the case of regions #1, #2 and #5. When the camera is shooting at high altitude with a vertical angle, the baseline and FPN are less discriminating for dense objects, while our model is always able to accurately discriminate and correctly categorise them. In addition, for occluded objects, such as the occluded van in region #4, the FPN misidentifies it as a small car, while the baseline considers it belongs to both categories, whereas the R-AFPN in this paper is able to discriminate its type well. All these phenomena reveal that the R-AFPN has good robustness to scale-varying differences.

Quantitative comparison

Table 6 shows the comparison results of this method with some other state-of-the-art methods on the TinyPerson training set. In order to evaluate the small-scale detection performance of each model in a richer way, this work adds additional model comparison results with a default input image size of 640 × 640 to the experimental preset

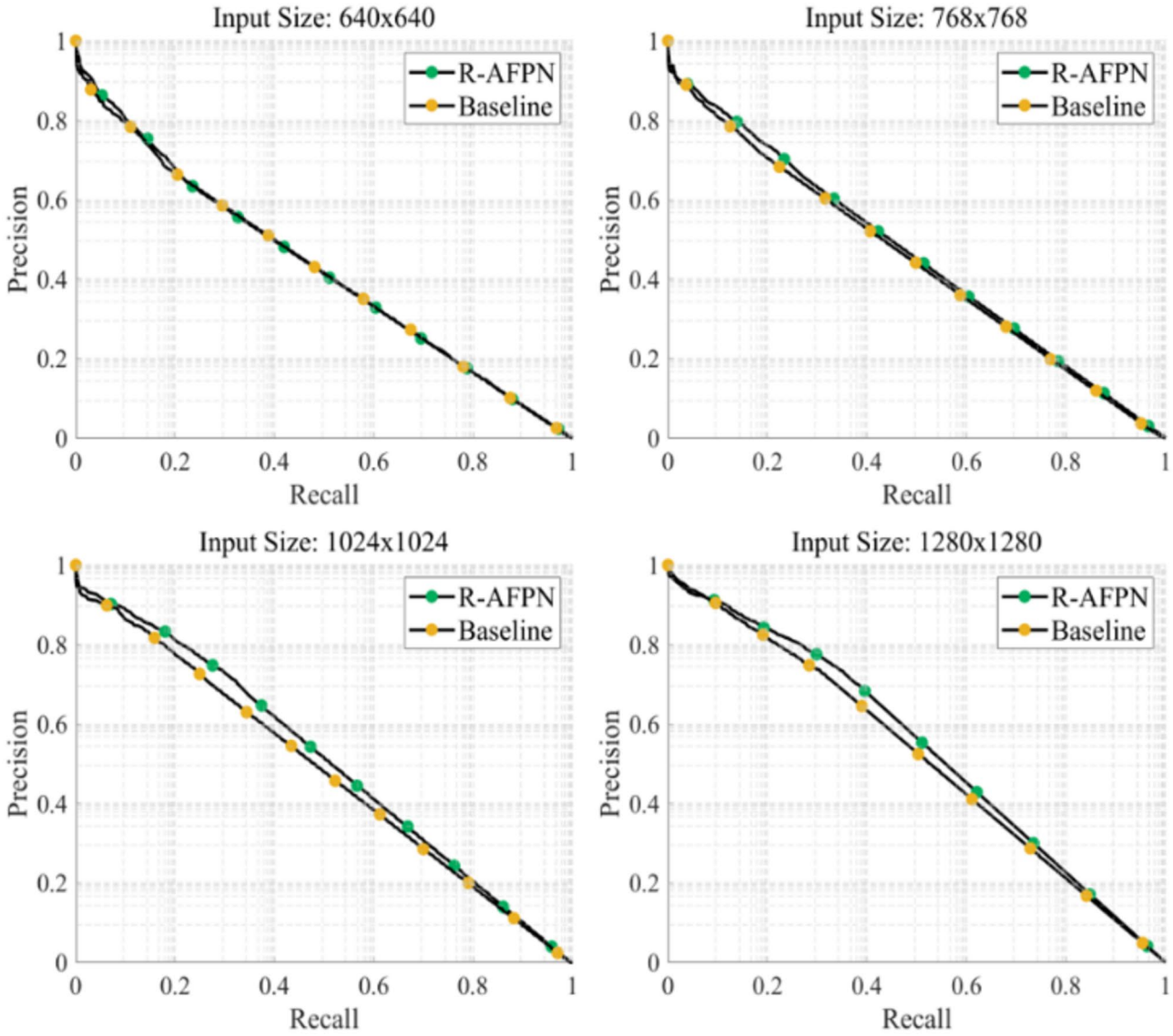

Fig. 7. Input Resolution Comparison.

Methods	Upsample mode	Recall↑	AP_{50}↑	GFLOPS↓	FPS↑
R-AFPN	Nearest	33.4	50.8	**257.8**	**37.6**
	Bicubic	35.9	50.8	260.5	37.3
	Bilinear	**36.0**	51.2	257.9	36.8
	Area	34.0	**51.5**	**257.8**	37.2

Table 5. Validation results of R-AFPN with different upsampling modes on the TinyPerson dataset. Significant values are in (bold).

of 1024 × 1024. The comparison reveals that R-AFPN achieves the highest scores on all metrics related to object recognition accuracy, outperforming the second-ranked AFPN by 1.4% on AP_{50}, and outperforming YOLOv11 by 2.2% and 2.8% on Recall on low- and high-resolution images, respectively. Unfortunately, the present method is not optimal in the comparison of Parameter and GFLOPS. Although R-AFPN does not excel in network model lightweighting, we should pay more attention to the detection accuracy of small-scale targets. This is because it is much more important to improve the detection accuracy of small targets that lack sufficient detail information compared to the lightweighting of the network model, and the present method is capable of meeting the real-time requirements of UAVs, as detailed in Table 5.

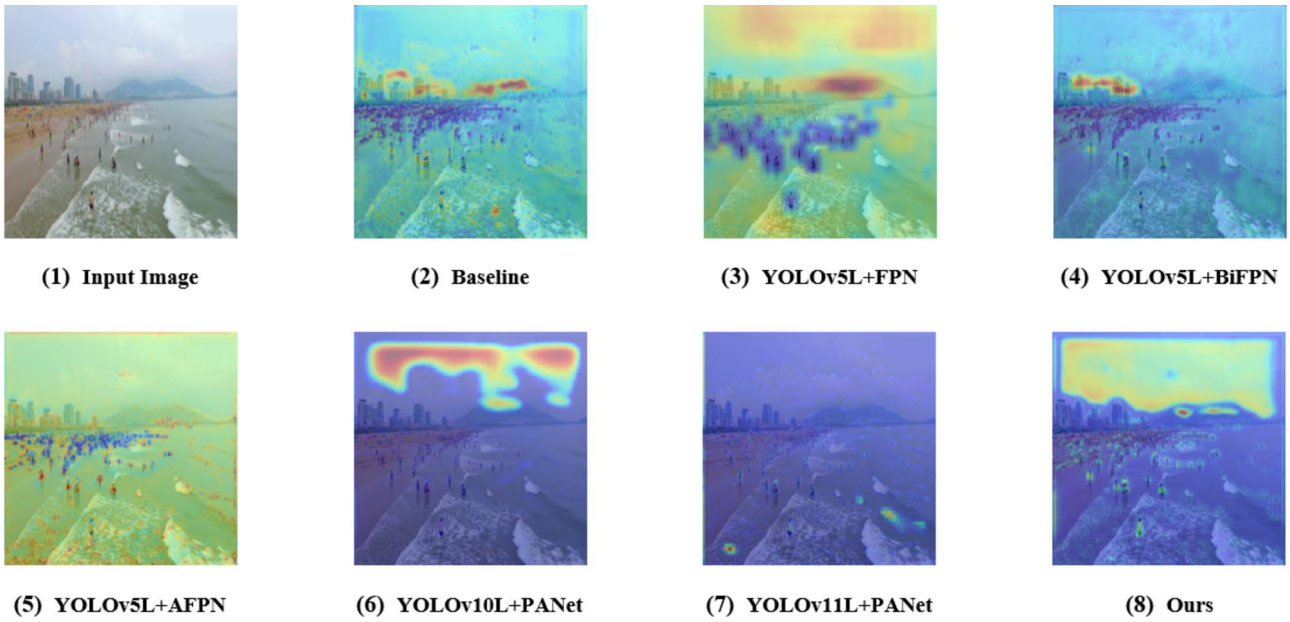

Fig. 8. Comparison of output feature maps on the TinyPerson dataset.

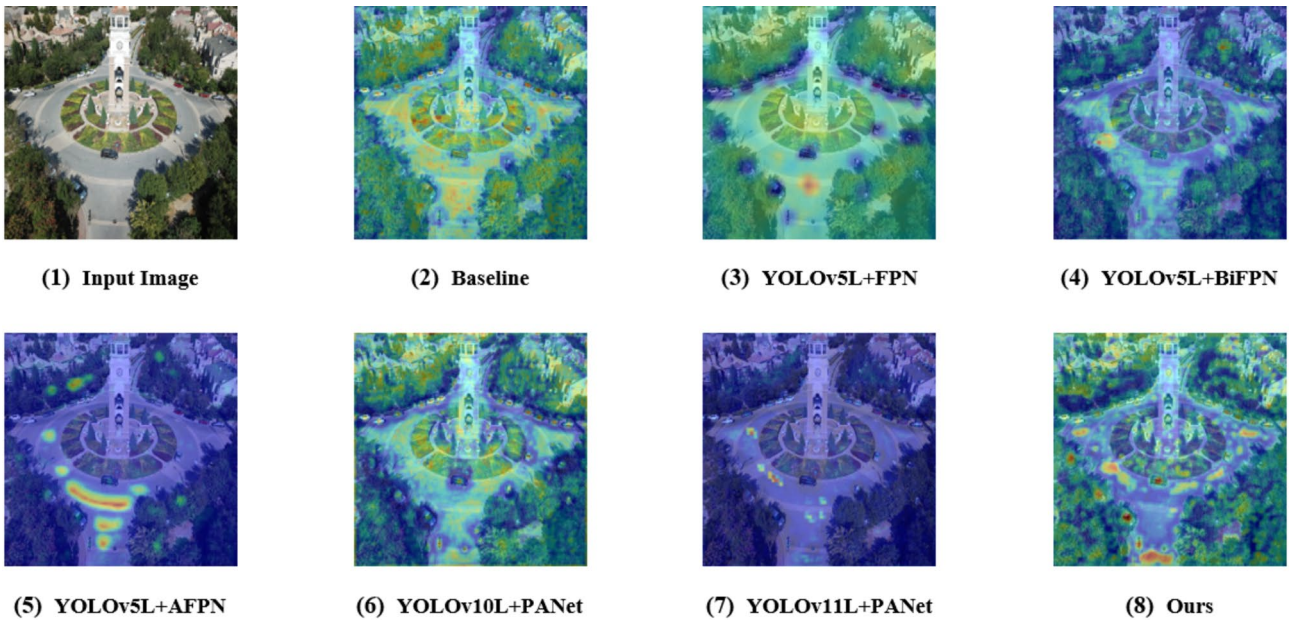

Fig. 9. Comparison of output feature maps on the VisDrone-DET 2019 dataset.

Table 7 shows the results of the comparison of the present method with some other state-of-the-art methods on the VisDrone-DET 2019 validation set. The comparison reveals that the present method reaches the optimum only on mAP50-95, while it is 5.8% lower than the best YOLOv10 on mAP50. Upon comprehensive analysis, it is found that the present method is still leading on tiny scale objects, and although it outperforms some models overall, the robustness to cope with large scale gaps is not optimal.

Conclusion and future work

This paper proposes a novel feature pyramid architecture, termed the Residual Asymptotic Feature Pyramid Network (R-AFPN), which utilizes multi-scale feature fusion and contextual learning to achieve robust detection. The main contributions of the proposed R-AFPN are as follows. Firstly, this paper designs a residual asymptotic feature fusion module (RAFF) based on AFPN and BiFPN to enhance the feature representation of the feature maps used to predict objects. Secondly, this paper focuses on more shallow features in conjunction

Fig. 10. Comparison of image detection results on the TinyPerson dataset.

Fig. 11. Comparison of image detection results on VisDrone-DET 2019 dataset.

with the contextual context as a way to improve the location information and detail information of global features. Therefore, a shallow information extraction module (SIE) and a progressive feature fusion module (HFF) are designed. Experiments were conducted on the challenging TinyPerson and VisDrone-DET 2019 datasets to demonstrate the optimal performance of the proposed R-AFPN. Experimental results show that R-AFPN significantly improves the detection results for small targets and achieves state-of-the-art performance on the single-category TinyPerson dataset. However, it still falls short compared to the state-of-the-art network on the multi-category VisDone-DET dataset, which is caused by the model's different focus on multi-category targets. As a result, the robustness of the model on such datasets is not optimized. Therefore, when deploying the application on UAVs, it is necessary to consider detecting for a single category of small targets and avoiding simultaneous detection of multiple categories of targets in order to maximize the best performance of this network. In the follow-up work, it will be considered that designing a new feature pyramid network to make up for this shortcoming, and extend the proposed method to the work of large target detection, so that it has a

Method	Backbone	Size	Recall↑	AP_{50}↑	AP_{50-95}↑	Params↓	GFLOPS↓
YOLOv5L + PANet	CSPDarkNet53	640 × 640	19.8	42.5	25.3	46.2 M	275.2
YOLOv8L + PANet	CSPDarkNet53	640 × 640	20.5	22.4	7.0	43.6 M	163.1
YOLOv10L + PANet	Enhanced-CSPNet	640 × 640	20.6	20.6	6.6	25.7 M	126.3
YOLOv11L + PANet	Enhanced-CSPNet	640 × 640	20.3	22.2	7.0	**25.3 M**	**86.6**
YOLOv5L + R-AFPN(our)	CSPDarkNet53	640 × 640	**22.5**	**42.8**	**25.6**	39.5 M	257.9
YOLOv5L + PANet	CSPDarkNet53	1024 × 1024	26.5	47.7	29.0	46.2 M	277.1
YOLOv5L + FPN	CSPDarkNet53	1024 × 1024	26.7	48.2	29.3	40.7 M	253.4
YOLOv5L + AFPN	CSPDarkNet53	1024 × 1024	29.6	49.3	30.4	58.5 M	331.8
YOLOv5L + BiFPN	CSPDarkNet53	1024 × 1024	23.8	47.4	29.4	46.7 M	283.5
YOLOv8L + PANet	CSPDarkNet53	1024 × 1024	32.7	36.5	12.5	43.6 M	165.4
YOLOv10L + PANet	Enhanced-CSPNet	1024 × 1024	31.8	32.5	11.1	25.8 M	127.2
YOLOv11L + PANet	Enhanced-CSPNet	1024 × 1024	33.4	35.1	12.0	**25.3** M	**87.3**
YOLOv5L + R-AFPN(our)	CSPDarkNet53	1024 × 1024	**36.2**	**50.7**	**31.2**	39.5 M	259.6

Table 6. Comparative experiments with other state-of-the-art networks on the TinyPerson Training Set. Significant values are in (bold).

Method	Backbone	Size	Recall↑	mAP50↑	mAP50-95↑	Params↓	GFLOPS↓
YOLOv5L + PANet	CSPDarkNet53	1024 × 1024	44.8	47.8	39.5	46.2 M	276.0
YOLOv5L + FPN	CSPDarkNet53	1024 × 1024	43.6	48.3	39.7	40.7 M	252.3
YOLOv5L + AFPN	CSPDarkNet53	1024 × 1024	44.9	49.1	**40.7**	58.5 M	329.5
YOLOv5L + BiFPN	CSPDarkNet53	1024 × 1024	45.0	48.5	40.2	46.8 M	282.4
YOLOv10L + PANet	Enhanced-CSPNet	1024 × 1024	51.7	**54.6**	34.2	25.8 M	126.4
YOLOv11L + PANet	Enhanced-CSPNet	1024 × 1024	**51.9**	54.5	33.9	**25.3 M**	**86.6**
R-AFPN (our)	CSPDarkNet53	1024 × 1024	43.8	48.8	**40.7**	39.5 M	258.5

Table 7. Comparative Experiments with Other State-of-the-art Networks on the VisDrone Validation Set. Significant values are in (bold).

stronger scene application effect. The experimental data and network models supporting the results of this study are available at this link: https://github.com/kaixinxiaobaobei/zuowen.

Data availability

The experimental data and network models supporting the results of this study are available at this link: https://github.com/kaixinxiaobaobei/zuowen.

Received: 22 February 2025; Accepted: 24 April 2025
Published online: 09 May 2025

References

1. Tetila, E. C. et al. Automatic recognition of soybean leaf diseases using UAV images and deep convolutional neural networks. *IEEE Geosci. Remote. Sens. Lett.* **17**, 903–907 (2020).
2. Arshad, M. A. et al. Drone navigation using region and edge exploitation-based deep CNN. *IEEE Access.* **10**, 95441–95450 (2022).
3. Liu, S. et al. Explainable attention-based UAV target detection for search and rescue scenarios. *IEEE Internet Things J.* https://doi.org/10.1109/JIOT.2024.3519158 (2024).
4. Wang, H., Liu, C., Cai, Y., Chen, L. & Li, Y. YOLOv8-QSD: An improved small object detection algorithm for autonomous vehicles based on YOLOv8. *IEEE Trans Instrum Meas.* **73**, 1–16 (2024).
5. Jiang, L. et al. MFFSODNet: Multiscale feature fusion small object detection network for UAV aerial images. *IEEE Trans Instrum Meas.* **73**, 1–14 (2024).
6. Sun, X., Liu, K., Chen, L., Cai, Y. & Wang, H. LLTH-YOLOv5: A real-time traffic sign detection algorithm for low-light scenes. *Automot. Innov.* **7**, 121–137 (2024).
7. Zhou, G. Q., Qian, L. H. & Gamba, P. A novel iterative self-organizing pixel matrix entanglement classifier for remote sensing imagery. *IEEE Trans. Geosci. Remote Sens.* **62**, 1–21 (2024).
8. Zhuang, J. et al. Infrared weak target detection in dual images and dual areas. *Remote Sens.* **16**, 3608 (2024).
9. Xu, K. C. et al. HiFusion: An unsupervised infrared and visible image fusion framework with a hierarchical loss function. *IEEE Trans. Instrum. Meas.* **74**, 1–16 (2025).
10. Wang, Z. H. et al. MLP-Net: Multilayer perceptron fusion network for infrared small target detection. *IEEE Trans. Geosci. Remote Sens.* **6**, 1–13 (2025).
11. Wang, B. Y. et al. A novel embedded cross framework for high-resolution salient object detection. *Appl. Intell.* **55**, 277 (2025).
12. Lin, T. Y. et al. Feature pyramid networks for object detection. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 936–944 (2017).
13. Wang, K., Liew, J. H., Zou, Y., Zhou, D. & Feng, J. PANet: Few-shot image semantic segmentation with prototype alignment. in *Proceedings of the IEEE/CVF International Conference on Computer Vision* 9196–9205 (2019).

14. Tan, M. Pang, R. & Le, Q. V. EfficientDet: Scalable and efficient object detection. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 10778–10787 (2020).
15. Yang, G., Lei, J., Tian, H., Feng, Z. & Liang, R. Asymptotic Feature pyramid network for labeling pixels and regions. *IEEE Trans. Circuits Syst. Video Technol.* **34**, 7820–7829 (2024).
16. Varghese, R. & Sambath, M. YOLOv8: A novel object detection algorithm with enhanced performance and robustness. in *ADICS* 1–6 (2024).
17. Wang, A. et al. YOLOv10: Real-time end-to-end object detection. Preprint at https://doi.org/10.48550/arXiv.2405.14458 (2024).
18. Khanam, R. Hussain, M. YOLOv11: An overview of the key architectural enhancements. Preprint at https://doi.org/10.48550/arXiv.2410.17725 (2024).
19. Girshick, R. Fast R-CNN. in *Proceedings of the IEEE International Conference on Computer Vision* 1440–1448 (2015).
20. Ren, S., He, K., Girshick, R. & Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**, 1137–1149 (2017).
21. Oliva, A. & Torralba, A. The role of context in object recognition. *Trends Cogn. Sci.* **11**, 520–527 (2007).
22. Zhang, S. et al. S^3FD: Single shot scale-invariant face detector. in *Proceedings of the IEEE International Conference on Computer Vision* 192–201 (2017).
23. Tang, X. Du, D. K. He, Z. & Liu, J. PyramidBox: A context-assisted single shot face detector. in *Proceedings of the European Conference on Computer Vision (ECCV)* 797–813 (2018).
24. Zhang, L. et al. Real-time lane detection by using biologically inspired attention mechanism to learn contextual information. *Cogn. Comput.* **13**, 1333–1344 (2021).
25. Xia, Z. Pan, X. Song, S. Li, L. E. & Huang, G. Vision transformer with deformable attention. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 4794–4803 (2022).
26. Pan, X. et al. On the integration of self-attention and convolution. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 815–825 (2022).
27. Liu, S. T. Huang, D. & Wang, Y. H. Learning spatial fusion for single-shot object detection. Preprint at https://doi.org/10.48550/arXiv.1911.09516 (2019).

Acknowledgements

This work was supported in part by the General Projects of Shaanxi Science and Technology Plan (No.2023-JC-YB-504), the National Natural Science Foundation of China (No. 62172338), the Xijing University special talent research fund (No.XJ17T03).

Author contributions

Z.C. wrote the main manuscript text and designed all the experiments. Y.M. and J.L. provided laboratory resources and financial support. M.C., and Y.Y. assisted in the initial experimental setup. Z.G., Z.W., and T.W. assisted in the code collection and reproduction of the comparison experiments. Y.L. completed the dataset collection.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Y.M.