Research on Feature Enhancement for Small Object Detection

Yuyao Tang¹, *, Bingxue Gu², b and Bo Jiang³, c*
¹East China Institute of Computing Technology, Shanghai, Shanghai, China
²East China Institute of Computing Technology, Shanghai, Shanghai, China
³East China Institute of Computing Technology, Shanghai, Shanghai, China
*aemail: ty909@mail.ustc.edu.cn, bemail: bingxueG@mail.ustc.edu.cn,
c* Corresponding author’s e-mail: b26jiang@126.com

Abstract. Small object detection is a challenging research direction in the field of computer vision, due to the low resolution and restricted information of small objects. At present, the general detectors only use appearance features to classify and locate objects, but they are prone to failure under the interference of background noise. On the other hand, the detector based on deep neural network has excellent performance on large scale, but it is difficult to extract enough information of small objects. This paper proposes a feature enhancement network (FENet), which contains two modules. The Residual feature enhancement (RFE) module combines residual learning and sub-pixel convolution to improve the resolution of input small objects and remove image noise. The Attention Feature Pyramid (AFP) module integrates the feature pyramid and attention mechanism, which can extract context information and filter redundant context information. At the same time, considering the imbalance of the contribution of large and small objects to the loss function during the training process, a feedback-driven function is introduced to solve the problem of uneven loss under multiple scales. Experimental results show that compared with the existing small object detection methods, our method has better performance.

1. Introduction

Object detection is an important research direction in the field of computer vision, and it has a wide range of applications in the fields of surveillance security, autonomous driving, traffic monitoring, drone remote sensing, and so on. Traditional object detection algorithms use manual feature design with low accuracy, while the development of deep learning has greatly improved the target detection performance. However, although most of the detectors based on deep learning can achieve accurate detection of large and medium-sized targets, the performance of accurate detection of small objects is poor.

Small object occupies a small proportion of the image, has a low resolution, and can extract very little effective information. There are three difficulties in the detection of small objects based on deep learning: (1) The features generated by the basic CNN from a single layer do not contain enough feature information for small object detection (2) For small objects in complex backgrounds, they are easily disturbed by similar objects. The model lacks contextual information to aid identification. (3) Most detectors do not consider image noise and other quality issues, but improving the image resolution and image denoising can greatly increase the performance of small object detection.

Regarding the lack of context information for small objects, research scholars often consider using feature pyramids [1] to extract context information. MPFPN [2] can prevent small targets from losing
information at a deep level. The literature [3] proposed a bi-directional stepped concatenation feature pyramid to avoid the loss of the current layer information during the pyramid construction process. The literature [4] improves the object detection performance by fusing the information between the two feature pyramids. However, the above-mentioned documents only consider the extraction of context information, and do not process the extracted redundant context information. Our method combines the attention mechanism and feature pyramid, which can filter the extracted context information, reduce redundant information, and improve detection performance.

As for the lack of sufficient small object detection information in a single layer, we usually think of multi-scale feature fusion methods. Especially for small targets, it is very important to combine the underlying feature information with the deep semantic information. FFDN (Feature Fusion Detection Network) [5] can achieve feature fusion of multiple resolutions. MFSOD (multi-scale feature fusion single shot object detector based on DenseNet) [6] could fuse multi-scale features of different levels. Our method integrates skip connection and up-sampling operations to propagate high-level semantic information back to low-level feature mapping, so that obtain the spatial and semantic information required for small object detection.

Considering the quality of the detected image, most studies would choose to introduce super-resolution into object detection, because small object detection always benefits from large scale. The literature [7] uses Perceptual GAN to enhance the feature information of small objects. The literature [8] uses sub-pixel convolution on the top layer of DenseNet to detect small targets, but it is limited by the limited information of a single feature map. The literature [9][10] refers to the texture and content of the image itself to achieve super-resolution images. However, none of them consider the noise of the image itself. Our method can not only use super-resolution operations to enhance the feature information of small targets, but also use residual learning to remove image noise. In addition, in the iterative learning process, compared with the large target, the loss of the small target accounts for a small proportion of the total loss. Therefore, in order to balance the loss distribution under different scales, we introduce the feedback-driven [11] loss function to reduce the leakage of the small target. The number of inspections and false inspections.

We conducted ablation studies and comparative experiments based on the data set we collected. The results show that our method is superior to other state-of-the-art methods, and each component can effectively improve the performance of the model. In summary, our contributions are as follows:

1) We designed an RFE module to obtain more small object feature information through residual learning and sub-pixel convolution, and improve the resolution of small object features and remove the noise in the image.
2) We designed an AFP module that integrates the feature pyramid and attention mechanism, and realizes the fusion of high-level semantic information and shallow spatial information to prevent the loss of small target information after deep convolution.
3) Considering that there is a significant gap between the loss provided by the small object and the loss provided by the large target in the iterative training process, we introduce a feedback-driven loss function to balance the loss of different scales, thereby reducing the number of missed detections of small objects.

2. Materials and Methods
We propose the network structure of FENet as shown in the Figure 1, which will be described in detail in this section. The RFE module can remove image noise and perform super-resolution processing. The AFF module introduces a feature pyramid and attention mechanism to solve the problem that a single detection layer does not contain enough information for small object detection. In addition, we have also introduced a feedback-driven function in yolo head, which can solve the problem of loss imbalance under multiple scales.
2.1 Image preprocessing
Small objects have small pixels and occupy a small proportion in the original image, so they are difficult to detect. On the other hand, in common object detection models, the general basic backbone neural network often has several down-sampling processing, resulting in the size of small objects in the feature map is basically only a single-digit pixel size. Therefore, we first crop the small objects in the picture, which is similar to enlarging the proportion of the small objects in the picture. We divide the input picture \( I_{hw} \) into a grid of \( n \times n \), and crop the picture to generate a picture \( I_{iPhw}^{iPhw} i,n = \) with the grid size and the order from top to bottom and from left to right. And \( /, / hh n ww n \equal \) . You can adjust the size of \( n \) as you want, and in our experiment, we set \( n=7 \).

2.2 Residual Feature Enhancement Module
The structure of the RFE module is shown in the Figure 1. First, the input image \( P_i \) generates a noisy residual image through the DnCNN network. The denoised image \( P_d \) extracts the main content features through the content extractor, and then generates a higher-resolution feature map through sub-pixel convolution. The texture extractor extracts the texture details of the \( P_d \), and finally integrates the results of the content extractor to output \( P_f \). The output \( Pf \) can be defined as:
\[
(P_f = E_c(P_i - D(P_i)) \oplus E_t(P_i - D(P_i)) \uparrow 2x + E_c(P_i - D(P_i)) \uparrow 2x
\]
where \( D(\cdot) \) represents DnCNN component, \( E_t(\cdot) \) denotes the texture extraction module and \( E_c(\cdot) \) denotes the content extraction module. \( \uparrow 2x \) is the operation to secondary enlargement by sub-pixel convolution. \( \oplus \) is the operation to concate. The content extractor and texture extractor are internally composed of residual blocks.

In this module, we mainly use sub-pixel convolution to achieve image super-resolution. With respect to the presence of many deconvolution region of 0, to enhance the convolution of pixel sub-pixel height and width dimensions of the transmission channel by the pixel dimensions. The feature map generated by the content extractor is \( F_c \in \mathbb{R}^{h\times w\times c} \), and the features are rearranged into \( kH\times kW\times kC \) after a pixel-shuffling operation. The mathematical formula for this operation is defined as follows:
\[
SC(F_c)_{h,y,z} = F_c(\lceil x/k \rceil,\lceil y/k \rceil,\mod(x,k) + C \cdot \mod(x,k) + z)
\]
where \( SC(F_c)_{h,y,z} \) denotes the output feature pixel on coordinate \( (x, y, z) \) after sub-pixel convolution and \( k \) is the magnification factor. \( \lceil \cdot \rceil \) represents rounding down operation. In the experiment, we set \( k=2 \).
to double the spatial scale.

In this module, we introduce the DnCNN network to remove the noise of the image, as shown in the Figure 1. The DnCNN network can gradually separate the image structure from the noise by combining residual learning and batch normalization, and has excellent denoising performance. DnCNN has three different types of layers with different colors. The convolutional layers in the Figure 1 all use 64 filters with a size of 3×3 to generate feature maps. DnCNN learns the mapping function through the residual to obtain a clean denoised image \( P_d \), \( P_d = P_i - v \). \( P_i \) is the input image with noise, and \( v \) is the noise obtained through residual learning. Formally, we define the noise loss function as the average square error between the expected residual image and the noise input estimated residual image:

\[
I_n = \frac{1}{2N} \sum_{r=1}^{N} \| \rho(y_r) - (y_r - x_r) \|^2_r
\]

where \( I_n \) represents the noise loss function and \((y_i, x_i)\) denote noisy-clean training image pairs.

2.3 Attention Feature Pyramid Module

As the depth of the convolutional neural network increases, the ability of the model to extract features also increases. However, since the effective features of small targets are less, as the number of layers increases, the target information is easily lost. Therefore, it is necessary to combine the texture information of the bottom layer with the semantic information extracted from the deep layer, and constructing FPN is the most common method to extract context information. As shown in the Figure 2, we constructed an AFP module that includes multiple up-sampling, multiple down-sampling, and skip connections. At the same time, we also introduce the attention mechanism to better extract effective features.

Attention module we selected CAM structure, as shown in Figure 3. We can use the formula 4 to express.

\[
Y_i = p_i + W \sum_{j=1}^{N} \text{softmax}(W_{ij} p_j) \otimes W_{ij} p_j
\]

The attention module establishes a paired dependency relationship between pixel pairs by sharing all query positions on the two feature maps. Finally, through a 1×1 convolution, the feature is extracted again and the residual connection is made with \( p_i \), and finally \( Y \) is output.

2.3 Attention Feature Pyramid Module

As the depth of the convolutional neural network increases, the ability of the model to extract features also increases. However, since the effective features of small targets are less, as the number of layers increases, the target information is easily lost. Therefore, it is necessary to combine the texture information of the bottom layer with the semantic information extracted from the deep layer, and constructing FPN is the most common method to extract context information. As shown in the Figure 2, we constructed an AFP module that includes multiple up-sampling, multiple down-sampling, and skip connections. At the same time, we also introduce the attention mechanism to better extract effective features.

Attention module we selected CAM structure, as shown in Figure 3. We can use the formula 4 to express.

\[
Y_i = p_i + W \sum_{j=1}^{N} \text{softmax}(W_{ij} p_j) \otimes W_{ij} p_j
\]

The attention module establishes a paired dependency relationship between pixel pairs by sharing all query positions on the two feature maps. Finally, through a 1×1 convolution, the feature is extracted again and the residual connection is made with \( p_i \), and finally \( Y \) is output.

2.3 Attention Feature Pyramid Module

As the depth of the convolutional neural network increases, the ability of the model to extract features also increases. However, since the effective features of small targets are less, as the number of layers increases, the target information is easily lost. Therefore, it is necessary to combine the texture information of the bottom layer with the semantic information extracted from the deep layer, and constructing FPN is the most common method to extract context information. As shown in the Figure 2, we constructed an AFP module that includes multiple up-sampling, multiple down-sampling, and skip connections. At the same time, we also introduce the attention mechanism to better extract effective features.

Attention module we selected CAM structure, as shown in Figure 3. We can use the formula 4 to express.

\[
Y_i = p_i + W \sum_{j=1}^{N} \text{softmax}(W_{ij} p_j) \otimes W_{ij} p_j
\]

The attention module establishes a paired dependency relationship between pixel pairs by sharing all query positions on the two feature maps. Finally, through a 1×1 convolution, the feature is extracted again and the residual connection is made with \( p_i \), and finally \( Y \) is output.
2.4 Loss Function

Based on yolo loss, the loss function of FENet is composed of noise loss, classification loss, confidence loss and positioning loss. As shown in formula 5.

\[
I_i = I_{noi} + I_{cls} + I_{cls} + I_{con}
\]

\[
= \lambda_{noi} \sum_{i=0}^{a} \sum_{j=0}^{b} Y_{ij}^{noi} \left[ (x_i - \hat{x}_j)^2 + (y_i - \hat{y}_j)^2 \right] + \lambda_{cls} \sum_{i=0}^{a} \sum_{j=0}^{b} Y_{ij}^{cls} \left[ (\sqrt{w_i} - \sqrt{\hat{w}_j})^2 + (\sqrt{h_i} - \sqrt{\hat{h}_j})^2 \right]
\]

\[
+ \lambda_{conf} \sum_{i=0}^{a} \sum_{j=0}^{b} Y_{ij}^{conf} \left( C_i - \hat{C}_j \right)^2 + \sum_{i=0}^{a} \sum_{j=0}^{b} \left( p_i(c) - \hat{p}_j(c) \right)^2 + I_i
\]

(5)

However, due to the iterative process, there is a significant gap between the loss provided by small objects and the loss provided by large objects. This seemingly fair loss function is unfair, and may lead to an unbalanced distribution of losses on different scales, and even small objects account for a small proportion of the total loss. Therefore, we introduced feedback-driven loss for positioning loss, as shown in the formula 6-8.

\[
I_i' = \lambda_{noi} \sum_{i=0}^{a} \sum_{j=0}^{b} Y_{ij}^{noi} f(w_i \times h_i) \left[ (x_i - \hat{x}_j)^2 + (y_i - \hat{y}_j)^2 \right] + \lambda_{cls} \sum_{i=0}^{a} \sum_{j=0}^{b} Y_{ij}^{cls} f(w_i \times h_i) \left[ (\sqrt{w_i} - \sqrt{\hat{w}_j})^2 + (\sqrt{h_i} - \sqrt{\hat{h}_j})^2 \right]
\]

\[
f(x) = \begin{cases} \frac{\text{Loss}(\text{total})}{\pi \text{Loss}(\text{small})} \cdot \arccos x, & x < \omega \\ 2 - x, & x \geq \omega \end{cases}
\]

(7)

\[
\frac{\text{Loss}(\text{total})}{\pi \text{Loss}(\text{small})} \cdot \arccos \omega = 2 - \omega
\]

(8)

Where \( \omega \) denotes the solution of the formula 8. Iterative learning of the loss ratio of small objects as the feedback coefficient can effectively improve the contribution of small object loss. In summary, the final loss function is as follows:

\[
I_{whole} = \lambda_{noi} \sum_{i=0}^{a} \sum_{j=0}^{b} Y_{ij}^{noi} f(w_i \times h_i) \left[ (x_i - \hat{x}_j)^2 + (y_i - \hat{y}_j)^2 \right] + \lambda_{cls} \sum_{i=0}^{a} \sum_{j=0}^{b} Y_{ij}^{cls} f(w_i \times h_i) \left[ (\sqrt{w_i} - \sqrt{\hat{w}_j})^2 + (\sqrt{h_i} - \sqrt{\hat{h}_j})^2 \right]
\]

\[
+ \lambda_{conf} \sum_{i=0}^{a} \sum_{j=0}^{b} Y_{ij}^{conf} \left( C_i - \hat{C}_j \right)^2 + \sum_{i=0}^{a} \sum_{j=0}^{b} \left( p_i(c) - \hat{p}_j(c) \right)^2 + \frac{1}{2} \sum_{i=0}^{a} \| y_i - y_i \|^2
\]

(9)

3. Results & Discussion

3.1 Datasets

The datasets used in the experiment are images of various objects extracted from the simulator, such as tanks, cars, planes and so on, with a total image size of 120K and a resolution of 1920*1080. There are 12 categories. The images are divided into three subsets: training set, verification set and test set with a ratio of about 4:2:1. According to the definition of COCO data set, pixels less than 32*32 are defined as small targets, pixels between 32 and 96 are defined as medium targets, and pixels greater than 96*96 are defined as large targets. The definition and statistics of the object are given in Table 1. Mean average precision (mAP), F1 score and other parameters were used to test the final performance of the model.

| Object Size       | Proportion |
|-------------------|------------|
| Large Object      | > (96×96)  | 42%         |
| Medium Object     | (32×32) ~ (96×96) | 38%         |
| Small Object      | < (32×32)  | 30%         |

3.2 Implementation Details

Our experiment is based on TITAN X GPU. The environment used is pytorch 1.7.1 and CUDA 11.0. The experiment uses a clustering algorithm to get the a priori frame of the data set, and uses yolo head for target classification and positioning. We choose ResNet as the backbone. The size of the input picture
is fixed at 608×608. Momentum is set to 0.9 and weight decay is set to 0.0005. We have trained for 50 epochs in all experiments. In order to speed up the convergence of the model, a linear warm-up learning strategy is used to adjust the learning rate during training. The initial learning rate is 1e-4. In addition, we also used the Kaiming method to initialize all the new layers, and used the random descent algorithm to optimize the loss function. We set the non-maximum threshold to 0.5, and before getting the final test result, we use NMS to filter to more redundant frames.

3.3 Ablation Study
As we all know, ablation research is an effective means to explore the performance of each module through the controlled variable method. Therefore, in order to prove that the proposed module is effective, we conducted ablation experiments in this part, and the experimental results are shown in the Table 2. The effect of using a certain component alone can be improved, but it is not as good as the overall effect of all components. At the same time, we can see that the RFE module is the most effective of the three components.

Table 2. Results of Ablation Studies

| RFE Module | AFP Module | Feedback-driven loss | mAP (for small objects) | mAP (for medium objects) | mAP (for large objects) |
|------------|------------|----------------------|--------------------------|--------------------------|------------------------|
| ✓          | ✓          |                      | 86.2%                    | 90.5%                    | 92.8%                  |
| ✓          | ✓          | ✓                    | 83.6%                    | 90.1%                    | 92.5%                  |
| ✓          | ✓          | ✓                    | 85.7%                    | 89.2%                    | 90.7%                  |
| ✓          | ✓          | ✓                    | 87.1%                    | 91.1%                    | 93.4%                  |

3.4 Comparison with State-of-the-Arts
We present the results of our model in the Table 3 and compare them with other latest research results. In the Table 3, our proposed method is superior to other methods not only on a small scale, but also on three scales. Compared with FAN, the accuracy of our network on large, medium and small objects can be improved by 0.9%, 0.5% and 0.6% respectively.

Table 3. Comparison Study Results with Different Detectors

| Method          | Small  | Medium | Large  |
|-----------------|--------|--------|--------|
|                 | mAP    | F1     | mAP    | F1     | mAP    | F1     |
| HR [12]         | 80.3   | 81.2   | 85.7   | 87.2   | 87.1   | 89.6   |
| SSH [13]        | 83.2   | 84.5   | 89.2   | 90.1   | 90.2   | 92.4   |
| SFD [14]        | 82.7   | 84.1   | 90.6   | 91.3   | 91.2   | 92.1   |
| Perceputal GAN [7] | 84.0   | 86.3   | 90.1   | 92.4   | 91.4   | 92.5   |
| SRFACE [15]     | 87.1   | 88.2   | 90.1   | 92.6   | 91.6   | 93.2   |
| FAN [16]        | 88.2   | 89.2   | 90.6   | 92.8   | 95.2   | 96.1   |
| FENet (ours)    | 89.1   | 90.3   | 91.1   | 93.7   | 95.8   | 97.4   |

4 Conclusions
In this paper, we propose a new model FENet, which can effectively solve the problem of small object detection. FENet mainly includes two modules, RFE and AFP module. The RFE module can not only use the underlying spatial information of the small target, but also the appearance texture information. At the same time, the residuals are used to extract the noise of the image, which can effectively improve the image quality. The AFP module can prevent the loss of information of small deep targets and extract useful context information. In addition, we also added a loss function to solve the unfairness of multi-scale loss to model training in the iterative process. Many experiments have proved the effectiveness of our method. Since this article does not consider the imbalance between the positive sample and the negative sample of the small target, we can continue to improve in the future. In addition, how to reduce the amount of model parameters is also one of the subsequent problems that need to be solved.

References
[1] Lin, Tsung-Yi, et al. "Feature pyramid networks for object detection." Proceedings of the IEEE
conference on computer vision and pattern recognition. 2017.

[2] Liu, Y., Yang, F., & Hu, P. (2020). Small-object detection in UAV-captured images via multi-branch parallel feature pyramid networks. IEEE Access, 8, 145740-145750.

[3] Zheng, Q., & Chen, Y. (2021). Feature pyramid of bi-directional stepped concatenation for small object detection. Multimedia Tools and Applications, 80(13), 20283-20305.

[4] Li, H., Lin, K., Bai, J., Li, A., & Yu, J. (2019). Small object detection algorithm based on feature pyramid-enhanced fusion SSD. Complexity, 2019.

[5] Zhang, W., Liu, X., Xu, L., Li, Z., Zhao, H., & Zhou, J. (2019). Feature Fusion Detection Network for Multi-scale Object Detection. In Cyberspace Data and Intelligence, and Cyber-Living, Syndrome, and Health (pp. 403-412). Springer, Singapore.

[6] Zhai, M., Liu, J., Zhang, W., Liu, C., Li, W., & Cao, Y. (2019, August). Multi-scale feature fusion single shot object detector based on DenseNet. In International Conference on Intelligent Robotics and Applications (pp. 450-460). Springer, Cham.

[7] Li, J., Liang, X., Wei, Y., Xu, T., Feng, J., & Yan, S. (2017). Perceptual generative adversarial networks for small object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1222-1230).

[8] Zhou, P., Ni, B., Geng, C., Hu, J., & Xu, Y. (2018). Scale-transferable object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 528-537).

[9] Zhang, Z., Wang, Z., Lin, Z., & Qi, H. (2019). Image super-resolution by neural texture transfer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 7982-7991).

[10] Zheng, H., Ji, M., Wang, H., Liu, Y., & Fang, L. (2018). Crossnet: An end-to-end reference-based super resolution network using cross-scale warping. In Proceedings of the European conference on computer vision (ECCV) (pp. 88-104).

[11] Liu, G., Han, J., & Rong, W. (2021). Feedback-driven loss function for small object detection. Image and Vision Computing, 111, 104197.

[12] Hu, P., & Ramanan, D. (2017). Finding tiny faces. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 951-959).

[13] Najibi, M., Samangouei, P., Chellappa, R., & Davis, L. S. (2017). Ssh: Single stage headless face detector. In Proceedings of the IEEE international conference on computer vision (pp. 4875-4884).

[14] Zhang, S., Zhu, X., Lei, Z., Shi, H., Wang, X., & Li, S. Z. (2017). S3fd: Single shot scale-invariant face detector. In Proceedings of the IEEE international conference on computer vision (pp. 192-201).

[15] Bai, Y., Zhang, Y., Ding, M., & Ghanem, B. (2018). Finding tiny faces in the wild with generative adversarial network. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 21-30).

[16] Wang, J., Yuan, Y., & Yu, G. (2017). Face attention network: An effective face detector for the occluded faces. arXiv preprint arXiv:1711.07246.