Traffic sign recognition algorithm based on feature pyramid attention

Pan Su¹, Wenju Li*, Liye Sha², Yanjiao Shi¹ and Tianzhen Dong¹

¹ College of Computer Science and Information Engineering, Shanghai University of Technology, Shanghai, 201418, China
² Shanghai Precision Dosing & Weighing System Co., Ltd. Shanghai, 201108, China
*Corresponding author’s e-mail: wjli@sit.edu.cn

Abstract. The traditional traffic sign recognition algorithm is easily affected by factors such as complex background and illumination in a real scene, which can easily lead to missed detection and misdetection, in the stage of traffic sign detection, the Feature Pyramid Spatial Attention (FPSA) module is proposed. In the process of generating the feature pyramid, the FPSA module uses high-level features as the attention information of low-level features. In order to solve the problem of sample imbalance in the dataset, the dataset is screened, and the category loss weight based on the effective number of samples is introduced when calculating the loss. Experiments conducted on 45 traffic sign categories on the TT-100K dataset prove that FPSA and the proposed category loss weights are beneficial to improve the performance of small object detection and recognition in complex backgrounds, and are obtained on the Faster R-CNN model 78.9% mAP accuracy.

1. Introduction
The detection and recognition of traffic signs is a typical application scenario of computer vision, which is mainly used in fields such as self-driving cars or car-assisted driving. Since the design of traffic signs should take into account the effective reminder of the road conditions, warnings or instructions ahead of the car driver under long-distance or foggy conditions, when designing traffic sign detection and recognition algorithms, colour and shape are mainly used as prior information [1].

With the rise of deep learning technology, neural networks gradually occupy a dominant position in the field of computer vision. In 2017, He Kaiming et al. [2] proposed a new object detection model Faster R-CNN, and then this model was quickly extended to various research fields of object detection. In the same year, Lin TY et al. [3] proposed Feature Pyramid Network (FPN), for object detection tasks. This model can extract multiple feature layers of different sizes from the backbone network to form a feature pyramid, which is especially suitable for small object detection tasks. In 2018, Li H et al. [4] used Feature Pyramid Attention (FPA) on the last layer of features extracted from the backbone network in the semantic segmentation task, and then used Global Attention Upsampling (GAU) Combining modules and low-level features, the entire network has a U-shaped structure. In 2020, Mei Y et al. [5] proposed a new method combining attention and feature pyramid in the image restoration task, using low-level features as the attention information of high-level features.

In 2019, Zhou S et al. [6] proposed a small traffic sign recognition model based on convolutional neural network based on the PVANet model, and achieved good performance. In 2020, Wan J et al. [7] proposed a small object recognition algorithm for traffic signs based on YOLOv3 [8]. In 2021, Guo F
et al. [9] proposed the YOLOv3-A algorithm by adding channel attention and spatial attention based on semantic segmentation to solve the problem of misdetection and missed detection for small objects and complex background conditions, which increases the detection accuracy and recall rate on the TT100K dataset by 1.9% and 2.8%, respectively. In the same year, Chu J et al. [10] proposed a non-anchor frame traffic sign recognition algorithm based on encoder-decoder structure, which further improved the recognition accuracy of traffic signs. In 2020, the proposal of YOLOv4 [11] makes the object detection model achieve the best detection performance at present, which is expected to further improve the performance of traffic sign recognition.

2. Network structure

2.1. Spatial attention module

There are two main types of self-attention mechanisms in the field of computer vision: spatial attention and channel attention. Spatial attention mainly by convolution to calculate the attention weight of the feature map in the two dimensions of height and width, as shown in figure 1, firstly calculate the maximum and average value of the input feature map in the channel dimension, And then stitch the two obtained features together according to the channel dimensions to obtain a feature map with a channel number of 2, then use the 7x7 convolution and Sigmoid activation function to obtain the attention weight, and finally get the output by multiply the attention weight and the input feature. The spatial attention module can be expressed by formula (1):

\[
M_s(F) = \partial\left(f^{7 \times 7}\left([\text{Avgpool}(F); \text{MaxPool}(F)]\right)\right)
\]

Among them, the \( F \) represents the input feature map, \( M_s(F) \) represents the output single-channel attention weight, \( f^{7 \times 7} \) represents the convolutional layer of the kernel \( 7 \times 7 \), \( \partial \) represents the sigmoid activation function, \([;]\) means that multiple feature maps in square brackets are spliced in the channel dimension.

Channel attention mainly calculates attention weight in the channel dimension of the feature map. Its structure is shown in figure 2. It was proposed by Jie H et al. [12] in 2017. First, it uses the global average pooling aggregation feature and get a feature map with a size of 1x1, and then use two fully connected layers to squeeze and expand the channel dimensions respectively, and use sigmoid activation to obtain the attention weight, and finally multiply the input feature map and the attention weight to get output.

The task of traffic sign recognition is similar to object detection and focuses more on the area of interest in the picture. Therefore, this paper combines the spatial attention module with the feature pyramid and proposes the FPSA module, which uses the high-level semantic features as the attention information of the low-level features, instruct the network to focus on the area of interest.
2.2. Feature Pyramid Spatial Attention module
In the field of computer vision, there have been some studies that combine the attention mechanism with the feature pyramid. Li H et al. [4] proposed the FPA module in the semantic segmentation task, using multiple convolution kernels of different sizes to convolve the last layer of features extracted by the backbone network to form a feature pyramid. Because the feature map is small, the amount of calculation increases can be ignored. Then, after 1×1 convolution processing, the original features extracted by CNN are multiplied pixel by pixel by the pyramid attention information. In addition, the global pooling branch structure is used to connect the output features to further improve the performance of the FPA module. Mei Y et al. [5] proposed to use the low-level features extracted by the backbone network as the attention information of the high-level features in the image restoration task, and proposed PANet.

In order to deal with the influence of complex background and illumination on the performance of traffic sign detection, we propose to combine the spatial attention module (Figure 1) with the feature pyramid to form the FPSA module, as shown in Figure 3. In the FPSA module, spatial attention no longer exists as self-attention, but as a cross-attention between different hierarchical features. At the same time, contrary to Mei Y et al. [5], the FPSA module using high-level semantic information to guide low-level information.

For the high-level features, first use the nearest neighbour interpolation method for upsampling, then use the spatial attention module to calculate the attention weight, and finally multiply the attention map and the low-level features to get the output.

![Figure 3. Feature pyramid spatial attention](image)

3. Preprocessing of TT-100K dataset
TT-100K dataset was used in the experiment, which is launched by the Tsinghua-Tencent joint laboratory. It contains more than 100,000 pictures and 221 traffic sign categories. Among them, the traffic sign subset contains 16,811 pictures and 182 traffic sign categories. Small object traffic signs with pixels in the range of (0, 32) account for about 42% of the total, and traffic signs with pixels in the range of (32, 96) traffic signs account for approximately 49% of the total, while large traffic signs within the range of (96, 400) account for only 9% of the total, as shown in Figure 5. The number of instances of each traffic sign category ranges from 1 to 2997. As shown in Figure 4, the sample imbalance is very serious. Currently, there are two main methods for sample imbalance. One is to use category loss weight method, and the other is to use re-sampling strategy based on category perception.

In order to alleviate the problem of sample imbalance, we screened the TT-100K dataset and selected the top 45 traffic sign categories based on the number of samples. Figure 6 shows the 45 categories with the highest number of instances after screening. The number of instances ranges from 69 to 2997, and the phenomenon of sample imbalance still exists.
In order to further solve the problem of samples imbalance, this paper uses a weighted cross-entropy loss function, which adds a weight to a category according to the number of instances in this category. We adopt the concept of effective sample size according to the method proposed by Cui Y et al. [13], and assumes that the effective sample size of each category in the training set can be calculated by formula (2):

$$E_n = \frac{1 - \beta^n}{1 - \beta}$$  \hspace{1cm} (2)

Among them, $\beta = 0.999$, $n$ represents the number of samples in each category. Therefore, the category loss weight can be calculated using formula (3):

$$w = \frac{1}{E_n} = \frac{1 - \beta}{1 - \beta^n}$$  \hspace{1cm} (3)

The weight of each category increases as the number of instances decreases. Through analysis, the category weight is mostly between 0.0001 and 0.02. Considering that too small loss will cause the
network to learn too slowly, the category weight is multiplied by 100. The final category weight can be calculated using formula (4):

$$\frac{100}{E_n} = 100 \times \frac{1 - \beta}{1 - \beta^*}$$  \hspace{1cm} (4)

4. Experiment and analysis

The experiment uses the Faster R-CNN model based on FPN [3] and RetinaNet [14] as the baseline, and conducts ablation experiments for category weights and spatial attention respectively to prove the effectiveness of the proposed method. The input picture size of Faster R-CNN is (800, 1000), and the input picture size of RetinaNet is (800, 1333). RetinaNet is a one-stage object detection model proposed by He K in 2017. Here we also added a feature pyramid to extract multi-scale features to facilitate the detection of small objects.

For the proposed category loss weight and FPSA module, the parameter settings in the model training process are shown in table 1.

| Parameters       | Value     |
|------------------|-----------|
| epochs           | 50        |
| batch size       | 8         |
| learning rate    | 0.005     |
| weight decay     | 0.0005    |
| optimizer        | SGD       |
| lr_scheduler     | Cosine    |

Among them, Cosine gradually decreasing the learning rate from 0.005 to 0.00005 during the training process. The experimental results are shown in table 2. Note that, "FPN" represents the Faster R-CNN model based on FPN, "*" represents the model using category loss weights, and "**" represents the model using both FPSA and category loss weights, "Test~" represents the inference speed during testing, and "Train~" represents the speed during training.

| Model            | AP   | mAP   | Small | Medium | Large | Test~ | Train~ | Small | Medium | Large |
|------------------|------|-------|-------|--------|-------|-------|--------|-------|--------|-------|
| FPN[2,3]         | 0.375| 0.641 | 0.199 | 0.470  | 0.472 | 20fps | 9fps   | 0.406 | 0.637  | 0.630 |
| RetinaNet[16]    | 0.349| 0.494 | 0.167 | 0.431  | 0.645 | 19fps | 9fps   | 0.373 | 0.716  | 0.876 |
| FPN*             | 0.521| 0.745 | 0.308 | 0.620  | 0.643 | 20fps | 9fps   | 0.469 | 0.744  | 0.743 |
| RetinaNet*       | 0.411| 0.565 | 0.173 | 0.502  | 0.742 | 19fps | 9fps   | 0.412 | 0.747  | 0.857 |
| FPN**            | 0.570| 0.789 | 0.348 | 0.667  | 0.714 | 20fps | 9fps   | 0.488 | 0.767  | 0.785 |
| RetinaNet**      | 0.434| 0.590 | 0.188 | 0.529  | 0.761 | 19fps | 9fps   | 0.413 | 0.741  | 0.861 |

According to table 2, it can be seen that due to the serious long tail effect of the TT-100K data set, the FPN-based Faster R-CNN model and RetinaNet finally obtained only 64.1% and 49.4% mAP
accuracy, respectively. The recall rate is only 40.6% and 37.3%, while the small objects in the TT-100K data set account for more than 40%, while the large objects account for less than 10%.

After adding category loss weights during training, the mAP accuracy of FPN* and RetinaNet* was 74.5% and 56.5%, respectively, which was 10.4% and 7.1% higher than the FPN and RetinaNet models, respectively. The impact of the long-tail data set on the model has been improved to a certain extent.

Furthermore, after adding spatial attention to the feature pyramid of the FPN* model and the RetinaNet* model, the accuracy of mAP reached 78.9% and 59%, and the recall rate of small objects reached 48.8% and 41.3%, respectively. This proves that spatial attention can strengthen the model's ability to extract object features to a certain extent, and help the model achieve more accurate predictions.

5. Conclusion
Aiming at the problems of small objects, colour and illumination changes in traffic sign recognition in real scenes that lead to low model recognition performance, the FPSA module is proposed, which combines the feature pyramid network and the spatial attention module. In the FPSA module, the high-level semantic features are transformed into the attention information of the low-level features through the spatial attention module, and finally a feature map is generated. At the same time, in order to deal with the problem of sample imbalance in the TT-100K data set, it is proposed to add category loss weights according to the number of samples in each category in the training set when training the model.

Experiments show that the addition of category weights and spatial attention modules can significantly improve the recognition accuracy of the model, and the impact on the training and testing speed is almost negligible.

Acknowledgments
This work had supported by the National Natural Science Foundation of China Youth Science Fund Project (No. 61806126)

References
[1] Qu, Z., Shao, Y., Deng, T., Zhu, J., & Song, X. (2019). Traffic sign detection and recognition under complicated lighting conditions. Laser & Optoelectronics Progress, 56(23), 231009. Podani, J. (1994) Multivariate Data Analysis in Ecology and Systematics. SPB Publishing, The Hague.
[2] Ren, S., He, K., Girshick, R., & Sun, J. (2017). Faster r-cnn: towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis & Machine Intelligence, 39(6), 1137-1149.
[3] Lin, T. Y., Dollar, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature Pyramid Networks for Object Detection. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE Computer Society.
[4] Li, H., Xiong, P., An, J., & Wang, L. (2018). Pyramid attention network for semantic segmentation.
[5] Mei, Y., Fan, Y., Zhang, Y., Yu, J., & Shi H. (2020). Pyramid attention networks for image restoration.
[6] Zhou, S., Zhi, X., Liu, D., Ning, H., Jiang, L., & Shi, F. (2019). A convolutional neural network-based method for small traffic sign detection. Journal of Tongji University (Natural Science).
[7] Wan, J., Ding, W., Zhu, H., Xia, M., Huang, Z., & Tian, L., et al. (2020). An efficient small traffic sign detection method based on yolov3. Journal of Signal Processing Systems, 1-13.
[8] Redmon J., & Farhadi A. (2018). Yolov3: an incremental improvement. arXiv e-prints.
[9] Guo, F., Zhang, Y., Tang, J & Li, W. (2021). YOLOv3-A: Traffic sign detection network based on attention mechanism. Journal of Communications (01), 87-99. doi:.
[10] Chu, J, Huang, H & Lu, W. (2021). An anchorless frame traffic sign recognition algorithm based on attention model. Progress in Laser and Optoelectronics (),. doi:10.3788/top58.1010008.
[11] Bochkovskiy, A., Wang, C. Y., & Liao, H. (2020). Yolov4: optimal speed and accuracy of object detection.
[12] Jie, H., Li, S., Gang, S., & Albanie, S. (2017). Squeeze-and-excitation networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, PP(99).

[13] Cui, Y., Jia, M., Lin, T. Y., et al. (2019). Class-Balanced Loss Based on Effective Number of Samples[J]. https://arxiv.org/pdf/1901.05555.pdf

[14] Lin, T. Y., Goyal, P., Girshick, R., He, K., & P, Dollár. (2017). Focal loss for dense object detection. IEEE Transactions on Pattern Analysis & Machine Intelligence, PP(99), 2999-3007.