Hindawi
Security and Communication Networks
Volume 2022, Article ID 6805526, 9 pages
https://doi.org/10.1155/2022/6805526

Research Article

A Novel One-Shot Object Detection via Multifeature Auxiliary Information

Yu Song, Min Li, Weidong Du, Yao Gou, Zhaoqing Wu, and Yujie He

Xi'an Institute of High Technology, Xi'an, Shaanxi 710025, China

Correspondence should be addressed to Yu Song; huogongdaoren@126.com

Received 14 December 2021; Accepted 17 January 2022; Published 21 February 2022

Academic Editor: Thippa Reddy G

With the advantage of using only a limited number of samples, few-shot learning has been developed rapidly in recent years. It is mostly applied in the object classification or detection of a small number of samples which is typically less than ten. However, there is not much research related to few-shot detection, especially one-shot detection. In this paper, the multifeature information-assisted one-shot detection method is proposed to improve the accuracy of one-shot object detection. Specifically, two auxiliary modules are applied to the detection algorithm: Semantic Feature Module (SFM) and Detail Feature Module (DFM), which, respectively, extract semantic feature information and detailed feature information of samples in the support set. Then these two kinds of information are then calculated with the feature image extracted from the query image to obtain the corresponding auxiliary information that is used to complete one-shot detection. Thanks to the two auxiliary modules, which can retain more semantic and detailed information of samples in the support set, the proposed method can enhance the utilization rate of sample feature information and improve object detection accuracy by 2.97% compared to the benchmark method.

1. Introduction

Deep neural networks have been widely used in computer vision, such as posture recognition [1] and plant disease recognition [2], and object detection is the research hotspot in this field. Generally speaking, object detection algorithms can be divided into two categories according to different training strategies: one-stage object detection and two-stage object detection. The popular algorithms are the YOLO algorithms [3–5] and R-CNN algorithms [6–8], which dramatically improve the object detection effect and enhance detection efficiency. However, those algorithms rely on object annotation information, which cannot be easily obtained. Therefore, researchers gradually focus on few-shot detection.

Few-shot detection is derived from few-shot learning, a particular case of meta-learning. At present, the learning methods can be roughly broken down into four categories: measurement learning-based learning, meta-learning-based learning, data enhancement-based learning, and multimodal approaches-based learning, among which the meta-learning-based learning method is the most popular. In the meta-training stage, by compositing several samples from different classes to take different meta-task, the model can learn the differences between examples of various categories and the similarities between samples of the same type. While in the meta-testing stage, the recognition task can be completed without retraining or only with a small amount of rapid training for a new category. However, few-shot learning is mainly used in few-shot classification rather than few-shot detection.

Few-shot object detection is used to complete detection for objects with very few samples in the dataset. The existing few-shot detection methods fall into three categories: fine-tuning, model structure-based learning, and metric-based learning. The few-shot detection training strategy generally contains two stages: meta-training stage and fine-tuning stage. In the meta-training, N categories were randomly chosen from the training set, each containing K samples to form the support set of the model, namely, a meta-task. Next, a small number of object samples were selected to fine-tune the model. The purpose was to train the model to detect

N classes of objects from $N \times K$ data and then generalize the knowledge to adapt to new classes. This task is called N-way K-shot. In few-shot learning, the K value is usually smaller than ten; when $K = 1$, it is called one-shot learning.

Existing metric-based few-shot detection mainly divides the dataset into the support set and the query set. It selects several image samples from the two sets to form the minimum training unit task (meta-task) and then trains the model through specific strategies. The detection algorithm first obtains the corresponding features of the images in the two sets, then measures the distance between the two features, and judges the object category according to the distance. According to the label's location information, the regression operation is performed to complete the object positioning. We note that multiple features with different scales will be generated when the convolutional neural network extracts features of the support images. However, as the current algorithm only conducts simple distance measurement, the utilization rate of object feature information is extremely low.

To solve this problem, this paper proposed a novel one-shot detection method on the basis of metric-based learning. The main contributions of this paper are as follows:

(1) This novel method integrated the Semantic Feature Module (SFM) and the Detail Feature Module (DFM), which generated features about the support images of two sizes (7×7 , 3×3). These features were then operated with query images' feature and generated the corresponding multifeature auxiliary information (MFAI) of Semantic Feature Auxiliary Information and Detailed Feature Auxiliary Information.

(2) Experimental results showed that both the SFM and the DFM could increase the accuracy of one-shot detection. A combination of the two modules could even increase the detection accuracy by 2.97% compared to the original algorithm.

2. Related Works

In recent years, research on few-shot learning has attracted a lot of interest, which can effectively solve the classification and detection task using only a few labeled samples. The recent related works of few-shot classification, few-shot object detection, and one-shot object detection are listed in Table 1.

In the general one-shot object detection method, the weight extracted from the image is mainly used to measure the object feature distance. However, the semantic feature information and detailed feature information of the support set object are not fully utilized. This paper introduced the SFM and the Object Detail Module based on one-shot object detection, inspired by literature [18]. By using more object features to train the deep neural network, the detection effect of our model was better.

3. Method

The semantic information and detailed information need to be generated separately. In theory, the support images can obtain feature images of different sizes through different modules. In our method, the 7×7 feature can retain more object details, while the 3×3 feature contains more semantic information about the object. The 7×7 feature and feature of the query image were used for dot product operation. The 3×3 feature is convolved with the feature of the query image. The corresponding MFAI of Semantic Feature Information and Detail Feature Information was generated through the above two operations.

3.1. Training Strategy. As mentioned above, in the training stage, assume that the dataset is D and divided into D_{base} and D_{novel}. D_{base} represents an object image dataset which contains a large number of annotated images, of which category is C_{base}; and D_{novel} represents a dataset containing a few of samples with annotations with category as C_{novel}. With $D_{\text{base}} \cap D_{\text{novel}} = \varnothing$, $C_{\text{base}} \cap C_{\text{novel}} = \varnothing$, so the ultimate goal of one-shot detection is to classify and locate the object of query image in D_{novel}. Similar to the literature [16], the whole training process is divided into two steps. Firstly, the data in C_{base} are used to train the model, so the model can learn the meta-features; then this trained model can have a good detection effect on the object of the base class. Lastly, the $D_{\text{base}} \cup D_{\text{novel}}$ dataset is utilized for model training and fine-tuning to adapt to the new category, then generalizing the knowledge learned in the first step to the new object category.

Based on few-shot learning, we innovatively utilized the semantic information and detailed information of the object to complete the one-shot detection. Specifically, we used the smallest training unit $T = \{(S_i, x_i)\}_{i=1}^{|T|}$ in the training stage. The data in T are all from the randomly sampled support set in D_{base}, S_i, and the query image, x_i. i represents the i – th task. S_i contains N categories, with K samples in each class. Through multitask training, we obtained an object model with a detection base class. Next, the fine-tuned model was continuously adapted with the $D_{\text{base}} \cup D_{\text{novel}}$ dataset, and the detection model $f_\theta(x|S)$ was fine-tuned to fit the new category, in which θ is the parameter that the model needs to learn. The final one-shot detection could be completed by f_θ tuned well.

3.2. Multifeature Information-Assisted Detection Method. Firstly, the feature extraction module was used to extract the 7×7 feature image of query image. Then the support images were input into the DFM, SFM, and Weighted Module (WM), respectively, to get the corresponding 7×7, 3×3, 1×1 feature maps, while the channels were consistent. Next, we applied dot product 7×7 feature of support images with 7×7 feature generated by query image, and finally generated 7×7 feature—Detail Feature Auxiliary Information (DFAI), then the 7×7 feature generated by query image was convolved with the 3×3 feature of support images as a filter, and finally generated 7×7 feature—Semantic Feature Auxiliary Information (SFAI). The average operation was carried out on the two kinds of auxiliary information. The next step was to convolve the processed averaged feature (7×7 feature) with the weight information (1×1 feature) generated by the support images; thus, the 7×7 MFAI to be detected was

Table 1: Recent related works.

Category	Ref.	Methods
Few-shot classification	[9]	The Mahalanobis distance in a state-of-the-art few-shot learning approach (CNAPS [10]) is adopted to improve performance
	[11]	Presents a novel network to learn and preserve the feature manifold's topology formed by different classes
	[12]	Proposes the similarity ratio as an indicator for the generalization performance of a few-shot model
	[13]	Takes advantage of the earth mover's distance (EMD) to measure the distance between dense image representations which determines image relevance
	[14]	Merges three learning methods: visual feature learning, knowledge inferring, and classifier learning, into a unified framework
	[15]	Introduces the oPen sEt mEta LEaRning (PEELER), which randomly selects a set of novel classes, maximizes the posterior entropy over every sample, and utilizes the Mahalanobis distance as a new metric
Few-shot object detection	[16]	Improves the CentreNet detector for the few-shot learning and a class-specific code generator is modeled by meta-learning
	[17]	Uses Attention-RPN, multirelation detector and contrastive training strategy to detect novel objects
	[18]	Proposes the model that uses labeled base categories and quickly improves to new categories, utilizing a meta-feature learner and a new upgraded module
One-shot object detection	[19]	Develops coattention and coexcitation framework (CoAE) that contributes to several technical aspects
	[20]	Develops a new algorithm to guide the parameter posterior towards its true distribution to remedy the posterior fading problem that compromises the effectiveness of shared weights

generated. Finally, the detection feature map was put into the detection network to generate the classification and location information. The process was described as shown in Figure 1.

To explain the function of each module in detail, we elaborated the function into two steps. The first step was to combine the Semantic Feature Auxiliary Module (SFAM) with the WM to output the feature to be detected; then combine the dot product information auxiliary module with the WM to output the feature to be detected. The former is illustrated in Figure 2. Firstly, query image and support images were input into both the SFM and the feature extraction module to extract their respective features. Then the feature extraction module outputs the $1024 \times 7 \times 7$ feature I_Q, the SFM outputs the $N \times 1024 \times 3 \times 3$ feature S_C, and the WM outputs the $N \times 1024 \times 1 \times 1$ feature S_W. After the convolution operation of S_C and I_Q, F_C was generated, as shown in

$$F_C = S_C * I_Q. \tag{1}$$

Then, S_W and F_C were convolved to get the $N \times 1024 \times 7 \times 7$ feature Y_C to be detected, as shown in

$$Y_C = S_W * F_C, \tag{2}$$

$$Y_C = S_W * \left(S_C * I_Q\right). \tag{3}$$

The DFM is illustrated in Figure 3. Similarly, the query image and support images were put into the feature extraction module and the DFM to extract their respective features. So, the feature extraction module outputs the $1024 \times 7 \times 7$ feature I_Q, the DFM outputs the $N \times 1024 \times 7 \times 7$ feature S_D, and the WM outputs the $N \times 1024 \times 1 \times 1$ feature S_W. After the dot production operation of S_D and I_Q, F_D was generated, as shown in

$$F_D = S_D \otimes I_Q. \tag{4}$$

Then, S_W and F_D were convolved to obtain the $N \times 1024 \times 7 \times 7$ feature Y_D to be detected, as shown in

$$Y_D = S_W * F_D, \tag{5}$$

$$Y_D = S_W * \left(S_D \otimes I_Q\right). \tag{6}$$

3.3. Loss Function. To handle the various objects which need to be detected in one-shot detection, the model in this paper adopted a softmax layer [18]. The predicted score on classification for the i-th class was represented by $\hat{c} = e^{c_i} / \sum_{j=1}^{N} e^{c_j}$. To get better model convergence, the cross-entropy loss over the calibrated scores $\hat{c}$ was adopted, as shown in

$$L_c = -\sum_{i=1}^{N} O(\cdot, i) \log\left(\hat{c}_i\right), \tag{7}$$

where $O(\cdot, i)$ is an indicator function. When the current anchor box fits into class i, its value is 1. Otherwise, the value is 0. For the bounding box regression calculation and object determination method, we followed the same detection way as YOLOV3. After anchors with different aspect ratios were preset, coordinate classification of anchors would be processed through calculation and finally predicted the object. In this paper, corresponding loss functions were adopted, such as the Mean Squared Error (MSE) loss and the Binary Cross-Entropy (BCE) loss. The loss function of multifeature information-assisted one-shot detection proposed in our model is shown in

$$L_{\text{det}} = L_c + L_{bbx} + L_{\text{obj}}, \tag{8}$$

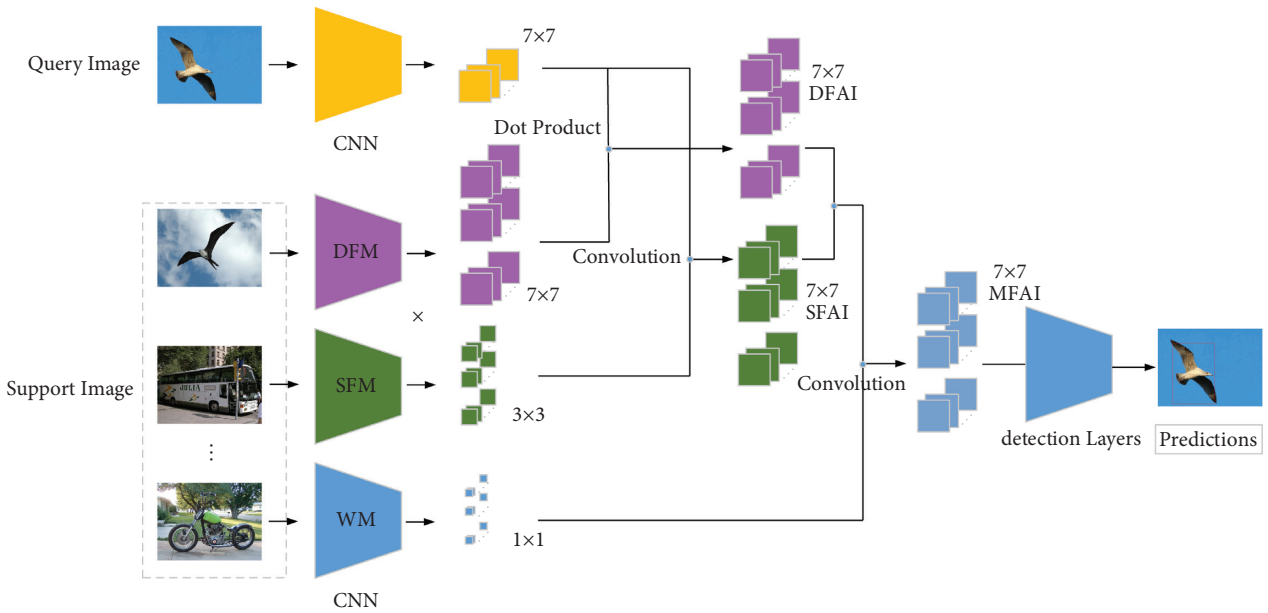

Figure 1: The structure of the one-shot detection assisted by multiple feature information.

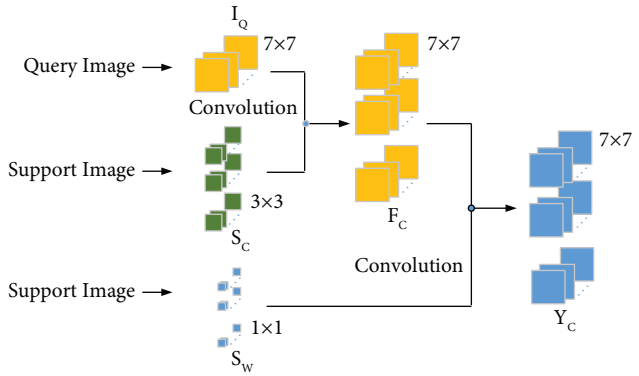

Figure 2: Schema of the semantic feature auxiliary module.

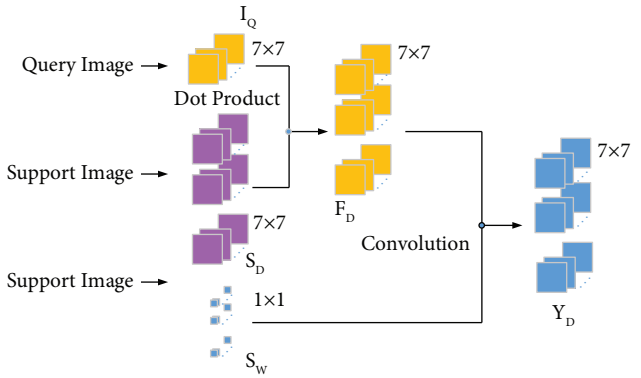

Figure 3: Schema of the detail feature auxiliary module.

where L_{bbox} is expressed as

$$L_{bbox} = \lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^{B} I_{ij}^{obj} (2 - w_i \times h_i) \left[(a_i - \hat{a}_i)^2 + (b_i - \hat{b}_i)^2 + (w_i - \hat{w}_i)^2 + (h_i - \hat{h}_i)^2 \right], \tag{9}$$

where S means the grid set in YOLO and S^2 represents 13×13, 26×26, and 52×52. B stands for prediction box. I_{ij}^{obj} means the box at i, j, which is 1 if it is an object; otherwise, it is 0. a_i, b_i, w_i, h_i represent the central point coordinates and the width and height of the object, respectively. $\hat{a}_i$, $\hat{b}_i$, $\hat{w}_i$, $\hat{h}_i$ represent the predicted values of the center point coordinates, width, and height, respectively.

L_{obj} can be expressed as in

$$L_{obj} = \lambda_{noobj} \sum_{i=0}^{S^2} \sum_{j=0}^{B} I_{ij}^{noobj} (c_i - \hat{c}_i)^2 + \lambda_{obj} \sum_{i=0}^{S^2} \sum_{j=0}^{B} I_{ij}^{obj} (c_i - \hat{c}_i)^2, \tag{10}$$

where I_{ij}^{noobj} means the box at i, j which is 1 if it is not an object; otherwise, it is 0. $\hat{c}_i$ is the prediction confidence on the object for class i. $\hat{c}_i$ is set to 1 if the object is the true value of a certain class; otherwise, it is 0.

4. Experiment

4.1. Experimental Environment. The running environment of the algorithm verification experiment is shown in Table 2.

4.2. Dataset. The datasets we adopted are VOC 2007 [21] and VOC 2012 [22], which are the widely used object detection benchmarks. Out of the 20 categories, we selected samples of five categories as novel datasets and the remaining 15 categories as the base datasets. The model training was divided into two stages: base training stage and one-shot fine-tuning stage. In the base training stage, the images of base samples were used to train the normal model in the supervised mode. And in the one-shot fine-tuning

Table 2: Runtime environment.

Item	Parameter
GPU	Titan Xp (12G video memory) CUDA 10.0 cuDNN 7.0
Operating system	Ubuntu 16.04
Python version	3.6
Iterations	60
Learning rate	0.001
Momentum	0.9
Momentum attenuation coefficient	0.00004

stage, the images of novel ones were used to ensure that each class of objects only had one annotated bounding box.

4.3. Experiment and Analysis. A large number of experiments have been done on these datasets. To illustrate the effectiveness of the different methods, several representative datasets were selected. In the test phase, we used five unseen categories of data in training: bird, bus, cow, motorbike, and sofa. Due to the space limitation, we only present the results data of bird and bus categories in Figures 4 and 5.

As can be seen from the above Figure 4, there are a total of 12 images arranged in four rows, with three object images in each row. All the object images used the same kind of detection algorithm. The first line to the fourth line, respectively, shows the detection results of the original algorithm [18], the detailed feature information auxiliary algorithm, the semantic feature information auxiliary algorithm, and the multifeature information auxiliary algorithm proposed in this paper. Different detection results of the same object are presented in each column for those four algorithms. For column (a), the object is conspicuous, so all four algorithms can correctly identify it. As the background of the object in column (b) is relatively complex, the original algorithm and the auxiliary algorithm of detailed feature information cannot detect the object well. In contrast, the algorithms in the third and fourth lines that integrate semantic information can detect the object more accurately. Since the objects in column (c) belong to multiobject detection in a complex background; the detection effect of the second row is not ideal. The algorithms in the first and third rows can completely detect conspicuous objects. While the fourth line algorithm can detect multiple objects, the second object detection is not complete because there is little difference between the background color and the object color.

Similarly, there are also 12 images in Figure 5, which are arranged in the same manner. The object image of each row uses the same kind of detection algorithm. The first row to the fourth row, respectively, represents the detection results of different buses by the four algorithms. For column (a), the object is prominent, so all the four algorithms can correctly identify it. In column (b), the background of images is relatively complex. Although the original algorithm and the detailed feature assist algorithm can locate the object more accurately, there is still a misjudgment in the classification of extracted features. As a result, the bus is misclassified as a train. The algorithms in the third and fourth rows incorporate semantic information to detect objects more accurately. Since the object in column (c) is similar to a train, the four algorithms misjudge the result. That is why the accuracy of bus detection results is low.

4.4. Ablation Experiments. The accuracy improvement of our proposed method was due to the multifeature auxiliary detection mechanisms, that is, SFM and DFM. To illustrate the importance of these modules, we implemented ablation experiments by disabling different modules.

4.4.1. Semantic Feature Auxiliary Detection Algorithm. Figure 6 shows the flow chart of the semantic feature auxiliary detection algorithm. Two modules were constructed, namely, the SFAM and the WM. The semantic feature-assisted detection algorithm was compared with the original algorithm in the experiment. As can be seen in Table 3, semantic feature assistance achieves better performance.

4.4.2. Detail Feature Auxiliary Detection Algorithm. The detail feature auxiliary detection algorithm was implemented, as shown in Figure 7. Two modules were also constructed: the detail feature auxiliary module (DFAM) and the WM. In the experiment, this detail feature-assisted detection algorithm was also compared with the original algorithm. As indicated in Table 3, the performance of the detail feature-assisted detection algorithm is only better than that of the basic weight network.

4.4.3. Multifeature Auxiliary Detection Algorithm. In the multifeature auxiliary detection algorithm, both the above modules were introduced into the network and fused with the basic weighted network structure.

The detection results of the above three algorithms and the original benchmark algorithm are shown in Table 3.

Table 3 shows the comparison results of those algorithms proposed in this paper and the original algorithm, from which we can see the performance of the designed modules. On the left side of the table are different algorithms, while on the right side are the corresponding experimental results. The average precision (AP) of objects in the five classifications has also been calculated separately for the bird, bus, cow, motorbike, and sofa. The mAP represents the mean of AP for these five novel classes. The first row is the original algorithm, and the other three are related algorithms proposed in this paper. The second and third rows are the results of ablation experiments. The second algorithm adds the DFAM to the original algorithm, which is called the detail feature auxiliary detection algorithm. It can be observed that the AP for the bird, cow, motorbike, and sofa is higher than

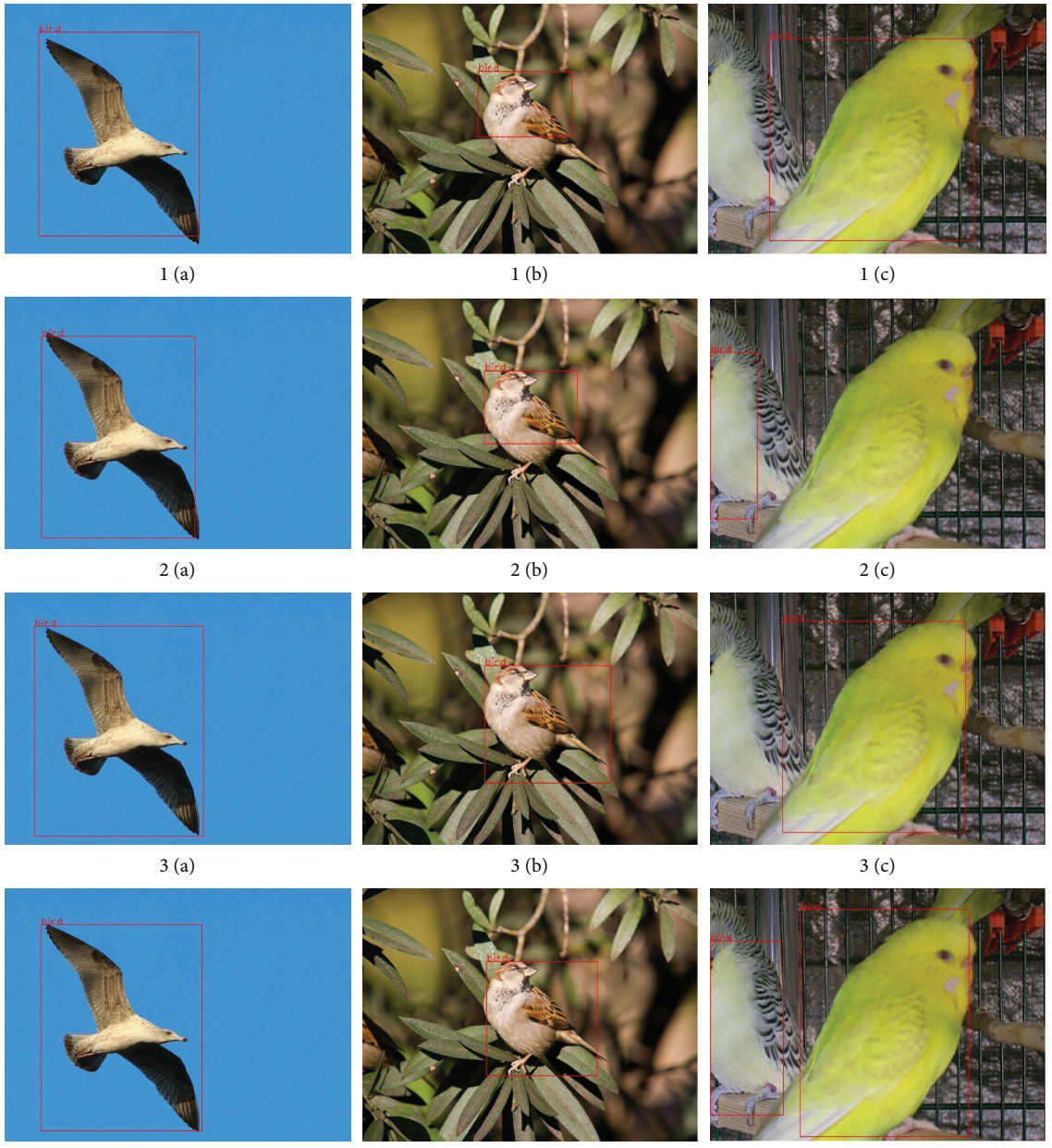

Figure 4: Bird object detection results.

that of the original algorithm, while it is not the case for the bus. Well, the performance is improved by less than 1%. The third row adds the SFAM to the original algorithm, which is called the semantic feature auxiliary detection algorithm. It can be observed that the AP of the algorithm is higher than that of the original algorithm, with an increase of 2.58% in AP and 1.75% in mAP for the bus. It can also be seen that the SFAM proposed in this paper does enhance the semantic information of objects and promotes classification accuracy and detection precision. The fourth row is the result of the multifeature auxiliary detection algorithm proposed in this paper. Notably, the algorithm has enhanced the detailed information and semantic information, and the detection results of all those five novel objects are superior to the original algorithm. In particular, the AP of the bus has been improved by 5.06%, and the mAP has been improved by 2.97%. The detection results in the fourth row show that the detail feature and semantic feature, two auxiliary information introduced in the algorithm, successfully improve the AP and mAP of one-shot object detection.

Figure 5: Bus object detection results.

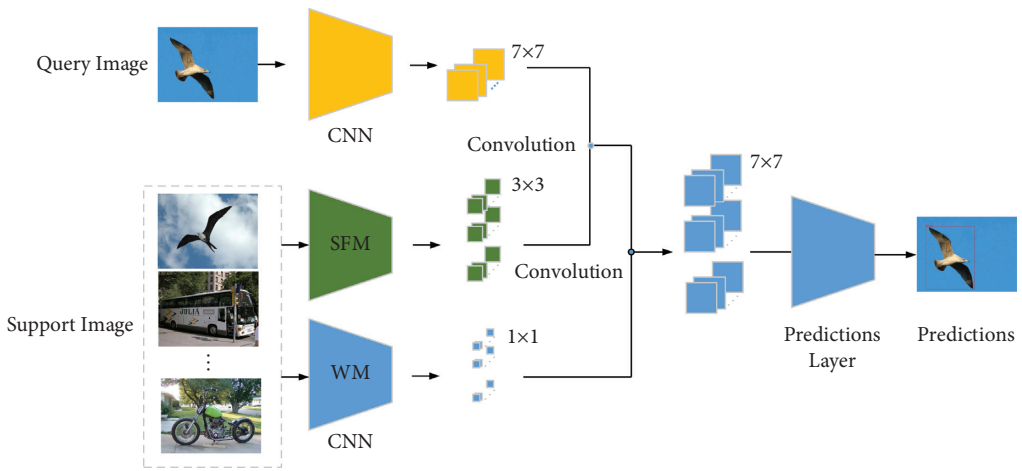

Figure 6: One-shot detection algorithm structure assisted by semantic information.

Table 3: Ablation experiment results.

Methods	Novel					
	Bird	Bus	Cow	Mbike	Sofa	mAP
Benchmark algorithm [16]	24.12	3.45	24.97	26.44	27.51	21.3
Detail feature auxiliary detection	**24.52**	3.27	**26.58**	**27.03**	**28.14**	**21.91**
Semantic feature auxiliary detection	**25.37**	**6.03**	**26.32**	**28.09**	**29.44**	**23.05**
Multifeature auxiliary detection	**27.02**	**8.51**	**27.8**	**28.67**	**29.35**	**24.27**

The bold values indicate that the performance of the proposed method is better than that of the Benchmark algorithm.

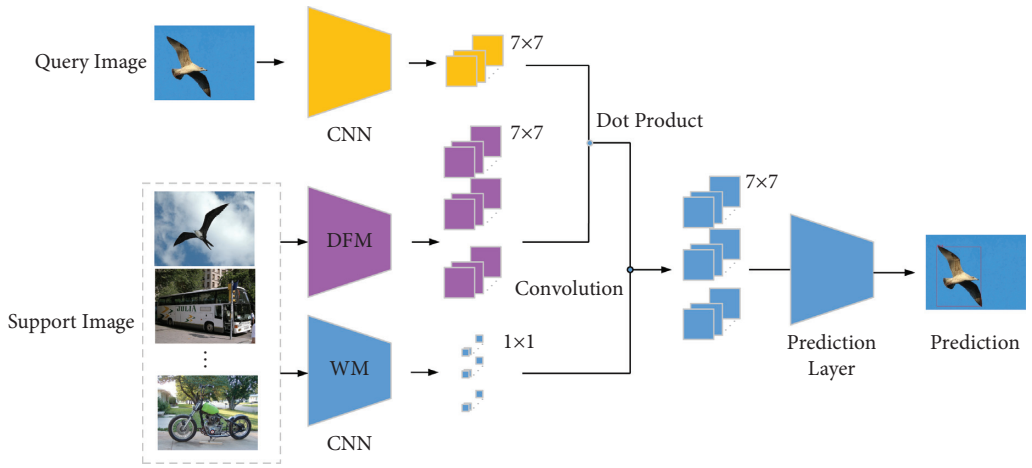

Figure 7: One-shot detection algorithm structure assisted by detailed information.

5. Conclusion

In this paper, a novel one-shot detection method based on multifeature auxiliary information was proposed. Compared to previous studies, this algorithm utilized two auxiliary mechanisms: the Semantic Feature Module and the Detail Feature Module, which significantly improved the detection effect of a single sample object. Experimental results on public datasets demonstrate that the new proposed algorithm has better one-shot detection performance than the original method. To further evaluate the advanced performance, an ablation experiment was conducted. Experiments showed that the two auxiliary modules play a positive role in the detection results. The combined detection accuracy of the two modules has been increased by 2.97% compared to the benchmark algorithm. In the future, we will apply this proposed method to other types of datasets, including infrared images and SAR images. Although this method is mainly for one-shot object detection, we also look forward to its application in few-shot object detection.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research was supported by the National Natural Science Foundation of China under Grant no. 62006240.

References

[1] T. R. Gadekallu, M. Alazab, R. Kaluri, and P. Reddy, "Hand gesture classification using a novel CNN-crow search algorithm," *Complex & Intelligent Systems*, vol. 7, no. 6, 2021.

[2] T. R. Gadekallu, D. S. Rajput, M. Reddy et al., "A novel PCA–whale optimization-based deep neural network model for classification of tomato plant diseases using GPU," *Journal of Real-Time Image Processing*, pp. 1–14, 2020.

[3] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 779–788, IEEE, Las Vegas, NV, USA, June 2016.

[4] J. Redmon and A. Farhadi, "YOLO9000: better, faster, stronger," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, pp. 6517–6525, IEEE, Honolulu, HI, USA, July 2017.

[5] J. Redmon and A. Farhadi, "YOLOv3: an incremental improvement," 2021, https://arxiv.org/pdf/1804.02767.pdf.

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

[7] R. Girshick, "Fast R-CNN," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 1440–1448, IEEE, Santiago, Chile, December 2015.

[8] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[9] P. Bateni, R. Goyal, V. Masrani, F. Wood, and L. Sigal, "Improved few-shot visual classification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14493–14502, IEEE, Seattle, WA, USA, June 2020.

[10] J. Requeima, J. Gordon, J. Bronskill, S. Nowozin, and R. E. Turner, "Fast and flexible multi-task classification using conditional neural adaptive processes," 2021, https://arxiv.org/abs/1906.07697v2.

[11] X. Tao, X. Hong, X. Chang, S. Dong, X. Wei, and Y. Gong, "Few-shot class-incremental learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12183–12192, IEEE, Seattle, WA, USA, June 2020.

[12] L. Zhou, P. Cui, X. Jia, S. Yang, and Q. Tian, "Learning to select base classes for few-shot classification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4624–4633, IEEE, Seattle, WA, USA, June 2020.

[13] C. Zhang, Y. Cai, G. Lin, and C. Shen, "DeepEMD: few-shot image classification with differentiable earth mover's distance and structured classifiers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12203–12213, IEEE, Seattle, WA, USA, June 2020.

[14] Z. Peng, Z. Li, J. Zhang, Y. Li, G. Qu, and J. Tang, "Few-shot image recognition with knowledge transfer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 441–449, IEEE, Seoul, South Korea, November 2019.

[15] B. Liu, H. Kang, H. Li, and G. Hua, "Few-shot open-set recognition using meta-learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8798–8807, IEEE, Seattle, WA, USA, June 2020.

[16] J.-M. Perez-Rua, X. Zhu, H. Timothy, and T. Xiang, "Incremental few-shot object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13846–13855, IEEE, Seattle, WA, USA, June 2020.

[17] Q. Fan, W. Zhuo, C.-K. Tang, and Y. Tai, "Few-shot object detection with attention-RPN and multi-relation detector," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4013–4022, IEEE, Seattle, WA, USA, June 2020.

[18] B. Kang, Z. Liu, X. Wang, F. Yu, J. Feng, and T. Darrell, "Few-shot object detection via feature reweighting," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8420–8429, IEEE, Seoul, South Korea, November 2019.

[19] T.-I. Hsieh, Y.-C. Lo, H.-T. Chen, and T. L. Liu, "One-shot object detection with co-attention and co-excitation," 2021, https://arxiv.org/abs/1911.12529.

[20] X. Li, C. Lin, C. Li et al., "Improving one-shot NAS by suppressing the posterior fading improving one-shot NAS by suppressing the posterior fading," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13836–13845, IEEE, Seattle, WA, USA, June 2020.

[21] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (VOC) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010.

[22] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: a retrospective," *International Journal of Computer Vision*, vol. 111, no. 1, pp. 98–136, 2015.