Few Shot Object Detection via a Generalized Feature Extraction Net

Dengyong Zhang[1], Huaijian Pu[1], Feng Li[1], R. Simon Sherratt[2], Se-Jung Lim[3]*

[1] School of Computer & Communication Engineering, Changsha University of Science & Technology, China
[2] School of Systems Engineering, the University of Reading, UK
[3] AI Liberal Arts Studies, Division of Convergence, Honam University, Republic of Korea
zhdy@csust.edu.cn, pujianjian1007@163.com, lif@csust.edu.cn, sherratt@ieee.org, limsejung@honam.ac.kr

Abstract

It is a new problem for deep learning to train a model on a small number of known targets to detect this object. Many recent studies are based on fine-tuning methods to solve. However, there is a lot of redundant information in the model during feature extraction, which will aggravate the difficulty of fine-tuning the model. Moreover, the neural network using the cross-entropy loss function classifier trained in few shots is prone to overfitting. We use the RS structure to reduce the number of channels in the model to reduce the repeated features in feature extraction. In addition, we use the Pearson distance function to calculate the classification loss of the model, to use the nonparametric method to reduce the number of parameters and prevent overfitting. Experimental results show that our method is better than the previous methods on Pascal VOC and FSOD datasets.

Keywords: Few shot object detection, Fine-tunning, Overfitting, Redundant information

1 Introduction

Deep learning achieves very good results in scenarios with sufficient data [1-2]. However, when there are few labeled data samples, it cannot achieve its desired effect [3]. The human brain and visual system can quickly learn and build conceptual models with a few examples, even for very young children with very little guidance. We believe that deep learning should also have this ability.

Inspired by human learning ability, few shot learning came into being and has been concerned and studied by many researchers in recent years. For example, Munkhdalai [4], Jamal [5], and Sun [6] are solved using meta learning methods. Learn general knowledge from other samples with a large number of samples, and then transfer this learning ability to samples that have not been contacted. In this process, the datasets is divided into base class and new class. The base class has a large number of samples, and the new class has only a small number of samples [7]. During the training process, the base class and the new class are divided into support sets and query sets at the same time. The purpose of the research is to enable the network to apply the learning ability in the base class to the new class.

Image classification is to output the category of the input picture [8-10]. Object detection is to output the category of the object and the location box corresponding to the object [11-12]. More requirements undoubtedly greatly increase the complexity of the model. Therefore, the previous few shot learning mainly focused on image classification tasks, while the research on few shot object detection came very late. The first paper on few shot detection was proposed by Kang [13] in 2019. In this paper, it is also the first time to use meta learning method to solve this problem. The object with only a few marker boxes is defined as new class, and the target with a large amount of sample data of other targets is defined as base class. The network trains a feature extractor on the basic classes that learn to learn, and then uses it to detect new classes.

In this method, we use a combination of fine tuning [14] and metric learning [15] to detect targets. The main idea of network design and training comes from transfer learning. Previously, fine-tuning was rarely applied to few shot object detection because it was almost impossible to fine-tune a large number of model parameters for a small number of samples. In this method, we first reduce the complexity of increasing duplicate features by reducing channel redundancy and then increase the diversity of features to increase the generalization of the base class training model. Finally, we use the measure function [16] to reduce the parameters that the model needs to be fine-tuned to achieve few shot object detection.

The nature of few shot object detection is still object detection, so our model design chooses two-stage object detection method [17]. One reason is that the two-stage object detection uses the region proposal network (RPN), which distinguishes the foreground from the background of the input picture first, without knowing the category of the object, which helps to find the object of the new class in the detection. Another reason is that after ROI Pooling, the extracted candidate box can be measured using a measure function in the metric space, bringing the same kind of samples closer and the different kinds of samples farther away. The detection method is modified based on the Faster RCNN [12] network. The entire detection model is trained on a data-rich base class, then the last layers of the network are fine-tuned in a small number of samples of new classes and base classes, while the other layers are frozen to adjust the model parameters.

In general, the main contributions of this paper are summarized below:

*Corresponding Author: Se-Jung Lim; E-mail: limsejung@honam.ac.kr
DOI: 10.53106/160792642023032402009

1. We propose a few shot object detection method that increases feature diversity and reduces channel redundancy. In this method, we increase feature diversity by flipping in the RS structure and use channel fusion to reduce channel redundancy. In the backbone network, we use the High-resolution network as the feature extraction network and add parallel spatial and channel attention modules for each multi-layer feature.

2. To increase the distance between different classes and reduce the distance between the same classes, we use Pearson metric to calculate the similarity between classes and reduce the complexity between classes.

2 Related Works

2.1 Object Detection

In 2014, R. Grirshick [18] proposed RCNN algorithm based on convolutional neural network, which brought object detection into the era of deep learning. Since then, the field of object detection began to develop at an unprecedented speed. The object detection based on Convolutional Neural Networks (CNN) can be divided into two categories: two-stage [19] and one-stage [20-23]. The main idea of two-stage object detection algorithm is to first extract high-quality region proposal using neural network, then classify and regression the extracted region proposal. Faster R-CNN proposed by the representative algorithm for S. Ren et al. Based on Fast R-CNN [11], this algorithm presents a region Proposal Network (RPN) based on neural network to extract the region recommendation box, improves the generation mechanism of the region recommendation box, and improves the speed and accuracy of the algorithm significantly. Specifically, the algorithm first uses the backbone of feature extraction to extract the feature map of the input image, then RPN generates a large number of region suggestion boxes based on the feature map, which is then mapped back to the feature map to get corresponding region suggestion feature boxes. Finally, the classification head of the network performs classification and regression operations on a large number of region suggestion feature boxes to get the final prediction results. That is, the category and corresponding position of objects on the input image. The two-stage object detection algorithm mainly focuses on the improvement of detection accuracy. The disadvantage is that the algorithm is more complex and the detection speed is slower.

2.2 Few Shot Object Detection

The application of few shot learning in image classification tasks has achieved relatively significant results [24]. Since the concept was proposed, there have been many excellent few shot image classification algorithms. However, due to the complexity of the object detection compared with the image classification, only a small number of few shot object detection algorithms have been proposed, and most of the algorithms have been improved based on the successful experience of the few shot classification algorithm. It mainly using the idea of metric learning, meta-learning [25-28] and fine-tune learning. Next, we will briefly introduce the algorithm based on fine-tuning.

In 2018, H. Chen et al. proposed LSTD [14] algorithm at AAAI conference, and the research upsurge of few shot object detection was officially raised. First, the LSTD algorithm integrates the advantages of SSD [29] and Faster R-CNN in one structure to reduce the difficulty of knowledge transfer in few shot scenarios. Secondly, LSTD presents a new regularization method to help knowledge migration from source domain to target domain. In 2020, the Fsdet [30] algorithm proposed by X. Wang et al. At the ICML international conference has achieved good results in few shot object detection in an intuitive and efficient way. As shown in Figure 1, fsdet algorithm proposes a two-stage fine tuning approach (TFA) based on transfer learning. TFA method divides the training of few shot object detection into two stages: the first stage is the base training phase. In this stage, the object detection algorithm only uses the samples and labels of the base class for training. After training, the base model is obtained, and then the weights of the classification layer and regression layer of the prediction head of the base model are randomly initialized. The second stage is the fine tuning phase. In this stage, the object detection algorithm uses the few shot data set with class balance, freezes the backbone network and RPN, and only allows the classification layer and regression layer of the prediction head to participate in the training. Using TFA method, we can simply and effectively make the two-stage object detection algorithm suitable for few shot scenarios.

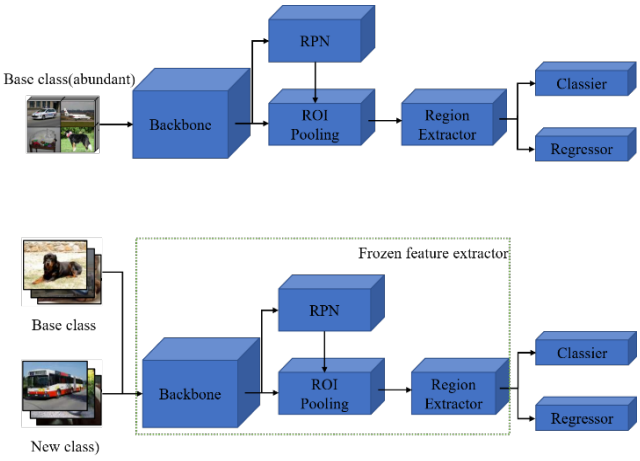

Figure 1. Fine-tuning method for few shot object detection
Note. The above figure is the basic training, and the following figure is the fine-tuning training

2.3 Metric Learning

In 2018 Karlinsky et. al [31] proposed a metric learning method based on representation for few shot classification. The authors assume that each category fits a Mixed Gaussian distribution in the feature space, and consider the majority of each component in the corresponding distribution for each category as a representation of that category [32-33]. By calculating the similarity between the features of the input image and the labels of each category, the probability of belonging to a class is obtained. Then the classification network is applied to the object detection model to achieve the object detection task based on few shot learning. Zhang et al. [34] think that the detection task based on metric learning is a conditional detection. The purpose of this task is to detect

the target that belongs to the same class in the given query image, essentially to find the most similar area in the two images. Therefore, the author introduces Bayesian conditional probability theory into RPN network in Faster R-CNN object detection model and proposes Comparison Net [34]. In 2019, Hsieh et al. [35] believed that the characteristics of the target image could provide a spatial context for distinguishing other foreground objects from backgrounds. Query images can provide a category context for finding more accurate detection targets. Therefore, in order to enhance the characteristics of new categories in query images and target images, a new object detection mechanism based on metric learning, co-attention and co-excitation (CoAE), is proposed to improve the detection accuracy.

3 The Proposed Approach

3.1 Methods and Settings

In this paper, we follow the LSTD [14] to set our data input. It is the first paper in this field. The input data is divided into base class C_b and new class C_n. Each class in the base class has a large number of samples, while each class in the new class has only k target instances. The size of k is generally taken as 50, 30, 10, 5, 3, 2, 1. For object detection dataset D ($D = \{x, y\}$, $x \in X$, $y \in Y\}$, where X represents input picture, Y represents its label) including category C_i and target box position R_i. This dataset D has class C, $C \in C_b \cup C_n$, $C_b \cap C_n = \phi$, C_b represents the base class category, and C_n represents the new class category. This paper conducts experimental results on Pascal VOC dataset and FOSD dataset. Taking VOC dataset as an example, VOC dataset has 20 categories, of which 15 categories are selected as the base category and the remaining 5 categories as the new category. We take all the targets in the 15 classes as the first stage of training, and in the fine-tuning stage, we select K targets in each class as input. Assuming that the new class in the second stage has n classes, the few shot object detection is regarded as N way-K shot. In the test, we hope that the network can get a good precision in both the base class and the new class, so we calculate the mean average precision (MAP) of all classes at the same time.

In this paper, we use the fine-tuning method proposed in FSdet [30] to solve the few shot object detection method. The overall design of the network is the same as the famous two-stage network Faster RCNN. This two-stage object detection method first needs to extract candidate boxes, and then classify and locate the extracted candidate boxes. So it is not an end-to-end detection model. In the process of extracting candidate boxes, the two-stage network only needs to distinguish between foreground and background, and does not need to distinguish the category of foreground. In this study, since both the new class and the base class are foreground, this can solve the problem that the region proposal network trained on the base class can be used on the new class. As shown in Figure 2, the network we use is mainly consists of backbone network, regional candidate box, ROI pooling, its classifier (C) and regressor (R). We use resnet50 as the backbone network to extract the features of the input image. The regional candidate box network is used to generate candidates, ROI pooling scales the candidate box to the same size, the classifier is used to get the category of the target, and the regressor gets the coordinate position of the target.

In the training process, this method first carries out the first stage training in the base class. This training process, like the conventional two-stage object detection. After that, the next stage training of fine tuning is carried out in the data consisting of new classes and base classes that only contain K targets. In the fine-tuning training, because the RPN network only distinguishes the foreground and background, the new class can also be selected by the RPN as the foreground. Therefore, the layer before the RPN is not trained, which can greatly reduce the network weight that needs to be adjusted in the fine-tuning training.

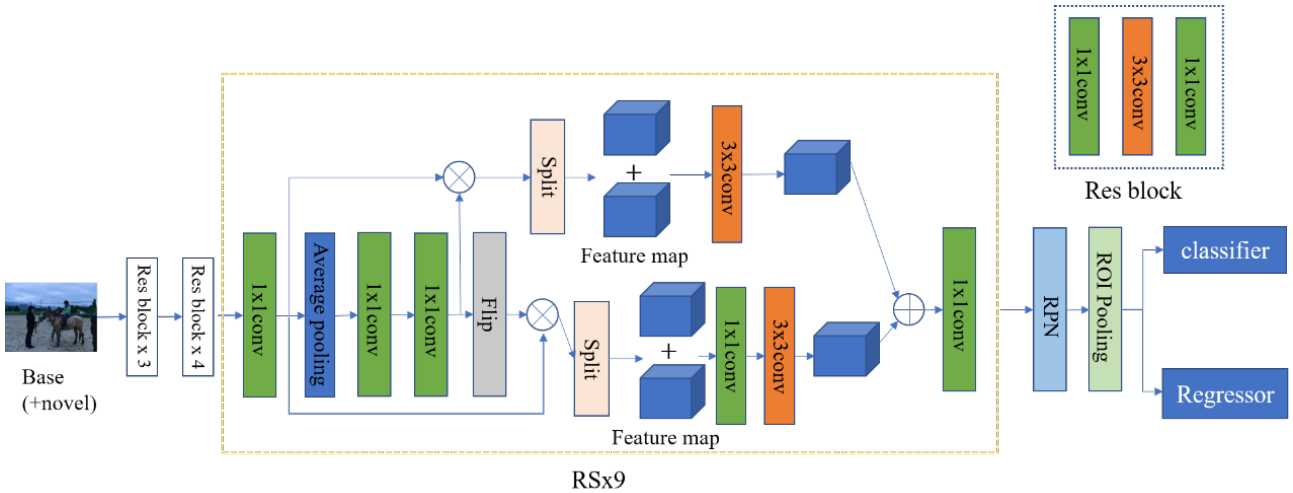

Figure 2. Few shot object detection method based on channel reduction and feature diversity

Note: The main framework of this network is based on Faster RCNN, and the backbone network is improved from Resnet50. The last several modules of Resnet50 are modified to RS structure to get a better few shot object detection structure.

3.2 RS Structure

In the process of feature extraction of deep learning, there are usually many overlapping feature information, which leads to a lot of redundant information. This redundant information will increase complexity for model training, and will lead to a lot of target noise when RPN generates candidate boxes. In addition, because the foreground and background of the target are obtained from the training of the base class, it can be applied to new classes, but there are often some differences between the new class and the base class, so we need to increase the diversity of the model. After the above consideration, we use an RS structure to increase the diversity of models and reduce the redundancy.

The RS structure is as shown in the Figure 2, in which the input feature (Called *F*), First, a 1x1 convolution is performed, and a vector is obtained by maximizing the pool of the obtained feature map. Then, the dimension of the obtained vector is reduced to 1/32 by 1x1 convolution. Finally, the dimension is increased to the original dimension by 1x1 convolution. Here, in order to increase the diversity of model features, we flip the obtained feature vectors. Then, the feature vectors that are not flipped and flipped are multiplied by the input features respectively to obtain the feature F_N and F_F. Through split, the number of channels obtained is divided into half of the original number, and then the corresponding positions are added to obtain F_{NA} and F_{FA}, and then reduce the dimension of F_{FA} to half through 3x3 convolution. So the number of channels F_{NA} is 1/2 of F and the number of channels F_{FA} is 1/4 of F. Then concat to get F_{out}. So the number of channels is 3/4 of F. In the process of splitting and fusion, the feature dimension is reduced, so as to reduce the redundancy. The flipping process increases the diversity of features. Then, the final characteristic graph is obtained through a 1x1 convolution.

3.3 Loss Function

The loss function in this method mainly includes the loss trained in the base class and the loss used in the fine tuning training.

In the first stage for base class training, we use the same loss function as in the fast RCNN, including the RPN loss function for extracting candidate boxs, and the classification and regression loss function for predicting target locations and categories in the whole network. The total loss is:

$$L = L_{rpn} + L_{cls} + L_{loc}. \quad \textbf{(1)}$$

L_{rpn} is the loss function of the RPN, which is a binary classification used to distinguish the foreground and background, and adjust the position of the anchor frame. L_{cls} is the category loss function of the classifier, and use cross entropy loss function. L_{loc} is the loss function of the regressor, and use L1 smooth loss function. In the fine-tuning stage, because there is no need to train the RPN network, RPN loss is no needed. However, in order to reduce the distance between classes and increase the distance between different classes, a Pearson correlation is used in the classifier. The correlation between the *i*th target of the input picture and the weight of each category can be expressed by the following formula.

$$S_{i,j} = \frac{\alpha E[(F(x)_i - \overline{F}(x)_i(w_j - \overline{w}_j)]}{\sqrt{\sum_{k=1}^{n}(F(x_k)_i - \overline{F}(x_k)_i)^2}\sqrt{\sum_{k=1}^{n}(w_{k,j} - \overline{w}_{k,j})^2}}. \quad \textbf{(2)}$$

Where $S_{i,j}$ is the distance between the ith target of input x and category j. $F(x)$ is the input feature map. $\overline{F}(x)_i$ is the average value of the instance target feature map. The weight matrix $W \in R^{d\times c}$ of each box category C can be written as $[w_1, w_2, ..., w_c]$, Where $w_c \in R^d$ is the weight vector for each category. Where α is a scale factor. In this experiment, we tried to use a multiple of 10 to 100, and finally found that 30 was the most appropriate.

4 Experimental Results and Analysis

4.1 Experimental Setting

We adopted the setting method similar to the training in mtfs [6]. As shown in Figure 1, set the data of the new class as *N* way *K* shot, where *N* is the number of categories of the sample image and K is the number of each sample. The experimental parameters are shown in Table 1. The values of N and *K* are different in different data sets. In the VOC dataset, the value of *N* is 5, and the value of *K* is 1, 3, 5, 10. And it is divided many times. For comparison with FSOD, we set *N*=5 and *K* to 5, that is, five pictures in each category. Using the learning rate decay strategy, set the initial learning rate of the entire model to 0.001. Each training contains 200 iterations. The step of learning rate attenuation is set to 48000. In recent years, the popularity of cloud computing and big data has provided high-performance computing power [36-37]. This experiment is conducted on the cloud platform. Our experiment is implemented under Ubuntu system, using python programming language and Pytorch deep learning framework. Three NVIDIA GTX 2080ti are used in the experiment.

Table 1. Simulation parameters and values

Parameters	Values
New class (N)	5
Number of each category (K)	1/3/5/10
Initial learning rate	0.001
The step of learning rate attenuation	48000
Epoch	200

4.2 Training Datasets

In order to detect the effect of the model in simple scenarios with a small number of categories and complex data scenarios with a large number of categories, VOC and fsod datasets are used in this paper to evaluate our method. Because many categories and pictures of fsod data are from coco data sets, and the detection results on VOC and fsod data sets can reflect the effect on coco data sets, this paper will not compare them on coco data sets.

4.2.1 PASCAL VOC

Since VOC has 20 categories, we have constructed three random splits. Each split has 15 categories as the base class and 5 categories as the new class. The categories in this ex-

periment are set as follows: the new categories are divided as follows for the first time: “bird”, “bus”, “cow”, “motorbike”, “sofa”. The new classes of the second split are as follows: “aeroplane”, “bottle”, “cow”, “horse”, “sofa”. The new classes of the third split are as follows: “boat”, “cat”, “motorbike”, “sheet”, “sofa”. In each division, select 1, 3, 5, and 10 pictures for each category of the new category.

4.2.2 Few Shot Object Detection (FSOD)

The FSOD dataset is proposed in the 2020 paper FSAM [7]. It is a dataset specially used for few shot object detection. The dataset contains 1000 categories, with more than 60K images and 182k bounding boxes. Each category contains approximately 100 samples.

4.3 Comparison of Experimental Results

4.3.1 Comparison of Experimental Results on VOC

The Pascal VOC dataset contains 20 categories. Follow most previous methods to divide the dataset three times, each time including the base class and the new class. Table 2 shows that the detection accuracy of our model basically exceeds that of the best detection model through the number of 1, 3, 5 and 10 samples on split1, split2 and split3. Obviously, our model shows better performance than previous methods. In particular, in split2, when there is only one picture, our model is 6.5% better than FSOD-KT. When k is set to 10, our model gets a competitive result. In split1, we get the best result of 57.8%. In split3, we achieved the best results in terms of the number of one picture, and achieved competitive results in terms of 5 pictures and 10 pictures. According to the overall results, our method can learn more general knowledge in few shot scenarios. In different cases (shots are 1, 3, 5, 10), our method has the ability of stability, which means that it can resist the instability caused by supporting image. As shown in Figure 3, it is the result of the first VOC Division.

As shown in Table 3, it is obvious that our model has achieved the most advanced results in the object detection tasks of single picture, three pictures and five pictures, with an average improvement rate of 3%, 2.1% and 1.1% respectively. In the task of 10 pictures, we also got a good result (52.1%).

4.3.2 Comparison of Experimental Results on FSOD

In Table 4, we compare our model with FSODAM, FRCNN and LSTD on the FSOD dataset. It can be seen that our method has certain advantages over these methods.

For fair comparison, we use the FSOD setting of 2-way 5 shot. We train the model on the training dataset of fsod, and then fine tune the model. It can be seen from the displayed results that our model has reached the most advanced level on AP50 and AP75. Surprisingly, our model exceeded the FSODAM by 6.2% on AP50, and even exceeded the FSODAM by 8.9% on the more stringent index AP75.

Figure 3. The new class detection results of our method in the first division of VOC dataset

Table 2. On VOC, divide the data set randomly three times, take 1, 2, 3, 5, and 10 for the divided data sets shot respectively, and get the AP50 result

	Split1				Split2				Split3			
Model/Shot	1	3	5	10	1	3	5	10	1	3	5	10
FSODFR [13]	14.8	26.7	33.9	47.2	15.7	22.7	30.1	39.2	19.2	25.7	40.6	41.3
MetaR-CNN	19.9	35.0	45.7	51.5	10.4	29.6	34.8	45.4	14.3	27.5	41.2	48.1
FSOD-KT [38]	27.8	46.2	55.2	56.8	19.8	38.7	38.9	41.5	29.5	38.6	43.8	45.7
Ours	29.5	48.6	53.4	57.8	26.3	42.6	43.5	52.1	30.5	40.3	44.3	46.4

Table 3. Mean average precision on different number of picture data on VOC

Model/Shot	1	3	5	10
FSODFR [13]	16.6	32.5	34.9	42.6
Meta R-CNN [26]	14.9	37.0	40.6	48.3
FSOD-KT [38]	25.7	41.7	46.0	48.0
Ours	28.7	43.8	47.1	52.1

Table 4. Results of AP50 and AP75 on FSOD dataset

Model	FSOD pretrain	Fine-tune	AP50	AP75
FRCNN [12]	no	yes	11.8	6.7
FRCNN [12]	yes	yes	23.0	12.9
LSTD [14]	yes	yes	24.2	13.5
FSODAM [7]	yes	no	27.5	19.4
Ours	yes	no	33.7	28.3

5 Conclusion

In this paper, a very simple small sample target detection network based on fine tuning is used. In the network, only the network parameters of the following layers need to be fine-tuned for other types of trained networks. RS structure is used in the network to reduce the redundancy between channels. In order to realize the generalization detection from base class to new class, RS structure also increases the diversity of features. In addition, in order to reduce the fine-tuning parameters, reduce the intra class distance and increase the inter class distance, a measurement coefficient is used in the network to optimize. In general, this method optimizes small sample target detection to a certain extent, and does not bring revolutionary changes to small sample target detection. At present, the accuracy of small sample target detection is still not high, and it needs a long way to go in the future.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under grant 62172059 and 62072055, Hunan Provincial Natural Science Foundations of China under Grant 2020JJ4626 and 2022JJ30621, Scientific Research Fund of Hunan Provincial Education Department of China under Grant 19B004.

References

[1] C. Szegedy, W. Liu, Y.-Q. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Boston, MA, USA, 2015, pp. 1-9.

[2] C. Chen, K.-L. Li, W. Wei, J.-T. Zhou, Z. Zeng, Hierarchical Graph Neural Networks for Few-Shot Learning, *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 32, No. 1, pp. 240-252, January, 2022.

[3] R.-X, Duan, D. Li, Q. Tong, T. Yang, X.-T. Liu, X.-L. Liu, A Survey of Few-Shot Learning: An Effective Method for Intrusion Detection, *Security and Communication Networks*, Vol. 2021, pp. 1-10, October, 2021.

[4] T. Munkhdalai, H. Yu, Meta networks, *International Conference on Machine Learning*, Sydney, Australia, 2017, pp. 2554-2563.

[5] M. A. Jamal, G.-J. Qi, Task agnostic meta-learning for few-shot learning, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, 2019, pp. 11719-11727.

[6] Q. Sun, Y. Liu, T.-S. Chua, B. Schiele, Meta-Transfer Learning for Few-Shot Learning, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, 2019, pp. 403-412.

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

[8] J.-G. Chen, K.-L. Li, K. Bilal, X. Zhou, K.-Q. Li, P. S. Yu, A Bi-layered Parallel Training Architecture for Large-Scale Convolutional Neural Networks, *IEEE Transactions on Parallel and Distributed Systems*, Vol. 30, No. 5, pp. 965-976, May, 2019.

[9] B. Pu, K.-L. Li, S.-L. Li, N.-B. Zhu, Automatic Fetal Ultrasound Standard Plane Recognition Based on Deep Learning and IIoT, *IEEE Transactions on Industrial Informatics*, Vol. 17, No. 11, pp. 7771-7780, November, 2021.

[10] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 2016, pp. 770-778.

[11] R. Girshick, Fast r-cnn, *Proceedings of the IEEE International Conference on Computer Vision*, Santiago, Chile, 2015, pp. 1440-1448.

[12] S.-Q. Ren, K.-M. He, R. Girshick, 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, June, 2017.

[13] B. Kang, Z. Liu, X. Wang, F. Yu, J. Feng, T. Darrell, Few-shot object detection via feature reweighting, *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Seoul, Korea, 2019, pp. 8420-8429.

[14] H. Chen, Y. Wang, G. Wang, Y. Qiao, Lstd: A low-shot transfer detector for object detection, *Proceedings of the AAAI Conference on Artificial Intelligence*, New Orleans, LA, USA, 2018, pp. 2836-2843.

[15] Y. Li, W. Feng, S. Lyu, Q. Zhao, X. Li, MM-FSOD: Meta and metric integrated few-shot object detection, December, 2020. https://arxiv.org/abs/2012.15159

[16] D. S. Trigueros, L. Meng, M. Hartnett, Face recognition: From traditional to deep learning methods, October, 2018. https://arxiv.org/abs/1811.00116

[17] J. Dai, Y. Li, K.-M. He, J. Sun, R-fcn: Object detection via region-based fully convolutional networks, *Advances in Neural Information Processing Systems*, Barcelona, Spain, 2016, pp. 379-387.

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

[19] K.-M. He, G. Gkioxari, P. Dollár, R. Girshick, Mask r-cnn, *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, 2017, pp. 2980-2988.

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

[21] J. Redmon, A. Farhadi, YOLO9000: better, faster, stronger, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, 2017, pp. 7263-7271.

[22] J. Redmon, A. Farhadi, Yolov3: An incremental improvement, April, 2018. https://arxiv.org/abs/1804.02767?e05802c1_page=1

[23] D. Cao, Z. Chen, L, Gao, An improved object detection algorithm based on multi-scaled and deformable convolutional neural networks, *Human-centric Computing and Information Sciences*, Vol. 10, No. 1, pp. 1-22, April, 2020.

[24] Y. Wang, Q. Yao, J. T. Kwok, L.-M. Ni, Generalizing from a few examples: A survey on few-shot learning, *Association for Computing Machinery Computing Surveys*, Vol. 53, No. 3, pp. 1-34, May, 2021.

[25] Y.-X. Wang, D. Ramanan, M. Hebert, Meta-learning to detect rare objects, *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Seoul, Korea, 2019, pp. 9925-9934.

[26] X. Yan, Z. Chen, A. Xu, X. Wang, X. Liang, L. Lin, Meta r-cnn: Towards general solver for instance-level low-shot learning, *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Seoul, Korea, 2019, pp. 9577-9586.

[27] K. Fu, T.-F. Zhang, Y. Zhang, M.-L. Yan, Z.-H. Chang, Z. Y. Zhang, X. Sun, Meta-SSD: Towards fast adaptation for few-shot object detection with meta-learning, *IEEE Access*, Vol. 7, pp. 77597-77606, June, 2019.

[28] G. Zhang, Z. Luo, K. Cui, S. Lu, Meta-detr: Few-shot object detection via unified image-level meta-learning, September, 2021. https://arxiv.org/abs/2103.11731

[29] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, Ssd: Single shot multibox detector, *European Conference on Computer Vision*, Amsterdam, The Netherlands, 2016, pp. 21-37.

[30] X. Wang, T.-E. Huang, T. Darrell, J. E. Gonzalez, F. Yu, Frustratingly simple few-shot object detection, *Proceedings of the 37th International Conference on Machine Learning*, Virtual Event, 2020, pp. 9919-9928.

[31] L. Karlinsky, J. Shtok, S. Harary, E. Schwartz, A. Aides, R. Feris, R. Giryes, A. M. Bronstein, RepMet: Representative-based metric learning for classification and few-shot object detection, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, 2019, pp. 5192-5201.

[32] D. Cao, K. Zeng, J. Wang, P. K. Sharma, X.-M. Ma, Y.-H. Liu, S.-Y. Zhou, BERT-based Deep Spatial-Temporal Network for Taxi Demand Prediction, *IEEE Transactions on Intelligent Transportation Systems*, Vol. 23, No. 7, pp. 9442- 9454, July, 2022.

[33] J. Zhang, S. Zhong, T. Wang, H.-C. Chao, J. Wang, Blockchain-Based Systems and Applications: A Survey, *Journal of Internet Technology*, Vol. 21, No. 1, pp. 1-14, January, 2020.

[34] T. Zhang, Y. Zhang, X. Sun, H. Sun, M. Yan, X. Yang, K. Fu, Comparison network for one-shot conditional object detection, January, 2020. https://arxiv.org/abs/1904.02317

[35] T.-I. Hsieh, Y.-C. Lo, H.-T. Chen, T.-L. Liu, One-shot object detection with co-attention and co-excitation, *Advances in Neural Information Processing Systems*, Vancouver, Canada, 2019, pp. 2725-2734.

[36] J. Wang, Y. Yang, T. Wang, R. Sherratt, J. Zhang, Big Data Service Architecture: A Survey, *Journal of Internet Technology*, Vol. 21, No. 2, pp. 393-405, March, 2020

[37] J. Wang, C.-Y Jin, Q. Tang, N.-X. Xiong, G. Srivastava, Intelligent Ubiquitous Network Accessibility for Wireless-Powered MEC in UAV-Assisted B5G, *IEEE Transactions on Network Science and Engineering*, Vol. 8, No. 4, pp. 2801-2813, October-December, 2021.

[38] G. Kim, H.-G. Jung, S.-W. Lee, Few-shot object detection via knowledge transfer, *2020 IEEE International Conference on Systems, Man, and Cybernetics*, Toronto, ON, Canada, 2020, pp. 3564-3569.

Biographies

Dengyong Zhang received the B.S. and M.S. degree from Changsha University of Science and Technology, Changsha, China, in 2003, 2006 respectively. He received Ph.D. degree from Hunan University, China, in 2018. Now, He is an associate professor at Changsha University of Science and Technology. His current research interests include digital media forensics and image processing

Huaijian Pu received the B.S. degree from Hunan Agricultural University, Changsha, China, in 2017. He is a postgraduate student in Changsha University of Science and Technology, Changsha, China. His research interests include computer vision, object detection, and deep learning.

Feng Li received the B.S. degree from Hunan Normal University, China, in 1984. He received M.S. degree from Zhejiang University, China, in 1988. He received Ph.D. degree from Sun Yat-sen University, China, in 2003. He is a professor at Changsha University of Science and Technology. His main research interests lie in computer vision and pattern recognition.

R. Simon Sherratt (M'97-SM'02-F'12) received the B.Eng. degree in Electronic Systems and Control Engineering from Sheffield City Polytechnic, UK in 1992, M.Sc. in Data Telecommunications in 1994 and Ph.D. in video signal processing in 1996 both from the University of Salford. Since 1996, he has been a Lecturer in Electronic Engineering at the University of Reading, currently a Senior Lecturer in Consumer Electronics and a Director for Teaching and Learning. His research topic is signal processing in consumer electronic devices concentrating on equalization and DSP architectures

Se-Jung Lim graduated from the department of Computer Engineering at Chonnam National University in 2008. She received her a master's degree and doctor's degree from Chonnam National University in 2010 and in 2016. She worked for Huneed Technologies CO., LTD as a senior researcher from 2013 to 2014. In April 2019, she joined Honam University and is currently an assistant professor of Liberal Arts & Convergence Studies Division of Convergence at Honam University. Her research interest includes Wireless Sensor Networks, Internet of Things and Big Data.