IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received 10 July 2023, accepted 12 August 2023, date of publication 23 August 2023, date of current version 30 August 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3307490

RESEARCH ARTICLE

DA-FSOD: A Novel Data Augmentation Scheme for Few-Shot Object Detection

JIAN YAO[1,2], TIANYUN SHI[2], XIAOPING CHE[3], (Member, IEEE), JIE YAO[4], AND LIUYI WU[2]

[1]Postgraduate Department, China Academy of Railway Sciences, Beijing 100082, China
[2]China Academy of Railway Sciences Corporation Ltd., Beijing 100081, China
[3]School of Software Engineering, Beijing Jiaotong University, Beijing 100044, China
[4]School of Information Management, Beijing Information Science and Technology University, Beijing 100101, China

Corresponding author: Tianyun Shi (shitianyun@sina.com)

This work was supported by the Young Elite Scientists Sponsorship Program by CAST under Grant 2020QNRC001.

ABSTRACT Deep learning techniques continue to be used in various applications in recent years. However, when it is difficult to obtain adequate training samples, the performance of the depth model will degrade. Although few-shot learning and data enhancement techniques can relieve this dilemma, the diversity of real data is too large to simulate. To tackle this challenge, we study a novel method, Data Augmentation Scheme For Few-Shot Object Detection (DA-FSOD), to improve the efficiency of model training on visual tasks. Specifically, to expand data augmentation space, we build a data augmentation operation pool (DAOP) based on several common-applied image process operations. Then we propose a novel data augmentation scheme, the series and parallel connection scheme, which superimposes the effects of different operations to generate diverse variants. To further explore and utilize the deep feature information, we leverage the semantic information of input image in model and propose imposed semantic data augmentation which augments training set semantically via deep features of augmented variants. The proposed method successfully enhanced the model performance. We validated our approach using extensive experiments on the domain of few-shot object detection. The results showed remarkable gains compared to state-of-the-art methods.

INDEX TERMS Few-shot learning, object detection, data augmentation, semantic information, image processing.

I. INTRODUCTION

With the development of technology, the performance of convolutional neural network [1] has improved a lot compared with before. This makes deep learning technology can better serve daily lives and bring convenience to all aspects of life, especially in visual-related applications. Among them, object detection [2] is a commonly applied computer technology related to computer vision and image processing that deals with detecting instances of semantic objects of a certain class in digital images and videos. Object detection has applications in many areas of computer vision, including image retrieval [3] and video surveillance [4].

With the continuous deepening of research work, a large number of detection frameworks [5], [6] have been proposed, and the accuracy and speed of object detection have been greatly improved. However, the object detection task usually consists of two sub-tasks, detecting the location of the target object and then recognizing it. This makes most methods of target detection consist of an impossible framework. These complex frameworks usually involve designing a specific network module, extracting features from a large amount of data, and processing the extracted features through a specific algorithm to obtain the final results. While these feature extraction modules are diverse, most of them rely on widely used convolutional neural networks. Therefore, the stability and feature extraction ability of the depth convolutional neural network will become one of the key factors affecting the target detection performance.

It is usually required abundant training data to get a well-trained model. With the development of scientific research, the training data for various tasks have been accumulated.

The associate editor coordinating the review of this manuscript and approving it for publication was Andrea F. Abate.

This makes it possible to use large-scale data training models, and the performance of computing resources is constantly improved, so that the model training efficiency has been greatly boosted. However, the model trained under a specific data set can only learn the characteristics of a specific distribution of data. This makes the performance of the model fragile, as other data distributions emerge from the test data. [7]. Therefore, a reliable and robust depth vision system becomes essential.

In order to solve the stability and efficiency problems of deep models, one of the methods is to expand the data distribution of the training set. More and more data sources are used to complete the training process. The model trained under this condition is considered to have acquired more knowledge of the characteristics of the data distribution. However, the process of model training is quite complicated, and some wrong attention to tiny details is more likely to lead to high sensitivity and instability of deep learning classifiers. On the other hand, while increasing the size of the training data is beneficial to the performance of the model used, it incurs substantial additional costs during the training process. One is that collecting and organizing data suitable for model training will bring huge manpower and material resources, and the other is that training the model based on large-scale data will bring additional time costs. In addition, during the data collection process, there may be situations where the data of some specific objects cannot be obtained in large quantities. These require the adopted model to be able to adequately learn object features from limited datasets.

Few-shot learning [8] is an effective way to improve models in data-constrained situations. Few-shot learning also referred to as low-shot learning in a few sources, is a type of machine learning method where the training dataset contains limited information. It aims to build accurate machine-learning models with less training data. Few-shot learning algorithms coupled with a data-centric approach to model development can help companies reduce data analysis or machine learning costs since the amount of input data is an important factor that determines resource costs. As the amount of training data available is insufficient, great quantities of prior knowledge are usually used in the process of constructing a few-shot learning algorithm. For example, machine learning models [9] exploit prior knowledge about the structure and variability of the data, which enables the construction of viable models from a few examples.

Besides, applying data augmentation techniques during model training is another effective strategy in few-shot learning. Data augmentation is a technique used to increase the amount of data by adding slightly modified copies of already existing data or newly created synthetic data from existing data. It acts as a regularizer and helps reduce overfitting when training a machine learning model. Some traditional image process operations, such as equalization, are embedded in various well-applied frameworks. An increasing number of methods have been proposed that aim to augment the training data. Recently, in view of the mismatched distribution of the training set and test set, a data enhancement technology named Augmix [10] is proposed, which uses the combination of different enhancement operations to coordinate the loss of consistency. Data augmentation is useful to improve the performance and outcomes of machine learning models by forming new and different examples to train datasets. Although data augmentation techniques can provide additional training samples for the model, the diversity of the data it generates is a key factor in the performance of the technique. In addition, in order to generate some more diverse data, additional computational overhead will be introduced into the model training process, which will reduce the efficiency of data augmentation methods.

To address these issues, we first build a data augmentation operation pool (DAOP) using several common operations. And we propose a novel data augmentation scheme, that is series and parallel connection scheme, to generate more diverse variants. Then we introduce semantic information as translating deep features along a specific direction corresponds to performing meaningful semantic transformations on the input image. And we propose imposed semantic data augmentation which augments the training set semantically via deep features of augmented variants.

To the best of our knowledge, this is the first study using semantic information to improve the performance of few-shot learning problem. We verify the effectiveness of the proposed method based on an important application in few-shot learning, that is few-shot object detection (FSOD).

Recently, Sun et al. [11] propose a few-shot object detection framework based on contrastive proposal encoding (FSCE) and achieve remarkable performance. We select this as our strong baseline in our experiment. The experimental results have proven that the proposed method boosts the performance of few-shot object detection compared with FSCE. The major contributions of this paper are summarized as follows.

- The series and parallel connection scheme is proposed based on a newly constructed data augmentation operation pool.
- Imposed semantic data augmentation is introduced to enrich the diversity of the training and generate the rationality of augmented variants.
- A large number of experiments are conducted to evaluate the performance of DA-FSOD in improving the data effectiveness of a given model.

II. RELATED WORK

A. DATA AUGMENTATION

Data augmentation involves techniques used for increasing the amount of data, based on different modifications, to expand the number of examples in the original dataset. Data augmentation not only helps to grow the dataset but also increases the diversity of the dataset. When training machine learning models, data augmentation acts as a regularizer and helps to avoid overfitting. Data augmentation techniques have been found useful in domains like NLP [12] and computer vision [13]. In NLP, data augmentation techniques can include swapping, deletion, and random insertion, among others. In computer vision, transformations like cropping, flipping, and rotation, have already been applied to many well-used deep-learning frameworks.

Numerous data augmentation methods have been proposed in recent years. DeVries and Taylor citedevries2017improved proposed a simple regularization technique of randomly masking out square regions of input during training, which is called cutout. Zhang et al. [15] analyzed the shortcomings of Empirical Risk Minimization (ERM) and illustrated that it is more proper to update the network according to the Vicinal Risk Minimization (VRM) principle by producing an element-wise convex combination of two images, which is called Mixup. Later, Yun et al. [16] proposed CutMix by cutting a patch and pasting a patch from the same place in another training image, and the corresponding label is also mixed in proportion to the size of the patch. Cubuk et al. [17] proposed RandAugment which has a significantly reduced search space and can be trained on the target task with no need for a separate proxy task. Furthermore, RandAugment can be used uniformly across different tasks and datasets and works out of the box, matching or surpassing all previous automated augmentation approaches on several datasets. Park et al. [18] proposed a data augmentation method for speech recognition, called SpecAugment, which can be applied directly to the feature inputs of a neural network. SpecAugment consists of warping the features, masking blocks of frequency channels, and masking blocks of time steps. Wei and Zou et al. [19] proposed a data augmentation technique for boosting performance on text classification tasks which consists of four operations: synonym replacement, random insertion, random swap, and random deletion. Recently, Wang et al. [20] proposed implicit semantic data augmentation (ISDA) to complement traditional augmentation schemes. Although data enhancement methods emerge in endlessly, these methods lack guidance on the specificity of the enhancement method, which leads to the limitation of the efficiency of the method.

B. OBJECT DETECTION

Object detection is the task of detecting instances of objects of a certain class within an image. The state-of-the-art methods can be categorized into two main types: one-stage methods and two stage-methods. One-stage methods prioritize inference speed, and example models include SSD [2] and YOLO [21]. Two-stage methods prioritize detection accuracy, and example models include Faster R-CNN [6] and Mask R-CNN [22]. The most popular benchmarks are the MSCOCO [11] and Pascal VOC [23] datasets. Models are typically evaluated according to a Mean Average Precision metric.

Liu et al. [2] proposed a single shot multiBox detector (SSD) for object detection. SSD discretizes the output space of bounding boxes into a set of default boxes over different aspect ratios and scales per feature map location. At prediction time, the network combines predictions from multiple feature maps with different resolutions to naturally handle objects of various sizes. Redmon et al. [21] proposed YOLO (You Only Look Once) by framing object detection as a regression problem to spatially separated bounding boxes and associated class probabilities. Later, Bochkovskiy et al. [24] updated the YOLO by utilizing new features: WRC (Weighted-Residual-Connections), CSP (Cross-Stage-Partial-connections), CmBN (Cross mini-Batch Normalization), SAT (Self-adversarial-training), Mish activation, Mosaic data augmentation, DropBlock regularization, and CIoU loss, and combining some of them to achieve remarkable results. Lin et al. [25] discovered that the extreme foreground-background class imbalance encountered during the training of dense detectors is the central cause. They proposed Focal Loss addressed this class imbalance by reshaping the standard cross entropy loss such that it down-weights the loss assigned to well-classified examples. Ren et al. [6] introduced a Region Proposal Network (RPN) that shares full-image convolutional features with the detection network, thus enabling nearly cost-free region proposals. Meanwhile, they merged RPN and Fast R-CNN [26] into a single network by sharing their convolutional features and proposed the Faster R-CNN object detection framework. He et al. [22] proposed a flexible, and general framework called Mask R-CNN for object instance segmentation. It extends Faster R-CNN by adding a branch for predicting an object mask in parallel with the existing branch for bounding box recognition.

Although these methods greatly improve the performance of the model on object detection, these methods are trained on large-scale datasets. When the training set is insufficient, the performance of the model cannot be effectively guaranteed.

C. FEW-SHOT LEARNING

Few-Shot Learning [8] is an example of meta-learning, where a learner is trained on several related tasks, during the meta-training phase, so that it can generalize well to unseen (but related) tasks with just few examples, during the meta-testing phase. An effective approach to the Few-Shot Learning problem is to learn a common representation for various tasks and train task-specific classifiers on top of this representation.

Apart from meta-learning, numerous methods are proposed focusing on few-shot learning in recent years. For example, Gao et al. [27] proposed a suite of techniques for fine-tuning language models on a small number of annotated examples, called LM-BFF (better few-shot fine-tuning of language models). LM-BFF consists of two parts: prompt-based fine-tuning together with a novel pipeline for automating prompt generation; and a refined strategy for dynamically and selectively incorporating demonstrations into each context. Sung et al. [28] proposed an end-to-end trainable framework, called the Relation Network (RN). During meta-learning, RN learns to learn a deep distance metric to compare a small number of images within episodes, each of which is designed to simulate the few-shot setting.

Few-shot object detection (FSOD) [29] is a typical application in few-shot learning. FSOD is about training a model on novel (unseen) object classes with little data, it still requires prior training on many labeled examples of base (seen) classes. A growing number of studies have been published to address this problem.

For example, Chen et al. [30] proposed a low-shot transfer detector (LSTD) to alleviate transfer difficulties in low-shot detection. LSTD leverages rich source-domain knowledge to construct an effective target-domain detector with very few training examples and integrates the advantages of both SSD and Faster RCNN in a unified deep framework To disentangle the learning of category-agnostic and category-specific components in a CNN based detection model, Wang et al. [5] proposed a framework that leverages meta-level knowledge about ''model parameter generation'' from base classes with abundant data to facilitate the generation of a detector for novel classes. Kang et al. [31] boosted the performance of FSOD by using a meta feature learner and a reweighting module within a one-stage detection architecture. The feature learner extracts meta features that are generalizable to detect novel object classes, using training data from base classes with sufficient samples. And the reweighting module transforms a few support examples from the novel classes to a global vector that indicates the importance or relevance of meta features for detecting the corresponding objects. Yan et al. [32] extended Faster /Mask R-CNN by proposing meta-learning over RoI (Region-of-Interest) features instead of a full image feature. This work disentangles multi-object information merged with the background, without bells and whistles, enabling Faster /Mask R-CNN turn into a meta-learner to achieve the tasks. Sun et al. [11] pointed out that object proposals with diverse intersection-of-union (IoR) score are enhanced variants of labeled objects. And they proposed FSCE to give deep model stronger ability to sense the differences between different proposals. These methods take full advantage of the input training data from different perspectives and achieve competitive performance. Take FSCE as an example. Based on the idea of contrastive learning, it constructs positive sample pairs with the detected regions of objects of the same category and their corresponding ground truth and constructs negative sample pairs with detected regions of different categories. It has achieved good performance in Few-shot object detection. Although the idea of this method can be flexibly applied to other problems, due to the unique data characteristics and processing flow of different problems, the method needs to be adjusted in the process of application, so some advanced methods might be complicated and difficult to apply to other problems.

III. METHODOLOGY

In this part, we demonstrate the detail of the proposed method. As shown in Fig. 1, we present our method in the following aspects: 1) introducing how to construct the data augmentation operation pool (DAOP); 2) building series and parallel connection rules to generate augmented variants; 3) developing the imposed semantic data augmentation using the deep features of augmented variants.

A. DAOP CONSTRUCTION

Data enhancement techniques can produce better generalization performance because the enhanced data generated by them can be used as supplementary data for model training. Since the experimental data is hard to obtain, in order to fully train the model, we construct a DAOP to enrich the data augmentation space. Inspired by [10], we build a finite set OP based on several common operations. So OP can be defined as Equation 1,

$$\begin{aligned} OP = \{&\varnothing, autocontrast, equalize, posterize, rotate,\\ &solarize, color, contrast, shear_x, shear_y,\\ &translate_x, translate_y, brightness, sharpness\} \end{aligned} \tag{1}$$

where $autocontrast, equalize, \cdots, sharpness$ are some widely-used image operation methods and control the underlying display properties of the image [17], [33]. The probability of each operation is equal. It should be noted that considering the proportion of clean images should be maintained. Therefore, the null operation $\varnothing$ is added in the DAOP, which can enhance the diversity of data. For example, there is a big difference between the images obtained by two non-null operations and the images obtained by three non-null operations.

B. SERIES AND PARALLEL CONNECTION

In order to generate higher diversity of augmented images, we build series and parallel connection scheme, as shown in Fig. 2.

For series connection, we select m ops from OP and used them to process input samples one after another. The procedure of series connection can be described as Equation 2,

$$x = op_m(op_{m-1}(\cdots op_i(\cdots op_1(x)))) \tag{2}$$

where op_i is randomly select from the DAOP OP, that is $op_i \in OP$. And x is the input sample from dataset D.

For parallel connection, We first generate n augmented variants in the series connection manner. Then we do a weighted blend of them, and the sum of the weights is 1. And we mix the vanilla image with the blended variant. The procedure of parallel connection can be described as Equation 3,

$$x = \lambda \sum_{j=1}^{n} (\beta_j x_{aug_j}) + (1 - \lambda) x_{ori} \tag{3}$$

where β_j is the weight of the augmented variant and $\sum \beta = 1$. x_{ori} is the vanilla image.

From these two connection methods, it can be found: these augmentation operations are sampled stochastically and layered to produce a high diversity of augmented images. These two connections allow us to generate diverse transformations, which are important for inducing robustness.

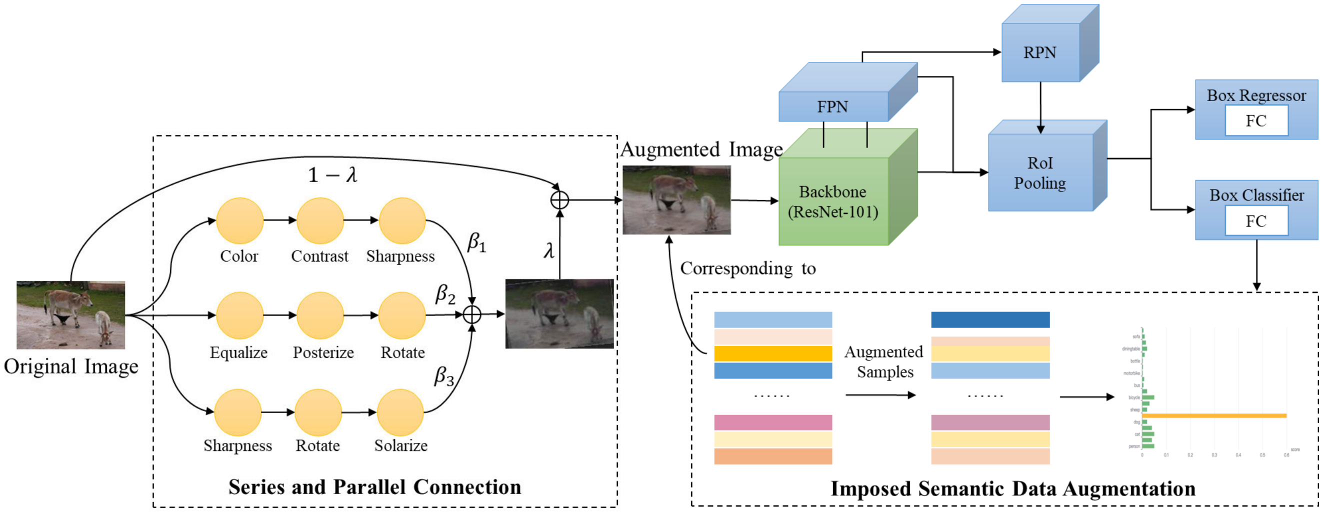


FIGURE 1. The process flow of DA-FSOD.

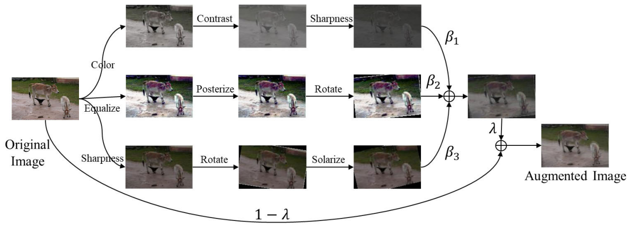


FIGURE 2. Example of series and parallel connection.

Meanwhile, preserving the information of the original image effectively controls the degree of image shift during the augmentation process, so that the model can fully learn the features in the data during the data augmentation process. Additionally, previous methods have attempted to increase diversity by directly composing augmentation primitives in a chain, but this can cause the image to quickly degrade and drift off the data manifold. Such image degradation can be mitigated and the augmentation diversity can be maintained by mixing together the results of several augmentation chains in convex combinations. A concrete account of the algorithm is given in the pseudocode below.

Algorithm 1 Series and Parallel Connection

```
Require: Image $x_{ori}$, Weight Array $\beta$, Fusion Coefficient $\lambda$,
    Series Coefficient $m$, Parallel Coefficient $n$, DAOP $OP = \{\varnothing, autocontrast, equalize, \cdots, sharpness\}$
Ensure: $\sum_{i=1}^{n} \beta_i = 1, \lambda \in [0, 1], OP \neq \varnothing$
    Fill $x_{aug}$ with zeros
    for $i \leftarrow 1$ to $n$ do
        Sample operations $op_1, \cdots, op_m \sim OP$
        $x_{aug} += \beta_i \cdot op_m(op_{m-1}(\cdots op_i(\cdots op_1(x_{ori}))))$
    end for
    $x_{out} = \lambda \cdot x_{aug} + (1 - \lambda) \cdot x_{ori}$
    return $x_{out}$
```

C. IMPOSED SEMANTIC DATA AUGMENTATION

After constructing DAOP and using series and parallel connection scheme to generate augmentated data, the complexity of the data has been significantly improved. However, it can be found from the design of the method that a double loop is introduced in the process of constructing the enhanced image through series and parallel connection. *m* and *n* jointly control the diversity of augmented images. In order to reduce additional computing overhead and improve the execution efficiency of the algorithm, in addition to constructing explicit data enhancement, we perform implicit data enhancement through high-dimensional information of deep features.

Deep networks have been known to excel at extracting high-level representations in the deep feature space, where the semantic relationships between samples can be captured by the spatial positions of their deep features [34]. It has been shown that translating deep features along specific directions corresponds to performing meaningful semantic transformations on input images [35]. Due to the small amount of training data, there is few meaningful semantic augmentation information that can be provided directly. Therefore, inspired by [20], we propose to augment the training set semantically via augmented deep features, that is, imposed semantic data augmentation.

Specifically, the samples of each category have their own characteristic distribution within the category. This data distribution implies the possible direction of change of this type of data. For this reason, we construct a zero-mean Gaussian distribution for each class by counting the intra-class covariance matrix of each class based on the augmented data, and then sample meaningful semantic transformation directions from them for data amplification in their respective categories, so as to approximate the manual labeling process and achieve a good balance between correctness, efficiency, and diversity. Which is defined as Equation 4,

$$\hat{a} \sim \mathcal{N}(a_{aug}, \alpha \sum\nolimits_{y}) \tag{4}$$

where a_{aug} is feature of the augmented input. y is the label of its corresponding input. $\sum_y$ is the covariance matrices. α is the coefficient factor. $\hat{a}$ is sampled result.

Instead of cross entropy loss, we use ISDA loss [20] to update the model based on $\hat{a}$, which is shown in Equation 5,

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N} -log(\frac{e^{\omega_{y_i}^{T}\hat{a}_i}}{\sum_{j=1}^{c} e^{\omega_j^{T}\hat{a}_i + \frac{1}{2}\alpha(\omega_j - \omega_{y_i})^{T}\sum_{y_i}(\omega_j - \omega_{y_i})}}) \tag{5}$$

where N is the total number of images in the mini-batch. C is the amount of class. $W = [\omega_1, \cdots, \omega_C]^T$ are the weight matrix corresponding to the final fully connected layer.

D. EXPERIMENTAL SETTING

1) DATASET

We followed the methods of Xing et al. [29] to carry out our expreiments. Extensive experiments are performed on the PASCAL VOC [23] benchmark. The data set consists of 20 categories, divided into 15 base categories and 5 novel categories. All base category data from the Pascal VOC 2007 + 2012 training set were considered available and k = 1,2,3,5, and 10 for randomly sampling novel instances from novel categories not previously seen. The same partitions of base and novel categories are used in this paper, which are referred to as novel splits 1, 2, and 3. [31] In addition, the average precision at 50 (AP50) of the novel predictions (nAP50) is selected as the evaluation indicator of the model performance on PASCAL VOC 2007 testing dataset. It is worth noting that since the training set contains a relatively small number of images, different choices for the training set may make the training set for the same task different. In order to reduce the influence caused by this difference, the average of the results of multiple independent repeated experiments is used as the final experimental results.

2) TRAINING SETTING

We use Faster-RCNN [6] as our basic detection model. We select ResNet-101 and feature pyramid network as the feature extraction network [36]. We use all base category data in PASCAL VOC 2007+2012 to carry out the data-abundant training stage. Then the model parameters trained by this stage are kept and used to initialize the network to fulfill the second training stage, which is the fine-tuning stage. Furthermore, we set the m in Equation 2 to 3, and set n and λ in Equation 3 to 3 and 0.4 as default. All experiments are carried out on 4 GPUs with 24GB of memory each. The batch size is set to 16 and the experiments are implemented based on PyTorch-1.0.1 and CUDA-10.0.

E. QUANTITATIVE RESULTS

The comparison results obtained for all three random novel splits from the PASCAL VOC dataset are shown in Table 1. Our method is significantly superior to all existing works with any number of shots and any splits. Hence, the validity of the proposed method is fully proved. Actually, we are the first work that has achieved nAP50 results surpassing 60% on novel split 1. Moreover, the average improvements of our method obtains compared to FSCE [11], which is our strong baseline, on three novel splits are 0.48%, 0.76%, and 1.96%. The standard deviations of each average improvement are 0.36, 0.26, and 0.92. These results illustrate that our method achieves better performance than SOTA methods. At the same time, compared with the improvement of the three novel splits, the improvement of our method on the second novel splits is more stable, and the improvement on the third novel splits is more obvious. More specifically, on novel split 1, our method obtains 0.5%, 1.1%, 0.3%, 0.2%, and 0.3% betterment under the 1-shot, 2-shot, 3-shot, 5-shot, and 10-shot cases, respectively. On novel split 2, our method achieve 0.5%, 0.8%, 1.1%, 0.9% and 0.5% improvements and achieves values of 1.9%, 2.7%, 2.8%, 1.9% and 0.5% on novel split 3, respectively. It is worth mentioning that there are obvious differences in the detection performance of models under different categories. It can be regarded that the recognition sensitivity of models to different objects is different. This is a very interesting finding. By studying the relationship between different objects, objects, and the background, it has the potential to provide guidance for the fine-grained recognition of the deep model.

TABLE 1. Performance evaluation (nAP 50) of existing FSOD methods on three PASCAL VOC Novel Splits.

Method / Shot	Backbone	Novel Split 1					Novel Split 2					Novel Split 3				
		1	2	3	5	10	1	2	3	5	10	1	2	3	5	10
LSTD [30]	VGG-16	8.2	1.0	12.4	29.1	38.5	11.4	3.8	5.0	15.7	31.0	12.6	8.5	15.0	27.3	36.3
YOLOv2-ft [5]	YOLO V2	6.6	10.7	12.5	24.8	38.6	12.5	4.2	11.6	16.1	33.9	13.0	15.9	15.0	32.2	38.4
FSRW [31]		14.8	15.5	26.7	33.9	47.2	15.7	15.3	22.7	30.1	40.5	21.3	25.6	28.4	42.8	45.9
MetaDet [31]		17.1	19.1	28.9	35.0	48.8	18.2	20.6	25.9	30.6	41.5	20.1	22.3	27.9	41.9	42.9
RepMet [37]	InceptionV3	26.1	32.9	34.4	38.6	41.3	17.2	22.1	23.4	28.3	35.8	27.5	31.1	31.5	34.4	37.2
FRCN-ft [31]	FRCN-R101	13.8	19.6	32.8	41.5	45.6	7.9	15.3	26.2	31.6	39.1	9.8	11.3	19.1	35.0	45.1
FRCN+FPN-ft [9]		8.2	20.3	29.0	40.1	45.5	13.4	20.6	28.6	32.4	38.8	19.6	20.8	28.7	42.2	42.1
MetaDet [31]		18.9	20.6	30.2	36.8	49.6	21.8	23.1	27.8	31.7	43.0	20.6	23.9	29.4	43.9	44.1
Meta R-CNN [32]		19.9	25.5	35.0	45.7	51.5	10.4	19.4	29.6	34.8	45.4	14.3	18.2	27.5	41.2	48.1
TFA [9]		25.3	36.4	42.1	47.9	52.8	18.3	27.5	30.9	34.1	39.5	17.9	27.2	34.3	40.8	45.6
FSIW [38]		24.2	35.3	42.2	49.1	57.4	21.6	24.6	31.9	37.0	45.7	21.2	30.0	37.2	43.8	49.6
FSCE [11]		32.9	44.0	46.8	52.9	59.7	23.7	30.6	38.4	43.0	48.5	22.6	33.4	39.5	47.3	54.0
Ours		33.4	45.1	47.1	53.1	60.0	24.2	31.4	39.5	43.9	49.0	24.5	36.1	42.3	49.2	54.5

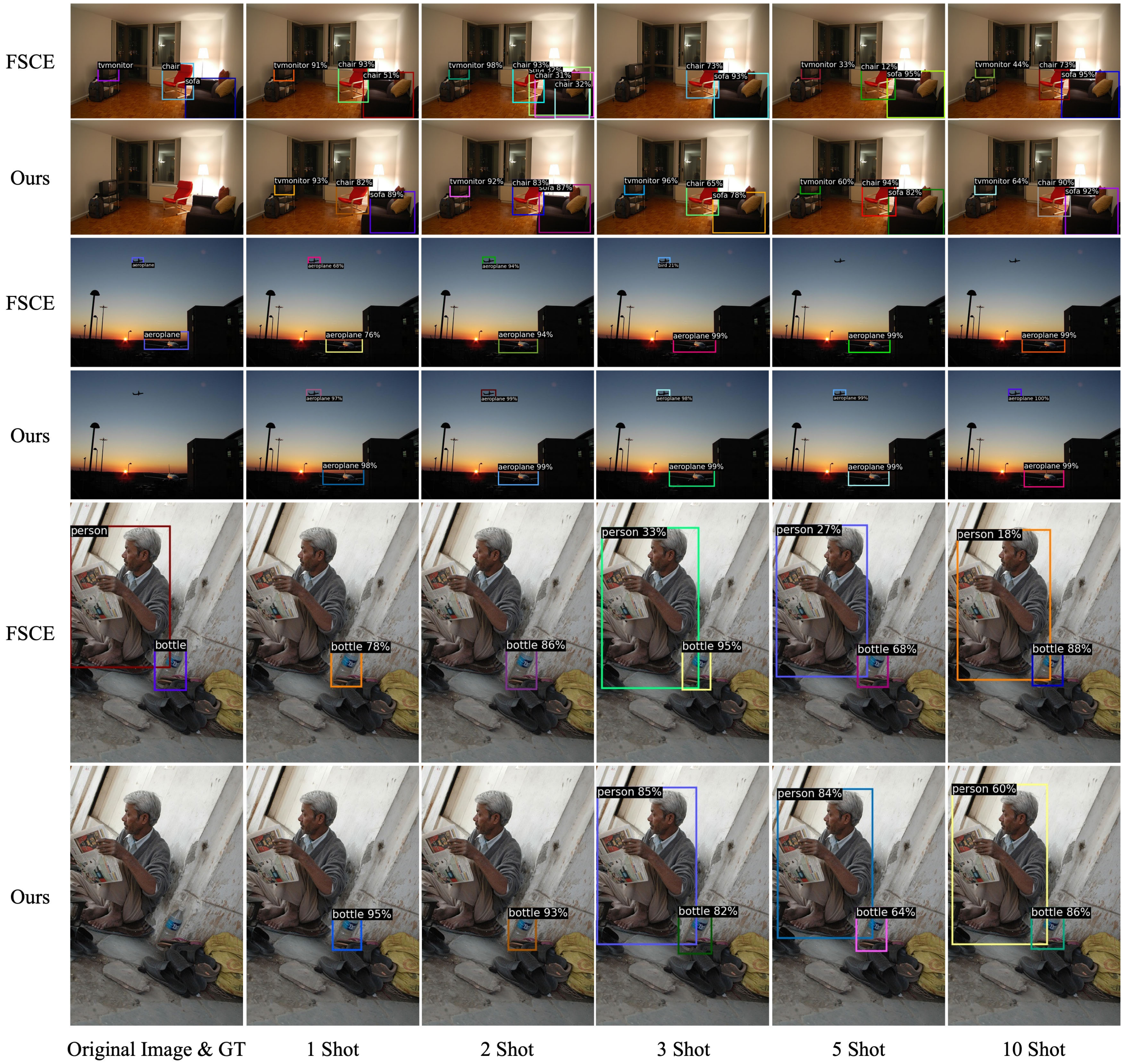


FIGURE 3. Visualization of the detection results.

We visualize the test results of the proposed method and FSCE. As shown in Fig.3, the first column represents the original images and their GT bounding box. The remaining columns represent the detection results under different shots. Moreover, as labeled on the left side of Fig.3, odd rows are the result of FSCE, and even rows are the result of the proposed method. It can be seen from the visualization results that with the increase in the number of shots, the recognition score and detection efficiency of the two methods gradually improve. However, when the number of shots is low, there appear some duplication detection problems and missing detection problems for some inputs whose context is complex. Moreover, when the input sample contains multiple targets, the detection efficiency and stability of the proposed method is remarkably better than that of FSCE.

For example, for the first sample, when the number of shots equals to 2, FSCE detects the tvmonitor successfully. However, it generates several extra areas and recognizes them as chairs. In addition, when the number of shots increases to 3, FSCE misses the tvmonitor. The same phenomenon appears in the second sample whose the numbers of the shots are 5 and 10. For the third test sample, the detection behavior seems the same for the FSCE and ours. While the confidence scores generated by our method are much better than that of FSCE.

During the testing process, a large number of detection areas will be generated, which will be filtered through the

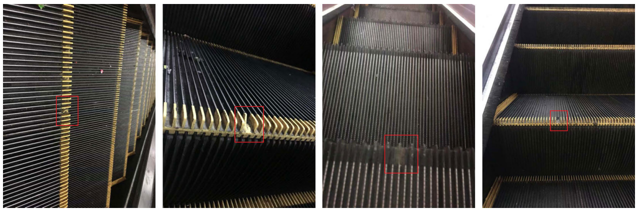


FIGURE 4. Samples of Elevator images.

IoU threshold, and the remaining areas will be classified and scored to produce the final detection result. It can be seen that the occurrence of the above detection problems indicates that the detector did not effectively correct the deviation area in the training process. Moreover, when the number of input images used for training remains low, the feature recognition capability of the model is limited. It is difficult for the model that is trained under such strict conditions to generate accurate activation responses for all objects in the image. According to the design idea of the proposed method, it can be found that our method effectively enhances the input data through the DAOP as well as the series and parallel connection scheme before the model training. Moreover, this data enhancement process does not affect each other between each epoch of the model training. Therefore, from the perspective of input data, the proposed method effectively expands the complexity of input data. On the other hand, we have utilized the deep feature information to sample new features which will replace the vanilla feature to fulfill the training process. As the input data will be modified during each training round, their features in the model will also be diverse, so the newly generated features will be disparate accordingly. In general, during the training process, the data utilization efficiency of the deep neural network is improved by the proposed method. According to this, it can be seen that our training model is more effective in making full use of limited training data.

F. ABLATION STUDY

In order to verify the generalization performance of the detection method proposed in this chapter in the real scene and verify the role of each module in improving the performance, we collected 1020 abnormal images of escalator steps in different environments in the actual scene of the passenger station and then conducted data cleaning and pre-processing operations. In the end, 600 escalator-level defect images were retained as a data set and divided into a training set and a test set according to a ratio of 8:2. We used the labelme annotation tool to manually annotate the sample images of the dataset.

We conducted three sets of comparative experiments to verify the validity of the method from two aspects: model input and semantic information. Each group of experiments was trained with the same number of samples. The test results under different module combinations are shown in Table 2.

From the experimental results in the first row and the third row in Table 2, it can be seen that the Series-Parallel Connection module brings a 2.4% accuracy improvement to the model, which is the most significant improvement in accuracy compared with another module, which effectively proves that the importance of data augmentation in Few-Shot Object Detection Tasks.

TABLE 2. Performance evaluation (nAP 50) of existing FSOD methods on three PASCAL VOC Novel Splits.

Series-Parallel Connection	Imposed Semantic Data Augmentation	nAP50
✓	×	60.4
×	✓	60.1
✓	✓	62.5

In addition, the visualization results of the Series-Parallel Connection module are shown in Figure 5. The first image is the original input picture, and the rest three pictures are augmented variants generated based on the original image. It can be seen that the defects of the escalator can be clearly revealed by using some augmentation operations, which proves the effectiveness of the proposed module.

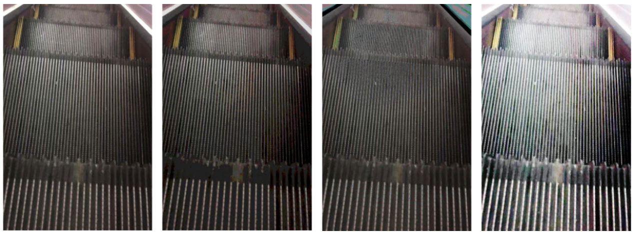


FIGURE 5. The visualization results of Series-Parallel Connection module. The first image is the original image and the rest three are its augmented variants.

Then, we input the enhanced samples into the network for testing, and visualize the detection results, as shown in Figure 6. As can be seen in Figure 6, the red bounding box is the ground truth, the green bounding box is the detection result after using the series-parallel connection module, and the yellow bounding box is the result without it. From the visualization results, it can be found that the green detection result of the data augmentation module has improved both in terms of confidence and coincidence. Experimental results show that DAOP significantly improves data richness. Since the escalator defect data has similar scene characteristics, using DAOP can introduce scene variables to effectively avoid model training overfitting. In addition, data enhancement can also provide some training samples containing occlusion and blurring, thereby further improving the robustness of the network model.

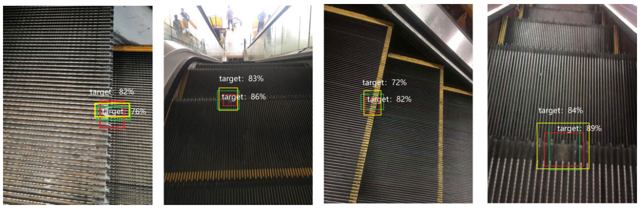


FIGURE 6. Comparison of introducing the data augmentation module.

On the other hand, from the experimental results in the second and third rows of Table 2, it can be seen that by introducing the imposed semantic data augmentation module, the model achieves a 2.1% accuracy improvement, which further proves the effectiveness of data augmentation strategy based on semantic information. Figure 7 shows the impact of the semantic information on the detection accuracy, where the red bounding represents the ground truth, the green box represents the detection result enhanced with the semantic information, and the yellow box represents the result without semantic information. It can be seen from the results in Figure 7 that after introducing the semantic information, the green bounding boxes are still significantly better than the results of the yellow bounding boxes in terms of confidence and coincidence, which proves that the imposed semantic data augmentation module designed in Section III enables the transformation of deep information along the direction of the semantic gradient, which effectively improves the richness of deep semantics, thereby further boosting the few-shot object detection performance of the model.

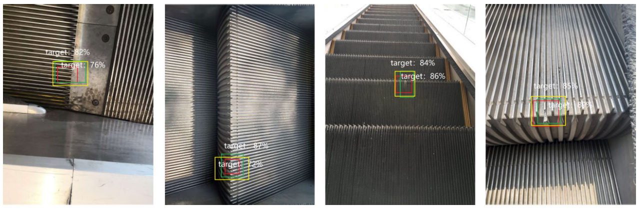
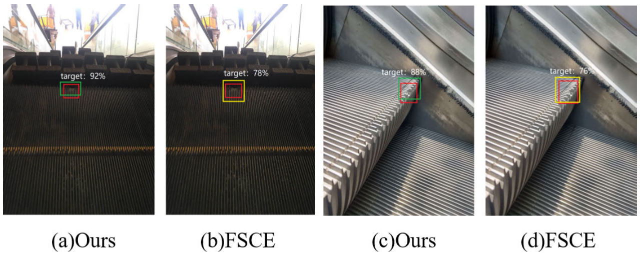


FIGURE 7. Comparison of introducing semantic loss function.

Based on the experimental results in Table 2, the two proposed modules have brought performance improvements to few-shot object detection.

In addition, we compared our method with the FSCE in the actual scene. The detection results are shown in Figure 8, where the red bounding boxes are the ground truth, the green bounding boxes are the detection results of the proposed method, and the yellow bounding boxes are the detection results of FSCE. The experimental results show that in the passenger station escalator samples, compared with the FSCE method, our method can detect escalator defects with higher confidence, which further verifies that the method has good adaptability and practical significance in object detection tasks of the actual scene.



(a)Ours (b)FSCE (c)Ours (d)FSCE

FIGURE 8. Comparison of the detection results of ours and FSCE.

IV. LIMITATION AND FUTURE WORK

With the continuous development of technology, deep learning techniques have been applied to various tasks. Among them, object detection is an essential application, which can be applied to miscellaneous scenarios, like moving target detection in the video monitoring system. Most recent detection frameworks [2], [5] are based on large amounts of data and complex modules. As a result, it usually takes a long time and a large amount of data to obtain a well-trained detection model. As deep learning technology continues to penetrate into all walks of life, the research problems become more and more refined. However, such granular data is usually not readily available in large quantities, which leads to model underfitting. In order to address this issue, we investigate a framework for deep object detection models with stronger performance and effectiveness by enhancing the data augmentation space and utilizing the deep feature of the input sample.

Although most of our experimental results show that compared with many SOTA methods, the proposed method can continuously improve performance, we also observe a limitation that might be associated with hyperparameters in series and parallel connections. Currently, the setting of these hyperparameters mainly relies on experimental experience. Sometimes finding a proper set of them could be time-consuming in a different scenario. Thus, it becomes necessary to design an end-to-end trainable framework that allows the network to determine the width and height of this connection scheme.

To the best of our knowledge, this is the first study that integrates the semantic information of deep features and combines them with data augmentation technology to focus on few-shot object detection. DA-FSOD improves the performance of deep neural network and can be embed to other related works that large amount of training data is hard to obtain, such as real-time target detection [39], multi-target vehicle detection and tracking [40], etc. These complex decision systems usually need to process multi-source data collected by different modules of sensors and carry out multi-stage analysis of these data. It is difficult to fully cover a real test environment through a limited data acquisition process. The proposed method has the potential to be applied to these scenarios as it can enhance the training data implicitly and facilitate the activation responses of features in the network.

V. CONCLUSION

In this paper, we propose a novel data augmentation scheme, that is DA-FSOD, to boost the data effectiveness of few-shot object detection. Specifically, we build a data augmentation operation pool based on several widely-used image process operations to enrich the data augmentation space. Then we propose the series and parallel connection scheme, which superimposes the effects of various operations and is able to generate more diverse augmented variants while maintain the core feature of original input image. To further explore the deep feature information, we utilize the semantic information of the input image and propose imposed semantic data augmentation. According to the experimental results, our method outperforms some typical SOTA methods in the domain of few-shot object detection. Moreover, based on our ablation studies, our method can obtain a better performance with an ordinary setting of the series and parallel connection, which illustrates the additional overhead of the proposed method is weeny compared to the performance gain.

REFERENCES

[1] J. Yao, D. Wang, H. Hu, W. Xing, and L. Wang, "ADCNN: Towards learning adaptive dilation for convolutional neural networks," *Pattern Recognit.*, vol. 123, Mar. 2022, Art. no. 108369.

[2] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot MultiBox detector," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 21–37.

[3] X. Zenggang, T. Zhiwen, C. Xiaowen, Z. Xue-Min, Z. Kaibin, and Y. Conghuan, "Research on image retrieval algorithm based on combination of color and shape features," *J. Signal Process. Syst.*, vol. 93, nos. 2–3, pp. 139–146, Mar. 2021.

[4] R. Sharma and A. Sungheetha, "An efficient dimension reduction based fusion of CNN and SVM model for detection of abnormal incident in video surveillance," *J. Soft Comput. Paradigm*, vol. 3, no. 2, pp. 55–69, May 2021.

[5] Y.-X. Wang, D. Ramanan, and M. Hebert, "Meta-learning to detect rare objects," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9924–9933.

[6] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 1–12.

[7] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *Proc. Int. Conf. Learn. Represent.*, 2014, pp. 1–13.

[8] A. Mehra and J. Hamm, "Penalty method for inversion-free deep bilevel optimization," in *Proc. Asian Conf. Mach. Learn.*, 2021, pp. 347–362.

[9] X. Wang, T. E. Huang, T. Darrell, J. E. Gonzalez, and F. Yu, "Frustratingly simple few-shot object detection," 2020, *arXiv:2003.06957*.

[10] D. Hendrycks, N. Mu, E. D. Cubuk, B. Zoph, J. Gilmer, and B. Lakshminarayanan, "AugMix: A simple data processing method to improve robustness and uncertainty," in *Proc. Int. Conf. Learn. Represent.*, 2020, pp. 1–12.

[11] B. Sun, B. Li, S. Cai, Y. Yuan, and C. Zhang, "FSCE: Few-shot object detection via contrastive proposal encoding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 7348–7358.

[12] S. Y. Feng, V. Gangal, J. Wei, S. Chandar, S. Vosoughi, T. Mitamura, and E. Hovy, "A survey of data augmentation approaches for NLP," 2021, *arXiv:2105.03075*.

[13] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *J. Big Data*, vol. 6, no. 1, pp. 1–48, Dec. 2019.

[14] T. DeVries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," 2017, *arXiv:1708.04552*.

[15] H. Zhang, M. Cissé, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–13.

[16] S. Yun, D. Han, S. Chun, S. J. Oh, Y. Yoo, and J. Choe, "CutMix: Regularization strategy to train strong classifiers with localizable features," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6022–6031.

[17] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "RandAugment: Practical automated data augmentation with a reduced search space," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 3008–3017.

[18] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," 2019, *arXiv:1904.08779*.

[19] J. Wei and K. Zou, "EDA: Easy data augmentation techniques for boosting performance on text classification tasks," 2019, *arXiv:1901.11196*.

[20] Y. Wang, G. Huang, S. Song, X. Pan, Y. Xia, and C. Wu, "Regularizing deep networks with semantic data augmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 7, pp. 3733–3748, Jul. 2022.

[21] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[22] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[23] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Jun. 2010.

[24] A. Bochkovskiy, C.-Y. Wang, and H.-Y. Mark Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.

[25] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2999–3007.

[26] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[27] T. Gao, A. Fisch, and D. Chen, "Making pre-trained language models better few-shot learners," 2020, *arXiv:2012.15723*.

[28] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. S. Torr, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1199–1208.

[29] W. Xing, J. Yao, Z. Liu, W. Liu, S. Zhang, and L. Wang, "Contrastive JS: A novel scheme for enhancing the accuracy and robustness of deep models," *IEEE Trans. Multimedia*, early access, Dec. 26, 2022, doi: 10.1109/TMM.2022.3232030.

[30] H. Chen, Y. Wang, G. Wang, and Y. Qiao, "LSTD: A low-shot transfer detector for object detection," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, 2018, pp. 1–8.

[31] B. Kang, Z. Liu, X. Wang, F. Yu, J. Feng, and T. Darrell, "Few-shot object detection via feature reweighting," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8419–8428.

[32] X. Yan, Z. Chen, A. Xu, X. Wang, X. Liang, and L. Lin, "Meta R-CNN: Towards general solver for instance-level low-shot learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9576–9585.

[33] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, "AutoAugment: Learning augmentation policies from data," 2018, *arXiv:1805.09501*.

[34] Y. Bengio, G. Mesnil, Y. Dauphin, and S. Rifai, "Better mixing via deep representations," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 552–560.

[35] P. Upchurch, J. Gardner, G. Pleiss, R. Pless, N. Snavely, K. Bala, and K. Weinberger, "Deep feature interpolation for image content changes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6090–6099.

[36] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 936–944.

[37] L. Karlinsky, J. Shtok, S. Harary, E. Schwartz, A. Aides, R. Feris, R. Giryes, and A. M. Bronstein, "RepMet: Representative-based metric learning for classification and few-shot object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5192–5201.

[38] Y. Xiao and R. Marlet, "Few-shot object detection and viewpoint estimation for objects in the wild," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 192–210.

[39] J. Yun, D. Jiang, Y. Liu, Y. Sun, B. Tao, J. Kong, J. Tian, X. Tong, M. Xu, and Z. Fang, "Real-time target detection method based on lightweight convolutional neural network," *Frontiers Bioeng. Biotechnol.*, vol. 10, Aug. 2022, Art. no. 861286.

[40] K. Zhang, H. Ren, Y. Wei, and J. Gong, "Multi-target vehicle detection and tracking based on video," in *Proc. Chin. Control Decis. Conf. (CCDC)*, Aug. 2020, pp. 3317–3322.

JIAN YAO received the B.S. and M.S. degrees from the School of Mathematical Sciences, Shanxi University, in 2012 and 2015, respectively. He is currently pursuing the Ph.D. degree with the China Academy of Railway Sciences. From 2015 to 2020, he was a Research Assistant with the Institute of Computing Technology, CARS. His current research interests include computer vision, image processing, and intelligent railway passenger stations.

TIANYUN SHI received the Ph.D. degree from the School of Automation, Beijing Institute of Technology, in 1998. He was the Ph.D. Supervisor and became a member of CCF, in 2006, and CAAI, in 2013. Currently, he is the Director of the Department of Science and Technology and Information Technology, China Academy of Railway Sciences Corporation Ltd. His current research interests include artificial intelligence applications, computer vision, intelligent railways, and intelligent railway passenger stations.

XIAOPING CHE (Member, IEEE) received the B.S. degree in network engineering from the Beijing University of Posts and Telecommunications, China, in 2009, the M.S. degree in computer and communication networks from Telecom SudParis, France, in 2011, and the Ph.D. degree from Institute MinesTelecom/Telecom SudParis, in 2014. From 2011 to 2014, he was with the French CNRS Laboratory SAMOVAR. He is currently an Associate Professor with the School of Engineering, Beijing Jiaotong University. His current research interests include virtual reality, user experience, software testing, and crowd sensing. He is a member of ACM.

JIE YAO received the B.S. and Ph.D. degrees in software engineering from Beijing Jiaotong University, in 2016 and 2022, respectively. From 2019 to 2020, he was a Visiting Student with the University of Central Florida. Currently, he is a Faculty Member with the School of Information Management, Beijing Information Science and Technology University. His current research interests include image processing, computer vision, and deep model robustness.

LIUYI WU received the M.S. degree from the School of Traffic and Transportation, Beijing Jiaotong University, in 2018. Currently, she is a Scientific Researcher with the Institute of Computing Technology, China Academy of Railway Sciences Corporation Ltd. Her current research interests include artificial intelligence applications and transportation planning.