TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 20/28 pp1237–1250
DOI: 10.26599/TST.2024.9010031
Volume 30, Number 3, June 2025

Few-Shot Object Detection via Dual-Domain Feature Fusion and Patch-Level Attention

Guangli Ren, Jierui Liu, Mengyao Wang, Peiyu Guan*, Zhiqiang Cao, and Junzhi Yu

Abstract: Few-shot object detection receives much attention with the ability to detect novel class objects using limited annotated data. The transfer learning-based solution becomes popular due to its simple training with good accuracy, however, it is still challenging to enrich the feature diversity during the training process. And fine-grained features are also insufficient for novel class detection. To deal with the problems, this paper proposes a novel few-shot object detection method based on dual-domain feature fusion and patch-level attention. Upon original base domain, an elementary domain with more category-agnostic features is superposed to construct a two-stream backbone, which benefits to enrich the feature diversity. To better integrate various features, a dual-domain feature fusion is designed, where the feature pairs with the same size are complementarily fused to extract more discriminative features. Moreover, a patch-wise feature refinement termed as patch-level attention is presented to mine internal relations among the patches, which enhances the adaptability to novel classes. In addition, a weighted classification loss is given to assist the fine-tuning of the classifier by combining extra features from FPN of the base training model. In this way, the few-shot detection quality to novel class objects is improved. Experiments on PASCAL VOC and MS COCO datasets verify the effectiveness of the method.

Key words: few-shot object detection; dual-domain feature fusion; patch-level attention

1 Introduction

In recent years, deep neural networks (DNN) have made significant progresses across a wide range of visual tasks such as image classification[1] and object detection[2–4], thanks to the large-scale labeled datasets. In the real world, the number of object classes is very huge, and it is challenging to construct the large-scale labeled dataset for all concerned object classes. Particularly in the field of robotics, the robot inevitably encounters new unknown objects outside existing public datasets. It is not an effective way to collect abundant data for each object class. Without adequate training data, the training effect becomes weak. This drives researchers to concern the few-shot object detection (FSOD), which detects a novel class object with a few samples. The network is firstly well trained on large-scale samples of base classes and then it is further learned with a few samples of novel classes. Nowadays, FSOD has become a research hotspot.

Mainstream FSOD methods are usually divided into

• Guangli Ren, Jierui Liu, Mengyao Wang, Peiyu Guan, and Zhiqiang Cao are with State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and also with School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China. E-mail: renguangli2018@ia.ac.cn; liujierui2019@ia.ac.cn; wangmengyao2023@ia.ac.cn; guanpeiyu2017@ia.ac.cn; zhiqiang.cao@ia.ac.cn.
• Junzhi Yu is with Department of Advanced Manufacturing and Robotics, Peking University, Beijing 100871, China. E-mail: junzhi.yu@ia.ac.cn.
* To whom correspondence should be addressed.

Manuscript received: 2023-11-13 ; revised: 2024-01-23; accepted: 2024-02-01

meta-learning-based[5] and transfer learning-based[6, 7] methods. The former trains the network with an episodic set of tasks to learn task-level meta-information and adapt to novel classes. In this solution, different meta-learners are used to extract meta-information for good generalization of models on novel class objects, where the meta-learners are built by means of weight generation[8] or feature reweighting[9, 10]. Different from the meta-learning-based methods with episodic training, the transfer learning-based solution firstly utilizes abundant data of base classes to train the network, and then the network is fine-tuned with a few samples of novel classes. Through the two-stage training, the knowledge and feature representations of the base classes are transferred to novel classes. The two-stage fine-tuning approach (TFA)[11] is a milestone method, which improves detection performance by simply freezing the backbone of network trained with base classes and fine-tuning the detector head to adapt to the novel objects. On the basis of TFA, a series of schemes, such as multi-scale refinement[12], contrastive learning[13], task separation[14], and gradient optimization[15], are presented to enhance the detection performance with better feature discrimination.

Notice that the backbone is usually initialized through a pre-trained ResNet-101 on ImageNet[16] with a large amount of object classes, which implies diverse features. However, the features tend to base-biased after base training, which reduces the feature diversity and thus affects the generalization on novel classes. How to retain the feature diversity is a key to improve the performance. Another noteworthy aspect is the valuable fine-grained features through feature refinement. Herein, attention mechanisms are preferable and it enables the network to focus on regions of interest. The attentions mainly include non-local and channel attentions, which respectively concern pixel-level or channel-level relation. It is worth noting that humans perceive objects in a more compositional manner, considering them as a combination of meaningful parts or components. The part-based perspective enhances the robustness and reliability to recognize objects. Constructing the part-level attention provides a promising direction to improve the semantic feature and thus promote the performance of FSOD.

The aforementioned analyses motivate us to build advanced few-shot object detection network. The main contributions are summarized as follows:

(1) A novel few-shot object detection method via dual-domain feature fusion and patch-level attention (DFFPA) is proposed. A base domain stream and an elementary domain stream are organized in a parallel way to enrich the feature diversity. The dual-domain features complement each other and feature pairs with the same size are respectively fused to extract discriminative features. Further, the feature map after region of interest (ROI) pooling is refined in a unit of patch and the relations among patches are mined. As a result, the adaptability to novel classes is enhanced.

(2) An elementary domain is superposed on the original base domain and more category-agnostic features are preserved for diversity of features. To better integrate various features, a dual-domain feature fusion (DFF) scheme is designed. By applying channel attention on the feature map of each domain, channel feature confidence scores are generated, which enables to perform adaptive weighting of channel-wise features. Combined with alternative mask between weighted features from distinct domains, valuable features are exploited, which are further fused by convolution-based channel adjustment. This strengthens the feature representation for recognition under complex scenarios.

(3) To make the network pay more attention to the crucial object parts in the feature map of ROI pooling, a patch-level attention (PA) mechanism is proposed. PA segments the feature map into multiple patches, which are further concatenated along channel dimension. On this basis, the channel attention is introduced to mine the internal relations among patches and the importance of patch is then adjusted through feature reweighting. The resultant refined features are fed into the detector head for effective prediction.

(4) A weighted classification loss is designed to assist the training of the classifier by leveraging extra features from FPN of the base training model into ROI pooling to synchronously predict the classification during the fine-tuning process. Experiments on PASCAL visual object classes (PASCAL VOC) and Microsoft common objects in context (MS COCO) verify the effectiveness of the proposed method.

The remainder of this paper is organized as follows. Section 2 presents the related work. Section 3 details the proposed method. The experiments are given in

Section 4 and Section 5 concludes the paper.

2 Related Work

In this section, we address the related work from two aspects: few-shot object detection and attention mechanism.

2.1 Few-shot object detection

Few-shot learning refers to generalize model from limited training data[17], which has attracted considerable attention in computer vision. As a typical task of few-shot learning, FSOD is dedicated to training a model to achieve simultaneous object classification and localization with few labelled samples of novel classes. To satisfy the limitation of novel class training samples, a simple strategy is to generate extra synthetic samples through data augmentation including geometric and pixel-level transformations, which directly provides more samples to alleviate overfitting. Xu et al.[18] proposed a few-shot object detection via sample processing (FSSP), which imposes background sparsity, multi-scale replication, and random clipping to enrich the scale distribution of few training samples. With these operations, each input image is augmented into multiple images with different scales for better object detection. However, such augmentation mainly focuses on internal variations of existing samples and it is impossible to cover all scenarios. In recent years, meta-learning-based and transfer learning-based solutions become the mainstream of FSOD. Meta-learning, also known as learning to learn, aims to enable models to acquire the ability to learn. A meta-learner is constructed to acquire meta-knowledge from an episodic set of tasks, which promotes the adaptability to new tasks. On the basis of one-stage detector you only look once version 2 (YOLOv2)[19], Kang et al.[10] integrated a meta feature learner and a reweighting module to generalize detection from base classes to novel ones. The reweighting module transforms a few support examples from the novel classes to corresponding global vectors, which are then used to reweight the meta features through channel-wise multiplication, enhancing the accuracy of detection. Wang et al.[8] introduced a meta-learning detection method MetaDet based on Faster region-based convolutional networks (R-CNN)[20]. It leverages the meta-level knowledge from base dataset to train a weight prediction meta-model, which achieves the meta-knowledge transfer from few-shot to category-specific parameters. Meta R-CNN[9] is presented for FSOD, where the R-CNN predictor head of Faster R-CNN is replaced using a predictor-head remodeling network (PRN). The PRN infers class-attentive vectors, which are utilized for channel-wise feature selection on ROI features, facilitating the detection or segmentation of novel class objects.

Different from the aforementioned solutions, transfer learning-based methods achieves FSOD of novel classes by transferring knowledge from base classes with abundant training data. In Ref. [21], Lu et al. analyzed the disagreement between classification and location, where the former emphasizes translational invariance and the latter concerns translational variance. Upon this insight, a decoupled metric network termed as DMNet is proposed based on single shot detector (SSD)[22]. By designing a two-branch representation transformation, the features are decoupled for classification and location, respectively. More works are rooted in two-stage Faster R-CNN framework. The two-stage fine-tuning approach (TFA)[11] is designed from a finding that fine-tuning on the part of detector is crucial for the performance, which facilitates the development of transfer learning-based methods. Sun et al. proposed a few-shot detection method based on contrastive proposal encoding (FSCE)[13], where contrastive learning is applied to promote instance level intra-class compactness and inter-class variance for better feature embedding. To handle the problem of scale variations in FSOD caused by the unique sample distribution, a multi-scale positive sample refinement (MPSR)[12] method is proposed. Considering the different feature preferences of classification and localization, Liu et al.[14] proposed an adaptive fully-dual network (AFD-Net). It uses dual query encoder and dual attention generator upon Faster R-CNN to separately extract features of classification and localization, and then dual aggregator is introduced for separate model reweighting. Besides, Guirguis et al. focused on the training optimization and proposed a constraint-based fine-tuning approach (CFA)[15]. On the gradient search, more constraints are imposed to address forgetting and enable better knowledge transfer from base to novel classes. It is worth mentioning that the transfer learning-based methods possibly suffer from insufficient feature diversity for generalization. In this paper, an elementary domain stream is added upon the base domain stream to provide more category-agnostic features for enhancing generalization.

2.2 Attention mechanism

The attention mechanism has demonstrated its outstanding performance in the fields of natural language processing[23, 24], object segmentation[25], and visual detection[9, 26–28]. It can adaptively assign weights to different parts of the input data, thereby enhancing the model's accuracy and generalization. A bidirectional long short-term memory[24] is proposed for event temporal relation extraction of natural language processing, where top-k attention mechanism is applied to select the important neighbor nodes and filter irrelevant noise. Geng et al.[28] combined skip connections with a channel attention to refine features in the task of grasp detection, which adaptively recalibrates the concatenated feature from multi-scale encoder outputs. Attention is also utilized in FSOD. The reweighting vectors in Ref. [10] are actually a form of channel attention. An extension of non-local self-attention mechanism is introduced in dense relation distillation with context-aware aggregation (DCNet)[26], where support features and query feature are densely matched. Then the pixel-level relevant information of co-existing objects between query and support samples is distilled to provide a robust feature. Different from the channel-level and pixel-level attentions, we focus on patch-level attention for patches of interest by mining the patch-level correlation of features.

3 Methodology

Figure 1 illustrates the proposed few-shot object detection network via DFF and PA, which is termed as DFFPA. It integrates a two-stream backbone, four dual-domain feature fusion modules, and a patch-level attention module into the framework of Faster RCNN with FPN[29]. Given an input image, feature extraction is first performed by the two-stream backbone, where a pre-trained elementary domain stream is additionally added parallel to original base domain stream. The multi-level feature maps from two streams are merged by the DFF module to obtain more discriminative feature maps. After ROI Pooling, the resultant features are refined through the PA module in a way of patch-level. Finally, the object detection is achieved with box classifier and regressor. It is noted that the elementary domain stream is pre-trained on the ImageNet dataset, and base domain stream takes the abovementioned trained weights as the initial values for further training on the base classes. Other parts of the whole network follow two-stage training process[11, 13, 30]. After training on the base classes with abundant labelled data, DFF, classifier, and regressor are further fine-tuned using a small balanced training set with k-shot per class comprising both base and novel classes. Also, a weighted classification loss is provided during fine-tuning to enhance the classifier, where the features from FPN of the base training model are extra introduced into ROI pooling to synchronously predict the class.

3.1 Two-stream backbone

FSOD methods typically use a single-stream backbone, which makes the features from the backbone usually base-biased and the generalization to novel classes is limited. To deal with this problem, a parallel backbone is additionally introduced for knowledge preservation and information enrichment. The processing of these two backbones is actually a type of data fusion, which involves the integration of data from different resources. Wang et al.[31] and Zhang et al.[32] have provided comprehensive reviews on data fusion. They categorize data fusion into different levels including pixel level, feature level, and decision level. Compared to the pixel level fusion, the feature level fusion focuses on region information and it is less affected by noises, making it more suitable to fuse the information of two backbones. As shown in Fig. 1, our two-stream backbone consists of a base domain stream and an elementary domain stream, both employing the ResNet-101 architecture[33]. The former is initialized with the weights of the latter, which not only accelerates the training of the detection network but also enables knowledge transfer in the thousand classes of ImageNet. In contrast, the elementary domain stream is directly from the pre-trained results on ImageNet and it implies low-level yet category-agnostic features. By integrating these two streams in parallel, a backbone with rich features is created. Specifically, in each stream, ResNet-101 is used with the layers of stem, res2, res3, res4, and res5. The four feature maps $F^{i}_{\text{res},j} (j = 2, 3, 4, 5)$ from the latter four layers of the stream i are selected as the multi-scale features, where $i = 1, 2$ denote the base domain stream and elementary domain stream, respectively.

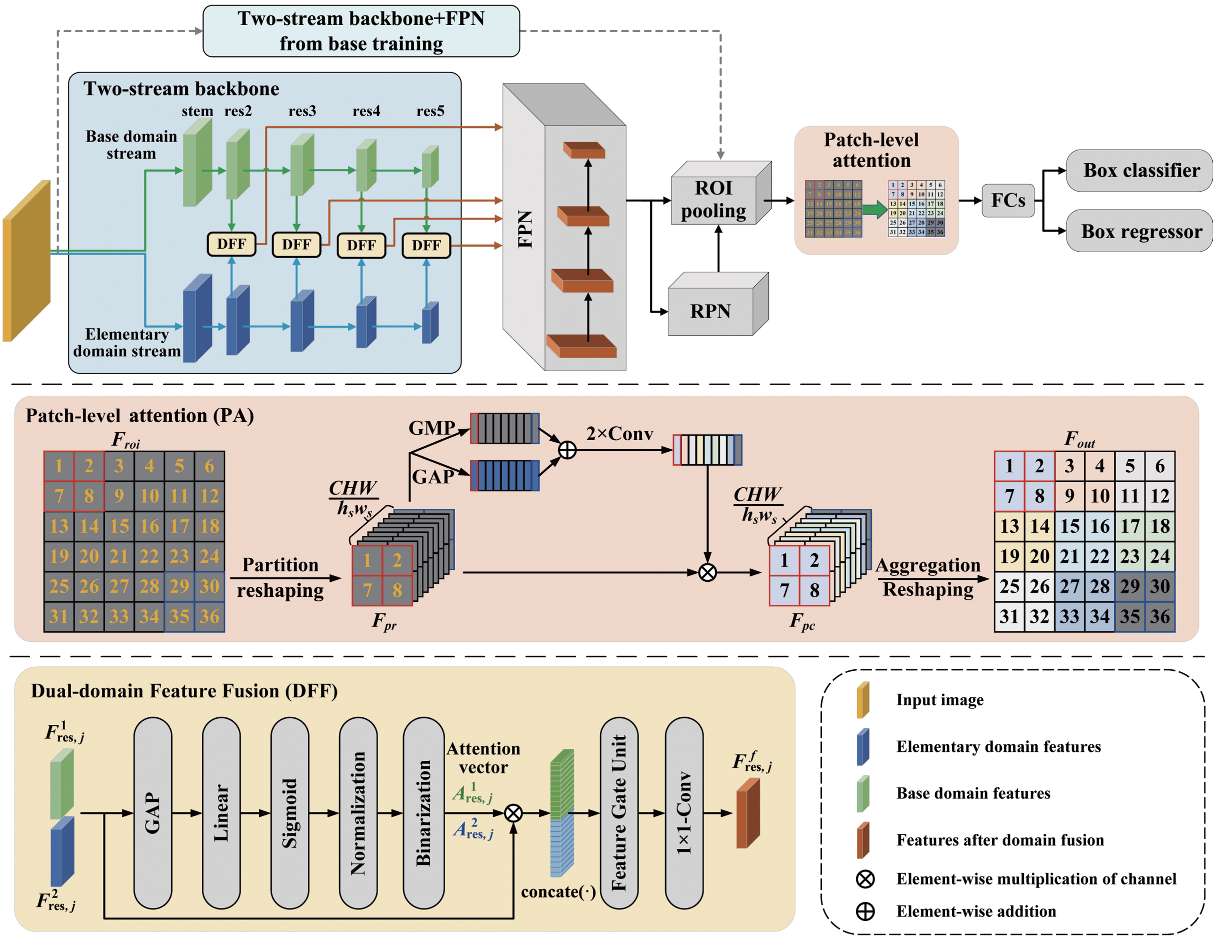


Fig. 1 Architecture of the proposed few-shot object detection via dual-domain feature fusion and patch-level attention (DFFPA). It incorporates a two-stream backbone, four dual-domain feature fusion (DFF) modules, and a patch-level attention (PA) module into the framework of Faster RCNN with FPN. The stem, res2, res3, res4, and res5 are the layers in ResNet-101[33]. The DFF module takes two feature maps $F^1_{\text{res},j}$ and $F^2_{\text{res},j}$ from the streams of base domain and elementary domain as inputs, and output the merged feature map $F^f_{\text{res},j}$, where $j = 2, 3, 4, 5$. The PA module comprises a partition reshaping, a channel attention, and an aggregation reshaping, which achieves the mapping from the input feature map $F_{\text{roi}} \in \mathbb{R}^{C\times H\times W}$ to F_{pr}, $F_{pc} \in \mathbb{R}^{\frac{CHW}{h_s w_s}\times h_s\times w_s}$, and then to the output feature map $F_{out} \in \mathbb{R}^{C\times H\times W}$. C, H, and W refer to the channel number, height, and width of the input feature map, respectively, and h_s, w_s denote the height and width of a patch, respectively.

To fuse the aforementioned feature pairs with the same size from two domains, the DFF module is designed. Due to the significant differences between the features from different domains, it is a challenge to complementarily fuse the features from two domains and further enhance the model's adaptability to complex scenarios. Without simple addition of multi-domain features, an attention block is introduced to the DFF module for adaptive feature fusion, where global average pooling (GAP), linear layer, Sigmoid activate function, normalization, and binarization constitute attention, as shown in Fig. 1. The corresponding features from different domains are processed firstly by the attention block to adaptively remove redundant channels and then concatenated as the input for domain selection based on feature gate unit (FGU). The resultant feature is sent to the convolution layer with 1×1 kernel for channel adjustment and the fused feature is then generated.

We consider the fusion of two feature maps $F^1_{\text{res},j}$, $F^2_{\text{res},j}$ from two streams. To achieve channel-wise feature selection, global average pooling is first applied on each feature map to obtain their respective global representations. Followed by a linear layer and a Sigmoid activation function, an initial channel-wise attention is acquired, which reflects the feature confidence score of each channel. For model stability, the initial attention is normalized to the range of 0-1, and we have $a^{i}_{\text{res},j}, i=1,2$. After a binarization operation, the channel attention vector $A^{i}_{\text{res},j}$ is obtained. Next, the outputted attention is used to reweight the corresponding input feature map $F^{i}_{\text{res},j}$ along the channel dimension for the refined feature map $F^{i,r}_{res,j}$. Finally, the concatenated result $[F^{1,r}_{\text{res},j}, F^{2,r}_{\text{res},j}]$ along the channel direction is adjusted in FGU and further fused through the convolution layer with 1×1 kernel to output the fused feature $F^{f}_{\text{res},j}$. The purpose of feature gate unit is to control the sparsity of features from the two streams.

The procedure of DFF is formulated as follows:

$$
\begin{aligned}
&F^{f}_{\text{res},j} = \\
&\text{Conv}\left(\text{FGU}\left(\text{concate}\left(\begin{array}{c} F^{1}_{\text{res},j}\cdot A^{1}_{\text{res},j}(a^{1}_{\text{res},j},p) \\ F^{2}_{\text{res},j}\cdot A^{2}_{\text{res},j}(a^{2}_{\text{res},j},p) \end{array}\right)\right)\right) = \\
&\text{Conv}\left(\text{concate}\left(\begin{array}{c} \varphi\cdot F^{1}_{\text{res},j}\cdot A^{1}_{\text{res},j}(a^{1}_{\text{res},j},p) \\ (1-\varphi)\cdot F^{2}_{\text{res,j}}\cdot A^{2}_{\text{res},j}(a^{2}_{\text{res},j},p) \end{array}\right)\right)
\end{aligned}
\tag{1}
$$

$$
\begin{aligned}
&A^{i}_{\text{res},j}(a^{i}_{\text{res},j},p) = \\
&\quad \text{Norm}\left(\text{Sigmoid}\left(\text{Linear}\left(\text{GAP}\left(F^{i}_{\text{res},j}\right)\right)\right)\right)\Big|_{p}
\end{aligned}
\tag{2}
$$

where p denotes the binarization threshold. φ is a parameter to control the proportion of the two concatenated features. During the training iterations, φ is 0 and 1 when the iteration number is odd and even, respectively. When inferring, φ is set to 1/2 for the joint involvement of two-stream features. If the element in $a^{i}_{\text{res},j}$ has a value smaller than p, it will be set to a preset value (0.1 in this paper), otherwise it is set to 1.0.

3.2 Patch-level attention

The humans tend to perceive objects with parts of interest. This inspires us to design a part-level attention mechanism. Unlike the previous attentions to capture pixel-wise long-range dependencies[34] or channel-wise relationship[35], our patch-level attention (see Fig. 1) combines channel attention and feature reshaping to capture the relationships among spatial patches. Specifically, the PA module is composed of a partition reshaping, a channel attention, and an aggregation reshaping, where partition and aggregation constitute feature reshaping. The partition reshaping operation aims to divide the feature map of single object into multiple spatial patches and then these patches are concatenated along the channel dimension. This provides the chance to mine the internal relationships among spatial patch features through channel attention. The channel attention processes the concatenated feature map using the channel adjustment based on pooling and convolution to reweight the confidence of each spatial patch, which is beneficial to capture key parts of channels. Afterwards, aggregation reshaping is used to ensure consistency in size of PA input and output. As a result, distinctive local regions are highlighted for better object detection.

The PA module takes the result $F_{\text{roi}}\in\mathbb{R}^{C\times H\times W}$ of ROI Pooling as the input, which is segmented into multiple patches with the size of $\dfrac{HW}{h_s w_s}\times h_s\times w_s$ in a channel-wise manner. These patches are concatenated in channel to generate a feature map $F_{pr}\in\mathbb{R}^{\frac{CHW}{h_s w_s}\times h_s\times w_s}$, where C, H, and W refer to the channel number, height, and width of the PA input, respectively, and h_s, w_s denote the height and width of a patch. It is important to mention that the height and width of the patches should be able to evenly divide the corresponding dimensions of the PA input. This ensures the smoothness of feature concatenation. On this basis, global max-pooling (GMP) and global average-pooling (GAP) are respectively imposed on the feature map F_{pr}, and two feature maps containing the maximum value and average value of each channel are generated. Followed by an element-wise addition, two cascaded convolutions with kernel size of 1×1 are employed to obtain the channel attention map, which is combined with F_{pr} in a way of element-wise multiplication to obtain the refined patch-level feature map F_{pc}. To restore the size in coincidence with PA input, the aggregation reshaping operation unfolds F_{pc} to output the feature map $F_{\text{out}}\in\mathbb{R}^{C\times H\times W}$, where the positions of the corresponding patches between the PA input and output remain unchanged.

3.3 Training loss function

Similar to the common transfer learning-based few-shot object detection[11, 13, 30], the training of the proposed DFFPA includes two stages: base training and few-shot fine-tuning. In the former stage, DFFPA are trained on the base classes with the following loss function[11].

$$
\mathcal{L}=\mathcal{L}_{\text{rpn}}+\mathcal{L}_{\text{cls}}+\mathcal{L}_{\text{loc}} \tag{3}
$$

where $\mathcal{L}_{\text{rpn}}$, $\mathcal{L}_{\text{cls}}$, and $\mathcal{L}_{\text{loc}}$ denote the losses of RPN, box classifier, and box regressor, respectively. As for the fine-tuning, the few samples of novel classes in the dataset is employed to further fine-tune DFFPA to adaptively accommodate novel classes, where DFF is specially adjusted to adapt to novel classes. Considering the parameters change of DFF affects the fused features and weakens the detection of base classes, the auxiliary feature F_{aux}^{f} from FPN of the base training model is provided for each proposal box, which is also fed into the subsequential network to compute corresponding box classification loss $\mathcal{L}_{\text{cls}}^{\text{aux}}$. The fine-tuning loss $\mathcal{L}_{\text{ft}}$ with weighted classification loss is given by

$$\mathcal{L}_{\text{ft}} = \mathcal{L}_{\text{rpn}} + \left(\alpha\mathcal{L}_{\text{cls}}^{\text{aux}} + \mathcal{L}_{\text{cls}}\right) + \mathcal{L}_{\text{loc}} \tag{4}$$

where α is the hyper-parameter to adjust the proportion of classification losses $\mathcal{L}_{\text{cls}}^{\text{aux}}$ and $\mathcal{L}_{\text{cls}}$. With $\mathcal{L}_{\text{ft}}$, the knowledge of base model could be better remained and the whole detection performance is improved. Algorithm 1 details the object detection process of the proposed DFFPA method. Preprocessing$(\cdot)$ represents data augmentation including RandomFlip and Expand on input images. $\text{scores}_{\text{cls}}$ and bboxes denote the classification scores and the predicted boxes for candidate objects. NMS$(\cdot)$ refers to the non-maximum suppression.

4 Experiment

4.1 Experimental setup

In this section, the proposed DFFPA is testified on two public datasets: PASCAL VOC[36] and MS COCO[37].

PASCAL VOC dataset with VOC 2007 and VOC 2012 is a well-known dataset with 20 categories. VOC 2007 and VOC 2012 are respectively divided into training and validation sets. There are three different base/novel splits, and in each split, 5 classes are chosen as novel classes while the remaining classes are designated as base classes. Novel split 1 contains the novel classes of bird, bus, cow, motorbike, and sofa, and other two splits are termed as novel split 2 (aeroplane, bottle, cow, horse, and sofa) and novel split 3 (boat, cat, motorbike, sheep, and sofa). For each novel class, *k*-shot examples are involved in training, where *k* represents the number of labeled instances with values of 1, 2, 3, 5, and 10. Evaluation is performed on the VOC 2007 test set. The average precision with 50% IoU threshold (AP_{50}) of the novel classes under different data split settings is reported.

Algorithm 1 DFFPA

```
Input: Input image $s$.
Output: Detection results det_res.
$s' \leftarrow \text{Preprocessing}(s)$;
$F_{\text{res},j}^{1}, F_{\text{res},j}^{2} \leftarrow \text{TwoStreamBackbone}(s')$;
$F_{\text{res},j}^{f} \leftarrow \text{DFF}\left(F_{\text{res},j}^{1}, F_{\text{res},j}^{2}\right)$;
$F_{\text{roi}} \leftarrow \text{ROI}\left(\text{RPN}\left(\text{FPN}\left(\bigcup_{j=2}^{5} F_{\text{res},j}^{f}\right)\right)\right)$;
$F_{\text{out}} \leftarrow \text{PA}(F_{\text{roi}})$;
$\text{scores}_{\text{cls}} \leftarrow \text{Classifier}(\text{FC}(F_{\text{out}}))$;
$\text{bboxes} \leftarrow \text{Regressor}(\text{FC}(F_{\text{out}}))$;
$\text{det\_res} \leftarrow \text{NMS}(\text{scores}_{\text{cls}}, \text{bboxes})$;
return det_res
```

MS COCO dataset comprises 80 classes. 20 classes are regarded as novel classes and the others serve as base classes. The proposed method is testified on MS COCO 2014 validation dataset with *k*=10 and 30 shots. AP, AP_{50}, AP_{75}, AP_{S}, AP_{M}, and AP_{L} are considered as evaluation metrics. Note that AP_{S}, AP_{M}, and AP_{L} refer to different average precision values at different scales (*Small*, *Medium*, and *Large*).

Implementation details: Referring to the setting of the existing methods including TFA[11], FSCE[13], and DMNet[21], the stochastic gradient descent[38] with the momentum of 0.9 and the parameter decay of 0.0001 is adopted to optimize our network end-to-end with a mini-batch size of 16. During the base training phase, the learning rate is set to 0.02, while for the few-shot fine-tuning, it is adjusted to 0.015. Our experiments are conducted on a device with NVIDIA TITAN RTX GPU and Intel(R) Xeon(R) CPU E5-2660 v4 CPU.

4.2 Ablation studies

1) Ablation of DFFPA: To testify the performance of our proposed DFFPA, its sixth variants DFFPA-I, DFFPA-II, DFFPA-III, DFFPA-IV, DFFPA-V, and DFFPA-VI are considered according to whether DFF, PA, and weighted classification loss (wcl) are involved. DFFPA-II uses addition fusion to replace DFF, where the inputted features are fused through simple element-wise addition. Table 1 presents the ablation results of different variants of our DFFPA on novel split 1 of PASCAL VOC in terms of AP_{50} with 5-shot and 10-shot. Compared with TFA, DFFPA-I and DFFPA-II respectively adopt the modules of addition fusion and DFF to refine the domain feature, and the results indicates the merit of the proposed DFF. From the results of TFA and DFFPA-IV, the solution with PA module significantly improves the performance,

Table 1 Comparison of different variants of our DFFPA on novel split 1 of PASCAL VOC.

Method	Module				AP_{50}(%)	
	Feature fusion		Patch-level Attention (PA)	Weighted classification loss (wcl)	5-shot	10-shot
	Addition fusion	Dual DFF				
TFA (Baseline)	×	×	×	×	55.7	56.0
DFFPA-I	✓	×	×	×	54.4	56.5
DFFPA-II	×	✓	×	×	59.1	64.1
DFFPA-III	×	✓	×	✓	60.1	65.1
DFFPA-IV	×	×	✓	×	60.9	63.9
DFFPA-V	×	×	✓	✓	61.9	64.3
DFFPA-VI	×	✓	✓	×	62.1	65.2
DFFPA	×	✓	✓	✓	62.3	66.8

benefiting from the capability to effectively capture key parts of objects and highlight distinctive local regions. On the basis of DFFPA-II, DFFPA-VI combines PA and the performance is further improved. Besides, one can see from the results of DFFPA-VI and DFFPA that the introduction of wcl is beneficial to the detection. Overall, the proposed DFFPA performs the best.

2) Ablation of DFF: The DFF module aims to fuse features, where the binarization threshold p is a key factor to control the distribution of features during fusion. The smaller the value of p, the more the input features are preserved. When p is 0, all input features remain unchanged. We consider five variants of DFF with different binarization threshold p and the comparison on novel split 1 of PASCAL VOC in terms of AP, AP_{75}, and AP_{50} are shown in Table 2. One can see that the value of p has a significant impact on the result, and a bigger value shall decrease the detection quality. p is chosen to 0.2 in our method to pursue the best performance.

3) Ablation of PA: The PA is used to capture the interest of patch of objects following the ROI pooling. Notice that the input size of PA will limit the receptive field of each patch and further affect the performance. Thus, to select the proper input size, we conduct the ablation of PA with different input sizes, and the results are reported in Table 3. It can be seen that the input size 8 × 8 is preferable with its best performance.

Table 2 Ablation of DFF on novel split 1 of PASCAL VOC.

Setting	AP(%)	AP_{75}(%)	AP_{50}(%)
DFF-Ⅰ (p=0.0)	38.9	41.3	65.0
DFF-Ⅱ (p=0.2)	**40.4**	**43.1**	**66.8**
DFF-Ⅲ (p=0.4)	37.8	39.2	63.1
DFF-Ⅳ (p=0.6)	32.5	31.6	57.6
DFF-Ⅴ (p=0.8)	30.8	30.4	53.7

Table 3 Ablation of PA on novel split 1 of PASCAL VOC.

Method	Input size	AP(%)	AP_{75}(%)	AP_{50}(%)
PA-Ⅰ	6×6	38.2	39.0	63.5
PA-Ⅱ	10×10	39.3	41.4	64.9
PA-Ⅲ	12×12	39.5	42.6	64.7
PA-Ⅳ	8×8	**40.4**	**43.1**	**66.8**

4) Ablation of wcl: In the classification loss, an additional loss $\mathcal{L}_{\text{cls}}^{\text{aux}}$ based on the auxiliary feature is added upon the common loss $\mathcal{L}_{\text{cls}}$, and a hyper-parameter α is used to weight these two losses. To explore the influence of α, different experiments are conducted, as shown in Table 4, where α is set to 0.25, 0.50, 0.75, and 1.0, respectively. With the increasing of α, the detection performance gradually improves. According to the results, α is selected to 1.0, which means equal treating of $\mathcal{L}_{\text{cls}}^{\text{aux}}$ and $\mathcal{L}_{\text{cls}}$.

4.3 Comparison with existing methods

We evaluate and compare the proposed DFFPA with existing methods including MetaDet[8], Meta R-CNN[9], YOLO-ft-full[10], FRCN-ft-full[9], FSRW[10], TFA[11], MPSR[12], FSCE[13], DCNet[26], and DMNet[21]. Table 5 shows the novel class detection results under different shots (k=1, 2, 3, 5, and 10) in terms of AP_{50} on three splits of PASCAL VOC. It is indicated that the proposed method performs well. In

Table 4 Ablation of wcl on novel split 1 of PASCAL VOC.

Setting	AP(%)	AP_{75}(%)	AP_{50}(%)
wcl-Ⅰ (α=0.25)	38.3	39.8	65.3
wcl-Ⅱ (α=0.50)	39.6	41.1	66.2
wcl-Ⅲ (α=0.75)	39.7	41.0	66.5
wcl-Ⅳ (α=1.0)	**40.4**	**43.1**	**66.8**

Table 5 Comparison of different methods on PASCAL VOC dataset in terms of AP_{50} (%).

Method	Novel split 1					Novel split 2					Novel split 3				
	1	2	3	5	10	1	2	3	5	10	1	2	3	5	10
YOLO-ft-full[10]	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
FRCN-ft-full[9]	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
FSRW[10]	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[8]	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
MetaR-CNN[9]	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[11]	39.8	36.1	44.7	55.7	56.0	23.5	26.9	34.1	35.1	39.1	30.8	34.8	42.8	49.5	49.8
MPSR[12]	41.7	-	51.4	55.2	61.8	24.4	-	39.2	39.9	47.8	35.6	-	42.3	48.0	49.7
FSCE[13]	44.2	43.8	51.4	61.9	63.4	27.3	29.5	43.5	44.2	50.2	37.2	41.9	47.5	54.6	58.5
DCNet[26]	33.9	37.4	43.7	51.1	59.6	23.2	24.8	30.6	36.7	46.6	32.3	34.9	39.7	42.6	50.7
DMNet[21]	39.0	48.9	50.7	58.6	62.5	31.2	32.4	40.3	47.6	52.0	41.7	41.8	42.7	50.3	52.1
Ours	44.0	52.0	54.4	62.8	66.8	31.5	36.4	46.5	49.4	53.1	39.9	44.9	48.9	56.4	57.1

the 1-shot setting, our results on novel split 1 and novel split 3 performs inferior to FSCE[13] and DMNet[21], respectively. This discrepancy possibly comes from the limited number of training samples available for the novel class.

Figure 2 visualizes features learned by DFFPA on a selected image from PASCAL VOC dataset. The heatmaps are obtained through Grad-CAM (Gradient-weighted class activation mapping)[39], which reflects the degree of attention on different regions of the image. For the original image in Fig. 2a, the features from the res5 layers of base and elementary domains are depicted in Figs. 2b and 2c, respectively. They are fused by DFF and the result is shown in Fig. 2d. Figures 2e and 2f provides the features outputted from FPN and PA, respectively. The final detection result is given in Fig. 2g. It is observed that the features processed by DFF, FPN, and PA can focus on the target object well, which provides good feature discriminability. In addition, Fig. 3 visualizes the detection results generated by the proposed DFFPA on PASCAL VOC dataset with 10-shot, where each row corresponds to a novel split.

For MS COCO dataset, it comprises more target object categories than PASCAL VOC dataset. This means more challenging in manifold scenarios and the detection results for the novel classes are presented in Table 6. Overall, our DFFPA demonstrate good performance. Figure 4 gives the detection results of the proposed DFFPA with 30-shot on MS COCO dataset and the novel objects can be well detected.

4.4 Robustness verification

To further verify the proposed method, different interferences are imposed on four images from novel split 1 of PASCAL VOC and the detection results with

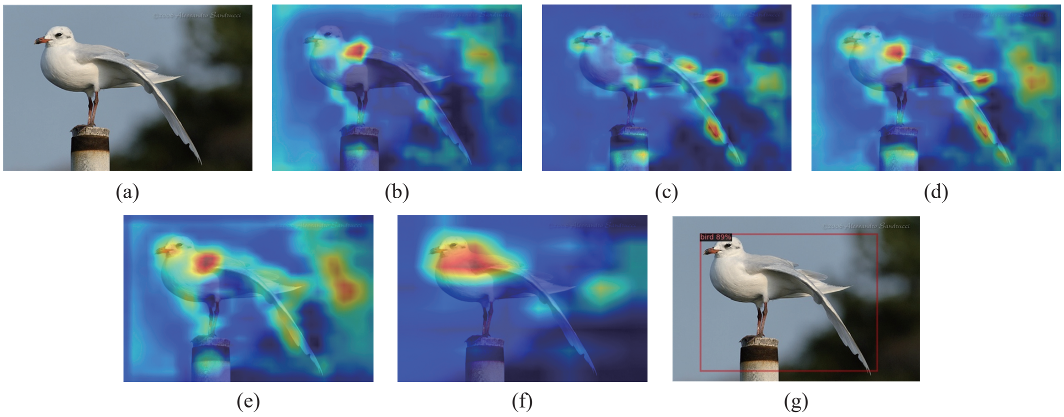

(a) (b) (c) (d)
(e) (f) (g)

Fig. 2 Visualization of features learned by DFFPA on the selected image from PASCAL VOC dataset. The heatmap is generated by Grad-CAM[39]. (a) Original image. (b) Feature from the res5 layer of base domain. (c) Feature from the res5 layer of elementary domain. (d) Fused result of (b) and (c). (e) Feature from FPN. (f) Output feature of PA. (g) Detection result.

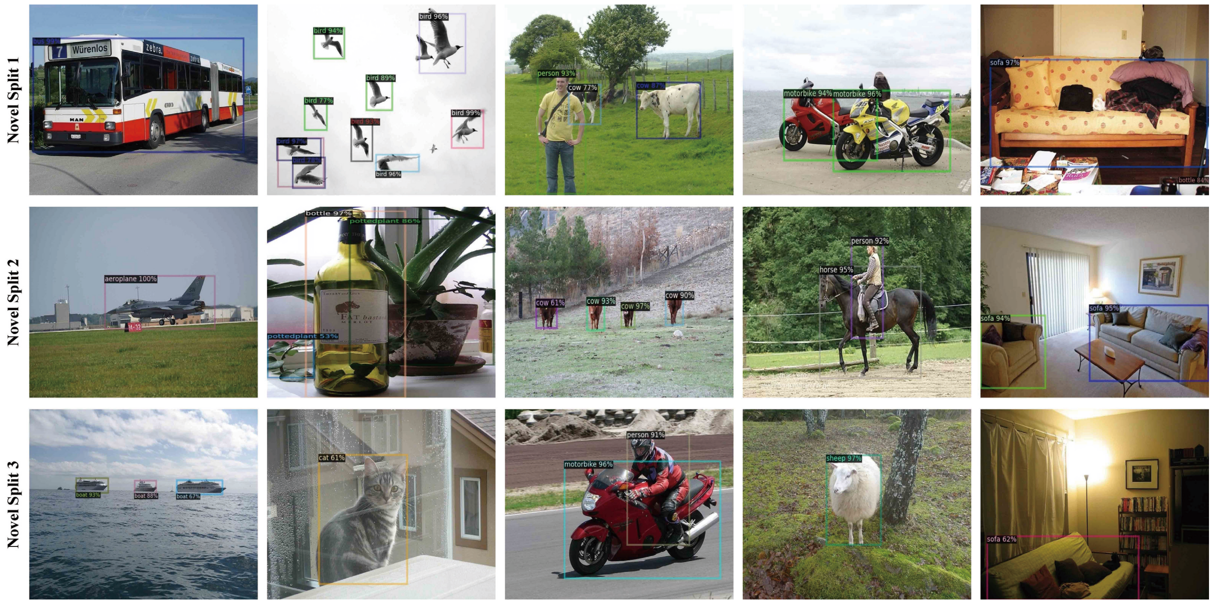


Fig. 3 Detection results of DFFPA with shot=10 on PASCAL VOC dataset. Results with confidence scores higher than 0.5 are visualized.

Table 6 Comparison of different methods on MS COCO dataset in terms of AP, AP_{50}, AP_{75}, AP_S, AP_M, and AP_L (%).

Method	10-shot						30-shot					
	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
YOLO-ft-full[10]	3.1	7.9	1.7	0.7	2.0	6.3	7.7	16.7	6.4	0.4	3.3	14.4
FRCN-ft-full[9]	6.5	13.4	5.9	1.8	5.3	11.3	11.1	21.6	10.3	2.9	8.8	18.9
FSRW[10]	5.6	12.3	4.6	0.9	3.5	10.5	9.1	19.0	7.6	0.8	4.9	16.8
MetaDet[8]	7.1	14.6	6.1	1.0	4.1	12.2	11.3	21.7	8.1	1.1	6.2	17.3
Meta R-CNN[9]	8.7	19.1	6.6	2.3	7.7	14.0	12.4	25.3	10.8	2.8	11.6	19.0
TFA[11]	10.0	-	9.3	-	-	-	13.7	-	13.4	-	-	-
MPSR[12]	9.8	17.9	9.7	3.3	9.2	16.1	14.1	25.4	14.2	4.0	12.9	23.0
FSCE[13]	11.9	-	10.5	-	-	-	16.4	-	16.2	-	-	-
DCNet[26]	12.8	23.4	11.2	4.3	13.8	21.0	18.6	32.6	17.5	6.9	16.5	27.4
DMNet[21]	10.0	17.4	10.4	3.4	8.3	16.1	17.1	29.7	17.7	4.8	14.7	26.5
Ours	13.0	24.0	12.4	3.2	14.1	21.3	18.6	33.7	18.7	6.9	20.9	27.7

confidence scores higher than 0.3 are illustrated in Fig. 5. The first row presents the detection results of original images. The second to seventh rows correspond to the results after brightness enhancement (25%), brightness reduction (25%), salt-pepper noise with intensity of 3%, GridMask with size 3×3, Gaussian blur with kernel size 5×5, and random mask. In spite of these interferences, the proposed method still achieves stable detection, which proves its robustness.

4.5 Verification on actual scene

To further testify the proposed method, the experiments on actual scene are conducted and the results are shown in Figs. 6a and 6b, where the concerned object classes include bottle, cup, teddy bear, dining table, apple, banana, orange, remote, cell phone, sports ball, and mouse. It is seen that the objects are correctly detected.

5 Conclusion

This paper proposes a few-shot object detection method with dual-domain feature fusion and patch-level attention. Following the transfer learning-based framework, a two-stream backbone is designed, and the feature diversity is retained through parallel base and elementary domains. The features from these two

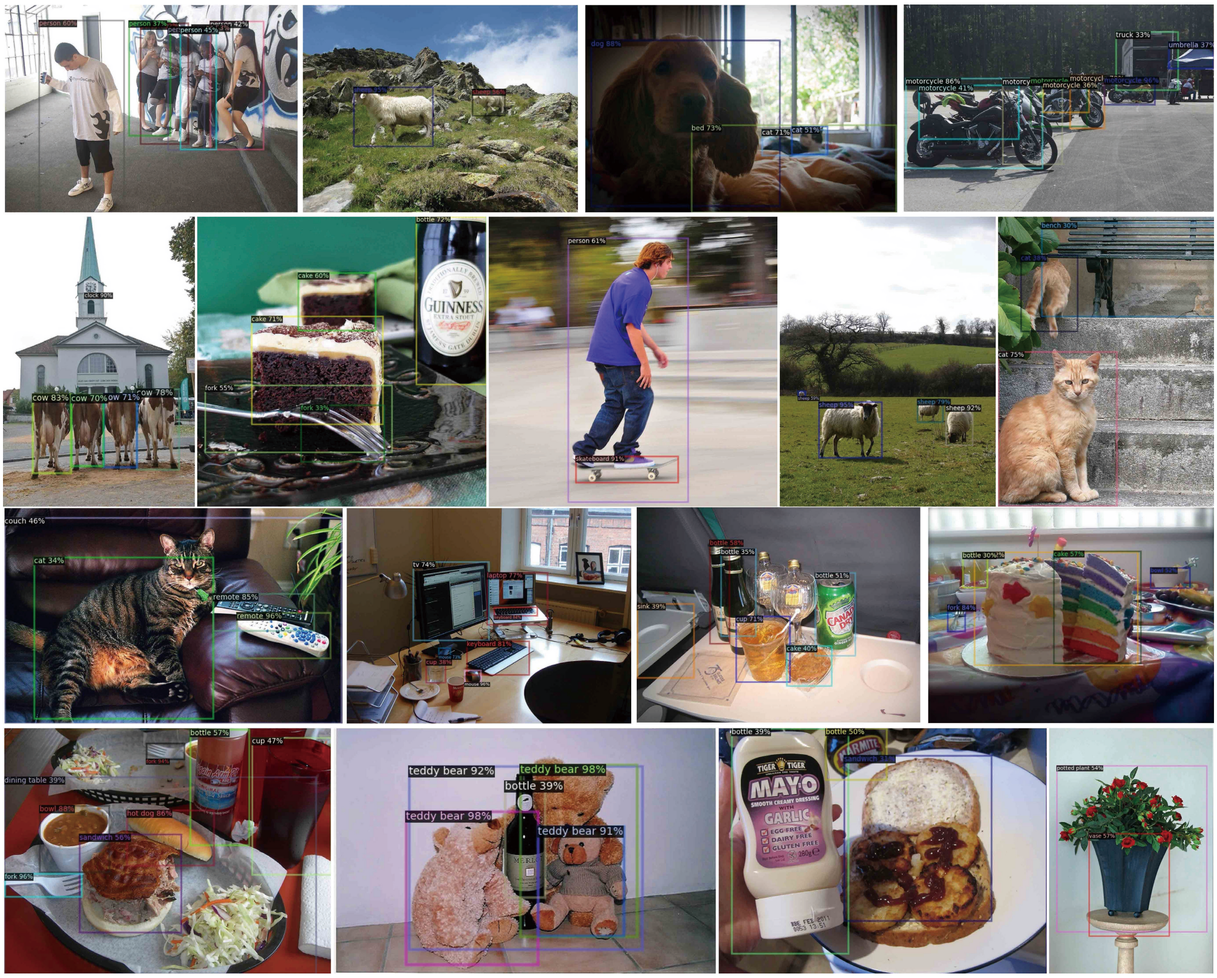


Fig. 4 Detection results of DFFPA with shot=30 on MS COCO dataset. Results with confidence scores higher than 0.3 are displayed.

domains are fused in the dual-domain feature fusion module, and adaptive feature refinement are achieved. Then, a patch-level attention is presented to capture the crucial parts of object features from ROI head for better feature discrimination. In addition, a weighted classification loss is given to assist the fine-tuning of classifier. Experimental results on PASCAL VOC and MS COCO datasets demonstrate the effectiveness of the proposed method.

Acknowledgment

This work was supported in part by Beijing Natural Science Foundation (Nos. L233030 and 2022MQ05), and in part by the National Natural Science Foundation of China (Nos. 62073322, 61836015, and 61633020).

References

[1] W. Rawat and Z. Wang, Deep convolutional neural networks for image classification: A comprehensive review, *Neural Comput.*, vol. 29, no. 9, pp. 2352–2449, 2017.

[2] J. Jiao, H. Pan, C. Chen, T. Jin, Y. Dong, and J. Chen, Two-stage lesion detection approach based on dimension-decomposition and 3D context, *Tsinghua Science and Technology*, vol. 27, no. 1, pp. 103–113, 2022.

[3] S. Guo, F. Liu, X. Yuan, C. Zou, L. Chen, and T. Shen, HSPOG: An optimized target recognition method based on histogram of spatial pyramid oriented gradients, *Tsinghua Science and Technology*, vol. 26, no. 4, pp. 475-483, 2021.

[4] X. Wu, D. Sahoo, and S. C. H. Hoi, Recent advances in deep learning for object detection, *Neurocomputing*, vol. 396, pp. 39–64, 2020.

[5] M. Andrychowicz, M. Denil, S. Gomez, M. W. Hoffman, D. Pfau, T. Schaul, B. Shillingford, and N. D. Freitas, Learning to learn by gradient descent by gradient descent, in *Proc. Int. Conf. on Neural Information Processing Systems*, Red Hook, NY, USA, 2016, pp. 3988–3996.

[6] W. Y. Chen, Y. C. Liu, Z. Kira, Y. C. Wang, and J. B. Huang, A closer look at few-shot classification, arXiv preprint arXiv: 1904.04232, 2019.

[7] Y. Tian, Y. Wang, D. Krishnan, J. B. Tenenbaum, and P.

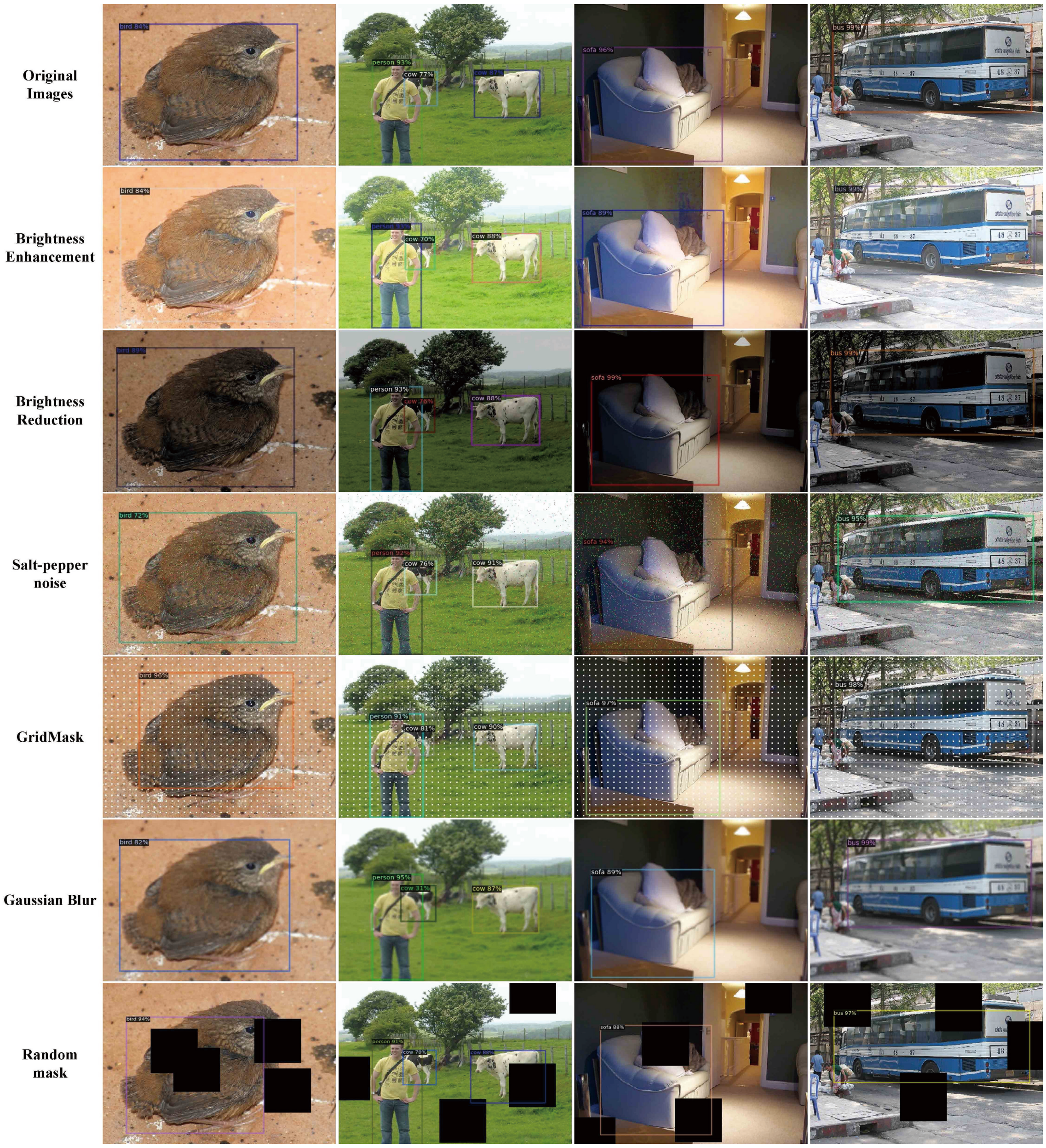


Fig. 5 Robustness verification of DFFPA on selected images from PASCAL VOC. Results with confidence scores higher than 0.3 are visualized.

Isola, Rethinking few-shot image classification: A good embedding is all you need? in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J. M. Frahm, eds. Cham, Switzerland: Springer, 2020, pp. 266–282.

[8] Y. X. Wang, D. Ramanan, and M. Hebert, Meta-learning to detect rare objects, in *Proc. IEEE/CVF Int. Conf. Computer Vision* (*ICCV*), Seoul, Republic of Korea, 2019, pp. 9924–9933.

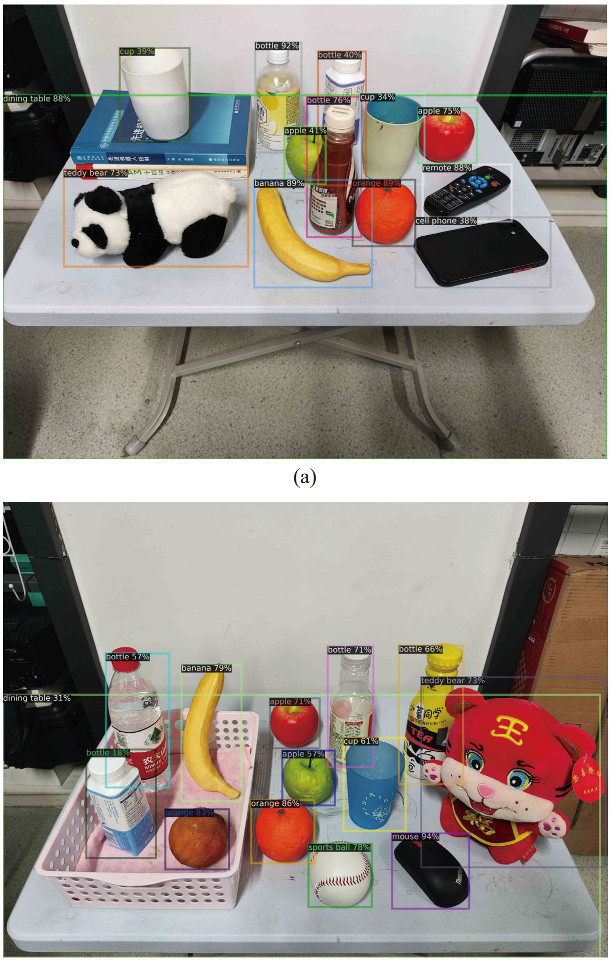
[9] X. Yan, Z. Chen, A. Xu, X. Wang, X. Liang, and L. Lin, Meta R-CNN: Towards general solver for instance-level low-shot learning, in *Proc. IEEE/CVF Int. Conf. Computer Vision* (*ICCV*), Seoul, Republic of Korea, 2019, pp. 9576–9585.

[10] 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. Computer Vision* (*ICCV*), Seoul, Republic of Korea, 2019, pp. 8419–8428.

[11] X. Wang, T. Huang, J. Gonzalez, T. Darrell, and F. Yu, Frustratingly simple few-shot object detection, in *Proc. Int. Conf. on Machine Learning*, Virtual, 2020, pp. 9919–9928.

[12] J. Wu, S. Liu, D. Huang, and Y. Wang, Multi-scale positive sample refinement for few-shot object detection, in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J. M. Frahm, eds. Cham, Switzerland: Springer, 2020, pp. 456–472.

[13] B. Sun, B. Li, S. Cai, Y. Yuan, and C. Zhang, FSCE: Few-shot object detection via contrastive proposal encoding, in



(a)

(b)

Fig. 6 Detection results of DFFPA on actual scene.

Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (*CVPR*), Nashville, TN, USA, 2021, pp. 7348–7358.

[14] L. Liu, B. Ma, Y. Zhang, X. Yi, and H. Li, AFD-net: Adaptive fully-dual network for few-shot object detection, in *Proc. 29th ACM Int. Conf. Multimedia*, Virtual Event, 2021, pp. 2549–2557.

[15] K. Guirguis, A. Hendawy, G. Eskandar, M. Abdelsamad, M. Kayser, and J. Beyerer, CFA: Constraint-based finetuning approach for generalized few-shot object detection, in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition Workshops* (*CVPRW*), New Orleans, LA, USA, 2022, pp. 4038–4048.

[16] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and F.-F. Li, ImageNet: A large-scale hierarchical image database, in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Miami, FL, USA, 2009, pp. 248–255.

[17] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni. Generalizing from a few examples: A survey on few-shot learning, *ACM Computing Surveys*, 2020, vol. 53, no. 3, pp. 1–34.

[18] H. Xu, X. Wang, F. Shao, B. Duan, and P. Zhang, Few-shot object detection via sample processing, *IEEE Access*, vol. 9, pp. 29207–29221, 2021.

[19] J. Redmon and A. Farhadi, YOLO9000: Better, faster, stronger, in *Proc. IEEE Conf. Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, HI, USA, 2017, pp. 6517–6525.

[20] S. Ren, K. He, R. Girshick, and J. Sun, Faster R-CNN: Towards real-time object detection with region proposal networks, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, 2017.

[21] Y. Lu, X. Chen, Z. Wu, and J. Yu, Decoupled metric network for single-stage few-shot object detection, *IEEE Trans. Cybern.*, vol. 53, no. 1, pp. 514–525, 2023.

[22] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, *SSD*: *Single Shot MultiBox Detector*. Cham, Switzerland: Springer, 2016, pp. 21–37.

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, Attention is all you need, in *Proc. 31st Int. Conf. on Neural Information Processing Systems* (*NIPS'17*), Red Hook, NY, USA, 6000–6010, 2017.

[24] X. Xu, T. Gao, Y. Wang, and X. Xuan, Event temporal relation extraction with attention mechanism and graph neural network, *Tsinghua Science and Technology*, vol. 27, no. 1, pp. 79–90, 2022.

[25] L. Zhang, K. Zhang, and H. Pan, SUNet++: A deep network with channel attention for small-scale object segmentation on 3D medical images, *Tsinghua Science and Technology*, vol. 28, no. 4, pp. 628–638, 2023.

[26] H. Hu, S. Bai, A. Li, J. Cui, and L. Wang, Dense relation distillation with context-aware aggregation for few-shot object detection, in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Nashville, TN, USA, 2021, pp. 10180–10189.

[27] G. Ren, W. Geng, P. Guan, Z. Cao, and J. Yu, Pixel-wise grasp detection via twin deconvolution and multi-dimensional attention, *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 8, pp. 4002–4010, 2023.

[28] W. Geng, Z. Cao, P. Guan, F. Jing, M. Tan, and J. Yu, Grasp detection with hierarchical multi-scale feature fusion and inverted shuffle residual, *Tsinghua Science and Technology*, vol. 29, no. 1, pp. 244–256, 2024.

[29] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, Feature pyramid networks for object detection, in *Proc. IEEE Conf. Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, HI, USA, 2017, pp. 936–944.

[30] J. Leng, T. Chen, X. Gao, Y. Yu, Y. Wang, F. Gao, and Y. Wang, A comparative review of recent few-shot object detection algorithms. arXiv preprint arXiv:2111.00201, 2021.

[31] S. Wang, M. E. Celebi, Y.-D. Zhang, X. Yu, S. Lu, X. Yao, Q. Zhou, M.-G. Miguel, Y. Tian, J. M. Gorriz, et al., Advances in data preprocessing for biomedical data fusion: An overview of the methods, challenges, and prospects, *Inf. Fusion*, vol. 76, pp. 376–421, 2021.

[32] Y.-D. Zhang, Z. Dong, S.-H. Wang, X. Yu, X. Yao, Q. Zhou, H. Hu, M. Li, C. Jiménez-Mesa, J. Ramirez, et al., Advances in multimodal data fusion in neuroimaging: Overview, challenges, and novel orientation, *Inf. Fusion*, vol. 64, pp. 149–187, 2020.

[33] K. He, X. Zhang, S. Ren, and J. Sun, Deep residual learning for image recognition, in *Proc. IEEE Conf. Computer Vision and Pattern Recognition* (*CVPR*), Las

Vegas, NV, USA, 2016, pp. 770–778.

[34] X. Wang, R. Girshick, A. Gupta, and K. He, Non-local neural networks, in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 7794–7803.

[35] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, Dual attention network for scene segmentation, in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, CA, USA, 2019, pp. 3141–3149.

[36] 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, 2010.

[37] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, *Microsoft COCO*: *Common Objects in Context*. Cham, Switzerland: Springer, 2014.

[38] K. Lu, H. Wang, H. Zhang, and L. Wang, Convergence in high probability of distributed stochastic gradient descent algorithms, *IEEE Trans. Automat. Contr.*, vol. 69, no. 4, pp. 2189–2204, 2024.

[39] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, Grad-CAM: Visual explanations from deep networks via gradient-based localization, in *Proc. IEEE Int. Conf. Computer Vision* (*ICCV*), Venice, Italy, 2017, pp. 618–626.

Guangli Ren received the BE degree in intelligent science and technology from Dalian Maritime University, Dalian, China, in 2015, and the ME degree in technology of computer application from the Capital Normal University, Beijing, China, in 2018. He is currently pursuing the PhD degree in control theory and control engineering from the Institute of Automation, Chinese Academy of Sciences, Beijing, China. His current research interests include visual SLAM and robotic manipulation.

Jierui Liu received the BE degree from South China University of Technology, Guangdong, China, in 2019. He is currently pursuing the PhD degree in control theory and control engineering from the Institute of Automation, Chinese Academy of Sciences, Beijing, China. His current research interests include visual measurement and service robot.

Mengyao Wang received the BE degree from Shandong University, Shandong, China, in 2023. She is currently pursuing the PhD degree in control theory and control engineering from the Institute of Automation, Chinese Academy of Sciences, Beijing, China. Her current research interests include deep learning and scene understanding.

Peiyu Guan received the BE degree in electronic information science and technology from Jilin University, Changchun, China, in 2017. In 2022, she received the PhD degree in control theory and control engineering from the Institute of Automation, Chinese Academy of Sciences, Beijing, China. She is currently an assistant professor in the Institute of Automation, Chinese Academy of Sciences. Her research interests include service robot and image processing.

Zhiqiang Cao received the BE degree in industrial automation and ME degree in control theory and control engineering from Shandong University of Technology, Jinan, China, in 1996 and 1999, respectively. In 2002, he received the PhD degree in control theory and control engineering from the Institute of Automation, Chinese Academy of Sciences, Beijing, China. He is currently a professor in the Institute of Automation, Chinese Academy of Sciences. His research interests include SLAM, navigation, and manipulation of service robot.

Junzhi Yu received the BE degree in safety engineering and the ME degree in precision instruments and mechanology from the North University of China, Taiyuan, China, in 1998 and 2001, respectively, and the PhD degree in control theory and control engineering from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2003. From 2004 to 2006, he was a postdoctoral researcher with the Center for Systems and Control, Peking University. He was an associate professor with the Institute of Automation, Chinese Academy of Sciences, in 2006, where he was a full professor in 2012. In 2018, he joined the College of Engineering, Peking University, as a tenured full professor. His current research interests include intelligent robots, motion control, and intelligent mechatronic systems.