开源

This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Universal-Prototype Enhancing for Few-Shot Object Detection

Aming Wu[1] Yahong Han[2,3,4*] Linchao Zhu[5] Yi Yang[5]
[1]School of Electronic Engineering, Xidian University, Xi'an, China
[2]College of Intelligence and Computing, Tianjin University, Tianjin, China
[3]Tianjin Key Lab of Machine Learning, Tianjin University, Tianjin, China
[4]Peng Cheng Laboratory, Shenzhen, China [5]ReLER Lab, AAII, University of Technology Sydney
amwu@xidian.edu.cn, yahong@tju.edu.cn, {Linchao.Zhu, yi.yang}@uts.edu.au

Abstract

Few-shot object detection (FSOD) aims to strengthen the performance of novel object detection with few labeled samples. To alleviate the constraint of few samples, enhancing the generalization ability of learned features for novel objects plays a key role. Thus, the feature learning process of FSOD should focus more on intrinsical object characteristics, which are invariant under different visual changes and therefore are helpful for feature generalization. Unlike previous attempts of the meta-learning paradigm, in this paper, we explore how to enhance object features with intrinsical characteristics that are universal across different object categories. We propose a new prototype, namely universal prototype, that is learned from all object categories. Besides the advantage of characterizing invariant characteristics, the universal prototypes alleviate the impact of unbalanced object categories. After enhancing object features with the universal prototypes, we impose a consistency loss to maximize the agreement between the enhanced features and the original ones, which is beneficial for learning invariant object characteristics. Thus, we develop a new framework of few-shot object detection with universal prototypes ($FSOD^{up}$) that owns the merit of feature generalization towards novel objects. Experimental results on PASCAL VOC and MS COCO show the effectiveness of $FSOD^{up}$. Particularly, for the 1-shot case of VOC Split2, $FSOD^{up}$ outperforms the baseline by 6.8% in terms of mAP.

1. Introduction

Recently, owing to the success of deep learning, great progress has been made on object detection [26, 11, 14, 12]. However, the outstanding performance [25, 21, 3, 18] depends on abundant annotated objects in training images for each category. As a challenging task, few-shot object detection (FSOD) [16, 35] mainly aims to improve the detection performance for novel objects that belong to certain categories but appear rarely in the annotated training images.

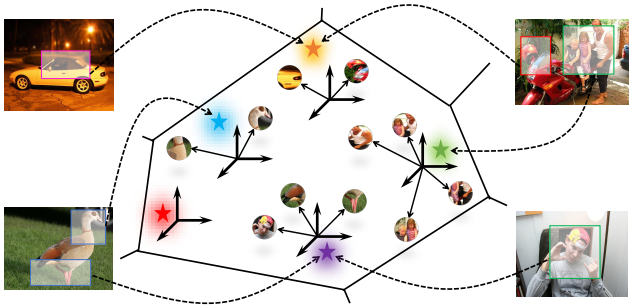


Figure 1. Universal prototypes (colorful stars) are learned from all object categories, which are not specific to certain object categories. Universal prototypes capture different intrinsical object characteristics via latent projection, e.g., the prototype ★ incorporates object characteristics of 'car' and 'motorbike'.

The main challenge of FSOD lies in how to learn generalized object features from both abundant samples in base categories and few samples in novel categories, which can simultaneously describe invariant object characteristics and alleviate the impact of unbalanced categories. Recently, meta-learning strategy [28, 30, 9] has been utilized in [38, 37, 35, 8] to adapt representation ability from base object categories to novel categories. However, the weak performance compared to basic fine-tuning methods [33, 36, 4, 5] shows the meta-learning technique fails to improve the generalization ability of object feature learning.

One possible reason is that the adaptation process in meta-learning mechanism could not capture the invariant characteristics across categories sufficiently. The invariance, i.e., invariant under different visual changes like textual variances or environmental noises, is always associated with the intrinsical object characteristics. As demonstrated in [23], the models that could extract invariant representations often generalize better than their non-invariant coun-

*Corresponding author

terparts. Therefore, in this paper, we explore how to enhance the generalization ability of object feature learning with the invariant object characteristics.

We devise universal prototypes (as shown in Fig. 1) to learn the invariant object characteristics. Different from the prototypes that are separately learned from each category [28, 20, 32], the proposed universal prototypes are learned from all object categories. The benefits are two-fold. On the one hand, prototypes from all categories capture rich information not only from different object categories but also from contexts of images. On the other hand, the universal prototypes reduce the impact of data-imbalance across different categories. Moreover, via fine-tuning, the universal prototypes can be effectively adapted to data-scarce novel categories. To this end, we develop a new framework of few-shot object detection with universal prototypes ($FSOD^{up}$). Particularly, we utilize a soft-attention of the learned universal prototypes to enhance the object features. Such a universal-prototype enhancement (i.e., each element of the enhanced features is a combination of prototypes) aims to simultaneously improve invariance and retain the semantic information of original object features. Here we employ a consistency loss to enable the maximum agreement between the enhanced and original object features. During training, we first train the model on data-abundant base categories. Then, the model is fine-tuned on a reconstructed training set that contains a small number of balanced training samples from both base and novel object categories. Experimental results on two benchmarks and extensive visualization analyses demonstrate the effectiveness of the proposed method. Our code will be available at `https://github.com/AmingWu/UP-FSOD`.

The contributions are summarized as follows:

(1) Towards FSOD, we devise a dedicated prototype and a new framework with universal-prototype enhancenment.

(2) We successfully demonstrate that, after fine-tuning with universal-prototype enhanced features, object detectors effectively adapt to novel categories.

(3) We obtain new performance on PASCAL VOC [7, 6] and MS COCO [19]. Enhancing invariance and generalization with the learned universal prototypes is empirically verified. Moreover, extensive visualization analyses also show that universal prototypes are capable of enhancing object characteristics, which is beneficial for FSOD.

2. Related Work

Few-shot image classification. Few-shot image classification [31, 24, 29, 13, 10] targets to recognize novel categories with only few samples in each category. Meta-learning is a widely used method to solve few-shot classification [22], which aims to leverage task-level meta knowledge to help the model adapt to new tasks with few labeled samples. Vinyals et al. [31] and Snell et al. [28] employed the meta-learning policy to learn the similarity metric that could be transferrable across different tasks. Particularly, based on the policy of meta-learning, prototypical network [28] is proposed to take the center of congener support samples' embeddings as the prototype of this category. The classification can be performed by computing distances between the representations of samples and prototype of each category. However, when the data is unbalanced or scarce, the learned prototypes could not represent the information of each category accurately, which affects the classification performance. Besides, during meta-learning, Gidaris et al. [10] and Wang et al. [34] introduced new parameters to promote the adaptation to novel tasks. However, these meta-learning methods for few-shot image classification could not be directly applied to object detection that requires localizing and recognizing objects.

Few-shot object detection. Most existing methods employ meta-learning [8, 17] or fine-tuning [39, 36] strategies to solve FSOD. Specifically, Wang et al. [35] developed a meta-learning based framework to leverage meta-level knowledge from data-abundant base categories to learn a detector for novel categories. Yan et al. [38] further extended Faster R-CNN [26] by performing meta-learning over RoI (Region-of-Interest) features. However, the weak performance compared to basic fine-tuning methods shows meta-learning based methods fail to improve the generalization ability of object detectors. For the method of fine-tuning and the model pre-trained on the base categories, Wang et al. [33] employed a two-stage fine-tuning process, i.e., fine-turning the last layers of the detector and freezing the other parameters of the detector, to make the object predictor adapt to novel categories. Wu et al. [36] proposed a method of multi-scale positive sample refinement to handle the problem of scale variations in object detection, which is similar to data augmentation [40].

Different from previous methods for FSOD, in this paper, we propose to learn universal prototypes from all object categories. And we develop a new framework of FSOD with universal-prototype enhancement. Experimental results and visualization analysis demonstrate the effectiveness of universal-prototype enhancement.

3. FSOD with Universal Prototypes

In this paper, we follow the same FSOD settings introduced in Kang et al. [16]. Annotated detection data are divided into a set of base categories that have abundant instances and a set of novel categories that have only few (usually less than 30) instances per category. The main purpose is to improve the generalization ability of detectors.

3.1. Learning of Universal Prototypes

Recently, many methods [28, 20, 32] construct a prototype for each category to solve few-shot image classifi-

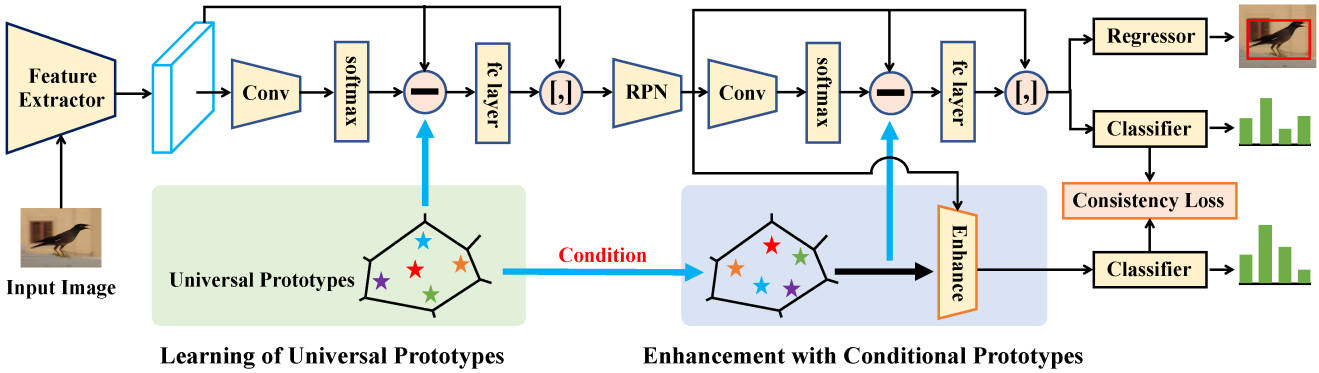


Figure 2. The architecture of few-shot object detection with universal-prototype enhancement. 'Conv' and 'fc layer' separately indicate convolution and fully-connected layer. The colorful stars are the learned universal prototypes. '⊖' and '[,]' denote the residual operation and concatenation operation, respectively. We focus on improving the generalization of detectors via learning invariant object characteristics. Firstly, universal prototypes are learned from all object categories. With the output of RPN (Region Proposal Network), we obtain the conditional prototypes via a conditional transformation of universal prototypes. Next, the enhanced object features are calculated based on conditional prototypes. Finally, a consistency loss is computed between the enhanced and original features.

cation. Though prototypes reflecting category information have been demonstrated to be effective for image classification, they could not be applied to FSOD. The reason may be that these category-specific prototypes represent image-level information and fail to capture object characteristics that are helpful for localizing and recognizing objects. Different from category-specific prototypes, based on all object categories, we attempt to learn universal prototypes that are beneficial for capturing intrinsical object characteristics that are invariant under different visual changes.

Concretely, the left part of Fig. 2 shows the learning process of universal prototypes. We adopt widely used Faster R-CNN [26], a two-stage object detector, as the base detection model. Given an input image, we first employ the feature extractor, e.g., ResNet [15], to extract corresponding features $F \in \mathbb{R}^{w \times h \times m}$, where w, h, and m separately denote width, height, and the number of channels. Then, the universal prototypes are defined as $C = \{c_i \in \mathbb{R}^m, i = 1, ..., D\}$. Next, based on the prototypical set C, we calculate descriptors that represent image-level information.

$$\mathcal{I} = W_g * F + b_g, \quad \text{共有D个UP，所以对F做3×3卷积转为D个channel再做channel-wise softmax，作用是感知F的每个location对每个UP的attention/weight是多少}$$

$$\mathcal{V}_i = \sum_{j=1}^{wh} \frac{e^{\mathcal{I}_{j,i}}}{\sum_{i=1}^{D} e^{\mathcal{I}_{j,i}}} (F_j - c_i), \quad (1)$$

共有D个UP，对于每个UP，F的每个location都要减去它。最终对所有location进行求和，以感知这张图片对每个UP的weight

where $W_g \in \mathbb{R}^{3 \times 3 \times m \times D}$ and $b_g \in \mathbb{R}^D$ are convolutional parameters. $\mathcal{V} \in \mathbb{R}^{D \times m}$ represents the output descriptors. '$F_j - c_i$' indicates the residual operation, by which the visual features can be assigned to the corresponding prototype. Finally, we take the concatenated result of F and $\mathcal{V}$ as the input of the RPN module.

$$P = \text{RPN}(\Psi([F,\ \mathcal{V}_r W_p + b_p])), \quad (2)$$

where $\mathcal{V}_r \in \mathbb{R}^{1 \times Dm}$ is the reshaped result of $\mathcal{V}$. Meanwhile, $W_p \in \mathbb{R}^{Dm \times m}$ and $b_p \in \mathbb{R}^m$ are parameters of the fully-connected layer. '[,]' is the concatenation operation. By the concatenation operation, the descriptors $\mathcal{V}$ can be fused into the original features F, which enhances the representation ability of F. Ψ consists of two convolutional layers with ReLU activation and is used to transform the concatenated result. Finally, $P \in \mathbb{R}^{n \times s \times s \times m}$ is the output of RPN with RoI Pooling [26, 14], where n and s separately indicate the number of proposals and the size of proposals. The feature dimension of P is the same as F.

3.2. Enhancement of Object Features

As shown in the right part of Fig. 2, we first compute conditional prototypes based on the universal prototypes C. Then, we conduct enhancement of object features with the conditional prototypes.

3.2.1 The Computation of Conditional Prototypes

Since the computation of Eq. (1) is based on the extracted features that represent the whole input image, the universal prototypes C mainly reflect image-level information. Here, the image-level information includes object-level information and other associated information about image content. Whereas, after RPN, the proposal features P mainly contain object-level information. The directly using of universal prototypes C may not accurately represent object-level information. Thus, we make an affine transformation to promote C to move towards the space of object-level features.

$$\mathcal{A} = \alpha \odot C + \beta, \quad (3)$$

where $\alpha \in \mathbb{R}^{D \times 1}$ and $\beta \in \mathbb{R}^{D \times 1}$ are the transformed parameters. $\odot$ is element-wise product. Finally, $\mathcal{A} \in \mathbb{R}^{D \times m}$ represents the conditional prototypes. Next, we employ the same processes as Eq. (1) to generate object-level descrip-

tors. The processes are shown as follows:

$$E = W_c * P + b_c,$$
$$O_{k,i} = \sum_{j=1}^{s^2} \frac{e^{E_{k,j,i}}}{\sum_{i=1}^{D} e^{E_{k,j,i}}} (P_{k,j} - a_i), \tag{4}$$

where $k = 1, \cdots, n$. $W_c \in \mathbb{R}^{3\times3\times m\times D}$ and $b_c \in \mathbb{R}^D$ are convolutional parameters. $a_i \in \mathbb{R}^{1\times m}$ is the i-th conditional prototype of $\mathcal{A}$. $O \in \mathbb{R}^{n\times D\times m}$ indicates the output descriptors. Finally, we take the concatenated result of P and O as the input of the classifier.

$$y = \mathrm{Clf}([\Psi_c(P),\ O_r W_r + b_r]), \tag{5}$$

where $O_r \in \mathbb{R}^{n\times Dm}$ is the reshaped result of O. Clf denotes the classifier. Meanwhile, $W_r \in \mathbb{R}^{Dm\times 2m}$ and $b_r \in \mathbb{R}^{2m}$ are parameters of the fully-connected layer. Ψ_c consists of two fully-connected layers and outputs a matrix with the dimension $n \times 2m$. Finally, y is the predicted probability. In the experiment, we find employing the descriptors O generated based on the conditional prototypes improves the performance of FSOD, which shows the effectiveness of conditional prototypes.

3.2.2 Enhancement with Conditional Prototypes

In order to improve the generalization of detectors, we explore to utilize conditional prototypes to enhance object features. Specifically, Fig. 3 shows the enhancement details. For proposal features $P \in \mathbb{R}^{n\times s\times s\times m}$ and conditional prototypes $\mathcal{A} \in \mathbb{R}^{D\times m}$, we separately employ a convolutional layer $\Phi_p \in \mathbb{R}^{1\times1\times m\times m}$ and fully-connected layer $\Phi_a \in \mathbb{R}^{m\times m}$ to project P and $\mathcal{A}$ into an embedding space, i.e., $e_p = \Phi_p(P)$ and $e_a = \Phi_a(\mathcal{A})$. Then, based on each element of e_p, we calculate the soft-attention of e_a to obtain enhancement of object features.

$$\lambda_k = \mathrm{softmax}(e_{p,k} e_a^T),$$
$$Enh_k = \mathrm{ReLU}(\Phi_t([e_{p,k},\ \lambda_k e_a]) + P_k), \tag{6}$$

where $k = 1, \cdots, n$. $e_{p,k} \in \mathbb{R}^{s^2\times m}$ indicates the k-th component of e_p. $\lambda_k \in \mathbb{R}^{s^2\times D}$ denotes attention weights. Φ_t consists of two convolutional layers with ReLU activation. And the output dimension of Φ_t is m. $P_k \in \mathbb{R}^{s\times s\times m}$ is the k-th component of P. Finally, $Enh \in \mathbb{R}^{n\times s\times s\times m}$ is the enhanced object features, which fuses the information of conditional prototypes and is helpful for improving the generalization on novel objects. Next, Enh is taken as the input of the classifier to output the predicted probability.

$$y_{enh} = \mathrm{Clf}([\Psi_c(Enh),\ \Psi_c(P)]), \tag{7}$$

where y_{enh} is the predicted probability. Besides, Eq. (5) and Eq. (7) share the same classifier. In the experiment, we find the enhanced operations (Eq. (6) and (7)) are beneficial for FSOD, which further indicates the learned prototypes contain object-level information.

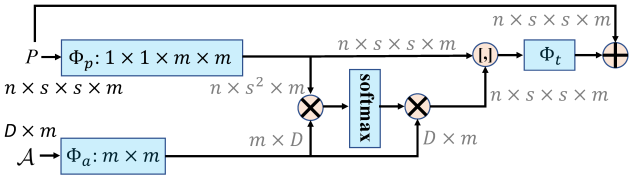


Figure 3. Enhancement of object features. Based on each element of RPN output P, we calculate the soft-attention of the conditional prototypes $\mathcal{A}$ to generate enhanced features. Each element of the enhanced features is a combination of conditional prototypes, which retains the semantic information of P.

3.3. Two-stage Fine-tuning Approach

Many semi-supervised learning methods [2, 1] rely on a consistency loss to enforce that the model output remains unchanged when the input is perturbed. Inspired by this idea, to learn invariant object characteristics, we compute the consistency loss between the prediction y from original features (see Eq. (5)) and the prediction y_{enh} from enhanced features. Particularly, the KL-Divergence loss is employed to enforce consistent predictions, i.e., $\mathcal{L}_{\mathrm{con}} = \mathcal{H}(y,\ y_{enh})$. The joint training loss is defined as follows:

$$\mathcal{L} = \mathcal{L}_{\mathrm{rpn}} + \mathcal{L}_{\mathrm{cls}} + \mathcal{L}_{\mathrm{loc}} + \gamma \mathcal{L}_{\mathrm{con}}, \tag{8}$$

where $\mathcal{L}_{\mathrm{rpn}}$ is the loss of the RPN to distinguish foreground from background and refine bounding-box anchors. $\mathcal{L}_{\mathrm{cls}}$ and $\mathcal{L}_{\mathrm{loc}}$ separately indicate classification loss and box regression loss. And γ is a hyper-parameter.

During training, we employ a two-stage fine-tuning approach (as shown in Fig. 4) to optimize $FSOD^{up}$ model. Concretely, in the base training stage, we employ the joint loss $\mathcal{L}$ to optimize the entire model based on the data-abundant base classes. After the base training stage, only the last fully-connected layer (for classification) of the detection head is replaced. The new classification layer is randomly initialized. Besides, during few-shot fine-tuning stage, different from the work [33], none of the network layers is frozen. And we still employ the loss $\mathcal{L}$ to fine-tune the entire model based on a balanced training set consisting of both the few base and novel categories.

3.4. Discussion

In this section, we further discuss universal prototypes for few-shot object detection.

Though prototypes have been demonstrated to be effective for few-shot image classification [28, 31], it is unclear how to build prototypes for FSOD [16]. (1) If we follow few-shot image classification and construct prototypes for each category, the computational costs increase for the case

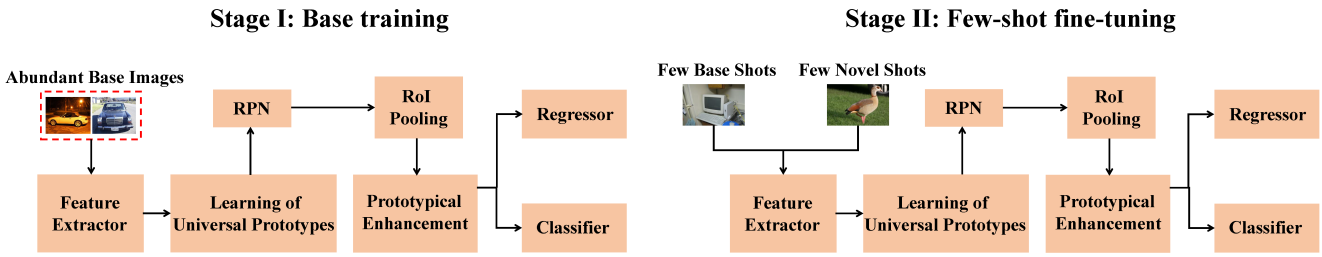


Figure 4. Illustration of two-stage fine-tuning approach for $FSOD^{up}$. In the base training stage, the entire detector, including the feature extractor, the module for learning of universal prototypes, and the module for enhancement based on conditional prototypes, are jointly trained on the data-abundant base categories. In the few-shot fine-tuning stage, the entire detector is fine-tuned on a balanced training set consisting of both the few base and novel categories.

of a large number of object categories. Meanwhile, due to the unbalanced object categories, the constructed prototypes may not accurately reflect category information. (2) Related to the above, detectors for certain object category can be affected by co-appearing objects in one image, and thus the quality of the constructed prototype for such category may be burdened. (3) More importantly, since the number of object categories in the stage of the base training is different from that of the few-shot fine-tuning, constructing a prototype for each object category makes it impossible to align the prototypes between the base training and the few-shot fine-tuning. That is to say, the prototypes pre-trained on base categories cannot be directly utilized in the fine-tuning stage. Therefore, for fine-tuning based methods, it is difficult to build a prototype for each category.

To solve FSOD, we propose to learn universal prototypes from all object categories. The universal prototypes are not specific to certain object categories and can be effectively adapted to novel categories via fine-tuning. In the experiments, we find that the universal prototypes are helpful for characterizing the regional information of different object categories. Meanwhile, with the help of universal-prototype enhancement, the performance of few-shot detection can be significantly improved.

4. Experiments

We first evaluate our method on PASCAL VOC [7, 6] and MS COCO [19]. For a fair comparison, we use the settings in [16, 38] to construct few-shot detection datasets. Concretely, for PASCAL VOC, the 20 classes are randomly divided into 5 novel classes and 15 base classes. Here, we follow the work [16] to use the same three class splits, where only K object instances are available for each novel category and K is set to 1, 2, 3, 5, 10. For MS COCO, the 20 categories overlapped with PASCAL VOC are used as novel categories with K = 10, 30. And the remaining 60 categories are taken as the base categories.

Implementation Details. Faster R-CNN [26] is used as the base detector. Our backbone is Resnet-101 [15] with the RoI Align [14] layer. We use the weights pre-trained on ImageNet [27] in initialization. For FSOD, the number of universal prototypes (see Eq. (1)) is set to 24. All these prototypes are randomly initialized. Next, the model is trained with a batchsize of 2 on 2 GPUs, 1 image per GPU. Meanwhile, to alleviate the impact of the scale issue, we employ the positive sample refinement [36]. The hyper-parameter γ (see Eq. (8)) is set to 1.0. All models are trained using SGD optimizer with a momentum of 0.9 and a weight decay of 0.0001. Finally, during inference, we take the output u of Eq. (5) as the classification result.

4.1. Performance Analysis of Few-Shot Detection

We compare $FSOD^{up}$ with two baseline methods, i.e., TFA [33] and MPSR [36]. These two approaches all use the two-stage fine-tuning method to solve FSOD.

Results on PASCAL VOC. Table 1 shows the results of PASCAL VOC. As the number of novel categories decreases, the performance degrades significantly. This indicates that addressing the few-shot problem is crucial to improve the generalization of detectors. We can see that the proposed $FSOD^{up}$ method consistently outperforms the two baseline methods. This shows that employing universal-prototype enhancement is helpful for learning invariant object characteristics and thus improves performance. Meanwhile, this also indicates that focusing on invariance plays a key role in solving FSOD.

In Fig. 5, we show the detection results of MPSR [36] and our method. 'bird' and 'bus' belong to the novel categories. We can see that our method can successfully detect objects existing in images. This further shows that the proposed universal-prototype enhancement is helpful for capturing invariant object characteristics, which improves the accuracy of detection.

Results on MS COCO. Table 2 shows the few-shot detection performance on MS COCO dataset. Compared with two baseline methods, i.e., TFA [33] and MPSR [36], our method consistently outperforms their performance. This further demonstrates the effectiveness of the proposed universal-prototype enhancement. Besides, FSOD-VE [37] is a recently proposed meta-learning based method, which

	Novel Set 1					Novel Set 2					Novel Set 3				
Method / Shot	1	2	3	5	10	1	2	3	5	10	1	2	3	5	10
Meta R-CNN [38]	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
RepMet [17]	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
FSOD-VE [37]	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
TFA w/fc [33]	36.8	29.1	43.6	**55.7**	57.0	18.2	29.0	33.4	35.5	39.0	27.7	33.6	42.5	48.7	50.2
TFA w/cos [33]	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
TFA$^\sharp$ w/fc [37, 33]	22.9	34.5	40.4	46.7	52.0	16.9	26.4	30.5	34.6	39.7	15.7	27.2	34.7	40.8	44.6
TFA$^\sharp$ w/cos [37, 33]	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
MPSR$^\sharp$ [36]	40.7	41.2	48.9	53.6	60.3	24.4	29.3	39.2	39.9	47.8	32.9	34.4	42.3	48.0	49.2
Ours ($FSOD^{up}$)	**43.8**	**47.8**	**50.3**	**55.4**	**61.7**	**31.2**	**30.5**	**41.2**	**42.2**	**48.3**	**35.5**	**39.7**	**43.9**	**50.6**	**53.5**

Table 1. Few-shot detection performance (mAP (%)) on PASCAL VOC dataset. We evaluate the performance on three different sets of novel categories. Resnet-101 [15] is used as the backbone. '♯' indicates that we directly run the released code to obtain the results.

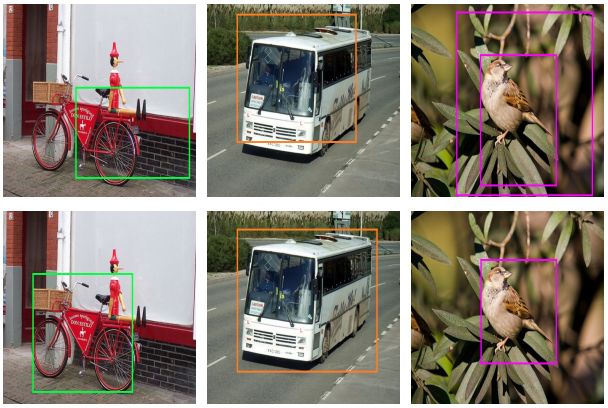


Figure 5. Detection results based on the 5-shot case. The first row shows the results of MPSR [36]. The second row is our detection results. Our method detects the objects accurately.

combines FSOD with a few-shot viewpoint estimation and follows Meta R-CNN [38] to optimize detectors. Though FSOD-VE's performance of the 10-shot case is higher than our method, our method outperforms FSOD-VE on the small objects. Meanwhile, compared with FSOD-VE, the training of our method is much easier. And we do not use the viewpoint information. These results further demonstrate that exploiting universal-prototype enhancement is helpful for improving detectors' generalization.

4.2. Ablation Analysis

In this section, based on the Novel Set 1 of PASCAL VOC, we make an ablation analysis of our method.

Conditional prototypes. In order to sufficiently represent object-level information, based on the universal prototypes C (see Eq. (1)), we make an affine transformation to obtain conditional prototypes $\mathcal{A}$ (see Eq. (3)). Next, we make an ablation analysis of conditional prototypes.

Table 3 shows the comparison results. We can see that utilizing the conditional operation improves detection performance significantly. Particularly, for the 2-shot case, our method separately outperforms 'No Condition' and 'New Prototype' by 4.0% and 3.2%. This shows that based on the universal prototypes, the conditional prototypes represent object-level information effectively, which improves the performance of detection.

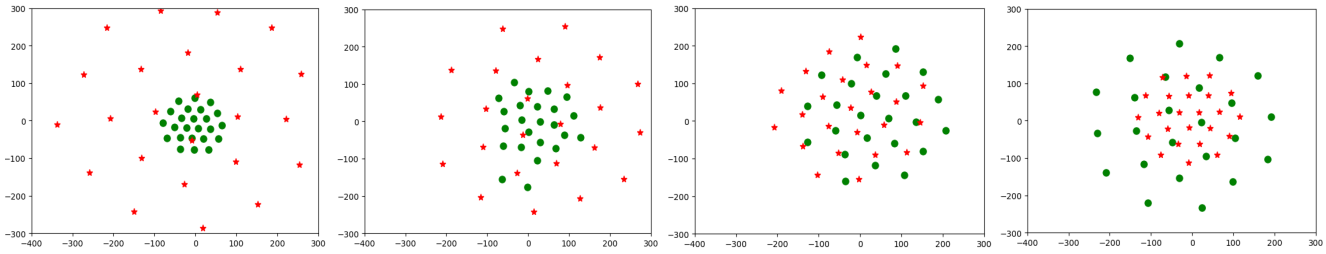
The number of universal prototypes. For our method,

Shots	Method	AP	AP75	AP_S	AP_M	AP_L
10	Meta R-CNN [38]	8.7	6.6	2.3	7.7	14.0
	FSOD-VE [37]	**12.5**	9.8	2.5	**13.8**	**19.9**
	TFA w/fc [33]	10.0	9.2	–	–	–
	TFA w/cos [33]	10.0	9.3	–	–	–
	TFA$^\sharp$ w/fc [37, 33]	9.1	8.5	–	–	–
	TFA$^\sharp$ w/cos [37, 33]	9.1	8.8	–	–	–
	MPSR$^\sharp$ [36]	9.5	9.5	3.3	8.2	15.9
	Ours ($FSOD^{up}$)	11.0	**10.7**	**4.5**	11.2	17.3
30	Meta R-CNN [38]	12.4	10.8	2.8	11.6	19.0
	FSOD-VE [37]	14.7	12.2	3.2	**15.2**	23.8
	TFA w/fc [33]	13.4	13.2	–	–	–
	TFA w/cos [33]	13.7	13.4	–	–	–
	TFA$^\sharp$ w/fc [37, 33]	12.0	11.8	–	–	–
	TFA$^\sharp$ w/cos [37, 33]	12.1	12.0	–	–	–
	MPSR$^\sharp$ [36]	13.8	13.5	4.0	12.9	22.9
	Ours ($FSOD^{up}$)	**15.6**	**15.7**	**4.7**	15.1	**25.1**

Table 2. Few-shot detection performance (%) on MS COCO dataset. Here, AP_S, AP_M, and AP_L separately indicate the mAP performance of the small, medium, and large objects.

method/shot	1	2	3	5	10
No Condition	38.1	43.8	48.9	55.6	60.6
New Prototype	42.1	44.6	48.8	**56.1**	60.1
Ours	**43.8**	**47.8**	**50.3**	55.4	**61.7**

Table 3. Analysis of conditional prototypes. Here, 'No Condition' indicates we do not use conditional operation in Eq. (3) and directly use the universal prototypes C to make enhancement. 'New Prototype' indicates we define a new set of prototypes to replace the conditional prototypes.



(a) 1-shot (b) 3-shot (c) 5-shot (d) 10-shot

Figure 6. The t-SNE plot of prototypes. We analyze the impact of employing different shots. Here, the number of prototypes is 24. ● and ★ separately denote the universal prototypes (see Eq. (1)) and conditional prototypes (see Eq. (3)). For novel categories, using a different number of samples affects the distribution of the universal and conditional prototypes. As the number of novel objects increases, the universal prototypes become more scattered, whereas the conditional ones become more concentrated.

number/shot	1	2	3	5	10
16	41.2	42.7	48.3	54.2	60.1
20	42.5	44.1	50.1	56.0	60.5
24	**43.8**	**47.8**	**50.3**	55.4	**61.7**
28	42.6	44.6	49.6	**56.7**	60.6
32	41.4	42.1	49.6	53.9	60.0

Table 4. The impact of the number of universal prototypes. Here, we only utilize a different number of prototypes and keep other components unchanged.

		Novel Classes					Mean	
Shot	Method	bird	bus	cow	mbike	sofa	Novel	Base
2	MPSR$^\sharp$ [36]	36.8	24.8	56.9	59.1	28.4	41.2	65.4
	Ours ($FSOD^{up}$)	**40.7**	**41.3**	**58.9**	**62.2**	**35.9**	**47.8**	**66.3**
5	MPSR$^\sharp$ [36]	44.1	**60.7**	54.3	**66.8**	42.1	53.6	69.5
	Ours ($FSOD^{up}$)	**47.0**	60.5	**57.3**	66.4	**46.1**	**55.4**	**69.7**

Table 5. AP (%) of each novel category on the 2-/5-shot case. We also present mAP (%) of novel and base categories.

the number of universal prototypes (see Eq. (1)) is an important hyper-parameter. If the number is small, these prototypes could not represent invariant object characteristics sufficiently. On the contrary, a large number of prototypes may increase parameters and computational costs.

Table 4 shows the performance of employing a different number of prototypes. We can see that the performance of utilizing 24 prototypes is the best. When the number is larger or fewer than 24, the performance degrades significantly. This shows the number of prototypes affects FSOD performance. In general, for the case of a large-scale dataset with a large number of categories, employing more prototypes could capture object-level characteristics sufficiently, which is helpful for improving detectors' generalization on novel object categories.

Visualization analysis of prototype distribution. In Fig. 6, based on different shots, we analyze the distribution of prototypes. Concretely, as the number of novel objects increases, in order to improve the detection performance, the universal prototypes (see Eq. (1)) will become more scattered to capture more image-level information. After RPN, the conditional prototypes are calculated to represent object-level information. And the features calculated based on the conditional prototypes are used for classification. Thus, as the number of novel objects increases, the distribution of the conditional prototypes will become more concentrated to focus on specific categories, which could improve the accuracy of detection. These analyses further show universal prototypes are capable of enhancing feature representations, which is beneficial for FSOD.

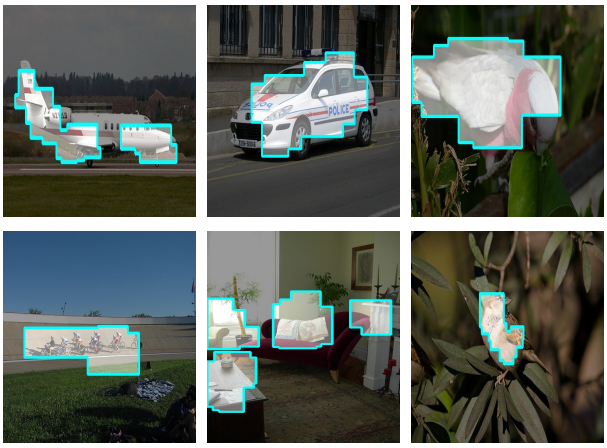
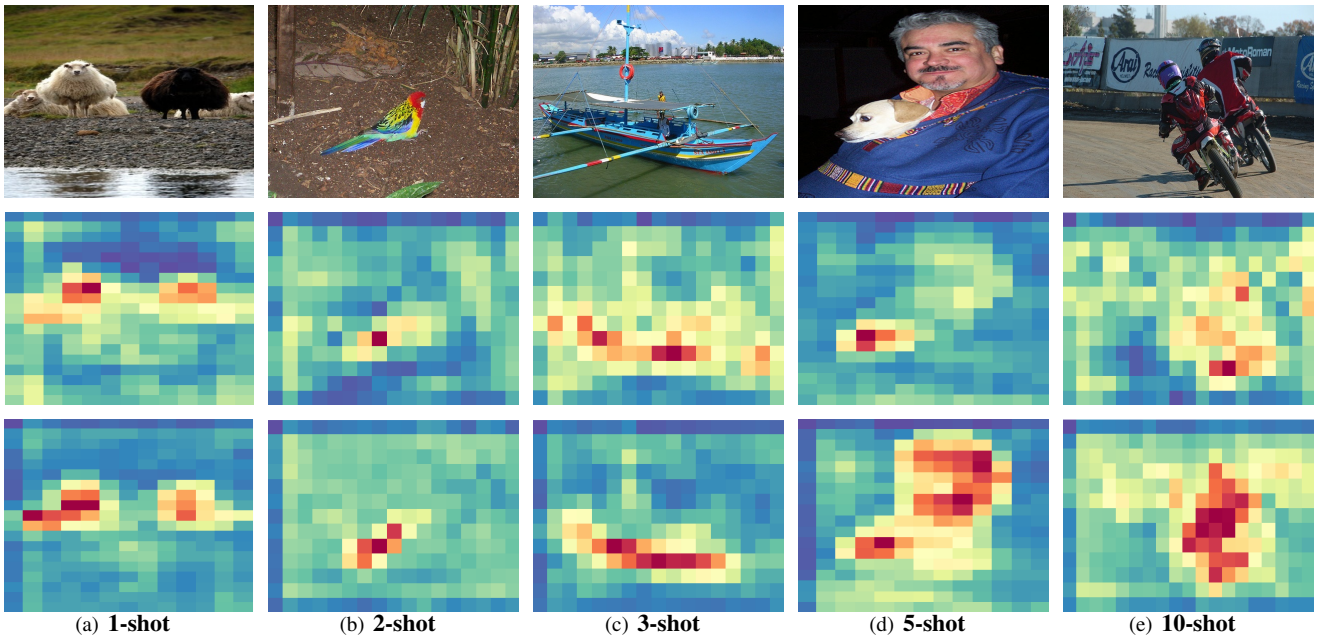


Figure 7. Assignment of image regions to universal prototypes based on the 5-shot case. The highlight regions in each image are assigned to one same prototype, respectively.

Visualization of assignment maps. In Fig. 7, we visualize the assignment maps of universal prototypes, i.e., the soft-assignment $\frac{e^{\mathcal{I}_{j,i}}}{\sum_{i=1}^{D} e^{\mathcal{I}_{j,i}}}$ in Eq. (1). For each image, we can see that different object regions are assigned to one same universal prototype. Particularly, for the second image of the second row, the object regions of 'sofa' and 'table' are all assigned to one same prototype. This indicates the universal prototypes are not specific to certain object categories. Moreover, the universal prototypes are helpful for characterizing the region information of different objects and could be effectively adapted to novel categories via fine-tuning.



(a) **1-shot** (b) **2-shot** (c) **3-shot** (d) **5-shot** (e) **10-shot**

Figure 8. Visualization of the feature map used for RPN based on different shots. The second and third row separately indicate F and the output of Ψ (see Eq. (2)). For each feature map, the channels corresponding to the maximum value are selected for visualization.

setting/shot	1	2	3	5	10
1.4	43.2	43.5	49.0	54.8	61.0
1.2	41.1	42.1	50.2	54.3	60.7
1.0	**43.8**	**47.8**	**50.3**	55.4	**61.7**
0.8	39.7	42.5	49.0	56.0	60.5
0.6	40.8	43.3	50.1	**56.9**	60.6

Table 6. Ablation analysis of the hyper-parameter γ.

The performance of base categories. Table 5 shows the performance of each novel and base categories. We can see that our method outperforms MPSR [36] on novel and base categories. Particularly, for the 'bus' and 'sofa' category of the 2-shot case, our method outperforms MPSR by 16.5% and 7.5%. This indicates our method could improve the generalization performance of the detector.

Analysis of Hyper-Parameter γ. For the joint training loss $\mathcal{L}$ (see Eq. (8))), we use a hyper-parameter γ to balance the consistency loss $\mathcal{L}_{\text{con}}$. Table 6 shows the results. We can see that different settings of the hyper-parameter γ affect the performance of FSOD. For our method, when γ is set to 1.0, the performance is the best.

Analysis of the output descriptors. In Eq. (2) and (5), the output descriptors are fused as the input of the RPN and classifier. Next, we analyze the impact of the descriptors. Concretely, for Eq. (2), we only take F as the input of RPN and keep other components unchanged. For the 1-shot and 5-shot case, fusing the descriptors improves the performance by 2.7% and 1.8%. For Eq. (5), we only take $\Psi_c(P)$ as the input of classifier and keep other components unchanged. For the 1-shot and 5-shot case, fusing the descriptors improves the performance by 2.1% and 1.2%. This shows fusing descriptors into the current features is helpful for improving the representation ability of the features.

In Fig. 8, based on different shots, we show visualization results of F and the output of Ψ (see Eq. (2)). Here, we separately take F and the output of Ψ as the input of RPN. We can see that for the base and novel categories, compared with F, the output of Ψ contains more object-related information. Taking the 5-shot result as an example, the output of our method (the fourth image of the third row) contains more information about 'Person' category. This further indicates fusing descriptors is helpful for enhancing the object-level information.

5. Conclusion

To solve FSOD, we propose to learn universal prototypes from all object categories. Meanwhile, we develop an approach of few-shot object detection with universal prototypes ($FSOD^{up}$). Concretely, after obtaining the universal and conditional prototypes, the enhanced object features are computed based on the conditional prototypes. Next, through a consistency loss, $FSOD^{up}$ enhances the invariance and generalization. Experimental results on two datasets show the effectiveness of the proposed method.

Acknowledgement

This work is supported by the NSFC (under Grant 61876130, 61932009).

References

[1] David Berthelot, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. *arXiv preprint arXiv:1911.09785*, 2019.

[2] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *Advances in Neural Information Processing Systems*, pages 5049–5059, 2019.

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. *European Conference on Computer Vision*, 2020.

[4] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. A closer look at few-shot classification. In *ICLR 2019 : 7th International Conference on Learning Representations*, 2019.

[5] Guneet Singh Dhillon, Pratik Chaudhari, Avinash Ravichandran, and Stefano Soatto. A baseline for few-shot image classification. In *ICLR 2020 : Eighth International Conference on Learning Representations*, 2020.

[6] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111(1):98–136, 2015.

[7] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[8] Qi Fan, Wei Zhuo, Chi-Keung Tang, and Yu-Wing Tai. Few-shot object detection with attention-rpn and multi-relation detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4013–4022, 2020.

[9] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML'17 Proceedings of the 34th International Conference on Machine Learning - Volume 70*, pages 1126–1135, 2017.

[10] Spyros Gidaris and Nikos Komodakis. Dynamic few-shot visual learning without forgetting. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4367–4375, 2018.

[11] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[12] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[13] Bharath Hariharan and Ross Girshick. Low-shot visual recognition by shrinking and hallucinating features. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3037–3046, 2017.

[14] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[16] Bingyi Kang, Zhuang Liu, Xin Wang, Fisher Yu, Jiashi Feng, and Trevor Darrell. Few-shot object detection via feature reweighting. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8420–8429, 2019.

[17] Leonid Karlinsky, Joseph Shtok, Sivan Harary, Eli Schwartz, Amit Aides, Rogerio Feris, Raja Giryes, and Alex M. Bronstein. Repmet: Representative-based metric learning for classification and few-shot object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[18] Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 936–944, 2017.

[19] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[20] Jinlu Liu, Liang Song, and Yongqiang Qin. Prototype rectification for few-shot learning. *European Conference on Computer Vision*, 2020.

[21] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.

[22] Jiang Lu, Pinghua Gong, Jieping Ye, and Changshui Zhang. Learning from very few samples: A survey. *arXiv preprint arXiv:2009.02653*, 2020.

[23] Clare Lyle, Marta Kwiatkowksa, and Yarin Gal. An analysis of the effect of invariance on generalization in neural networks. In *International conference on machine learning Workshop on Understanding and Improving Generalization in Deep Learning*, 2019.

[24] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *ICLR 2017 : International Conference on Learning Representations 2017*, 2017.

[25] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7263–7271, 2017.

[26] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[27] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[28] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, pages 4077–4087, 2017.

[29] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip H.S. Torr, and Timothy M. Hospedales. Learning to compare: Relation network for few-shot learning. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1199–1208, 2018.

[30] Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B. Tenenbaum, and Phillip Isola. Rethinking few-shot image classification: a good embedding is all you need? *arXiv preprint arXiv:2003.11539*, 2020.

[31] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *Advances in neural information processing systems*, pages 3630–3638, 2016.

[32] Kaixin Wang, Jun Hao Liew, Yingtian Zou, Daquan Zhou, and Jiashi Feng. Panet: Few-shot image semantic segmentation with prototype alignment. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9197–9206, 2019.

[33] Xin Wang, Thomas E Huang, Trevor Darrell, Joseph E Gonzalez, and Fisher Yu. Frustratingly simple few-shot object detection. *International Conference on Machine Learning*, 2020.

[34] Xin Wang, Fisher Yu, Ruth Wang, Trevor Darrell, and Joseph E Gonzalez. Tafe-net: Task-aware feature embeddings for low shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1831–1840, 2019.

[35] Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. Meta-learning to detect rare objects. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9925–9934, 2019.

[36] Jiaxi Wu, Songtao Liu, Di Huang, and Yunhong Wang. Multi-scale positive sample refinement for few-shot object detection. *European Conference on Computer Vision*, 2020.

[37] Yang Xiao and Renaud Marlet. Few-shot object detection and viewpoint estimation for objects in the wild. *European Conference on Computer Vision*, 2020.

[38] Xiaopeng Yan, Ziliang Chen, Anni Xu, Xiaoxi Wang, Xiaodan Liang, and Liang Lin. Meta r-cnn: Towards general solver for instance-level low-shot learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9577–9586, 2019.

[39] Ze Yang, Yali Wang, Xianyu Chen, Jianzhuang Liu, and Yu Qiao. Context-transformer: Tackling object confusion for few-shot detection. In *AAAI 2020 : The Thirty-Fourth AAAI Conference on Artificial Intelligence*, volume 34, pages 12653–12660, 2020.

[40] Barret Zoph, Ekin D Cubuk, Golnaz Ghiasi, Tsung-Yi Lin, Jonathon Shlens, and Quoc V Le. Learning data augmentation strategies for object detection. *arXiv preprint arXiv:1906.11172*, 2019.