Few-Shot Object Detection by Knowledge Distillation Using Bag-of-Visual-Words Representations

Wenjie Pei[2,†], Shuang Wu[2,†], Dianwen Mei[2], Fanglin Chen[2], Jiandong Tian[3], and Guangming Lu[1,2,*]

[1] Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies
[2] Harbin Institute of Technology, Shenzhen, China
[3] Shenyang Institute of Automation, Chinese Academy of Sciences
{wenjiecoder, wushuang9811}@outlook.com, {178mdw, linwers}@gmail.com, luguangm@hit.edu.cn, tianjd@sia.cn

Abstract. While fine-tuning based methods for few-shot object detection have achieved remarkable progress, a crucial challenge that has not been addressed well is the potential class-specific overfitting on base classes and sample-specific overfitting on novel classes. In this work we design a novel knowledge distillation framework to guide the learning of the object detector and thereby restrain the overfitting in both the pre-training stage on base classes and fine-tuning stage on novel classes. To be specific, we first present a novel Position-Aware Bag-of-Visual-Words model for learning a representative bag of visual words ($BoVW$) from a limited size of image set, which is used to encode general images based on the similarities between the learned visual words and an image. Then we perform knowledge distillation based on the fact that an image should have consistent $BoVW$ representations in two different feature spaces. To this end, we pre-learn a feature space independently from the object detection, and encode images using $BoVW$ in this space. The obtained $BoVW$ representation for an image can be considered as distilled knowledge to guide the learning of object detector: the extracted features by the object detector for the same image are expected to derive the consistent $BoVW$ representations with the distilled knowledge. Extensive experiments validate the effectiveness of our method and demonstrate the superiority over other state-of-the-art methods.

Keywords: Few-Shot Object Detection, Bag of Visual Words, Knowledge Distillation

1 Introduction

Few-shot object detection aims to learn effective object detectors on a set of base classes with sufficient samples, which can be generalized efficiently to novel

[†] Equal contribution.
[*] Corresponding author.

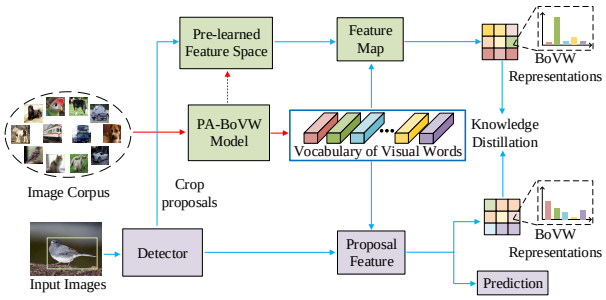


Fig. 1. We learn a representative bag of visual words ($BoVW$) using the proposed $PA\text{-}BoVW$ model. For an extracted positive proposal, we encode it with $BoVW$ in a pre-learned feature space and the feature space of the object detector, respectively. Then we perform knowledge distillation by matching two $BoVW$ representations to guide the learning of the object detector.

classes with only a few samples available. Thus, few-shot object detection eliminates exhaustive label annotation of massive data on novel classes. Compared to general object detection [27,31,32,40], few-shot object detection [3,19] is much more challenging due to the difficulty of learning generalizable features that can be transferred from base classes to novel classes.

A classical type of methods for few-shot object detection is fine-tuning based methods [1,9,26,30,38,44,48,49,57], which first train the object detector using the samples from base classes, then fine-tune the model on novel classes. A prominent example is TFA [44], which first adopts such two-stage training strategy to transfer knowledge from base classes to novel classes. Based on such fine-tuning framework, many methods are proposed to deal with various challenges of few-shot object detection. Typical methods include FSCE [38] aiming to facilitate the separability among similar classes, MPSR [49] which seeks to rectify the sample distribution for novel classes, and HallucFsDet [57] which is designed to tackle the problem of data scarcity.

A crucial challenge of fine-tuning based framework for few-shot object detection is the potential class-specific overfitting on base classes and sample-specific overfitting on novel classes. On the one hand, although sufficient samples are provided for base classes, the object detector is still prone to overfitting on base classes during the first stage of training process. In this case, the detector learns the class-specific features instead of the class-agnostic features, which cannot be transferred to novel classes and would adversely affect object detection for novel classes. On the other hand, owing to the scarcity of training samples for novel classes in the fine-tuning stage, the object detector tends to be overfitting on these individual samples and thus learns sample-specific features that cannot be generalized across different samples for a same novel class.

To address above limitation, we propose to perform knowledge distillation to guide the learning process of few-shot object detection and thus restrain the potential overfitting on both base classes and novel classes. As shown in Figure 1,

we propose the novel Position-Aware Bag-of-Visual-Words ($PA\text{-}BoVW$) model, which is able to learn a bag of visual words ($BoVW$) from a limited size of image set. The learned visual words are representative and comprehensive to be capable of encoding general images based on the similarities between the learned visual words and an image. Then we can perform knowledge distillation based on the intuition that an image should have consistent $BoVW$ representations in two different feature spaces, provided that the image is encoded properly, namely not overfitted, in both feature spaces. Concretely, we first pre-learn a feature space and derive a $BoVW$ representation for an image in this space. The obtained $BoVW$ representation can be considered as distilled knowledge to guide the learning of object detector: the extracted features by the object detector for the same image are expected to derive the consistent $BoVW$ representation with the distilled knowledge.

Unlike typical way that identifies visual words as the clustering centroids in the deep feature space [11,18], we learn visual words as learnable vectorial embeddings. To be specific, our proposed $PA\text{-}BoVW$ model first constructs an effective deep embedding space for learning the visual words by training a backbone network in a self-supervised way. Then the visual words is learned in this embedding space in a supervised way employing image classification as a pretext task. Besides, we employ DeCov loss [5] as an auxiliary loss to reduce the inter-word redundancy and encourage the diversity of visual words. As a result, the $PA\text{-}BoVW$ model is learned in an independent way from the task of object detection. Thus the encoded $BoVW$ representation in its embedding space can be used as distilled knowledge, which can be transferred to the learning process of the detector to restrain potential overfitting on both base and novel classes.

To conclude, we make following contributions: 1) We propose the novel $PA\text{-}BoVW$ model, which constructs an effective embedding space to learn a representative vocabulary of visual words; 2) Based on the $PA\text{-}BoVW$ model, we design a knowledge distillation framework to guide the learning of the object detector and thereby restrain the potential overfitting on both base classes and novel classes; 3) Extensive experiments validate the effectiveness of our method and demonstrate the advantages of our method over state-of-the-art methods for few-shot object detection.

2 Related Work

Few-shot learning. Early works of few-shot learning focus on the task of image classification. Metric-based methods learn a suitable embedding space, where samples can be categorized correctly via a nearest neighbor classifier with Euclidean distance [37], cosine similarity [4,41] or graph distance [21,36,53]. Initialization-based methods aim to learn good initialization so that the model can adapt to novel tasks by a few optimization steps [10,23]. Hallucination-based methods alleviate data scarcity issue via learning generators to augment novel classes [14,45]. However, these approaches could not be directly applied to few-shot object detection which requires both classification and localization.

Few-shot object detection. Few-shot object detection aims to detect objects with few annotated training examples provided. There are several early methods adopting the idea of meta-learning [8,13,17,19,24,50,52,56]. FSRW [19] is a novel few-shot detector based on YOLOv2 [31], which re-weights the features with channel-wise attention and leverages these features to detect novel objects. Meta R-CNN [52] applies similar feature re-weighting scheme to Faster R-CNN [32] and performs meta-learning over RoI features. These methods usually suffer from a complicated training process and fail to learn generalizable features that can be transferred from base classes to novel classes. Recently, several fine-tuning based methods [1,9,26,30,38,44,48,49,57] achieve higher performance compared to meta-learning based methods. TFA [44] performs a simple two-stage fine-tuning approach which fine-tunes only the last layer on novel classes. MPSR [49] proposes to generate multi-scale positive samples to solve the problem of scale variations. FSCE [38] provides a strong baseline which fine-tunes feature extractors during the fine-tuning stage and employs a contrastive branch to rescue misclassifications. However, all these fine-tuning based methods suffer from overfitting on both base classes and novel classes. In this work, we design a novel knowledge distillation framework to tackle the problem.

Bag of visual words. Bag-of-Visual-Words is a popular technique for image recognition. Many variants of $BoVW$ have been proposed in the past[6,22,42] and they continue to be widely used in recent deep learning approaches[11,12,18,34]. VWE[34] designs a visual words learning module to generate CAMs[58] for weakly-supervised semantic segmentation. BoWNet[11] and OBOW[12] apply $BoVW$ to self-supervised learning. QuEST[18] introduces to distill the quantized feature maps from the teacher to the student. In this work, we learn a $BoVW$ model via a self-supervised task and image classification task.

Knowledge distillation for object detection. Knowledge distillation [16] is an effective way to transfer knowledge acquired in teacher network to student network. Early works focus on the task of image classification [33,39,55]. Recently, there are several works which propose to transfer knowledge for object detection. Chen et al. [2] distill knowledge from the teacher detector to the student detector in all components (i.e., feature extraction, RPN, classification and regression networks). Wang et al. [43] design a fine-grained feature imitation method which distills the features from foreground area. In this work, we propose a novel knowledge distillation method which transfers knowledge from a $BoVW$ model to a few-shot object detector, aiming to suppress potential overfitting on both base and novel classes.

3 Method

To deal with the potential overfitting in few-shot object detection based on deep learning networks, we propose to perform knowledge distillation to guide the learning process of few-shot object detection. Specifically, we learn a bag of visual words ($BoVW$) for encoding images. The learned visual words are presumably representative, hence an image should have consistent $BoVW$ representations in

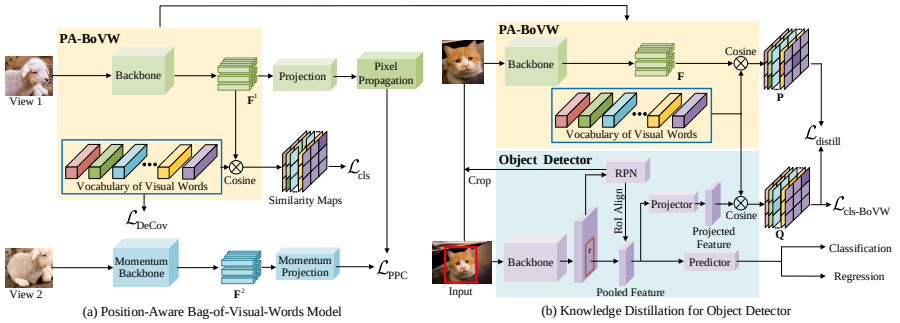


Fig. 2. The overall architecture of our method. We first train the proposed *PA-BoVW* model for learning a bag of visual words (*BoVW*) via two pretext tasks: pixel-to-propagation consistency [51] and base-class recognition task. During the training procedure of the detector, given a positive region proposal, we crop it from original images and fed it into our pre-learned *PA-BoVW* model to obtain the *BoVW* representation. Then we use the obtained *BoVW* representation as distilled knowledge to guide the learning of the detector.

two independent feature spaces. We first pre-learn a feature space and the derived *BoVW* representation in this space can be considered as distilled knowledge that is transferred to the learning of the few-shot object detector. The extracted features by the object detector for each positive proposal are expected to derive consistent *BoVW* representation with the distilled knowledge, thereby avoiding overfitting during supervised learning.

In this section, we will first elaborate on the proposed Position-Aware Bag-of-Visual-Words model (*PA-BoVW*) for learning a bag of visual words and encoding images based on these visual words. Then we will show how to perform knowledge distillation to guide the learning of few-shot object detection.

3.1 Position-Aware Bag-of-Visual-Words Model

A typical way of constructing a bag of visual words from an image corpus is to cluster the image patches in the corpus in deep feature space and select the clustering centroids as the visual words [11,18]. While such an unsupervised method is straightforward and feasible given a sufficiently large image corpus, it shows limited effectiveness when the size of image corpus is limited. This is mainly because such a method tends to focus on the statistically frequent image patches which may not be semantically representative.

In this work we present a novel Position-Aware Bag-of-visual-Words (*PA-BoVW*) model, which is able to learn a representative vocabulary of visual words from a limited size of image set. As shown in Figure 2 (a), we view each visual word as a learnable embedding and learn the parameters of all word embeddings in a supervised way based on two pretext tasks. The first pretext task is Pixel-to-Propagation Consistency [51], which trains the backbone network

to construct the embedding space for learning visual words in a self-supervised learning framework. Then we perform image classification on the representations encoded by our *PA-BoVW* model. Hence, the task of image classification serves as the second pretext task for optimizing the parameter learning of both the word embeddings and the backbone network. To this end, we first construct an iconic-object image dataset $\mathcal{D}$ as image corpus by extracting all base class objects from the detection dataset according to their ground-truth bounding boxes and labels. We then train our *PA-BoVW* model on $\mathcal{D}$.

Learnable word embeddings. Unlike the typical way that identifies visual words as the clustering centroids in the deep feature space, we learn visual words as learnable vectorial embeddings. A prominent benefit of such way is that the visual words do not necessarily correspond to image patches in the corpus. Instead, our model can learn an effective vocabulary of word embeddings freely from the whole embedding space under the optimization of the designed supervision (i.e., the classification pretext-task in our case).

Self-supervised learning of embedding space via Pixel-to-Propagation Consistency. To construct an effective embedding space for learning visual words, we employ a backbone network to encode the input image into a latent feature space and optimize the backbone network using Pixel-to-Propagation Consistency (PPC) [51] in a self-supervised way. PPC optimizes the backbone to make each pixel distinguishable from other pixels in the embedding space.

Formally, given an image corpus $\mathcal{D}$, for an image $I \in \mathcal{D}$, two views (I^1, I^2) are generated by typical data augmentations (random cropping, color distortion, etc.). They are then fed into a self-supervised framework including a regular encoding network and a paired momentum encoding network to extract features respectively, as shown in Figure 2 (a). The encoding network consists of a backbone network and a projection network. The features which pass the backbone networks are denoted as $(\mathbf{F}^1, \mathbf{F}^2)$. Then the projection networks convert them to $(\mathbf{E}^1, \mathbf{E}^2)$. Each pixel in the feature maps that pass the regular encoding network are processed by a pixel propagation module [51] to enrich its feature by attending to all other pixels in the same view according to their similarities, for instance, the pixel $\mathbf{x}_i^1$ in $\mathbf{E}^1$ is processed as:

$$q(\mathbf{x}_i^1) = \sum\nolimits_{j \in \mathbf{E}^1} \max\left(\frac{(\mathbf{x}_i^1)^\top \cdot \mathbf{x}_j^1}{\|\mathbf{x}_i^1\|_2 \left\|\mathbf{x}_j^1\right\|_2}, 0\right)^2 \cdot \Psi(\mathbf{x}_j^1), \forall \mathbf{x}_i^1 \in \mathbf{E}^1. \tag{1}$$

$\Psi(\cdot)$ is a feature transformation module comprising 2 convolution layers with a batch normalization layer and a ReLU layer. The obtained feature $q(\mathbf{x}_i^1)$ is then used to maximize its cosine similarities with its corresponding pixel $\mathbf{x}_i^2$ in the other view $\mathbf{E}^2$ passing the momentum backbone:

$$\mathcal{L}_{\text{PPC}} = 2 - \frac{q(\mathbf{x}_i^1)^\top \cdot \mathbf{x}_i^2}{\|q(\mathbf{x}_i^1)\|_2 \|\mathbf{x}_i^2\|_2} - \frac{q(\mathbf{x}_i^2)^\top \cdot \mathbf{x}_i^1}{\|q(\mathbf{x}_i^2)\|_2 \|\mathbf{x}_i^1\|_2}. \tag{2}$$

Note that $\mathbf{x}_i^1$ and $\mathbf{x}_i^2$ are the corresponding pixels in two views $(\mathbf{E}^1, \mathbf{E}^2)$, but they could have different positional coordinates. Since only the feature maps going

through the regular backbone are processed by PPC [51], both views have chance to go through the regular backbone and the momentum backbone respectively.

Since the embedding space is constructed independently from the learning of the object detector, the embedding space and the feature space of the object detector are two different feature spaces. Thus, the constructed embedding space can be used to not only learn the vocabulary of visual words, but encode $BoVW$ representations based on the learned visual words for an image as distilled knowledge. Such distilled knowledge is further used to guide the learning of the object detector.

Position-aware encoding of *BoVW* representation. Typical way of representing images using the visual words is to divide an image into patches and calculate the histogram over the visual words [22]. However, such encoding scheme loses the position information which is crucial for object detection. To address this limitation, we encode an image using the vocabulary of visual words by calculating the Cosine similarity between each pixel of the image and each visual word in the embedding space while retaining the positional relationship among pixels. Formally, given an image I, its features which pass the backbone network are denoted as $\mathbf{F} \in \mathbb{R}^{C\times H\times W}$ with C feature maps of size $H \times W$. The learned vocabulary of visual words are denoted as $\mathbf{V} \in \mathbb{R}^{K\times D}$ in which K is the number of visual words and D is the feature dimension for each word. The Cosine similarity between the pixel (h, w) in $\mathbf{F}$ and the j-th word in $\mathbf{V}$ is calculated as:

$$\mathbf{P}_{j,h,w} = \frac{\mathbf{V}_j^\top \cdot \mathcal{F}_{\text{conv}}(\mathbf{F}_{h,w})}{\left\|\mathbf{V}_j^\top\right\|_2 \left\|\mathcal{F}_{\text{conv}}(\mathbf{F}_{h,w})\right\|_2}, \tag{3}$$

where $\mathcal{F}_{\text{conv}}$ is a transformation function implemented by a convolutional layer to project $\mathbf{F}$ from C channels to D channels. Consequently, we obtain a similarity map $\mathbf{P} \in \mathbb{R}^{K\times H\times W}$ as the encoded $BoVW$ representation for the image I, which retains the position information for each pixel.

Supervised learning of visual words by the pretext task of image classification. We employ image classification as a pretext task to guide the learning of the visual words based on two considerations: 1) the visual words should be discriminative for object recognition in that our goal is object detection; 2) the learned visual words should have well generalizability and can be used for knowledge distillation to restrain potential overfitting for object detection, thus the pretext task should be independent of the task of object detection.

Given an image $I \in \mathcal{D}$, we first encode it with $BoVW$. Then the obtained $BoVW$ representation is fed into a simple classification head consisting of a pooling layer and a fully-connected layer. Formally, we perform average pooling over the $BoVW$ representation of I, namely the similarity map $\mathbf{P}$, along the H and W dimension and thus achieve a vectorial representation whose dimension is equal to the size of the vocabulary $\mathbf{V}$:

$$\mathbf{P}_{\text{avg}} = \frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W}\mathbf{P}_{j,h,w}. \tag{4}$$

The obtained $\mathbf{P}_{\text{avg}}$ are further fed into a fully-connected layer $\mathcal{F}_{\text{fc}}$ and Softmax function $\mathcal{F}_{\text{softmax}}$ for classification prediction. Cross Entropy (CE) loss is used for optimization:

$$\mathcal{L}_{\text{cls}} = \text{CE}(y_I, \mathcal{F}_{\text{softmax}}(\mathcal{F}_{\text{fc}}(\mathbf{P}_{\text{avg}}))), \tag{5}$$

where y_I is the groundtruth label for image I.

Reducing inter-word correlation. To encourage the diversity of visual words and reduce the redundancy among words, we add DeCov loss [5] as an auxiliary loss to minimize the correlation between different visual words:

$$\mathcal{L}_{\text{DeCov}} = \frac{1}{2}(\|\Sigma(\mathbf{V})\|_F^2 - \|\text{diag}(\Sigma(\mathbf{V}))\|_2^2), \tag{6}$$

where $\Sigma(\cdot)$ denotes the covariance matrix and $\text{diag}(\cdot)$ extracts the diagonal elements of a matrix.

The whole Bag-of-Visual-Words model can be trained under the supervision by above three loss functions jointly:

$$\mathcal{L}_{\text{BoVW}} = \mathcal{L}_{\text{PPC}} + \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{DeCov}}, \tag{7}$$

Note that although our *PA-BoVW* model is learned independently from the object detector in an extra step, it can be trained quite efficiently due to small size of object images and relatively simple supervision tasks compared to the task of object detection.

3.2 Knowledge Distillation for Object Detection

Our Position-Aware Bag-of-Visual-Words (*PA-BoVW*) model learns the embedding space for visual words using PPC [51] as the pretext task, and learns the vocabulary of visual words based on the pretext task of image classification. Thus, our *PA-BoVW* model is optimized in a completely independent way from the task of object detection. As a result, the encoded *BoVW* representation in the embedding space of the *PA-BoVW* model for an image can be viewed as distilled knowledge, which can be transferred to the learning process of few-shot object detection to suppress potential overfitting on this image. The rationale behind this design is that a well learned (non-overfitting) feature representation for an object by a detector should bear consistent similarity distribution over the learned visual words with the corresponding *BoVW* representation by our *PA-BoVW* model. Thus, we can derive consistent *BoVW* representations from the learned features by the object detector and our *PA-BoVW* model, respectively.

As shown in Figure 2 (b), we adopt the typical object detection framework, which is built upon Faster R-CNN [32]. Actually, our proposed method of knowledge distillation can be readily integrated into any classical object detection framework. Given a positive region proposal r generated by RPN (Region Proposal Network), which is assigned with one of the ground-truth labels and bounding boxes during training, we crop the corresponding region from the original input image and resize it to a fixed size by bilinear interpolation, then fed it into our *PA-BoVW* model to obtain its *BoVW* representation $\mathbf{P}(r) \in \mathbb{R}^{K \times H \times W}$ by

Eq. 3. Meanwhile, we calculate the Cosine similarities between the features $\mathbf{G}(r)$ for r, obtained from the RoI pooling layer of the object detector, and the vocabulary of visual words $\mathbf{V}$. Note that $\mathbf{G}(r)$ and the visual words are not learned in the same feature space, thus we project them into the same feature space first and then compute the Cosine similarities in the same way as Eq.3:

$$\begin{aligned} &\mathbf{Q}_{j,h,w}(r) = \frac{g(\mathbf{V}_j)^\top \cdot \phi(\mathbf{G}_{h,w}(r))}{\|g(\mathbf{V}_j)\|_2 \, \|\phi(\mathbf{G}_{h,w}(r))\|_2}, \\ &j = 1, \ldots, K, h = 1, \ldots, H, w = 1, \ldots, W. \end{aligned} \tag{8}$$

Herein, $g(\cdot)$ is the project function implemented as a fully connected layer for $\mathbf{V}$, while $\phi(\cdot)$ denotes the project function for features $\mathbf{G}(r)$, which is formulated as a 1×1 convolutional layer. K, H, W are the size of the vocabulary $\mathbf{V}$, the height and the width of $\mathbf{G}(r)$, respectively.

The obtained $\mathbf{Q}(r) \in \mathbb{R}^{K \times H \times W}$ is equivalent to the $BoVW$ representation encoded on the learned features $\mathbf{G}$ of the object detector. If $\mathbf{G}(r)$ is learned well and not overfitting on the input data, it should result in consistent $BoVW$ representation as our pre-trained $PA\text{-}BoVW$ model. Thus, we minimize the L1-norm distance between these two $BoVW$ representations to guide the learning process of the object detector:

$$L_{\text{distill}} = \frac{1}{RHW} \sum_{r=1}^{R} \sum_{h=1}^{H} \sum_{w=1}^{W} \left\| \mathbf{P}_{h,w}(r) - \mathbf{Q}_{h,w}(r) \right\|_1, \tag{9}$$

where R is the number of positive proposals.

Knowledge distillation on both base classes and novel classes. As most fine-tuning based methods [30,38,44,49] do, we first train the object detector on base classes and then fine-tune the model on the novel classes. We perform the knowledge distillation process in both training stages to restrain the potential overfitting on base classes and novel classes, respectively.

Collaborative object detection with *BoVW* representations. The obtained $BoVW$ representation $\mathbf{Q}(r)$ can also be used for object classification for the region proposal r. Thus, we perform classification by fusing the predicted scores from $\mathbf{Q}(r)$ and the original features respectively:

$$\begin{aligned} p &= \eta \cdot p_{\text{orig}} + (1 - \eta) \cdot p', \\ p' &= \mathcal{F}_{\text{softmax}}(\mathcal{F}_{\text{fc}}(\mathbf{Q}(r))), \end{aligned} \tag{10}$$

where p_{orig} and p' are predicted scores from the original features and from the $BoVW$ representation respectively. Here p' is obtained by performing a linear transformation $\mathcal{F}_{\text{fc}}$ and softmax function on $\mathbf{Q}(r)$. η is a hyper-parameter to fuse two scores. During training, Cross Entropy loss is used as an auxiliary loss to guide the optimization:

$$\mathcal{L}_{\text{cls-BoVW}} = \text{CE}(y, p'), \tag{11}$$

where y is the groundtruth label for the region proposal r.

Consequently, the object detector is trained under the supervision of three losses jointly:

$$\mathcal{L}_{\text{obj}} = \mathcal{L}_{\text{det}} + \mathcal{L}_{\text{distill}} + \mathcal{L}_{\text{cls-BoVW}}, \tag{12}$$

where $\mathcal{L}_{det}$ corresponds to the standard Faster R-CNN [32] losses for object detection, including the losses for RPN, classification and box regression.

4 Experiments

4.1 Experimental Setup

Benchmarks. We evaluate our approach on PASCAL VOC [7] dataset and MS COCO [28] dataset. We follow the previous work [19] for data construction to have a fair comparison. PASCAL VOC comprises 15 base classes and 5 novel classes. We utilize the same three class splits introduced in [19], where each novel class has $k = 1, 2, 3, 5, 10$ instances sampled from the combination of VOC 2007 and VOC 2012 trainval sets. VOC 2007 test set is used for evaluation. As for MS COCO, the 60 categories disjoint with PASCAL VOC are selected as base classes, and the remaining 20 categories are used as novel classes with $K = 10, 30$. For evaluation metrics, we report AP50 of novel classes (nAP50) for PASCAL VOC and COCO-style AP of the novel classes for MS COCO.

Implementation Details. We evaluate our approach by building it upon two state-of-the-art methods: TFA++ [38] and DeFRCN [30]. TFA++ [38] is a strong baseline which jointly fine-tunes the feature extractors and box predictors during the fine-tuning stage. DeFRCN [30] is a simple yet effective architecture which is the current state of the art.

For *PA-BoVW* model, we use an ImageNet [35] pre-trained ResNet101 [15] as the backbone. The input size is 224×224. The feature dimension of visual word is 512. The number of visual words is set to 256 for PASCAL VOC and 1024 for MS COCO. We follow the same data augmentation strategy in PPC [51], where two random patches of an image are independently sampled, followed by random horizontal flip, color distortion, gaussian blur, and solarization. We use AdamW optimizer to optimize the *PA-BoVW* model with the initial learning rate of 1e-4 for 24 epochs. We decay the learning rate by ratio 0.1 at epoch 18 and 22. The total batch size is set to 256. The object detector is trained on 8 GPUs with a batch size of 16. The η is uniformly set to 0.5. All other training settings are the same as that in TFA++[38] and DeFRCN[30].

4.2 Comparison with State-of-the-art Methods

Results on PASCAL VOC. Table 1 presents the results on PASCAL VOC, which show that our approach improves the performance of TFA++[38] by a large margin in all cases including different numbers of training shots in different splits. Particularly, for the 2-shot case of Novel Split 1, 5-shot case of Novel Split 2 and 2-shot case of Novel Split 3, our approach is 8.1%, 6.5%, 5.5% higher than the baseline. When applying our approach to DeFRCN[30], which is the

Table 1. Comparison with existing few-shot object detection methods using nAP50 as evaluation metric on three PASCAL VOC Novel Split sets. 'Ours (KD-TFA++)' denotes our method using TFA++ [38] as the baseline. † indicates that model is evaluated using the released code.

Method / Shots	Novel Split 1					Novel Split 2					Novel Split 3				
	1	2	3	5	10	1	2	3	5	10	1	2	3	5	10
LSTD [3]	8.2	1.0	12.4	29.1	38.5	11.4	3.8	5.0	15.7	31.0	12.6	8.5	15.0	27.3	36.3
MetaDet [46]	18.9	20.6	30.2	36.8	49.6	21.8	23.1	27.8	31.7	43.0	20.6	23.9	29.4	43.9	44.1
Meta R-CNN [52]	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
FSRW [19]	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
RepMet [20]	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
NP-RepMet [54]	37.8	40.3	41.7	47.3	49.4	**41.6**	43.0	43.4	47.4	49.1	33.3	38.0	39.8	41.5	44.8
MPSR [49]	41.7	-	51.4	55.2	61.8	24.4	-	39.2	39.9	47.8	35.6	-	42.3	48.0	49.7
TFA w/cos [44]	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
HallucFsDet [57]	47.0	44.9	46.5	54.7	54.7	26.3	31.8	37.4	37.4	41.2	40.4	42.1	43.3	51.4	49.6
Retentive R-CNN[9]	42.4	45.8	45.9	53.7	56.1	21.7	27.8	35.2	37.0	40.3	30.2	37.6	43.0	49.7	50.1
FSCE [38]	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
FADI [1]	50.3	54.8	54.2	59.3	63.2	30.6	35.0	40.3	42.8	48.0	45.7	49.7	49.1	55.0	59.6
CME [25]	41.5	47.5	50.4	58.2	60.9	27.2	30.2	41.4	42.5	46.8	34.3	39.6	45.1	48.3	51.5
UP-FSOD [47]	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	53.3
QA-FewDet [13]	42.4	51.9	55.7	62.6	63.4	25.9	37.8	46.6	48.9	51.1	35.2	42.9	47.8	54.8	53.5
TFA++† [38]	43.4	42.1	47.3	57.2	60.8	24.3	27.7	42.0	42.0	48.5	38.0	41.0	45.8	54.0	56.2
Ours (KD-TFA++)	47.0	50.2	52.5	62.1	64.2	29.7	32.9	45.9	48.5	51.1	42.6	46.5	48.8	56.8	57.4
DeFRCN [30]	57.0	58.6	64.3	67.8	67.0	35.8	42.7	51.0	54.5	52.9	52.5	56.6	55.8	60.7	**62.5**
Ours (KD-DeFRCN)	**58.2**	**62.5**	**65.1**	**68.2**	**67.4**	37.6	**45.6**	**52.0**	**54.6**	**53.2**	**53.8**	**57.7**	**58.0**	**62.4**	62.2

current state of the art, our method still improves the performance in most cases, especially in the extremely-few-shot regimes such as 1-shot and 2-shot.

Results on MS COCO. Table 2 shows the results on MS COCO. Applying our approach to two baselines achieves 1.1% and 0.3% nAP performance gain for 10-shot, 0.6% and 0.1% in terms of novel AP performance gain for 30-shot, respectively. There is no as large performance gain as on PASCAL VOC, which is probably because MS COCO has much more training images and thus has a lower risk of overfitting. To validate this speculation, we further evaluate our method by only using a small subset of base-class data for training. Specifically, we randomly select 10% samples from base classes to form a training set. For novel classes, we keep the same setting in standard few-shot object detection. Table 3 shows that the performance gains are larger than using all training data.

4.3 Ablation Studies

In this section, we conduct ablation studies on the Novel Split 1 of PASCAL VOC using TFA++ [38] as the baseline.

Effect of each functional component. Table 5 shows the efficacy of each functional components for few-shot object detection on novel classes, including distillation on base classes, novel classes and score fusion for classification in Equation 10. With the knowledge distillation for base classes, the performance gain is 4.2%/2.5%/1.9% for 3/5/10-shot, respectively. By performing distillation on novel class during the fine-tuning stage, the performance gain increases

Table 2. Few-shot object detection performance on MS COCO.

Method	nAP		nAP75	
	10	30	10	30
LSTD [3]	3.2	6.7	2.1	5.1
MetaDet [46]	7.1	11.3	6.1	8.1
Meta R-CNN [52]	8.7	12.4	6.6	10.8
FSRW [19]	5.6	9.1	4.6	7.6
TFA w/cos [44]	10.0	13.7	9.3	13.4
MPSR [49]	9.8	14.1	9.7	14.2
SRR-FSD [59]	11.3	14.7	9.8	13.5
Retentive R-CNN [9]	10.5	13.8	–	–
FSCE [38]	11.9	16.4	10.5	16.2
FADI [1]	12.2	16.1	11.9	15.8
CME [25]	15.1	16.9	16.4	17.8
UP-FSOD [47]	11.0	15.6	10.7	15.7
QA-FewDet [13]	11.6	16.5	9.8	15.5
TFA++† [38]	11.7	16.0	10.3	15.3
Ours (KD-TFA++)	12.8	16.6	11.5	16.1
DeFRCN [30]	18.6	22.5	17.6	22.3
Ours (KD-DeFRCN)	**18.9**	**22.6**	**17.8**	**22.6**

Table 3. Results on MS COCO with 10% labeled base-class samples.

Method	nAP		nAP75	
	10	30	10	30
DeFRCN [30]	12.1	14.9	8.5	11.5
Ours (KD-DeFRCN)	**13.0**	**16.0**	**9.7**	**12.6**

Table 4. Effect of each loss function.

L_{DeCov}	L_{distill}	$L_{\text{cls-BoVW}}$	3-shot	5-shot	10-shot
✓			47.3	57.2	60.8
✓	✓		51.2	59.9	62.6
	✓	✓	50.8	60.1	61.3
✓	✓	✓	**51.6**	**60.6**	**63.1**

0.1%/0.9%/0.4% respectively, which indicates the improved generalization ability on novel classes by our method. Finally, fusing the predicted scores from the original features and from the BoVW representation for classification yields the extra performance gain by 0.9%/1.5%/1.1%.

Effect of each loss function. Table 4 shows the effect of each loss function. Both the distillation loss L_{distill} and collaborative detection loss $\mathcal{L}_{\text{cls-BoVW}}$ improve the performance distinctly. Comaring the results in the last two rows, the $\mathcal{L}_{\text{DeCov}}$ which is designed to encourage the diversity of visual words and reduce the redundancy among words, also improves the performance.

Quantification for the overfitting on base classes. Fine-tuning training strategy tends to make models overfit on base classes. Since most parameters of the feature extractors are freezed or just fine-tuned slightly during the fine-tuning stage, most model capacity is allocated to fitting the base samples. To quantify such overfitting on base classes, we perform three experiments: 1) Using fine-tuning training strategy, we first pre-train the baseline TFA++ on base classes, then fine-tune it with sufficient novel-class samples instead of k shot per class, which is denoted as two-stage training mode; 2) Similar to the setting in 1), we fine-tune our model with sufficient novel-class samples after pre-training; 3) we train the baseline using sufficient samples for both base and novel classes together in one-stage mode. We compare both nAP50 and the number of misclassified samples from novel to base classes to measure the overfitting. The results in Table 6 show that 1) the baseline trained in two-stage mode performs much worse than training itself in one-stage mode and has a larger number of misclassified samples, indicating that the model is heavily biased towards base classes; 2) using the same two-stage training mode, our method achieves 3.5% of performance gain than baseline and substantially decreases misclassified cases, which demonstrates the effectiveness of our method for suppressing the overfitting on base classes.

Table 5. Ablation studies of key components.

Baseline	Distillation for Base	Distillation for Novel	Score Fusion	nAP50 3-shot	nAP50 5-shot	nAP50 10-shot
✓				47.3	57.2	60.8
✓	✓			51.5	59.7	62.7
✓	✓	✓		51.6	60.6	63.1
✓	✓	✓	✓	**52.5**	**62.1**	**64.2**

Table 6. Quantification for the overfitting.

Methods	nAP50	Misclassified cases (novel→base)
TFA++ (two-stage)	73.2	1021
Ours (KD-TFA++) (two-stage)	76.7	723
TFA++ (one-stage)	85.8	631

Table 7. Effect of different pretext tasks.

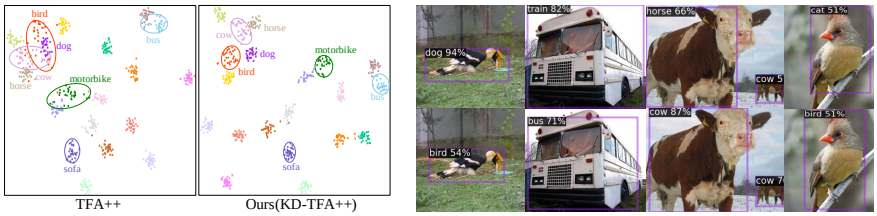
Pretext tasks	nAP50 3	5	10
w/o distillation	47.3	57.2	60.8
Cluster	49.6	57.4	61.1
Cls	52.3	60.8	63.6
Cls + PPC	**52.5**	**62.1**	**64.2**

Table 8. Distillation on deep features vs on $BoVW$ representations.

Distillation methods	nAP50 1	2	3	5	10
Baseline	43.4	42.1	47.3	57.2	60.8
On deep features	33.2	37.9	36.5	47.1	48.3
On BoVW	**47.0**	**50.2**	**52.5**	**62.1**	**64.2**

Effect of different pretext tasks for learning *BoVW* models. We conduct experiments to compare our *PA-BoVW* model and the typical method for learning the visual words (denoted as *'Cluster'*), which trains a classification network on base classes and selects the clustering centroids of the feature vectors as visual words. Then we evaluate the performance our *PA-BoVW* optimized using only the image classification as the pretext task (denoted as *'Cls'*). Table 7 shows that the performance gains from *'Cluster'* is smaller than that of our *PA-BoVW* using only the pretext task of image classification. Using the pretext task of PPC further boosts the performance substantially, which implies the importance of learning the embedding space by PPC in a self-supervised way.

Distillation on deep features vs on *BoVW* representations. A classical way to perform knowledge distillation is to learn an independent feature space and perform distillation between two feature spaces. We conduct such experiment to compare between distillation on deep features and on $BoVW$ representations. Specifically, we directly distill the pooled deep features in our trained *PA-BoVW* model to the feature space of the detector. The results of such method shown in Table 8 are much worse than our method. This is reasonable since distillation between feature space relies heavily on 1) the quality of the referenced features from which the distillation is performed and 2) well modeling of mapping between two feature space. By contrast, our method distills knowledge based on the similarity distribution over learned visual words, which benefits from similar merits of feature representations as Bag-of-Word representations used in NLP.



(a) t-SNE. (b) Detection results.

Fig. 3. (a) The t-SNE visualization of proposal embeddings of baseline with and without distillation. (b) Detection results based on the 10-shot case. The first row shows the results of the baseline and the second row shows the results of our approach.

Qualitative Evaluation. Figure 3 (a) shows the t-SNE [29] visualization of proposal embeddings from randomly selected 30 instance bounding boxes per category. The baseline (TFA++ [38]) tends to misclassify some samples of novel classes as similar base classes. For instance, the samples from novel classes 'bird' and 'cow' cannot be clearly separated from other base classes like 'dog' and 'horse'. In contrast, applying our approach to the baseline model leads to more accurate boundaries. Figure 3 (b) shows the detection results of the baseline and our approach. We can observe that our method can successfully detect the novel objects while the baseline tends to misclassify these objects as base classes.

5 Conclusion

To solve the potential overfitting in few-shot object detection, we propose a knowledge distillation framework. We first learn a *PA-BoVW* model using two pretext tasks, namely Pixel-to-Propagation Consistency and image classification. Based on the *PA-BoVW* model, then we perform distillation to guide the learning of detector. As an orthogonal component, our approach can be easily combined with other methods and significantly improve the performance.

Acknowledgements This work was supported in part by the NSFC fund (U2013210, 62006060, 62176077), in part by the Guangdong Basic and Applied Basic Research Foundation under Grant (2019Bl515120055, 2021A1515012528, 2022A1515010306), in part by the Shenzhen Key Technical Project under Grant 2020N046, in part by the Shenzhen Fundamental Research Fund under Grant (JCYJ20210324132210025), in part by the Shenzhen Stable Support Plan Fund for Universities (GXWD20201230155427003-20200824125730001, GXWD20201230155427003-20200824164357001), in part by the Medical Biometrics Perception and Analysis Engineering Laboratory, Shenzhen, China, and in part by the Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies (2022B1212010005).

References

1. Cao, Y., Wang, J., Jin, Y., Wu, T., Chen, K., Liu, Z., Lin, D.: Few-shot object detection via association and discrimination. In: NeurIPS (2021)
2. Chen, G., Choi, W., Yu, X., Han, T., Chandraker, M.: Learning efficient object detection models with knowledge distillation. In: NeurIPS (2017)
3. Chen, H., Wang, Y., Wang, G., Qiao, Y.: Lstd: A low-shot transfer detector for object detection. In: AAAI (2018)
4. Chen, W.Y., Liu, Y.C., Kira, Z., Wang, Y.C.F., Huang, J.B.: A closer look at few-shot classification. In: ICLR (2018)
5. Cogswell, M., Ahmed, F., Girshick, R., Zitnick, L., Batra, D.: Reducing overfitting in deep networks by decorrelating representations. arXiv preprint arXiv:1511.06068 (2015)
6. Csurka, G., Dance, C., Fan, L., Willamowski, J., Bray, C.: Visual categorization with bags of keypoints. In: ECCVW (2004)
7. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International journal of computer vision **88**(2), 303–338 (2010)
8. Fan, Q., Zhuo, W., Tang, C.K., Tai, Y.W.: Few-shot object detection with attention-rpn and multi-relation detector. In: CVPR (2020)
9. Fan, Z., Ma, Y., Li, Z., Sun, J.: Generalized few-shot object detection without forgetting. In: CVPR (2021)
10. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. In: ICML (2017)
11. Gidaris, S., Bursuc, A., Komodakis, N., Pérez, P., Cord, M.: Learning representations by predicting bags of visual words. In: CVPR (2020)
12. Gidaris, S., Bursuc, A., Puy, G., Komodakis, N., Cord, M., Perez, P.: Obow: Online bag-of-visual-words generation for self-supervised learning. In: CVPR (2021)
13. Han, G., He, Y., Huang, S., Ma, J., Chang, S.F.: Query adaptive few-shot object detection with heterogeneous graph convolutional networks. In: ICCV (2021)
14. Hariharan, B., Girshick, R.B.: Low-shot visual recognition by shrinking and hallucinating features. In: ICCV (2017)
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
16. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
17. Hu, H., Bai, S., Li, A., Cui, J., Wang, L.: Dense relation distillation with context-aware aggregation for few-shot object detection. In: CVPR (2021)
18. Jain, H., Gidaris, S., Komodakis, N., Pérez, P., Cord, M.: Quest: Quantized embedding space for transferring knowledge. In: ECCV (2020)
19. Kang, B., Liu, Z., Wang, X., Yu, F., Feng, J., Darrell, T.: Few-shot object detection via feature reweighting. In: ICCV (2019)
20. Karlinsky, L., Shtok, J., Harary, S., Schwartz, E., Aides, A., Feris, R.S., Giryes, R., Bronstein, A.M.: Repmet: Representative-based metric learning for classification and few-shot object detection. In: CVPR (2019)
21. Kim, J., Kim, T., Kim, S., Yoo, C.D.: Edge-labeling graph neural network for few-shot learning. In: CVPR (2019)
22. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: CVPR (2006)

23. Lee, K., Maji, S., Ravichandran, A., Soatto, S.: Meta-learning with differentiable convex optimization. In: CVPR (2019)
24. Li, A., Li, Z.: Transformation invariant few-shot object detection. In: CVPR (2021)
25. Li, B., Yang, B., Liu, C., Liu, F., Ji, R., Ye, Q.: Beyond max-margin: Class margin equilibrium for few-shot object detection. In: CVPR (2021)
26. Li, Y., Zhu, H., Cheng, Y., Wang, W., Teo, C.S., Xiang, C., Vadakkepat, P., Lee, T.H.: Few-shot object detection via classification refinement and distractor retreatment. In: CVPR (2021)
27. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: ICCV (2017)
28. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV (2014)
29. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of Machine Learning Research **9**(11) (2008)
30. Qiao, L., Zhao, Y., Li, Z., Qiu, X., Wu, J., Zhang, C.: Defrcn: Decoupled faster r-cnn for few-shot object detection. In: ICCV (2021)
31. Redmon, J., Farhadi, A.: Yolo9000: Better, faster, stronger. In: CVPR (2017)
32. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: NeurIPS (2015)
33. Romero, A., Ballas, N., Kahou, S.E., Chassang, A., Gatta, C., Bengio, Y.: Fitnets: Hints for thin deep nets. arXiv preprint arXiv:1412.6550 (2014)
34. Ru, L., Du, B., Zhan, Y., Wu, C.: Weakly-supervised semantic segmentation with visual words learning and hybrid pooling. International Journal of Computer Vision **130**(4), 1127–1144 (2022)
35. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International journal of computer vision **115**(3), 211–252 (2015)
36. Satorras, V.G., Estrach, J.B.: Few-shot learning with graph neural networks. In: ICLR (2018)
37. Snell, J., Swersky, K., Zemel, R.S.: Prototypical networks for few-shot learning. In: NeurIPS (2017)
38. Sun, B., Li, B., Cai, S., Yuan, Y., Zhang, C.: Fsce: Few-shot object detection via contrastive proposal encoding. In: CVPR (2021)
39. Tian, Y., Krishnan, D., Isola, P.: Contrastive representation distillation. In: ICLR (2019)
40. Tian, Z., Shen, C., Chen, H., He, T.: Fcos: Fully convolutional one-stage object detection. In: ICCV (2019)
41. Vinyals, O., Blundell, C., Lillicrap, T., Kavukcuoglu, K., Wierstra, D.: Matching networks for one shot learning. In: NeurIPS (2016)
42. Wang, J., Yang, J., Yu, K., Lv, F., Huang, T., Gong, Y.: Locality-constrained linear coding for image classification. In: CVPR (2010)
43. Wang, T., Yuan, L., Zhang, X., Feng, J.: Distilling object detectors with fine-grained feature imitation. In: CVPR (2019)
44. Wang, X., Huang, T., Gonzalez, J., Darrell, T., Yu, F.: Frustratingly simple few-shot object detection. In: ICML (2020)
45. Wang, Y.X., Girshick, R.B., Hebert, M., Hariharan, B.: Low-shot learning from imaginary data. In: CVPR (2018)
46. Wang, Y.X., Ramanan, D., Hebert, M.: Meta-learning to detect rare objects. In: ICCV (2019)
47. Wu, A., Han, Y., Zhu, L., Yang, Y.: Universal-prototype enhancing for few-shot object detection. In: ICCV (2021)

48. Wu, A., Zhao, S., Deng, C., Liu, W.: Generalized and discriminative few-shot object detection via svd-dictionary enhancement. In: NeurIPS (2021)
49. Wu, J., Liu, S., Huang, D., Wang, Y.: Multi-scale positive sample refinement for few-shot object detection. In: ECCV (2020)
50. Xiao, Y., Marlet, R.: Few-shot object detection and viewpoint estimation for objects in the wild. In: ECCV (2020)
51. Xie, Z., Lin, Y., Zhang, Z., Cao, Y., Lin, S., Hu, H.: Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. In: CVPR (2021)
52. Yan, X., Chen, Z., Xu, A., Wang, X., Liang, X., Lin, L.: Meta r-cnn: Towards general solver for instance-level low-shot learning. In: ICCV (2019)
53. Yang, L., Li, L., Zhang, Z., Zhou, X., Zhou, E., Liu, Y.: Dpgn: Distribution propagation graph network for few-shot learning. In: CVPR (2020)
54. Yang, Y., Wei, F., Shi, M., Li, G.: Restoring negative information in few-shot object detection. In: NeurIPS (2020)
55. Yim, J., Joo, D., Bae, J., Kim, J.: A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In: CVPR (2017)
56. Zhang, L., Zhou, S., Guan, J., Zhang, J.: Accurate few-shot object detection with support-query mutual guidance and hybrid loss. In: CVPR (2021)
57. Zhang, W., Wang, Y.X.: Hallucination improves few-shot object detection. In: CVPR (2021)
58. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: CVPR (2016)
59. Zhu, C., Chen, F., Ahmed, U., Shen, Z., Savvides, M.: Semantic relation reasoning for shot-stable few-shot object detection. In: CVPR (2021)