PNPDet: Efficient Few-shot Detection without Forgetting via Plug-and-Play Sub-networks

Gongjie Zhang[1] Kaiwen Cui[1] Rongliang Wu[1] Shijian Lu*[1] Yonghong Tian[2]
[1]Nanyang Technological University [2]Peking University
{gongjiezhang,shijian.lu,rlwu}@ntu.edu.sg kaiwen001@e.ntu.edu.sg yhtian@pku.edu.cn

Abstract

The human visual system can detect objects of unseen categories from merely a few examples. However, such capability remains absent in state-of-the-art detectors. To bridge this gap, several attempts have been proposed to perform few-shot detection by incorporating meta-learning techniques. Such methods can improve detection performance on unseen categories, but also add huge computational burden, and usually degrade detection performance on seen categories. In this paper, we present PNPDet, a novel Plug-and-Play Detector, for efficient few-shot detection without forgetting. It introduces a simple but effective architecture with separate sub-networks that disentangles the recognition of base and novel categories and prevents hurting performance on known categories while learning new concepts. Distance metric learning is further incorporated into sub-networks, consistently boosting detection performance for both base and novel categories. Experiments show that the proposed PNPDet can achieve comparable few-shot detection performance on unseen categories while not losing accuracy on seen categories, and also remain efficient and flexible at the same time.

1. Introduction

Deep convolutional neural networks based models have achieved state-of-the-art performance on many visual understanding tasks, such as image classification and object detection. The success of such models relies heavily on large-scale and fully-annotated datasets. On the other hand, such large-scale and annotated datasets are not always available for various specific applications [25, 60] due to expensive human labelling costs and/or difficulty in data acquisition. In addition, it is difficult to directly apply a deep model trained on large-scale datasets to deal with new concepts with limited training data available. Despite poor performance on new concepts, it usually requires heavy fine-tuning efforts and usually degrades the performance on learnt concepts undesirably, as reported by [11, 16, 57, 55], et al. These limitations restrict further applications of existing visual understanding models in many real-world scenarios. In contrast, by leveraging previous knowledge, we humans have a much more powerful and flexible visual system that can handle unseen object categories easily by spotting just one or only a few examples without causing negative effects on seen categories.

Researches to bridge the gap of learning new concepts from a small amount of data are usually termed as few-shot learning, which primarily focus on few-shot classification [19, 45, 34, 47, 49, 31, 11, 2, 44, 21]. Recently, there are also several attempts to investigate the task of few-shot detection [16, 1, 44, 30, 57, 54, 8]. The mainstream of current few-shot detection approaches are based on integration of meta-learning techniques and general detectors like Faster R-CNN [41]. Such methods need to perform a separate feed-forward process for each category, thus are computationally expensive and slow.

Another gap between existing visual models and human visual system is the ability to learn unseen categories (novel categories) from a small amount of data without sacrificing performance on categories that the visual models are initially trained on (base categories). Spyros Gidaris and Nikos Komodakis [11] make a first attempt to achieve learning without forgetting in the task of few-shot classification. However, current state-of-the-art approaches on few-shot object detection either completely lose the ability to detect seen categories [1], or cause a dramatic degrade on detection performance of seen categories after learning new concepts [16, 44, 30, 54, 57, 8].

An ideal few-shot object detector should be efficient, flexible, accurate and is able to perform few-shot object detection on novel categories without impeding the performance on base categories. This goal of learning novel categories without sacrificing accuracy on base categories remains an open challenge due to the following difficulties. First, few-shot detection is naturally a much more difficult task compared with few-shot classification, as it needs

*Corresponding author.

to perform dense classification and localization simultaneously. Second, the classification task in object detection involves an additional 'background' category that almost always outnumbers positive categories, restricting the direct application of few-shot classification methods on the detection task. More importantly, under certain scenarios, objects in few-shot detection that are initially classified into 'background' can become positive object instances after adding novel categories. Such ambiguity will cause heavy alternation in network parameters, which costs a lot of time in optimization and even leads to performance drop on base categories. Third, unlike classification models, most existing object detectors adopt deep network architectures that are more prone to over-fitting to limited number of training samples of novel categories.

To address the challenges mentioned above, we propose PNPDet – a single-stage Plug-and-Play Detector for efficient few-shot detection without forgetting. Its architecture is illustrated in Fig. 1. Unlike most existing few-shot detection methods, one unique feature of the proposed PNPDet is that learning novel categories is as simple as attaching an additional sub-network to existing model without interfering the the performance over base categories or requiring high extra computational cost. Specifically, PNPDet introduces a novel single-stage object detection architecture into few-shot detection where feature extractor and bounding box generator are shared among base and novel categories, and the recognition of base and novel categories is disentangled with separate sub-networks. Inspired by the success distance metric learning (DML) in few-shot classification tasks, we incorporate Cosine Similarity Comparison Head (CosHead) into the PNPDet which helps to improve the object recognition and object localization performance significantly. By adopting cosine distance (cosine similarity) as the distance metric, PNPDet can better model objects' relationship with corresponding prototypes and background as 'similar' and 'irrelevant', respectively, and this brings consistent improvements for both object detection with abundant training data and few-shot detection. To handle various intra-category variances of feature representations, we propose Adaptive CosHead to learn a scale factor for each novel category to normalize the intra-category variance of feature representations, further boosting few-shot detection accuracy.

Our contributions are fourfold: **(1)** We proposed a novel PNPDet, which performs few-shot object detection on novel categories without degrading the accuracy on base categories by disentangling their recognition processes with separate sub-networks. **(2)** We introduce Cosine Similarity Comparison Head (CosHead), bringing distance metric learning (DML) into the proposed PNPDet, which consistently improves object detection performance either with abundant training data or with few-shot training data. Besides, we extend it to Adaptive CosHead that further improves few-shot detection by learning a scale factor for each novel category to normalize the intra-category variance of feature representations. **(3)** Despite its simplicity, flexibility and efficiency, experiments demonstrate that PNPDet can achieve comparable few-shot detection performance on novel categories and even superior overall detection performance, compared with existing state-of-the-art meta-detectors. **(4)** We pose an alternative direction for few-shot detection – simply attaching sub-networks to learn new concepts other than mainstream approaches of meta-learning.

2. Related Works

General Object Detection. Over the past few years, deep neural networks have achieved remarkable improvements in object detection. State-of-the-art object detection methods can be broadly classified into two categories, namely, two-stage detectors and single-stage detectors. Two-stage detectors mainly refer to Region-based Convolutional Neural Networks (R-CNN), including the original R-CNN [13], Fast R-CNN [12] and the most commonly used Faster R-CNN [41]. Two-stage detectors are generally more accurate and robust, leading to many variants and applications [61, 14, 29, 51, 15, 10, 56, 50]. Single-stage detectors mainly include SSD [27], YOLO [38, 39, 40], RetinaNet [23], RefineDet [62], etc. The recently proposed anchor-free detectors such as CornerNet [20], CenterNet [64] and FCOS [48] can also be classified as single-stage detectors. Single-stage detectors do not require proposal generation stage, and directly predict bounding boxes and detection confidences of multiple categories, thus are conceptually simpler and significantly faster compared with two-stage detectors. In this work, the proposed PNPDet follows the fashion of CenterNet [64], a simple yet effective single-stage object detector.

Few-shot Learning. Few-shot learning refers to learning from just a few training samples for each novel category.

There exist many research works on this topic, mostly concentrating on few-shot classification [35, 58, 6, 31, 43, 53, 9, 3, 21, 46, 37, 17, 22, 11, 32, 36, 42, 63, 26, 25]. Most of these works [9, 3, 21, 46, 37, 17, 22, 11, 32, 36, 42, 26] adopt meta-learning techniques to attack few-shot classification task. Besides, some other approaches [49, 45, 47, 44, 2] are based on distance metric learning (DML), which try to learn feature representations that preserve the category neighborhood structure under certain distance metric, where features corresponding to the same category are closer than features from different categories. However, they still lack the ability to attend to novel categories with no drop of accuracy on base categories. To enable few-shot classification with this property, [11] further introduces an

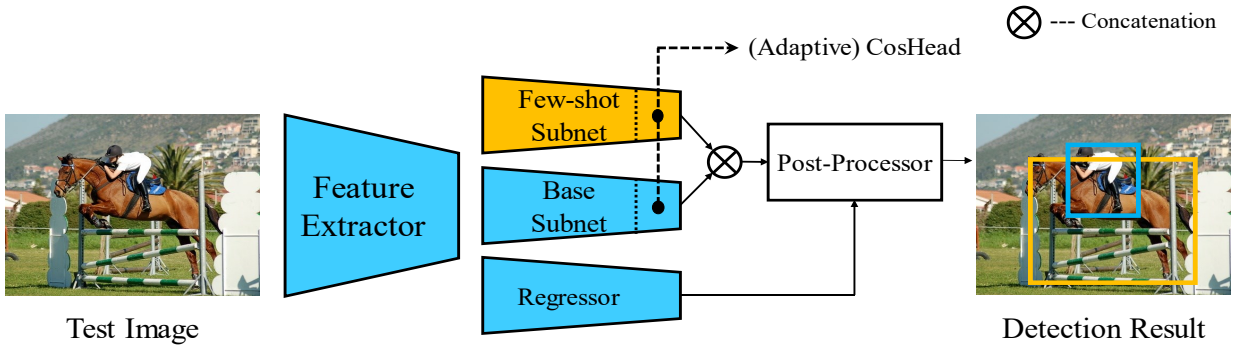

Figure 1. Illustration of the proposed PNPDet architecture: Feature extractor $F(\cdot)$ generates feature maps that are shared by all downstream sub-networks. Sub-networks generate heatmaps for object recognition and localization, while regressor predicts heights and widths of bounding boxes at different locations for both base and novel categories. Parameters of blue-colored components are learned from large-scale datasets with bounding box annotations, and parameters of orange-colored components are learned from few-shot training samples during the few-shot training stage.

attention-based mechanism to generate weights for novel categories.

On the other hand, there are only several attempts to attack the problem of few-shot object detection. Low-Shot Transfer Detector (LSTD) [1] is the first work to attack the problem of few-shot detection. It integrates SSD [27] and Faster R-CNN [41] into a unified detector and achieves few-shot detection in transfer learning manner. However, LSTD has limitations in terms of low inference speed, low adaptation speed, and completely losing the ability to detect objects of base categories after fine-tune. Most other works on few-shot detection are based on integration of meta-learning techniques and general detectors like Faster R-CNN [41], YOLO [39], etc. Using YOLO v2 [39] as backbone, FeatReweight [16] proposes to meta-learn a group of weights to re-weight the importance of features, generating category-specific feature maps for both base and novel categories. In addition, [30, 57, 54, 8] all extend Faster R-CNN [41] by meta-learning over Region of Interest (RoI) features and/or image features. Such meta-learning based approaches need to perform a separate feed forward computation for each category to be detected, thus require large amount of computational resources, and are inefficient. Besides, such approaches still severely deteriorate detection performance on base categories after adding novel categories into consideration.

We notice that two concurrent works on few-shot detection adopt several similar ideas as ours. ONCE [33] also adopt CenterNet [64] as backbone to achieve incremental few-shot detection without forgetting. However, it still falls into the fashion of meta-learning, while our proposed PNPDet simply attaches sub-networks on existing networks to achieve few-shot detection. FsDet [52] also incorporates distance metric learning into Faster-RCNN with FPN, and achieves state-of-the-art performance on few-shot detection. However, it still cannot guarantee no forgetting on base categories. And it is based on a very deep backbone architecture (ResNet101 + FPN), thus also requires heavy computational resources.

By taking ideas from distance metric learning (DML) and combine them with a simple yet effective architecture with multiple sub-networks, our proposed PNPDet achieves comparable few-shot detection performance on unseen categories while preserving high detection performance of seen categories as well as efficient and flexible inference.

3. Task Definition

Given two sets of categories $\mathcal{C}_{base}$ and $\mathcal{C}_{novel}$ which are mutually exclusive to each other, the model is initially trained to detect objects of base categories $\mathcal{C}_{base}$ on dataset $\mathcal{D}_{base}$, which is a large-scale dataset with bounding box annotations over $\mathcal{C}_{base}$. Few-shot detection without forgetting requires the model to learn to detect categories from both $\mathcal{C}_{base}$ and $\mathcal{C}_{novel}$, provided with only a few annotated examples $\mathcal{D}_{novel}$ for novel categories $\mathcal{C}_{novel}$. We define the task of N-shot object detection as detecting objects from $\mathcal{C}_{novel}$ with exactly N object instances as training examples for each novel category.

4. Method

In this section, we present details of our proposed Plug-and-Play Detector (PNPDet). First, we introduce the basic architecture of PNPDet. Then, we introduce Cosine Similarity Comparison Head (CosHead) and Adaptive CosHead, which effectively incorporate distance metric learning (DML) into the architecture of object detection for object recognition. Finally, we describe the training and inference strategy of the proposed few-shot detection framework.

4.1. Basic Architecture

We first review the architecture of CenterNet [64], which will be used as the underlying architecture in our proposed PNPDet. CenterNet is a recently proposed anchor-free object detection network, which performs object detection in keypoint detection manner. Precisely, CenterNet consists of three parts: a feature extractor to generate high-resolution feature maps, a class-specific heatmap prediction branch, and a class-agnostic regressor for bounding box regression. Each category corresponds to one heatmap, and the peaks in the heatmap represent object centers. Widths and heights of bounding boxes are directly regressed through the bounding box regressor. For more detailed introduction of CenterNet, readers are strongly encouraged to read the original paper.

The proposed PNPDet is built upon the architecture of CenterNet. The overview of the proposed PNPDet is illustrated in Fig. 1. We follow most of the settings as the original CenterNet [64]: modified DLA-34 [59] with deformable convolutional networks [4, 65] pre-trained on ImageNet [5] as backbone, loss functions and target generations stated in [64], etc. As is shown in Fig. 1, to make predictions on novel categories, the proposed PNPDet adds a parallel heatmap prediction sub-network to generate heatmaps for novel categories. Heatmaps generated by both base and few-shot sub-networks are then merged to produce final detection results, together with bounding box regression information provided by class-agnostic regressor. In Fig. 1, blue-colored components are trained with abundant data of base categories, and few-shot sub-network in orange is trained with few-shot training samples of novel categories. Each sub-network is made up of only four standard convolutional layers with ReLU as activation function, such a lightweight design allows fast inference and adaptation speed, and prevents potential over-fitting.

Such architecture has several advantages under few-shot detection settings. First, it models the 'background' category implicitly based on similarity to the prototypes, which simplifies the network and makes the model easy to be modified for novel categories. Second, base and novel categories are disentangled by incorporating two separate sub-networks for heatmap prediction, so that the updating from novel categories will not interfere with base sub-network, which achieves few-shot detection without forgetting. Third, the class-agnostic bounding box regression branch can be directly applied to novel categories through a transfer learning manner, effectively transferring knowledge learned from base categories.

However, although such architecture ensures no forgetting over base categories when learning new concepts, it fails to produce satisfactory results for novel categories with limited training samples. To address this problem, we incorporate distance metric learning (DML) by replacing the last layer of sub-networks with the proposed CosHead or Adaptive CosHead, more details to be described in Section 4.2. Detailed training and inference strategy for PNPDet will be discussed in Section 4.3.

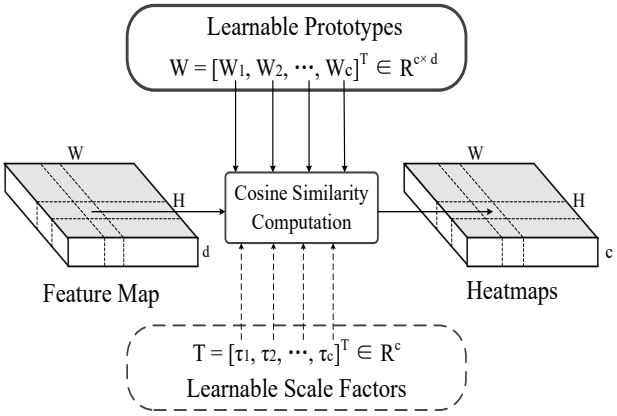

Figure 2. Illustration of the proposed CosHead and Adaptive CosHead in generating heatmaps for object recognition and localization in distance metric learning manner. CosHead is illustrated with solid parts while Adaptive CosHead is illustrated with both solid and dashed parts. Learnable scale factors of Adaptive CosHead act to normalize intra-category feature variances.

4.2. Metric Learning for Few-shot Detection

The key difficulty of few-shot object detection lies in the recognition of objects from novel categories. The object recognition (classification) task in few-shot detection is even more challenging than the task of few-shot classification. Few-shot classification task is usually evaluated in an episodic manner, where the model only needs to determine the category of reference images that the testing image is most similar to. Distance metric learning (DML) is a widely adopted technique to achieve satisfactory few-shot learning for classification, which aims to learn feature representations that preserve the category neighborhood structure. It is very natural to apply DML to search for the closest reference category within an episode. However, in the task of object detection, there exists a special 'background' category that almost always outnumbers samples from positive categories. And it is hard to find an appropriate prototype for the background category due to the large intra-class variations, which restricts the ability to apply existing DML based methods in image classification to the objection detection task.

To better solve the object recognition problem in few-shot detection, we introduce Cosine Similarity Comparison Head (CosHead) into the detection architecture mentioned above. We show that with simple modification, distance metric learning can be incorporated into our detection architecture, and greatly boost detection performance over novel categories under low-shot setting.

The intuition behind the introduced CosHead is very simple: rather than directly learning discriminative feature representations, we learn a feature space where features corresponding to the same category are closer than features from different categories under certain distance metric. To achieve this, the model needs to learn a prototype for each category as the best representative of the corresponding category and performs dense comparisons between prototypes and features from different locations. In the context of detection, the 'background' category is modeled implicitly and considered as 'irrelevant' objects with the learned category prototypes. To model the relationship of 'resemblance' and 'irrelevance', we adopt cosine distance (cosine similarity) as our pre-defined distance metric.

The proposed CosHead is illustrated in Fig. 2, and can be formulated as follows. It takes as input a feature map $\mathcal{F} \in \mathbf{R}^{d \times H \times W}$, and generates heatmaps $\mathcal{S} \in \mathbf{R}^{C \times H \times W}$ ranging from 0 to 1 indicating the similarities between features and the learned prototypes at all locations. Here d denotes dimension of input feature representations, and C denotes number of categories. Heatmaps are computed as:

$$\mathcal{S}_{c,x,y} = \sigma(\tau \cdot \frac{{W_c}^T \cdot \mathcal{F}_{x,y}}{|W_c| \cdot |\mathcal{F}_{x,y}|}) \tag{1}$$

where $\sigma(\cdot)$ denotes sigmoid function, $|\cdot|$ denotes L2-norm, $S_{c,x,y} \in \mathbf{R}$ indicates the similarity score at location (x, y) with category c, $\mathcal{F}_{x,y} \in \mathbf{R}^d$ is the feature vector extracted by the feature extractor at location (x, y), $W_c \in \mathbf{R}^d$ denotes the learnable feature prototype for category c, $\tau \in \mathbf{R}$ is a scalar to extend the range of cosine similarity, which is fixed at 10.0 in all experiments. The learnable prototypes $W \in \mathbf{R}^{C \times d}$ can be learned directly through back propagation.

Besides, since the network is not trained on novel categories with abundant samples, the feature representations for novel categories are not as compact as features for base categories. In addition, unlike few-shot classification where each image is a good representative of its corresponding category, in the task of few-shot object detection, each bounding box serves as a representation of its category. In this context, objects have larger-scale variations, and could even be vague or occluded. As a result, there exists larger intra-category variance for each category, making it even harder for novel category recognition. Different few-shot training samples can also exhibit different intra-category variances.

To alleviate this problem, we further propose **Adaptive CosHead**, as shown in Fig. 2. In addition to the prototype for each category, Adaptive CosHead further learns an adaptive scale factor for each category in order to normalize different intra-category variances. Adaptive CosHead can be formulated as:

$$\mathcal{S}_{c,x,y} = \sigma(\tau \cdot \tau_c \cdot \frac{{W_c}^T \cdot \mathcal{F}_{x,y}}{|W_c| \cdot |\mathcal{F}_{x,y}|}) \tag{2}$$

where τ_c is a learnable scalar for category c, which is also directly learned through back propagation.

The proposed CosHead and Adaptive CosHead have several advantages in the context of low-shot detection. First, feature representations learned by distance metric learning have better generalization ability for novel categories. Second, using cosine similarity can effectively model the relationship of 'irrelevance', which perfectly models the relationship between learned prototypes and background regions. By doing this, the model does not need to model 'background' specifically, making it easier to adapt to novel categories which is initially classified as 'background'. Third, under low-shot setting, there is no guarantee that the support set of novel categories are representative enough to generate high responses for objects of novel categories. Incorporating cosine distance with CenterNet architecture can alleviate this issue as it only deals with peak values and cosine distance tends to generate smooth responses. This can be further supported by Fig. 3.

We clarify that incorporating cosine distance metric with neural networks is not our contribution, as this has been investigated in [49, 2, 11, 34], etc. to solve few-shot classification tasks. Instead, we investigate the potential of combining distance learning metric into the detection framework. We demonstrate that with simple modifications, cosine distance metric learning can be incorporated into the detection framework, and can greatly boost detection performance under low-shot settings.

4.3. Training and Inference Procedure

The training procedure for our proposed PNPDet consists of two stages. First, we perform **initial training** on the large-scale dataset for object detection with bounding box annotations over base categories. During this initial training stage, the detector is trained to detect objects of base categories. In addition to a sub-network to generate heatmaps for base categories, it learns a class-agnostic bounding box regression branch, a class-agnostic center point regression branch, and a feature extractor which preserves the category neighborhood structure under cosine distance metric. Everything is trained end-to-end during this stage. Then, we perform **few-shot training** on the small-scale dataset to train a sub-network to generate heatmaps for novel categories. During this stage, all parameters except the sub-network for novel categories are set fixed. The sub-network for novel category is initialized using weights from the sub-network to generate heatmaps for base categories in order to reuse the high-level knowledge learned from large-scale base categories.

During inference, the sub-networks for both base and novel categories are stacked on the feature extractor in parallel to produce heatmaps for all categories. Regressor produces bounding box guidance for all categories. Detection

results for both base and novel categories are generated at one go with marginal extra computational expenses.

5. Experiments

5.1. Experimental Setup

We follow the setups of previous works for few-shot object detection [16, 54, 57, 33]. Specifically, widely adopted object detection datasets Pascal VOC [7] and MS COCO [24] are used for few-shot detection setups.

Pascal VOC dataset consists of images covering 20 object categories. We follow previous works [41, 39, 16] to use Pascal VOC 07 and 12 train/val images for training and use Pascal VOC 07 test set for evaluation. Following [16], we use 3 novel/base category split settings, i.e., ("bird", "bus", "cow", "motorbike", "sofa"/ others); ("aeroplane", "bottle","cow","horse","sofa" / others) and ("boat", "cat", "motorbike","sheep", "sofa"/ others). The number of shots is set to 1, 3 and 10, following previous settings. Mean average precision (mAP) at IoU threshold 0.5 is used as evaluation metrics. Results are aggregate over 10 randomly selected support sets of novel categories.

MS COCO dataset covers 80 object categories including the 20 categories in PASCAL VOC. We use the 20 shared categories as novel categories, and use the remaining 60 categories in COCO dataset as base categories. We use train2017 set as the training set, and perform evaluation on the val2017 set. Evaluation metrics defined by COCO [24] are adopted. Results are aggregate over 10 randomly selected support sets of novel categories.

5.2. Implementation Details

Our model is trained on a single NVIDIA GeForce 2080Ti GPU with 11GB memory. Images are resized to 384×384 before feeding into the networks. Standard data augmentation techniques, including horizontal flipping, random crop and color jittering are adopted in both initial and few-shot training stages. The detailed architecture for sub-networks is $Conv3\times3-ReLU-Conv3\times3-ReLU-Conv3\times3-ReLU-Head$, where $Head$ is either a plain convolutional layer, or the proposed (Adatpive) CosHead. For initial training stage, we follow the training scheme of [64] to train the model until convergence. For few-shot training stage, initial learning rate is set to 1.0×10^{-5} and decays to 1.0×10^{-6} with cosine annealing decay [28] to convergence; Network is optimized using Adam optimizer [18]. During inference, Non-Maximum Suppression (NMS) with an IoU threshold of 0.6 is adopted to generate final results.

5.3. Ablation Experiments

We first conduct ablation experiments over the design choice of sub-networks, i.e., plain convolutional layer, CosHead or Adaptive CosHead for base and novel sub-networks.

Table 1. Detection performance (mAP@0.5) after initial training on base categories at different category splits on Pascal VOC.

	Split 1	Split 2	Split 3
Plain Conv	74.4	72.2	73.8
CosHead	**75.5**	**73.1**	**74.6**
Adaptive CosHead	74.4	72.3	74.0

Before presenting few-shot detection performance, we first report detection performance on base categories of different models after the initial training stage in Table 1. For fair comparison, the only difference among three methods is the last layer of sub-networks. Bold numbers indicate the best performance obtained within each category split. As Table 1 shows, CosHead is able to consistently outperform Plain Conv, which potentially demonstrates the advantages of DML as stated in Section 4.2. Incorporating CosHead can thus be deemed as an improvement over the original CenterNet [64] for object detection. Surprisingly, Adaptive CosHead performs slightly worse than CosHead. We believe it is because when training data is sufficient, feature representations of the same category under the learned feature space are already compact, so it is unnecessary to learn an extra scale factor to normalize the intra-category variance of each category.

Based on experiment results in Table 1, we use CosHead by default in all the following experiments for base category sub-networks.

We further present ablative experiment results in few-shot detection settings on Pascal VOC dataset. As shown in Table 2, simply fine-tuning CenterNet to fully convergence will cause dramatic performance drop on base categories. Unlike naive fine-tune strategy, since the proposed PNPDet disentangle the recognition of different category split with different sub-networks, the performance on base categories will not drop even after learning to recognize new concepts. However, without introducing distance metric learning, detection accuracy of PNPDet on novel categories is still unsatisfactory, especially at extremely low-shot settings (1 shot and 3 shot). With CosHead and Adaptive CosHead introduced, without adding any extra computational burden, the few-shot detection performance on novel categories increase a lot, especially at extremely low-shot settings. This demonstrate the effectiveness of introducing cosine distance as distance metric in few-shot detection. We also notice that Adaptive CosHead is able to outperform CosHead under low-shot setting, especially at 1 and 3 shots. In 10 shot experiments, Adaptive CosHead can only boost marginal performance gain on novel categories. This aligns with experiments in Table 1, which also indicates Adaptive CosHead is able to effectively help to normalize the intra-category

Table 2. Few-shot detection performance (mAP@0.5) of some variants of PNPDet on both base and novel categories under Pascal VOC.

			Split 1			Split 2			Split 3			Avg		
			1 shot	3 shot	10 shot	1 shot	3 shot	10 shot	1 shot	3 shot	10 shot	1 shot	3 shot	10 shot
CenterNet-ft-full		base	68.2	65.0	59.8	66.0	66.2	61.0	64.6	62.9	58.3	66.3	64.7	59.7
		novel	8.5	14.4	32.5	9.0	11.6	32.9	9.0	14.0	26.4	8.8	13.3	30.6
		overall	53.3	52.4	53.0	51.8	52.6	54.0	50.7	50.7	50.3	51.9	51.9	52.4
PNPDet	Plain Conv	base	75.5	75.5	75.5	73.1	73.1	73.1	74.6	74.6	74.6	74.4	74.4	74.4
		novel	8.3	10.7	29.4	6.9	9.4	27.6	8.5	11.0	28.2	7.9	10.4	28.4
		overall	58.7	59.3	64.0	56.6	57.2	61.7	58.1	58.7	63.0	57.8	58.4	62.9
	CosHead	base	75.5	75.5	75.5	73.1	73.1	73.1	74.6	74.6	74.6	74.4	74.4	74.4
		novel	15.7	25.3	41.8	15.9	24.6	34.9	17.8	25.9	36.3	16.5	25.3	37.7
		overall	60.6	63.0	**67.1**	58.8	61.0	63.6	60.4	62.4	**65.0**	59.9	62.1	65.2
	AdaptiveCosHead	base	75.5	75.5	75.5	73.1	73.1	73.1	74.6	74.6	74.6	74.4	74.4	74.4
		novel	18.2	27.3	41.0	16.6	26.5	36.4	18.9	27.2	36.2	17.9	27.0	37.9
		overall	**61.2**	**63.5**	66.9	**59.0**	**61.5**	**63.9**	**60.7**	**62.8**	**65.0**	**60.3**	**62.6**	**65.3**

variance of each category when training data is extremely scarce for the convolutional feature extractor to do so. But when training data for each category is efficient for the feature extractor to adjust the intra-category variance, its performance gain will be marginal.

We further visualize and compare heatmaps for novel categories generated by (1) Plain Conv and (2) Adaptive CosHead in Fig. 3. Heatmap responses of CosHead is very similar with Adaptive CosHead visually, therefore we do not show them. Models of both settings are trained using the same 3-shot training samples. As is shown in Fig. 3, the first row contains input images that contain objects of novel categories (motorbike, cow, bus, and bird from left to right). The second and third rows show heatmaps generated by (1) Plain Conv and (2) Adaptive CosHead, respectively. It is obvious that heatmaps generated by Adaptive CosHead contain clearer peak responses at center points, and contain fewer artifacts. This further demonstrates the superiority of heatmap generation in cosine distance metric learning manner.

Through these ablation study, for the following experiments, we choose Adaptive CosHead as the default setting for novel category sub-networks, and choose CosHead as the default as the default setting for base category sub-networks.

5.4. Benchmark Experiments

We further benchmark the proposed PNPDet with state-of-the-art meta-learning based few-shot detectors as well as naive fine-tune approaches. Results under commonly used Pascal VOC 15 $\Rightarrow$ 5 is shown in Table 3. Note that listed methods are based on different backbone networks and detection architectures.

Table 3 suggests the incompetence of naive fine-tune strategy over both two-stage and single-stage detectors. Specifically, detection performance on novel categories is low due to over-fitting, and base category performance suffers from catastrophic forgetting. Meanwhile, by incorporating meta-learning techniques with Faster R-CNN [41] and YOLO v2 [39] respectively, Meta R-CNN [57] and FeatReweight [16] tend to produce better detection performance on novel categories and suffer less from forgetting on base categories. This can be attributed to meta-learning techniques that acquire meta-level knowledge, which is more powerful to generalize to other unseen categories. However, such state-of-the-art meta-detectors still have a large performance gap on base categories after few-shot learning, compared with detection performance by general object detectors. By adopting a novel strategy of attaching sub-networks for few-shot detection, our proposed PNPDet is able to achieve comparable performance over novel categories without forgetting, achieving superior overall performance after learning new concepts.

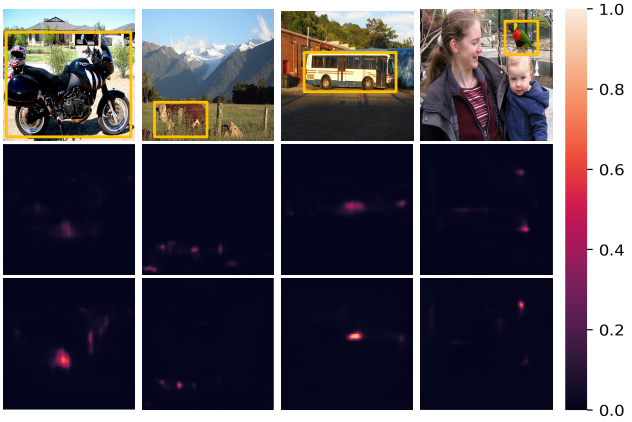

Figure 3. Heatmaps for novel categories generated by different methods: The heatmaps generated with Adaptive CosHead in row 3 exhibit stronger and cleaner responses at the center of objects as compared with the heatmaps generated with Plain Conv in row 2.

We specifically focus on comparison between PNPDet and FeatReweight [16], since it is the state-of-the-art meta-detector with efficient **single-stage** architecture. Compared with it, the proposed PNPDet achieves superior overall few-

Table 3. Few-shot detection performance (mAP@0.5) benchmark on both base and novel categories under Pascal VOC dataset. RED and BLUE indicate state of the art and the second best. N/A indicates no published performance available.

		Split 1			Split 2			Split 3			Avg		
		1 shot	3 shot	10 shot	1 shot	3 shot	10 shot	1 shot	3 shot	10 shot	1 shot	3 shot	10 shot
FRCNN-ft-full w/ Res101 FPN	base	62.6	61.3	59.8	63.2	61.0	59.8	63.7	62.1	60.5	63.2	61.5	60.0
	novel	9.9	21.6	35.6	9.4	17.4	29.8	8.1	19.0	31.0	9.1	19.3	32.1
	overall	49.4	51.4	53.8	49.7	50.1	52.3	49.8	51.3	53.1	49.6	51.0	53.1
Meta R-CNN [57] w/ Res101	base	N/A	64.8	67.9	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	novel	19.9	35.0	51.5	10.4	29.6	45.4	14.3	27.5	48.1	14.9	30.7	48.3
	overall	N/A	57.3	63.8	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
CenterNet-ft-full w/ DLA34	base	68.2	65.0	59.8	66.0	66.2	61.0	64.6	62.9	58.3	66.3	64.7	59.7
	novel	8.5	14.4	32.5	9.0	11.6	32.9	9.0	14.0	26.4	8.8	13.3	30.6
	overall	53.3	52.4	53.0	51.8	52.6	54.0	50.7	50.7	50.3	51.9	51.9	52.4
FeatReweight [16] w/ YOLO v2	base	66.4	64.8	63.6	68.2	66.0	64.7	65.9	65.0	63.1	66.8	65.3	63.8
	novel	14.8	26.7	47.2	15.7	22.7	40.5	21.3	28.4	45.9	17.3	25.9	44.5
	overall	53.5	55.3	59.5	55.1	55.2	58.7	54.8	55.9	58.8	54.5	55.5	59.0
PNPDet (Ours) w/ DLA34	base	75.5	75.5	75.5	73.1	73.1	73.1	74.6	74.6	74.6	74.4	74.4	74.4
	novel	18.2	27.3	41.0	16.6	26.5	36.4	18.9	27.2	36.2	17.9	27.0	37.9
	overall	61.2	63.5	66.9	59.0	61.5	63.9	60.7	62.8	65.0	60.3	62.6	65.3

Table 4. 10-shot detection performance (AP, AR maxDet=1) benchmark under COCO dataset. N/A indicates no published performance available.

	Base Class		Novel Class		Overall	
	AP	AR	AP	AR	AP	AR
CenterNet-ft-full	20.7	23.4	1.4	8.2	15.8	19.6
FeatReweight [16]	N/A	N/A	5.6	10.1	N/A	N/A
Meta R-CNN [57]	N/A	N/A	**8.7**	**12.6**	N/A	N/A
ONCE [33]	22.9	**29.9**	5.1	9.5	18.4	**24.8**
PNPDet (Ours)	**25.8**	25.5	5.5	**12.6**	**20.7**	22.3

shot detection performance. More specifically, as PNPDet learns to detect novel categories, it does not lose accuracy on base categories, and is able to generate superior detection performance over novel categories under extremely low-shot settings (1 shot and 3 shot). However, PNPDet detection performance on novel categories seems to be saturated when more training samples of novel categories are provided, compared with FeatReweight. We believe this is because when more samples are provided, meta-learning based approaches are able to use such samples to train a better feature extractor for such novel categories. Updating feature extractor comes at a price of catastrophic forgetting of base category knowledge. However, our proposed PNPDet can only update its corresponding sub-networks. We expect adopting deeper sub-networks with PNPDet can further boost few-shot detection performance with larger shots.

10-shot detection results under COCO 60 $\Rightarrow$ 20 setting are presented in Table 4. Under this setting, the proposed PNPDet can also achieve superior overall performance among all categories, and comparable novel category detection performance with state-of-the-art few-shot meta-detectors. Note our proposed PNPDet outperforms ONCE [33] – a meta-learning version of CenterNet for few-shot detection, for AP in both base and novel categories.

The proposed PNPDet is also very efficient. In our Pascal VOC setting with input image size of 384 $\times$ 384, it runs at 47 FPS on a NVIDIA 2080Ti, only adding marginal extra computational cost compared with CenterNet, which runs at 52 FPS. The additional sub-network for novel categories detection only involves $\sim$0.7M extra parameters. The adaptation speed is also very fast, which only takes about 140s under our setting. On the other hand, meta-detectors need to perform one feed forward for each category to be detected, adding large extra computational expenses. Besides, Meta R-CNN [57] adopts two-stage architecture with deep backbone, which is slow for inference. Such properties make PNPDet a preferable few-shot detector that is suitable to be deployed in many real-world scenarios, especially those requires frequent modifications.

However, we also note that the proposed PNPDet's performance on novel categories under larger shots (e.g., 10 shot) is significantly inferior to those meta-learning few-shot detectors. Further in-depth researches are encouraged to bridge this gap.

6. Conclusion

This paper presents PNPDet – a Plug-and-Play Detector for efficient few-shot detection without forgetting. By disentangling recognition of base and novel categories via sub-networks, the proposed PNPDet can detect novel categories without degrading the accuracy of base categories. We further incorporate metric learning in base and novel category recognition sub-networks, boosting detection performance of both base and novel categories. Experimental results demonstrate the superiority of PNPDet, which achieves superior overall performance and comparable novel category performance, and possesses the merits of fast adaptation, fast inference and flexibility.

References

[1] Hao Chen, Yali Wang, Guoyou Wang, and Yu Qiao. LSTD: A low-shot transfer detector for object detection. In *AAAI*, 2018.

[2] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Wang, and Jia-Bin Huang. A closer look at few-shot classification. In *ICLR*, 2019.

[3] Yinbo Chen, Xiaolong Wang, Zhuang Liu, Huijuan Xu, and Trevor Darrell. A new meta-baseline for few-shot learning, 2020.

[4] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *ICCV*, pages 764–773, 2017.

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009.

[6] Matthijs Douze, Arthur Szlam, Bharath Hariharan, and Hervé Jégou. Low-shot learning with large-scale diffusion. In *CVPR*, pages 3349–3358, 2018.

[7] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The Pascal Visual Object Classes (VOC) Challenge. *IJCV*, 88(2):303–338, June 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 *CVPR*, 2020.

[9] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. *arXiv preprint arXiv:1703.03400*, 2017.

[10] Jiyang Gao, Jiang Wang, Shengyang Dai, Li-Jia Li, and Ram Nevatia. Note-rcnn: Noise tolerant ensemble rcnn for semi-supervised object detection. In *ICCV*, pages 9508–9517, 2019.

[11] Spyros Gidaris and Nikos Komodakis. Dynamic few-shot visual learning without forgetting. In *CVPR*, pages 4367–4375, 2018.

[12] Ross Girshick. Fast R-CNN. In *ICCV*, pages 1440–1448, 2015.

[13] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, pages 580–587, 2014.

[14] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *ICCV*, pages 2961–2969, 2017.

[15] Zhenwei He and Lei Zhang. Multi-adversarial faster-rcnn for unrestricted object detection. In *ICCV*, pages 6668–6677, 2019.

[16] Bingyi Kang, Zhuang Liu, Xin Wang, Fisher Yu, Jiashi Feng, and Trevor Darrell. Few-shot object detection via feature reweighting. In *ICCV*, pages 8420–8429, 2019.

[17] Siavash Khodadadeh, Ladislau Boloni, and Mubarak Shah. Unsupervised meta-learning for few-shot image classification. In *NeurIPS*, pages 10132–10142, 2019.

[18] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[19] Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. Siamese neural networks for one-shot image recognition. In *ICML Deep Learning Workshop*, volume 2, 2015.

[20] Hei Law and Jia Deng. CornerNet: Detecting objects as paired keypoints. In *ECCV*, pages 734–750, 2018.

[21] Kwonjoon Lee, Subhransu Maji, Avinash Ravichandran, and Stefano Soatto. Meta-learning with differentiable convex optimization. In *CVPR*, 2019.

[22] Xinzhe Li, Qianru Sun, Yaoyao Liu, Qin Zhou, Shibao Zheng, Tat-Seng Chua, and Bernt Schiele. Learning to self-train for semi-supervised few-shot classification. In *NeurIPS*, pages 10276–10286, 2019.

[23] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, pages 2980–2988, 2017.

[24] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014.

[25] Suiyi Ling, Andreas Pastor, Jing Li, Zhaohui Che, Junle Wang, Jieun Kim, and Patrick Le Callet. Few-shot pill recognition. In *CVPR*, 2020.

[26] Bo Liu, Hao Kang, Haoxiang Li, Gang Hua, and Nuno Vasconcelos. Few-shot open-set recognition using meta-learning. In *CVPR*, 2020.

[27] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. SSD: Single shot multibox detector. In *ECCV*, pages 21–37, 2016.

[28] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *arXiv preprint: 1608.03983*, 2016.

[29] Pengyuan Lyu, Minghui Liao, Cong Yao, Wenhao Wu, and Xiang Bai. Mask textspotter: An end-to-end trainable neural network for spotting text with arbitrary shapes. In *ECCV*, pages 71–88, 2018.

[30] Claudio Michaelis, Ivan Ustyuzhaninov, Matthias Bethge, and Alexander S. Ecker. One-shot instance segmentation. In *arXiv preprint: 1811.11507*, 2018.

[31] Nikhil Mishra, Mostafa Rohaninejad, Xi Chen, and Pieter Abbeel. A simple neural attentive meta-learner. In *ICLR*, 2018.

[32] Boris Oreshkin, Pau Rodríguez López, and Alexandre Lacoste. Tadam: Task dependent adaptive metric for improved few-shot learning. In *NeurIPS*, pages 721–731, 2018.

[33] Juan-Manuel Perez-Rua, Xiatian Zhu, Timothy M Hospedales, and Tao Xiang. Incremental few-shot object detection. In *CVPR*, pages 13846–13855, 2020.

[34] Hang Qi, Matthew Brown, and David G Lowe. Low-shot learning with imprinted weights. In *CVPR*, pages 5822–5830, 2018.

[35] Siyuan Qiao, Chenxi Liu, Wei Shen, and Alan L. Yuille. Few-shot image recognition by predicting parameters from activations. In *CVPR*, 2018.

[36] Siyuan Qiao, Chenxi Liu, Wei Shen, and Alan L Yuille. Few-shot image recognition by predicting parameters from activations. In *CVPR*, pages 7229–7238, 2018.

[37] Alireza Rahimpour and Hairong Qi. Class-discriminative feature embedding for meta-learning based few-shot classification. In *WACV*, pages 3179–3187, 2020.

[38] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You Only Look Once: Unified, real-time object detection. In *CVPR*, pages 779–788, 2016.

[39] Joseph Redmon and Ali Farhadi. YOLO 9000: Better, faster, stronger. In *CVPR*, pages 6517–6525, 2017.

[40] Joseph Redmon and Ali Farhadi. YOLO v3: an incremental improvement. In *arXiv preprint: 1804.02767*, 2018.

[41] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NeurIPS*, pages 91–99, 2015.

[42] Andrei A Rusu, Dushyant Rao, Jakub Sygnowski, Oriol Vinyals, Razvan Pascanu, Simon Osindero, and Raia Hadsell. Meta-learning with latent embedding optimization. *arXiv preprint arXiv:1807.05960*, 2018.

[43] Edgar Schonfeld, Sayna Ebrahimi, Samarth Sinha, Trevor Darrell, and Zeynep Akata. Generalized zero-and few-shot learning via aligned variational autoencoders. In *CVPR*, pages 8247–8255, 2019.

[44] Eli Schwartz, Leonid Karlinsky, Joseph Shtok, Sivan Harary, Mattias Marder, Sharathchandra Pankanti, Rogerio Feris, Abhishek Kumar, Raja Giries, and Alex M Bronstein. Repmet: Representative-based metric learning for classification and one-shot object detection. In *CVPR*, pages 5197–5206, 2019.

[45] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *NeurIPS*, pages 4077–4087, 2017.

[46] Qianru Sun, Yaoyao Liu, Tat-Seng Chua, and Bernt Schiele. Meta-transfer learning for few-shot learning. In *CVPR*, 2019.

[47] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *CVPR*, pages 1199–1208, 2018.

[48] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. FCOS: Fully convolutional one-stage object detection. In *ICCV*, pages 9627–9636, 2019.

[49] Oriol Vinyals, Charles Blundell, Tim Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *NeurIPS*, pages 3630–3638, 2016.

[50] Tan Wang, Jianqiang Huang, Hanwang Zhang, and Qianru Sun. Visual commonsense r-cnn. In *CVPR*, pages 10760–10770, 2020.

[51] Tao Wang, Xiaopeng Zhang, Li Yuan, and Jiashi Feng. Few-shot adaptive faster r-cnn. In *CVPR*, pages 7173–7182, 2019.

[52] Xin Wang, Thomas E. Huang, Trevor Darrell, Joseph E Gonzalez, and Fisher Yu. Frustratingly simple few-shot object detection.

[53] Xin Wang, Fisher Yu, Ruth Wang, Trevor Darrell, and Joseph E Gonzalez. Tafe-net: Task-aware feature embeddings for low shot learning. In *CVPR*, pages 1831–1840, 2019.

[54] Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. Meta-learning to detect rare objects. *ICCV*, pages 9924–9933, 2019.

[55] Xiongwei Wu, Doyen Sahoo, and Steven C. H. Hoi. Meta-RCNN: Meta learning for few-shot object detection, 2020.

[56] Mengmeng Xu, Yancheng Bai, Sally Sisi Qu, and Bernard Ghanem. Semantic part rcnn for real-world pedestrian detection. In *CVPR Workshops*, 2019.

[57] 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 *ICCV*, 2019.

[58] Han-Jia Ye, Hexiang Hu, De-Chuan Zhan, and Fei Sha. Few-shot learning via embedding adaptation with set-to-set functions. In *CVPR*, pages 8808–8817, 2020.

[59] Fisher Yu, Dequan Wang, Evan Shelhamer, and Trevor Darrell. Deep layer aggregation. In *CVPR*, pages 2403–2412, 2018.

[60] Gongjie Zhang, Kaiwen Cui, Tzu-Yi Hung, and Shijian Lu. Defect-GAN: High-fidelity defect synthesis for automated defect inspection. In *WACV*, 2021.

[61] Gongjie Zhang, Shijian Lu, and Wei Zhang. CAD-Net: A context-aware detection network for objects in remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 57(12):10015–10024, 2019.

[62] Shifeng Zhang, Longyin Wen, Xiao Bian, Zhen Lei, and Stan Z Li. Single-shot refinement neural network for object detection. In *CVPR*, pages 4203–4212, 2018.

[63] Linjun Zhou, Peng Cui, Xu Jia, Shiqiang Yang, and Qi Tian. Learning to select base classes for few-shot classification. In *CVPR*, 2020.

[64] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. In *arXiv preprint: 1904.07850*, 2019.

[65] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable Convnets v2: More deformable, better results. In *CVPR*, pages 9308–9316, 2019.