Exploring Effective Knowledge Transfer for Few-shot Object Detection

Zhiyuan Zhao  
State Key Laboratory of Virtual Reality Technology and Systems,  
Beihang University  
Beijing, China  
zhaozhiyuan@buaa.edu.cn

Qingjie Liu∗  
State Key Laboratory of Virtual Reality Technology and Systems,  
Beihang University  
Beijing, China  
qingjie.liu@buaa.edu.cn

Yunhong Wang  
State Key Laboratory of Virtual Reality Technology and Systems,  
Beihang University  
Beijing, China  
yhwang@buaa.edu.cn

ABSTRACT

Recently, few-shot object detection (FSOD) has received much attention from the community, and many methods are proposed to address this problem from a knowledge transfer perspective. Though promising results have been achieved, these methods fail to achieve shot-stable: methods that excel in low-shot regimes are likely to struggle in high-shot regimes, and vice versa. We believe this is because the primary challenge of FSOD changes when the number of shots varies. In the low-shot regime, the primary challenge is the lack of inner-class variation. In the high-shot regime, as the variance approaches the real one, the main hindrance to the performance comes from misalignment between learned and true distributions. However, these two distinct issues remain unsolved in most existing FSOD methods. In this paper, we propose to overcome these challenges by exploiting rich knowledge the model has learned and effectively transferring them to the novel classes. For the low-shot regime, we propose a distribution calibration method to deal with the lack of inner-class variation problem. Meanwhile, a shift compensation method is proposed to compensate for possible distribution shift during fine-tuning. For the high-shot regime, we propose to use the knowledge learned from ImageNet as guidance for the feature learning in the fine-tuning stage, which will implicitly align the distributions of the novel classes. Although targeted toward different regimes, these two strategies can work together to further improve the FSOD performance. Experiments on both the VOC and COCO benchmarks show that our proposed method can significantly outperform the baseline method and produce competitive results in both low-shot settings (shot<5) and high-shot settings (shot≥5). Code is available at https://github.com/JulioZhao97/EffTrans_Fsdet.git.

KEYWORDS

few-shot object detection, knowledge transfer, distribution calibration, distribution regularization

ACM Reference Format:

Zhiyuan Zhao, Qingjie Liu, and Yunhong Wang. 2022. Exploring Effective Knowledge Transfer for Few-shot Object Detection. In Proceedings of the 30th ACM International Conference on Multimedia (MM ’22), October 10–14, 2022, Lisboa, Portugal. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3503161.3548062

1 INTRODUCTION

Today, deep neural networks (DNNs) have achieved outstanding performance in a wide variety of data-intensive applications. Unfortunately, due to the data-driven nature of deep models, their performance is severely hampered in data-limited scenarios. On the other hand, humans have an incredible capacity for learning from a few examples and generalizing to new concepts from a small amount of information. For instance, by showing a photograph of a stranger to a child once, he/she can rapidly identify this stranger from a pile of pictures. To acquire this ability, researchers endeavor to empower the deep models with the quick and robust learning ability in data-limited scenarios, termed few-shot learning [4, 19, 39, 48].

As a fundamental task in computer vision, object detection has witnessed tremendous progress in the last few years, yet it still suffers from the data curse. Precedented by few-shot recognition, efforts have been devoted to addressing few-shot object detection (FSOD), which is a much more challenging task. Earlier attempts [2] inherit ideas from approaches for few-shot classification [3, 25] and adapt them to FSOD. For instance, following the meta-learning paradigm [33, 38], meta-detectors [10, 47, 47] are trained on the base classes to learn prior knowledge of the base classes and then are updated on the novel classes to make predictions. Another line of work involves the fine-tuning framework. Fine-tuning based FSOD methods normally consist of two steps: (1) firstly, base detectors are pre-trained on abundant base classes; (2) secondly, the detectors are fine-tuned on novel classes for adaption. These methods intend to transfer knowledge learned on base classes to the novel classes and are known as transfer learning based methods. However, due to the rarity of the target data, fine-tuning all parameters of the models is inefficient. Wang et al. [40] propose to fine-tune only the classification and regression branches of the detector while freezing the feature extractor. Their approach yields competitive results with this simple strategy and...
reinvigorates researchers’ interest in transfer learning methodology [5, 54].

Despite the progress made, the performance of existing transfer learning based FSOD methods are still far from satisfying. We notice that most of these methods that perform well in low-shot regimes are likely to be inferior in high-shot regimes, and vice versa. In other words, these methods are incapable of achieving shot-stable. We believe that this is because the primary challenges in the low-shot regime and the high-shot regime are very different from each other (as shown in Figure 1). In the low-shot regimes, the primary difficulty is the lack of inner-class variation. In the high-shot regimes, to improve the detection performance, the key factor is to tackle misalignment between learned distributions and true distributions. However, in most existing transfer learning based methods, these two distinct issues are ignored thus their performance is unstable across different shots.

To overcome these unsolved issues, a key point is to exploit the rich knowledge the model has learned in previous stages. To this end, we propose a calibration and regularization based method that enables effective knowledge transfer. Specifically, we propose a distribution calibration method to deal with the lack of variation issue for the low-shot regime. We calibrate the biased novel class distributions with the base class distributions. Then synthetic training features are sampled from calibrated novel distributions and added to training subsequently. In this calibration-and-generation manner, the inner-class variation of novel classes can be greatly enriched. Besides, we notice that there exist distribution shift of base classes due to the fine-tuning process, which may mislead novel class distributions and limit the effectiveness of calibration. To overcome this limitation, we present a strategy of compensating for possible distribution shift, namely shift compensation.

In the high-shot regime, we propose to use the knowledge learned from ImageNet as guidance for feature learning in the fine-tuning stage, which will implicitly align the distributions of the novel classes. This is inspired by the recent studies that ImageNet features are stable and expressive and thus can be used as teachers for downstream task learning [8, 23]. We have tried the base detector model as the teacher, which also adopts the ImageNet pre-trained backbone, however, receiving unsatisfactory results. We suspect the ImageNet features are corrupted by the base class training. In this paper, the knowledge transfer from ImageNet features is achieved with a regularization loss. Although the two solutions are targeted for different regimes of FSOD, they can be combined together for performance boost. Our contributions are as follows:

1. We investigate the fundamental cause of the shot-unstable problem of FSOD: the key issue for FSOD in the low-shot regime and high-regime are different from each other and needs targeted solution.

2. We propose two effective knowledge transfer strategies targeted for the low-shot and high-shot regimes of FSOD, respectively. The two strategies can be combined together for further improvement of FSOD.

3. Experiments on VOC and COCO benchmarks show that our method significantly outperforms the baseline method and achieves competitive results in both low-shot (shot<5) and high-shot (shot≥5) regimes.
2 RELATED WORK

2.1 Object Detection

Modern object detectors are built on top of deep neural networks. These methods can be broadly divided into single-stage detectors and two-stage detectors. Single-stage detectors are usually with high detection efficiency however a relatively low detection accuracy [18, 20, 26, 27]. Two-stage detectors mostly refer to Faster-RCNN [29] and its derivatives [1, 9, 22, 45]. They usually attain higher performance than single-stage detectors thanks to the stage-wise refining pipeline. Also, this flexible architecture makes them easily adaptable to extended tasks such as FSOD. In addition to these two families, some anchor-free detection methods [13, 37, 50, 56, 57] are proposed to release detectors from burdensome anchor settings.

2.2 Few-shot Object Detection

Few-shot object detection approaches can be broadly grouped into three branches: transfer learning based methods, metric-learning based methods, and meta-learning based methods. Transfer learning based methods mainly focus on transferring knowledge from base classes to novel classes and unleashing the potential of fine-tuning [14, 47, 53]. Chen et al. [2] combine the advantages of Faster RCNN and SSD to alleviate the transfer difficulties from the source domain to the target domain. Recently, Wang et al. [40] rekindle the interest in transfer learning by showing its potential in improving FSOD, which inspires a lot of follow-up works [35, 54]. Their proposed method, named TFA, freezes the parameters of the model trained on the base classes and only fine-tunes the detection head with the novel classes. Meta-learning based methods follow the ideas of meta-learning in few-shot classification task [6, 15, 21, 24, 28, 31, 33, 38], which intend to learn generic knowledge across base classes and then generalize to novel classes. Specifically, Yan et al. [47] propose to conduct meta-learning over RoI regions. Kang et al. [10] design a few-shot detection method using a meta feature learner and a reweighting module.

Metric-learning based methods focus on learning better representations [12, 30, 36, 55]. Leonid et al. [11] propose a sub-network to learn an embedding space and apply it for novel class detection. Sun et al. [35] adopt supervised contrastive learning to learn better representations, which is implemented with an additional contrastive branch to guide RoI feature learning. Except for these methods, there are some other interesting works proposed for FSOD. Zhang et al. [54]’s work demonstrates that hallucination is helpful for few-shot detection. Wu et al. [43] believe that there is a universal prototype across all categories, with which the features learned from base classes can be generalized well to novel classes. Qiao et al. [23] propose DeFRCN, which largely improves few-shot detection performance through multi-stage and multi-task decoupling.

2.3 Distribution Calibration

The goal of Distribution Calibration (DC) is to align a target distribution to a reference distribution. The idea has been used for solving many unbalanced distribution problems. In [52], Zhang et al. investigate the performance bottleneck of the two-stage learning framework and proposed a unified distribution alignment strategy for long-tail visual recognition. Distribution calibration is also adopted in regression tasks [34]. Their work shows that predictions from previously trained regression models can be improved through distribution calibration. In [32], Shen et al. propose to learn a binary network by calibrating the latent representation through a teacher-student paradigm. Recently, Yang et al. [49] propose to reuse the statistics from many-shot classes and transfer them to better estimate the distributions of the few-shot classes according to their class similarities.

3 OUR METHOD

Before delving into the details of our proposed method, we first review the basic problem setting of FSOD and then take a brief look at our baseline method TFA.

3.1 Problem Setting

We follow the FSOD setting introduced in [10]. Given two sets of classes: base classes $C_b$ and novel classes $C_n$, the learning procedure of few-shot detection is normally divided into two stages. The first stage is training a base model on sufficient training data of base classes $C_b$. On the second stage, model is fine-tuned on novel classes $C_n$. $K$ samples (normally $K \leq 10$) for each class are used in the second stage fine-tuning. It is worth noting that there is no overlap between $C_b$ and $C_n$, that is $C_b \cap C_n = \emptyset$. In the second stage, to preserve the performance on base classes, model is fine-tuned on a balanced set containing training samples from both base classes $C_b$ and novel classes $C_n$.

3.2 Review of TFA

TFA is built on top of a two-stage detector Faster-RCNN. It consists of two learning stages. In the first stage, the Faster-RCNN is trained on base classes $C_b$. In the second stage, the Faster-RCNN model is fine-tuned on novel classes $C_n$, where each class contains $K$ training samples. Notably, during the fine-tuning stage, parameters of the backbone network and RoI feature extractor are fixed, only parameters of the prediction head, i.e., the classification branch.

Figure 2: Illustration of baseline method TFA. The learning procedure of TFA consists of 2 stages: (1) base training and (2) few-shot fine-tuning. During fine-tuning, only parameters of the detection head are updated.
3.3 Base Training and Distribution Calculation

In this section, we present our proposed method in detail. The learning procedure of our proposed method is divided into two steps: (1) base training and distribution calculation and (2) few-shot fine-tuning with distribution calibration. In the first stage, a base detector (e.g., Faster-RCNN) is trained on base classes \(C_b\) and compute statistic information of base classes. The only component in the detection model we should concern with is the task-relevant prediction head, and we fine-tune it to adapt the model to the novel classes.

Distribution Calculation. In a Faster-RCNN model, RoI feature \(x^c_i \in \mathbb{R}^d\) is extracted by the RoI feature extractor for each proposal. Then we use them to calculate the distributions of all base classes.

For simplicity, we assume that the class distribution \(\Omega_c\) of each category is a multidimensional Gaussian distribution. Formally, it can be written as \(\{\Omega_c = (\mu_c, \sigma_c) | c \in C_b, \mu_c \in \mathbb{R}^d, \sigma_c \in \mathbb{R}^{d \times d}\}\). \(\mu_c\) and \(\sigma_c\) are mean and covariance of \(\Omega_c\). We use all RoI features that are considered as positive samples \((IOU \geq 0.5)\) to calculate the distributions of the base classes. The mean and variance are calculated as follows \((c \in C_b\) denotes a base class, \(N_c\) denotes the number of positive RoI features of class \(c\) in calculation):\

\[
\mu_c = \frac{1}{N_c} \sum_{i} x^c_i 
\]

\[
\sigma_c = \frac{1}{N_c - 1} \sum_{i} (x^c_i - \mu_c)(x^c_i - \mu_c)^T 
\]

3.4 Distribution Calibration

In this section, we present our proposed distribution calibration method targeted at the low-shot regime. In distribution calibration, biased novel classes distributions are calibrated using distributions and the regression branch are fine-tuned. The two-stage learning procedure and model construction of TFA are illustrated in Figure 2. The essential idea of TFA is that it provides a shared backbone for base and novel classes, which makes it feasible to transfer knowledge from base classes to novel classes. The only component in the detection model we should concern with is the task-relevant prediction head, and we fine-tune it to adapt the model to the novel classes.
of the base classes, which are learned by sufficient samples. Meanwhile, shift compensation is proposed to obtain a better estimation of base class distributions.

**Shift Compensation.** After the first stage training, we have a detector well-trained on abundant base class data and the distributions of the base classes. In the second stage, the detection model is fine-tuned on novel classes. Considering that when fine-tuning the model on new categories (novel classes), the distributions of base classes may shift [16, 51]. Therefore, we propose to compensate for this shift to make sure that the base class distributions would not mislead the novel class distributions.

Formally, we use $t$ to denote the current training iteration. All base class positive RoI features at iteration $t$ are denoted as $X_t = \{x_i^t\}$ and base distributions are denoted as $\{\Omega_c = (\mu_c, \sigma_c)\}_{c=1}^{C_b}$. For each positive base class RoI features $x_i^t$ in $X_t$ (class label of $x_i^t$ is denoted as $c$), distribution of base class $c$ is compensated as follows:

$$\mu_c = \theta \times \mu_c + (1 - \theta) \times x_i^t$$  \hspace{0.5cm} (3)

where $\theta$ is a hyper-parameter controlling the compensation degree.

**Distribution Calibration.** After compensation, we use the compensated base class distributions to calibrate the biased novel class distributions. Specifically, for a novel RoI feature $x$ of class $n$ in the fine-tuning stage, we find its $k$ nearest base classes and use them to calibrate the novel distribution. Formally, we have:

$$D = \{d_c = \|x - u_c\|_2, c \in C_b\}$$  \hspace{0.5cm} (4)

$$c^\text{cal}_b = \{c|d_c \in \text{min}_k D, c \in C_b\}$$  \hspace{0.5cm} (5)

where $c^\text{cal}_b$ stands for the $k$ nearest base classes. The closer the centers of two classes to each other, the more similar they are. Therefore, the knowledge is more likely to be shared between them. Here, the biased distribution of novel class $n$ is calibrated with its $k$ nearest base classes as follows:

$$\mu_n = x + \frac{1}{k} \sum_c^{c^\text{cal}_b} \mu_c$$  \hspace{0.5cm} (6)

$$\sigma_n = \frac{1}{k} \sum_c^{c^\text{cal}_b} \sigma_c + \alpha$$  \hspace{0.5cm} (7)

$\mu_n$ and $\sigma_n$ denotes the calibrated novel distribution of class $n$. $\alpha$ is a hyper-parameter which controls the dispersion of the calibrated distribution. After calibration, we sample $M$ synthetic RoI features $\{\hat{x}\}$ from the calibrated novel distribution. This procedure can be denoted as:

$$\{\hat{x}\} = \{\hat{x}|\Omega_n = (\mu_n, \sigma_n)\}$$  \hspace{0.5cm} (8)

Finally, we feed both real sample $x$ and synthetic samples $\{\hat{x}\}$ into the detection head for training. The loss values in the fine-tuning stage come from both real and synthetic samples. Benefit from this calibration-and-generation process, the inner-class variation of the novel classes is enriched, and the over-fitting will be alleviated.

### 3.5 Distribution Regularization

Through distribution calibration, the lack of inner-class variation issue is addressed. However, the primary challenge in the high-shot regime remains unsolved: the misalignment between the learned distributions and the true distributions. Commonly, the ImageNet pre-trained network is used to initialize the base detector. Considering that ImageNet pre-trained networks are rich in knowledge, they can provide useful guidance for feature learning during fine-tuning[23]. Therefore, we propose to further fine-tune the detector with distribution regularization.

Our proposed distribution regularization strategy is shown in Figure 4. The ImageNet pre-trained network, which is used in parameter initializing in the base model training, is used for regularizing and guiding the fine-tuning. Specifically, let $X$ denote an image during fine-tuning. We feed $X$ and all positive foreground regions (output by RPN in the few-shot detector) into the ImageNet pre-trained network to extract RoI features $\{x_i\}$. Let $\{x_i\}$ denote the positive RoI features produced by the few-shot detector, we use a distance-based loss function to enforce the closeness between $\{x_i\}$ and $\{\hat{x}_i\}$ during fine-tuning:

$$L_{\text{reg}} = \frac{1}{N} \sum_{i=1}^{N} ||x_i - \hat{x}_i||_2$$  \hspace{0.5cm} (9)

where $N$ denotes the number of foreground regions. The joint loss is a combination of prediction loss and regularization loss:

$$L = L_{\text{cls}} + L_{\text{bbox}} + \lambda L_{\text{reg}}$$  \hspace{0.5cm} (10)

where $\lambda$ balances the detection loss and regularization loss. To make sure RoI features are learnable, the RoI feature extractor is not fixed during fine-tuning. Through our proposed distribution regularization, the semantic information contained in the ImageNet-pretrained network is fully utilized, and the biased learned distributions are implicitly aligned.

### 4 EXPERIMENTS

We conduct extensive experiments on PASCAL VOC (07+12) and MS COCO benchmarks. Our proposed method yields new state-of-the-art performance in most low-shot settings (shot<5) and also achieves competitive results in high-shot settings (shot≥5). We
Table 2: Detection results of novel classes on COCO benchmark under low-shot settings (shot=1/2/3). "DC" indicates distribution calibration and "DR" denotes distribution regularization. Our proposed method can achieve new state-of-the-art performance.

| Type             | Method                          | Split1          | Split2          | Split3          |
|------------------|---------------------------------|-----------------|-----------------|-----------------|
|                  |                                 | 1   | 2   | 3   | 5   | 10  | 1   | 2   | 3   | 5   | 10  | 1   | 2   | 3   | 5   | 10  |
| Meta Learning    | FSRW [10]_{ECCV19}              | 14.8| 15.5| 26.7| 33.9| 47.2| 5.7 | 15.3| 22.7| 30.1| 40.5| 21.3| 25.6| 28.4| 42.8| 45.9|
|                  | MetaDet [42]_{ECCV19}           | 18.9| 20.6| 30.2| 36.8| 49.6| 21.8| 23.1| 27.8| 31.7| 43  | 20.6| 23.9| 29.4| 43.9| 44.1|
|                  | Meta-RNN [47]_{ECCV19}          | 19.9| 25.5| 35  | 45.7| 51.5| 10.4| 19.4| 29.6| 34.8| 45.4| 14.3| 18.2| 27.5| 41.2| 48.1|
|                  | FsDetView [46]_{ECCV20}         | 24.2| 35.3| 42.4| 49.1| 57.4| 21.6| 24.6| 31.9| 37.0| 45.7| 31.2| 30.0| 37.2| 43.8| 49.6|
|                  | CME [16]_{CVPR21}               | 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|
|                  | Wu et al. [43]_{Arxiv21}        | 43.8| 47.8| 50.3| 55.4| 61.7| 31.2| 30.5| 41.2| 42.2| 48.3| 35.5| 39.7| 43.9| 50.6| 53.5|
| Metric Learning  | RepMet [11]_{CVPR19}            | 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|
|                  | FSCE [35]_{CVPR21}              | 44.2| 43.8| 51.4| 61.9| 63.4| 23.7| 30.6| 38.4| 43.0| 48.5| 37.2| 41.9| 47.5| 54.6| 58.5|
| Transfer Learning| MPRSR [35]_{ECCV20}             | 41.7| 43.1| 51.4| 55.2| 61.8| 24.4| 29.3| 39.2| 39.9| 47.8| 35.6| 40.6| 42.3| 48.0| 49.7|
|                  | FRN-R [40]_{ICML20}             | 8.2 | 20.3| 29.0| 40.1| 45.5| 13.4| 20.6| 28.6| 32.4| 38.8| 19.6| 20.8| 28.7| 42.2| 42.1|
|                  | TFA w/cos [40]_{ICML20}         | 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|
|                  | Halluc [54]_{CVPR21}            | 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|
|                  | Fan et al. [5]_{CVPR21}         | 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|
|                  | **Our Method**                  | **46.7**| **53.1**| **53.8**| **61.0**| **62.1**| **30.1**| **34.2**| **41.6**| **41.9**| **44.8**| **41.0**| **46.0**| **47.2**| **55.4**| **55.6**|

Table 1: Novel class AP50 (nAP50) results on PASCAL VOC. Red and Blue indicate the best and the second best results respectively. Our proposed method significantly outperforms baseline TFA and achieves competitive results compared with other methods.

Table 2: Detection results of novel classes on COCO benchmark under low-shot settings (shot=1/2/3). "DC" indicates distribution calibration and "DR" denotes distribution regularization. Our proposed method can achieve new state-of-the-art performance.

| Methods                  | nAP  | nAP50 | nAP75 | nAP  | nAP50 | nAP75 | nAP  | nAP50 | nAP75 |
|--------------------------|------|-------|-------|------|-------|-------|------|-------|-------|
| MPRSR [35]_{ECCV20}     | 2.3  | 4.1   | 2.3   | 3.5  | 6.3   | 3.4   | 5.2  | 9.5   | 5.1   |
| FsDetView [46]_{ECCV20}  | 3.2  | 8.9   | 1.4   | 4.9  | 13.3  | 2.3   | 6.7  | 18.6  | 2.9   |
| Hallucination [54]_{CVPR21} | 4.4  | 7.5   | 4.9   | 5.6  | 9.9   | 5.9   | 7.2  | 13.3  | 7.4   |
| TFA w/cos [40]_{ICML20} | 3.4  | 5.8   | 3.8   | 4.6  | 8.3   | 4.8   | 6.6  | 12.1  | 6.5   |
| TFA w/cos + DC (ours)   | 5.1  | 9.4   | 5.0   | 6.0  | 11.4  | 6.0   | 7.3  | 14.5  | 7.0   |
| TFA w/cos + DC + DR (ours) | 5.7  | 10.6  | 5.8   | 7.1  | 13.3  | 7.0   | 8.6  | 17.2  | 7.9   |

employ the same setting (base/novel split, base/novel samples) as in [40] for fair comparisons. Before presenting experimental results, we first elaborate on benchmark datasets and implementation details.

### 4.1 Benchmark Datasets

**PASCAL VOC.** For PASCAL VOC, all classes are divided into 15 base classes and 5 novel classes. VOC07 and 12 trainval-sets are used for base model training. Following TFA, the same three base/novel class splits (namely split 1, 2, 3) and base/novel instances are used in experiments. We report AP50 of both base and novel classes in all settings.

**COCO.** On COCO benchmark, 80 categories are divided into 60 base classes and 20 novel classes. Evaluations are performed under shot=10/30 settings on COCO 2017 val-set. To further validate the effectiveness of our proposed method, we conduct experiments in shot=1/2/3 settings. Experimental setup is the same as in previous works [40, 54]. COCO-style metrics are reported.

### 4.2 Implementation Details

For base model training on PASCAL VOC, we train model 50k iterations on 4 GPUs with base-size 4. The initial learning rate is set to 0.02 and reduced by 0.1 at 40k iterations. SGD optimizer with momentum 0.9 and weight decay 1e-4 is used. Positive RoI contains to 10 samples, namely \( M \). Numbers of new generated samples are similar.

For few-shot fine-tuning, the initial learning rate is set to 0.002, and fine-tuning steps are adjusted according to the number of shots. Model will be fine-tuned on a balanced set consisting of data from both base and novel classes where each class contains \( K \) samples, namely \( K \) shot setting. We firstly fine-tune the detector with distribution calibration and then we fine-tune the detector with distribution regularization. To maintain a balanced base-novel set during fine-tuning, we also sample \( M \) base class synthetic samples. Implementation details on COCO benchmark are similar.

As for hyper-parameters, \( \alpha \) in Equation 7 is experimentally set to \( 10^{-4} \) or \( 10^{-3} \). Numbers of new generated sample \( M \) is set to
### 4.3 Main Results

**PASCAL VOC.** Detection results of our proposed method are shown in Table 1. Our proposed method yields significant performance gain in both low-shot settings (shot=5) and high-shot settings (shot≥5) compared to baseline TFA. Meanwhile, compared with other FSOD methods, our proposed distribution calibration method can achieve new state-of-the-art performance in most low-shot settings (shot<5). In high-shot settings, our method can also achieve comparable results. More importantly, our proposed method can achieve more stable results: compared with other FSOD methods whose performance are unstable across different shots, our method can achieve competitive performance in not only low-shot but also high-shot settings.

**COCO.** Experimental results on COCO benchmark under low-shot settings are shown in Table 2. Our method can largely outperform baseline TFA. Compared with existing methods, the proposed method achieves new state-of-the-art performance. Experimental results on COCO benchmark under high-shot settings are shown in Table 3. Our method can achieve competitive results compared with existing state-of-the-art FSOD detectors, our method achieves competitive performance.

### 4.4 Ablation Studies

**Ablation for shift compensation.** Ablation studies of our proposed shift compensation on PASCAL VOC benchmark is shown in Table 4. "w" and "w/o" indicates distribution calibration with and without shift compensation, respectively. Results show that when DC and DR are combined together, the performance comes to the best, which demonstrates that DC and DR are not mutually exclusive but can reinforce each other.

| Type            | Methods                          | shot=10          | shot=30          | shot=50          | shot=75          | shot=100         |
|-----------------|----------------------------------|------------------|------------------|------------------|------------------|------------------|
| Meta Learning   | FSRW [10]_{ICCV19}               | 5.6              | 12.3             | 4.6              | 0.9              | 3.5              | 10.5             |
|                 | Meta-Det [42]_{ICCV19}           | 7.1              | 14.6             | 6.1              | 1.0              | 4.1              | 12.2             |
|                 | Meta-RCNN [47]_{ICCV19}          | 8.7              | 19.1             | 6.6              | 2.3              | 7.7              | 14.0             |
|                 | Wu et al. [43]_{Arxiv21}         | 11.0             | -                | 10.7             | 4.5              | 11.2             | 17.3             |
|                 | +0.7                             |                  |                  |                  |                  |                  |                  |
| Metric Learning | FSC [35]_{CVPR21}               | 11.9             | -                | 10.5             | -                | -                | -                |
|                 | +0.5                             |                  |                  |                  |                  |                  |                  |
| Transfer Learning| MPSR [44]_{ICCV20}              | 9.8              | 17.9             | 9.7              | 3.3              | 9.2              | 16.1             |
|                 | Li et al. [17]_{CVPR21}          | 11.3             | 20.3             | -                | -                | -                | -                |
|                 | TFA w/cos [40]_{ICML20}          | 10.0             | 19.1             | 9.3              | -                | -                | -                |
|                 | TFA w/cos + DC (ours)            | 9.9              | 19.0             | 9.0              | 4.3              | 8.9              | 15.9             |
|                 | TFA w/cos + DR (ours)            | 12.1             | 23.9             | 11.1             | 4.0              | 11.7             | 19.7             |
|                 | TFA w/cos + DC + DR (ours)       | 12.5             | 25.3             | 11.1             | 5.0              | 12.2             | 19.6             |
|                 | +0.7                             |                  |                  |                  |                  |                  |                  |

| Type            | Methods                          | shot=10          | shot=30          | shot=50          | shot=75          | shot=100         |
|-----------------|----------------------------------|------------------|------------------|------------------|------------------|------------------|
| Novel Classes   | + DC (w/o SC)                    | 44.6             | 50.9             | 51.3             | 58.2             | 58.0             | 58.0             |
|                 | + DC (w SC)                      | 45.8             | 52.4             | 52.9             | 58.7             | 58.6             | 58.6             |
|                 | Improvement†                     | +1.2             | +1.5             | +1.6             | +0.5             | +0.6             | +0.6             |
|                 |                                  |                  |                  |                  |                  |                  |                  |
| Base Classes    | TFA w/cos [40]                   | 79.6             | 78.9             | 79.1             | 79.3             | 78.4             | 78.5             |
|                 | + DC (w/o SC)                    | 78.6             | 78.1             | 78.8             | 78.6             | 79.0             | 79.0             |
|                 | + DC (w SC)                      | 79.1             | 78.4             | 79.3             | 79.0             | 79.7             | 79.7             |
|                 | Improvement†                     | +0.5             | +0.3             | +0.5             | +0.4             | +0.7             | +0.5             |
|                 |                                  |                  |                  |                  |                  |                  |                  |

Table 3: Performance comparisons on COCO benchmark. "DC" and "DR" denote distribution calibration and distribution regularization, respectively. Compared with existing state-of-the-art FSOD detectors, our method achieves competitive performance.

Table 4: Ablation experiments on shift compensation. "DC" and "SC" denote distribution calibration and shift compensation, respectively. Our proposed shift compensation can boost performance of distribution calibration on both novel and base classes.

1. $\lambda$ in regularization-based fine-tuning is set to $10^{-4}$, $\theta$ in shift compensation is set to 0.9999. As for $k$ in distribution calibration, we give more details in ablation study.
Ablation for distribution calibration and distribution regularization. We conduct ablation studies on distribution calibration and distribution regularization on PASCAL VOC benchmark. Results are shown in Table 5. We can see that in low-shot settings the improvements mainly come from distribution calibration, while in high-shot settings the improvements led by distribution regularization is higher. This validates our assumption that the primary issues for low-shot and high-shot regimes are different from each other. Besides, for base classes, detection performance is slightly compromised in only few cases, our proposed method can achieve better performance compared with baseline TFA in most cases.

| Ablation | Novel Classes | Base Classes |
|----------|---------------|--------------|
| TFA w/cos [40] | 39.8 | 79.6 |
| + DC | 45.8 | 79.1 |
| Improvement↑ | +6.0 | +8.3 |
| | +6.2 | +8.4 |
| + DC + DR | 46.7 | 79.3 |
| Improvement↑ | +0.9 | +3.2 |
| | +0.7 | +0.9 |

Table 5: Ablation experiments of the distribution calibration and distribution regularization on PASCAL VOC. “DC” indicates distribution calibration and “DR” denotes distribution regularization.

Figure 5: The effect of different k in distribution calibration.

5 CONCLUSIONS
In this paper, we reveal the fundamental cause of shot-unstable in FSOD: the problem nature of few-shot detection is different in low-shot and high-shot cases. In the low-shot regime the primary challenge is the lack of inner-class variation, while in the high-shot regime the primary issue is a misalignment problem. Meanwhile, we propose to address these distinct two issues by transferring the learned knowledge more effectively. Our proposed method achieves competitive and more stable results.

ACKNOWLEDGMENTS
This work was supported by National Natural Science Foundation of China (Grant No. 62176017 and No.41871283).
