Transformation Invariant Few-Shot Object Detection

Aoxue Li Zhenguo Li
Huawei Noah's Ark Lab, China
lax@pku.edu.cn, Li.Zhenguo@huawei.com

Abstract

Few-shot object detection (FSOD) aims to learn detectors that can be generalized to novel classes with only a few instances. Unlike previous attempts that exploit meta-learning techniques to facilitate FSOD, this work tackles the problem from the perspective of sample expansion. To this end, we propose a simple yet effective Transformation Invariant Principle (TIP) that can be flexibly applied to various meta-learning models for boosting the detection performance on novel class objects. Specifically, by introducing consistency regularization on predictions from various transformed images, we augment vanilla FSOD models with the generalization ability to objects perturbed by various transformation, such as occlusion and noise. Importantly, our approach can extend supervised FSOD models to naturally cope with unlabeled data, thus addressing a more practical and challenging semi-supervised FSOD problem. Extensive experiments on PASCAL VOC and MSCOCO datasets demonstrate the effectiveness of our TIP under both of the two FSOD settings.

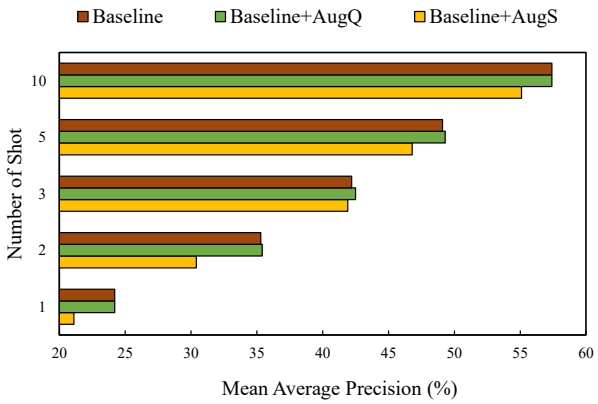

Figure 1. Results with naive data augmentation. The evaluation metrics are the 1, 2, 3, 5, 10-shot detection performance (i.e., mean Average Precision, mAP) on the first novel class set of PASCAL VOC dataset. Notations: 'Baseline' – A meta-learning-based approach [23] that has achieved the state-of-the-art results; 'Baseline + AugS' – Augmenting support images when training the baseline model. 'Baseline + AugQ' – Augmenting query images when training the baseline model.

1. Introduction

While the availability of large number of labeled data has enabled deep neural networks to dominate the computer vision community. they struggle in addressing problems with scarce labeled data [6, 15]. In contrast, humans can rapidly learn new concepts with only a few examples available. This big gap between humans and deep neural networks provides fertile ground for developing deep learning techniques. Due to this fact, few-shot learning, which learns algorithms that allow for better generalization on tasks with a few labeled training samples, has become topical. Different from most previous works designed for few-shot classification, we focus on a more challenging and more practical case – few-shot object detection (FSOD) [5, 9, 2]. Specifically, given a set of base classes with rich labeled data per class and a set of novel class with a few labeled data per class, FSOD aims to learn a model to detect objects from both base and novel classes.

Recent progress of FSOD has featured meta-learning strategy [10, 24, 23]. It uses a pool of auxiliary detection tasks generated from base class training set to perform transfer learning to novel class tasks with only a few examples available. Here, each auxiliary task is constructed to simulate the few-shot scenario: given a small training set (called support set) with a labeled instance per class, and a small test set (called query set), a meta learner trains the target detector in a guided manner: For each class, its support sample is used to extract class-wise representative features and embedded into a guidance vector. Then the guidance vector is incorporated into query feature learning to facilitate the query sample features suitable for detecting objects of the target class.

Orthogonal to the design of meta-learning strategy, we tackle this challenging FSOD problem from the perspective of sample expansion and further improve its performance. A naive solution is to involve transformed images

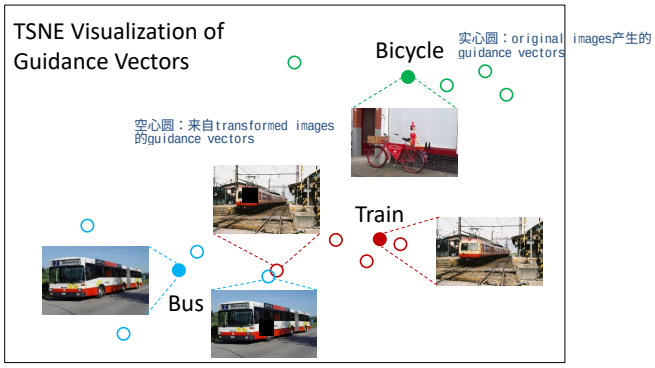

Figure 2. The TSNE Visualization of guidance vectors obtained by naive data augmentation solution. In this figure, guidance vectors from different classes are illustrated in different colors. The guidance vectors generated by using transformed images are denoted by hollow circles, while the guidance vectors of original images are denoted by solid circles. These original images and its transformed variants are also provided. The guidance vectors of transformed images generated by using the same original image are shown to be separable in the guidance embedding space. This means that the learned guidance vectors didn't effectively encode the representative features which should be invariant to image transformation.

into the training process. However, it is impressive that simply adding the data augmentation techniques leads to very limited improvement or even performance drop, as shown in Figure 1. To analysis this phenomenon, we provide a TSNE visualization of the guidance vectors extracted by using transformed images, as shown in Figure 2. Here, guidance vectors from different classes are illustrated in different colors. The guidance vectors of original images are denoted by solid circles, while the guidance vectors of their transformed images are denoted by hollow circles. These original images and their transformed variants are also provided. We can observe that guidance vectors of transformed images generated by using the same original image are separable in the guidance embedding space. This means that the learned guidance vectors didn't effectively encode the class-wise representative features which should be invariant to image transformation.

To overcome this issue, a novel approach named by Transformation Invariant Principle (TIP) is proposed. The TIP applies consistency regularization on guidance vectors from various transformed images to provide additional supervision for guidance learning. As illustrated in Figure 3, the TIP for guidance extraction branch is implemented by adding a Transformed Guidance Consistency (TGC) Loss on the top of the guidance vectors of original images and their transformed variants. The TGC Loss computes the difference between guidance vectors generated from an original image and its transformed variants. Moreover, the TIP introduces the proposal consistent regularization into query image prediction to generate transformation invariant query features. This is implemented as a proposal detection network that takes transformed images as inputs and outputs suitable Region of Interest (RoI) proposals for their original images. The prediction of bounding boxes is conditioned on these RoI proposals and the transformation invariant guidance vectors learned by TGC Loss. The proposed TIP can be used to cope with unlabeled images and thus extend our approach to a more realistic yet more challenging scenario, i.e. semi-supervised FSOD. In this way, both fully-supervised and semi-supervised FSOD problems can be handled in a unified detection framework. Experimental results on PASCAL VOC and MSCOCO datasets demonstrate that our approach is effective for both of the two FSOD settings.

In summary, our contributions are three folds:

- To the best of our knowledge, this is the first work to address the challenging FSOD problem from the perspective of sample expansion.
- We propose a simple yet effective approach, named by TIP, to improve the generalization ability over transformed images, and experimental results demonstrate that our method achieves state-of-the-art results on two benchmark datasets.
- Our approach can be easily extended to a more realistic yet more challenging semi-supervised FSOD scenario, with superior performance obtained. This further validates the effectiveness of the proposed approach.

2. Related Work

Few-shot learning is a fundamental yet unsolved problem in machine learning and computer vision [1, 19, 13, 14, 11, 25, 12]. Most of these existing work is developed in the context of classification, while we focus on the more challenging object detection task in the few-shot scenario. Meta-learning is a popular solution to address FSOD [22, 10, 23, 24]. Wang et al. developed a weight prediction meta-model to produce the parameters of category-specific components from few examples [22]. A detector with these learned features is used for novel class objects detection. Kang et al. proposed to reweigh pre-trained features by a meta-model and facilitated these features suitable for detecting novel classes [10]. Yan et al. extended Kang et al's method and developed a meta-learning model over more fine-grained RoI features, and thus achieve better performance on novel classes [24]. Xiao et al. propose a unified meta-learning framework that can tackle both 2D detection task and 3D viewpoint estimation task [23]. In this

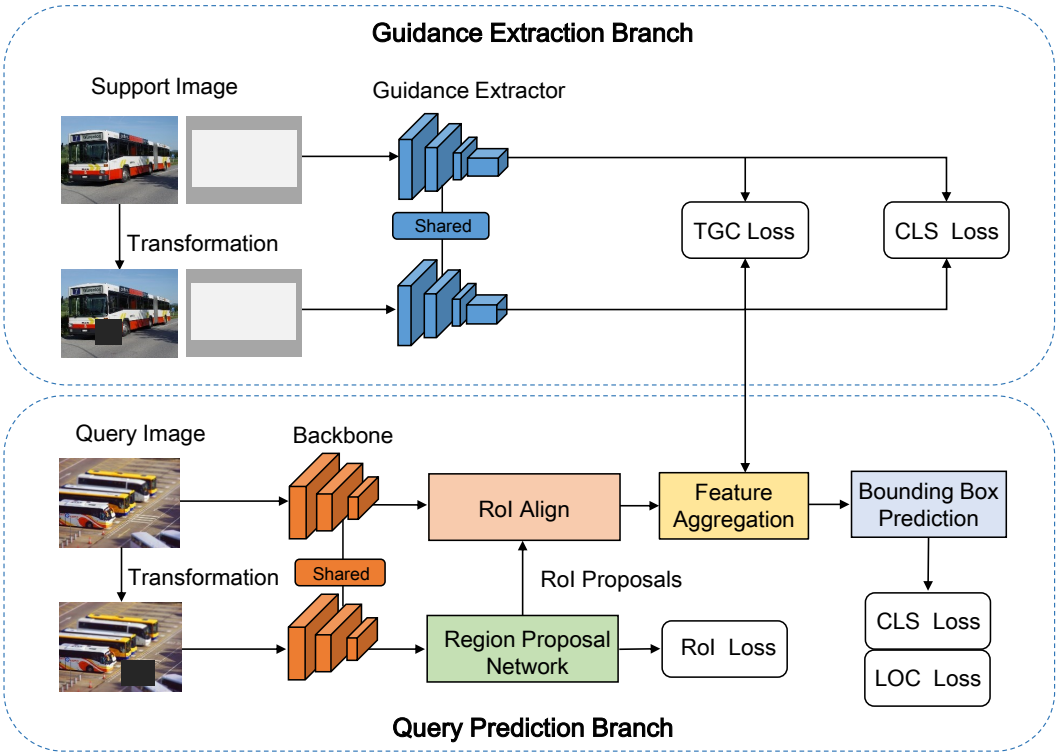

Figure 3. Overview of the proposed approach. To improve the generalization ability of current meta-learning-based FSOD approaches, our TIP approach applies consistency regularization on both guidance extraction and query prediction branches. For guidance extraction branch, we develop a TGC Loss over guidance vectors of original images and their transformed variants to impose consistency between them. For query prediction branch, we introduce a proposal consistent regularization by first predicting the RoI proposals of transformed images and then predicting bounding boxes of query samples conditioned on these RoI proposals as well as the transformation-invariant guidance vectors extracted by the guidance extraction branch.

paper, we propose to exploit TIP to improve the generalization ability of these meta-learning-based FSOD approaches. This will benefit object detection on novel classes with only a few labeled samples. The proposed TIP approach is orthogonal to the design of meta-learning approaches and can be applied to improve any meta-learning-based approaches for FSOD.

In addition to meta-learning-based approaches, some data synthesis approaches [8, 21, 18, 26, 7, 3] have been proposed for few-shot image classification scenario, but rare works are proposed for FSOD. These data synthesis approaches designed for few-shot classification can alleviate the severe lack of novel class data in some degree. Note that, although our approach is developed from the perspective of sample expansion, we focus on learning invariant representations by adding consistency regularization between transformed images, rather than generating fake data which cannot be distinguished from real images. Therefore, our approach is flexible to integrate previous data synthesis approaches.

3. Preliminary

We start by defining precisely the current meta-learning paradigm for FSOD. Recent progress of FSOD has been made possible by following an episodic meta-learning paradigm [10, 24, 22, 23]. Here, we are given a set of base classes C_{base} with sufficient labeled samples per class and a set of novel classes C_{novel} with a few labeled instances per class. A meta learner randomly samples many auxiliary training tasks from the whole training set to simulate the few-shot situation in novel classes.

Specifically, in each episode, we randomly select N class from base class set C_{base} to form an episodic class set C_e. For each class $c_i \in C_e$, we randomly select an image with an object of this class and thus form a small-scale training set, which is called support set $S = \{I_i^s\}_{i=1}^K$, each element denoting a support image. Then, we randomly select several samples from the remain base class images that also contain objects from C_e and form a test set, which is called query set Q. The meta learning framework consists of two

branches: a guidance extraction branch and a query prediction branch. In the guidance extraction branch, a guidance encoder $\mathcal{G}$ takes each support image $I_i^s \in S$ as input, and generates a guidance vector g_i that captures representative features related to its class c_i. A classification loss of these guidance vectors forces them to be separable in the guidance embedding space. In the query prediction branch, each query image is encoded by a query encoder $\mathcal{R}$ to obtain its visual features. Then, we aggregate each guidance vector and query features to facilitate more effective features to detect objects from its class. After that, a detection module $\mathcal{P}$ takes these aggregated features as inputs and outputs potential locations of bounding boxes and class probabilities. We optimize the parameters of the full model by minimizing the detection loss of all query samples and the classification loss of guidance vectors. Since this strategy can learn an effective learning algorithm that produces detector with good generalization on novel class test examples, we call this paradigm as meta-learning.

4. Transformation Invariant Principle

4.1. Motivation

Although the meta-learning-based approaches have achieved promising results, the generalization gap on unlabeled test samples from novel classes between fully-supervised setting and few-shot setting is still very large. In this work, we propose to narrow the gap from the perspective of sample expansion. However, due to the guidance inconsistency between different transformed images, simply adding data augmentation techniques fails to improve detection performance, as illustrated in Figures 1&2. To overcome this issue, we propose a TIP approach which imposes invariant consistency among transformed images. Our TIP approach introduces consistency regularization to the guidance extraction branch and query prediction branch, respectively. The proposed TIP approach can be inserted into the current meta-learning-based FSOD approaches and further improve their detection precision.

4.2. Transformation Invariant Guidance Extraction

To effectively encode representative features related to each class, the guidance vector of each class should be transformation invariant. To obtain such invariant guidance vectors, we design a Transformed Guidance Consistency (TGC) Loss, where the difference between guidance vectors of original images and their transformed images are computed.

Specifically, given a support image I_i^s, we feed it into a data transformation module $\mathcal{T}$ and thus generate a transformed image $\mathcal{T}(I_i^s)$. The transformation module can be implemented as non-parametric data augmentation techniques, e.g. Gaussian noise and cutout, or parametric data synthesis approaches. Then, we feed a support image $I_i^s \in S$ as well as its bounding box annotation L_i^s into the guidance encoder $\mathcal{G}$ and produce its guidance vector g_i. $\mathcal{G}$ is implemented as a convolutional neural network, e.g. ResNet-101 in our paper. The bounding box annotation L_i^s is converted to a binary mask and concatenated with the image to construct a four-channel input for $\mathcal{G}$. Given a support image I_i^s (with its annotation L_i^s), the formulation of its guidance vector is given as follows:

$$g_i = \mathcal{G}(I_i^s, L_i^s). \tag{1}$$

Similarly, we feed the transformed image $\mathcal{T}(I_i^s)$ into $\mathcal{G}$ and produce its guidance vector $\hat{g}_i$. The difference between the two vectors is then computed with a measurement function $\mathcal{M}$ and a transformed guidance consistency loss $\mathcal{L}_{TGC}$ formulated in Equation (2) is designed to minimize the difference.

$$\mathcal{L}_{TGC}(c_i) = \mathcal{M}(g_i, \hat{g}_i). \tag{2}$$

The measure function $\mathcal{M}$ is implemented by an L_2-norm over the difference of the two guidance vectors. Its formulation is given in Equation (3).

$$\mathcal{M}(g_i, \hat{g}_i) = \|g_i - \hat{g}_i\|. \tag{3}$$

We also analyze the forms of the measurement function to validate the effectiveness of the proposed approach (see ablation study in Section 5.1.3). In addition, we follow [23, 22] and add a cross entropy classification loss $\mathcal{L}_{CLS}^{G}$ that encouraging guidance vectors to be diverse for different classes.

By exploiting the consistency between transformed images, our TGC Loss forces guidance vectors of the same image to be more clustered and thus captures more representative features shared among the same objects in transformed images. These refined guidance vectors will help to produce more effective features for detecting novel class objects and thus benefit FSOD.

4.3. Transformation Invariant Query Prediction

For query prediction, we assume that results of images with spatial invariant transformation (such as Gaussian noise) should be consistent with those of original images. To achieve this, we first use transformed images to extract RoI proposals and then detect objects in original images conditioned on these transformed RoI proposals. We minimize the detection loss of original images to force the detector to perform consistent detection results on both original images and their transformed variants.

Specifically, given a query image I^q, we first feed it into the transformation module $\mathcal{T}$ to produces its transformed variant $\mathcal{T}(I^q)$. Then we feed the transformed image into a backbone network $\mathcal{B}$ followed by a Region Proposal Network (RPN) $\mathcal{RPN}$ [17] to extract a set of RoI proposals

$\hat{R} = \{\hat{r}_{ij}\}_{i=1}^{N_r}$, where N_r denotes the total number of RoIs. The formulation of these RoI proposals is given as follows:

$$\hat{R} = \{\hat{r}_j\}_{j=1}^{N_r} = \mathcal{RPN}((\mathcal{B}(\mathcal{T}(I^q))). \quad (4)$$

In the meanwhile, the original image I^q is fed into the backbone $\mathcal{B}$ to extract base feature maps. A RoI align layer $\mathcal{A}$ is exploited to align the RoI proposals of transformed images with feature map of original image and produces a set of RoI features conditioned on $\hat{R}$. These conditioned RoI features are formulated in Equation (5).

$$\hat{F} = \{\hat{f}_j\}_{j=1}^{N_r} = \mathcal{A}(\mathcal{B}(I^q), \hat{R}), \quad (5)$$

where $\hat{F} = \{\hat{f}_j\}_{j=1}^{N_r}$ denotes the set of conditioned RoI features.

After that, as the method described in Section 3, these RoI features are aggregated with the transformation invariant guidance vectors learned by the method proposed in Section 4.2 to produce features related to specific class object detection. The detection module $\mathcal{P}$ takes these aggregated features as inputs, and then outputs potential locations of bounding boxes and class probabilities.

To optimize the parameters in this branch, we use the same detection loss of query images as in [23, 24]. The detection loss is formulated in Equation (6).

$$\mathcal{L}_{Query} = \mathcal{L}_{LOC} + \mathcal{L}_{CLS}^{Q} + \mathcal{L}_{RoI}, \quad (6)$$

where $\mathcal{L}_{LOC}$ denotes the smooth L_1 loss over predicted bounding box locations, $\mathcal{L}_{CLS}^{Q}$ denotes the cross entropy loss over predicted class probabilities, and $\mathcal{L}_{RoI}$ is applied to the output of the RPN to distinguish foreground from background and refine the proposals.

By combining the classification loss and TGC Loss for guidance extraction, we minimize the loss function formulated in Equation (7) to train the whole detection model.

$$\mathcal{L}_{overall} = \mathcal{L}_{Query} + \mathcal{L}_{TGC} + \mathcal{L}_{CLS}^{G}. \quad (7)$$

4.4. Extension to Semi-Supervised Scenario

Although the TIP approach is originally designed for fully-supervised FSOD, it can be easily extended to cope with unlabeled images, thus leading to a more challenging yet realistic scenario, i.e., semi-supervised FSOD. Specifically, our approach exploits the similar framework as in the fully-supervised scenario. The only difference is adding a transformation consistency loss over unlabeled images to impose the consistency between guidance vectors of unlabeled images and its variants obtained by spatial invariant transformations. For each unlabeled image as well as its transformed image, we compute their guidance vectors by using the guidance extractor $\mathcal{G}$ and leverage the transformation consistency loss $\mathcal{L}_u$ to minimize their differences. The loss $\mathcal{L}_u$ for unlabeled images is formulated in Equation (8).

$$\mathcal{L}_u(I_i^u) = \|\mathcal{G}(\mathcal{T}(I_i^u), L_u) - \mathcal{G}(I_i^u, L_u)\|, \quad (8)$$

where $L_u = \phi$ denotes an empty annotation set. The bounding box annotation L_u is converted to an all-one mask and concatenated with the unlabeled image to construct a four-channel input for $\mathcal{G}$. Experimental results show that our method not only outperforms baseline methods and even outperforms fully-supervised approaches on some few-shot cases (see Table 5).

5. Experimental Evaluation

In this work, we evaluate our approach in two scenarios: 1) Standard FSOD where labeled instances are used for model training; 2) Semi-supervised FSOD where a large number of unlabeled samples are available during model training.

5.1. Standard Few-Shot Object Detection

In this section, two benchmark datasets, i.e., PASCAL VOC [4] and MSCOCO [16] are used to evaluate the effectiveness of our approach.

5.1.1 Experimental Setup

PASCAL VOC 2007 and 2012 datasets [4] consists of a total of 16.5k train-val images and 5k test images from 20 different categories. Consistent with the standard few-shot setup in [23], we use VOC 07 and 12 train-val sets for training and VOC 07 test set for testing. 15 of the 20 categories are considered as base classes, and the remaining 5 categories as novel classes. As in [23, 10, 22, 20], we use the same three base-novel class splits. Each novel class has only K bounding box annotations for model training. Here, K is set to be 1, 2, 3, 5, 10.

MSCOCO [16] is a more challenging benchmark for object detection. This dataset involves 80k training, 40k validation, and 20k test images over 80 object categories. Following the setting in [23, 24], 5k images from the validation set (denoted as minval) are used for evaluation, and the remaining train-val images for training. We use the 20 categories that also present in PASCAL VOC as novel classes, and the other 60 categories are used as base classes. Similar to PASCAL VOC, we randomly select a few bounding box annotations for novel classes and K is set to be 10 and 30. The training details and implementation details of our TIP approaches are given in the supplementary material. As in [23], we report results averaged over multiple random runs. We compare our approach with recent and representative FSOD approaches [23, 24, 20] that have achieved the state of-the-art results.

5.1.2 Results and Discussions

We provide the competitive results on PASCAL VOC and MSCOCO datasets in Tables 1&2. For PASCAL VOC

Model	Novel Class Set 1					Novel Class Set 2					Novel Class Set 3				
	1-shot	2-shot	3-shot	5-shot	10-shot	1-shot	2-shot	3-shot	5-shot	10-shot	1-shot	2-shot	3-shot	5-shot	10-shot
MRCNN [24]	19.9	25.5	35.0	45.7	51.5	10.4	19.4	29.6	34.8	45.4	14.3	18.2	27.5	41.2	48.1
TFA/w.fc [20]	22.9	34.5	40.4	46.7	52.0	16.9	26.4	30.5	34.6	39.7	15.7	27.2	34.7	40.8	44.6
TFA/w.cos [20]	25.3	36.4	42.1	47.9	52.8	18.3	27.5	30.9	34.1	39.5	17.9	27.2	34.3	40.8	45.6
FA [23]	24.2	35.3	42.2	49.1	57.4	21.6	24.6	31.9	37.0	45.7	21.2	30.0	37.2	43.8	49.6
TIP(ours)	**27.7**	**36.5**	**43.3**	**50.2**	**59.6**	**22.7**	**30.1**	**33.8**	**40.9**	**46.9**	**21.7**	**30.6**	**38.1**	**44.5**	**50.9**

Table 1. Comparative results for standard FSOD on the PASCAL VOC dataset. We evaluate the performance on three different sets of novel categories. The mean average precision (%) on the novel classes is used as the evaluation metrics of this dataset. The reported results are averaged over multiple runs. Our approach consistently outperforms competing models across all number of shots and all three novel class sets.

No.Shot	Model	Average Precision						Average Recall					
		0.5:0.95	0.5	0.75	S	M	L	1	10	100	S	M	L
10	MRCNN [24]	8.7	19.1	6.6	2.3	7.7	14.0	12.6	17.8	17.9	7.8	15.6	27.2
	TFA/w.fc [20]	9.1	17.3	8.5	-	-	-	-	-	-	-	-	-
	TFA/w.cos [20]	9.1	17.1	8.8	-	-	-	-	-	-	-	-	-
	FA [23]	12.5	27.3	9.8	2.5	13.8	19.9	20.0	25.5	25.7	7.5	27.6	38.9
	TIP(ours)	**16.3**	**33.2**	**14.1**	**5.4**	**17.5**	**25.8**	**23.6**	**30.2**	**30.5**	**12.7**	**32.3**	**43.8**
30	MRCNN [24]	12.4	25.3	10.8	2.8	11.6	19.0	15.0	21.4	21.7	8.6	20.0	32.1
	TFA/w.fc [20]	12.0	22.2	11.8	-	-	-	-	-	-	-	-	-
	TFA/w.cos [20]	12.1	22.0	12.0	-	-	-	-	-	-	-	-	-
	FA [23]	14.7	30.6	12.2	3.2	15.2	23.8	22.0	28.2	28.4	8.3	30.3	42.1
	TIP(ours)	**18.3**	**35.9**	**16.9**	**6.0**	**19.3**	**29.2**	**25.2**	**32.0**	**32.3**	**14.1**	**34.6**	**45.1**

Table 2. Comparative results for standard FSOD on the MSCOCO dataset. We evaluate the performance on 20 novel classes that also present in PASCAL VOC dataset. The evaluation metrics are average precision and average recall under different IoU thresholds, different number of predicted bounding boxes and different object scales, as in [16]. The reported results are averaged over multiple runs. Our approach achieves significant performance improvement, comparing with the state-of-the-art approach [23].

dataset, our TIP approach outperforms state-of-the-arts results across different number of shots. This indicates that the proposed approach can effectively detect objects with very limited bounding box annotations. Moreover, our TIP approach yields the best results on different novel class sets on the PASCAL VOC dataset, thanks to the transformation invariant guidance vectors and query features learned by our approach.

For MSCOCO dataset, our approach achieves big performance gains over the state-of-the art models, i.e., 16.3% vs. 12.5 % (FA) mAP on 10 shots scenario and 18.3 % vs. 14.7% (FA) mAP on 30 shots scenario. The significant performance improvement mainly comes from the cooperation of transformation invariant principle on query prediction branch and guidance extraction branch, which produces more effective guidance vector and query features for detecting objects from novel classes. Th significant performance improvement also demonstrates the scalability of our approach for large-scale FSOD problems. In particular, our approach outperforms the state-of-the-art approach with a large margin on different size of objects simultaneously. This demonstrates that our TIP approach is effective for detecting objects in different scales.

5.1.3 Ablation Study

We investigate the contributions of different proposed transformation invariant modules and summarize the results in Table 3. We compare our full model with three stripped-down versions : 1) 'Baseline' – the meta-learning model [23], without any transformation invariant regularization. 'TIGE' – the model similar to [23] except for the guidance extraction by the transformation invariant model proposed in Section 4.2; 'TIQP' – the model similar to [23] except for predicting bounding boxes of query images by the transformation invariant model proposed in Section 4.3. These ablations are based on 1, 2, 3, 5, 10-shot object detection performances on PASCAL VOC in the first base/novel split setting. We can observe that the model without any transformation invariant regularization gets the worst performance. Moreover, adding either transformation invariant regularization on guidance extraction or query prediction can improve the performance. Applying both modules further provides more performance gains. This implies that our TIP applied to both guidance extraction and query prediction branches is crucial to improve performance of FSOD models.

提高也不算太大？

Model	1-shot	2-shot	3-shot	5-shot	10-shot
Baseline	24.2	35.3	42.2	49.1	57.4
TIGE	26.5	35.9	42.8	49.5	59.2
TIQP	26.3	35.8	42.8	49.7	59.0
Full Model	**27.7**	**36.5**	**43.3**	**50.2**	**59.6**

Table 3. Ablation study on contribution of different transformation invariant modules in our approach. The evaluation metric is the 1, 2, 3, 5, 10-shot detection performance (mAP) on the first novel class set of the PASCAL VOC dataset. Notations: 'Baseline' – the meta-learning model [23], without any transformation invariant regularization; 'TIGE' – the model similar to [23] except for the guidance extraction by the transformation invariant model proposed in Section 4.2; 'TIQP' – the model similar to [23] except for predicting bounding boxes of query images by the transformation invariant model proposed in Section 4.3.

Moreover, we conduct a diagnosis experiments over the forms of measurement function formulated in Equation (2). In addition the L_2 distance given in Equation (3), we also provide another two measurement function: one is smooth L_1 loss function over the difference between normalized guidance vectors of original images and its transformed variants ; the other is the 'parametric L_2 distance' which is formulated in the following equation.

$$\|g_i - \theta(\hat{g}_i)\|, \tag{9}$$

where $\theta(\cdot)$ is an embedding function whose input dimension and output dimension are the same. The results of these measurement functions are illustrated in Table 4. We can observe that our L_2 distance solution yields better results than two alternative solutions. This indicates that our L_2 distance solution is more suitable than the other two alternatives.

5.1.4 Qualitative Results

We provide qualitative visualizations of the detected novel objects of PASCAL VOC in Figure 4. We show our model achieves much better detection result than baseline model, thanks to the transformation invariant guidance vectors and query features learned by our TIP approach.

5.2. Semi-Supervised Few-Shot Object Detection

5.2.1 Experimental Setup

To further evaluate the effectiveness of our approach, we test our approach in a more challenging yet practical setting, i.e., semi-supervised FSOD, where a small subset of labeled base class data and additional unlabeled samples is available for model training. The experiments are conducted on PASCAL VOC dataset. We first create two additional splits to separate the images of each class into disjoint labeled and unlabeled sets. Specifically, we randomly select 50% labeled samples from base classes to form a labeled set and the remain 50% samples are used to form an unlabeled set. Similarly, we creat a more challenging semi-supervised setting where base class set is split into 25% labeled samples and 75% unlabeled samples. For novel classes, we follow the setting in standard FSOD and only provide K bounding box annotations per novel class for training, where K is set to be 1, 2, 3, 5, and 10. For testing, we use the 5k samples from VOC 2007 test as in standard setting. The training strategy is the same as that in standard FSOD. The mean average precision on novel classes is used as evaluation metric for this dataset.

Model	1-shot	2-shot	3-shot	5-shot	10-shot
Smooth L_1	25.9	35.4	42.5	49.6	58.1
Parametric L_2	26.9	35.7	42.6	49.3	59.0
L_2 distance (ours)	**27.7**	**36.5**	**43.3**	**50.2**	**59.6**

Table 4. Diagnosis experiment over the forms of measurement function formulated in Equation (2). The evaluation is the same as in Table 3. Notations: 'Parametric L_2' – formulated in Equation (9). 'Smooth L_1' – smooth L_1 loss function over the difference between normalized guidance vectors of original images and its transformed variants. 'L_2 distance (ours)' – Our solution formulated in Equation (3). Our solution is shown to be more effective than two alternative solutions.

Since works on semi-supervised FSOD are rare, we use two striped-down versions of our approach, i.e., 'TIGE' and 'TIQP' in ablation study, as baselines for the semi-supervised scenario. In addition, we also compare our approach with fully-supervised FSOD approaches [23, 20, 24] which used the whole base class training set for model training. This comparison can measure the gap between semi-supervised FSOD approaches and fully-supervised ones.

5.2.2 Results and Discussions

Table 5 provides the comparative results on PASCAL VOC dataset under semi-supervised FSOD scenario. We can observe that: 1) Our full model outperforms the two baselines under the semi-supervised FSOD scenario. This validates the effectiveness of our approach in semi-supervised FSOD scenario. 2) Our semi-supervised FSOD approach outperforms some fully-supervised FSOD approaches in some few-shot cases. In particularly, our approach with 50% labeled samples for training achieves comparable results or even outperforms the fully-supervised state-of-the-arts (i.e. FA [23]) in some cases. The superior performance of our approach demonstrates its generalization ability in semi-supervised FSOD. This also confirms that, through the the proposed TIP, the model can learn to obtain a better novel class detector which copes with both labeled and unlabeled data to facilitate FSOD.

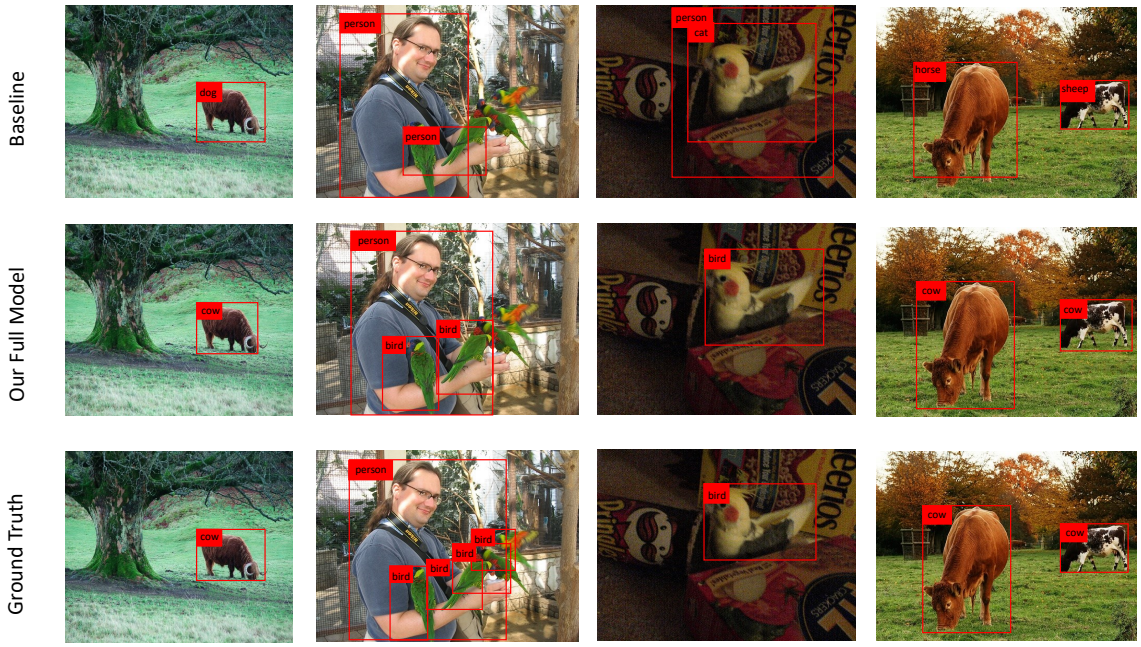

Figure 4. Qualitative visualizations of the detected novel objects obtained by baseline models and our approach on the PASCAL VOC dataset. The results of baseline model and our approach are given in the first and second rows, respectively. Their ground truth is given in the last row. Our approach yields much better detection results than baseline.

Supervision	Model	Novel Class Set 1					Novel Class Set 2				
		1-shot	2-shot	3-shot	5-shot	10-shot	1-shot	2-shot	3-shot	5-shot	10-shot
Fully	MRCNN [24]	19.9	25.5	35.0	45.7	51.5	10.4	19.4	29.6	34.8	45.4
	TFA/w.fc [20]	22.9	34.5	40.4	46.7	52.0	16.9	26.4	30.5	34.6	39.7
	TFA/w.cos [20]	25.3	36.4	42.1	47.9	52.8	18.3	27.5	30.9	34.1	39.5
	FA [23]	24.2	35.3	42.2	49.1	57.4	21.6	24.6	31.9	37.0	45.7
Semi-25%labeled	TIGE(ours)	20.4	30.7	34.6	43.5	50.6	16.9	20.4	28.1	34.0	41.9
	TIQP(ours)	20.5	30.2	34.8	43.1	50.3	17.6	21.0	27.5	34.3	41.2
	TIP(ours)	**21.3**	**31.6**	**35.9**	**44.0**	**51.5**	**18.1**	**21.9**	**28.7**	**35.9**	**42.3**
Semi-50%labeled	TIGE(ours)	25.3	32.4	38.8	46.7	54.3	18.7	27.4	31.1	36.5	44.9
	TIQP(ours)	24.3	31.1	42.4	45.2	52.1	19.4	26.4	31.7	36.5	44.7
	TIP(ours)	**25.6**	**33.2**	**43.1**	**46.6**	**55.7**	**21.1**	**27.8**	**32.2**	**38.0**	**45.6**

Table 5. Comparative results for semi-supervised FSOD on the PASCAL VOC dataset. We evaluate the performance on three different sets of novel categories. The mean average precision (%) on the novel classes is used as the evaluation metrics of this dataset. The reported results are averaged over multiple runs. Due to the limited text space, we show the results of the first two novel class sets. For the results of the third novel class set, please refer to the supplementary material.

6. Conclusion

In this paper, we propose a transformation invariant principle to address the challenging FSOD problem from the perspective of sample expansion. By introducing the consistency regularization on both guidance extraction and query prediction branches, our approach facilitates vanilla FSOD models invariant to various image transformations. In particular, our approach can cope with unlabeled image and thus be extended to semi-supervised FSOD scenario. Extensive experiments on the PASCAL VOC and MSCOCO datasets show that the proposed approach achieves state-of-the-art results under both fully-supervised and sem-supervised FSOD scenarios.

Acknowledgement. We also want to use TIP on MindSpore[1], which is a new deep learning computing framework. These problems are left for future work.

[1]https://www.mindspore.cn/

References

[1] Luca Bertinetto, Joao F Henriques, Philip HS Torr, and Andrea Vedaldi. Meta-learning with differentiable closed-form solvers. In *International Conference on Learning Representations,ICLR*, 2019. 2

[2] Hao Chen, Yali Wang, Guoyou Wang, and Yu Qiao. LSTD: A low-shot transfer detector for object detection. In *AAAI Conference on Artificial Intelligence*, pages 2836–2843, 2018. 1

[3] Zitian Chen, Yanwei Fu, Kaiyu Chen, and Yu-Gang Jiang. Image block augmentation for one-shot learning. In *AAAI Conference on Artificial Intelligence*, pages 3379–3386, 2019. 3

[4] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision,IJCV*, 88(2):303–338, 2010. 5

[5] Qi Fan, Wei Zhuo, Chi-Keung Tang, and Yu-Wing Tai. Few-shot object detection with attention-rpn and multi-relation detector. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition,CVPR*, pages 4012–4021, 2020. 1

[6] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pages 1126–1135, 2017. 1

[7] Hang Gao, Zheng Shou, Alireza Zareian, Hanwang Zhang, and Shih-Fu Chang. Low-shot learning via covariance-preserving adversarial augmentation networks. In *Advances in Neural Information Processing Systems,NeurIPS*, pages 983–993, 2018. 3

[8] Bharath Hariharan and Ross B. Girshick. Low-shot visual recognition by shrinking and hallucinating features. In *IEEE International Conference on Computer Vision, ICCV*, pages 3037–3046, 2017. 3

[9] Ting-I Hsieh, Yi-Chen Lo, Hwann-Tzong Chen, and Tyng-Luh Liu. One-shot object detection with co-attention and co-excitation. In *Advances in Neural Information Processing Systems,NeurIPS*, pages 2721–2730, 2019. 1

[10] Bingyi Kang, Zhuang Liu, Xin Wang, Fisher Yu, Jiashi Feng, and Trevor Darrell. Few-shot object detection via feature reweighting. In *IEEE Conference on Computer Vision,ICCV*, pages 8420–8429, 2019. 1, 2, 3, 5

[11] Kwonjoon Lee, Subhransu Maji, Avinash Ravichandran, and Stefano Soatto. Meta-learning with differentiable convex optimization. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 10657–10665, 2019. 2

[12] Aoxue Li, Weiran Huang, Xu Lan, Jiashi Feng, Zhenguo Li, and Liwei Wang. Boosting few-shot learning with adaptive margin loss. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, pages 12573–12581, 2020. 2

[13] Aoxue Li, Tiange Luo, Zhiwu Lu, Tao Xiang, and Liwei Wang. Large-scale few-shot learning: Knowledge transfer with class hierarchy. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 7212–7220, 2019. 2

[14] Aoxue Li, Tiange Luo, Tao Xiang, Weiran Huang, and Liwei Wang. Few-shot learning with global class representations. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV*, pages 9714–9723, 2019. 2

[15] Zhenguo Li, Fengwei Zhou, Fei Chen, and Hang Li. Meta-sgd: Learning to learn quickly for few shot learning. *arXiv preprint arXiv:1707.09833*, 2017. 1

[16] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *European Conference on Computer Vision, ECCV*, pages 740–755, 2014. 5, 6

[17] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems,NIPS*, pages 91–99, 2015. 4

[18] Eli Schwartz, Leonid Karlinsky, Joseph Shtok, Sivan Harary, Mattias Marder, Abhishek Kumar, Rogério Schmidt Feris, Raja Giryes, and Alexander M. Bronstein. Delta-encoder: an effective sample synthesis method for few-shot object recognition. In *Advances in Neural Information Processing Systems,NeurIPS*, pages 2850–2860, 2018. 3

[19] Oriol Vinyals, Charles Blundell, Tim Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems, NIPS*, pages 3630–3638, 2016. 2

[20] Xin Wang, Thomas E. Huang, Trevor Darrell, Joseph E. Gonzalez, and Fisher Yu. Frustratingly simple few-shot object detection. In *International Conference on Machine Learning,ICML*, pages 9919–9928, 2020. 5, 6, 7, 8

[21] Yu-Xiong Wang, Ross Girshick, Martial Hebert, and Bharath Hariharan. Low-shot learning from imaginary data. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 7229–7238, 2018. 3

[22] Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. Meta-learning to detect rare objects. In *IEEE Conference on Computer Vision, ICCV*, pages 9925–9934, 2019. 2, 3, 4, 5

[23] Yang Xiao and Renaud Marlet. Few-shot object detection and viewpoint estimation for objects in the wild. In *European conference on computer vision, ECCV*, pages 192–210, 2020. 1, 2, 3, 4, 5, 6, 7, 8

[24] 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 *IEEE/CVF International Conference on Computer Vision*, pages 9576–9585, 2019. 1, 2, 3, 5, 6, 7, 8

[25] Sung Whan Yoon, Jun Seo, and Jaekyun Moon. Tapnet: Neural network augmented with task-adaptive projection for few-shot learning. In *International Conference on Machine Learning, ICML*, pages 7115–7123, 2019. 2

[26] Hongguang Zhang, Jing Zhang, and Piotr Koniusz. Few-shot learning via saliency-guided hallucination of samples. In *IEEE Conference on Computer Vision and Pattern Recognition,CVPR*, pages 2770–2779, 2019. 3