DiGeo: Discriminative Geometry-Aware Learning for Generalized Few-Shot Object Detection

Jiawei Ma Yulei Niu* Jincheng Xu Shiyuan Huang Guangxing Han Shih-Fu Chang
Columbia University
{jiawei.m, yulei.niu, jx2467, shiyuan.h, gh2561, sc250}@columbia.edu

Abstract

Generalized few-shot object detection aims to achieve precise detection on both base classes with abundant annotations and novel classes with limited training data. Existing approaches enhance few-shot generalization with the sacrifice of base-class performance, or maintain high precision in base-class detection with limited improvement in novel-class adaptation. In this paper, we point out the reason is insufficient Discriminative feature learning for all of the classes. As such, we propose a new training framework, DiGeo, to learn Geometry-aware features of inter-class separation and intra-class compactness. To guide the separation of feature clusters, we derive an offline simplex equiangular tight frame (ETF) classifier whose weights serve as class centers and are maximally and equally separated. To tighten the cluster for each class, we include adaptive class-specific margins into the classification loss and encourage the features close to the class centers. Experimental studies on two few-shot benchmark datasets (VOC, COCO) and one long-tail dataset (LVIS) demonstrate that, with a single model, our method can effectively improve generalization on novel classes without hurting the detection of base classes. Our code can be found here.

1. Introduction

Recent years have witnessed the tremendous growth of object detection through deep neural models and large-scale training [2, 13–15, 40, 42, 45, 60, 65]. However, the success of detection models heavily relies on the amount and quality of annotations, which requires expensive annotation cost and time. In addition, traditional object detection models perform worse on the classes with a limited number of annotations [11, 52, 56], while human are able to learn from few observations. In order to close the gap between human vision system and detection models, recent studies have investigated how to generalize well on rare classes under

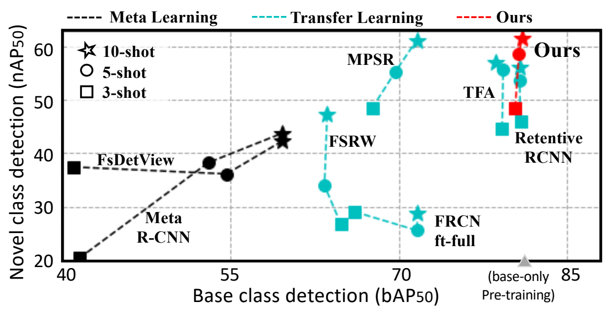

Figure 1. Performance on few-shot object detection on Pascal VOC [3]. Previous transfer-learning approaches (blue) balancing the training data by aggressively down-sampling the base set and may result in overfitting. Instead, we (red) use the full train set, aiming to both maintain precise base detection but learn discriminative features from the limited annotations for few-shot classes.

the few-shot object detection (FSOD) setting. Specifically, given many-shot (*base*) classes with plenty of training data and few-shot (*novel*) classes with extremely limited training data (*e.g*., 5 annotated instances per class), FSOD expects the model to detect the objects in the novel classes well.

To improve the generalization ability on novel-class detection, recent studies [6, 44, 52] conduct transfer learning in a two-step manner. In detail, the model is pre-trained on the whole set of base classes, and then fine-tuned on the union of the set of novel classes and an aggressively down-sampled base subset. However, the efficient few-shot adaptation is often achieved at the expense of sacrificing precision on base detection (Fig. 1). Being aware of this limitation, Fan *et al*. [6] proposed to evaluate the performance of both base and novel classes in the generalized few-shot object detection (GFSOD) setting. In addition, they proposed a consistency regularization to emphasize the pre-trained base knowledge during fine-tuning and employed an ensembling strategy. However, they design different classifiers for base and novel classes, and the adaptation on novel classes is impeded due to a complex ensembling process.

In this paper, we pointed out that the devil is in in-

*Corresponding Author.

sufficient discriminative feature learning for few-shot object detection, including inefficient knowledge adaptation to novel classes and unexpected knowledge forgetting of base classes. First, as the novel instances are extremely limited during training, it is hard to capture the representative visual information of novel classes and adapt the knowledge learned from base classes to novel classes. As a result, the model cannot distinguish between the novel classes, which weakens the few-shot adaptation. Secondly, balanced training strategies such as down-sampling fail to utilize the diverse training samples from base set. Thus, it is hard to preserve the complete knowledge of base classes, which leads to overfitting and further decreases the detection scores.

To tackle these challenges, we proposed a new training framework, **DiGeo**, to make the best of both worlds for generalized few-shot object detection, *i.e*., improving generalization on novel classes without hurting the detection of base classes. Our motivation is to learn **Di**scriminative **Geo**metry-aware features via *inter-class separation* and *intra-class compactness*. For inter-class separation, we expect the class centers [53] to be well distinct from each other. Motivated by the symmetric geometry of simplex equiangular tight frame (ETF) [36], we proposed to use ETF as classifier to guide the separation of features. To be specific, we derive an offline ETF whose weights are maximally & equivalently separated (*i.e*., independent from the training data distribution) and are assigned as fixed centers for all classes. For intra-class compactness, we expect the features to be closed to the class centers for a clear decision boundary. In practice, we add class-specific margins to output logits during training to push the features close to the class centers. The margins are based on instance distribution prior and are then adaptively adjusted though self-distillation. Meanwhile, we consider the huge imbalance between *base* set and *novel* set, and up-sample the *novel* set to facilitate the feature extraction.

We validate the effectiveness of DiGeo under the GFSOD setting on Pascal VOC [3, 4] and MS COCO [27]. Compared to existing methods, we can both achieve precise detection on base classes and sufficiently improve the adaptation efficiency on novel classes using a single model. Furthermore, our DiGeo can be intuitively extended to long-tailed object detection. Experimental results on LVIS datasets demonstrate the generalizibility of our approach. Our contributions are summarized as follows:

- We revisit few-shot object detection from a perspective of discriminative feature learning, and point out that existing methods fail in knowledge adaptation to novel classes and suffer from knowledge forgetting of base classes.
- We propose DiGeo to pursue an desired feature geometry, *i.e*., inter-class separation and intra-class compactness, which consistently improves the performance on both base and novel classes.
- We conduct extensive experiments on three benchmark datasets for few-shot object detection and long-tailed object detection to verify the generalizability of DiGeo.

2. Related Work

Few-shot object detection (FSOD) aims to detect objects of few-shot (*novel*) classes at instance-level. To improve the adaptation efficiency, the approaches based on the meta-learning and the transfer-learning are investigated. The *meta-learning approaches [5, 9–12, 18]* learns a class-agnostic meta-learner to align instances of the same class from different images. Under the Faster-RCNN framework, the attention-based meta-RPN [5] and meta-detector [10] are proposed to generate class-relevant proposals and improve the instance alignment. In addition, approaches based on Transformer [11] and YoLo [18] are proposed to extract features jointly and align features at multiple scales. The *transfer-learning approaches [30, 44, 52, 54]* performs finetuning for few-shot adaptation. Specifically, TFA [52] pre-trains an base detector from plenty of *base* samples and finetune it for novel classes. To improve the adaptation efficiency, multi-scale feature extraction [54] and regularization such as contrastive loss [44], margin equilibrium [21] and transformation invariance [20] are employed. Recently, DeFRCN [39] adjusts gradients back-propagated from different losses and achieve superior novel detection scores.

Generalized Few-Shot Object Detection. For all FSOD approaches mentioned above, the precision on *base* detection is sacrificed after few-shot adaptation. This phenomenon has also been observed in various vision tasks where models forget the base knowledge due to domain gap or distribution gap [33–35, 38, 48, 63, 64]. As pointed out by Fan *et al*. [6], different from the classification [7, 17, 24, 29, 31, 43, 49, 59, 62], an image may contain instances from both novel and base classes and base detection is also important. Then, they propose a consistency regularization few-shot fine-tuning and employ an model ensembling technique to preserve the precision of base detection. However, the few-shot adaptation efficiency is inevitably limited. In a more general case, *long-tail object detection* (LTOD) has been studied where techniques such as resampling [41, 61], decoupling [23, 50] and reweighting [22, 61] are studied. Also, ACSL [51] revisits LTOD from a statistic-free perspective and propose the adaptive suppression loss.

Feature Distribution on a Balanced Set has been studied in classification. To be specific, the weights in the last linear layer is treated as class centers where the geometry property of feature output by pernuminate layer is analyzed. Recently, Papyan *et al*. [36] summarized it as neural collapse (NC) and observed that 1) the features in the same class are maximally concentrated towards the class mean and different feature clusters are maximally separated [57]. 2) The class means and the class centers converge to each other.

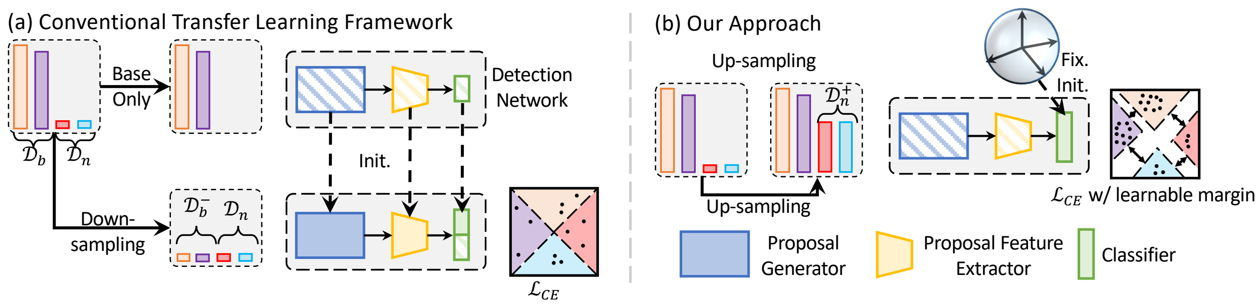

Figure 2. Comparison of training frameworks. (a) Conventional approaches first pre-train a base detector among the *base* set $\mathcal{D}_b$ and then finetune on the union of novel set $\mathcal{D}_n$ and down-sampled subset of base classes $\mathcal{D}_b^-$. (b) Instead, we choose to up-sample $\mathcal{D}_n$ and directly train the detector on the full set. We derive a fixed classifier offline with maximally & equally separated weights and learn the adaptive margins to tighten the feature clusters. The margins are estimated from priors of instance distribution and learned though self-distillation. The block with shading means training from scratch. We use the same design for localization and omit it for simplicity.

3. Background

We first introduce the few-shot object detection (FSOD) task, and analyze the limitations of existing FSOD methods.

3.1. Few-shot Object Detection

In this paper, we focus on the task of few-shot object detection (FSOD). The training data consists of a *base* set $\mathcal{D}_b$ and a *novel* set $\mathcal{D}_n$, where the *base* classes $\mathcal{C}_b$ have plenty of annotated object instances while *novel* classes $\mathcal{C}_n$ has limited annotations. In an N_n-way K-shot FSOD task with $|\mathcal{C}_n| = N_n$, each novel class has K annotated instances. Note that an image may contain multiple instances from different classes with associated bounding boxes, which is more challenging than the few-shot classification where each image contains one object to be recognized. Then, we follow [6, 52] to validate the robustness of detection model under the generalized few-shot object detection (GFSOD) setting, where the test samples come from both base and novel classes, and the models are evaluated on all classes.

Commonly, object detection models consist of a proposal generation module to generate a set of region candidates, and a detection module to localize & classify objects on the extracted proposals [2, 40, 42, 45, 65]. For the classification part, an additional background class should be considered to recognize the proposal with no foreground objects. We regard the last linear layer as classifier, and its weights $W = \{\mathbf{w}_i\}_{i=1}^{N_b+N_n+1}$ as class centers where $N_b = |\mathcal{C}_b|$. Without loss of generality, we set $W_b = \{\mathbf{w}_i\}_{i=1}^{N_b}$, $W_n = \{\mathbf{w}_i\}_{i=N_b+1}^{N_b+N_n}$, and $W_- = \{\mathbf{w}_{N+1}\}$ as weights for base classes $\mathcal{C}_b$, novel classes $\mathcal{C}_n$, and background c_-.

3.2. Analysis of Existing Methods

As a representative *transfer-learning* approach shown in Fig. 2(a), TFA [52] first trains a *Base* detector on $\mathcal{D}_b$ for $\mathcal{C}_b$ as initialization. Then, in an N_n-way K-shot GFSOD task, K instances for each *base* class $c \in \mathcal{C}_b$ from $\mathcal{D}_b$ are selected to a subset $\mathcal{D}_b^-$. The detector is fine-tuned on $\mathcal{D}_b^- \cup \mathcal{D}_n$ with balanced training data distribution over $\mathcal{C}_b \cup \mathcal{C}_n$. However, for each class, as the training data is extremely limited, overfitting to $\mathcal{D}_b^-$ is unignorable and results in the drop of base detection. As such, Retentive RCNN [6] proposes to ensemble the detector adapted for $\mathcal{C}_b \cup \mathcal{C}_n$ and the *Base* detector by combining their outputs as final prediction. However, the novel detection performance on $\mathcal{C}_n$ is limited.

Nevertheless, training among $\mathcal{D}_b \cup \mathcal{D}_n$ makes the model favor $\mathcal{C}_b$. As shown in Fig. 3(a), the novel weights W_n are not well-learned and close to weights of other foreground classes. With such a classifier, the proposal features (*i.e.*, input feature of classifier) cannot be separated. Thus, as shown in Fig. 2(b), we obtain a classifier offline with well-separated weights. For each class, the features are trained to be compact and close to the centers using learnable margins.

4. Approach

Considering the limitations mentioned above, we aim to achieve the best of both worlds using a single model, *i.e.*, improve the few-shot adaptation performance on novel classes without hurting the precision on base detection. Our motivation is to enhance the discriminative feature learning of detection models, *i.e.*, clear boundaries on the feature space to discriminate all classes. We realize this idea from two aspects, inter-class separation between all classes and intra-class compactness for each class.

4.1. Inter-Class Separation

We realize inter-class separation by maximizing the pairwise distances between class centers. Specifically, for each $\mathbf{w}_i$, we maximize its minimum distance with all other weights $W \setminus \{\mathbf{w}_i\}$:

$$W^* = \text{argmax}_W \sum_{i=1}^{N_c} \min_{j, i \neq j} \|\mathbf{w}_i - \mathbf{w}_j\|_2^2 \quad (1)$$
$$s.t.\ \|\mathbf{w}_i\| = 1, \ \ \forall \mathbf{w}_i \in W$$

where $N_c = N_b + N_n + 1$ and all weight vectors are of the same norm (*e.g.*, 1). When the feature dimension $d \geq N_c - 1$, the distances of all class center pairs in W^* should be the same. Also, the angle between any two of the class centers has the same value given $\|\mathbf{w}_i\| = 1$. In this way, we expect the class centers to be evenly distributed in the feature space. In this case, W^* is equivalent to simplex equiangular tight frame (ETF) [36]. Furthermore, we have the following theorem for ETF.

Theorem *Suppose the vector space is d-dimensional and the number of vectors is N. When $d \geq N - 1$, we can always derive a simplex ETF whose vectors are maximally and equally separated from each other.*

The above theorem guarantees the existence of ETF in application when $d \geq N - 1$. For $d < N_c - 1$, *e.g.*, the number of classes is large while the feature dimension is compact, we can project the d-dim feature to a d'-space space with $d' \geq N_c - 1$. Then, we can always obtain a Simplex ETF classifier in the mapped feature space.

We have two options to obtain the Simplex ETF classifier. The *online* solution is to use Eq. (1) as a regularization loss to learn the classifier during training. The *offline* solution is to manually set the Simplex ETF classifier for all classes $\mathcal{C} = \mathcal{C}_b \cup \mathcal{C}_n \cup \{c_-\}$ and fix it during training. We experimentally find the offline solution is more stable and better than the online solution (details discussed in Sec. 5.4) and thus use the offline solution in implementation.

4.2. Intra-Class Compactness

We realize intra-class compactness by tightening the clusters of features and push the samples close to the assigned center in W^*. The challenges are two folds. First, the number of training samples in base and novel classes extremely are imbalanced, which makes it hard to determine the boundaries of novel classes in the feature space. Second, as the number of novel classes is much smaller than that of base classes, *i.e.*, $|\mathcal{D}_n| \ll |\mathcal{D}_b|$, the network receives less positive gradients for novel classes [51], which makes the features of instances in novel classes farther to the class centers and thus less discriminative.

Inspired by the success of logit adjustment in long-tailed recognition [32], we apply class-specific margins on logits to modify the classification loss and balance the optimization between base and novel classes. Specifically, we calculate the class-specific margins based on the frequencies of instance (*i.e.*, bounding box annotations) as priors:

$$m_c = \begin{cases} -\log(p_c) & \text{, if } c \in \mathcal{C}_b \cup \mathcal{C}_n \\ -\log(p_-) & \text{, if } c = c_- \end{cases}, \tag{2}$$

where p_c is the frequency of bounding box annotations for class c, and p_- is an estimated probability of background boxes to train the classifier, and $p_- + \sum_{c \in \mathcal{C}_b \cup \mathcal{C}_n} p_c = 1$. Intuitively, the class with fewer data is assigned with a larger margin to guarantee the learning of this class.

Suppose that the logit outputs for sample x are $\mathbf{v} = \{v_c\}_{c \in \mathcal{C}}$, we use the following prior-margin cross-entropy loss by adding the margins to the logits:

$$\mathcal{L}_{\text{prior}}(x) = -\sum_{c \in \mathcal{C}} y_c \cdot \log \frac{\exp(v_c - m_c)}{\sum_{c' \in \mathcal{C}} \exp(v_{c'} - m_{c'})}. \tag{3}$$

where y_c equals to 1 if c is the ground-truth label, otherwise $y_c = 0$. Note that our prior-margin loss reduce to vanilla cross-entropy loss if all margins m_c are set as 0. As the margins are obtained based on prior distribution and fixed during training, we term this baseline as Prior.

Though the margin-based loss is calculated over all the proposals, precisely calculating the margins from the proposals is time-consuming. Thus, we obtain the prior margins over all annotated bounding box instances. In this case, there is a misalignment between proposal-based loss and instance-based margin. To mitigate this gap, we proposed to adaptively learn the margins based on the priors. Motivated by the success of self distillation [49] in knowledge transfer, we use the detection module learned from $\mathcal{L}_{\text{prior}}$ in Eq. (3) as teacher model, and distill its knowledge to a student model to adaptively learn and update the margins through soft labels, which has the same architecture as teacher model but different parameters. For sample x, the ground-truth label is y, the adaptive-margin distillation objective is:

$$\mathcal{L}_{\text{adapt}}(x) = -\sum_{c \in \mathcal{C}} \frac{p_c^t + y_c}{2} \log \frac{\exp(v_c^s - m_c^s)}{\sum_{c' \in \mathcal{C}} \exp(v_{c'}^s - m_{c'}^s)}, \tag{4}$$

where the predicted probability for class c of the teacher model p_c^t is obtained by $\frac{\exp(v_c - m_c)}{\sum_{c' \in \mathcal{C}} \exp(v_{c'} - m_{c'})}$, v_c^s denotes the logit output for class c of the student model, and m_c^s denotes the adaptive learnable margin for class c. The teacher model is fixed during self distillation, and the student detection head uses the same ETF classifier weights W^* with other parts in the detection module to be learned. Finally, we use the student model for evaluation.

Even though the margins are added during training, the extreme imbalance between base set and novel set still makes the detector favors more on base set. Considering this limitation and the challenge that the number of novel classes is very limited to provide the gradients for network updating, we proposed to up-sample the images containing annotations of novel classes ($\mathcal{D}_n^+$). Specifically, we use repeated factor sampling (RFS) [8] and the repeating times is set by a hyper-parameter threshold in RFS. We experimentally found that using up-sampling itself can achieve marginal improvement, but can clearly improve the novel detection precision combined with our approach. This observation demonstrates that the up-sampling strategy works closely with our hypothesis rather than just a trivial trick.

Table 1. Performance comparison of AP_{50} on the PASCAL VOC dataset on all classes $\mathcal{C}_b \cup \mathcal{C}_n$. The **best** and second-best are highlighted.

Approach	split 1					split 2					split 3					Avg
	1	2	3	5	10	1	2	3	5	10	1	2	3	5	10	
Meta-Learning Approaches																
Meta RCNN [56]*	17.5	30.5	36.2	49.3	55.6	19.4	33.2	34.8	44.4	53.9	20.3	3.0	41.2	48.0	55.1	38.0
FSRW [37]	53.5	50.2	55.3	56.0	59.5	55.1	54.2	55.2	57.5	58.9	54.2	53.5	54.7	58.6	57.6	55.6
FsDetView [55]*	36.4	40.3	40.1	50.0	55.3	36.3	43.7	41.6	45.8	54.1	37.0	39.5	40.7	50.7	54.8	44.4
Transfer-Learning Approaches																
TFA w/ fc [52]	69.3	66.9	70.3	73.4	73.2	64.7	66.3	67.7	68.3	68.7	67.8	68.9	70.8	72.3	72.2	69.5
TFA w/ cos [52]	69.7	68.2	70.5	73.4	72.8	65.5	65.0	67.7	68.0	68.6	67.9	68.6	71.0	72.5	72.4	69.5
FRCN-ft-full [56]*	55.4	57.1	56.8	60.1	60.9	50.1	53.7	53.6	55.9	55.5	58.5	59.1	58.7	61.8	60.8	57.2
MPSR [54]	56.8	60.4	62.8	66.1	69.0	53.1	57.6	62.8	64.2	66.3	55.2	59.8	62.7	66.9	67.7	62.1
Retentive R-CNN [6]†	**71.3**	**72.3**	72.1	74.0	74.6	66.8	**68.4**	70.2	70.7	71.5	**69.0**	**70.9**	72.3	73.9	74.1	71.5
DiGeo (Ours)	69.7	70.6	**72.4**	**75.4**	**76.1**	**67.5**	**68.4**	**71.4**	**71.6**	**73.6**	68.6	**70.9**	**72.9**	**74.4**	**75.0**	**71.9**

*: results reported by Retentive R-CNN [6] and TFA [52]. †: Model ensembling.

Table 2. Comparison of nAP_{50} and bAP_{50} on the PASCAL VOC.

Approach	nAP_{50} (Avg. on splits for each shot)					bAP_{50}
	1	2	3	5	10	Avg.
Meta R-CNN [56]*	11.2	15.3	20.5	29.8	37.0	43.1
FSRW [37]	16.6	17.5	25.0	34.9	42.6	65.0
FsDetView [55]*	26.9	20.4	29.9	31.6	37.1	49.5
TFA w/ fc [52]	27.6	30.6	39.8	46.6	48.7	79.6
TFA w/ cos [52]	31.4	32.6	40.5	46.8	48.3	79.3
FRCN-ft-full [56]*	16.1	20.6	28.8	33.4	36.5	67.2
MPSR [54]	**36.2**	**37.2**	44.6	49.1	53.2	68.1
Retentive RCNN [6]†	31.4	37.1	41.4	46.8	48.8	**81.6**
DiGeo (Ours)	31.6	36.1	**45.8**	**51.2**	**55.1**	81.3

*: results are reported by Retentive R-CNN [6] and TFA [52].
†: Model ensembling. Full tables can be found in Supp.

5. Experiment

We mainly conduct experiments on the few-shot object detection (FSOD) benchmark datasets Pascal VOC and MS COCO to validate the effectiveness of our proposed DiGeo. We further apply DiGeo on long-tailed object detection and conduct experiments on LVIS to show its generalizability.

5.1. Datasets & Training Details

Pascal VOC [3, 4] consists of 20 classes where the class split for $\mathcal{C}_b$ and $\mathcal{C}_n$ are 15 and 5 separately. The train set $\mathcal{D}_b \cup \mathcal{D}_n$ are from Pascal VOC 07+12 trainval sets [3, 4] where $\mathcal{D}_n$ is randomly sampled with K in $\{1, 2, 3, 5, 10\}$. Following TFA [52], we conduct experiments on three base-novel class partitions marked as $\{1, 2, 3\}$. In each partition, for fair comparison, we use the same sampled novel instances and report the detection precision for $\mathcal{C}_n$ (nAP_{50}), $\mathcal{C}_b$ (bAP_{50}) and $\mathcal{C}_b \cup \mathcal{C}_n$ (AP_{50}) on Pascal VOC 07 test set [3].
MS COCO [27] is derived from COCO14 [27] consisting of 80 classes where $|\mathcal{C}_b| = 60$, $|\mathcal{C}_n| = 20$ and $\mathcal{C}_n$ are in common with Pascal VOC. The $\mathcal{D}_b$ and $\mathcal{D}_n$ are from train set with $K = \{10, 30\}$. The detection precision of $\mathcal{C}_n$ (nAP), $\mathcal{C}_b$(bAP) and $\mathcal{C}_b \cup \mathcal{C}_n$ (AP) on COCO 14 val set are reported.
LVIS [8] is derived from COCO17 [27] and contains $\sim$0.7M training instances of 1230 classes. The classes are divided into three groups w.r.t. the amount of annotation, rare (1-10), common (11-100), and frequent (>100). Following [52], we report the precision for all classes (AP) and class groups (AP_r, AP_c, and AP_f) on the val set.
Implementation Details. We instanlize our approach on Faster-RCNN [42,52] which employs a region proposal network (RPN) to generate region candidates. For fair comparison, we use ResNet-101 with FPN [25] as backbone to extract image feature maps where the Resnet-101 backbone is initialized by ImageNet [19]-pretrained model. As the outputs of penultimate layer in original classification module are non-negative and does not meet the property of the ETF classifier, we add a linear layer (projector) with the same input and output dimension on top of the penultimate layer. The projector output is then used for classification. For RFS [16], we set the up-sampling threshold as 0.01 for PASCAL VOC and MS COCO and 0.001 for LVIS. During distillation, we share and fix the parameters of ResNet101 and FPN and only learn a new detection head. We follow the setup in TFA [52] baseline such as SGD optimizer [46]. More details can be found in Supp.

5.2. Comparison with FSOD Methods

We show the comparisons between our methods and state-of-the-art few-shot object detection approaches on PASCAL VOC and MSCOCO. We follow previous works to conduct experiments on three data splits with different shots of novel classes. As for the performance AP_{50} over all classes in Table 1, our DiGeo achieves the best performances for 12 out of 15 cases. Compared to the baseline method TFA [52], our DiGeo outperformed TFA consistently in all shots & splits. Compared to the state-of-the-art Retentative RCNN model, our DiGeo achieves better AP_{50} when the number of shots is larger than 2, and obtains com-

Table 3. Performance comparison of MS COCO dataset.

Approach	10-shot								30-shot							
	AP	bAP	nAP	nAP_{50}	nAP_{75}	nAPs	nAPm	nAPl	AP	bAP	nAP	nAP_{50}	nAP_{75}	nAPs	nAPm	nAPl
FRCN-ft-full [56]*	18.1	21.0	9.2	17.0	9.2	3.4	8.3	15.1	18.6	20.6	12.5	23.0	12.0	3.1	12.0	21.0
FRCN-BCE [56]*	29.2	36.8	6.4	-	-	-	-	-	30.2	36.8	10.3	-	-	-	-	-
TFA w/ fc [52]	27.9	33.9	10.0	19.2	9.2	3.9	8.4	16.3	29.3	34.5	13.5	24.9	13.2	5.0	12.6	21.7
TFA w/ cos [52]	28.4	34.6	9.8	18.7	9.0	**4.5**	8.8	15.8	29.9	35.3	13.6	25.0	13.4	**5.9**	12.2	21.3
MPSR [54]	15.3	17.1	9.7	17.9	9.7	3.3	9.2	16.1	17.1	18.1	14.1	25.4	14.2	4.0	12.9	23.0
Meta R-CNN [56]	5.4	5.2	6.1	19.1	6.6	2.3	7.7	14.0	7.8	7.1	9.9	25.3	10.8	2.8	11.6	19.0
FsDetView [55]	6.7	6.4	7.6	-	-	-	-	-	10.0	9.3	12.0	-	-	-	-	-
Retentive R-CNN [6]	**32.1**	**39.2**	**10.5**	**19.5**	9.3	3.9	8.5	16.3	32.9	39.3	13.8	22.9	13.8	4.6	11.9	22.6
DiGeo	32.0	**39.2**	10.3	18.7	**9.9**	**4.5**	**10.0**	**16.8**	**33.1**	**39.4**	**14.2**	**26.2**	**14.8**	5.3	**13.1**	**23.9**

*: results are reported by Retentive R-CNN [6] and TFA [52].

parable performances for extremely few-shot cases. As for the detailed comparisons over novel classes (nAP_{50}) and base classes (bAP_{50}) in Table 2, our DiGeo still consistently outperforms the baseline TFA method for all the settings.

In addition, our DiGeo achieves a better trade-off between base-class performance and novel-class generalization. On the one hand, although MPSR achieved higher performance for extremely few-shot settings (*e.g.*, 36.2 vs. 31.6 for 1-shot) and competitive performances with $\{3, 5, 10\}$-shots, its performance drops by large margins (*e.g.*, 68.1 vs. 81.3 for bAP_{50}). This observation indicates that MPSR improves the few-shot generalization with the sacrifice of base-class knowledge. On the other hand, as Retentive RCNN [6] includes the base detector & RPN through max ensembling at test time, the adapted detector can be trained to specifically detect novel instances, where the bAP_{50} is slightly higher than ours (*i.e.*, 81.6 vs. 81.3). However, its few-shot generalization is not satisfying when the number of shots is larger than 2 (*e.g.*, 6.3 lower than ours with 10 shots). In contrast, our approach only train a single detector and achieves stable and consistent gain. Similarly, for the results on MSCOCO shown in Table 3, our DiGeo outperforms Retentive RCNN for the novel-class metrics including nAP_{75}, nAPs, nAPm and nAPl. These comparison demonstrate that our DiGeo has a strong few-shot generalization ability without base-class knowledge forgetting.

5.3. Analysis of Inter-Class Separation

Revisit the conventional adaptation strategy from the perspective of separation between classes. Recall that existing few-shot object detection methods follow in TFA [52] and employs a two-step strategy, *i.e.*, first pre-train on the base train set $\mathcal{D}_b$ to learn a *Base* detector, and then fine-tune on the union of the downsampled base set and the novel set, *i.e.*, $\mathcal{D}_b^- \cup \mathcal{D}_n$. We take TFA [52] as the baseline and consider the following settings for the second step: (1) full set $\mathcal{D}_b \cup \mathcal{D}_n$, (2) balanced set $\mathcal{D}_b^- \cup \mathcal{D}_n$, (3) only novel set $\mathcal{D}_n$.

As shown in Fig. 3, we visualize the separation of classifier weights based on their pair-wise cosine similarities. By comparing Fig. 3(a-c), fine-tuning among a balanced set is vital to learn the well-separated classifier weights for all classes $\mathcal{C}_b \cup \mathcal{C}_n$. Instead, using the full set would make the novel classes entangled in Fig. 3(a) due to the extremely imbalanced class distribution (*i.e.*, $|\mathcal{D}_n| \ll |\mathcal{D}_b|$). Although only using $\mathcal{D}_n$ can maximally separate the weights of W_n, as no training data of $\mathcal{C}_b$ is seen, the separation between *base* weights in W_b is hurt and each *novel* class center $\mathbf{w}_i \in W_n$ may still close to some *base* weight, *e.g.*, the similarity between classes "cat" and "cow" is relatively high in Fig. 3(c).

As summarized in Table 4, the novel detection fails when the full set is used in fine-tuning ($\text{Row}_{\{1\}}$). The detection precision on both $\mathcal{C}_b \cup \mathcal{C}_n$ is sub-optimal when no annotation of base class is provided ($\text{Row}_{\{3\}}$) and the detector can easily overfit to the small $\mathcal{D}_n$. Then, finetuning on the balanced set ($\text{Row}_{\{2\}}$) can preserve the base knowledge, maximize the few-shot adaptation effect, and achieve the highest score among the three settings. However, such a balanced set has discarded the diverse training samples of $\mathcal{C}_b$ and the performance drop in base detection is inevitable.

Training on the union of whole base set and upsampled novel set. In contrast, we propose to train from $\mathcal{D}_b \cup \mathcal{D}_n^+$ directly. Note that $\mathcal{D}_n^+$ is a duplication of $\mathcal{D}_n$ with same images but more copies. To properly separate the features of different classes, we use the data-independent optimization target in Eq. 1 to derive a ETF classifier weights W^* offline. As mentioned in Sec. 4.1, Eq. 1 can still be used as a regularization loss to supervised the learning of the last linear layer during training (online). However, as shown in Fig. 3(d), it is still hard to get a perfect ETF classifier shown in Fig. 3(e). After all, the update of classifier weights is also impacted by the weight decay regularization and classification loss, and the learning of weights is not stable, in particular, on an extremely imbalanced dataset. As the classifier weights are kept being updated, the optimization direction of each feature cluster is not stable, which then impede the adaptation efficiency. As compared in Table. 5, the performance by online optimization is slightly worse, in particular when $K = 1$. Though the classifier weights are fixed in

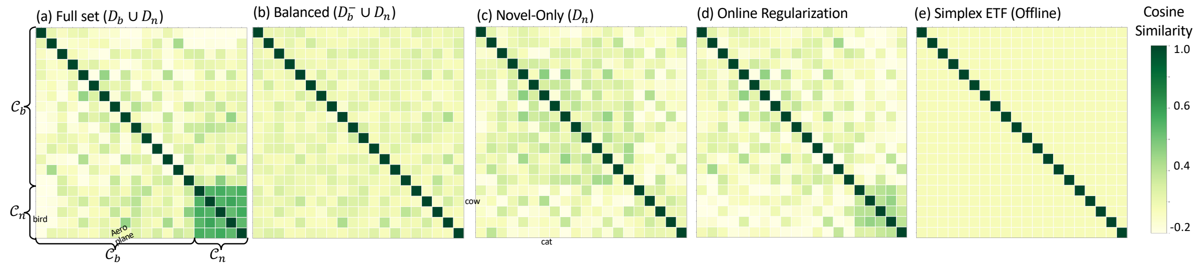

Figure 3. Comparison of pair-wise similarities of classifier weights. We follow the TFA [52] baseline but alter the train set at adaptation step between (a) full set $\mathcal{D}_b \cup \mathcal{D}_n$, (b) a balanced set $\mathcal{D}_b^- \cup \mathcal{D}_n$, or (c) *novel* set $\mathcal{D}_n$. For our approach, to maximally separate the classifier weights, we can (d) apply Eq. (1) as a regularization term during training (online) or (e) derive the classifier weights offline and fix them. The weights for all classes are arranged w.r.t a decreasing order of the number of annotations per class, *i.e.*, from top (left) to bottom (right), from many-shot to few-shot.

Table 4. Comparison between adaptation strategies.

Approach	VOC 5-shot			COCO 10-shot		
	AP_{50}	bAP_{50}	nAP_{50}	AP	bAP	nAP
$\mathcal{D}_b \cup \mathcal{D}_n$	60.6	80.8	0	29.4	39.2	0
$\mathcal{D}_n$	43.6	74.8	44.1	24.1	29.5	7.8
$\mathcal{D}_b^- \cup \mathcal{D}_n$	73.4	79.3	55.7	27.9	33.9	10
Prior	74.9	81.0	56.4	31.5	38.8	9.6
DiGeo	**75.4**	**81.0**	**58.7**	**32.0**	**39.2**	**10.3**

Table 5. Comparison between Offline and Online Classifiers

Approach	1-shot			5-shot		
	AP_{50}	bAP_{50}	nAP_{50}	AP_{50}	bAP_{50}	nAP_{50}
Offline	**68.9**	79.9	**35.8**	**74.9**	**81.0**	**56.4**
Online	68.4	**80.2**	33.1	74.6	80.8	55.9

Table 6. Ablation study of the Prior baseline.

Idx.	ETF	Margin	RFS	AP_{50}	bAP_{50}	nAP_{50}
1				60.6	80.8	0.0
2	✓			61.8	81.2	3.6
3		✓		63.9	81.1	12.2
4			✓	62.8	80.8	9.3
5	✓	✓		69.6	81.0	35.8
6	✓		✓	65.2	**81.2**	17.0
7	✓	✓	✓	**74.9**	81.0	**56.4**

ETF, as the pair-wise angles between weights are the same, we can equivalently assign the weights to all classes $\mathcal{C}$.

Next, as compared in Table 6 Row$_{\{1,3,4\}}$, though adding margins or performing RFS may help with inter-class separation and improve nAP_{50} on $\mathcal{C}_n$, since the weights W_n are still not well-learned due to the extreme imbalance between $\mathcal{D}_b$ and $\mathcal{D}_n$, the performance gain is limited. In contrast, fixing the weights as ETF (Row$_{\{(3,5),(4,6)\}}$) can improve the novel detection, in particular, the nAP_{50} is boosted from 12.2 to 35.8 in Row$_{\{3,5\}}$, which shows that the inter-class separation is essential for distinguishing objects in GFSOD.

Table 7. Comparison of Initialization.

Init.	Method	VOC 5-shot			COCO 10-shot		
		AP_{50}	bAP_{50}	nAP_{50}	AP	bAP	nAP
Base	Prior	74.9	80.9	56.7	31.7	39.0	9.7
Base	DeFRCN [39]	74.1	77.1	65.1	30.1	34.4	17.3
Prior	DeFRCN [39]	74.8	78.0	65.3	30.7	35.1	17.4

Furthermore, being orthogonal to the previous FSOD approaches, our model can be intuitively used as initialization for their adaptation. For the sake of simplicity, we only consider Prior and use a strong baseline DeFRCN [39] for comparison. The PCB calibration [39] is removed to better demonstrate the effect of Prior. As reported in Table 7, though DeFRCN has improved *novel* detection (nAP_{50}) significantly, it still sacrifices the performance on base set. Then, comparing with using *Base* detector as initialization, on both datasets, using our Prior can both help with the adaptation on $\mathcal{C}_n$ and mitigate the drop in $\mathcal{C}_b$ (bAP_{50}). Finally, comparing Table 7 Row$_{\{1\}}$ and Prior in Table 6, adding the step of *Base* detector initialization can only provide marginal improvement for Prior. As Prior has already outperformed TFA, we skip pre-training step for simplicity.

5.4. Analysis of Intra-Class Compactness

Even though the classifier weights have been maximally and equally separated in Fig. 3(e), as the training data is limited, it is still necessary to effectively push the features towards the assigned weight. As compared in Fig. 4, when ETF is used, for each $c \in \mathcal{C}_n$, as $|\mathcal{D}_n| \ll |\mathcal{D}_b|$, the mean of the its features is still distant from the assigned weights. However, our Prior baseline clearly push the features to the assigned weights to facilitate the novel detection. Similarly, in Table 6, only using the ETF classifier can introduce limited gain (Row$_{\{1,2\}}$). Though the ETF classifier with dot-regression loss has been used for long-tail classification [57, 58], we note the efficiency in dealing with hugely imbalanced datasets is limited. By adding margins

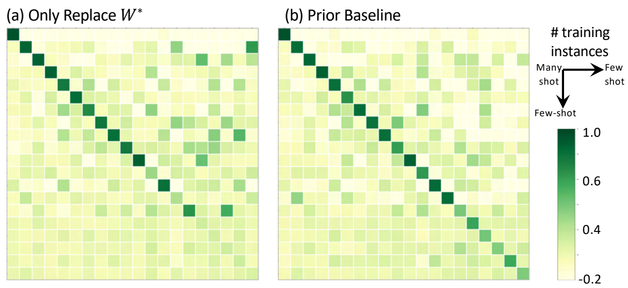

Figure 4. Pair-wise cosine similarity between class means and weights. The class mean is obtained by averaging of features for each class. When the detector is trained on full set $\mathcal{D}_b \cup \mathcal{D}_n$, we first (a) replace the linear classifier with ETF and (b) then use our Prior baseline. Our Prior can push the features close to the assigned weights effectively.

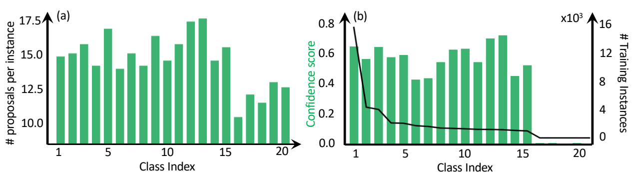

Figure 5. With a detector trained on the $\mathcal{D}_b \cup \mathcal{D}_n$ with vanilla $\mathcal{L}_{CE}$. For each class, we calculate the mean of a) the number of proposals per instance and b) classification confidence score where the number of annotations per class is plotted for reference.

to tighten each cluster and/or up-sampling novel instances in RFS to ensure that sufficient features of $\mathcal{C}_n$ are used for training, the nAP$_{50}$ can then be improved (Row$_{\{2,5,6,7\}}$).

Obtaining effective margins is essential to train on an extremely imbalanced dataset. As discussed in [32], the margins to be added should meet conditions such as *Fisher consistency* [1, 28] to balance the error among different classes. As $|\mathcal{D}_n| \ll |\mathcal{D}_b|$, directly learning the margins for each class individually from scratch (Table 8 Row$_{\{1\}}$) is difficult and may suffer from training instability such as gradient explosion. By sharing margins for classes in the same group, *i.e.*, $\mathcal{C}_b$, $\mathcal{C}_n$, and c_{-}, the nAP$_{50}$ can be improved slightly.

As summarized in Fig. 5, for each class, the number of proposals used to train the detection module ranges from 11 to 17 per instance on average. As such, using the prior of instance distribution $\{p_c\}_{c \in \mathcal{C}_b \cup \mathcal{C}_n}$ can help estimate good margins (Prior). However, as the number of proposals for $\mathcal{C}_n$ (11$\sim$14) is still slightly less than that of $\mathcal{C}_b$ (13$\sim$17) and the margin m_{c_-} for c_- is roughly estimated, it is still necessary to learn margins adaptively. As no stronger prior knowledge can be used, directly learning the margins initialized by $\{-\log(p_c)\}$ does not help clearly (Prior+). However, through self-distillation, the logits output by pre-trained Prior baseline model can be used to indicate the relationship between the proposals features and all class centers, which is then used as supervision signal in our DiGeo.

Table 8. Ablation study of learning effective margins.

Approach	Fixed*	Init.	AP$_{50}$	bAP$_{50}$	nAP$_{50}$
individual		Random	56.4	69.5	17.1
group		Random	60.4	73.2	21.9
Prior	✓	$-\log(p_c)$	74.9	**81.0**	56.4
Prior+		$-\log(p_c)$	**74.8**	80.8	**56.7**

*: The margins are fixed during training.

Table 9. Performance comparison of LVIS.

Approach	ResNet-50				ResNet-101
	AP	APr	APc	Apf	AP
Baseline [8]	22.7	10.6	22.0	28.0	24.5
TFA w/ fc [52]	24.1	14.9	23.9	27.9	-
TFA w/ cos [52]	24.4	16.9	24.3	27.7	-
DiGeo (Ours)	**24.9**	**17.3**	**24.6**	**28.5**	**26.8**
RFS [8]	24.9	14.4	24.5	29.5	-
Focal Loss [26]	22.0	10.5	22.4	25.9	-
EQL [47]	25.1	11.9	26.0	29.1	26.1
BAGS [23]	26.0	17.7	25.8	**29.5**	26.4
ACSL [51]	26.4	18.6	26.4	29.4	27.5
DiGeo (Ours)	**26.7**	**18.9**	**27.0**	29.0	**27.9**

Full table can be found in the Supp.

5.5. Extension to Long-tailed Object Detection

As compared in Table 9, we use TFA [52] and ACSL [51] as two baselines. By employing our design, our DiGeo can achieve higher detection precision on both cases. For comparison with ACSL, we follow the training procedure in ACSL and our approach can benefit from the prior of data distribution to learn discriminative features. More detailed explanation & results can be found in Supp.

6. Conclusion

In this paper, we revisit generalized few-shot object detection from a perspective of discriminative feature learning. We further proposed a simple but effective framework, Discriminative Geometry-aware (DiGeo) learning, for inter-class separation and intra-class compactness. Experiments demonstrates that our DiGeo improves generalization on novel classes without hurting the detection of base classes, and can be extended to long-tail object detection. In the future, we will keep investigating the desired properties of features in object detection and adapted it more realistic scenarios such as domain adaptation.

Acknowledgement This material is based on research sponsored by Air Force Research Laboratory (AFRL) under agreement number FA8750-19-1-1000. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation therein. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of Air Force Laboratory, DARPA or the U.S. Government.

References

[1] Peter L Bartlett, Michael I Jordan, and Jon D McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006. 8

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 1, 3

[3] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html. 1, 2, 5

[4] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html. 2, 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4013–4022, 2020. 2

[6] Zhibo Fan, Yuchen Ma, Zeming Li, and Jian Sun. Generalized few-shot object detection without forgetting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4527–4536, 2021. 1, 2, 3, 5, 6

[7] 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. PMLR, 2017. 2

[8] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5356–5364, 2019. 4, 5, 8

[9] Guangxing Han, Yicheng He, Shiyuan Huang, Jiawei Ma, and Shih-Fu Chang. Query adaptive few-shot object detection with heterogeneous graph convolutional networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3263–3272, October 2021. 2

[10] Guangxing Han, Shiyuan Huang, Jiawei Ma, Yicheng He, and Shih-Fu Chang. Meta faster r-cnn: Towards accurate few-shot object detection with attentive feature alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 780–789, 2022. 2

[11] Guangxing Han, Jiawei Ma, Shiyuan Huang, Long Chen, and Shih-Fu Chang. Few-shot object detection with fully cross-transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5321–5330, June 2022. 1, 2

[12] Guangxing Han, Jiawei Ma, Shiyuan Huang, Long Chen, Rama Chellappa, and Shih-Fu Chang. Multimodal few-shot object detection with meta-learning based cross-modal prompting. *arXiv preprint arXiv:2204.07841*, 2022. 2

[13] Guangxing Han, Xuan Zhang, and Chongrong Li. Revisiting faster r-cnn: A deeper look at region proposal network. In *International Conference on Neural Information Processing*, pages 14–24, 2017. 1

[14] Guangxing Han, Xuan Zhang, and Chongrong Li. Single shot object detection with top-down refinement. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 3360–3364. IEEE, 2017. 1

[15] Guangxing Han, Xuan Zhang, and Chongrong Li. Semi-supervised dff: Decoupling detection and feature flow for video object detectors. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 1811–1819, 2018. 1

[16] Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009. 5

[17] Shiyuan Huang, Jiawei Ma, Guangxing Han, and Shih-Fu Chang. Task-adaptive negative envision for few-shot open-set recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7171–7180, 2022. 2

[18] Bingyi Kang, Zhuang Liu, Xin Wang, Fisher Yu, Jiashi Feng, and Trevor Darrell. Few-shot object detection via feature reweighting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8420–8429, 2019. 2

[19] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017. 5

[20] Aoxue Li and Zhenguo Li. Transformation invariant few-shot object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3094–3102, 2021. 2

[21] Bohao Li, Boyu Yang, Chang Liu, Feng Liu, Rongrong Ji, and Qixiang Ye. Beyond max-margin: Class margin equilibrium for few-shot object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7363–7372, 2021. 2

[22] Bo Li, Yongqiang Yao, Jingru Tan, Gang Zhang, Fengwei Yu, Jianwei Lu, and Ye Luo. Equalized focal loss for dense long-tailed object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6990–6999, 2022. 2

[23] Yu Li, Tao Wang, Bingyi Kang, Sheng Tang, Chunfeng Wang, Jintao Li, and Jiashi Feng. Overcoming classifier imbalance for long-tail object detection with balanced group softmax. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10991–11000, 2020. 2, 8

[24] Han Lin, Guangxing Han, Jiawei Ma, Shiyuan Huang, Xudong Lin, and Shih-Fu Chang. Supervised masked knowledge distillation for few-shot transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[25] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the*

IEEE conference on computer vision and pattern recognition, pages 2117–2125, 2017. 5

[26] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 8

[27] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 2, 5

[28] Yi Lin. A note on margin-based loss functions in classification. *Statistics & probability letters*, 68(1):73–82, 2004. 8

[29] Bin Liu, Yue Cao, Yutong Lin, Qi Li, Zheng Zhang, Mingsheng Long, and Han Hu. Negative margin matters: Understanding margin in few-shot classification. In *European conference on computer vision*, pages 438–455. Springer, 2020. 2

[30] Jiawei Ma, Guangxing Han, Shiyuan Huang, Yuncong Yang, and Shih-Fu Chang. Few-shot end-to-end object detection via constantly concentrated encoding across heads. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVI*, pages 57–73. Springer, 2022. 2

[31] Jiawei Ma, Hanchen Xie, Guangxing Han, Shih-Fu Chang, Aram Galstyan, and Wael Abd-Almageed. Partner-assisted learning for few-shot image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10573–10582, 2021. 2

[32] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *International Conference on Learning Representations*, 2020. 4, 8

[33] Yulei Niu, Long Chen, Chang Zhou, and Hanwang Zhang. Respecting transfer gap in knowledge distillation. *arXiv preprint arXiv:2210.12787*, 2022. 2

[34] Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12700–12710, 2021. 2

[35] Yulei Niu and Hanwang Zhang. Introspective distillation for robust question answering. *Advances in Neural Information Processing Systems*, 34:16292–16304, 2021. 2

[36] Vardan Papyan, XY Han, and David L Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):24652–24663, 2020. 2, 4

[37] Juan-Manuel Perez-Rua, Xiatian Zhu, Timothy M Hospedales, and Tao Xiang. Incremental few-shot object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13846–13855, 2020. 5

[38] Jiaxin Qi, Yulei Niu, Jianqiang Huang, and Hanwang Zhang. Two causal principles for improving visual dialog. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10860–10869, 2020. 2

[39] Limeng Qiao, Yuxuan Zhao, Zhiyuan Li, Xi Qiu, Jianan Wu, and Chi Zhang. Defrcn: Decoupled faster r-cnn for few-shot object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8681–8690, 2021. 2, 7

[40] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 1, 3

[41] Jiawei Ren, Cunjun Yu, Xiao Ma, Haiyu Zhao, Shuai Yi, et al. Balanced meta-softmax for long-tailed visual recognition. *Advances in neural information processing systems*, 33:4175–4186, 2020. 2

[42] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 1, 3, 5

[43] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017. 2

[44] Bo Sun, Banghuai Li, Shengcai Cai, Ye Yuan, and Chi Zhang. Fsce: Few-shot object detection via contrastive proposal encoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7352–7362, 2021. 1, 2

[45] Peize Sun, Rufeng Zhang, Yi Jiang, Tao Kong, Chenfeng Xu, Wei Zhan, Masayoshi Tomizuka, Lei Li, Zehuan Yuan, Changhu Wang, et al. Sparse r-cnn: End-to-end object detection with learnable proposals. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14454–14463, 2021. 1, 3

[46] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pages 1139–1147. PMLR, 2013. 5

[47] Jingru Tan, Changbao Wang, Buyu Li, Quanquan Li, Wanli Ouyang, Changqing Yin, and Junjie Yan. Equalization loss for long-tailed object recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11662–11671, 2020. 8

[48] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3716–3725, 2020. 2

[49] Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B Tenenbaum, and Phillip Isola. Rethinking few-shot image classification: a good embedding is all you need? In *European Conference on Computer Vision*, pages 266–282. Springer, 2020. 2, 4

[50] Tao Wang, Yu Li, Bingyi Kang, Junnan Li, Junhao Liew, Sheng Tang, Steven Hoi, and Jiashi Feng. The devil is in classification: A simple framework for long-tail instance segmentation. In *European conference on computer vision*, pages 728–744. Springer, 2020. 2

[51] Tong Wang, Yousong Zhu, Chaoyang Zhao, Wei Zeng, Jinqiao Wang, and Ming Tang. Adaptive class suppression loss for long-tail object detection. In *Proceedings of*

the IEEE/CVF conference on computer vision and pattern recognition, pages 3103–3112, 2021. 2, 4, 8

[52] Xin Wang, Thomas Huang, Joseph Gonzalez, Trevor Darrell, and Fisher Yu. Frustratingly simple few-shot object detection. In *International Conference on Machine Learning*, pages 9919–9928. PMLR, 2020. 1, 2, 3, 5, 6, 7, 8

[53] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *European conference on computer vision*, pages 499–515. Springer, 2016. 2

[54] Jiaxi Wu, Songtao Liu, Di Huang, and Yunhong Wang. Multi-scale positive sample refinement for few-shot object detection. In *European conference on computer vision*, pages 456–472. Springer, 2020. 2, 5, 6

[55] Yang Xiao and Renaud Marlet. Few-shot object detection and viewpoint estimation for objects in the wild. In *European conference on computer vision*, pages 192–210. Springer, 2020. 5, 6

[56] Xiaopeng Yan, Ziliang Chen, Anni Xu, Xiaoxi Wang, Xiaodan Liang, and Liang Lin. Meta r-cnn: Towards general solver for instance-level low-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9577–9586, 2019. 1, 5, 6

[57] Yibo Yang, Liang Xie, Shixiang Chen, Xiangtai Li, Zhouchen Lin, and Dacheng Tao. Do we really need a learnable classifier at the end of deep neural network? *arXiv preprint arXiv:2203.09081*, 2022. 2, 7

[58] Yibo Yang, Haobo Yuan, Xiangtai Li, Zhouchen Lin, Philip Torr, and Dacheng Tao. Neural collapse inspired feature-classifier alignment for few-shot class incremental learning. *arXiv preprint arXiv:2302.03004*, 2023. 7

[59] Nikolaos-Antonios Ypsilantis, Noa Garcia, Guangxing Han, Sarah Ibrahimi, Nanne Van Noord, and Giorgos Tolias. The met dataset: Instance-level recognition for artworks. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. 2

[60] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022. 1

[61] Songyang Zhang, Zeming Li, Shipeng Yan, Xuming He, and Jian Sun. Distribution alignment: A unified framework for long-tail visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2361–2370, 2021. 2

[62] An Zhao, Mingyu Ding, Zhiwu Lu, Tao Xiang, Yulei Niu, Jiechao Guan, and Ji-Rong Wen. Domain-adaptive few-shot learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1390–1399, 2021. 2

[63] Beier Zhu, Yulei Niu, Yucheng Han, Yue Wu, and Hanwang Zhang. Prompt-aligned gradient for prompt tuning. *arXiv preprint arXiv:2205.14865*, 2022. 2

[64] Beier Zhu, Yulei Niu, Xian-Sheng Hua, and Hanwang Zhang. Cross-domain empirical risk minimization for unbiased long-tailed classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3589–3597, 2022. 2

[65] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 1, 3