FSCE: Few-Shot Object Detection via Contrastive Proposal Encoding

Bo Sun[1], Banghuai Li[2]*, Shengcai Cai[2], Ye Yuan[2], and Chi Zhang[2]
[1]University of Southern California
[2]MEGVII Technology
bos@usc.edu, {libanghuai,caishengcai,yuanye,zhangchi}@megvii.com

Abstract

*Emerging interests have been brought to recognize previously unseen objects given very few training examples, known as few-shot object detection (FSOD). Recent researches demonstrate that good feature embedding is the key to reach favorable few-shot learning performance. We observe object proposals with different Intersection-of-Union (IoU) scores are analogous to the intra-image augmentation used in contrastive visual representation learning. And we exploit this analogy and incorporate supervised contrastive learning to achieve more robust objects representations in FSOD. We present **F**ew-**S**hot object detection via **C**ontrastive proposals **E**ncoding (**FSCE**), a simple yet effective approach to learning contrastive-aware object proposal encodings that facilitate the classification of detected objects. We notice the degradation of average precision (AP) for rare objects mainly comes from misclassifying novel instances as confusable classes. And we ease the misclassification issues by promoting instance level intra-class compactness and inter-class variance via our contrastive proposal encoding loss (CPE loss). Our design outperforms current state-of-the-art works in any shot and all data splits, with up to* $+8.8\%$ *on standard benchmark PASCAL VOC and* $+2.7\%$ *on challenging COCO benchmark. Code is available at: https://github.com/MegviiDetection/FSCE.*

1. Introduction

Development of modern convolutional neural networks (CNNs) [1, 2, 3] give rise to great advances in general object detection [4, 5, 6]. Deep detectors demand a large amount of annotated training data to saturate its performance [7, 8]. In few-shot learning scenarios, deep detectors suffer severer over-fitting and the gap between few-shot detection and general object detection is larger than the corresponding gap in few-shot image classification [9, 10, 11]. On the contrary, a child can rapidly comprehend new visual concepts and recognize objects from a newly learned category given very few examples. Closing such gap is therefore an important step towards more successful machine perception [12].

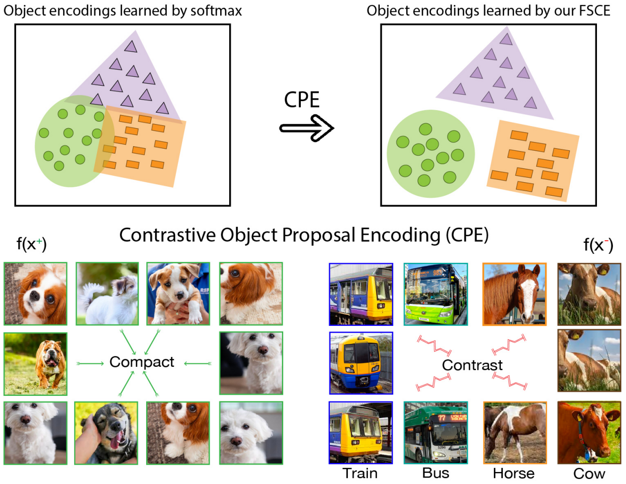

Object encodings learned by softmax
Object encodings learned by our FSCE
CPE
Contrastive Object Proposal Encoding (CPE)
f(x⁺)
f(x⁻)
Compact
Contrast
Train
Bus
Horse
Cow

Figure 1. Conceptualization of our contrastive object proposals encoding. We introduce a score function which measures the semantic similarity between region proposals. Positive proposals (x^+) refer to region proposals from the same category or the same object. Negative proposals (x^-) refer to proposals from different categories. We encourage the object encodings to have the property that $score(f(x), f(x^+)) >> score(f(x), f(x^-))$, such that our contrastively learned object proposals have smaller intra-class variance and larger inter-class difference

Precedented by few-shot image classification, earlier attempts in few-shot object detection utilize meta-learning strategy [13, 14, 15]. Meta-learners are trained with an episode of individual tasks, meta-task samples from common objects (base class) to pair with rare objects (novel class) to simulate few-shot detection tasks. Recently, the two-stage fine-tune based approach (TFA) reveals more potential in improving few-shot detection. Baseline TFA [16] simply freeze all base class trained parameters and fine-tune

*Corresponding author: libanghuai@megvii.com

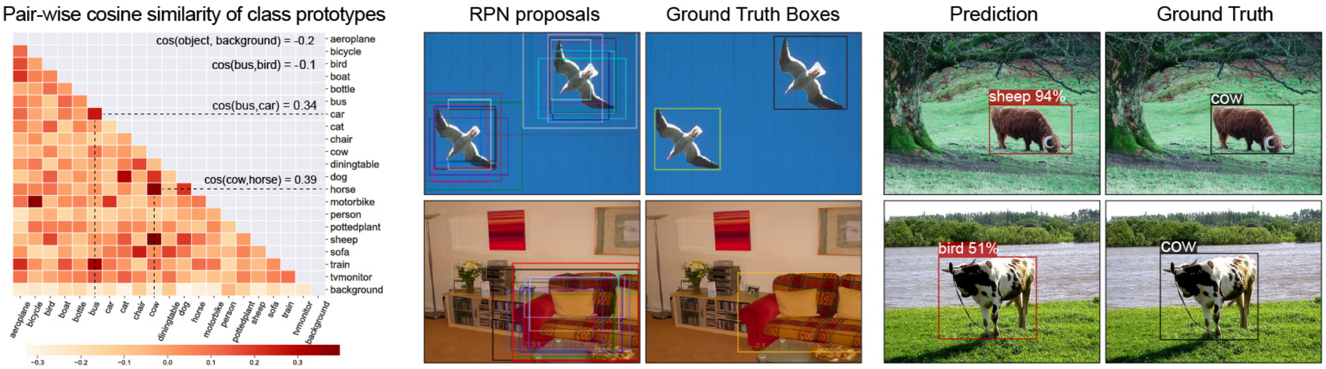


Figure 2. We find in fine-tuning based few-shot object detector, classification is more error-prone than localization. In the fine-tuning stage, RPN is able to make good enough foreground proposals for novel instances, hence novel objects are often accurately localized but mis-classified as confusable base classes. Here shows 20 top-scoring RPN proposals and example detection results from PASCAL VOC Split 1, wherein *bird*, *sofa* and *cow* are novel categories. The left panel shows the pair-wise cosine similarity between the class prototypes learned in the bounding box classifier. For example, the similarity between *bus* and *bird* is -0.10, but the similarity between *cow* and *horse* is 0.39. Our goal is to decrease the instance-level similarity between **similar** objects that are from **different** categories.

only box classifier and box regressor with novel data, yet outperforms previous meta-learners. MPSR [17] improves upon TFA by alleviating the scale bias inherent to few-shot dataset, but their positive refinement branch demands manual selection, which is somewhat less neat. In this work, we observe and address the essential weakness of the fine-tuning based approach – constantly mislabeling novel instances as confusable categories, and improve the few-shot detection performance to the new state-of-the-art (SOTA).

Object detection involves localization and classification of appeared objects. In few-shot detection, one might naturally conjecture the localization of novel objects is going to under-perform its base categories counterpart, with the concern that rare objects would be deemed as background [14, 13, 18]. However, based on our experiments with Faster R-CNN [4], the commonly adopted detector in few-shot detection, class-agonistic region proposal network (RPN) is able to make foreground proposals for novel instances, and the final box regressor can localize novel instances quite accurately. In comparison, as demonstrated in Figure 2, misclassifying detected novel instances as confusable base classes is indeed the main source of error. We visualize the pairwise cosine similarity between class prototypes [19, 20, 21] of a Faster R-CNN box classifier trained with PASCAL VOC [22, 23]. The cosine similarity between prototypes from resembled categories can be 0.39, whereas the similarity between objects and background is on average -0.21. In few-shot setting, the similarity between cluster centers can go as high as 0.59, *e.g.*, between *sheep* and *cow*, *bicycle* and *motorbike*, making classification for similar objects error-prone. We make a calculation upon baseline TFA, manually correcting misclassified yet accurately localized box predictions can increase novel class average precision (nAP) by over 20 points.

A common approach to learn well-separated decision boundary is to use a large margin classifier [24], but with our trials, category-level positive-margin based classifiers does not work in this data-hunger setting [20, 25]. To learn instance-level discriminative feature representations, contrastive learning [26, 27] has demonstrated its effectiveness in tasks including recognition [28], identification [29] and the recent successful self-supervised models [30, 31, 32, 33]. In supervised contrastive learning for image classification [34], intra-image augmentations of images from the same class are used to enrich the positive example pairs. We think region proposals with different Intersection-over-Union (IoU) for an object are naturally analogous to the intra-image augmentation *cropping*, as illustrated in Figure 1. Therefore in this work, we explore to extend the supervised batch contrastive approach [34] to few-shot object detection. We believe the contrastively learned object representations aware of the intra-class compactness and the inter-class difference can ease the misclassification of unseen objects as similar categories.

We present **F**ew-**S**hot object detection via **C**ontrastive proposals **E**ncoding (**FSCE**), a simple yet effective fine-tune based approach for few-shot object detection. When transfer the base detector to few-shot novel data, we augment the primary Region-of-Interest (RoI) head with a contrastive branch, the contrastive branch measures the similarity between object proposal encodings. A supervised contrastive objective with specific considerations for detection will be optimized to reduce the variance of object proposal embeddings from the same category, while pushing different-category instances away from each other. The proposed contrastive objective, contrastive proposal encoding (CPE) loss, is employed to the original classification and localization objective in a multi-task fashion. The end-to-end training of our proposed method is identical to vanilla Faster R-CNN.

To our best knowledge, we are the first to bring contrastive learning into few-shot object detection. Our simple design sets the new state-of-the-art in any shot (1, 2, 3, 5, 10, and 30), with up to $+8.8\%$ on the standard PASCAL VOC benchmark and $+2.7\%$ on the challenging COCO benchmark.

2. Related Work

Few-shot learning. Few-shot learning aims to recognize new concepts given limited labeled examples. Meta-learning approaches aim at training a meta-model on episodes of individual tasks such that it can adapt to new tasks with few samples [35, 11, 36, 10, 37, 38, 39], known as "learning-to-learn". Deep metric-learning based approaches emphasize learning good feature representation embeddings that facilitate downstream tasks. The most intuitive metrics including cosine similarity [20, 40, 41, 21], euclidean distance to class center [19], and graph distances [42]. Interestingly, hallucinator-based methods solve the data deficiency via learning to generate fake-data [9]. Existing few-shot learners are mostly developed in the context of classification. In comparison, few-shot detection is more challenging as it involves both classification and localization, yet under-researched.

Few-shot object detection. There are two lines of work addressing the challenging few-shot object detection (FSOD) problem. First, meta-learning based approaches devise a stage-wise and periodic meta-training paradigm to train a meta-learner to help knowledge transfer from base classes. Meta R-CNN [13] meta-learns channel-wise attention layer for remodeling the RoI head. MetaDet [14] applies a weight prediction meta-model to dynamically transfer category-specific parameters from the base detector. FSIW [15] improves upon Meta R-CNN and FSRW [43] by more complex feature aggregation and meta-training on a balanced dataset. With the balanced dataset introduced in TFA [16], fine-tune based detectors are rowing over meta-learning based methods in performance, MPSR [17] sets the current state-of-the-art by mitigating the scale scarcity in few-shot datasets, but its generalizability is limited because the positive refinement branch contains manual decisions. RepMet [44] attaches an embedding sub-net in RoI head to model a posterior class distribution. It utilizes advanced tricks including OHEM [45] and SoftNMS [46] but fails to catch up with current SOTA. We criticize complex algorithms as they can easily overfit and exhibit poor test results in FSOD. Instead, our insight here is that the degeneration of average precision (AP) for novel categories mainly comes from misclassifying novel instances as confusable categories, and we resort to contrastive learning to learn discriminative object proposal representations without complexing the model.

Contrastive learning The recent success of self-supervised models can be attributed to the renewed interest in exploring contrastive learning. [47, 30, 48, 49, 32, 50, 33, 51]. Optimizing the contrastive objectives [48, 20, 21, 34] simultaneously maximize the agreement between similar instances defined as positive pairs and encourage the difference among dissimilar instances or negative pairs. With contrastive learning, the algorithm learns to build representations that do not concentrate on pixel-level details, but encoding high-level features effective enough to distinguish different images [33, 32, 50, 51]. Supervised contrastive learning [34] extends the batch contrastive approach to supervised setting, but for image classification.

To our best knowledge, this work is the first to integrate supervised contrastive learning [29, 34] into few-shot object detection. The state-of-the-art few-shot detection performance in any shot and all benchmarks demonstrate the effectiveness of our proposed method.

3. Method

Our proposed method FSCE involves a simple two-stage training. First, the standard Faster R-CNN detection model is trained with abundant base-class data ($D_{train} = D_{base}$). Then, the base detector is transferred to novel data through fine-tuning on a balanced dataset [8] with novel instances and randomly sampled base instances ($D_{train} = D_{novel} \cup D_{base}$). The backbone feature extractor is frozen during fine-tuning while the RoI feature extractor is supervised by a contrastive objective. We jointly optimize the contrastive proposal encoding (CPE) loss we proposed with the original classification and regression objectives in a multi-task fashion. Overview of our method is shown in Figure 3.

3.1. Preliminary

Rethink the two-stage fine-tuning approach. Original TFA [16] only fine-tunes the last two fc layers–box classifier and box regressor–with novel data, the rest structures

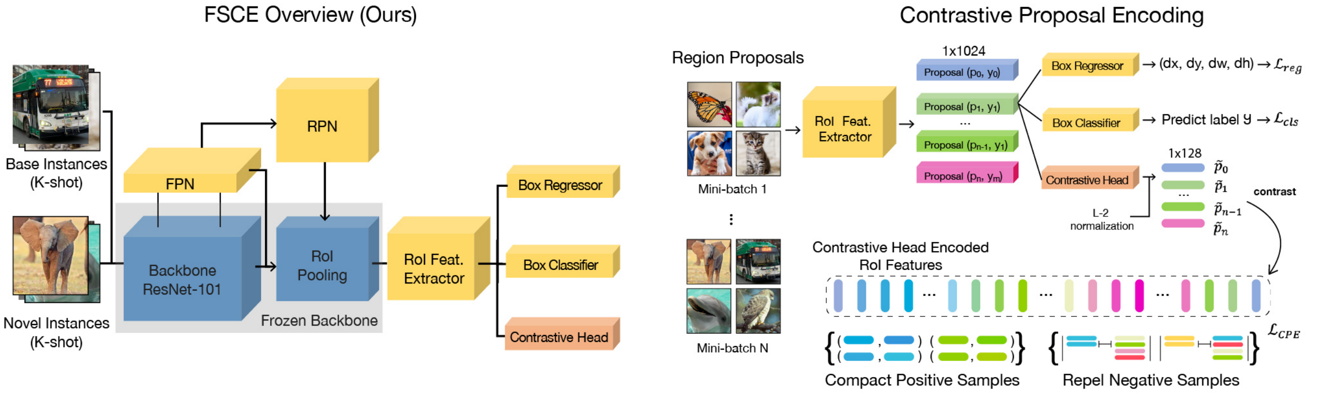


Figure 3. Overview of our proposed FSCE. In our method, we jointly fine-tune the FPN pathway and RPN while fixing the backbone. We find this is effective in coordinating backbone feature maps to activate on novel objects yet still avoid the risk of overfitting. To learn contrastive object proposal encodings, we introduce a contrastive branch to guide the RoI features to learn contrastive-aware proposal embeddings. We design a contrastive objective to maximize the within-category agreement and cross-category disagreement.

are frozen and taken as a fixed feature extractor. This could be viewed as an approach to counter the over-fitting of limited novel data. However it is counter-intuitive that Feature Pyramid Network (FPN [52]), RPN, especially the RoI feature extractor which contain semantic information learned from base classes only, could be transferred directly to novel classes without any form of training. In baseline TFA, unfreezing RPN and RoI feature extractor leads to degraded results for novel classes. However, we find this behavior is reversible and can benefit novel detection results if trained properly. We propose a stronger baseline which adapts much better to novel data with jointly fine-tuned feature extractors and box predictors

Strong baseline. We establish our strong baseline from the following observations. Initially, the detection performance for novel classes decreases as more network components are fine-tuned with novel shots. However, we notice a significant gap in the key RPN and RoI statistics between the data-abundant base training stage and the novel fine-tuning stage. As shown in Figure 4, the number proposals from positive anchors in novel fine-tuning is only $\frac{1}{4}$ of its base training counterpart and the number of foreground proposals decreases consequently. We observe, especially at the beginning of fine-tuning, the positive anchors for novel objects receive comparatively low scores from RPN. Due to the low objectness scores, less positive anchors can pass non-max suppression (NMS) and become proposals that provide actual learning opportunities in RoI head for novel objects. Our insight is to rescue the low objectness positive anchors that are suppressed. Besides, re-balancing the foreground proposals fraction is also critical to prevent the diffusive yet easy backgrounds from dominating the gradient descent for novel instances in fine-tuning.

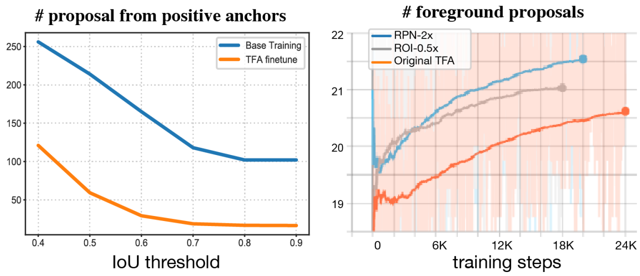


Figure 4. Key detection statistics. Left shows the average number of positive anchors per image in RPN in base training and novel fine-tuning stage. Right shows the average number of foreground proposals per image during fine-tuning. In the left, orange line shows the original TFA setting, which use the same specs as base training. In the right, the blue line shows double the number of anchors kept after NMS in RPN, the gray line shows reducing RoI head batch size by half.

Method	Fine-tune	Refinement		Novel AP50	
	FPN	RPN	ROI	3	10
TFA w/ cos [16]	-	-	-	44.7	56.0
Strong baseline (Ours)	✓	✗	✗	45.3	57.1
	✓	✓	✗	47.2	59.8
	✓	✓	✓	49.7	61.4

Table 1. Novel detection performance of our strong baseline on PASCAL VOC Novel Split 1.

We use unfrozen RPN and ROI with two modifications, (1) double the maximum number of proposals kept after NMS, this brings more foreground proposals for novel instances, and (2) halving the number of sampled proposals in RoI head used for loss computation, as in fine-tuning stage the discarded half contains only backgrounds (standard RoI batch size is 512, and the number of foreground proposals are far less than half of it). As shown in Table 1, our strong baseline boosts the baseline TFA by non-trivial margins. Moreover, the tunable RoI feature extractor opens up room for realizing our proposed contrastive object proposal encoding.

3.2. Contrastive object proposal encoding

In two-stage detection frameworks, RPN takes backbone feature maps as inputs and generates region proposals, RoI head then classifies each region proposal and regresses a bounding box if it is predicted to contain an object. In Faster R-CNN pipeline, RoI head feature extractor first pools the region proposals to fixed size and then encodes them as vector embeddings $x \in \mathbb{R}^{D_R}$ known as the RoI features. Typically $D_R = 1024$ in Faster R-CNN w/ FPN. General detectors fail to establish robust feature representations for region proposals from limited shots, resulting in mislabeling localized objects and low average precision. The idea is to learn more discriminative object proposal embeddings, but according to our experiments, the category-level positive-margin classifier [20, 25] does not work in this data-hungry setting. In order to learn more robust object feature representations from fewer shots, we propose to apply batch contrastive learning [34] to explicitly model instance-level intra-class similarity and inter-class distinction [29, 26] of object proposal embeddings.

To incorporate contrastive representation learning into the Faster R-CNN framework, we introduce a contrastive branch to the primary RoI head, parallel to the classification and regression branches. The RoI feature vector x contains post-ReLU [53] activations thus is truncated at zero, so the similarity between two proposals embeddings can not be measured directly. Therefore, the contrastive branch applies a 1-layer multi-layer-perceptron (MLP) head with negligible cost to encode the RoI feature to contrastive feature $z \in \mathbb{R}^{D_C}$, by default $D_C = 128$. Subsequently, we measure similarity scores between object proposal representations on the MLP-head encoded RoI features and optimize a contrastive objective to maximize the agreement between object proposals from the same category and promote the distinctiveness of proposals from different categories. The proposed contrastive loss for object detection is described in the next section.

We adopt a cosine similarity based bounding box classifier, where the logit to predict i-th instance as j-th class is computed by the scaled cosine similarity between the RoI feature x_i and the class weight w_j in the hypersphere,

$$logit_{\{i,j\}} = \alpha \frac{x_i^\top w_j}{||x_i|| \cdot ||w_j||} \tag{1}$$

α is a scaling factor to enlarge the gradient. We empirically fix $\alpha = 20$ in our experiments. The proposed contrastive branch guides the RoI head to learn contrastive-aware object proposal embeddings which ease the discrimination between different categories. In the cosine projected hypersphere, our contrastive object proposal embeddings form tighter clusters with enlarged distances between different clusters, therefore increasing the generalizability of the detection model in the few-shot setting.

3.3. Contrastive Proposal Encoding (CPE) Loss

Inspired by supervised contrastive objectives in classification [34] and identification [29], our CPE loss is defined as follows with considerations tailored for detection. Concretely, for a mini-batch of N RoI box features $\{z_i, u_i, y_i\}_{i=1}^N$, where z_i is contrastive head encoded RoI feature for i-th region proposal, u_i denotes its Intersection-over-Union (IOU) score with matched ground truth bounding box, and y_i denotes the label of the ground truth,

$$\mathcal{L}_{CPE} = \frac{1}{N} \sum_{i=1}^{N} f(u_i) \cdot L_{z_i} \tag{2}$$

$$L_{z_i} = \frac{-1}{N_{y_i} - 1} \sum_{j=1, j \neq i}^{N} \mathbb{I}\{y_i = y_j\} \cdot \log \frac{\exp(\tilde{z}_i \cdot \tilde{z}_j / \tau)}{\sum_{k=1}^{N} \mathbb{I}_{k \neq i} \cdot \exp(\tilde{z}_i \cdot \tilde{z}_k / \tau)} \tag{3}$$

N_{y_i} is the number of proposals with the same label as y_i, and τ is the hyper-parameter temperature as in InfoNCE [48].

In the above formula, $\tilde{z}_i = \frac{z_i}{||z_i||}$ denotes normalized features hence $\tilde{z}_i \cdot \tilde{z}_j$ measures the cosine similarity between the i-th and j-th proposal in the projected hypersphere. The optimization of the above loss function increases the instance-level similarity between object proposals with the same label and spaces proposals with different labels apart in the projection space. As a result, instances from each category will form a tighter cluster, and the margins around the periphery of the clusters are enlarged. The effectiveness of our CPE loss has been confirmed by t-SNE visualization, as shown in Figure 5 (a) and (b).

Proposal consistency control. Unlike image classification where semantic information comes from the entire image, classification signals in detection come from region proposals. We propose to use an IoU threshold to assure the consistency of proposals that are used to be contrasted, with the consideration that low IoU proposals deviate too much from the center of regressed objects, therefore might contain irrelevant semantics. In the formula above, $f(u_i)$ controls the

Method / Shot	Backbone	Novel Split 1					Novel Split 2					Novel Split 3				
		1	2	3	5	10	1	2	3	5	10	1	2	3	5	10
LSTD *AAAI 18* [54]	VGG-16	8.2	1.0	12.4	29.1	38.5	11.4	3.8	5.0	15.7	31.0	12.6	8.5	15.0	27.3	36.3
YOLOv2-ft *ICCV19* [14]	YOLO V2	6.6	10.7	12.5	24.8	38.6	12.5	4.2	11.6	16.1	33.9	13.0	15.9	15.0	32.2	38.4
†FSRW *ICCV 19* [43]		14.8	15.5	26.7	33.9	47.2	15.7	15.3	22.7	30.1	40.5	21.3	25.6	28.4	42.8	45.9
†MetaDet *ICCV 19* [14]		17.1	19.1	28.9	35.0	48.8	18.2	20.6	25.9	30.6	41.5	20.1	22.3	27.9	41.9	42.9
‡RepMet *CVPR 19* [44]	InceptionV3	26.1	32.9	34.4	38.6	41.3	17.2	22.1	23.4	28.3	35.8	27.5	31.1	31.5	34.4	37.2
FRCN-ft *ICCV 19* [14]	FRCN-R101	13.8	19.6	32.8	41.5	45.6	7.9	15.3	26.2	31.6	39.1	9.8	11.3	19.1	35.0	45.1
FRCN+FPN-ft *ICML* 20[16]		8.2	20.3	29.0	40.1	45.5	13.4	20.6	28.6	32.4	38.8	19.6	20.8	28.7	42.2	42.1
†MetaDet *ICCV 19* [14]		18.9	20.6	30.2	36.8	49.6	21.8	23.1	27.8	31.7	43.0	20.6	23.9	29.4	43.9	44.1
†Meta R-CNN *ICCV 19* [13]		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 *ICML 20* [16]	FRCN-R101	36.8	29.1	43.6	55.7	57.0	18.2	29.0	33.4	35.5	39.0	27.7	33.6	42.5	48.7	50.2
TFA w/ cos *ICML 20* [16]		39.8	36.1	44.7	55.7	56.0	23.5	26.9	34.1	35.1	39.1	30.8	34.8	42.8	49.5	49.8
MPSR *ECCV 20* [17]		41.7	-	51.4	55.2	61.8	24.4	-	39.2	39.9	47.8	35.6	-	42.3	48.0	49.7
FSCE (Ours)		**44.2**	**43.8**	**51.4**	**61.9**	**63.4**	**27.3**	**29.5**	**43.5**	**44.2**	**50.2**	**37.2**	**41.9**	**47.5**	**54.6**	**58.5**
TFA w/ cos* *ICML 20* [16]	FRCN-R101	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
†FSIW* *ECCV 20* [15]		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
FSCE* (Ours)		**32.9**	**44.0**	**46.8**	**52.9**	**59.7**	**23.7**	**30.6**	**38.4**	**43.0**	**48.5**	**22.6**	**33.4**	**39.5**	**47.3**	**54.0**

Table 2. Performance evaluation (nAP 50) of existing few-shot detection methods on three PASCAL VOC Novel Split sets. † marks meta-learning based methods. $\star$ represents average over 10 random seeds. ‡ marks methods use N-way K-Shot meta-testing, which is a different evaluation protocol, see in Sec. 4.1.

consistency of proposals, defined with proposal consistency threshold ϕ, and a re-weighting function $g(\cdot)$,

$$f(u_i) = \mathbb{I}\{u_i \geqslant \phi\} \cdot g(u_i) \tag{4}$$

$g(\cdot)$ assigns different weight coefficients for object proposals with different level of IoU scores. We find ϕ=0.7 is a good cut-off such that the contrastive head is trained with most centered object proposals. Ablations regarding ϕ and g are shown in Sec. 4.3.

Training objectives. In the first stage, the base detector is trained with a standard Faster R-CNN loss [4], a binary cross-entropy loss $\mathcal{L}_{rpn}$ to make foreground proposals from anchors, a cross-entropy loss $\mathcal{L}_{cls}$ for bounding box classifier, and a smoothed-$L1$ loss $\mathcal{L}_{reg}$ for box regression deltas. When transfer to novel data in the fine-tuning stage, we find the contrastive loss can be added to the primary Faster R-CNN loss in a multi-task fashion without destabilizing the training,

$$\mathbb{L} = \mathcal{L}_{rpn} + \mathcal{L}_{cls} + \mathcal{L}_{reg} + \lambda \mathcal{L}_{CPE} \tag{5}$$

λ is set to 0.5 to balance the scale of the losses.

4. Experiments

Extensive experiments are performed in both PASCAL VOC [22, 23] and COCO [55] benchmarks. Our FSCE forms an upper envelope for all fine-tuning based methods and memory-inefficient meta-learns with large margins in any shots in all data splits. We strictly follow the consistent few-shot detection data construction and evaluation protocol [43, 16, 17, 15] to ensure fair and direct comparison. In this section, we first describe the few-shot detection settings, then provide complete comparisons of contemporary few-shot detection works on PASCAL VOC and COCO benchmarks, and provide ablation studies.

Method	Year	Novel AP		Novel AP75	
		10	30	10	30
LSTD [54]	AAAI *18*	3.2	6.7	-	-
†FSRW [43]	ICCV *19*	5.6	9.1	4.6	7.6
†MetaDet [14]	ICCV *19*	7.1	11.3	5.9	10.3
†Meta-RCNN [13]	ICCV *19*	8.7	12.4	6.6	10.8
MPSR [17]	ECCV 20	9.8	14.1	9.7	14.2
TFA w/ cos [16]	ICML *20*	10.0	13.7	9.3	13.4
Ours	N/A	**11.9**	**16.4**	**10.5**	**16.2**
TFA w/ cos* [16]	ICML *20*	9.1	12.1	8.8	12.0
†FSIW* [15]	ECCV *20*	**12.5**	14.7	9.8	12.2
Ours*	N/A	11.1	**15.3**	**9.8**	**14.2**

Table 3. Few-shot detection evaluation results on COCO. $\star$ represents average over 10 random seeds. † marks meta-learning based methods.

Implementation Details. For the detection model, we use Faster-RCNN [4] with Resnet-101 [1] and Feature Pyramid Network [52]. All experiments are run on 8 GPUs with standard batch-size 16. The solver is standard SGD with momentum 0.9 and weight decay 1e-4. Naturally, we scale the training steps when training number of shots. Every detail will be open-sourced in a self-contained codebase to facilitate future research.

4.1. Few-shot detection benchmarks

PASCAL VOC. The overall 20 categories in PASCAL VOC are divided into 15 base categories and 5 novel categories. All base category data from PASCAL VOC 07+12 trainval sets are considered available, and K-shot of novel instances are randomly sampled from previously unseen novel classes for $K = 1, 2, 3, 5$ and 10. Following existing works [16, 43, 15], we consider the same three random partitions of base and novel categories and samplings introduced in [43], referred as Novel Split 1, 2, and 3. And we report AP50 for novel predictions (nAP50) on PASCAL VOC 2007 test set. Note, this is different from the N-Way K-shot settings commonly used in meta-learning based methods [44]. The huge variance between different random runs make the N-Way K-shot evaluation protocol unsuitable for few-shot object detection. For methods that provide results over 10 random seeds, we provide the corresponding results to compare with.

MS COCO. Similarly, for the 80 categories in COCO, 20 categories in common with PASCAL VOC are reserved as novel classes, the rest 60 categories are used as base classes. The $K = 10$ and 30 shots detection performance are evaluated on 5K images from COCO 2014 val dataset, COCO-style AP and AP75 for novel categories are reported by convention.

4.2. Few-shot detection results

PASCAL VOC Results. Results for all three random novel splits from PASCAL VOC are shown in Table 2. Our FSCE outperforms all existing works in any shot and all splits. The effectiveness of our method is fully demonstrated. We are the first to achieve >50 nAP50 on split 2 and split 3, with up to +8.8 nAP50 above current SOTA on split 3. At the same time, our contrastive proposal encodings powered FSCE persists the less base forgetting property as in TFA. Demonstrated below in Table 4.

Method	Base AP50			Novel AP50		
	1	3	5	1	3	5
Baseline-FPN [17]	56.9	66.2	67.9	25.5	41.1	49.6
MPSR [17]	59.4	67.8	68.4	41.7	51.4	55.2
TFA w/ cos (Our impl.)	**79.1**	**77.3**	**77.0**	39.8	44.6	55.6
FSCE (Ours)	78.9	74.1	76.6	**44.2**	**51.4**	**61.9**

Table 4. Base forgetting comparisons on PASCAL VOC Split 1. Before fine-tuning, the base AP50 in base training is 80.8.

COCO Results. Few-shot detection results for COCO are shown in Table 3. Our FSCE set new state-of-the-art for all shots, under the same testing protocol and same metrics. Our proposed methods gain +1.7 nAP and +2.7 nAP75 above current SOTA, which is more significant than the gaps between any previous advancements.

4.3. Ablation

Components of our proposed FSCE. First, with our modified training specification for fine-tune stage, the class-agnostic RPN and RoI head can be directly transferred to novel data and incur huge performance gain, this is because we utilize more low-quality RPN proposals that would normally be suppressed by NMS and provide more foregrounds to learn given the limited optimization opportunity in few-shot setting. And the jointly fine-tuned FPN top-down convolution and RoI feature extractor opens up room for better representation learning. Second, our CPE loss guides the RoI feature extractor to establish contrastive-aware objects embeddings, intra-class compactness and inter-class variance ease the classification task and rescue misclassifications. The whole system benefits from the proposal consistency control by employing only high-IoU region proposals that are less deviated from objects center to contrast. All ablation studies are done with PASCAL VOC Novel Split 1 unless otherwise specified.

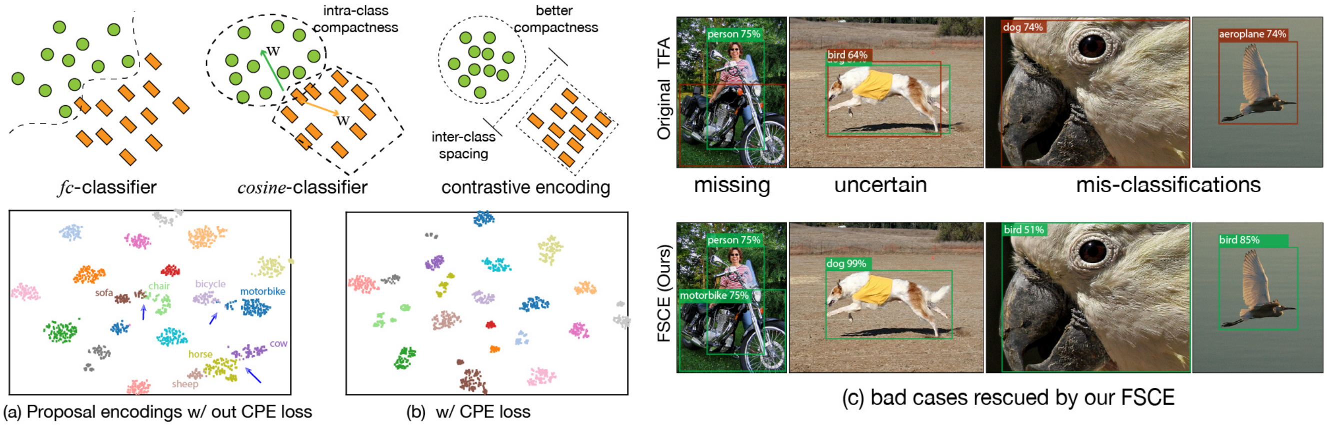
Method	Refinement		CPE loss	Proposal Consistency	Novel AP50		
	RPN	ROI			3	5	10
TFA w/ cos	✗	✗	-	-	44.7	55.7	56.0
FSCE (Ours)	✓	✗	✗	✗	47.2	56.9	59.8
	✓	✓	✗	✗	49.7	58.6	61.4
	✓	✓	✓	✗	50.6	60.7	62.7
	✓	✓	✓	✓	**51.4**	**61.9**	**63.4**

Table 5. Ablation for key components proposed in FSCE.

Ablation for contrastive branch hyper-parameters. Primary RoI feature vector contains post-ReLU activations truncated at zero, we therefore encode the RoI feature with a contrastive head to $z \in \mathbb{R}^{D_C}$ such that similarity can be meaningfully measured. Based on our ablations, the few-shot detection performance is not sensitive to the contrastive head dimension. And among the commonly used temperature τ used in contrastive objectives [34, 32, 50], a medium temperature $\tau = 0.2$ works better than relatively small value 0.07 and large value 0.5.

Contrast Head Dimension	Temperature (τ)		
	0.07	0.2	0.5
$D_C = 128$	63.1	**63.4**	62.9
$D_C = 256$	62.4	**63.4**	63.3

Table 6. Ablation for contrastive hyper-parameters, results from 10 shot of PASCAL VOC Split 1.



(a) Proposal encodings w/ out CPE loss

(b) w/ CPE loss

(c) bad cases rescued by our FSCE

Figure 5. Conceptually and t-SNE visualization of the object proposal embeddings learned with and without our CPE loss, our CPE loss explicitly model the within-class similarity and cross-class distance. t-SNE here shows the proposal encodings from randomly selected 200 PASCAL VOC images. Right panel shows bad cases rescued by our contrastive-aware representations.

Ablation for Proposal Consistency Control. In equation (3) and (4), we propose a compound proposal consistency control mechanism, comprised of an indicator function with an IoU cut-off threshold ϕ, and a function $g(\cdot)$ for re-weighting proposals with different level of IoU. Turns out a re-weighting is not necessary and a simple high-IoU cut-off works the best for 5 and 10 shots, but when number of shots is low, simply filtering out proposals with IoU less than ϕ becomes less favorable as the data sparsity is too severe. In low-shot cases, keeping all proposals but down-weight low-IoU ones make more sense, and empirically, exponential decay (easy mining) does worse than a simple linear weighting.

Option	Threshold	Reweight function	Novel AP		
			3	5	10
Hard Clip	$\phi = 0.5$	$g(x) = 1$	50.5	60.7	62.1
	$\phi = 0.7$	$g(x) = 1$	50.8	**61.9**	**63.4**
Weighting	$\phi = 0$	$g(x) = x$	**51.4**	59.7	61.1
	$\phi = 0$	$g(x) = e^x - 1$	50.8	59.6	61.6

Table 7. Ablation for proposal consistency control in FSCE.

Visual inspections and analysis. Figure 5 shows visual inspections of our proposed FSCE. We find in data-abundant general detection, the saturated performance of fc classifier and $cosine$ classifier are essentially equal. fc layer can learn sophisticated decision boundary from enough data. Existing literature and we all confirm that $cosine$ box classifier excels in few shot object detection, this can be attributed to the explicitly modeled similarity helps form tighter instances clusters on the projected unit hypersphere. The intuition to spacing different categories is trivial, but per our experiments well-established margin-based classifiers [20, 21] does not work in this data-hunger setting (-2 nAP compared to FSCE in 10 shots and worse in lower shots). Instead of adding a margin to classifier, FSCE models the instance-level intra-class similarity and inter-class via CPE loss and guide RoI head to learn contrastive-aware object proposal representations. t-SNE [56] visualization of objects proposal embeddings affirms the effectiveness of our CPE loss in reducing intra-class variance and form more defined decision boundaries, this aligns well with our proposition. Figure 5 (c) shows example bad cases from TFA that are rescued by our FSCE including, missing detection for novel instances, low confidence scores for novel instances, and the pervasive misclassifications.

5. Conclusion

In this work, we propose a new perspective of solving FSOD via contrastive proposals encoding. Effectively saving accurately localized objects from being misclassified, our method achieves state-of-the-art results in any shot and both benchmarks, with up to +8.8% on PASCAL VOC and +2.7% on COCO. Our proposed contrastive proposal encoding head has a negligible cost and is generally applicable. It can be chipped into any two-stage detectors without interfering with the training pipeline. Also, we provide a strong baseline comparable to contemporary SOTA to facilitate future research in FSOD. For a broader impact, FSOD is of great worth considering the vast amount of objects in the real world. Our work proves the plausibility of incorporating contrastive learning into object detection frameworks. We hope our work can inspire more researches in contrastive visual embedding and few-shot object detection.

Acknowledgement. This work was supported by grants from the National Key R & D Program of China. Grant number: 52019YFB1600500.

References

[1] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 1, 6

[2] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1

[3] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1

[4] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28, pages 91–99. Curran Associates, Inc., 2015. 1, 2, 6

[5] T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):318–327, 2020. 1

[6] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 1

[7] Mingxing Tan, Ruoming Pang, and Quoc V. Le. Efficientdet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1

[8] Chien-Yao Wang, Hong-yuan Liao, Yuen-Hua Wu, Ping-Yang Chen, Jun-Wei Hsieh, and I-Hau Yeh. Cspnet: A new backbone that can enhance learning capability of cnn. pages 1571–1580, 06 2020. 1, 3

[9] Y. Wang, R. Girshick, M. Hebert, and B. Hariharan. Low-shot learning from imaginary data. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7278–7286, 2018. 1, 3

[10] Siavash Khodadadeh, Ladislau Boloni, and Mubarak Shah. Unsupervised Meta-Learning for Few-Shot Image Classification. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d\textquotesingle Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 10132–10142. Curran Associates, Inc., 2019. 1, 3

[11] Mengye Ren, Eleni Triantafillou, Sachin Ravi, Jake Snell, Kevin Swersky, Joshua B. Tenenbaum, Hugo Larochelle, and Richard S. Zemel. Meta-learning for semi-supervised few-shot classification. In *Proceedings of 6th International Conference on Learning Representations ICLR*, 2018. 1, 3

[12] Christina M. Funke, Judy Borowski, Karolina Stosio, Wieland Brendel, Thomas S. A. Wallis, and Matthias Bethge. Five Points to Check when Comparing Visual Perception in Humans and Machines. *arXiv:2004.09406 [cs, q-bio, stat]*, October 2020. arXiv: 2004.09406. 1

[13] X. Yan, Z. Chen, A. Xu, X. Wang, X. Liang, and L. Lin. Meta r-cnn: Towards general solver for instance-level low-shot learning. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9576–9585, 2019. 1, 2, 3, 6

[14] Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. Meta-Learning to Detect Rare Objects. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9924–9933, Seoul, Korea (South), October 2019. IEEE. 1, 2, 3, 6

[15] Yang Xiao and Renaud Marlet. Few-shot object detection and viewpoint estimation for objects in the wild. In *European Conference on Computer Vision (ECCV)*, 2020. 1, 3, 6, 7

[16] 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)*, July 2020. 1, 3, 4, 6, 7

[17] 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*, 2020. 2, 3, 6, 7

[18] 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

[19] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, pages 4077–4087. Curran Associates, Inc., 2017. 2, 3

[20] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu. Cosface: Large margin cosine loss for deep face recognition. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5265–5274, 2018. 2, 3, 5, 8

[21] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4685–4694, 2019. 2, 3, 8

[22] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 2, 6

[23] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111(1):98–136, 2015. 2, 6

[24] Gamaleldin F. Elsayed, Dilip Krishnan, Hossein Mobahi, Kevin Regan, and Samy Bengio. Large margin deep networks for classification. In *NeurIPS*, pages 850–860, 2018. 2

[25] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2, 5

[26] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE, 2006. 2, 5

[27] S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 539–546 vol. 1, 2005. 2

[28] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 2

[29] Yi Sun, Yuheng Chen, Xiaogang Wang, and Xiaoou Tang. Deep learning face representation by joint identification-verification. In *Advances in neural information processing systems*, pages 1988–1996, 2014. 2, 3, 5

[30] Zhirong Wu, Yuanjun Xiong, X Yu Stella, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2, 3

[31] Jiahao Xie, Xiaohang Zhan, Ziwei Liu, Yew Soon Ong, and Chen Change Loy. Delving into inter-image invariance for unsupervised visual representations. *arXiv preprint arXiv:2008.11702*, 2020. 2

[32] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2, 3, 7

[33] Chen Ting, Kornblith Simon, Norouzi Mohammad, and Hinton Geoffrey. A simple framework for contrastive learning of visual representations. *ICML 2020*, 2020. 2, 3

[34] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33, 2020. 2, 3, 5, 7

[35] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. volume 70 of *Proceedings of Machine Learning Research*, pages 1126–1135, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR. 3

[36] Qianru Sun, Yaoyao Liu, Tat-Seng Chua, and Bernt Schiele. Meta-transfer learning for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 3

[37] Sachin Ravi and Hugo Larochelle. OPTIMIZATION AS A MODEL FOR FEW-SHOT LEARNING. page 11, 2017. 3

[38] Alex Nichol and John Schulman. Reptile: a scalable metalearning algorithm. *arXiv preprint arXiv:1803.02999*, 2(3):4, 2018. 3

[39] Andrei A. Rusu, Dushyant Rao, Jakub Sygnowski, Oriol Vinyals, Razvan Pascanu, Simon Osindero, and Raia Hadsell. Meta-learning with latent embedding optimization. In *International Conference on Learning Representations*, 2019. 3

[40] Yi Sun, Yuheng Chen, Xiaogang Wang, and Xiaoou Tang. Deep learning face representation by joint identification-verification. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 1988–1996. Curran Associates, Inc., 2014. 3

[41] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Wang, and Jia-Bin Huang. A closer look at few-shot classification. In *International Conference on Learning Representations*, 2019. 3

[42] Victor Garcia Satorras and Joan Bruna Estrach. Few-shot learning with graph neural networks. In *International Conference on Learning Representations*, 2018. 3

[43] B. Kang, Z. Liu, X. Wang, F. Yu, J. Feng, and T. Darrell. Few-shot object detection via feature reweighting. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8419–8428, 2019. 3, 6, 7

[44] L. Karlinsky, J. Shtok, S. Harary, E. Schwartz, A. Aides, R. Feris, R. Giryes, and A. M. Bronstein. Repmet: Representative-based metric learning for classification and few-shot object detection. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5192–5201, 2019. 3, 6, 7

[45] Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 3

[46] N. Bodla, B. Singh, R. Chellappa, and L. S. Davis. Soft-nms — improving object detection with one line of code. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5562–5570, 2017. 3

[47] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *International Conference on Learning Representations*, 2019. 3

[48] Oriol Vinyals Aaron van den Oord, Yazhe Li. Representation learning with contrastive predictive coding. *Advances in Neural Information Processing Systems*, 31, 2018. 3, 5

[49] Ching-Yao Chuang, Joshua Robinson, Lin Yen-Chen, Antonio Torralba, and Stefanie Jegelka. Debiased contrastive learning. In *Advances in neural information processing systems*, 2020. 3

[50] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 3, 7

[51] Kevin Swersky Mohammad Norouzi Geoffrey Hinton Ting Chen, Simon Kornblith. Big self-supervised models are strong semi-supervised learners. *Advances in Neural Information Processing Systems*, 33, 2020. 3

[52] T. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 936–944, 2017. 4, 6

[53] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *ICML*, 2010. 5

[54] Hao Chen, Yali Wang, Guoyou Wang, and Yu Qiao. Lstd: A low-shot transfer detector for object detection. In *AAAI*, 2018. 6

[55] 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. 6

[56] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008. 8