Few-Shot Object Detection via Classification Refinement and Distractor Retreatment

Yiting Li[1] *, Haiyue Zhu[2] † *, Yu Cheng[1] *, Wenxin Wang[1]
Chek Sing Teo[2], Cheng Xiang[1], Prahlad Vadakkepat[1], and Tong Heng Lee[1]
[1] National University of Singapore
[2] SIMTech, Agency for Science, Technology and Research
elelyit@nus.edu.sg, zhu_haiyue@simtech.a-star.edu.sg, chengyu996@gmail.com
wenxin.wang@u.nus.edu, csteo@simtech.a-star.edu.sg, elexc@nus.edu.sg, prahlad@nus.edu.sg, eleleeth@nus.edu.sg

Abstract

We aim to tackle the challenging Few-Shot Object Detection (FSOD), where data-scarce categories are presented during the model learning. The failure modes of Faster-RCNN in FSOD are investigated, and we find that the performance degradation is mainly due to the classification incapability (false positives) caused by category confusion, which motivates us to address FSOD from a novel aspect of classification refinement. Specifically, we address the intrinsic limitation from the aspects of both architectural enhancement and hard-example mining. We introduce a novel few-shot classification refinement mechanism where a decoupled Few-Shot Classification Network (FSCN) is employed to improve the final classification of a base detector. Moreover, we especially probe a commonly-overlooked but destructive issue of FSOD, i.e., the presence of distractor samples due to the incomplete annotations where images from the base set may contain novel-class objects but remain unlabelled. Retreatment solutions are developed to eliminate the incurred false positives. For FSCN training, the distractor is formulated as a semi-supervised problem, where a distractor utilization loss is proposed to make proper use of it for boosting the data-scarce classes, while a confidence-guided dataset pruning (CGDP) technique is developed to facilitate the few-shot adaptation of base detector. Experiments demonstrate that our proposed framework achieves state-of-the-art FSOD performance on public datasets, e.g., Pascal VOC and MS-COCO.

1. Introduction

Deep learning based object detection [13, 4, 2] have achieved remarkable performance outperforming traditional approaches [24, 5]. However, deep learning detection relies on the availability of a large number of training samples. In many practical applications such as robotics [22, 23], labeling a large amount of data is often time-consuming and labor-intensive. This paper focuses on a practically desired but rarely explored area, i.e., Few-Shot Object Detection (FSOD). With the aid of data-abundant base classes, the object detector is trained to additionally detect novel classes through very limited samples. Existing approaches are mainly built on top of Faster-RCNN [13]. For example, the current state-of-the-art approach TFA [17] is presented that employs a classifier rebalancing strategy for registering novel classes. During finetuning, the backbone pre-trained on the base set is reused and being frozen, while only the box classifier and regressor are trained with novel data. Despite its' remarkable progress, its performance on challenging benchmarks such as MS-COCO, is still far away from satisfaction compared with those general data-abundant detection tasks, which deserves more research efforts as data-efficiency is practically preferred in most real-world applications.

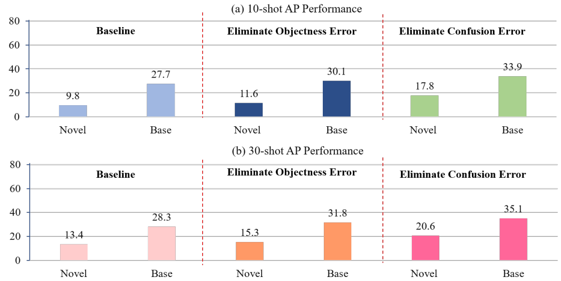

Figure 1. FSOD performance gain by eliminating classification false positives.

To make a step towards the challenging FSOD task, it crucial to find out the major cause of performance degrada-

*∗ indicates equal contribution (Yiting Li, Haiyue Zhu and Yu Cheng).
† indicates corresponding author: Haiyue Zhu.

tion in novel classes. Regarding the architecture of Faster-RCNN, its localization branch is typically class-agnostic with satisfactory performance. Thus our insight is to tackle the limitations of the classification branch for FSOD in this work. Specifically, we evaluate TFA from two important aspects on few-shot classes: 1) IOU awareness, i.e., robustness to hard negatives and 2) category discriminability, i.e., robustness to category confusion. Models that are weak in the first aspect often predict poorly localized hard negatives as “confident” foregrounds of the same category, while those that are weak in the second aspect may suffer from classification confusion between categories that share similar visual features or appear in similar contexts. Next, we analyze the potential performance gain by eliminating these two types of errors separately. For example, given the classification of a poorly localized box (e.g., IOU=0.4) from category “dog”, the effect of the first type false positives (objectness error) can be eliminated by erasing the prediction score for its corresponding semantic category “dog”, while scores for other categories are preserved. To eliminate the second type false positives (confusion error), scores for all other categories except “dog” are erased. Results are shown in Fig. 1. For the 10-shot case, eliminating the objectness error only provides 1.7 points performance gain in mAP, while eliminating the confusion error can dramatically boost the performance gain to 8.0 points, which indicates that classification results of TFA is IOU-aware but less discriminative to confusable categories.

Maintaining IOU-awareness during finetuning is not supervising, as the objectness knowledge gained from a large dataset is usually universal and generalizable, thus can be reliably generalized into unseen novel classes as well. However, lacking inter-class separability often leads to the issue of category confusion. We conjecture possible reasons from the following aspects of architectural limitation: The classification branch of Faster-RCNN based detectors is not purposely designed for few-shot adaptation. For example, the shared feature representation for both classification and localization is shown to be suboptimal for learning category discriminative representations since classification requires translation-invariant features while localization prefers translation-covariant features. Such mismatched learning goals degrade the quality of category-specific translation invariance features [4], thus pose a tough challenge to learn discriminative classifiers when samples of novel classes are scarce.

In this work, we propose a unified approach for addressing the above limitation. Given the fact that TFA is IOU-aware but less semantic discriminative, our key insight is to enhance the original classification results by injecting additional category-discriminative information. In this work, a novel Few-Shot Classification Refinement mechanism is proposed to handle both objectness estimation and category discriminability simultaneously. Our framework consists of two branches, named as the “IOU-aware classification branch” and the “discriminability enhancement branch”,which separately perform their efforts on estimating objectness and alleviating category confusion, respectively. As the name suggests, the IOU-aware classification branch is responsible for producing accurate IOU estimation for each object proposal, which is implemented by the original TFA. At the same time, the enhancement branch is designed as a translation-invariant classifier to produce category-discriminative classification results. After that, outputs of these two complementary branches are aggregated together to produce less confused yet IOU-aware confidence.

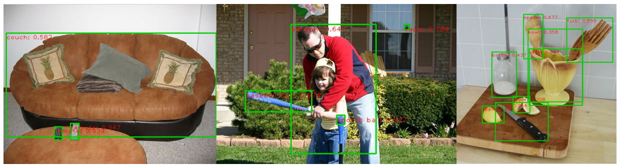

Figure 2. Some samples to show the co-occurrence of both base and novel classes in the same image according to the commonly used dataset setting for MS-COCO, “couch”, “person” and “bottle” are novel classes while the others are base classes. Due to incomplete annotations, those novel-class objects can be unlabelled in base set.

For exhaustively preserving the classification-preferred translation-invariant features, we design the enhancement branch as a decoupled classifier that does not share any parameters with the base detector, where we call it a Few-Shot Correction Network (FSCN). It segments region proposal from image space and provides extra classification refinement to the base detector. Therefore, the classification and localization tasks are decoupled in Faster-RCNN, which naturally solves the issue of shared feature representation. To further improve the semantic discriminability of FSCN, we train it by sampling misclassified false positives from TFA, so as to drive its focus towards the weakness of the base detector and enhance its capacity for eliminating category confusion.

Moreover, we focus on a unique but practically-existed problem of FSOD in this work, i.e., the presence of distractor samples due to the incomplete annotations, where objects belonging to novel classes can exist in the base set but remain unlabelled. As shown in Fig. 2, such a situation is quite realistic for most real-world applications, e.g., in autonomous driving, FSOD is to extend the detection for a novel object “scooter”. However, “scooter” may also exist in the previous images for training the base classes with no annotations, so that such “scooter” distractors will be false emphasized as “background” continuously, which introduces destructive noise. Obviously, completely annotating all novel objects requires to repeatedly review the whole dataset upon the arrival of each novel classes, which is against the motivation of FSOD that dramatically increases the annotation cost especially when the detection tasks are evolving frequently. Hence, “distractor” is defined as those unlabelled novel-class objects in the base set, where proposals corresponding to those unlabelled novel objects are falsely supervised as negative examples. As a result, the positive gradients provided by the few-shot training samples could be easily overwhelmed by the discouraging gradients produced by the distractors during the detector finetuning, so that the resultant detector often inclines to predict novel classes with lower probabilities thus suffers catastrophic performance degradation. To the best of our knowledge, such the distractor phenomenon has not been treated in existing FSOD works without any attention to address it properly.

In this work, we purposely tackle such distractor phenomenon by designing delicate retreatment approaches for both base detector and FSCN correspondingly. For the few-shot adaptation of base detector, a Confidence-Guided Dataset Pruning (CGDP) technique is proposed in this work, which utilizes the self-supervision to exclude the potential distractors to the greatest extent and form a cleaner and balanced training set for few-shot adaptation. Moreover, to sample enough hard examples, the training of FSCN has to be performed on the whole dataset, which exists distractors similarly. However, instead of eliminating the distractors, we specially propose a distractor utilization loss to make proper use of those potential unlabelled novel-class objects in the base set through a semi-supervised manner. In view of the data scarcity of the novel classes, such extra samples help to improve the final detection performance with zero additional annotation cost [15, 14]. Here, we summarize our main contributions as follow:

- We explore the limitations of the classifier rebalancing method (TFA) for FSOD problems and propose a novel few-shot classification refinement framework for exhaustively boosting its FSOD performance. A novel few-shot correction network is developed to achieve great semantic discriminability so as to eliminate false positives caused by category confusion.
- We are the first to address the destructive distractor issue for FSOD. Instead of blindly treating it, a confidence-guided filtering strategy is proposed to exclude the distractors for base detector fine-tuning.
- A semi-supervised distractor utilization strategy is proposed to cooperate with FSCN, which not only stabilizes the training process but also significantly promotes the learning on data-scarce novel classes with no extra annotation cost.
- Our proposed FSOD framework achieves the state-of-the-art results in various datasets with remarkable few-shot performance and knowledge retention ability.

2. Related Works

2.1. Decoupled Classification Refinement for Object Detection

Regarding the misaligned learning goals between the proposal classification and bounding box regression tasks, many effective techniques are proposed to address this issue by introducing various detection refinement strategies. Decoupled Classification Refinement (DCR) [4] proposes to improve detection performance through a decoupled classification correction network, which is the most related work to our research.However, our application is significantly different from DCR. We specially targets the problem of FSOD, which has the additional challenge of localizing novel objects from just a few training samples, unlike the DCR limited to the data-abundant applications. Moreover, we propose the systematic approach to exploit the unique distractor phenomenon of FSOD in a semi-supervised manner for the refinement mechanism. To the best of our knowledge, we are the first to adapt the hard example mining strategy to address the FSOD problem.

2.2. Few-Shot Object Detection (FSOD)

Most of the recent few-shot detection approaches are adapted from the few-shot recognition paradigms. A distillation-based approach is proposed in [3] with less-forgotten constraint and background depression regularization. [7] emphasizes the class-specific information by reweighting top-layer feature maps with channel-wise attentions, so that the obtained features can be used to detect novel object effectively. YOLO-LS [7] and Meta-RCNN [19] propose to emphasize the class-specific feature informative via a meta-learning based channel-attention generator. Metric learning approaches are adopted for the detection classification [8], and [17] proposes a cosine-similarity based Faster-RCNN (TFA) with a category-balance fine-tuning set and achieves the state-of-the-art performance on public datasets. Context-transformer [20] proposes to leverage the rich source-domain knowledge and exploit useful context cues from the target-domain to tackle the challenging object confusion. ONCE [11] proposes a new research area of incremental few-shot object detection, where novel classes are added incrementally without using the samples from base classes. MPSR [18] focus on issue of scale variations caused by annotation scarcity, which generates multi-scale object pyramids to refine the prediction at various scales.

3. Our Approach

3.1. Problem Definition

Our FSOD setting follows the classical formulation [7, 17]. Given a base set $\mathcal{D}_{bs} = \{(\boldsymbol{I}_i^{bs}, \boldsymbol{y}_i^{bs})\}$ with sufficient

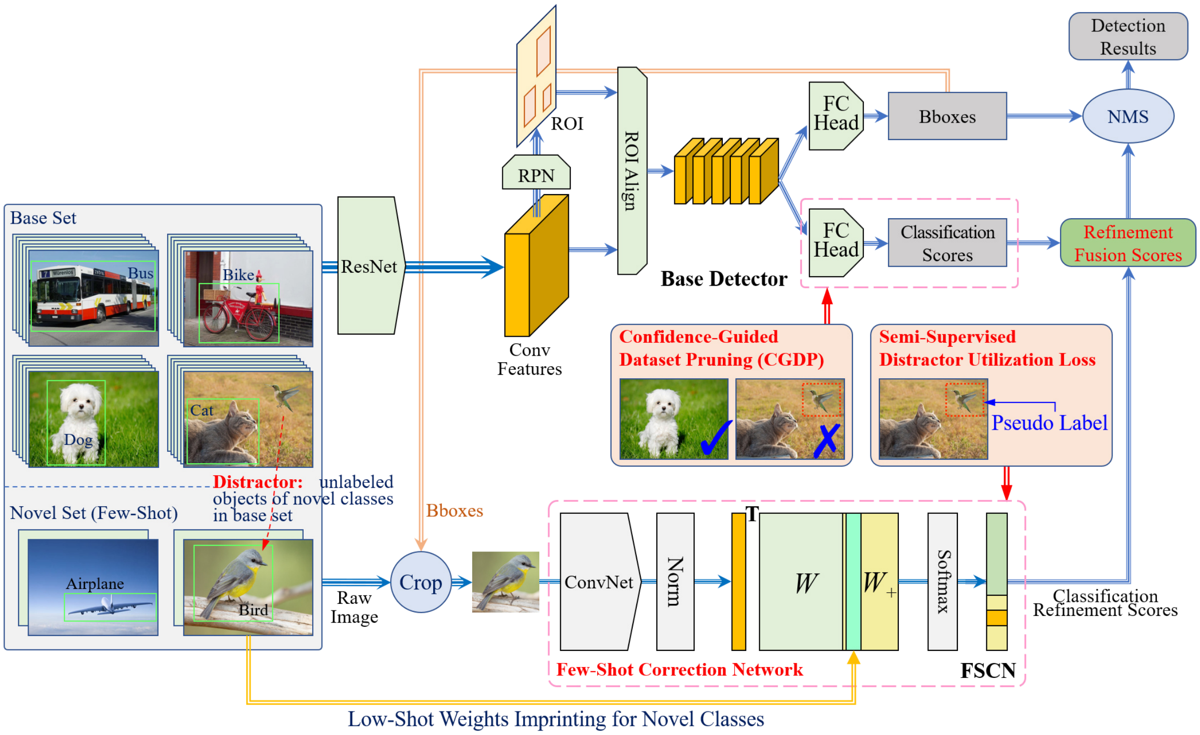
Detection Results
NMS
Bboxes
FC Head
FC Head
Classification Scores
Refinement Fusion Scores
ROI
RPN
ROI Align
Conv Features
ResNet
Base Detector
Base Set
Bus
Bike
Dog
Cat
Distractor: unlabeled objects of novel classes in base set
Novel Set (Few-Shot)
Airplane
Bird
Raw Image
Crop
Bboxes
Confidence-Guided Dataset Pruning (CGDP)
✓
✗
Semi-Supervised Distractor Utilization Loss
Pseudo Label
ConvNet
Norm
T
W
W+
Softmax
Classification Refinement Scores
Few-Shot Correction Network
FSCN
Low-Shot Weights Imprinting for Novel Classes

Figure 3. Illustration of the proposed FSOD framework, where the FSCN provides extra classification refinement to eliminate the false positives. When adapting to new few-shot tasks, separate strategies are proposed for the base detector and FSCN. For the fine-tuning of the base detector, CGDP is proposed to filter out those base-set images that may contain unlabeled novel-class objects, e.g., the “bird” here. In contrast, FSCN requires to train on the whole dataset for sampling enough false positives, thus a semi-supervised distractor utilization loss is proposed to encourage the FSCN to learn from those confident unlabeled distractor proposals to boost the data-scarce classes, instead of falsely treating them as negatives.

annotated samples for each class, where $\boldsymbol{I}_i^{bs} \in \mathcal{I}$ denotes an input image and $\boldsymbol{y}_i^{bs} = \{(c_j^{bs}, \boldsymbol{l}_j)\}_{j=1}^{N_i}$ denotes a list of N_i bounding-box annotations containing box location $\boldsymbol{l}_j$ and category $c_j^{bs} \in \mathcal{C}_{bs}$. $\mathcal{C}_{bs}$ is the space of base categories, $N_{bs} = |\mathcal{C}_{bs}|$ is the category number in $\mathcal{D}_{bs}$. During the initial pre-training phase, an object detector $\boldsymbol{\mathcal{F}}(\cdot|\theta_b)$ is trained on $\mathcal{D}_{bs}$ for detecting objects in $\mathcal{C}_{bs}$ with parameters θ_b. The FSOD task is performed on a k-shot novel set $\mathcal{D}_{nv} = \{(\boldsymbol{I}_i^{nv}, \boldsymbol{y}_i^{nv})\}$ with novel categories $\mathcal{C}_{nv}$, where $\mathcal{C}_{bs} \cap \mathcal{C}_{nv} = \emptyset$ and $|\mathcal{C}_{nv}| = N_{nv}$. The objective of FSOD is to adapt the pre-trained detector parameters from θ_b to θ_* through both sets $\mathcal{D}_{bs} \cup \mathcal{D}_{nv}$, so that $\boldsymbol{\mathcal{F}}(\cdot|\theta_*)$ can effectively detect the objects from all classes in $\mathcal{C}_{bs} \cup \mathcal{C}_{nv}$.

The definition of the distractor phenomenon in FSOD is that some images $\{\boldsymbol{I}_i^{bs}\}$ in $\mathcal{D}_{bs}$ may possibly contain unlabel objects belonging to $\mathcal{C}_{nv}$. According to previous works, those objects are unlabeled in $\mathcal{D}_{bs}$ and will be treated as the background during detector fine-tuning, which introduces dramatic confusion for detector training. However, in real-world scenarios, revisiting the massive $\mathcal{D}_{bs}$ to label out all objects belonging to $\mathcal{C}_{nv}$ is not affordable, and more importantly, conflicts the main purpose of few-shot learning. Therefore, handling the distractor through the delicate algorithm is of great significance to avoid the huge annotation cost and improve the FSOD performance.

3.2. Framework Overview with Few-Shot Classification Refinement

In view of the scarce training samples available for FSOD problem, the learning difficulty is significantly enlarged due to the intrinsic architecture limitation of detector, which often results in less discriminative classifier and leads to category confusion. Essentially, for object detectors, such issue actually comes down to the overwhelming number of misclassified false positives. Motivated by this, we aim to tackle the challenging FSOD problem from the view of hard example mining. Specifically, our framework exploits to alleviate the burden of differentiating false positives by leveraging a powerful few-shot classification refinement mechanism. A decoupled correction network is employed to further refine and enhance the proposal classification, which is trained from the hard false positives sampled from the box regressor of the base detector. Such error-oriented perspective plus the additional architecture-level enhancement also provide a unified way to jointly address the few-shot adaption and category confusion.

The overall architecture of the proposed FSOD framework is shown as in Fig. 3, which consists of two parallel networks, i.e., the base detector $\boldsymbol{\mathcal{F}}_d(\cdot)$ and the FSCN $\boldsymbol{\mathcal{F}}_r(\cdot)$. In this work, $\boldsymbol{\mathcal{F}}_d(\cdot)$ takes Faster-RCNN as a example, the input image is processed by $\boldsymbol{\mathcal{F}}_d(\cdot)$ first to obtain the primary proposal information. The proposed FSCN $\boldsymbol{\mathcal{F}}_r(\cdot)$ takes the proposals of box regressor as inputs, which are cropped from original image according to the proposal location, denoted as $\boldsymbol{I}_p = Cr(\boldsymbol{I}, \boldsymbol{p})$, where $Cr(\cdot)$ denotes the crop function, and $\boldsymbol{I}$ and $\boldsymbol{p}$ denotes input image and proposal boxes predicted by $\boldsymbol{\mathcal{F}}_d(\cdot)$, respectively. Similar as the Faster-RCNN proposal classifier in $\boldsymbol{\mathcal{F}}_d(\cdot)$, FSCN $\boldsymbol{\mathcal{F}}_r(\cdot)$ outputs a classification distribution vector $\boldsymbol{s}_r$ with $N_t + 1$ elements, where $N_t = N_{bs} + N_{nv}$ is the number of all base+novel classes and the additional $+1$ is the background class. Therefore, the proposed FSCN $\boldsymbol{\mathcal{F}}_r(\cdot)$ can be represented as

$$\boldsymbol{s}_r = \boldsymbol{\mathcal{F}}_r(\boldsymbol{I}_p) = \boldsymbol{\mathcal{F}}_r\big(Cr(\boldsymbol{I}, \boldsymbol{p})\big), \tag{1}$$

where $\boldsymbol{s}_r = \{s_r^j\}_{j=1}^{N_t+1}$ is the classification confidence vector for all $N_t + 1$ categories. The key idea is to augment the base detector $\boldsymbol{\mathcal{F}}_d(\cdot)$ with FSCN $\boldsymbol{\mathcal{F}}_r(\cdot)$ in parallel to enhance the proposal classification capability. Since $\boldsymbol{\mathcal{F}}_r(\cdot)$ is trained with false positives sampled from $\boldsymbol{\mathcal{F}}_d(\cdot)$, the proposed FSOD architecture,

$$\boldsymbol{\mathcal{F}}(\cdot) = \boldsymbol{\mathcal{F}}_d(\cdot) \oplus \boldsymbol{\mathcal{F}}_r(\cdot), \tag{2}$$

is endowed with stronger discriminative capability to eliminate the false positives, which is crucial for FSOD performance.

3.3. Few-Shot Correction Network (FSCN)

3.3.1 Network Description

The proposed FSCN $\boldsymbol{\mathcal{F}}_r(\cdot)$ consists of two components: a feature extractor $\boldsymbol{\phi}_\vartheta$ and a linear classifier φ_w. The feature extractor

$$\boldsymbol{z}_p = \boldsymbol{\phi}_\vartheta(\boldsymbol{I}_p|\boldsymbol{\vartheta}), \tag{3}$$

maps a 2D input image $\boldsymbol{I}_p$ to a feature embedding $\boldsymbol{z}_p \in \mathbf{R}^d$, where $\boldsymbol{\vartheta}$ denotes its network parameters. The linear classifier

$$\boldsymbol{s}_r = \varphi_w(\boldsymbol{z}_p|\boldsymbol{w}), \tag{4}$$

calculates the similarities to all classes followed by softmax, where $\boldsymbol{w} = \{\boldsymbol{w}_j\}_{j=1}^{N_t+1}$ and $\boldsymbol{w}_j \in \mathbf{R}^d$. In addition, unlike image classification task where a single large object is in the center of an image, objects in detection tasks may appear from a wide range of scales or appear at an arbitrary position. However, the effective receptive field of traditional CNNs is usually small and spatially biased to the central region. As a result, objects located at the outer area of the receptive field are more likely to be ignored. Hence, a good correction network is required to have a sufficiently large receptive field that can handle such complex appearance of region proposals. In this work, a Compact Generalized Non-Local (CGNL) module [21] is equipped with FSCN to achieve global receptive field.

The key point of few-shot learning is to use a good similarity metric that can be easily generalized to unseen classes. In this work, we introduce cosine similarity metric into FSCN, which can well encourage the unified recognition over all classes. Specifically, we use a zero-bias fully-connection layer in φ_w followed by softmax. Given a proposal image input $\boldsymbol{I}_p \in \mathcal{I}_p$, the classification confidence s_r^j for category j can be calculated as

$$s_r^j = \kappa \frac{\boldsymbol{\phi}_\vartheta(\boldsymbol{I}_p)}{||\boldsymbol{\phi}_\vartheta(\boldsymbol{I}_p)||_2} \cdot \frac{\boldsymbol{w}_j^T}{||\boldsymbol{w}_j||_2}, \tag{5}$$

where $\cdot$ denotes Frobenius inner product and $||\cdot||_2$ denotes the L_2-normalization, κ is a learnable scale parameter used to ensure the convergence of training [16].

3.3.2 Weight Imprinting for Novel Classes

To adapt the FSCN $\boldsymbol{\mathcal{F}}_r(\cdot)$ from base classes to novel classes, we introduce a weight imprinting technique [12] for FSOD to directly initialize its parameters for sequential learning. Consider the $\boldsymbol{\mathcal{F}}_r(\cdot)$ trained from base categories $\mathcal{C}_{bs}$ to be adapted to novel categories $\mathcal{C}_{nv}$, the weights $\boldsymbol{w}$ in φ_w is augmented from $\{\boldsymbol{w}_j\}_{j=1}^{N_{bs}+1}$ to $\{\boldsymbol{w}_j\}_{j=1}^{N_{bs}+N_{nv}+1}$. Hence, for those new-coming classes, an intuitive way to set their weights is to average the corresponding normalized feature vectors $\boldsymbol{z}_p$,

$$\hat{\boldsymbol{z}}_j = \frac{1}{N_j} \sum_{i=1}^{N_j} \frac{\boldsymbol{\phi}_\vartheta(\boldsymbol{I}_{pi})}{||\boldsymbol{\phi}_\vartheta(\boldsymbol{I}_{pi})||_2}, \ \boldsymbol{I}_{pi} \in \mathcal{I}_p^j, \tag{6}$$

where $j = N_{bs} + 1, N_{bs} + 2, \ldots, N_{bs} + N_{nv}$; $\mathcal{I}_p^j$ denotes the j-th class set of foreground region-proposal images extracted from $\mathcal{D}_{nv}$, and $|\mathcal{I}_p^j| = N_j$. The final weights $\boldsymbol{w}_j$ is calculated by normalizing the averaged features, where $\boldsymbol{w}_j = \hat{\boldsymbol{z}}_j / ||\hat{\boldsymbol{z}}_j||_2$. Note that there is one special background class for the detection classification, which is shared among both the base and novel sets. Similarly, its weights are inferred by sampling background region proposals uniformly from $\mathcal{D}_{bs} \cup \mathcal{D}_{nv}$ similar as for those novel classes.

3.4. Semi-Supervised Distractor Utilization Loss

Under the low-data constraint on novel classes, to tackle the above mentioned issue of category confusion, the abundant number of false positives sampled from base set become particularly valuable, especially for those producing high response to novel classes due to the shared visually-similar appearances. However, without complete annotations, those unlabeled objects presented in base set (distractors) are often falsely emphasized as negatives, which is destructive to FSCN. With the commonly-used cross-entropy loss, the encouraging gradients provided from the few-shot training samples are easily suppressed by the discouraging gradients produced from those unlabeled distractors. Inspired by this, we delve into the most fundamental but important issue for training FSCN, i.e., how to avoid blindly learning from distractors and even make proper use of them. To address this, we proposed a semi-supervised distractor utilization loss, which employs the confident unlabeled data to promote the learning of few-shot classes through a semi-supervised manner, thus improves the final detection performance.

We notice that for each background proposal sampled from the base set, there is a probability for it to be an unlabeled foreground of novel class. The idea of the proposed semi-supervised distractor utilization loss is to assign positive gradients to those potential foreground novel classes as well as the original background class, so that the network training is more motivated for those under-represented novel classes. Consequently, the key issue is to determine the possible foreground class for each background proposal sampled from $\mathcal{D}_{bs}$, we formulate it as a semi-supervised learning problem and tackle it with the pseudo hard labeling technique. Specifically, given a background proposal $\boldsymbol{I}_{bp}$ from $\mathcal{D}_{bs}$, its pseudo label can be determined according to the prediction confidence of the current FSCN,

$$\begin{aligned} \boldsymbol{s}_r &= \boldsymbol{\mathcal{F}}_r(\boldsymbol{I}_{bp}) = \{s_r^j\}_{j=1}^{N_{bs}+N_{nv}+1}, \\ C_{pl} &= \arg\max\left(\{s_r^j\}_{j=N_{bs}+1}^{N_{bs}+N_{nv}}\right),\ C_{pl} \in \mathcal{C}_{nv}. \end{aligned} \tag{7}$$

However, if all the background proposals are labeled as positive samples of novel classes, there will be no negative samples for FSCN training, which leads the FSCN to produce a highly biased prediction. Therefore, such predicted pseudo label can not be directly employed to train the network. To address this issue, we further introduce a new concept of background augmentation, which defines a *Augmented Background* class by merging the original background class C_b with the generated pseudo class C_p, denoted as set $\mathcal{C}_b^+$,

$$\mathcal{C}_b^+ = C_b \cup C_p. \tag{8}$$

For example, assuming there are 60 base class in $\mathcal{C}_{bs}$, 20 novel class in $\mathcal{C}_{nv}$ and one background class C_b. For a background proposal that produces high activation to one novel classes "human" (pseudo labeled class C_{pl}), we merge the class "human" C_{pl} and the background class C_b into a new augmented background class $\mathcal{C}_b^+$. Thus the new label space Γ_{bs} have 60 base class $\mathcal{C}_{bs}$, 19 novel class $(\mathcal{C}_{nv} \backslash C_{pl})$ with one augmented background class $\mathcal{C}_b^+$. As a result, the overall prediction score for the Augmented Background class $\mathcal{C}_b^+$ is to aggregate the softmax confidence from both C_b and C_{pl} according to Eq. 8,

$$P(\mathcal{C}_b^+|\boldsymbol{I}_{bp}) = P(C_b|\boldsymbol{I}_{bp}) + P(C_{pl}|\boldsymbol{I}_{bp}). \tag{9}$$

Based on this modification, an improved cross-entropy loss, which we termed as distractor utilization loss, is proposed to not only alleviate the negative influence of distractors but also exploit the distractors to boost the training for those data-scarce novel classes,

$$L_{dul} = \frac{1}{N_{bs} + N_{nv}} \sum_{c \in \Gamma_{bs}} -\log P(c|\boldsymbol{I}_p), \tag{10}$$

where Γ_{bs} is the reformulated category space on $\mathcal{D}_{bs}$, defined as $\Gamma_{bs} = \mathcal{C}_b^+ \cup \mathcal{C}_{bs} \cup (\mathcal{C}_{nv} \backslash C_{pl})$, and $\mathcal{C}_b^+$ is treated as one Augmented Background class in Γ_{bs}. The proposed distractor utilization loss L_{dul} assigns encouraging gradients to the potential corresponding novel classes to boost the few-shot performance when facing distractors. In the meanwhile, the original gradients to background class persist as well in regardless of distractor or not. Note that the distractor utilization loss is only needed to the proposals sampled from $\mathcal{D}_{bs}$, since for $\mathcal{D}_{nv}$, the standard cross-entropy loss is enough as full annotations are available under the common FSOD setting.

However, only a small portion of backgrounds sampled from base set are to be truly unlabeled objects of novel classes, while the major portion are just hard negatives of base classes. In practice, when applying the above L_{dul} loss, if the merging strategy is applied to all backgrounds sampled from base set, each will contribute an encouraging gradient to novel class, and the accumulated gradient is too strong and lead to a biased prediction towards novel classes. To avoid this, we propose an unlabelled object mining (UOM) strategy to automatically select the high-possibility unlabelled objects. Without considering the object occlusion, for a background proposal to be considered as unlabeled objects, it should at least intersect not too much with any known objects. Inspired by this, a spatial metric M_{sp} is developed for performing effective training sample selection, which measures the maximal spatial intersection between the candidate proposal and all known ground-truth objects. Specifically, for a background proposal $\boldsymbol{p_b}$ with location $\boldsymbol{l_b}$ sampled from a base set image, we calculate the spatial metric M_{sp} as,

$$M_{sp}(\boldsymbol{p_b}) = \max_{\text{all } j}\left(\frac{Area(\boldsymbol{l_b}) \cap Area(\boldsymbol{l_j})}{Area(\boldsymbol{l_b})}\right), \tag{11}$$

where $Area(\cdot)$ denotes the box area specified by $\boldsymbol{l}$, and $\boldsymbol{l_j}$ is the annotated box location of j-th base object in the current image. Different from the conventional IOU metric, the proposed intersection ratio represents a normalized measure that focuses on the area of empty volume contained in each region proposal. Thus, L_{dul} loss is only applied on those high-possibility background proposals to be novel classes, which exploits the distractor for few-shot classes learning effectively but avoid the unwanted prediction bias.

3.5. Confidence-Guided Dataset Pruning (CGDP)

While the effective training of FSCN has been addressed in Section 3.4, we focus on the few-shot adaptation of base detector $\mathcal{F}_d(\cdot)$ in this subsection. The motivation for CGDP is to form a small clean subset with less distractors from $\mathcal{D}_{bs}$ to facilitate the base detector few-shot adaptation. Our approach is mainly motivated by the recent pool-based active learning technique[1]. However, unlike the traditional active learning that usually focuses on picking out the most informative unlabeled samples for human annotation, we aim at developing an automatic pipeline by taking the advantage of self-supervision to effectively clean the distractor samples. Basically, our proposed CGDP is a two-stage process which consists of the indicator learning stage and the dataset pruning stage. Here, the indicator is to indicate the possibility to have distractors for a image from $\mathcal{D}_{bs}$.

For the first stage, we look for a simple yet effective way to develop an efficient query function. Specifically, given a base detector $\mathcal{F}_b(\cdot)$ that is pre-trained from $\mathcal{D}_{bs}$, an indicator $\mathcal{F}^c_{ind}(\cdot)$ is the classification branch obtained by fine-tuning $\mathcal{F}_b(\cdot)$ on $\mathcal{D}_{ind}$ using normal cosine-similarity classification loss without considering the issues of distractors. Here, $\mathcal{D}_{ind}$ is a balanced training set made of the whole $\mathcal{D}_{nv}$ and a small portion of $\mathcal{D}_{bs}$, and only the remaining portion of $\mathcal{D}_{bs}$ will be used for dataset pruning in next stage, denoted as $\mathcal{D}_{pru}$, i.e., $\mathcal{D}_{ind} \cup \mathcal{D}_{pru} = \mathcal{D}_{bs} \cup \mathcal{D}_{nv}$. Given an input image $\boldsymbol{I}_{bs}$, the classification confidences of all region proposals are predicted by $\mathcal{F}^c_{ind}(\cdot)$ as,

$$\{\{s_i^j\}_{j=1}^{N_{bs}+N_{nv}+1}\}_{i=1}^{N_p} = \mathcal{F}^c_{ind}(\boldsymbol{I}_{bs}), \tag{12}$$

where s_i^j is the confidence score of j-th novel class for i-th proposal, and N_p is total number of proposals. The proposed query function $Q(\cdot)$ that estimates the likelihood of an image to have distractors is defined as,

$$Q(\boldsymbol{I}_{bs}) = \max_i \max_j \{\{s_i^j\}_{j=N_{bs}+1}^{N_{bs}+N_{nv}}\}_{i=1}^{N_p}. \tag{13}$$

In the second pruning stage, $Q(\cdot)$ is used to select the samples from $\mathcal{D}_{pru}$ in order to form a clean subset $\mathcal{D}_{cln}$. Specially, we construct a class-specific data pool for each category in $\mathcal{C}_{bs}$ by sampling images from $\mathcal{D}_{pru}$. Suppose there are total N_{c_i} images in $\mathcal{D}_{pru}$ that contains the objects of class $c_i \in \mathcal{C}_{bs}$, the data pool for class c_i is constructed as

$$Pool(c_i) = \{\boldsymbol{I}_j^{c_i}, Q(\boldsymbol{I}_j^{c_i})\}_{j=1}^{N_{c_i}},\ \boldsymbol{I}_j^{c_i} \in \mathcal{D}_{pru},\ c_i \in \mathcal{C}_{bs}, \tag{14}$$

where $\boldsymbol{I}_j^{c_i}$ is the j-th image that contains object from class c_i and $Q(\boldsymbol{I}_j^{c_i})$ is its corresponding likelihood of being with distractors. From each data pool, we select its top m samples which have the lowest likelihood in order to form the clean balanced training set $\mathcal{D}_{cln}$. It is also noted that $\mathcal{D}_{cln}$ follows the original distributions of base classes in $\mathcal{D}_{pru}$ as well as $\mathcal{D}_{bs}$. Overall, the proposed CGDP pipeline can be formulated as

$$\left(\mathcal{F}^c_{ind}(\cdot), \mathcal{D}_{ind}, Q(\cdot), \mathcal{D}_{pru}\right) \rightarrow \mathcal{D}_{cln}. \tag{15}$$

Table 1. Evaluation on MS-COCO Novel Set

Shots	Methods	Novel AP	Novel AP50
10	YOLO Low-Shot [7]	5.60	12.3
	Meta-RCNN [19]	8.70	19.1
	ONCE-NL [11]	5.1	-
	MPSR [18]	9.8	17.9
	TFA [17]	10.0	-
	cos-FRCN [17]	9.8	15.3
	cos-FRCN + CGDP	10.6	17.8
	cos-FRCN + FSCN	11.1	18.5
	cos-FRCN + CGDP + FSCN	**11.3**	**20.3**
30	YOLO Low-Shot [7]	9.10	19.0
	Meta-RCNN [19]	12.4	25.3
	ONCE-NL [11]	-	-
	MPSR [18]	14.1	25.4
	TFA [17]	13.7	-
	cos-FRCN [17]	13.4	25.1
	cos-FRCN + CGDP	14.3	27.8
	cos-FRCN + FSCN	14.6	28.5
	cos-FRCN + CGDP + FSCN	**15.1**	**29.4**

4. Experiments

4.1. Implementation Details

We implement the proposed FSCN by using an ImageNet pre-trained ResNet50-CGNL model [21]. Given a mini-batch which contains n_{bs} base-set images and n_{nv} novel-set images, for each image feed into $\mathcal{F}^*_d(\cdot)$, we only reserve its top 300 candidate boxes and divide them into three groups which are the foreground, false positives. Given the classification confidence of a region proposal, if the network's response to one of its negative class is larger than a pre-defined threshold 0.1, we consider this region proposal as a false positive detection. We then sample a total number of m boxes from these three groups uniformly [4]. In our experiments, we set $n_{bs} = 6$, $n_{nv} = 2$ and $m = 32$. Finally, a ROI-Align layer is used to crop the selected boxes from the original image and reshape them into the size of 224×244. The threshold for unlabelled object mining(UOM) is set to be 0.2. Due to space limitation, more training details are included in appendix.

Table 2. Evaluation on Pascal VOC Novel Set

Method/Shots	Novel Split 1					Novel Split 2					Novel Split 3				
	1	2	3	5	10	1	2	3	5	10	1	2	3	5	10
YOLO Low-Shot [7]	14.8	15.5	26.7	33.9	47.2	15.7	15.3	22.7	30.1	39.2	19.2	21.7	25.7	40.6	41.3
Meta-RCNN [19]	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
Context Transformer [20]	34.2	-	-	44.2	-	26.0	-	-	36.3	-	29.3	-	-	40.8	-
MPSR [18]	**41.7**	-	**51.4**	55.2	61.8	24.4	-	39.2	39.9	47.8	**35.6**	-	42.3	48.0	49.7
TFA [17]	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
cos-FRCN [17]	37.1	41.9	43.6	49.1	53.3	23.0	26.7	35.5	36.1	40.6	25.9	31.7	40.6	45.7	49.7
cos-FRCN+CGDP	37.3	41.8	43.9	49.6	53.4	23.2	26.7	36.1	36.3	41.7	26.6	33.0	40.5	45.9	50.6
cos-FRCN+FSCN	40.3	45.0	46.7	56.9	**62.5**	27.1	30.8	40.5	42.1	46.1	30.5	35.1	43.5	49.6	55.3
cos-FRCN+CGDP+FRCN	40.7	**45.1**	46.5	**57.4**	62.4	**27.3**	**31.4**	**40.8**	**42.7**	**46.3**	31.2	**36.4**	**43.7**	**50.1**	**55.6**

4.1.1 Results on MS-COCO

We provide the mAP performance of the novel classes on MS-COCO [10] in Table 1, and compare our approach with the other six baselines, which are YOLO low-shot [7], Meta-RCNN [19] , ONCE no-incremental [11], TFA [17], MPSR [18], and context transformer [20]. Our approach uses ResNet50 [6] as the backbone for the base detector, which is similar to Meta-RCNN. However, we notice that a stronger backbone FPN [9] is employed by TFA. For a fair comparison, we re-implement its proposed cosine similarity classification and balanced fine-tuning strategy with our backbone (ResNet50), which roughly matches the original results in TFA. Here, we denote this re-implemented baseline as "cos-FRCN".

Regarding the results, we have several observations. 1), In all different numbers of training shots, our approach is able to outperform the previous methods by large margins, which achieve the state-of-the-art results. As we can see, it almost doubles the performance of the previous meta-learning approaches (YOLO low-shot) under the 10-shot case, which validates the effectiveness of our approach. 2), As the distractor is a unique and extremely-challenging problem for FSOD, the proposed CGDP can be seen as a simple yet effective solution which brings significant improvement of nearly 1 point on novel classes. 3), The absolute performance gain of FSCN is even larger than CGDP, which indicates that the intrinsic architecture limitations of Faster-RCNN is severer than the issue of distractors.

4.1.2 Results on Pascal VOC

We further present the evaluation results on Pascal VOC as shown in Table. 2. Experiments are conducted under the k-shot setting with three different dataset splits, where $k = 1, 2, 3, 5, 10$. Our approach consistently outperform the existing approaches with significant margin in nearly all different splits/shots, which demonstrates that the effectiveness of the proposed few-shot classification refinement mechanism. It also worth to note that there is no significant performance gain when introducing CGDP into the training of cos-FRCN, which is quite different from the results on MS-COCO. We conjecture this is because Pascal VOC contains much less unlabeled objects than MS-COCO, which makes the problem of distractors less obvious.

5. Conclusions

This paper casts a new viewpoint to address the challenging FSOD problem from both the architecture limitation and destructive distractor phenomenon, where a two-level learning approach is proposed to jointly address the above issues in a unified manner. First, we propose a architecture-level enhancement, where a novel few-shot correction network is introduced to alleviate the burden of category confusion. Second, instead of blindly treating distractor samples, the data-level learning strategies are proposed to separately address the few-shot adaption for both the base detector and FSCN. CGDP effectively excludes the distractors for the base detector adaptation by a confidence-guided filtering strategy, while the semi-supervised distractor utilization loss make use the distractors for boosting the data-scarce classes in FSCN. Remarkably, through fusing the proposed CGDP with the FSCN, we are the first to propose an integrated FSOD framework with excellent few-shot performance and incredible knowledge retention ability.

References

[1] Hamed H Aghdam, Abel Gonzalez-Garcia, Joost van de Weijer, and Antonio M López. Active learning for deep detection neural networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3672–3680, 2019. 7

[2] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. YOLOv4: Optimal speed and accuracy of object detection, 2020. 1

[3] Hao Chen, Yali Wang, Guoyou Wang, and Yu Qiao. LSTD: A low-shot transfer detector for object detection. In *AAAI*, 2018. 3

[4] Bowen Cheng, Yunchao Wei, Honghui Shi, Rogerio Feris, Jinjun Xiong, and Thomas Huang. Revisiting RCNN: On awakening the classification power of Faster RCNN. In *ECCV*, September 2018. 1, 2, 3, 7

[5] Ross Girshick. Fast R-CNN. In *ICCV*, pages 1440–1448, 2015. 1

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 8

[7] Bingyi Kang, Zhuang Liu, Xin Wang, Fisher Yu, Jiashi Feng, and Trevor Darrell. Few-shot object detection via feature reweighting. In *ICCV*, 2019. 3, 7, 8

[8] Leonid Karlinsky, Joseph Shtok, Sivan Harary, Eli Schwartz, Amit Aides, Rogério Schmidt Feris, Raja Giryes, and Alexander M. Bronstein. RepMet: Representative-based metric learning for classification and few-shot object detection. In *CVPR*, 2019. 3

[9] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection, 2017. 8

[10] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. *ArXiv*, 2014. 8

[11] Juan-Manuel Perez-Rua, Xiatian Zhu, Timothy Hospedales, and Tao Xiang. Incremental few-shot object detection, 2020. 3, 7, 8

[12] Hang Qi, Matthew Brown, and David G. Lowe. Low-shot learning with imprinted weights. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5822–5830, 2018. 5

[13] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems 28*, pages 91–99. 2015. 1

[14] Peng Tang, Xinggang Wang, Song Bai, Wei Shen, Xiang Bai, Wenyu Liu, and Alan Yuille. Pcl: Proposal cluster learning for weakly supervised object detection. *IEEE transactions on pattern analysis and machine intelligence*, 42(1):176–191, 2018. 3

[15] Peng Tang, Xinggang Wang, Xiang Bai, and Wenyu Liu. Multiple instance detection network with online instance classifier refinement. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2843–2851, 2017. 3

[16] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition, 2018. 5

[17] Xin Wang, Thomas E. Huang, Trevor Darrell, Joseph E. Gonzalez, and Fisher Yu. Frustratingly Simple Few-Shot Object Detection. *arXiv e-prints*, 2020. 1, 3, 7, 8

[18] Jiaxi Wu, Songtao Liu, Di Huang, and Yunhong Wang. Multi-scale positive sample refinement for few-shot object detection, 2020. 3, 7, 8

[19] Xiaopeng Yan, Ziliang Chen, Anni Xu, Xiaoxi Wang, Xiaodan Liang, and Liang Lin. Meta R-CNN: Towards general solver for instance-level low-shot learning. In *ICCV*, pages 9576–9585, 10 2019. 3, 7, 8

[20] Ze Yang, Yali Wang, Xianyu Chen, Jianzhuang Liu, and Yu Qiao. Context-transformer: Tackling object confusion for few-shot detection, 2020. 3, 8

[21] Kaiyu Yue, Ming Sun, Yuchen Yuan, Feng Zhou, Errui Ding, and Fuxin Xu. Compact generalized non-local network. In *NIPS*, page 6511–6520, 2018. 5, 7

[22] Haiyue Zhu, Xiong Li, Wenjie Chen, Xiaocong Li, Jun Ma, Chek Sing Teo, Tat Joo Teo, and Wei Lin. Weight imprinting classification-based force grasping with a variable-stiffness robotic gripper. *IEEE Transactions on Automation Science and Engineering*, 2021. 1

[23] Haiyue Zhu, Yiting Li, Fengjun Bai, Wenjie Chen, Xiaocong Li, Jun Ma, Chek Sing Teo, Pey Yuen Tao, and Wei Lin. Grasping detection network with uncertainty estimation for confidence-driven semi-supervised domain adaptation. In *IROS*, Oct. 2020. 1

[24] C. Lawrence Zitnick and Piotr Dollár. Edge boxes: Locating object proposals from edges. In *ECCV*, pages 391–405, 2014. 1