LEDetection: A Simple Framework for Semi-Supervised Few-Shot Object Detection

Phi Vu Tran
LexisNexis Risk Solutions

Abstract

Few-shot object detection (FSOD) is a challenging problem aimed at detecting novel concepts from few exemplars. Existing approaches to FSOD all assume abundant base labels to adapt to novel objects. This paper studies the new task of *semi-supervised FSOD* by considering a realistic scenario in which both base and novel labels are simultaneously scarce. We explore the utility of unlabeled data within our proposed label-efficient detection framework and discover its remarkable ability to boost semi-supervised FSOD by way of region proposals. Motivated by this finding, we introduce SoftER Teacher, a robust detector combining pseudo-labeling with consistency learning on region proposals, to harness unlabeled data for improved FSOD without relying on abundant labels. Rigorous experiments show that SoftER Teacher surpasses the novel performance of a strong supervised detector using only 10% of required base labels, without catastrophic forgetting observed in prior approaches. Our work also sheds light on a potential relationship between semi-supervised and few-shot detection suggesting that a stronger semi-supervised detector leads to a more effective few-shot detector.

1 INTRODUCTION

Modern object detection systems have enjoyed tremendous progress in recent years, with many successful applications across diverse industries. Their success can be mainly attributed to the availability of large-scale, well-annotated datasets such as the MS-COCO benchmark (Lin et al., 2014). However, the demand for more powerful detection models requires considerable investments in the hand-labeling of massive amounts of data, which are costly to scale. Thus, there is a growing trend in the community to shift toward a more *label-efficient* paradigm, one that can enable detection models to do more with less hand-labeled data. Such emerging research directions include semi-supervised detection (SSOD) and few-shot detection (FSOD), which have shown great promise in alleviating the dependency on large amounts of instance-level annotations.

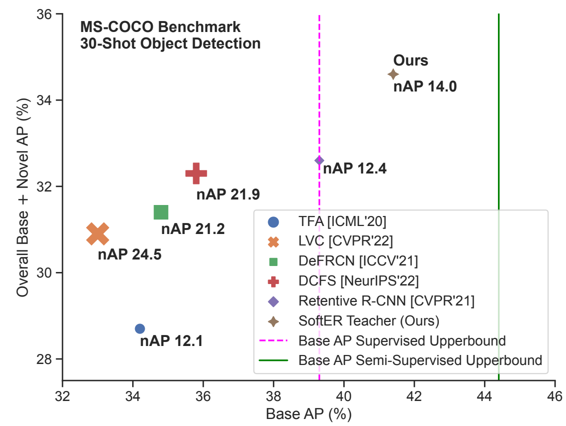


Figure 1: The evaluation of *generalized* FSOD is characterized by the trade-off between novel performance and base forgetting. We leverage unlabeled data to optimize for semi-supervised FSOD on both base + novel classes (top right). Our approach significantly expands base class AP, 39.3 → 44.4, while incurring less than 9% in base degradation (*vs.* 19% for LVC) and also improving on novel detection (nAP). Our SoftER Teacher is the best model on the `Overall AP` metric, leading the next best Retentive R-CNN by +2.0 AP.

This paper focuses on the intersection of SSOD and FSOD, which are essentially two sides of the same coin in the context of label-efficient detection. On one side, SSOD investigates the detection problem with a *small fraction of images* containing ground truth labels. On the other side, FSOD addresses the objective of

Proceedings of the 27[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

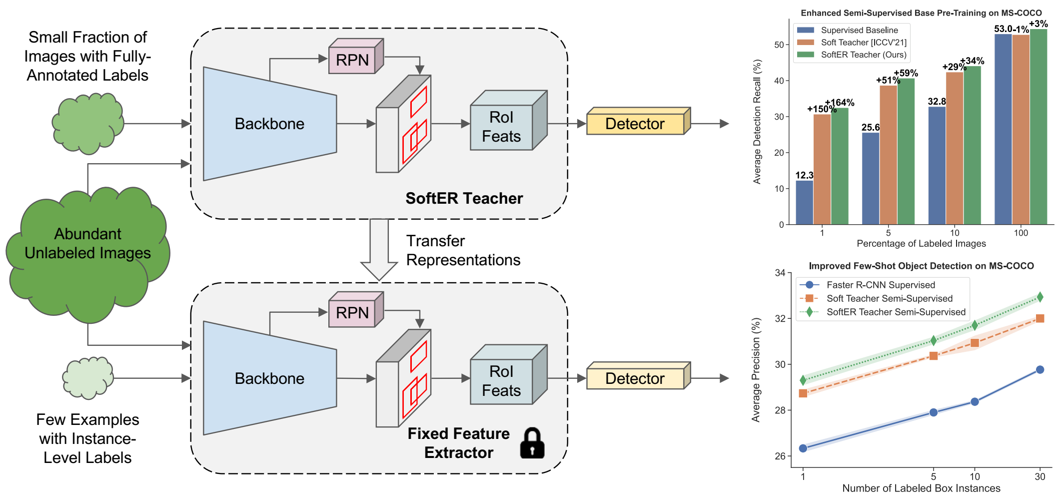


Figure 2: We present the Label-Efficient Detection framework to harness supplementary unlabeled data for generalized semi-supervised few-shot detection. At the core of the framework is our proposed SoftER Teacher with Entropy Regression for improved semi-supervised representation learning (**upper right**). Extensive comparative experiments show that SoftER Teacher is also a more label-efficient few-shot detector (**lower right**).

adapting a *base* detector to learn *novel* concepts from *few instance-level* exemplars. Existing approaches to FSOD all assume abundant base classes to train the base detector. However, such assumption is not ideal in practical scenarios where labels may be limited for both base and novel classes, giving rise to the research question: ***can we advance FSOD given the available unlabeled data without additional hand-labeling?***

We answer this question by introducing the unique framework of *semi-supervised few-shot detection*, in which we explore the utility of unlabeled data for improving detection with label scarcity for both base and novel classes. Inspired by recent advances in SSOD (Xu et al., 2021; Liu et al., 2022; Wang et al., 2023) and FSOD (Fan et al., 2021; Qiao et al., 2021; Gao et al., 2022), our approach is two-fold: (1) we leverage unlabeled data to improve detection with a small fraction of labeled images; and (2) we generalize the resulting semi-supervised detector into a label-efficient few-shot detector by way of transfer learning. Our chief motivation is to not necessarily depend on an abundance of labels for robust few-shot detection, which increases the versatility of our approach in realistic applications.

Moreover, our approach to semi-supervised FSOD adapts a base detector to learn novel concepts with *reduced performance degradation to base classes*, a desirable result missing in most prior approaches. Figure 1 illustrates that while recent work (Qiao et al., 2021; Kaul et al., 2022; Gao et al., 2022) achieve impressive detection on novel categories, *they all ignore the importance of preserving base class accuracy.* For *generalized* FSOD (Fan et al., 2021), the goal is to expand the learned vocabulary of the base detector with novel concepts. As such, base and novel class performances are equally important, since samples at test time may contain instances of both objects. Therefore, the more realistic evaluation metric for FSOD is not only novel AP, but the combined base and novel AP, for which our approach establishes a new state of the art.

We measure the utility of unlabeled data within our integrated semi-supervised few-shot framework, and discover an insightful empirical finding linking the effectiveness of unlabeled data to semi-supervised FSOD by way of region proposals. Without bells and whistles, by simply adding unlabeled data to a supervised detector, we show a marked improvement on both base and novel class performances while also mitigating catastrophic base forgetting (Lopez-Paz & Ranzato, 2017).

Main Contributions First, we introduce a simple and versatile framework to enable semi-supervised FSOD without depending on abundant labels. At the heart of the framework is our proposed SoftER Teacher to combine the strengths of pseudo-labeling with representation learning on unlabeled images, with the unique benefit of enhancing the quality of region proposals to substantially boost semi-supervised FSOD. Moreover, Section 3.1 contributes a new empirical analysis on the relationship between unlabeled data and region proposals, extending earlier results on proposal evaluation beyond supervised detection (Hosang et al., 2016).

Second, our work sheds insight into a potential relationship suggesting that a strong semi-supervised detector is also a label-efficient few-shot detector (Figure 2), an interesting and non-trivial empirical observation linking the two disparate domains. On the task of semi-supervised FSOD, our SoftER Teacher model exceeds the novel class performance of a strong supervised detector (Fan et al., 2021) using less than 10% of required base labels, while exhibiting less than 9% in base forgetting. When trained on 100% of available labels with supplementary unlabeled data, SoftER Teacher sets a new standard on semi-supervised few-shot performance given varying amounts of bounding box annotations.

Third, we establish the Label-Efficient Detection benchmark to quantify the utility of unlabeled data for generalized semi-supervised FSOD. We hope that our benchmark serves as a strong baseline, and a blueprint, to inspire future research toward this new problem setting in the community.

2 RELATED WORK

Semi-Supervised Detection The current state of the art on SSOD belongs to a family of pseudo-labeling methods, which trains a pair of detectors on pseudo labels along with human labels in the student-teacher framework (Tarvainen & Valpola, 2017). One such method is Soft Teacher (Xu et al., 2021) which vastly improves upon its counterparts STAC (Sohn et al., 2020b) and Unbiased Teacher (Liu et al., 2021) by enabling end-to-end pseudo-labeling on unlabeled images. In these methods, the teacher model is an exponential moving average (EMA) of its student counterpart and is used to predict pseudo labels on unlabeled data.

We extend the strong performance of Soft Teacher by incorporating a new module for Entropy Regression to learn additional representations from unlabeled images by way of region proposals. Our model, aptly named *SoftER Teacher*, combines the attractive benefits of pseudo-labeling with supplementary proposal learning to establish a stronger baseline for SSOD.

Few-Shot Detection The simple yet effective two-stage transfer learning approach following TFA (Wang et al., 2020) is currently leading the FSOD literature, which comprises an initial stage of base class pre-training followed by a second stage of novel category fine-tuning. While recent two-stage methods, such as DeFRCN (Qiao et al., 2021) and DCFS (Gao et al., 2022), extend TFA to achieve impressive performance on novel categories, they suffer from considerable base class forgetting, making them sub-optimal in real-world applications requiring accurate detection on test samples containing instances of both classes. Conversely, the methods of Retentive R-CNN (Fan et al., 2021) and DiGeo (Ma et al., 2023) have proven successful in preserving base class performance, without forgetting, but they both have much room for improvement with regards to novel class detection.

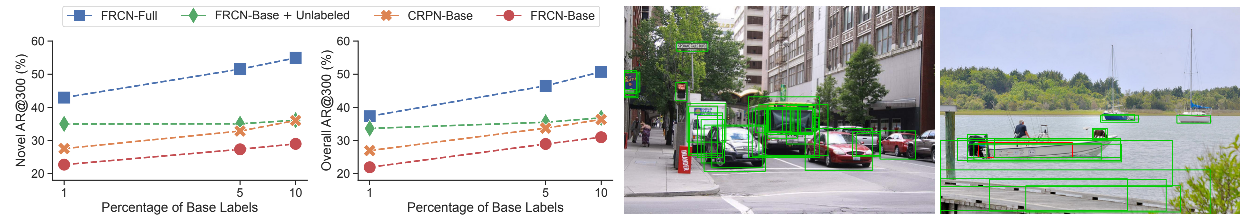
This work takes an important step toward balancing the base-novel performance trade-off by incorporating unlabeled images into the two-stage learning procedure. Our semi-supervised approach demonstrates that both base and novel performances can be further improved, with reduced base degradation, leading to a new standard on multiple challenging FSOD benchmarks.

Semi-Supervised Few-Shot Detection There have been few attempts at leveraging unlabeled data to improve FSOD, but to our knowledge none directly addressed the task of optimizing for semi-supervised FSOD, in which setting both base and novel labels are simultaneously scarce. LVC (Kaul et al., 2022) and MINI (Cao et al., 2022) mine novel targets from the *base training dataset* via pseudo-labeling to boost novel class detection, but come at the cost of base performance. UniT (Khandelwal et al., 2021) obtains impressive results on any-shot detection, but assumes abundant image-level labels for the base and novel targets. And SSFOD (Xiong et al., 2021) performs semi-supervised FSOD within a complicated episodic meta training and N-way k-shot evaluation framework (Karlinsky et al., 2019) while also requiring abundant base classes. Our approach is fundamentally different in that we do not strictly depend on abundant base labels, but make effective utilization of *external unlabeled data*, for robust semi-supervised FSOD with reduced base degradation.

3 APPROACH

We propose to combine the available (limited) labeled examples with supplementary unlabeled images to boost semi-supervised FSOD. We begin with the fully supervised scenario in which we have access to a set of labeled image-target pairs $(x_l, y_l) \in \mathcal{D}_{\text{sup}}$. The supervised FSOD setting assumes an abundant base dataset $C_{\text{base}} \in \mathcal{D}_{\text{sup}}$ to be fully annotated with bounding boxes covering all instances of interest and a novel dataset $C_{\text{novel}} \in \mathcal{D}_{\text{sup}}$ with only a few k (*e.g.*, $k \in \{1, 5, 10\}$) randomly labeled instances, or "shots", per category. The goal of FSOD is to expand the base detector by adapting it to learn new target concepts such that the resulting detector is optimized for accuracy on a test set comprising both classes in $C_{\text{base}} \cup C_{\text{novel}}$.

To maintain parity with the current state of the art, we adopt the two-stage learning procedure consisting of an initial stage of base class pre-training followed by a second stage of novel category fine-tuning. We consider the modern Faster R-CNN (FRCN) (Ren et al., 2015) system as our supervised base detector, which consists of a ResNet (He et al., 2016) *backbone* and a feature pyramid network (FPN) (Lin et al., 2017) *neck* for



(a) The impact of unlabeled data on proposal quality. **(b)** RPN with 1% labels. **(c)** RPN with 10% labels.

Figure 3: We analyze the effectiveness of the RPN as a function of base labels. **(a)** Unlabeled data provide a convincing boost in proposal quality, closing the gap between the `Base` and `Full` detectors, which should lead to better discovery of novel categories during fine-tuning. **(b–c)** In low-label regimes, unlabeled data can help produce diverse proposals (green boxes) on novel unseen objects {`boat`, `bus`, `car`, `dog`}, whereas the vanilla supervised `FRCN-Base` fails to capture comparable foreground objects with only one red box. Best viewed digitally.

feature extraction, a region proposal network (RPN), an RoIAlign (He et al., 2017) operation for mapping proposals to region-of-interest (RoI) features, and a fully-connected *head* for RoI classification and regression. Let FRCN be represented by f_θ, a stochastic function parametrized by a set of learnable weights θ. Formally, the base pre-training step is subjected to the standard supervised objective, over a mini-batch of labeled examples b_l, given by:

$$\mathcal{L}_{\text{sup}} = \frac{1}{|b_l|} \sum_{i \in b_l} \mathcal{L}_{\text{cls}}^{\text{roi}}(f_\theta(x_i), y_i) + \mathcal{L}_{\text{reg}}^{\text{roi}}(f_\theta(x_i), y_i). \quad (1)$$

Here, $f_\theta(x_i)$ denotes a forward pass on the i-th image to produce box classification and localization predictions from class-agnostic proposals, y_i is the i-th ground truth annotation containing box labels and coordinates, and $\left(\mathcal{L}_{\text{cls}}^{\text{roi}}, \mathcal{L}_{\text{reg}}^{\text{roi}}\right)$ are the cross-entropy and L_1 losses, respectively, for the RoI head. Henceforth for simplicity, we develop our approach only on the RoI head and omit the presentation on the losses of the RPN, which remain constant without changes, to predict and localize the "objectness" of region proposals.

3.1 What Makes for Effective FSOD?

We examine this question from the perspective of maximizing representation learning while minimizing base forgetting. In two-stage detectors (*e.g.*, Faster R-CNN), the quality of region proposals is a strong predictor of supervised detection performance (Hosang et al., 2016), since they focus the detector head on candidate RoIs. This is especially true for FSOD approaches based on transfer learning, in which the established procedure is to freeze the RPN during few-shot fine-tuning. Intuitively, if we can incorporate methods and/or data to boost representation learning by way of the RPN, then the detector should have a higher chance of discovering novel categories to improve few-shot performance.

We perform a motivating empirical analysis on the COCO dataset to verify our intuition. We split the dataset into disjoint 60 base and 20 novel categories and pre-train three variants of the FRCN detector on the base classes: (i) FRCN-Base, (ii) FRCN-Base + Unlabeled, which is augmented with `COCO-unlabeled2017` images leveraging the Soft Teacher formulation, and (iii) FRCN-Full using both base and novel classes to represent the upper-bound performance. We also experiment with CRPN-Base, a method specifically designed to improve proposal quality and detection performance using a two-stage Cascade RPN (Vu et al., 2019).

Figure 3a quantifies the "class-agnosticism" of various RPNs, using the standard metric `AR@300` proposals, given varying fractions of base labels. Surprisingly, unlabeled data have the remarkable ability to boost proposal recall on novel-only categories, even in the extremely low 1% label limit. Somewhat unsurprising is the ability of `CRPN-Base` to propose novel objects competitive with `FRCN-Base + Unlabeled` when more labels are available. Consistent with previous findings (Sun et al., 2021), Figures 3b–c show the vanilla supervised `FRCN-Base` has a strong tendency to reject novel objects as background, due to the lack of annotations, resulting in the worst recall on novel classes.

As alluded in Section 1, the contribution of unlabeled data to FSOD can also help mitigate catastrophic base forgetting. We find analogous effectiveness of `FRCN-Base + Unlabeled` on the combined `Overall AR@300` metric, for both base + novel objects, suggesting the RPN, when trained with unlabeled data, has the ability to retain base proposals and help combat base degradation during few-shot fine-tuning.

Discussion This paper rethinks a different and more versatile way to improve the RPN for FSOD while avoiding catastrophic forgetting. The previous LVC approach proposed to unfreeze the RPN during fine-tuning to obtain large performance gains on novel classes, but comes at the cost of significant base degradation (up to 19%). Similarly, FSCE (Sun et al., 2021) proposed to unfreeze the RPN while also doubling the number of proposals to encourage novel foreground detection during fine-tuning. However, this method increases the detection overhead and remains unclear whether it helps mitigate base forgetting. We illustrate that simply adding unlabeled data to the base detector leads to a compelling boost in proposal quality, without the need for any *ad hoc* modifications to the RPN.

We attribute this unique advantage of our approach to the potential base-novel object interactions found in abundant images. When learning with unlabeled data, the base detector can obtain semantically similar cues of novel objects to inform the RPN on foreground detection. Sun et al. (2021) showed that visually analogous objects have high cosine similarity scores (*e.g.*, $\text{sim}(\texttt{cow}, \texttt{horse}) = 0.39$). With 1% of labels, these base-novel interactions are limited, resulting in a recall of 22.7%. Given a sizable unlabeled dataset, the base detector improves its representations to yield a large gain of +12.3 AR points. Extensive experiments in Section 4.2 demonstrate that proposal recall indeed has a strong correlation with FSOD performance.

3.2 Semi-Supervised Base Pre-Training

Encouraged by the promising utility of unlabeled data, we now relax the strict assumption on having abundant base classes for FSOD and introduce a more general setting of having a small fraction of base labels given abundant unlabeled images. We approach the task of semi-supervised base pre-training by formulating an unsupervised loss computed on an unlabeled dataset $\mathcal{D}_{\text{unsup}}$ to be jointly trained with the supervised loss on $\mathcal{D}_{\text{sup}}$. We consider the following canonical optimization objective widely adopted as part of the framework for semi-supervised learning (Bachman et al., 2014; Sajjadi et al., 2016; Yang et al., 2023):

$$\min_{\theta} \mathcal{L}_{\text{sup}}(\mathcal{D}_{\text{sup}}, \theta) + \lambda \mathcal{L}_{\text{unsup}}(\mathcal{D}_{\text{unsup}}, \theta), \tag{2}$$

where $\lambda > 0$ is a hyper-parameter controlling the contribution of the unsupervised component. Next, we describe the unsupervised criterion on $\mathcal{D}_{\text{unsup}}$ to make FRCN into a semi-supervised detector.

Soft Teacher We adopt Soft Teacher (Xu et al., 2021) as the baseline SSOD formulation for its simplicity but strong performance. Soft Teacher trains FRCN in a student-teacher fashion on both labeled and unlabeled data. The student is trained on labeled examples in the standard supervised manner per Eq. (1). For unlabeled images, the teacher is treated as a fixed detector to generate thousands of box candidates, most of which are eliminated for redundancy with non-maximum suppression. Additionally, box candidates are thresholded for foreground objects and go through an iterative jittering-refinement procedure to reduce localization variance, resulting in a set of high-quality pseudo boxes to be jointly trained with ground truth annotations.

As is common practice (Sohn et al., 2020a), the teacher's parameters $\bar{\theta}$ are updated from the student's via $\bar{\theta} = \text{EMA}(\theta)$ at each training step. Integral to the success of Soft Teacher is a student-teacher data augmentation strategy inspired by STAC (Sohn et al., 2020b). The student trains on unlabeled images subjected to complex random perturbations, akin to RandAugment (Cubuk et al., 2020), including affine transforms. Separately, the teacher receives weakly augmented images with simple random resizing and flipping. This multi-stream augmentation design allows the teacher to generate reliable unsupervised targets on easy images to guide the student's learning trajectory on difficult images for better generalizability.

At the time of box classification and regression in the RoI head, we have a set of unlabeled images along with teacher-generated pseudo labels $(x_u, \hat{y}_u) \in \mathcal{D}_{\text{unsup}}$. The unsupervised loss for Soft Teacher on a mini-batch of unlabeled images b_u is defined as:

$$\mathcal{L}_{\text{soft}} = \frac{1}{|b_u|} \sum_{i \in b_u} \mathcal{L}_{\text{cls}}^{\text{roi}}(f_\theta(x_i), \hat{y}_i) + \mathcal{L}_{\text{reg}}^{\text{roi}}(f_\theta(x_i), \hat{y}_i). \tag{3}$$

SoftER Teacher The design of Soft Teacher employs class confidence thresholding and box jittering to select high-quality pseudo candidates for unsupervised classification and regression. However, Xu et al. (2021) showed that it uses an aggressive threshold of 0.9, resulting in a trade-off between low recall and high precision at 33% and 89%, respectively. We observe that low recall can result in poor detection performance on small and ambiguous objects (Li et al., 2018), especially in low-label regimes where the teacher has insufficient confidence about its predicted pseudo labels. We aim to extend Soft Teacher and improve its detection recall by learning additional representations from abundant region proposals. We show in Section 4.2 that recall is key to an effective few-shot detector.

Given a set of proposals p generated by the student's RPN on a batch of unlabeled images, we apply the student-teacher data augmentation pipeline described above to obtain (p_s, p_t), denoting transformed student and teacher proposals, which are related to each other by a matrix M. We then forward pass FRCN twice, as the student f_θ and teacher $f_{\bar{\theta}}$, to obtain two sets of RoI outputs for predicted box classification logits (z_s, z_t) and localization coordinates (r_s, r_t). Let g_c be the softmax function over the channel dimension c. We define an auxiliary unsupervised criterion for proposal box

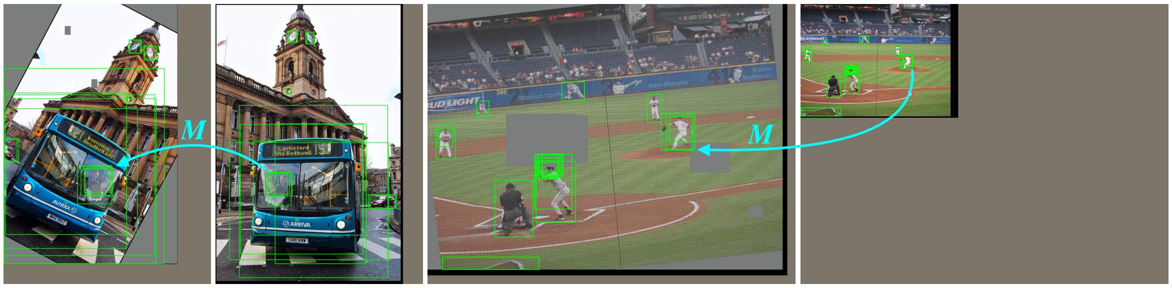


Figure 4: Visualizations of student-teacher proposals with confidence scores ≥ 0.99. As illustrated by the arrow, a pair of student-teacher proposals is related by a transformation matrix M, which is used to align proposals between student and teacher images for enforcing box classification similarity and localization consistency.

similarity based on a cross-entropy measure $H(z_s, z_t)$:

$$\mathcal{L}_{\text{cls}}^{\text{ent}} = \frac{1}{\sum_i w_i} \sum_{i \in p} w_i \cdot H(z_{is}, z_{it}), \quad \text{where} \quad H(z_{is}, z_{it}) = -\frac{1}{C} \sum_{c \in C} g_c(z_{it}) \log g_c(z_{is}). \tag{4}$$

Here, g_c outputs a distribution over C classes and w_i is the Boolean weight for the predicted foreground class: $w_i = 1$ if $\text{argmax}(z_{it}) \neq \textit{background}$, else $w_i = 0$.

Analogously for proposal box regression, we constrain the predicted box coordinates (r_s, r_t) to be close. Since there are complex geometric distortions between the two, we first map teacher proposal coordinates r_t to the student space using the transformation M. Then, we align the proposal boxes via the intersection-over-union (IoU) loss criterion (Rezatofighi et al., 2019) to compute their differences:

$$\mathcal{L}_{\text{reg}}^{\text{iou}} = 1 - \frac{1}{|p|} \sum_{i \in p} w_i \cdot \text{IoU}(r_{is}, M(r_{it})). \tag{5}$$

Recall that we have two different transformation pipelines operating on each proposal, so we have two augmented views of each proposal. Figure 4 illustrates that by enforcing these randomly augmented views, and their box coordinates, to be *similar*, we enable the student to tap into abundant region proposals to learn diverse feature representations across a spectrum of scale, color, and geometric perturbations. Our formulation draws inspiration from recent research on self-supervised representation learning with multi-augmented views (Grill et al., 2020; Chen & He, 2021). Note the cross-entropy similarity between the student and teacher predictions, Eq. (4), can be interpreted as a form of entropy regularization (Grandvalet & Bengio, 2004), which has been proven to work well in various semi-supervised classification scenarios (Miyato et al., 2017; Oliver et al., 2018). Combining Eqs. (1) and (3) to (5), the total optimization objective at the RoI head for our SoftER Teacher model is computed as:

$$\mathcal{L}_{\text{softer}}^{\text{total}} = \mathcal{L}_{\text{sup}} + \alpha \mathcal{L}_{\text{soft}} + \beta \left(\mathcal{L}_{\text{cls}}^{\text{ent}} + \mathcal{L}_{\text{reg}}^{\text{iou}} \right), \tag{6}$$

where we set $\alpha = \frac{|b_u|}{|b_l|}$ following Soft Teacher and find $\beta = 2\alpha$ works reliably well across all experiments.

3.3 Semi-Supervised Few-Shot Fine-Tuning

We propose a simple two-step approach to harness unlabeled data for semi-supervised few-shot fine-tuning. First, we initialize the few-shot detector, $f'_{\bar{\theta}} \leftarrow f_{\bar{\theta}}$, with parameters copied from the base *teacher* detector pre-trained with unlabeled data per Eq. (6). And second, we further train the RoI head of $f'_{\bar{\theta}}$ on novel classes using the available few-shot and unlabeled examples while freezing the base backbone, FPN, and RPN components. Then, we fine-tune the few-shot detector on a balanced training set of k shots per class containing both base and novel instances. We update only the RoI box classifier while freezing all other components, including the box regressor, since it is the main source of error (Sun et al., 2021). Rigorous experiments show that our approach provides a compelling boost to novel performance while enjoying substantial gains on base accuracy with reduced base forgetting. We present detailed ablation studies in Section 5 and Appendix A.6 to validate our approach and design choices.

4 EXPERIMENTS

Datasets We evaluate our approach on the challenging PASCAL VOC (Everingham et al., 2010) and MS-COCO 2017 (Lin et al., 2014) detection benchmarks. For VOC, we use the combined VOC07+12 `trainval` splits as the labeled training set and evaluate on the VOC2007 `test` set. For COCO, we utilize the `train2017` split as labeled data and test on `val2017`.

Implementation Details We conduct our few-shot experiments following the TFA benchmark (Wang

Table 1: FSOD results evaluated on COCO `val2017`. We report the mean and 95% confidence interval over 5 random samples for our models. SoftER Teacher with ResNet-101 is the best model on the combined `Overall AP` metric, incurring less than 9% in base forgetting *vs.* 11%–DCFS, 17%–DeFRCN, 18%–TFA, and 19%–LVC.

COCO `val2017`	Backbone	Base $AP_{50:95}$	Base $AP_{50:95}$ (60 Classes)			Novel $AP_{50:95}$ (20 Classes)			Overall $AP_{50:95}$ (80 Classes)		
Method			5-Shot	10-Shot	30-Shot	5-Shot	10-Shot	30-Shot	5-Shot	10-Shot	30-Shot
LVC* [CVPR'22]	R-50	–	–	29.7	33.3	–	**17.6**	**25.5**	–	26.7	31.4
SoftER Teacher (Ours)	R-50	**42.0**	**37.4 ± 0.2**	**37.4 ± 0.2**	**38.7 ± 0.2**	**8.2 ± 0.3**	10.3 ± 0.5	12.9 ± 0.6	**30.1 ± 0.1**	**30.7 ± 0.2**	**32.3 ± 0.1**
LVC* [CVPR'22]	R-101	39.3	–	31.9	33.0	–	17.8	**24.5**	–	28.4	30.9
TFA [ICML'20]	R-101	39.3	32.3 ± 0.6	32.4 ± 0.6	34.2 ± 0.4	7.0 ± 0.7	9.1 ± 0.5	12.1 ± 0.4	25.9 ± 0.6	26.6 ± 0.5	28.7 ± 0.4
DeFRCN [ICCV'21]	R-101	39.3	32.6 ± 0.3	34.0 ± 0.2	34.8 ± 0.1	13.6 ± 0.7	16.8 ± 0.6	21.2 ± 0.4	27.8 ± 0.3	29.7 ± 0.2	31.4 ± 0.1
DCFS [NeurIPS'22]	R-101	39.3	35.0 ± 0.2	35.7 ± 0.2	35.8 ± 0.2	**15.7 ± 0.5**	**18.3 ± 0.4**	21.9 ± 0.3	30.2 ± 0.2	31.4 ± 0.2	32.3 ± 0.2
Retentive R-CNN [CVPR'21]	R-101	39.3	39.3 ± n/a	39.2 ± n/a	39.3 ± n/a	7.7 ± n/a	9.5 ± n/a	12.4 ± n/a	31.4 ± n/a	31.8 ± n/a	32.6 ± n/a
SoftER Teacher (Ours)	R-101	**44.4**	**40.3 ± 0.2**	**40.2 ± 0.3**	**41.4 ± 0.2**	8.7 ± 0.6	11.0 ± 0.4	14.0 ± 0.6	**32.4 ± 0.2**	**32.9 ± 0.1**	**34.6 ± 0.1**

* indicates results were reported on a single sample run (hence the lack of error bars), which may have been over-estimated due to the high variance of few-shot training samples.

Table 2: FSOD results evaluated on VOC07 `test`. We report the mean and 95% confidence interval over 10 samples for our models. SoftER Teacher with ResNet-50 surpasses the supervised MPSR (Wu et al., 2020), TFA, and Retentive R-CNN models with ResNet-101 by a large margin on most metrics under study, while being more parameter-efficient with a smaller backbone. Results for the other two splits are given in Appendix A.1.

VOC07 `test` **– Split 1**	Backbone	Base AP_{50}	Base AR_{50}	Base AP_{50} (15 Classes)			Novel AP_{50} (5 Classes)			Overall AP_{50} (20 Classes)		
Method				1-Shot	5-Shot	10-Shot	1-Shot	5-Shot	10-Shot	1-Shot	5-Shot	10-Shot
MPSR* [ECCV'20]	R-101	80.8	–	61.5	69.7	71.6	**42.8**	55.3	61.2	56.8	66.1	69.0
Retentive R-CNN* [CVPR'21]	R-101	80.8	–	80.9	80.8	80.8	42.4	53.7	56.1	71.3	74.0	74.6
TFA [ICML'20]	R-101	80.8	–	77.6 ± 0.2	77.4 ± 0.3	77.5 ± 0.2	25.3 ± 2.2	47.9 ± 1.2	52.8 ± 1.0	64.5 ± 0.6	70.1 ± 0.4	71.3 ± 0.3
SoftER Teacher (Ours)	R-101	**85.0**	**93.1**	**83.3 ± 0.4**	**84.6 ± 0.2**	**84.9 ± 0.2**	35.5 ± 3.7	**60.3 ± 2.8**	**65.9 ± 1.6**	**71.3 ± 1.2**	**78.5 ± 0.8**	**80.2 ± 0.4**
Faster R-CNN (Our Impl.)	R-50	81.7	88.0	82.0 ± 0.2	82.4 ± 0.1	82.3 ± 0.1	27.9 ± 3.2	52.1 ± 2.1	58.2 ± 1.6	68.5 ± 0.8	74.9 ± 0.5	76.2 ± 0.4
Soft Teacher (Our Impl.)	R-50	**85.0**	90.9	**84.6 ± 0.4**	**85.2 ± 0.1**	**85.0 ± 0.1**	29.5 ± 4.4	53.8 ± 2.6	60.8 ± 1.5	**70.8 ± 1.2**	77.4 ± 0.7	79.0 ± 0.3
SoftER Teacher (Ours)	R-50	84.7	**92.3**	83.1 ± 0.2	84.4 ± 0.2	84.6 ± 0.2	**31.3 ± 4.3**	**57.8 ± 2.6**	**62.9 ± 1.7**	70.2 ± 1.2	**77.8 ± 0.7**	**79.2 ± 0.4**

* indicates results were reported on a single sample run (hence the lack of error bars), which may have been over-estimated due to the high variance of few-shot training samples.

et al., 2020). The VOC dataset is randomly partitioned into 15 base and 5 novel classes, in which there are $k \in \{1, 5, 10\}$ labeled boxes per category sampled from the joint VOC07+12 `trainval` splits. And the COCO dataset is divided into 60 base and 20 novel classes having the same VOC category names with $k \in \{5, 10, 30\}$ shots. We leverage `COCO-train2017` and `COCO-unlabeled2017` as external unlabeled sources to augment base pre-training and novel fine-tuning on VOC and COCO, respectively. We adopt the ResNet-101 backbone pre-trained on ImageNet-1K (Deng et al., 2009) for a direct comparison with existing work. We assess detection performance using average precision (AP) and recall (AR) metrics, following established protocols. Complete details are given in Appendix B.

4.1 SoftER Teacher is a Label-Efficient Few-Shot Detector

Table 1 compares the effectiveness of SoftER Teacher against other two-stage learning methods representing the state of the art on COCO, for the evaluation of both base and novel performances. We report the ideal supervised base AP of 39.3 (Fan et al., 2021) along with our substantially improved semi-supervised base AP of 44.4 to measure the extent of base forgetting. Recall that the more realistic evaluation metric for *generalized* FSOD is not only novel AP, but the combination of base and novel AP. We summarize the following key takeaways: **(a)** SoftER Teacher with ResNet-101 trained with supplementary unlabeled data is the best model on the combined `Overall AP` metric for 80 classes, leading the next best Retentive R-CNN by up to +2.0 AP; **(b)** SoftER Teacher with ResNet-50 surpasses Retentive R-CNN on novel performance while being more parameter-efficient; and **(c) SoftER Teacher achieves the state of the art while being more efficient with respect to parameters and labels**.

Table 2 shows comparative VOC results. We report the supervised base AP of 80.8 (Fan et al., 2021) along with our vastly expanded base AP of 85.0 using unlabeled data. SoftER Teacher with ResNet-50 incurs negligible base forgetting of less than 2% while exceeding the competition with ResNet-101 on most metrics. When equipped with ResNet-101, SoftER Teacher further improves on both base and novel classes by a notable margin. To our knowledge, MPSR, Retentive R-CNN, and LVC did not perform repeated runs over multiple random seeds per established protocol, the results of which are not comparable to ours. In general, the previous works of TFA, DeFRCN, DCFS, and ours all report marked reduction in novel performances with repeated sample runs when compared to a single trial.

Discussion Tables 1 and 2 corroborate our observation on the trade-off between novel performance and base forgetting, for which our approach aims to jointly optimize. We accomplish this goal with the help of our Entropy Regression module to learn with external unlabeled data "in the wild", without assuming the presence

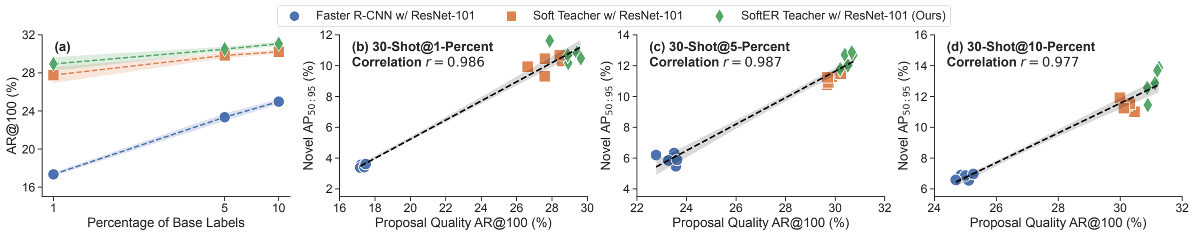


Figure 5: Proposal quality is highly correlated with semi-supervised FSOD. SoftER Teacher produces the best proposals among the comparisons **(a)**, which yield the strongest 30-shot performances **(b–d)**. Shaded regions denote standard deviation over 5 samples. Appendix A.2 gives similar trends for 5-shot and 10-shot results.

Table 3: We introduce the Label-Efficient Detection benchmark for generalized semi-supervised FSOD on COCO `val2017`. All models are equipped with ResNet-101. We report the mean and standard deviation over 5 samples. Using only 10% of base labels **(bottom row)**, SoftER Teacher surpasses the supervised novel performance of Retentive R-CNN trained with 100% of base labels **(top row)** while incurring less than 9% in base degradation.

Method	% Labeled	Base $AP_{50:95}$	Base $AR_{50:95}$	Base $AP_{50:95}$ (60 Classes)			Novel $AP_{50:95}$ (20 Classes)			Overall $AP_{50:95}$ (80 Classes)		
				5-Shot	10-Shot	30-Shot	5-Shot	10-Shot	30-Shot	5-Shot	10-Shot	30-Shot
Retentive R-CNN	100	39.3	–	39.3	39.2	39.3	7.7	9.5	12.4	31.4	31.8	32.6
SoftER Teacher		44.4	56.1	40.3 ± 0.2	40.2 ± 0.3	41.4 ± 0.2	8.7 ± 0.6	11.0 ± 0.4	14.0 ± 0.6	32.4 ± 0.2	32.9 ± 0.1	34.6 ± 0.1
Faster R-CNN		8.7 ± 0.3	12.3 ± 0.5	9.8 ± 0.3	10.0 ± 0.4	10.8 ± 0.3	1.9 ± 0.3	2.7 ± 0.1	3.5 ± 0.1	7.8 ± 0.2	8.2 ± 0.3	9.0 ± 0.2
Soft Teacher	1	19.9 ± 1.0	30.7 ± 1.1	19.4 ± 0.7	19.9 ± 0.8	21.2 ± 0.7	5.9 ± 0.8	7.9 ± 0.7	10.1 ± 0.5	16.0 ± 0.6	16.9 ± 0.7	18.4 ± 0.6
SoftER Teacher		19.8 ± 0.9	32.5 ± 1.0	19.2 ± 0.6	19.8 ± 0.5	21.1 ± 0.5	6.7 ± 0.3	8.8 ± 0.2	10.8 ± 0.5	16.1 ± 0.5	17.1 ± 0.4	18.5 ± 0.5
Faster R-CNN		19.1 ± 0.3	25.6 ± 0.4	18.5 ± 0.5	18.9 ± 0.3	20.0 ± 0.5	3.5 ± 0.2	4.6 ± 0.2	5.9 ± 0.3	14.8 ± 0.4	15.3 ± 0.2	16.5 ± 0.4
Soft Teacher	5	29.6 ± 0.3	38.7 ± 0.3	27.5 ± 0.4	27.8 ± 0.5	29.2 ± 0.5	6.7 ± 0.7	8.9 ± 0.4	11.1 ± 0.3	22.3 ± 0.4	23.1 ± 0.3	24.7 ± 0.4
SoftER Teacher		30.2 ± 0.2	40.7 ± 0.3	27.5 ± 0.4	27.9 ± 0.4	29.3 ± 0.2	7.9 ± 0.4	10.1 ± 0.5	12.4 ± 0.5	22.6 ± 0.3	23.4 ± 0.3	25.1 ± 0.2
Faster R-CNN		24.7 ± 0.2	32.8 ± 0.3	22.6 ± 0.4	22.8 ± 0.1	24.2 ± 0.2	3.8 ± 0.5	5.3 ± 0.2	6.8 ± 0.2	17.9 ± 0.3	18.4 ± 0.1	19.9 ± 0.2
Soft Teacher	10	33.3 ± 0.2	42.4 ± 0.2	30.5 ± 0.5	30.7 ± 0.4	32.1 ± 0.3	6.8 ± 0.3	9.0 ± 0.6	11.4 ± 0.3	24.6 ± 0.4	25.3 ± 0.4	26.9 ± 0.3
SoftER Teacher		33.4 ± 0.4	44.1 ± 0.2	30.3 ± 0.5	30.6 ± 0.5	32.0 ± 0.4	7.9 ± 1.3	10.4 ± 1.1	12.9 ± 1.0	24.6 ± 0.1	25.6 ± 0.3	27.2 ± 0.3

of abundant novel instances in such sources. Our VOC experiments yield state-of-the-art results by harnessing uncurated images from the `COCO-train2017` dataset, which exhibits strong domain mismatch and contains many classes outside of VOC. By stark contrast, LVC mines novel targets as auxiliary samples from the *base training set*, thereby making an unrealistic assumption that abundant novel classes necessarily be present at training time. While such assumption enables impressive novel performance, LVC suffers significant base forgetting, making it sub-optimal in practical settings.

Similarly, DeFRCN and DCFS extend the base FRCN architecture with complex auxiliary modules specially designed to promote strong novel performance, but also come at the cost of considerable base forgetting. We view these methods as complementary to ours, which can be further extended to take advantage of unlabeled data. Future work would explore how our SoftER Teacher formulation could be integrated with Retentive R-CNN, DeFRCN, and DCFS to push the performance envelope of FSOD without forgetting.

4.2 How Does Proposal Quality Impact Semi-Supervised Few-Shot Detection?

In this section, we continue our discussion from Section 3.1 by analyzing semi-supervised FSOD as a function of proposal quality in Figure 5. We measure proposal quality using the metric AR@p, for $p \in \{100, 300, 1000\}$ proposals, averaged over thresholds between 0.5 and 0.95. We arrive at the following conclusions: **(a)** SoftER Teacher produces better proposals than the comparisons across varying fractions of base labels; and **(b–d)** proposal quality is strongly correlated with semi-supervised FSOD, an insightful empirical finding that extends existing results beyond supervised detection (Hosang et al., 2016).

Although the strong Soft Teacher baseline is also effective at harnessing unlabeled data for semi-supervised FSOD, *our approach demonstrates superior learning by addressing a key shortcoming of Soft Teacher.* SoftER Teacher boosts object recall via our proposed Entropy Regression module, improving on Soft Teacher by +1.7 base $AR_{50:95}$, which yields a gain of +1.5 novel AP for the `30-Shot@10-Percent` setting. These results further bolster our empirical observation that a stronger semi-supervised detector leads to a more label-efficient few-shot detector. In principle, our versatile semi-supervised FSOD framework can generalize any semi-supervised detector to the few-shot setting. Future work would investigate if our findings can be extended to a more general case with other SSOD formulations, including one-stage detectors.

Table 4: Ablation experiments quantifying the effectiveness of each component in our semi-supervised approach using 1% of COCO labels. The first row corresponds to the Soft Teacher baseline and the last row is our SoftER Teacher configuration.

Proposal Similarity Measure	Proposal IoU Regression	$AP_{50:95}$	$AR_{50:95}$
None	✗	22.4	30.8
KL-Divergence	✗	22.8	31.5
Cross-Entropy (Eq. (4))	✗	22.7	31.6
None	✓	22.3	30.8
KL-Divergence	✓	22.9	31.8
Cross-Entropy (Eq. (4))	✓	**23.0**	**32.0**

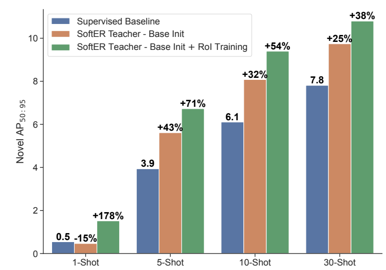


Figure 6: Unlabeled data improve semi-supervised few-shot fine-tuning on COCO `val2017`.

4.3 A New Benchmark for Generalized Semi-Supervised Few-Shot Detection

We present our Label-Efficient Detection benchmark on COCO `val2017` in Table 3. Our protocol for semi-supervised FSOD is as follows. In the first stage, we pre-train the base detector on the disjoint 60 base categories using $\{1, 5, 10\}$ percent of labels per Eq. (6). In the second stage, we transfer the parameters of the base teacher detector to the few-shot detector and fine-tune its RoI box classifier, keeping other components frozen, on a balanced training set of $k \in \{5, 10, 30\}$ shots per class containing both base and novel examples. In both stages, we supplement base pre-training and novel fine-tuning with images from `COCO-unlabeled2017`.

We report $AP_{50:95}$ performances on both base and novel classes along with the aggregated overall metric. We also report the ideal $AP_{50:95}$ and $AR_{50:95}$ metrics obtained from the first stage of base pre-training to measure the potential for base forgetting during the few-shot fine-tuning step. We encourage future work to follow suit as we emphasize the importance of optimizing for accuracy on both base and novel classes, a desideratum of generalized few-shot object detection.

5 Ablation Studies

5.1 SoftER Teacher System Design

Table 4 shows an ablation study on 1% of COCO labels to assess the key elements in our SoftER Teacher approach for SSOD. Compared to the Soft Teacher baseline (first row), the addition of the cross-entropy or KL-divergence measure to enforce proposal consistency leads to a boost in both AP and AR, although the performance difference between the two measures is immaterial. Interestingly, the addition of the IoU regression loss alone does not produce a performance improvement over the Soft Teacher baseline. However, when we couple IoU regression with the cross-entropy similarity measure, we obtain the best performing configuration (last row). SoftER Teacher improves on both precision and recall over the strong Soft Teacher baseline via our proposed Entropy Regression module for proposal learning with complex affine transformations.

5.2 Semi-Supervised Few-Shot Fine-Tuning with Unlabeled Data

As discussed in Section 3.3, we explore two ways of leveraging unlabeled data to fine-tune the few-shot detector on novel classes: (1) we initialize the few-shot detector with parameters copied from the base *teacher* detector pre-trained with unlabeled data per Eq. (6); and (2) we further train the RoI box classifier and regressor on novel classes using the available few-shot and unlabeled examples while freezing the base backbone, FPN, and RPN components. Figure 6 illustrates semi-supervised base initialization boosts novel AP by as much as 43% on COCO `val2017`, compared to the supervised baseline. In addition to semi-supervised base initialization, training the RoI head on few-shot novel classes with unlabeled images further amplifies the novel AP margin of SoftER Teacher.

6 CONCLUSION

This paper presented the Label-Efficient Detection framework to quantify the utility of unlabeled data for generalized semi-supervised FSOD. Central to the framework is our proposed SoftER Teacher model to boost semi-supervised FSOD, by way of consistency learning on diverse region proposals, without relying on an abundance of labels. Our simple and versatile LEDetection framework can inspire exciting future research directions to combine the latest advances in both SSOD and FSOD, thereby helping to unify the two disparate domains.

Acknowledgments

The author is grateful to Victor Palmer for insightful discussions, Brian Keller and Cole Winans for their continued support, and anonymous reviewers for their constructive feedback on this paper.

References

Bachman, P., Alsharif, O., and Precup, D. Learning with Pseudo-Ensembles. In *NeurIPS*, 2014. 5

Cao, Y., Wang, J., Lin, Y., and Lin, D. MINI: Mining Implicit Novel Instances for Few-Shot Object Detection. https://arxiv.org/abs/2205.03381, 2022. 3

Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., et al. MMDetection: Open MMLab Detection Toolbox and Benchmark. https://arxiv.org/abs/1906.07155, 2019. 17

Chen, X. and He, K. Exploring Simple Siamese Representation Learning. In *CVPR*, 2021. 6

Cubuk, E. D., Zoph, B., Shlens, J., and Le, Q. V. RandAugment: Practical Automated Data Augmentation with a Reduced Search Space. In *NeurIPS*, 2020. 5, 16

Deng, J., Dong, W., Socher, R., Li, L., Li, K., and Fei-Fei, L. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848. 7

DeVries, T. and Taylor, G. W. Improved Regularization of Convolutional Neural Networks with Cutout. https://arxiv.org/abs/1708.04552, 2017. 16

Everingham, M., Gool, L. V., Williams, C. K., Winn, J., and Zisserman, A. The PASCAL Visual Object Classes (VOC) Challenge. *IJCV*, 88(2):303–338, 2010. 6

Fan, Z., Ma, Y., Li, Z., and Sun, J. Generalized Few-Shot Object Detection without Forgetting. In *CVPR*, 2021. 2, 3, 7, 12

Gao, B.-B., Chen, X., Huang, Z., Nie, C., Liu, J., Lai, J., Jiang, G., Wang, X., and Wang, C. Decoupling Classifier for Boosting Few-Shot Object Detection and Instance Segmentation. In *NeurIPS*, 2022. 2, 3, 15

Grandvalet, Y. and Bengio, Y. Semi-Supervised Learning by Entropy Minimization. In *NeurIPS*, 2004. 6

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P. H., Buchatskaya, E., Doersch, C., Pires, B. A., et al. Bootstrap Your Own Latent: A New Approach to Self-Supervised Learning. In *NeurIPS*, 2020. 6

He, K., Zhang, X., Ren, S., and Sun, J. Deep Residual Learning for Image Recognition. In *CVPR*, 2016. 3

He, K., Gkioxari, G., Dollár, P., and Girshick, R. Mask R-CNN. In *ICCV*, 2017. 4

Hosang, J., Benenson, R., Dollár, P., and Schiele, B. What Makes for Effective Detection Proposals? *IEEE TPAMI*, 38(4):814–830, 2016. 2, 4, 8, 13

Karlinsky, L., Shtok, J., Harary, S., Schwartz, E., Aides, A., Feris, R., Giryes, R., and Bronstein, A. M. RepMet: Representative-Based Metric Learning for Classification and Few-Shot Object Detection. In *CVPR*, 2019. 3

Kaul, P., Xie, W., and Zisserman, A. Label, Verify, Correct: A Simple Few Shot Object Detection Method. In *CVPR*, 2022. 2, 3

Khandelwal, S., Goyal, R., and Sigal, L. UniT: Unified Knowledge Transfer for Any-shot Object Detection and Segmentation. In *CVPR*, 2021. 3

Li, Z., Peng, C., Yu, G., Zhang, X., Deng, Y., and Sun, J. DetNet: A Backbone Network for Object Detection. In *ECCV*, 2018. 5

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft COCO: Common Objects in Context. In *ECCV*, 2014. 1, 6

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., and Belongie, S. Feature Pyramid Networks for Object Detection. In *CVPR*, 2017. 3

Liu, Y.-C., Ma, C.-Y., He, Z., Kuo, C.-W., Chen, K., Zhang, P., Wu, B., Kira, Z., and Vajda, P. Unbiased Teacher for Semi-Supervised Object Detection. In *ICLR*, 2021. 3, 16

Liu, Y.-C., Ma, C.-Y., and Kira, Z. Unbiased Teacher v2: Semi-Supervised Object Detection for Anchor-Free and Anchor-Based Detectors. In *CVPR*, 2022. 2

Lopez-Paz, D. and Ranzato, M. Gradient Episodic Memory for Continual Learning. In *NeurIPS*, 2017. 2

Ma, J., Niu, Y., Xu, J., Huang, S., Han, G., and Chang, S.-F. DiGeo: Discriminative Geometry-Aware Learning for Generalized Few-Shot Object Detection. In *CVPR*, 2023. 3

Miyato, T., Maeda, S., Koyama, M., and Ishii, S. Virtual Adversarial Training: A Regularization Method for Supervised and Semi-Supervised Learning. *IEEE TPAMI*, 41:1979–1993, 2017. 6

Oliver, A., Odena, A., Raffel, C., Cubuk, E. D., and Goodfellow, I. J. Realistic Evaluation of Deep Semi-Supervised Learning Algorithms. In *NeurIPS*, 2018. 6

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *NeurIPS*, pp. 8024–8035. Curran Associates, Inc., 2019. 17

Qiao, L., Zhao, Y., Li, Z., Qiu, X., Wu, J., and Zhang, C. DeFRCN: Decoupled Faster R-CNN for Few-Shot Object Detection. In *ICCV*, 2021. 2, 3

Ren, S., He, K., Girshick, R., and Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *NeurIPS*, 2015. 3

Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., and Savarese, S. Generalized Intersection over Union: A Metric and A Loss for Bounding Box Regression. In *CVPR*, 2019. 6

Sajjadi, M., Javanmardi, M., and Tasdizen, T. Regularization with Stochastic Perturbations for Deep Semi-Supervised Learning. In *NeurIPS*, 2016. 5

Sohn, K., Berthelot, D., Li, C.-L., Zhang, Z., Carlini, N., Cubuk, E. D., Kurakin, A., Zhang, H., and Raffel, C. Fixmatch: Simplifying Semi-Supervised Learning with Consistency and Confidence. In *NeurIPS*, 2020a. 5, 17

Sohn, K., Zhang, Z., Li, C.-L., Zhang, H., Lee, C.-Y., and Pfister, T. A Simple Semi-Supervised Learning Framework for Object Detection. https://arxiv.org/abs/2005.04757, 2020b. 3, 5, 16

Sun, B., Li, B., Cai, S., Yuan, Y., and Zhang, C. FSCE: Few-Shot Object Detection via Contrastive Proposal Encoding. In *CVPR*, 2021. 4, 5, 6, 15

Tang, Y., Chen, W., Luo, Y., and Zhang, Y. Humble Teachers Teach Better Students for Semi-Supervised Object Detection. In *CVPR*, 2021. 13, 16, 17

Tarvainen, A. and Valpola, H. Mean Teachers are Better Role Models: Weight-Averaged Consistency Targets Improve Semi-Supervised Deep Learning Results. In *NeurIPS*, 2017. 3, 17

Vu, T., Jang, H., Pham, T. X., and Yoo, C. D. Cascade RPN: Delving into High-Quality Region Proposal Network with Adaptive Convolution. In *NeurIPS*, 2019. 4, 13

Wang, X., Huang, T. E., Darrell, T., Gonzalez, J. E., and Yu, F. Frustratingly Simple Few-Shot Object Detection. In *ICML*, 2020. 3, 6, 12, 15, 16, 18

Wang, X., Yang, X., Zhang, S., Li, Y., Feng, L., Fang, S., Lyu, C., Chen, K., and Zhang, W. Consistent-Teacher: Towards Reducing Inconsistent Pseudo-Targets in Semi-Supervised Object Detection. In *CVPR*, 2023. 2

Wu, J., Liu, S., Huang, D., and Wang, Y. Multi-Scale Positive Sample Refinement for Few-Shot Object Detection. In *ECCV*, 2020. 7

Xiong, W., Cui, Y., and Liu, L. Semi-Supervised Few-Shot Object Detection with a Teacher-Student Network. In *BMVC*, 2021. 3

Xu, M., Zhang, Z., Hu, H., Wang, J., Wang, L., Wei, F., Bai, X., and Liu, Z. End-to-End Semi-Supervised Object Detection with Soft Teacher. In *ICCV*, 2021. 2, 3, 5, 16

Yang, S., Dong, Y., Ward, R., Dhillon, I. S., Sanghavi, S., and Lei, Q. Sample Efficiency of Data Augmentation Consistency Regularization. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 206, pp. 3825–3853. PMLR, 2023. 5

Zhong, Z., Zheng, L., Kang, G., Li, S., and Yang, Y. Random Erasing Data Augmentation. In *AAAI*, 2020. 16

Table 5: Generalized FSOD results evaluated on VOC07 `test` over three random partitions. We compare our SoftER Teacher against its Soft Teacher counterpart and strong supervised baselines. We report the mean and 95% confidence interval over 10 random samples for our models. SoftER Teacher with ResNet-50 exceeds the supervised models with ResNet-101 by a large margin across most metrics under consideration.

VOC07 `test` – Split 1	Backbone	Base AP_{50}	Base AR_{50}	Base AP_{50} (15 Classes)			Novel AP_{50} (5 Classes)			Overall AP_{50} (20 Classes)		
Method				1-Shot	5-Shot	10-Shot	1-Shot	5-Shot	10-Shot	1-Shot	5-Shot	10-Shot
MPSR* [ECCV'20]	R-101	80.8	–	61.5	69.7	71.6	**42.8**	55.3	61.2	56.8	66.1	69.0
Retentive R-CNN* [CVPR'21]	R-101	80.8	–	80.9	80.8	80.8	42.4	53.7	56.1	71.3	74.0	74.6
TFA [ICML'20]	R-101	80.8	–	77.6 ± 0.2	77.4 ± 0.3	77.5 ± 0.2	25.3 ± 2.2	47.9 ± 1.2	52.8 ± 1.0	64.5 ± 0.6	70.1 ± 0.4	71.3 ± 0.3
SoftER Teacher (Ours)	R-101	**85.0**	**93.1**	**83.3 ± 0.4**	**84.6 ± 0.2**	**84.9 ± 0.2**	35.5 ± 3.7	**60.3 ± 2.8**	**65.9 ± 1.6**	**71.3 ± 1.2**	**78.5 ± 0.8**	**80.2 ± 0.4**
Faster R-CNN (Our Impl.)	R-50	81.7	88.0	82.0 ± 0.2	82.4 ± 0.1	82.3 ± 0.1	27.9 ± 3.2	52.1 ± 2.1	58.2 ± 1.6	68.5 ± 0.8	74.9 ± 0.5	76.2 ± 0.4
Soft Teacher (Our Impl.)	R-50	**85.0**	90.9	**84.6 ± 0.4**	**85.2 ± 0.1**	**85.0 ± 0.1**	29.5 ± 4.4	53.8 ± 2.6	60.8 ± 1.5	**70.8 ± 1.2**	77.4 ± 0.7	79.0 ± 0.3
SoftER Teacher (Ours)	R-50	84.7	**92.3**	83.1 ± 0.2	84.4 ± 0.2	84.6 ± 0.2	**31.3 ± 4.3**	**57.8 ± 2.6**	**62.9 ± 1.7**	70.2 ± 1.2	**77.8 ± 0.7**	**79.2 ± 0.4**
VOC07 `test` – Split 2	Backbone	Base AP_{50}	Base AR_{50}	Base AP_{50} (15 Classes)			Novel AP_{50} (5 Classes)			Overall AP_{50} (20 Classes)		
Method				1-Shot	5-Shot	10-Shot	1-Shot	5-Shot	10-Shot	1-Shot	5-Shot	10-Shot
MPSR* [ECCV'20]	R-101	81.9	–	60.8	71.2	72.7	**29.8**	**43.2**	47.0	53.1	64.2	66.3
Retentive R-CNN* [CVPR'21]	R-101	81.9	–	81.8	81.9	81.9	21.7	37.0	40.3	66.8	70.7	71.5
TFA [ICML'20]	R-101	81.9	–	73.8 ± 0.8	76.2 ± 0.4	76.9 ± 0.3	18.3 ± 2.4	34.1 ± 1.4	39.5 ± 1.1	59.9 ± 0.8	65.7 ± 0.5	67.6 ± 0.4
SoftER Teacher (Ours)	R-101	**85.9**	**93.1**	**84.9 ± 0.2**	**85.6 ± 0.2**	**85.9 ± 0.2**	23.8 ± 4.7	40.9 ± 2.4	**48.4 ± 2.1**	**69.6 ± 1.1**	**74.5 ± 0.7**	**76.5 ± 0.5**
Faster R-CNN (Our Impl.)	R-50	82.9	88.7	83.1 ± 0.1	83.5 ± 0.1	83.3 ± 0.1	18.3 ± 4.3	34.9 ± 1.5	40.6 ± 1.7	66.9 ± 1.1	71.4 ± 0.4	72.6 ± 0.4
Soft Teacher (Our Impl.)	R-50	85.3	91.5	**85.1 ± 0.2**	**85.3 ± 0.1**	**85.3 ± 0.1**	20.3 ± 4.7	37.9 ± 2.1	44.0 ± 1.8	**68.9 ± 1.2**	73.5 ± 0.6	75.0 ± 0.5
SoftER Teacher (Ours)	R-50	**85.3**	**93.1**	84.2 ± 0.2	85.1 ± 0.2	85.2 ± 0.2	**22.4 ± 4.3**	**41.1 ± 2.3**	**47.3 ± 2.1**	68.7 ± 1.1	**74.1 ± 0.7**	**75.8 ± 0.6**
VOC07 `test` – Split 3	Backbone	Base AP_{50}	Base AR_{50}	Base AP_{50} (15 Classes)			Novel AP_{50} (5 Classes)			Overall AP_{50} (20 Classes)		
Method				1-Shot	5-Shot	10-Shot	1-Shot	5-Shot	10-Shot	1-Shot	5-Shot	10-Shot
MPSR* [ECCV'20]	R-101	82.0	–	61.6	72.9	73.2	**35.9**	48.9	51.3	55.2	66.9	67.7
Retentive R-CNN* [CVPR'21]	R-101	82.0	–	81.9	82.0	82.1	30.2	**49.7**	50.1	69.0	73.9	74.1
TFA [ICML'20]	R-101	82.0	–	78.7 ± 0.2	78.5 ± 0.3	78.6 ± 0.2	17.9 ± 2.0	40.8 ± 1.4	45.6 ± 1.1	63.5 ± 0.6	69.1 ± 0.4	70.3 ± 0.4
SoftER Teacher (Ours)	R-101	**86.7**	**93.4**	**85.4 ± 0.3**	**86.4 ± 0.2**	**86.5 ± 0.1**	23.5 ± 2.8	46.6 ± 2.7	**53.8 ± 1.6**	**70.0 ± 0.8**	**76.5 ± 0.7**	**78.3 ± 0.4**
Faster R-CNN (Our Impl.)	R-50	82.6	88.0	83.1 ± 0.2	83.6 ± 0.1	83.3 ± 0.1	19.6 ± 1.9	44.1 ± 1.8	51.2 ± 1.3	67.3 ± 0.5	73.7 ± 0.4	75.3 ± 0.3
Soft Teacher (Our Impl.)	R-50	85.1	91.0	84.7 ± 0.2	85.2 ± 0.1	85.0 ± 0.1	22.2 ± 3.0	49.2 ± 2.2	55.6 ± 1.4	69.0 ± 0.9	76.2 ± 0.5	77.7 ± 0.3
SoftER Teacher (Ours)	R-50	**85.9**	**92.8**	**84.7 ± 0.2**	**85.4 ± 0.2**	**85.4 ± 0.2**	**24.4 ± 2.2**	**49.5 ± 1.8**	**55.9 ± 1.6**	**69.6 ± 0.7**	**76.4 ± 0.5**	**78.0 ± 0.4**

* indicates results were reported on a single sample run (hence the lack of error bars), which may have been over-estimated due to the high variance of few-shot training samples.

A Additional Quantitative Results

A.1 Generalized Few-Shot Detection on PASCAL VOC

We present the generalized FSOD results on VOC in Table 5, which comprises three random partition splits. We leverage `COCO-train2017` as *uncurated* unlabeled data to augment our experiments, which exhibits strong domain mismatch and contains many object classes outside of VOC. We report the ideal supervised base AP from previous work (Wang et al., 2020; Fan et al., 2021) along with our substantially improved semi-supervised base AP to measure the extent of base forgetting. These results further support our observation on the trade-off between novel performance and base forgetting, for which our approach aims to simultaneously optimize. We summarize the following key takeaways.

Base Performance Our re-implementation of the supervised Faster R-CNN baseline *does not* degrade base performance compared to the TFA benchmark across all three partitions. Base degradation is negligible with SoftER Teacher at less than 2%. We attribute this apparent improvement in base performance to our modified procedure of fine-tuning only the RoI box classifier and to our proposed Entropy Regression module enabling SoftER Teacher to achieve superior learning with unlabeled data.

SoftER Teacher *vs.* Supervised Baselines SoftER Teacher with ResNet-50 surpasses the supervised MPSR, TFA, and Retentive R-CNN models with ResNet-101 by a large margin on the combined overall base + novel AP metric across most experiments under consideration, while being more parameter-efficient with a smaller backbone. Although MPSR achieves impressive few-shot performance on novel categories, it suffers catastrophic base forgetting by as much as 26%. Retentive R-CNN does not exhibit base class degradation, but generally falls behind on novel class performance. To our knowledge, MPSR and Retentive R-CNN did not perform repeated experiments over multiple random seeds per established protocol, the results of which may have been over-estimated due to the high variance of few-shot training samples, and thus are not directly comparable to ours. In general, the previous work of TFA and ours observe marked reduction in novel performances with repeated sample runs (*e.g.*, 10 or 30) when compared to a single trial.

Table 6: Proposal quality is highly correlated with semi-supervised few-shot detection. SoftER Teacher produces the best proposal quality AR@p, for $p \in \{100, 300, 1000\}$, among the comparisons, which in turn yields the strongest novel k-shot performances given varying fractions of base labels. All models are equipped with the ResNet-101 backbone. We report the mean and standard deviation over 5 random samples.

Method	% Labeled	AR@100	AR@300	AR@1000	Base $AP_{50:95}$ (60 Classes)			Novel $AP_{50:95}$ (20 Classes)			Overall $AP_{50:95}$ (80 Classes)		
					5-Shot	10-Shot	30-Shot	5-Shot	10-Shot	30-Shot	5-Shot	10-Shot	30-Shot
Faster R-CNN	1	17.3 ± 0.1	22.0 ± 0.2	27.0 ± 0.4	9.8 ± 0.3	10.0 ± 0.4	10.8 ± 0.3	1.9 ± 0.3	2.7 ± 0.1	3.5 ± 0.1	7.8 ± 0.2	8.2 ± 0.3	9.0 ± 0.2
Soft Teacher	1	27.8 ± 0.8	32.4 ± 0.8	38.1 ± 0.9	19.4 ± 0.7	19.9 ± 0.8	21.2 ± 0.7	5.9 ± 0.8	7.9 ± 0.7	10.1 ± 0.5	16.0 ± 0.6	16.9 ± 0.7	18.4 ± 0.6
SoftER Teacher	1	28.9 ± 0.7	33.7 ± 0.6	39.4 ± 0.6	19.2 ± 0.6	19.8 ± 0.5	21.1 ± 0.5	6.7 ± 0.3	8.8 ± 0.2	10.8 ± 0.5	16.1 ± 0.5	17.1 ± 0.4	18.5 ± 0.5
Faster R-CNN	5	23.3 ± 0.3	28.7 ± 0.4	34.9 ± 0.5	18.5 ± 0.5	18.9 ± 0.3	20.0 ± 0.5	3.5 ± 0.2	4.6 ± 0.2	5.9 ± 0.3	14.8 ± 0.4	15.3 ± 0.2	16.5 ± 0.4
Soft Teacher	5	29.8 ± 0.2	35.2 ± 0.2	41.4 ± 0.3	27.5 ± 0.4	27.8 ± 0.5	29.2 ± 0.5	6.7 ± 0.7	8.9 ± 0.4	11.1 ± 0.3	22.3 ± 0.4	23.1 ± 0.3	24.7 ± 0.4
SoftER Teacher	5	30.5 ± 0.2	35.9 ± 0.2	42.0 ± 0.2	27.5 ± 0.4	27.9 ± 0.4	29.3 ± 0.2	7.9 ± 0.4	10.1 ± 0.5	12.4 ± 0.5	22.6 ± 0.3	23.4 ± 0.3	25.1 ± 0.2
Faster R-CNN	10	25.0 ± 0.2	30.7 ± 0.3	37.5 ± 0.3	22.6 ± 0.4	22.8 ± 0.1	24.2 ± 0.2	3.8 ± 0.5	5.3 ± 0.2	6.8 ± 0.2	17.9 ± 0.3	18.4 ± 0.1	19.9 ± 0.2
Soft Teacher	10	30.2 ± 0.2	35.9 ± 0.2	42.4 ± 0.2	30.5 ± 0.5	30.7 ± 0.4	32.1 ± 0.3	6.8 ± 0.3	9.0 ± 0.6	11.4 ± 0.3	24.6 ± 0.4	25.3 ± 0.4	26.9 ± 0.3
SoftER Teacher	10	31.1 ± 0.2	36.7 ± 0.2	43.1 ± 0.3	30.3 ± 0.5	30.6 ± 0.5	32.0 ± 0.4	7.9 ± 1.3	10.4 ± 1.1	12.9 ± 1.0	24.6 ± 0.1	25.6 ± 0.3	27.2 ± 0.3

Table 7: SSOD results on VOC07 `test`. `VOC0712` is the combined VOC07+12 `trainval` splits. `COCO-20` is the subset of `COCO-train2017` having the same 20 classes as VOC. SoftER Teacher outperforms Humble Teacher and Soft Teacher by a convincing margin.

Method	# Labels	Unlabeled	AP_{50}	$AP_{50:95}$	AR_{50}	$AR_{50:95}$
Supervised (Tang et al., 2021)	VOC07 (5K)	None	76.30	42.60	–	–
Supervised (Our Impl.)	VOC07 (5K)	None	**79.34**	**49.20**	**85.38**	**57.50**
Supervised (Tang et al., 2021)	VOC0712 (16K)	None	82.17	54.29	–	–
Supervised (Our Impl.)	VOC0712 (16K)	None	**84.53**	**57.77**	**89.73**	**65.73**
Humble Teacher [CVPR'21]	VOC07 (5K)	VOC12	80.94	**53.04**	–	–
Soft Teacher (Our Impl.)	VOC07 (5K)	VOC12	82.37	51.10	88.44	59.49
SoftER Teacher (Ours)	VOC07 (5K)	VOC12	**83.10**	51.26	**89.74**	**60.19**
Humble Teacher [CVPR'21]	VOC07 (5K)	VOC12 + COCO-20	81.29	54.41	–	–
Soft Teacher (Our Impl.)	VOC07 (5K)	VOC12 + COCO-20	82.50	54.47	87.14	62.45
SoftER Teacher (Ours)	VOC07 (5K)	VOC12 + COCO-20	**84.09**	**55.34**	**88.90**	**63.58**

Table 8: SSOD results on COCO `val2017`. The † setting refers to self-augmented supervised training without unlabeled data, and ‡ refers to the use of extra `unlabeled2017` images. We report the mean and standard deviation computed over 5 random samples.

COCO `val2017`	Average Precision ($AP_{50:95}$)				
Method	1%	5%	10%	†100%	‡100%
Supervised (Our Impl.)	10.57 ± 0.32	21.33 ± 0.40	26.80 ± 0.26	41.96	41.96
Humble Teacher [CVPR'21]	16.96 ± 0.38	27.70 ± 0.15	31.61 ± 0.28	–	42.37
Soft Teacher (Our Impl.)	21.38 ± 1.02	30.65 ± 0.19	33.95 ± 0.13	43.51	44.08
SoftER Teacher (Ours)	$\mathbf{21.93 \pm 0.90}$	$\mathbf{31.15 \pm 0.29}$	$\mathbf{34.08 \pm 0.05}$	**44.05**	**44.22**
Method	Average Recall ($AR_{50:95}$)				
	1%	5%	10%	†100%	‡100%
Supervised (Our Impl.)	15.87 ± 0.45	29.07 ± 0.47	36.80 ± 0.46	55.64	55.64
Soft Teacher (Our Impl.)	29.85 ± 0.89	38.68 ± 0.28	43.48 ± 0.25	55.66	56.18
SoftER Teacher (Ours)	$\mathbf{30.90 \pm 0.88}$	$\mathbf{39.60 \pm 0.41}$	$\mathbf{43.90 \pm 0.55}$	**56.06**	**56.22**

SoftER Teacher *vs.* Soft Teacher Although the strong Soft Teacher baseline is effective at harnessing unlabeled data for semi-supervised FSOD, our SoftER Teacher approach demonstrates superior learning by addressing a key shortcoming of Soft Teacher. SoftER Teacher consistently boosts object recall via our proposed Entropy Regression module, improving on Soft Teacher by as much as +1.8 base AR_{50}, which yields a gain of up to +4.0 novel AP_{50}. These results further bolster our empirical observation that a stronger semi-supervised detector leads to a more label-efficient few-shot detector.

A.2 The Impact of Proposal Quality on Semi-Supervised Few-Shot Detection

We present expansive results on proposal quality and its relationship with semi-supervised few-shot detection in Table 6. Following existing literature (Hosang et al., 2016; Vu et al., 2019), we measure proposal quality using the metric AR@p, for $p \in \{100, 300, 1000\}$ proposals, averaged over 10 overlap thresholds between 0.5 and 0.95. Proposal quality AR@p is not to be confused with the detection metric $AR_{50:95}$, which is used to evaluate object coverage computed on a per-category basis and averaged over categories.

A.3 SoftER Teacher Improves Precision and Recall for Semi-Supervised Detection

We present SSOD results for VOC and COCO in Tables 7 and 8, respectively. We leverage `COCO-20` and `COCO-unlabeled2017` as unlabeled images on VOC and COCO experiments, respectively, to compare with existing work. On both datasets, we re-implement and re-train the supervised and Soft Teacher models for a direct comparison with SoftER Teacher. As part of our re-implementation, we make a conscientious effort to obtain high-quality supervised and Soft Teacher baselines by maximizing their performance output. This is to ensure that any performance boost demonstrated by SoftER Teacher is directly attributed to our entropy regression module for proposal learning with complex affine transforms.

In Table 7, we compare our best-case supervised baselines to those trained by Humble Teacher (Tang et al., 2021) and show that ours achieve significantly better detection accuracy. Even in the presence of strong supervised and

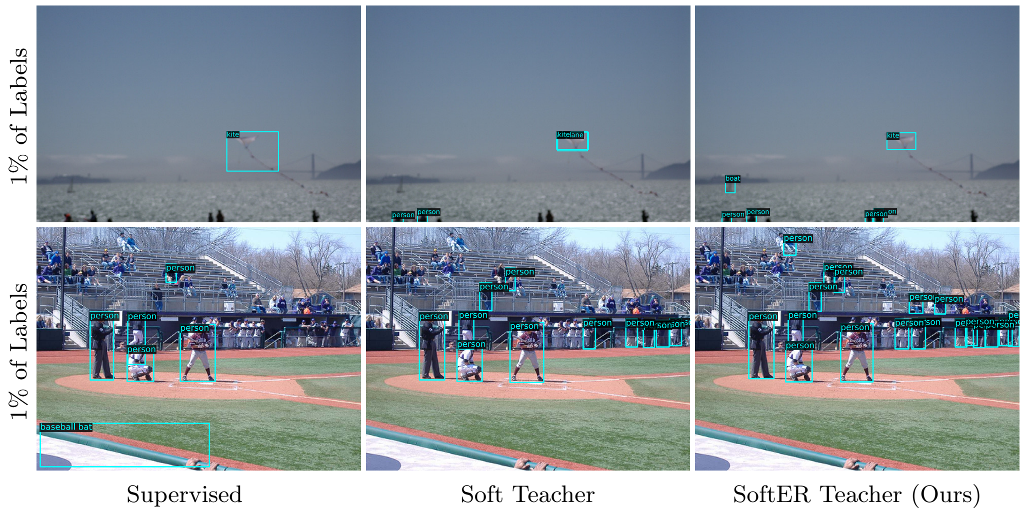


Figure 7: Qualitative detections on COCO `val2017` from models trained on 1% of labels. SoftER Teacher improves on both precision and recall, by recovering more missed objects while making fewer false positive detections, over its corresponding supervised and Soft Teacher counterparts. Best viewed digitally.

Table 9: FSOD results evaluated on COCO `val2017`. We report the mean and 95% confidence interval over 5 random samples for our models. SoftER Teacher with ResNet-50 surpasses TFA with ResNet-101 on both base and novel performances while also uniformly outperforming its Soft Teacher counterpart across all experiments.

COCO `val2017`	Backbone	Base $AP_{50:95}$	Base $AR_{50:95}$	Base $AP_{50:95}$ (60 Classes)				Novel $AP_{50:95}$ (20 Classes)			
Method				1-Shot	5-Shot	10-Shot	30-Shot	1-Shot	5-Shot	10-Shot	30-Shot
TFA w/cos	R-101	39.3	–	31.9 ± 0.7	32.3 ± 0.6	32.4 ± 0.6	34.2 ± 0.4	1.9 ± 0.4	7.0 ± 0.7	9.1 ± 0.5	12.1 ± 0.4
Faster R-CNN (Our Impl.)	R-50	39.3	53.0	34.4 ± 0.6	33.1 ± 0.2	33.2 ± 0.2	35.1 ± 0.3	1.0 ± 0.3	5.1 ± 0.4	7.2 ± 0.4	9.6 ± 0.2
Soft Teacher (Our Impl.)	R-50	41.3	52.8	37.6 ± 0.4	37.0 ± 0.1	36.8 ± 0.3	38.2 ± 0.3	1.7 ± 0.9	6.7 ± 0.4	8.8 ± 0.5	11.2 ± 0.4
SoftER Teacher (Ours)	R-50	42.0	54.4	38.0 ± 0.4	37.4 ± 0.2	37.4 ± 0.2	38.7 ± 0.2	2.4 ± 0.6	8.2 ± 0.3	10.3 ± 0.5	12.9 ± 0.6
SoftER Teacher (Ours)	R-101	**44.4**	**56.1**	$\mathbf{40.7 \pm 0.3}$	$\mathbf{40.3 \pm 0.2}$	$\mathbf{40.2 \pm 0.3}$	$\mathbf{41.4 \pm 0.2}$	$\mathbf{2.8 \pm 0.7}$	$\mathbf{8.7 \pm 0.6}$	$\mathbf{11.0 \pm 0.4}$	$\mathbf{14.0 \pm 0.6}$

semi-supervised baselines, our SoftER Teacher model continues to improve upon its counterparts across almost all AP and AR metrics. Notably, our approach demonstrates superior learning with unlabeled data by narrowing the gap to less than 0.5 AP_{50} between the fully supervised model trained on VOC07+12 (16K labels) and SoftER Teacher trained on VOC07 (5K labels) augmented with unlabeled images from VOC12+COCO-20.

In Table 8, our model consistently outperforms its Soft Teacher counterpart over varying fractions of labeled data, although the impact of proposal learning in SoftER Teacher diminishes as the percentage of labeled data increases. We also experiment with 100% labels, *i.e.*, the entire `train2017` set, in two settings. In the first setting without unlabeled data, referred to as *self-augmented supervised training*, we use the `train2017` set as the source of "unlabeled data" to generate pseudo targets. And in the second setting, we supplement supervised training with `unlabeled2017` images. We observe that even without unlabeled data, SoftER Teacher improves on the supervised baseline by +2.09 AP, suggesting that more representations can still be learned from `train2017` alone. In the setting with additional unlabeled data, our model further boosts accuracy by another +0.17 AP.

Figure 7 visualizes exemplar detections from models trained on 1% of COCO labels, wherein our SoftER Teacher improves on both precision and recall over the comparisons.

A.4 Generalized Few-Shot Detection on MS-COCO

We present additional FSOD results on the COCO dataset to include 1-shot detection in Table 9, leveraging `COCO-unlabeled2017` as supplementary unlabeled data. Here, we observe more supporting evidence to strengthen our empirical finding on the potential relationship between SSOD and FSOD to suggest that a stronger semi-supervised detector leads to a more label-efficient few-shot detector. SoftER Teacher uniformly outperforms Soft Teacher across all metrics and experiments under consideration, most notably on novel class detection.

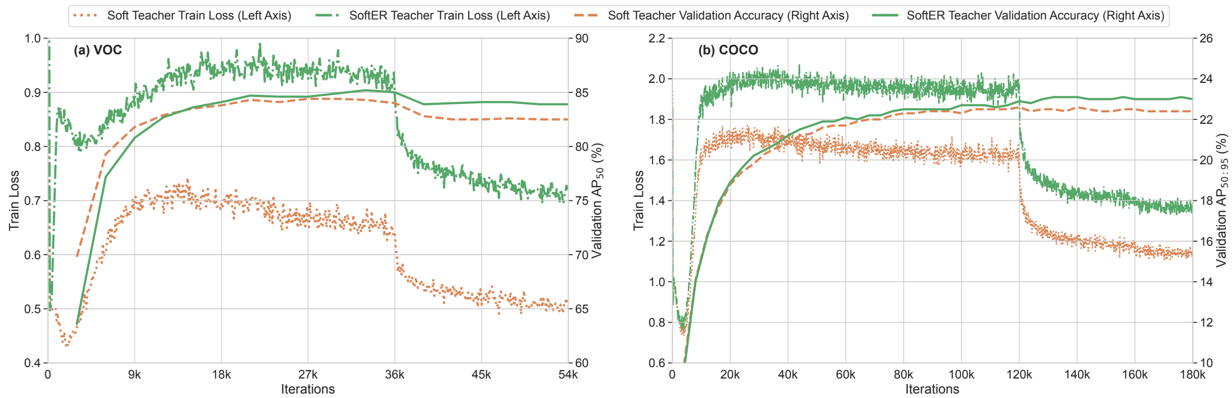


Figure 8: Visualization of training and validation behavior of Soft Teacher and SoftER Teacher on **(a)** VOC07 and **(b)** 1% of COCO labels. **Left:** The validation curve of Soft Teacher tends to overfit more than SoftER Teacher toward the end of training on VOC. **Right:** SoftER Teacher continues to improve even when Soft Teacher has reached its validation performance plateau at the 120k iterations mark.

Table 10: Ablation experiments evaluated on COCO `val2017` showing the standard procedure of fine-tuning both box classification and regression heads degrades base performance by as much as 21%. Our modified protocol of fine-tuning only the box classifier, while keeping the box regressor fixed, helps retain base detection accuracy with a performance drop of less than 11% for Faster R-CNN and 9% for SoftER Teacher.

Method	Base $AP_{50:95}$	Base $AP_{50:95}$ (60 Classes)				Novel $AP_{50:95}$ (20 Classes)			
		1-Shot	5-Shot	10-Shot	30-Shot	1-Shot	5-Shot	10-Shot	30-Shot
Faster R-CNN (fine-tune `cls+reg`)	39.3	31.2 (↓ 21%)	34.7 (↓ 12%)	34.8 (↓ 11%)	36.7 (↓ 7%)	0.6	3.9	6.0	7.9
Faster R-CNN (fine-tune `cls` only)	39.3	34.9 (↓ 11%)	35.8 (↓ 9%)	35.8 (↓ 9%)	37.1 (↓ 6%)	0.5	3.9	6.1	7.8
SoftER Teacher (fine-tune `cls+reg`)	42.0	33.6 (↓ 20%)	37.8 (↓ 10%)	38.1 (↓ 9%)	39.9 (↓ 5%)	1.5	6.7	9.4	10.8
SoftER Teacher (fine-tune `cls` only)	42.0	38.3 (↓ 9%)	39.1 (↓ 7%)	39.1 (↓ 7%)	40.2 (↓ 4%)	1.5	6.7	9.4	10.8

A.5 SoftER Teacher is Less Prone to Overfitting

We analyze the training behavior of Soft Teacher and SoftER Teacher for semi-supervised detection in Figure 8. For VOC, we train both models on VOC07 `trainval` labels with supplementary unlabeled images from VOC12+COCO-20. We observe from the validation curves that Soft Teacher seems to train faster than SoftER Teacher at the beginning, but has the propensity to overfit more than SoftER Teacher toward the end of training. For COCO, we train on 1% of labels sampled from `train2017` with the remaining 99% as unlabeled data. Similarly, we see from the validation curves that SoftER Teacher continues to improve even when Soft Teacher has reached its performance plateau. We attribute these characteristics to our entropy regression module for proposal learning, which provides SoftER Teacher a degree of robustness against overfitting.

A.6 To Freeze or Not to Freeze Box Regressor

The standard two-stage transfer learning procedure (Wang et al., 2020) fine-tunes the few-shot detector by updating both the RoI box classifier and regressor while keeping everything else frozen. Intuitively, we expect the RPN to produce accurate object regions during base pre-training, especially in the semi-supervised setting where it is further boosted by supplementary unlabeled images. We postulate that only the box classifier needs to be updated during fine-tuning to adapt base representations to novel concepts, and that fine-tuning the regression head is not necessary and may even hurt base performance. Table 10 verifies our intuition that fine-tuning both box classification and regression heads degrades base performance by as much as 21% on COCO `val2017`. By comparison, our modified protocol of fine-tuning only the box classifier helps retain base performance with a drop of less than 11%. Novel performance is unaffected between the two configurations. Our results are corroborated by existing work confirming that the main source of error with FSOD is indeed associated with the box classifier (Sun et al., 2021; Gao et al., 2022). Recall our goal for FSOD is to maximize novel detection accuracy while minimizing base performance degradation; keeping the box localization parameters fixed during fine-tuning is a simple and straight-forward way to help maintain base class accuracy.

Table 11: Summary of the data augmentation pipelines used to train Soft Teacher and SoftER Teacher. **Left:** transformations applied to the student trained on labeled data. **Middle:** strong augmentation pipeline, including complex affine transforms and cutout, applied to the student trained on unlabeled data. **Right:** weak augmentation pipeline applied to the teacher trained on unlabeled data.

Augmentation	Student Labeled Branch	Student Unlabeled Branch (Strong)	Teacher Unlabeled Branch (Weak)
Resize	short edge $\in [400, 1200]$	short edge $\in [400, 1200]$	short edge $\in [400, 1200]$
Flip	$p = 0.5$, horizontal	$p = 0.5$, horizontal	$p = 0.5$, horizontal
Identity	$p = 1/9$	$p = 1/9$	
AutoContrast	$p = 1/9$	$p = 1/9$	
Equalize	$p = 1/9$	$p = 1/9$	
Solarize	$p = 1/9$	$p = 1/9$	
Color	$p = 1/9$	$p = 1/9$	
Contrast	$p = 1/9$	$p = 1/9$	
Brightness	$p = 1/9$	$p = 1/9$	
Sharpness	$p = 1/9$	$p = 1/9$	
Posterize	$p = 1/9$	$p = 1/9$	
Translation		$p = 1/3$, $(x, y) \in (-0.1, 0.1)$	
Shearing		$p = 1/3$, $(x, y) \in (-30°, 30°)$	
Rotation		$p = 1/3$, angle $\in (-30°, 30°)$	
Cutout		$n \in [1, 5]$, size $\in [0.0, 0.2]$	

B Implementation Details

B.1 Few-Shot Datasets and Benchmarks

In this work, we evaluate our approach on the PASCAL VOC and MS-COCO 2017 detection benchmarks. For COCO experiments, the established evaluation protocol is on COCO 2014 following the TFA benchmark. However, both COCO 2014 and 2017 share the same images. The only difference between the two is the number of validation images (41k images for COCO `val2014` and 5k for COCO `val2017`). Wang et al. (2020) created the TFA benchmark by sampling from COCO 2014 a random subset of 5k images for validation and used the rest in the training split. Thus, both train/val splits from COCO 2014 and 2017 should effectively be the same, with minor variance due to the sampling process. Our preliminary experiments on both COCO 2014 (following the TFA splits) and the official COCO 2017 splits verified that the difference is indeed small with less than ± 0.3 AP. The benefit of using the official COCO 2017 splits is to remove the dependency on the random train/val splits created by the TFA benchmark and to maintain uniformity with our proposed semi-supervised FSOD benchmark in Table 3. For VOC experiments, we use the same sample splits taken from the TFA benchmark without changes.

B.2 Data Augmentation

We summarize the data augmentation strategy used to train Soft Teacher and SoftER Teacher in Table 11. There are three pipelines or branches of augmentation. The labeled branch uses random resizing and horizontal flipping along with color transformations. The student detector of the unlabeled branch undergoes the full complement of augmentations including strong affine geometric transformations and cutout (DeVries & Taylor, 2017; Zhong et al., 2020), akin to RandAugment (Cubuk et al., 2020), whereas the teacher detector leverages only weak resizing and horizontal flipping. At test time, we resize all instances to the shorter side of 800 resolution, but otherwise do not perform any test-time augmentation, following standard supervised and semi-supervised protocols.

B.3 Supervised and Semi-Supervised Training

High-Quality Baselines Following existing literature (Sohn et al., 2020b; Liu et al., 2021; Tang et al., 2021; Xu et al., 2021), we evaluate our approach for semi-supervised detection on VOC and COCO 2017 datasets. On both datasets, we re-implement and re-train the supervised Faster R-CNN and Soft Teacher[1] models for a direct comparison with SoftER Teacher. As part of our re-implementation, we make a conscientious effort to

[1] We leverage the original authors' source code made publicly available at `https://github.com/microsoft/SoftTeacher`.

Table 12: Supervised and semi-supervised training protocols on PASCAL VOC. `COCO-20` is the subset of `COCO-train2017` containing objects with the same 20 category names as VOC objects. `Sample Ratio` denotes the blend of (labeled, unlabeled) examples in a mini-batch. All settings are configured for 8× multi-GPU training.

Method	Labeled	Unlabeled	Batch Size	Sample Ratio	*lr*	*lr* Step	Iterations
Supervised	VOC07	None	16	(16, 0)	0.02	(12k, 16k)	18k
Supervised	VOC0712		16	(16, 0)	0.02	(36k, 48k)	54k
Soft Teacher	VOC07	VOC12	64	(32, 32)	0.01	(12k, 16k)	18k
SoftER Teacher			64	(32, 32)	0.01	(12k, 16k)	18k
Soft Teacher	VOC07	VOC12+ COCO-20	64	(32, 32)	0.01	(36k, 48k)	54k
SoftER Teacher			64	(32, 32)	0.01	(36k, 48k)	54k
Soft Teacher	VOC0712	COCO-20	64	(32, 32)	0.01	(40k, 52k)	60k
SoftER Teacher			64	(32, 32)	0.01	(40k, 52k)	60k

Table 13: Supervised and semi-supervised training protocols on COCO 2017. The † setting refers to self-augmented supervised training without unlabeled data, and ‡ corresponds to the use of supplementary `unlabeled2017` images. `Sample Ratio` denotes the blend of (labeled, unlabeled) examples in a mini-batch. All settings are configured for 8× multi-GPU training.

% Labeled	Method	Batch Size	Sample Ratio	*lr*	*lr* Step	Iterations
1	Supervised	8	(8, 0)	0.01	(120k, 160k)	180k
	Soft Teacher	40	(8, 32)	0.01	(120k, 160k)	180k
	SoftER Teacher	40	(8, 32)	0.01	(120k, 160k)	180k
5	Supervised	8	(8, 0)	0.01	(120k, 160k)	180k
	Soft Teacher	40	(8, 32)	0.01	(120k, 160k)	180k
	SoftER Teacher	40	(8, 32)	0.01	(120k, 160k)	180k
10	Supervised	8	(8, 0)	0.01	(120k, 160k)	180k
	Soft Teacher	40	(8, 32)	0.01	(120k, 160k)	180k
	SoftER Teacher	40	(8, 32)	0.01	(120k, 160k)	180k
†100	Supervised	16	(16, 0)	0.02	(480k, 640k)	720k
	Soft Teacher	64	(32, 32)	0.01	(480k, 640k)	720k
	SoftER Teacher	64	(32, 32)	0.01	(480k, 640k)	720k
‡100	Supervised	16	(16, 0)	0.02	(480k, 640k)	720k
	Soft Teacher	64	(32, 32)	0.01	(480k, 640k)	720k
	SoftER Teacher	64	(32, 32)	0.01	(480k, 640k)	720k

obtain the best-case supervised and Soft Teacher baselines by maximizing their performance output. We train the strong supervised baseline by using a longer training schedule (see Tables 12 and 13) and applying diverse color augmentations in addition to random resizing and horizontal flipping (see Table 11). And we re-train Soft Teacher exactly as is according to the authors' source code. This is to ensure that any performance boost demonstrated by SoftER Teacher is directly attributed to our entropy regression module for learning representations from region proposals, and not to changes in model configuration and training protocol.

VOC Evaluation We experiment with two supervised settings: (1) using VOC07 `trainval` split as labeled data, and (2) utilizing the joint VOC07+12 labeled set as an upper bound for supervised detection performance. We also have two semi-supervised settings: (1) augmenting supervised training on VOC07 with VOC12 as unlabeled data, and (2) leveraging the combined VOC12+COCO-20 as unlabeled data. `COCO-20` is the subset of `COCO-train2017` having the same 20 category names as VOC objects. Model performance is evaluated on the VOC07 `test` set. Detailed comparative results are given in Table 7.

COCO Evaluation There are three experimental settings: (1) *Partially labeled*, where we train on $\{1, 5, 10\}$ percent of labels randomly sampled from the `train2017` split while treating the remaining images as unlabeled data. (2) *Fully labeled*, where we leverage the extra 123k images from the `unlabeled2017` set to supplement supervised training on the entire `train2017`. And (3) *Self-augmented supervised training*, where we apply the `train2017` set, discarding all label information, as the source of "unlabeled" data to generate unsupervised pseudo targets. For each setting, we also train on the labeled portion alone, without using unlabeled data, to establish the lower-bound supervised baseline. Model performance is evaluated on the `val2017` set. See Table 8 for detailed comparative results.

Top-N Proposals To learn representations on region proposals, we extract the top 512 proposals, after non-maximum suppression, from each unlabeled image as generated by the student's RPN. Our motivation for selecting the top 512 proposals is to balance the trade-off among accuracy performance, memory requirements, and training duration. Moreover, our choice of $N = 512$ is consistent with $N = 640$ proposals empirically found by Humble Teacher (Tang et al., 2021) to be an optimal number with regards to detection accuracy.

Training Parameters We summarize our training protocols on VOC and COCO in Tables 12 and 13 for the supervised, Soft Teacher, and SoftER Teacher models. In general, Soft Teacher and our SoftER Teacher share the same configuration to ensure we can directly measure the impact of proposal learning and its contribution to detection accuracy. All hyper-parameters related to Soft Teacher remain the same, including the EMA momentum, which defaults to 0.999 following common practice in the semi-supervised classification literature (Tarvainen & Valpola, 2017; Sohn et al., 2020a). We implement our models in MMDetection (Chen et al., 2019) and PyTorch (Paszke et al., 2019), and train them using vanilla SGD optimization with momentum and weight decay set to 0.9 and 0.0001, respectively. We train on 8× A6000 GPUs each with 48GB of memory. One experiment takes between 12 hours and 10 days to complete, depending on the scope. At test time, we extract the teacher model from the final check-point for evaluation.

B.4 Semi-Supervised Few-Shot Training

In this section, we expound on our protocol for semi-supervised few-shot training on VOC and COCO datasets. We conduct our few-shot experiments following the TFA benchmark (Wang et al., 2020). The VOC dataset is randomly partitioned into 15 base and 5 novel classes, where there are $k \in \{1, 5, 10\}$ labeled boxes per category sampled from the combined VOC07+12 `trainval` splits. This process is repeated three times to create three partitions. And the COCO dataset is divided into 60 base and 20 novel classes having the same VOC category names with $k \in \{1, 5, 10, 30\}$ shots. We leverage `COCO-train2017` as the source of external unlabeled data to supplement few-shot training on VOC, and `COCO-unlabeled2017` images to augment experiments on COCO.

Semi-Supervised Base Pre-Training In the first stage, we train a base detector on base classes, along with the available unlabeled data, according to the formulation described in Section 3.2. For the supervised base detector, we equip Faster R-CNN with the ResNet-101 backbone. For the semi-supervised base detectors, we experiment with Soft Teacher and our proposed SoftER Teacher using the same ResNet-101 backbone. In some experiments, we also employ ResNet-50 to explore parameter-efficient learning with SoftER Teacher. The incorporation of unlabeled data in the base pre-training step achieves two benefits of practical significance in real-world settings. First, our versatile approach helps make the deployment of few-shot applications easier by not strictly depending on an abundance of labels, which is a fundamental limitation of prior work. Second, unlabeled images have the remarkable ability to advance FSOD by way of region proposals, as demonstrated in Sections 3.1 and 4.2, enabling SoftER Teacher to learn more efficiently with reduced labels for both base and novel classes.

Semi-Supervised Few-Shot Fine-Tuning In the second stage, we combine the parameters of the (semi-supervised) base detector with those of the novel detector into the overall few-shot detector and fine-tune it on a small balanced training set of k shots per class containing both base and novel examples. Before fine-tuning, we obtain the parameters of the novel detector in two ways. For VOC, we initialize the parameters of the novel classifier and regressor with normally distributed random values, analogous to TFA. For the COCO dataset, we reuse the base model pre-trained in the first stage, but further train the detector head from scratch on novel classes. We optimize the novel detector on both few-shot and unlabeled examples according to the semi-supervised protocols. At the fine-tuning step, we update only the RoI box classifier of the few-shot detector while freezing all other components, including the box regressor. We justify our approach and design choices with detailed ablation studies presented in Section 5. Table 14 summarizes our few-shot fine-tuning protocol.

Table 14: Protocol for few-shot fine-tuning on VOC and COCO datasets. All settings are configured for 8× multi-GPU training.

# Shot	Parameter	VOC07+12	COCO 2017
1	Batch Size	16	16
	lr	0.001	0.001
	lr Step	9k	14k
	Iterations	10k	16k
	Fine-Tune Layer	`cls+reg`	`cls`
5	Batch Size	16	16
	lr	0.001	0.001
	lr Step	18k	72k
	Iterations	20k	80k
	Fine-Tune Layer	`cls+reg`	`cls`
10	Batch Size	16	16
	lr	0.001	0.001
	lr Step	36k	144k
	Iterations	40k	160k
	Fine-Tune Layer	`cls+reg`	`cls`
30	Batch Size	–	16
	lr	–	0.001
	lr Step	–	216k
	Iterations	–	240k
	Fine-Tune Layer	–	`cls`

C Additional Qualitative Results

Figure 9: Visualizations of student-teacher proposals with confidence scores ≥ 0.99. A pair of student-teacher proposals is aligned between student and teacher images for the purpose of enforcing classification similarity and localization consistency. The student images are subjected to a wide spectrum of complex scale, color, and geometric distortions, whereas the teacher images undergo simple random resizing and horizontal flipping transformations as the basis for generating reliable unsupervised pseudo targets to regularize the student's learning trajectory. This multi-stream data augmentation strategy enables the student to tap into abundant region proposals to capture diverse feature representations that would otherwise be lost with aggressive confidence thresholding associated with pseudo-labeling. Best viewed digitally.

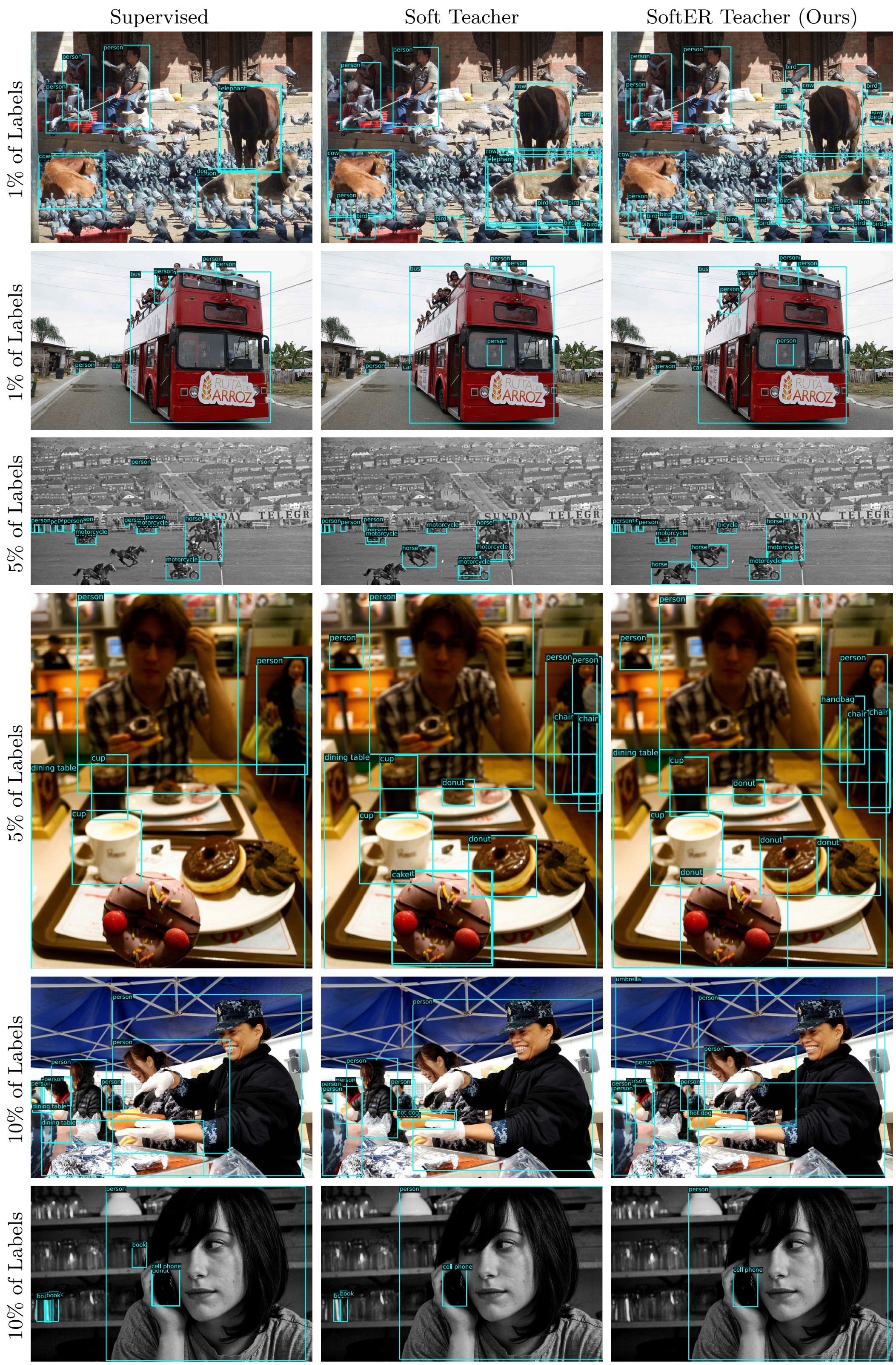


Figure 10: Exemplar detections from models trained on $\{1, 5, 10\}$ percent of labels sampled from COCO `train2017` and visualized on `val2017`. SoftER Teacher captures more object coverage while making fewer false positive detections than its supervised and Soft Teacher counterparts. The enhancements over Soft Teacher are especially pronounced in low-label settings and in crowded scenes with small and ambiguous objects. Best viewed digitally.