FoxInst: A Frustratingly Simple Baseline for Weakly Few-shot Instance Segmentation

Anonymous authors
Paper under double-blind review

Abstract

We propose the first weakly-supervised few-shot instance segmentation task and a frustratingly simple but strong baseline model, FoxInst. Our work is distinguished from other approaches in that our method is trained with weak annotations, *i.e.,* class and box annotations, during all phases, which leads to further data efficiency and practicality. Considering the challenging regime of our problem, we design the network to be an anchor-free architecture to avoid anchor box restriction, and train the network in a simple and stable way that first trains the whole network on the base classes, and then only fine-tunes the heads partially with few novel class data. To establish the foundation as a strong baseline, we carefully design evaluation setups by correcting the existing problems in the evaluation metric and test set, so that the effects of each component are well revealed. We show that FoxInst achieves comparable or even higher performance with the prior fully-supervised FSIS networks on COCO and PASCAL VOC datasets. We will release the code if accepted for reproduction.

1 Introduction

Instance segmentation (Hariharan et al., 2014; He et al., 2017) is a fundamental and versatile computer vision task that jointly tackles detection, classification, and semantic segmentation of each instance in an image. This provides a comprehensive analysis of a scene and can be usefully leveraged in subsequent applications (Zhang et al., 2016; Chen et al., 2016; Waqas Zamir et al., 2019; Ge et al., 2019). However, its applicability is limited because most instance segmentation methods (He et al., 2017; Tian et al., 2020; Bolya et al., 2019; Lee & Park, 2020; Chen et al., 2020) that have been proposed focus only on the limited closed-set of classes and are trained in a fully-supervised way with abundant labeled data. It is often impractical to apply them to a real-world deployment scenario where new classes need to be dealt with only few data but even without mask annotations, *i.e.,* a data-hungry setting.

Few-shot instance segmentation (FSIS) (Michaelis et al., 2018; Yan et al., 2019; Fan et al., 2020; Ganea et al., 2021; Nguyen & Todorovic, 2021) is proposed to address those challenges, which predicts instance-wise masks of novel objects by adapting a pre-trained network with scarce data. This appears to be annotation-efficient but still requires mask annotations, which are often costly to obtain. In a different context, few-shot semantic segmentation methods (Raza et al., 2019; Wang et al., 2019; Zhang et al., 2019) attempt to overcome this problem by using weak labels as supervision, *i.e.,* image-level tags or box annotations, without semantic mask labels. However, those approaches are still limited in data efficiency in that they leverage weak annotations only during the test phase and require full supervision in the base training phase, which uses much more data than the test phase. In addition, it is non-trivial to extend those approaches to few-shot instance segmentation.

In this regard, we first define a data-efficient task, weakly-supervised FSIS, and propose its frustratingly simple yet strong baseline, called FoxInst (See Fig. 1). Distinctiveness of our FoxInst is to require only weak labels, i.e., class label and bounding box per instance, for the entire training and test phase. With these weak labels, we train a few-shot instance segmentation network in a simple way that first trains the whole network on the base classes, and then only fine-tunes the prediction heads (the red parts in Fig. 1) with few novel class data. This simple design choice is motivated by the recent analysis by Raghu et al. (2020). Raghu et al. systematically analyze a popular few-shot meta-learning algorithm, MAML (Finn et al., 2017), and find that its success is due to feature reuse,

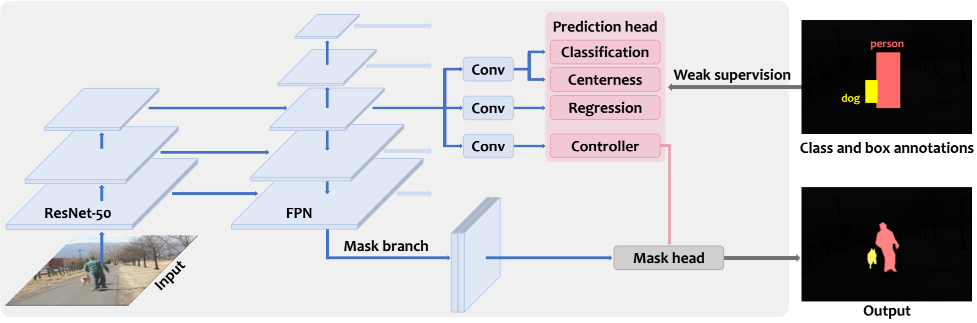


Figure 1: The architecture of FoxInst. In the base training phase, the entire network is trained on the training set of the base classes. In the few-shot fine-tuning stage, the backbone and mask branch (blue) are frozen while the four prediction heads (red) are fine-tuned with few data points sampled from the novel classes. Note that FoxInst is trained with class and box annotations as weak supervision in the whole training and fine-tuning phases.

not fast model adaption capability to novel tasks. This allows us to simplify the meta-learning and the model adaptation on novel classes to pre-training on the base classes and simple fine-tuning on novel classes, respectively, *i.e.,* two-stage fine-tuning (TFA; Wang et al., 2020). Hence, our method can be regarded as its extension to the weakly-supervised regime with mask prediction.

In addition, weak labels we use can be regarded as noisy labels; thus, our task is more challenging than FSIS and weakly-supervised instance segmentation. To derive a reliable and robust model, we adopt an anchor-free architecture. It is discussed that the anchor-based networks are often prone to overfitting and biased to specific sizes (Tian et al., 2019; Nguyen & Todorovic, 2021). This effect can be stand out in the few-shot regime, where only few training samples are given for the target object. FoxInst can avoid this problem with the anchor-free property. With these design choices, despite using only box annotations, FoxInst achieves comparable or even superior performance to the prior FSIS models trained with full-supervision, i.e., instance-wise masks, on the various experimental settings: COCO benchmark and PASCAL VOC. In particular, in the cross-dataset setting using COCO and VOC, FoxInst also outperforms the fully-supervised state-of-the-art with a large gap.

Lastly, since the weakly-supervised FSIS baselines are proposed for the first, it is important to analyze and identify the current bottleneck and status of development in the weakly-supervised FSIS regime, so that we can suggest the right future directions for subsequent research. We find out two issues on the evaluation metric and test data. For the metric, the widely used metrics in instance segmentation are entangled with the classification and mask (or localization) errors. Also, the existing test split in the popular benchmarks, COCO (Lin et al., 2015), is designed to have notable false positive bias leading to incorrect evaluation. To assess the details of components, we devise an alternative metric to assess separate effects and propose a new split of the original test set.

To summarize, our main contributions are as follows:

- We first define a data-efficient task, weakly-supervised few-shot instance segmentation, and propose its first baseline called FoxInst, which is a simple, efficient, and anchor-free model.
- For more analytical evaluation, we deploy a metric, FG-AP, to separate the measurement of detection and segmentation performance from the classification dependency. We also construct a refined test split which only includes images having at least one novel object instance.

2 Related Work

Few-shot instance segmentation (FSIS). Few-shot learning is an adaptation method that uses few examples belonging to unseen classes during training. Approaches to tackle few-shot learning can be categorized into metric-based (Vinyals et al., 2016; Snell et al., 2017; Sung et al., 2018; Gidaris & Komodakis, 2018) and optimization-based methods (Finn et al., 2017; Nichol & Schulman, 2018; Rusu et al., 2019; Antoniou et al., 2019; Grant et al., 2019). Recently, Chen et al. (2019) and Dhillon et al. (2020) show that simple fine-tuning is surprisingly effective in few-shot tasks.

Since the aforementioned techniques have been primarily developed for simple classification problems, extending the ideas to the complicated FSIS tasks is non-trivial. Only a few works (Michaelis et al., 2018; Yan et al., 2019; Fan et al., 2020; Ganea et al., 2021; Nguyen & Todorovic, 2021) tackle few-shot regimes of instance segmentation. FGN (Fan et al., 2020) proposes attention-based guide modules to deal with FSIS by episodic training (Vinyals et al., 2016) but because of the heavy memory requirement of episodic training, it can only consider a limited number of novel classes. MTFA (Ganea et al., 2021) is free from the memory issue by virtue of the fine-tuning adaptation to the novel classes. We take this strategy to leverage its simple and stable learning properties. Also, this line of approaches requires full supervision, *i.e.,* instance segmentation masks. In this respect, improving annotation efficiency using weak labels is a consequential research direction.

Weakly-supervised instance segmentation. Obtaining delicate mask annotations is laborious in terms of both time and cost. To reduce the dependency on this costly annotation, weakly-supervised instance segmentation (Hu et al., 2018; Zhou et al., 2016; 2018; Cholakkal et al., 2019; Hsu et al., 2019; Arun et al., 2020; Liu et al., 2020; Hwang et al., 2021; Lee et al., 2021) is proposed, which leverages image-level tags or box annotations as weak supervision. Specifically, a bounding box can be exploited to compute weakly-supervised loss functions and refine the localization using local color similarity (Tian et al., 2021). Our method further extends the annotation efficiency beyond these weakly-supervised methods to few-shot regimes.

Object localization and instance segmentation. A key module of instance segmentation is object localization which separates each instance from multiple objects and background. Object localization has been developed in either an anchor-based or anchor-free way. The most famous anchor-based model is Faster R-CNN (Ren et al., 2015) that predicts the bounding box locations based on the region proposal network. It is extended to instance segmentation by adding the mask branch (He et al., 2017). However, using anchor boxes is known to be computationally inefficient, and the existence of anchor boxes limits the flexibility of proposals, *e.g.,* scale and aspect ratio, which may lead to performance degradation or overfitting (Tian et al., 2019; Yang et al., 2018; Zhang et al., 2020).

On the other hand, anchor-free networks (Tian et al., 2019; Bolya et al., 2019; Chen et al., 2020; Lee & Park, 2020) predict multiple objects directly without an external proposal module. CondInst (Tian et al., 2020) employs a dynamic mask head and obtains instance-wise masks from a single feature map. In this work, we use the anchor-free architecture to avoid the problem of anchors in the FSIS setup, where it is more prone to overfitting and requires generalization ability.

3 Method

We first define the problem that this work aims to tackle in Sec. 3.1. In Sec. 3.2, we introduce how we build a novel weakly-supervised few-shot instance segmentation method.

3.1 Problem Definition

We mainly follow the few-shot instance segmentation setting introduced by (Ganea et al., 2021). The entire dataset $\mathcal{D}$ is divided into three sets: training $\mathcal{D}_{\text{base}}$, fine-tuning $\mathcal{D}_{\texttt{fine}}$, and query $\mathcal{D}_q$ sets. The set $\mathcal{D}_{\text{base}}$ contains a large number of examples of base classes $\mathcal{C}_{\text{base}}$, $\mathcal{D}_{\texttt{fine}}$ has few NK number of data of novel classes $\mathcal{C}_{\texttt{novel}}$, where N is the number of novel classes and K the number of examples in each class (*i.e., K*-shot and typically K is small in few-shot regimes), and the data in $\mathcal{D}_q$ are also sampled from $\mathcal{C}_{\texttt{novel}}$. Each input image x in both $\mathcal{D}_{\text{base}}$ and $\mathcal{D}_{\texttt{fine}}$ has a set of annotations $\{(c_i, b_i)\}_{i=1}^{I}$, where I is the number of instances in x, c_i is the category of the i-th instance, and $b_i \in \mathbb{R}^4$ is the bounding box coordinates of the i-th instance.

Note that, as described, both the training $\mathcal{D}_{\text{base}}$ and fine-tuning $\mathcal{D}_{\texttt{fine}}$ datasets do not have mask supervision, and the data in $\mathcal{D}_{\texttt{fine}}$ is scarce; thus, our problem of interest is a weakly-supervised setting as well as few-shot regime. Given this data setting, our goal is to find a model that performs well on the query set $\mathcal{D}_q$ of novel classes that are not overlapped with the base classes $\mathcal{C}_{\text{base}}$

3.2 FoxInst: Few-shot Box-based Instance Segmentation Network

Our approach deals with a few-shot instance segmentation task in a weakly-supervised way with fine-tuning method. The training process of FoxInst is as follows: base training and fine-tuning. We use only the box annotations as weak supervision to predict the mask during the whole phase. In this section, we elaborate the architecture design and the details of each process.

Architecture. We adopt the architecture of BoxInst (Tian et al., 2021) for the weakly-supervised FSIS setting. BoxInst is comprised of ResNet-based (He et al., 2016) feature pyramid network (FPN; Lin et al., 2017a) backbone, mask branch, dynamic mask head, and prediction heads. We follow FCOS (Tian et al., 2019) for the prediction branch design, where the branches are composed of classification, regression, and centerness heads. All the heads are convolutional layers, instead of fully-connected layers, for per-pixel and anchor-free prediction. BoxInst has an additional head called controller, which dynamically predicts the weights of the mask head. This instance-aware controller enables the mask generation to be class-agnostic.

Weakly base training. In the base training phase, we train our whole network (the blue and red parts in Fig. 1) with the large amount of training data $D_{\texttt{base}}$ sampled from the base classes $C_{\texttt{base}}$. Since only weak supervisions of class and box annotations $\{(c_i, b_i)\}_{i=1}^{I}$ are given in our problem, we obtain mask supervision signals by leveraging the weakly-supervised loss (Tian et al., 2021) for mask $L_{mask} = L_{proj.} + L_{pair.}$, where

$$\begin{aligned} &L_{proj.} = L(\mathrm{Proj}_x(\tilde{m}), \mathrm{Proj}_x(b)) + L(\mathrm{Proj}_y(\tilde{m}), \mathrm{Proj}_y(b)), \\ &L_{pair.} = -\tfrac{1}{N}\textstyle\sum_{e \in E_{in}} \mathbb{1}[S_e \geq \tau] \log P(y_e = 1), \end{aligned} \tag{1}$$

where $\tilde{m}$ denotes the prediction mask, b its associated ground-truth bounding box, $\texttt{Proj}(\cdot)$ the binary pixel projection operation onto each axis, and $L(\cdot, \cdot)$ the Dice loss (Sudre et al., 2017). e denotes an edge between axis-aligned neighbor pixels, $\mathbb{1}[S_e \geq \tau]$ the indicator function which becomes 1 if $S_e \geq \tau$ and 0 otherwise, S_e the color similarity between two connected pixels, $P(y_e = 1)$ the probability of an edge connecting two pixels belonging to the same class given the predicted mask, and N is the number of the edges in the set E_{in}, which contains edges with at least one pixel inside the box. The projection loss term $L_{proj.}$ enforces the predicted pixel-level mask to be the tight box when projected on the x-axis and y-axis aligning to the bounding box supervision. The pairwise loss term $L_{pair.}$ propagates the predicted mask along the regions having similar colors. $L_{mask}(\cdot)$ consists of these two loss terms which enable the model to predict finer masks aligned with object boundaries from bounding box supervision. This loss trains the mask branch and controller.

For the other heads, the given weak labels are sufficient to provide full-supervision for the classification, centerness, and bounding box regression heads. We can use the standard full-supervision loss for these components. Thus, the total loss for our method is summarized as:

$$L = L_{cls.} + \lambda_1 L_{cen.} + \lambda_2 L_{reg.} + \lambda_3 L_{mask}, \tag{2}$$

where $L_{cls.}$ denotes the focal loss (Lin et al., 2017b) for classification, $L_{cen.}$ the binary cross entropy loss for the object centerness, $L_{reg.}$ IoU loss (Yu et al., 2016) for box regression. $\{\lambda\}$ are the balancing parameters between the loss terms, and we set them to 1 in this work.

Weakly few-shot fine-tuning. Given the fine-tuning dataset $\mathcal{D}_{\texttt{fine}}$ of the novel classes $C_{\texttt{novel}}$, we adapt the model trained in the weakly base training phase to novel objects. We freeze the backbone and mask branch (the blue parts in Fig. 1), and fine-tune the prediction heads (the red parts in Fig. 1), *i.e.,* classification, centerness, bounding box regression, and controller heads, on the dataset $D_{\texttt{fine}}$, which has the NK number of data sampled from the N number of $C_{\texttt{novel}}$ with the K-shot samples for each class. The weights of classification head associated with the $C_{\texttt{novel}}$ are randomly initialized due to the mismatch between the number of $C_{\texttt{base}}$ and N. During this phase, FoxInst also exploits only weak supervisions $\{(c_i, b_i)\}_{i=1}^{NK}$ and is trained with the same loss in Eq. (2) used in the base training phase but with lower learning rates and the small number of iterations to prevent overfitting.

To the best of our knowledge, our FoxInst is the first few-shot instance segmentation network that does not require mask annotations at all including the base training phase. Compared to the weakly few-shot semantic segmentation literature (Raza et al., 2019) as the closest work, our method is far more annotation-efficient because it calls for full supervision in the base training phase, where the number of $D_{\texttt{base}}$ is large.

We can regard the backbone as a feature extractor and fine-tune the novel class-specific prediction heads using the features extracted from the backbone. This has an analogy to Raghu et al. (2020)

that discovers few-shot adaptation of the last layers (heads) far more heavily affects the performance than that of the earlier layers (body). Similar findings are reported in few-shot classification (Chen et al., 2019) and detection (Wang et al., 2020) studies. Our base training followed by the fine-tuning can be regarded as a variant of the previous works to the few-shot instance segmentation task with challenging noisy weak supervision.

4 EXPERIMENT

In this section, we evaluate our FoxInst and compare with the previous works and a baseline on various settings. The models are trained and evaluated on a single NVIDIA Quadro RTX 8000.

Evaluation metric. The standard evaluation metric in instance segmentation is the average precision (AP). However, AP evaluates both classification and detection (or segmentation) in a single value making it hard to disentangle the performance of individual tasks, *e.g.,* low AP can be either due to bad localization with correct classification or accurate localization with the wrong classification. In order to analyze the localization performance independent of the class prediction, we additionally adopt foreground AP (FG-AP) to evaluate FSIS methods. FG-AP disentangles the effects of each head, so that we can readily identify which part is the performance bottleneck in a system. This provides a systematic way to analyze results, which is not possible by AP.[1]

Datasets. We use the large-scale MS COCO benchmark (Lin et al., 2015) (COCO) and PASCAL VOC (Everingham et al., 2010) (VOC) to evaluate our method. COCO has 80 classes, and VOC has 20 classes that are a subset of COCO classes. The overlapped 20 classes are used as novel classes $\mathcal{C}_{\texttt{novel}}$, and the remaining 60 classes are assigned as base classes $\mathcal{C}_{\texttt{base}}$. Following the setting suggested by Ganea et al. (2021), we use the train set of 80k images and the validation set of 35k images in COCO for training. The remaining 5k images in COCO are used for testing. Additionally, we evaluate our model trained on COCO with the VOC test set, which contains 1,449 test images and is referred to as COCO2VOC.

We also construct and evaluate on a new data split, the COCO Novel-only test set, consisting of 3,992 images where each image includes at least one instance of the novel classes $\mathcal{C}_{\texttt{novel}}$. We find out that the remaining 1,008 images of the total 5k test set of COCO do not contain any instance belonging to $\mathcal{C}_{\texttt{novel}}$. Since the original test set configuration can bias to numerous false positives, we further evaluate our FoxInst on the Novel-only setting. We randomly sample $K = 1, 5, 10$ shots for each novel class to form $\mathcal{D}_{\texttt{fine}}$, and the average results from the total of 10 tests are reported for all experiments. More details of the dataset setting can be found in the supplementary material.

Baselines. Since we are the first to propose weakly-supervised FSIS, there is no directly comparable method; thus, we develop a baseline. GrabCut (Rother et al., 2004), which extracts the foreground region of an object given a bounding box covering the object, has been used as a strong baseline on several weakly-supervised segmentation studies (Khoreva et al., 2017; Kulharia et al., 2020). Following these works, we design the GrabCut baseline by replacing the mask branch of FoxInst to GrabCut, named FoxInst+GrabCut. In other words, FoxInst+GRabCut generates mask prediction from predicted bounding box using GrabCut instead of mask branch.

We also juxtapose the recent fully-supervised FSIS models for the reference purpose, including a fully-converged Mask R-CNN fine-tuned on the novel classes (MRCN+ft-full; Yan et al. 2019, where ft stands for fine-tuning), Meta R-CNN (Yan et al., 2019), and MTFA (Ganea et al., 2021).[2] Thereby, we can sense the performance gap between full-supervision and weakly-supervision.

4.1 COMPARISON WITH OTHER FSIS METHODS

Results on the COCO novel classes. We evaluate our FoxInst on the novel classes $\mathcal{C}_{\texttt{novel}}$ of COCO and report the results in Table 1. FoxInst achieves about two times higher segmentation performance than the GrabCut baseline (2.99% vs. 6.09% mask AP for $K = 10$). FoxInst also

[1] In few-shot semantic segmentation literature (Wang et al., 2019; Dong & Xing, 2018; Rakelly et al., 2018), FG-IoU has been used to measure mask quality independently to other factors, but we are the first to suggest to use FG-AP in the FSIS context.

[2] There is another recent fully-supervised FSIS work, FAPIS (Nguyen & Todorovic, 2021), but we cannot include it because the source code and the specific data composition of FAPIS have not been released.

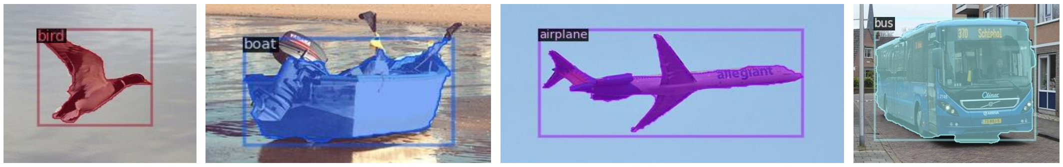


Figure 2: Qualitative results of FoxInst on the COCO novel classes. As shown in the figure, FoxInst predicts the mask well with only 5-shot samples for each class without pixel-level mask annotations.

Table 1: Performance on the COCO novel classes. FoxInst (det.) + GrabCut (seg.) represents the baseline model that replaces the mask branch of FoxInst to GrabCut.

Shots	**Mask sup.**	**Method**	**Detection**				**Segmentation**			
			AP	**AP50**	**FG-AP**	**FG-AP50**	**AP**	**AP50**	**FG-AP**	**FG-AP50**
K=1	$\mathcal{F}$	MTFA	2.47	4.85	2.89	6.14	2.66	4.56	2.40	4.15
	$\mathcal{B}$	FoxInst (det.) + GrabCut (seg.)	1.96	3.49	2.49	5.04	0.74	1.63	0.88	2.21
	$\mathcal{B}$	FoxInst (det. + seg.)					1.62	3.05	1.72	3.71
K=5	$\mathcal{F}$	MRCN+ft-full	1.3	3.0	-	-	1.3	2.7	-	-
	$\mathcal{F}$	Meta R-CNN	3.5	9.9	-	-	2.8	6.9	-	-
	$\mathcal{F}$	MTFA	6.61	12.32	6.30	12.97	6.62	11.58	5.11	9.88
	$\mathcal{B}$	FoxInst (det.) + GrabCut (seg.)	5.58	10.02	6.57	13.23	2.39	5.19	2.45	5.89
	$\mathcal{B}$	FoxInst (det. + seg.)					4.56	8.52	4.31	9.39
K=10	$\mathcal{F}$	MRCN+ft-full	2.5	5.7	-	-	1.9	4.7	-	-
	$\mathcal{F}$	Meta R-CNN	5.6	14.2	-	-	4.4	10.6	-	-
	$\mathcal{F}$	MTFA	8.52	15.53	7.21	14.49	8.39	14.64	5.88	11.29
	$\mathcal{B}$	FoxInst (det.) + GrabCut (seg.)	7.55	13.62	8.30	16.39	2.99	6.66	3.05	7.42
	$\mathcal{B}$	FoxInst (det. + seg.)					6.09	11.63	5.61	12.22

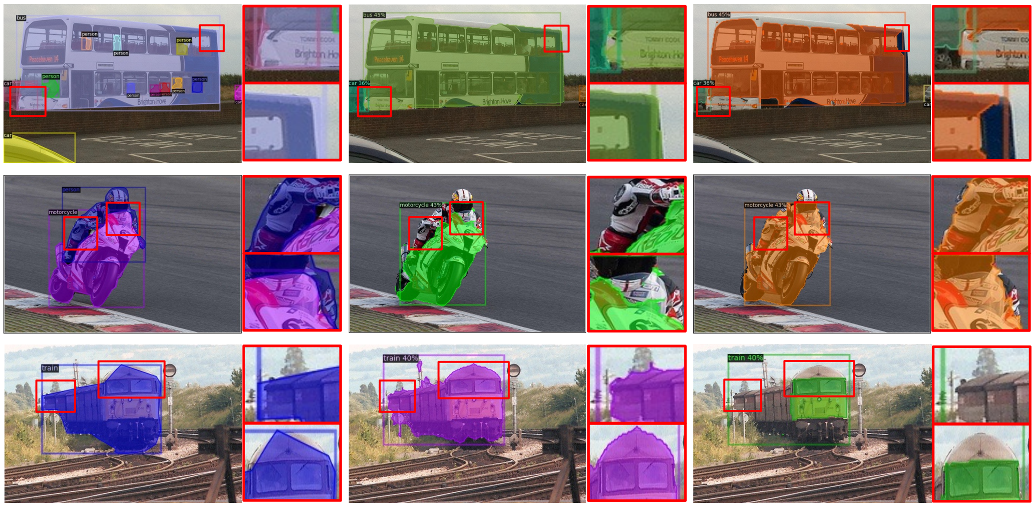
$\mathcal{F}$–Full supervision, $\mathcal{B}$–Weak supervision with bounding boxes.

has comparable performance to the fully-supervised networks and even surpasses the recent fully-supervised FSIS methods, MRCN+ft-full and Meta R-CNN. These results imply that our simple fine-tuning approach in FoxInst is indeed strong and our FoxInst system integration is non-trivial.

For the all K-shot settings, the performance gaps between FoxInst and MTFA are smaller in FG-AP than in AP, e.g., for $K = 10$, 2.3% gap in mask AP (6.09% vs. 8.39%) and 0.27% gap in mask FG-AP (5.61% vs. 5.88%). This demonstrates that FoxInst can generate a segmentation mask as delicate as the powerful fully-supervised method, MTFA, and the bottleneck of FoxInst is not segmentation but classification. This analysis is useful to identify a future direction to further develop in this new area of weakly-supervised FSIS. Besides, this fact is well demonstrated in Fig. 2 that shows the high-quality mask predictions of FoxInst despite few training examples without any mask annotation. In addition, FoxInst not only outperforms the GrabCut baseline (6.66% vs. 11.63% mask AP50 for $K = 10$) but also predicts more accurate masks than the ground-truth in some cases, *e.g.,* the back of the bus and the top of the train in Fig. 3.

We postulate that this high-performance mask prediction of FoxInst is by virtue of the class-agnostic mask head. (Pinheiro et al., 2015) demonstrate that the class-agnostic mask model can be generalized well to unseen classes if trained with abundant full annotations. Our results show that its effectiveness is extended and generalized in our weakly few-shot regime as well.

Results on both base and novel COCO classes. In Table 2, we evaluate FoxInst fine-tuned on both $\mathcal{C}_{\texttt{base}}$ and $\mathcal{C}_{\texttt{novel}}$ of COCO, following the protocol suggested by Ganea et al. (2021), and summarize the performance on the overall and individual class compositions. Through this experiment, we can observe how much the performance of the model trained on $\mathcal{C}_{\texttt{base}}$ is maintained even after fine-tuning, *i.e.,* forgetting effects (Gidaris & Komodakis, 2018). In this experiment, we compare our method to the closest fully-supervised work, MTFA, and the weakly-supervised GrabCut baseline, and the models are trained on $\mathcal{D}_{\texttt{fine}}$, which is a balanced subset consisting of both $\mathcal{C}_{\texttt{base}}$ and $\mathcal{C}_{\texttt{novel}}$. As performance upper bound references, we also compare FoxInst with the models trained only on $\mathcal{C}_{\texttt{base}}$ in the base training phase, which are tagged with Base-only.



(a) Ground-truth (b) FoxInst (ours) (c) GrabCut baseline

Figure 3: Qualitative comparisons between our FoxInst and the GrabCut baseline. The enlarged views (red box) show in detail that FoxInst predicts a higher quality mask than the GrabCut baseline.

Table 2: Performance on both base and novel COCO classes. In this setting, fine-tuning and query sets are sampled from both base and novel COCO classes while base training is the same as before. Base-only tag means the model is trained only on the base classes without fine-tuning phase.

Shots	Mask sup.	Method	Detection						Segmentation					
			Overall		Base		Novel		Overall		Base		Novel	
			AP	AP50	AP	AP50	AP	AP50	AP	AP50	AP	AP50	AP	AP50
-	$\mathcal{F}$	MTFA Base-only	28.67	43.53	38.22	58.04	-	-	26.34	41.55	35.12	55.40	-	-
	$\mathcal{B}$	FoxInst Base-only	30.74	45.24	38.43	56.55	-	-	23.69	40.85	29.61	51.06	-	-
K=1	$\mathcal{F}$	MTFA	24.32	39.64	31.73	51.49	2.10	4.07	22.98	37.48	29.85	48.64	2.34	3.99
	$\mathcal{B}$	FoxInst (det.) + GrabCut (seg.)	25.37	38.31	33.10	49.79	2.19	3.86	6.25	11.63	8.03	14.82	0.94	2.06
	$\mathcal{B}$	FoxInst (det. + seg.)							19.57	34.13	25.52	44.41	1.73	3.30
K=5	$\mathcal{F}$	MTFA	26.39	41.52	33.11	51.49	6.22	11.63	25.07	39.95	31.29	49.55	6.38	11.14
	$\mathcal{B}$	FoxInst (det.) + GrabCut (seg.)	28.66	42.66	36.22	53.43	6.00	10.36	7.56	14.06	9.24	16.88	2.52	5.60
	$\mathcal{B}$	FoxInst (det. + seg.)							22.09	38.14	27.90	47.89	4.66	8.86
K=10	$\mathcal{F}$	MTFA	27.44	42.84	33.83	52.04	8.28	15.25	25.97	41.28	31.84	50.17	8.36	14.58
	$\mathcal{B}$	FoxInst (det.) + GrabCut (seg.)	28.69	42.78	35.73	52.64	7.59	13.19	7.64	14.31	9.22	17.02	2.89	6.17
	$\mathcal{B}$	FoxInst (det. + seg.)							22.12	38.22	27.56	47.24	5.80	11.18

$\mathcal{F}$–Full supervision, $\mathcal{B}$–Weak supervision with bounding boxes.

The results show that FoxInst surpasses the fully-supervised MTFA in the detection task on $\mathcal{C}_{\texttt{base}}$ and achieves comparable performance on $\mathcal{C}_{\texttt{novel}}$, which leads to an outstanding result on the overall AP. In the detection task, the Base-only cases of both MTFA and FoxInst show the best results for $\mathcal{C}_{\texttt{base}}$ as expected, but the gap between FoxInst fine-tuned on both base and novel classes and the FoxInst Base-only tends to be smaller than that of MTFA. For $K = 5, 10$, those gaps are further small, which are 2.21 (36.22% vs. 38.43% mask AP) and 2.7 (35.73% vs. 38.43% mask AP), respectively. This implies that FoxInst can lead off the information obtained during the base training phase while being successfully adapted to the novel classes in the detection task.

In the segmentation task, FoxInst outperforms the GrabCut baseline and the gap in $K = 10$ is prominent as 30.22% (from 17.02% to 47.24%) and 5.01% (from 6.17% to 11.18%) mask AP50 for $\mathcal{C}_{\texttt{base}}$ and $\mathcal{C}_{\texttt{novel}}$, respectively. Moreover, our inference time is much faster than the GrabCut baseline, where FoxInst takes about 0.24 seconds per image and GrabCut requires about 1.2 seconds per object (note that per object, not per image). Also, compared to the fully supervised MTFA, FoxInst perform favorably against MTFA despite being only with bounding box annotations.

Results on the novel COCO classes for Novel-only setting. As mentioned, in the query set $\mathcal{D}_q$ of COCO, there are images containing no novel instance, which can bring in false positives in evaluation procedure. For the more fair and right experiment on $\mathcal{C}_{\texttt{novel}}$, we construct a Novel-only

Table 3: Performance on the COCO novel classes for Novel-only setting. In the Novel-only setting, the query set contains 3,992 images that have at least one novel object among the total 5,000 COCO query images.

Shots	Mask sup.	Method	Detection				Segmentation			
			AP	AP50	FG-AP	FG-AP50	AP	AP50	FG-AP	FG-AP50
K=1	$\mathcal{F}$	MTFA	2.82	5.55	3.92	8.33	2.97	5.16	3.26	6.14
	$\mathcal{B}$	FoxInst (det.) + GrabCut (seg.)	2.28	4.04	3.14	6.30	0.85	1.87	1.11	2.76
	$\mathcal{B}$	FoxInst (det. + seg.)					1.88	3.53	2.19	4.66
K=5	$\mathcal{F}$	MTFA	7.27	13.59	7.98	16.42	7.21	12.69	6.46	12.52
	$\mathcal{B}$	FoxInst (det.) + GrabCut (seg.)	6.15	11.08	8.04	16.30	2.56	5.57	2.92	7.04
	$\mathcal{B}$	FoxInst (det. + seg.)					4.95	9.36	5.22	11.52
K=10	$\mathcal{F}$	MTFA	9.27	16.99	8.77	17.78	9.02	15.88	7.10	13.78
	$\mathcal{B}$	FoxInst (det.) + GrabCut (seg.)	8.22	14.90	9.66	19.31	3.18	7.13	3.47	8.51
	$\mathcal{B}$	FoxInst (det. + seg.)					6.54	12.62	6.45	14.27

$\mathcal{F}$–Full supervision, $\mathcal{B}$–Weak supervision with bounding boxes.

Table 4: One-shot performance on the COCO2VOC novel classes. COCO2VOC is a cross-dataset using the COCO as base training and fine-tuning set, and VOC as query set.

Mask sup.	Method	Detection				Segmentation			
		AP	AP50	FG-AP	FG-AP50	AP	AP50	FG-AP	FG-AP50
$\mathcal{F}$	MRCN+ft+full	-	6.0	-	-	-	0.4	-	-
$\mathcal{F}$	Siamese Mask R-CNN	-	23.9	-	-	-	13.8	-	-
$\mathcal{F}$	Meta R-CNN	-	20.1	-	-	-	12.5	-	-
$\mathcal{F}$	FGN	-	30.8	-	-	-	16.2	-	-
$\mathcal{F}$	MTFA	9.99	21.68	9.37	21.19	9.51	19.28	8.38	17.10
$\mathcal{B}$	FoxInst (det.) + GrabCut (seg.)	**17.79**	**33.08**	**17.76**	**34.54**	6.85	15.62	6.98	16.32
$\mathcal{B}$	FoxInst (det. + seg.)					**12.17**	**26.24**	**12.14**	**26.70**

$\mathcal{F}$–Full supervision, $\mathcal{B}$–Weak supervision with bounding boxes.

setting and evaluate the models on it in Table 3, where the refined set consists of only images that include at least one novel class instance. The results are improved by and large compared with Table 1, which means that the original $\mathcal{D}_q$ of COCO causes a lot of false positive bias and disturbs further analysis. Based on our findings, we recommend future few-shot learning works consider our new test split, the Novel-only, as an evaluation dataset.

Results on the COCO2VOC. In Table 4, we further evaluate our FoxInst on the cross-dataset setting called COCO2VOC, referring to Fan et al. (2020), where the training dataset $\mathcal{D}_{\texttt{base}}$ and $\mathcal{D}_{\texttt{fine}}$ are sampled from the associated data in COCO, and the query set $\mathcal{D}_q$ from the associated ones with $\mathcal{C}_{\texttt{novel}}$ in VOC. The evaluation procedure follows the ground-truth only evaluation (GTOE) setting (Fan et al., 2020), which eliminates the predictions for the classes that do not exist in the ground-truth by zeroing the corresponding softmax outputs. The results show that our FoxInst surpasses all counterpart models, including fully-supervised models, by a large margin. We postulate that all the models except our FoxInst are anchor-based ones; thus, they may suffer from overfitting to a specific scale of anchor box suitable for $\mathcal{D}_{\texttt{base}}$ of COCO, and it may lead to performance degradation for $\mathcal{D}_q$ of VOC. Contrastively, FoxInst as an anchor-free model is inherently free from overfitting problems and robust to cross-datasets.

4.2 Ablation Study

We perform a couple of ablation studies on the 5-shot setting for the COCO novel classes.

Fine-tuning strategy. To analyze the importance of each component in FoxInst during the fine-tuning phase, we evaluate the combinations of the components to be fine-tuned as follows: (1) the classification head alone, (2) the classification and the box regression heads, (3) all of the prediction heads, and (4) the whole model. The results are summarized in Table 5. When comparing (3) and (4), the result supports that high performance is achieved from the prediction head fine-tuning by

Table 5: Ablation study of FoxInst according to fine-tuning components. The components of FoxInst are classifier, box regressor, controller, and backbone and mask branch.

	Classifier	Box regressor	Controller	Backbone & Mask branch	Detection		Segmentation	
					AP	**AP50**	**AP**	**AP50**
(1)	✓				5.20	9.15	4.27	7.84
(2)	✓	✓			5.21 (+0.01)	9.26 (+0.11)	4.33 (+0.06)	8.02 (+0.18)
(3)	✓	✓	✓		**5.58** (+0.38)	**10.02** (+0.87)	**4.56** (+0.29)	**8.52** (+0.68)
(4)	✓	✓	✓	✓	5.12 (-0.08)	8.33 (-0.82)	4.01 (-0.26)	7.28 (-0.56)

Table 6: Initialization comparison of class-agnostic prediction heads. Class-agnostic prediction heads include the centerness, bounding box regression, and controller heads. Warm-start parameters are obtained from the base trained model.

Initialization	Detection		Segmentation	
	AP	**AP50**	**AP**	**AP50**
Random	5.17	9.79	0.11	0.26
Warm-start	**5.58** (+0.41)	**10.02** (+0.23)	**4.56** (+4.45)	**8.58** (+8.32)

feature reuse (Raghu et al., 2020), not fast model adaptation (Finn et al., 2017). It implies that our weakly-supervised two-phase fine-tuning method, which freezes the feature extraction layers and fine-tunes only the prediction branches, is suitable for the weakly-supervised FSIS task. Next, comparing (1), (2), and (3), we can figure out the importance of each prediction head through the gain as each component is added. The results imply that the controller head is a key component of FoxInst during the fine-tuning phase.

Weight initialization. We compare weight initialization on our FoxInst's class-agnostic prediction heads; the bounding box regression, centerness, and controller heads. There are two ways of initialization: (1) random initialization and (2) warm-start with the weights pre-trained on the base classes. Table 6 shows that the gap between the two methods is subtle in detection (9.79% vs. 10.02% mask AP) but rather noticeable in segmentation (0.26% vs. 8.58% mask AP). It implies that a warm initial point gives a positive effect on the controller, which is the segmentation-related one. Bounding box regression and centerness heads, which are the detection-related ones, are less affected by initialization and easier to be trained with only few samples.

5 Conclusion

We first define a data-efficient instance segmentation task, weakly-supervised few-shot instance segmentation (weakly-supervised FSIS), and propose a frustratingly simple baseline, called FoxInst, which is the first weakly-supervised FSIS network. FoxInst only uses class labels and bounding box annotations as supervision during both the train and test phase, and fast adapt to novel classes with only few data. This is more data efficient than the previous FSIS setup and models. FoxInst outperforms or performs favorably against the strong baseline and the fully-supervised FSIS methods on various data settings. In addition, FoxInst even outperforms the state-of-the-art of the fully-supervised FSIS in the cross-dataset setting, COCO2VOC.

Moreover, we systematically analyze which parts are dominant components by proposing a metric, FG-AP which better disentangles the effects of each component. Also, we found a critical evaluation bias in FSIS, and refine the existing test dataset by removing images with no novel class instance. This allows for more precise evaluation. Our analysis suggests that future research should deal with the current bottleneck, *i.e.,* classification, in focus. We believe that FoxInst can be a solid foundation in the weakly-supervised FSIS task and open up many new subsequent applications to be derived.

Ethics Statement

As many studies have focused only on the accuracy of deep learning (Xie et al., 2020; Devlin et al., 2019; Dosovitskiy et al., 2021), the impact on the environment has been overlooked and an astonishing amount of carbon footprint is still being generated. The concept of Green AI (Schwartz et al., 2020) is emerging and the attitude of studying artificial intelligence while considering their effect on the environment is being emphasized, *e.g.,* reproducibility of the trained model, understanding the principle of deep learning, and data efficiency. We believe that adopting the mechanism of FoxInst to train a deep learning network with few data samples and a small training schedule would contribute to decrease a carbon footprint in terms of data-efficiency.

REFERENCES

Antreas Antoniou, Harrison Edwards, and Amos Storkey. How to train your maml. In *International Conference on Learning Representations (ICLR)*, 2019.

Aditya Arun, C. V. Jawahar, and M. Pawan Kumar. Weakly supervised instance segmentation by learning annotation consistent instances. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (eds.), *European Conference on Computer Vision (ECCV)*, 2020.

Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. Yolact: Real-time instance segmentation. In *IEEE International Conference on Computer Vision (ICCV)*, 2019.

Hao Chen, Xiaojuan Qi, Lequan Yu, Qi Dou, Jing Qin, and Pheng-Ann Heng. Dcan: Deep contour-aware networks for object instance segmentation from histology images. *Medical Image Analysis (MedIA)*, 36, 2016.

Hao Chen, Kunyang Sun, Zhi Tian, Chunhua Shen, Yongming Huang, and Youliang Yan. BlendMask: Top-down meets bottom-up for instance segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

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

Hisham Cholakkal, Guolei Sun, Fahad Shahbaz Khan, and Ling Shao. Object counting and instance segmentation with image-level supervision. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019.

Guneet Singh Dhillon, Pratik Chaudhari, Avinash Ravichandran, and Stefano Soatto. A baseline for few-shot image classification. In *International Conference on Learning Representations (ICLR)*, 2020.

Nanqing Dong and Eric P. Xing. Few-shot semantic segmentation with prototype learning. In *British Machine Vision Conference (BMVC)*, 2018.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision (IJCV)*, 88(2), 2010.

Zhibo Fan, Jin-Gang Yu, Zhihao Liang, Jiarong Ou, Changxin Gao, Gui-Song Xia, and Yuanqing Li. Fgn: Fully guided network for few-shot instance segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning (ICML)*, 2017.

Dan Andrei Ganea, Bas Boom, and Ronald Poppe. Incremental few-shot instance segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Yuying Ge, Ruimao Zhang, Xiaogang Wang, Xiaoou Tang, and Ping Luo. Deepfashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Spyros Gidaris and Nikos Komodakis. Dynamic few-shot visual learning without forgetting. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Erin Grant, Chelsea Finn, Sergey Levine, Trevor Darrell, and Thomas Griffiths. Recasting gradient-based meta-learning as hierarchical bayes. In *International Conference on Learning Representations (ICLR)*, 2019.

Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Simultaneous detection and segmentation. In *European Conference on Computer Vision (ECCV)*, 2014.

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

Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask r-cnn. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.

Cheng-Chun Hsu, Kuang-Jui Hsu, Chung-Chi Tsai, Yen-Yu Lin, and Yung-Yu Chuang. Weakly supervised instance segmentation using the bounding box tightness prior. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Ronghang Hu, Piotr Dollár, Kaiming He, Trevor Darrell, and Ross Girshick. Learning to segment every thing. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Jaedong Hwang, Seohyun Kim, Jeany Son, and Bohyung Han. Weakly supervised instance segmentation by deep community learning. In *IEEE Winter Conf. on Applications of Computer Vision (WACV)*, 2021.

Anna Khoreva, Rodrigo Benenson, Jan Hosang, Matthias Hein, and Bernt Schiele. Simple does it: Weakly supervised instance and semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Viveka Kulharia, Siddhartha Chandra, Amit Agrawal, Philip Torr, and Ambrish Tyagi. Box2seg: Attention weighted loss and discriminative feature learning for weakly supervised segmentation. In *European Conference on Computer Vision (ECCV)*, 2020.

Jungbeom Lee, Jihun Yi, Chaehun Shin, and Sungroh Yoon. Bbam: Bounding box attribution map for weakly supervised semantic and instance segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Youngwan Lee and Jongyoul Park. Centermask: Real-time anchor-free instance segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context. *arXiv*, 2015.

Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017a.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. In *IEEE International Conference on Computer Vision (ICCV)*, 2017b.

Yun Liu, Yu-Huan Wu, Peisong Wen, Yujun Shi, Yu Qiu, and Ming-Ming Cheng. Leveraging instance-, image- and dataset-level information for weakly supervised instance segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2020.

Claudio Michaelis, Ivan Ustyuzhaninov, Matthias Bethge, and Alexander S. Ecker. One-shot instance segmentation. *arXiv*, 2018.

Khoi Nguyen and Sinisa Todorovic. Fapis: A few-shot anchor-free part-based instance segmenter. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Alex Nichol and John Schulman. Reptile: a scalable metalearning algorithm. *arXiv*, 2018.

Pedro O. Pinheiro, Ronan Collobert, and Piotr Dollar. Learning to segment object candidates. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015.

Aniruddh Raghu, Maithra Raghu, Samy Bengio, and Oriol Vinyals. Rapid learning or feature reuse? towards understanding the effectiveness of maml. In *International Conference on Learning Representations (ICLR)*, 2020.

Kate Rakelly, Evan Shelhamer, Trevor Darrell, Alyosha Efros, and Sergey Levine. Conditional networks for few-shot semantic segmentation. In *International Conference on Learning Representations Workshops (ICLRW)*, 2018.

Hasnain Raza, Mahdyar Ravanbakhsh, Tassilo Klein, and Moin Nabi. Weakly supervised one shot segmentation. In *IEEE International Conference on Computer Vision (ICCV)*, 2019.

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 (NeurIPS)*, 2015.

Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. ”grabcut”: interactive foreground extraction using iterated graph cuts. *ACM Transactions on Graphics (SIGGRAPH)*, 2004.

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

Roy Schwartz, Jesse Dodge, Noah Smith, and Oren Etzioni. Green ai. *Communications of the ACM (CACM)*, 63, 2020.

Jake Snell, Kevin Swersky, and Richard S Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Carole H. Sudre, Wenqi Li, Tom Kamiel Magda Vercauteren, Sébastien Ourselin, and M. Jorge Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. *Deep learning in medical image analysis and multimodal learning for clinical decision support*, 2017.

Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *IEEE International Conference on Computer Vision (ICCV)*, 2019.

Zhi Tian, Chunhua Shen, and Hao Chen. Conditional convolutions for instance segmentation. In *European Conference on Computer Vision (ECCV)*, 2020.

Zhi Tian, Chunhua Shen, Xinlong Wang, and Hao Chen. Boxinst: High-performance instance segmentation with box annotations. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

Kaixin Wang, Jun Hao Liew, Yingtian Zou, Daquan Zhou, and Jiashi Feng. Panet: Few-shot image semantic segmentation with prototype alignment. In *IEEE International Conference on Computer Vision (ICCV)*, 2019.

Xin Wang, Thomas E. Huang, Trevor Darrell, Joseph E Gonzalez, and Fisher Yu. Frustratingly simple few-shot object detection. In *International Conference on Machine Learning (ICML)*, 2020.

Syed Waqas Zamir, Aditya Arora, Akshita Gupta, Salman Khan, Guolei Sun, Fahad Shahbaz Khan, Fan Zhu, Ling Shao, Gui-Song Xia, and Xiang Bai. isaid: A large-scale dataset for instance segmentation in aerial images. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019.

Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. `https://github.com/facebookresearch/detectron2`, 2019.

Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V. Le. Self-training with noisy student improves imagenet classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Xiaopeng Yan, Ziliang Chen, Anni Xu, Xiaoxi Wang, Xiaodan Liang, and Liang Lin. Meta r-cnn: Towards general solver for instance-level low-shot learning. In *IEEE International Conference on Computer Vision (ICCV)*, 2019.

Tong Yang, Xiangyu Zhang, Zeming Li, Wenqiang Zhang, and Jian Sun. Metaanchor: Learning to detect objects with customized anchors. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Jiahui Yu, Yuning Jiang, Zhangyang Wang, Zhimin Cao, and Thomas S. Huang. Unitbox: An advanced object detection network. *Proceedings of the 24th ACM international conference on Multimedia*, 2016.

Chi Zhang, Guosheng Lin, Fayao Liu, Rui Yao, and Chunhua Shen. Canet: Class-agnostic segmentation networks with iterative refinement and attentive few-shot learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Shifeng Zhang, Cheng Chi, Yongqiang Yao, Zhen Lei, and Stan Z. Li. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Ziyu Zhang, Sanja Fidler, and Raquel Urtasun. Instance-level segmentation for autonomous driving with deep densely connected mrfs. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Yanzhao Zhou, Yi Zhu, Qixiang Ye, Qiang Qiu, and Jianbin Jiao. Weakly supervised instance segmentation using class peak response. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Supplementary material

In this supplementary material, we present additional implement details, qualitative results, and experiments that are not included in the main paper due to the page limit.

A Implement Details

Our framework is implemented based on Detectron2 (Wu et al., 2019). For the COCO benchmark, the base training phase runs for 90k iterations with a learning rate of 0.01, decaying with a factor of 10 at 60k and 80k iterations. We fine-tune the network on the novel classes with a learning rate of 0.0005 for 1k iterations. When we consider both base and novel classes as in Table 2, we reuse the classification heads trained on the base training phase and the fine-tuning on the novel classes. In other words, we fine-tune FoxInst on a balanced subset of both base and novel classes, the warm-starting with weights from the base training and fine-tuning on the novel classes.

However, we find out that using a unified learning rate for all convolution filters of classification head significantly degrades the performance on the base classes, *i.e.,* forgetting effects (Gidaris & Komodakis, 2018). Therefore, the convolution filters for the base classes are fine-tuned with a small learning rate of 0.00001 $(= 1e^{-5})$, while a relatively large value of 0.005 is used for the convolution weights linked to the novel classes. The other prediction heads are trained with a learning rate of 0.0005 as same as the original setting. For $K = 1, 5$, we fine-tune FoxInst only for 500 iterations, because the longer training schedule causes overfitting to the novel classes, injuring the performance on the base classes. The models are trained for 1k iterations on 10-shot setting with the assist of comparatively enough samples to avoid the problem.

B Qualitative Results

In this section, we give some qualitative results with 5-shot setting.

B.1 Qualitative Results on the COCO novel classes.

Figure 4 is the visualization of FoxInst trained with 5-shot setting on the COCO novel classes. The result shows the high-quality mask predictions of FoxInst despite few training examples without any mask annotation

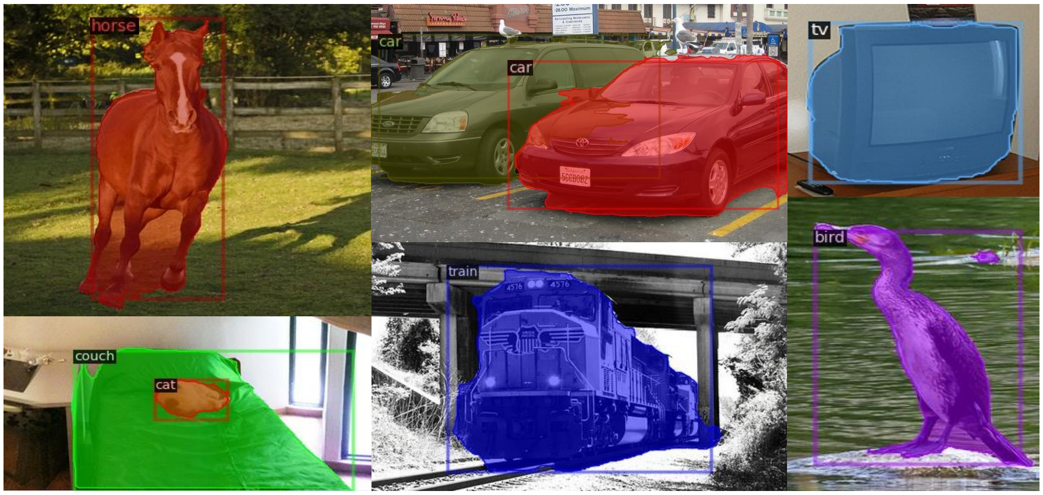


Figure 4: Qualitative results of FoxInst on the COCO novel classes.

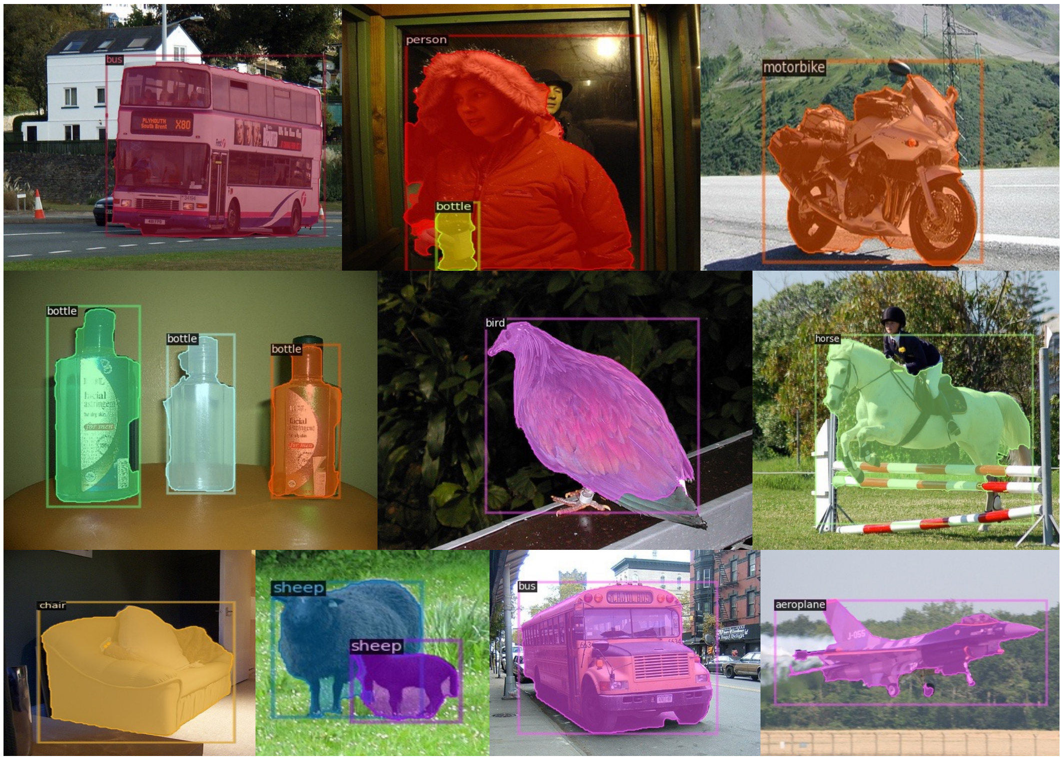


Figure 5: Qualitative results of FoxInst on the COCO2VOC setting.

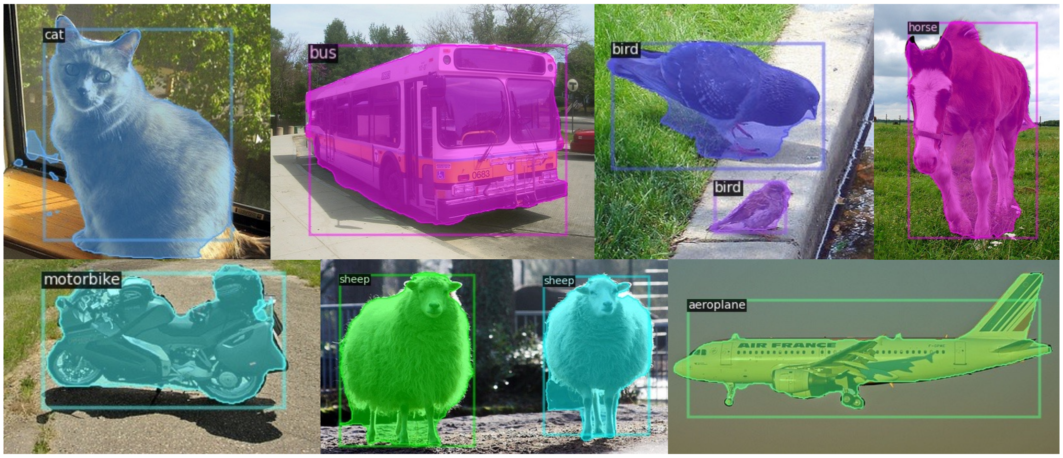


Figure 6: Qualitative results of VOC2VOC.

B.2 Qualitative Results on COCO2VOC.

Figure 5 is the visualization of FoxInst trained with 5-shot setting on COCO2VOC. This result demonstrates that FoxInst can obtain high-quality masks only with few samples and box annotation, even for the cross-dataset, where the training and query dataset are different. We believe that FoxInst can be extended to a more complicated domain adaptation task with its powerful and efficient method.

Table 7: Performance of FoxInst and the GrabCut baseline on the VOC2VOC novel classes.

Shots	**Method**	**Novel-class Setup 1**				**Novel-class Setup 2**				**Novel-class Setup 3**			
		Detection		**Segmentation**		**Detection**		**Segmentation**		**Detection**		**Segmentation**	
		AP	**AP50**	**AP**	**AP50**	**AP**	**AP50**	**AP**	**AP50**	**AP**	**AP50**	**AP**	**AP50**
K=1	FoxInst (det.) + GrabCut (seg.)	6.96	14.30	3.00	6.73	4.66	11.89	2.99	7.23	5.37	11.67	1.84	4.47
	FoxInst (det. + seg.)			**4.37**	**10.47**			**3.34**	**8.87**			**2.52**	**6.87**
K=5	FoxInst (det.) + GrabCut (seg.)	16.78	32.35	6.29	14.30	11.21	25.33	5.81	13.41	13.08	28.12	5.09	12.05
	FoxInst (det. + seg.)			**10.52**	**23.64**			**6.88**	**17.18**			**6.64**	**18.22**
K=10	FoxInst (det.) + GrabCut (seg.)	22.43	42.63	8.30	18.46	14.66	31.52	6.78	16.05	17.88	38.16	6.72	15.87
	FoxInst (det. + seg.)			**14.09**	**32.26**			**8.36**	**21.07**			**9.24**	**25.56**

C RESULTS ON VOC2VOC.

We further evaluate our FoxInst on the VOC2VOC setting, which assesses the model trained on the small number of base training data. In this setting, the VOC dataset is used in base training, fine-tuning, and query evaluation. Fifteen classes are used as the base classes, and the remaining five classes are assigned to the novel classes. Three configurations of the novel classes are used as following TFA (Wang et al., 2020); {"bird", "bus", "cow", "motorbike", "sofa"}, {"aeroplane", "bottle", "cow", "horse", "sofa"}, and {"boat", "cat", "motorbike", "sheep", "sofa"}. Evaluation for VOC2VOC follows the Novel-only setting, which leaves 420, 421, and 402 images out of the total 1,449 test images for each configuration.

We compare against the GrabCut baseline[3] and the results are shown in Table 7. In all K-shot settings, FoxInst surpasses the GrabCut baseline. For the first setup, FoxInst achieves 32.26% mask AP50 with 10-shot samples, which is 13.80% higher than the performance of GrabCut. Figure 6 is the qualitative results of our FoxInst trained with 5-shot samples on VOC2VOC. FoxInst can detect minute regions, *e.g.,* the front wheel of the airplane, and multiple instances in a single image. These results demonstrate that FoxInst can perform successfully with the small base training dataset.

[3] Although FGN (Fan et al., 2020) also considers the VOC2VOC setting, we cannot compare FoxInst with it because they randomly sample the novel classes and the specific composition has not been released.