FGN: Fully Guided Network for Few-Shot Instance Segmentation

Zhibo Fan[1], Jin-Gang Yu[1,2,*], Zhihao Liang[1], Jiarong Ou[1],
Changxin Gao[3], Gui-Song Xia[4], Yuanqing Li[1,2]

[1]South China University of Technology [2]Guangzhou Laboratory
[3]Huazhong University of Science and Technology [4]Wuhan University

{zanefan0323,zhliang19980922}@gmail.com, {jingangyu,yqli}@scut.edu.cn,
au_jaring@mail.scut.edu.cn, cgao@hust.edu.cn, guisong.xia@whu.edu.cn

Abstract

Few-shot instance segmentation (FSIS) conjoins the few-shot learning paradigm with general instance segmentation, which provides a possible way of tackling instance segmentation in the lack of abundant labeled data for training. This paper presents a Fully Guided Network (FGN) for few-shot instance segmentation. FGN perceives FSIS as a guided model where a so-called support set is encoded and utilized to guide the predictions of a base instance segmentation network (i.e., Mask R-CNN), critical to which is the guidance mechanism. In this view, FGN introduces different guidance mechanisms into the various key components in Mask R-CNN, including Attention-Guided RPN, Relation-Guided Detector, and Attention-Guided FCN, in order to make full use of the guidance effect from the support set and adapt better to the inter-class generalization. Experiments on public datasets demonstrate that our proposed FGN can outperform the state-of-the-art methods.

1. Introduction

Instance segmentation [10, 12] is a fundamental computer vision task which aims to simultaneously localize, classify and estimate the segmentation masks of object instances from a given image. The past few years have witnessed notable advances on instance segmentation thanks to the prosperity of convolutional neural networks (CNN) [12, 19, 4, 3], as well as its success in a variety of real-world applications [33, 31, 9]. Existing CNN-based approaches to instance segmentation are mostly fully-supervised, which require abundant labeled data for model training [12, 24, 11]. Such a data-hungry setting however may be impractical.

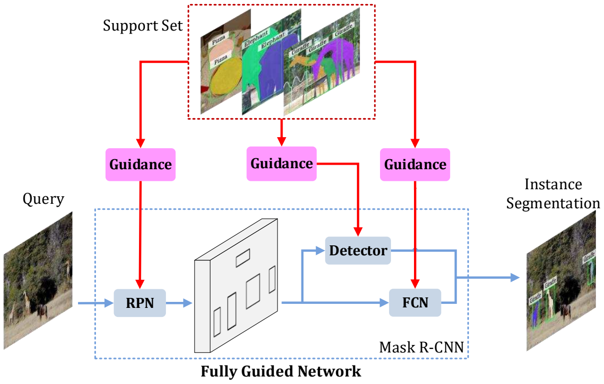

Figure 1. Illustration of few-shot instance segmentation using the proposed Fully Guided Network (FGN). To adapt better to the inter-class generalization, FGN introduces different guidance mechanisms for the various key components in Mask R-CNN.

Inspired by the remarkable ability of human to learn with limited data, few-shot learning (FSL) has recently received a lot of research attention [29, 27, 16, 28, 8]. Assuming the availability of a large amount of labeled data belonging to certain classes (base classes) for training, FSL aims at making predictions on data from other different classes (novel classes) given only a handful of labeled exemplars for each [29, 27]. Instead of fine-tuning an ordinary model pre-trained on base classes with the very limited novel-class samples, or conducting data augmentation, FSL learns a conditional model that makes predictions conditioned on a support set, so as to adapt to the inter-class generalization.

The majority of existing FSL models focus on visual classification, and a minority on semantic segmentation [32, 22, 26, 5]. Nevertheless, it has been rarely explored so far in the context of instance segmentation, the task of our concern termed as few-shot instance segmentation (FSIS). While we argue the FSL paradigm should be effective as well for addressing instance segmentation with limited data, it is by no means trivial to couple the two practically. Crucial to any FSL approach is an appropriate mechanism for encoding and utilizing the support set to guide the base net-

[*]Corresponding author

work (*e.g.*, ResNet [13] for classification or FCN [20] for semantic segmentation). In comparison with the tasks of visual classification or semantic segmentation, designing such a guidance mechanism for instance segmentation becomes far more challenging, which is mainly because instance segmentation networks usually have more complex structures.

In previous attempts [21, 30], the authors proposed to establish guided networks upon Mask R-CNN [12], probably the most representative model for general instance segmentation. Mask R-CNN is a two-stage network, where the first-stage region proposal network (RPN) generates class-agnostic object proposals, and the second-stage subnet consists of three heads for classification, bounding-box (bbox) regression and mask segmentation respectively. Previous works achieve guidance by simply introducing a single guidance module at a certain location in Mask R-CNN. Michaelis *et al.* [21] proposed to make Siamese the backbone network in the first stage to encode the guidance from support set. Consequently, all subsequent components for different tasks (including RPN and the three heads) undesirably have to share the same guidance. In [30], guidance is injected into Mask R-CNN at the front of the second stage by taking class-attentive vectors extracted from support set to reweight the feature maps, which enforces all second-stage components to share the same guidance and totally ignores the first-stage RPN.

In this paper, we present a Fully Guided Network (FGN) to address few-shot instance segmentation, as conceptually demonstrated in Fig. 1. FGN conjoins the few-shot learning paradigm with Mask R-CNN to establish a guided network. Different from prior works [21, 30], the key philosophy of FGN is that, *components for different tasks in Mask R-CNN should be guided differently to achieve full guidance* (which gives reason to the name of "Fully Guided Network"). Our intuition is that, the problem setting of FSIS brings different challenges to the various components in Mask R-CNN, which are difficult to be addressed by the use of a single guidance mechanism. Towards this end, FGN introduces three guidance mechanisms into Mask R-CNN, namely, the Attention-Guided RPN (AG-RPN), the Relation-Guided Detector (RG-DET) and the Attention-Guided FCN (AG-FCN), respectively. AG-RPN encodes the support set by class-aware attention, which is then utilized to guide RPN so that it can focus on the novel classes of concern and generate class-aware proposals. RG-DET guides the detector branch by an explicit comparison scheme to adapt to the inter-class generalization in FSIS. AG-FCN also takes attentional information from the support set to guide the mask segmentation procedure. Specific guidance modules are carefully designed and effective training strategy is suggested for model learning (see Figure 2 and Section 3 for details). Experimental results on public datasets demonstrate the proposed FGN can outperform the state-of-the-art FSIS approaches. In summary, the main contributions of our work are two-fold:

- We propose the Fully Guided Network, a novel framework for few-shot instance segmentation.
- We suggest three effective guidance mechanisms, *i.e.,* AG-RPN, RG-DET and AG-FCN, leading to superior performance.

2. Related Work

In this section, we briefly review the related literature.

Instance Segmentation. Instance segmentation can be viewed as a task at the intersection of semantic segmentation and object detection, which has made significant advances in recent years [10, 12, 24, 11, 19, 4, 3], benefited from deep CNN. Existing instance segmentation approaches are either proposal-based or proposal-free. The most representative work of the former category may be Mask R-CNN [12], which utilizes an RPN to generate class-independent object candidates in the first stage, and the second-stage procedure deals with these candidates only. Other influential works include [14, 19, 3]. The latter category of methods directly performs instance segmentation without relying on RPN, to balance between performance and computational efficiency. Representative works include [17, 7]. Instance segmentation has been mainly explored under the fully supervised setting so far, which may be impractical for certain applications.

Few-Shot Classification. FSL [29, 27] has recently emerged as a promising paradigm for learning predictive models from very limited training data (typically a handful of training samples only for each class). An external dataset with a large number of labeled data (but of different classes from the target ones) is usually necessitated, from which a set of episodes are sampled to simulate the target task. A conditional classifier is then learned from these episodes, which makes predictions conditioned on a support set. The conditional classifier is expected to be generalized well to the target task (on novel classes). A number of few-shot classification models have been proposed recently, including Matching Networks [29], Prototypical Networks [27], Relation Networks [28], the models based on Siamese CNN [16], graph CNN [8], *etc*. These models can be distinguished by how they encode and utilize the support set to guide the base network.

Few-Shot Semantic Segmentation. It is natural to consider adapting the FSL paradigm to other computer vision tasks, like semantic segmentation, object detection, *etc*. In light of the spirit of few-shot classification, Shaban *et al.* [1] proposed to utilize a conditioning branch to encode the support set and modulate an FCN-based segmentation branch to achieve one-shot semantic segmentation. Following a similar structure, some authors suggested differ-

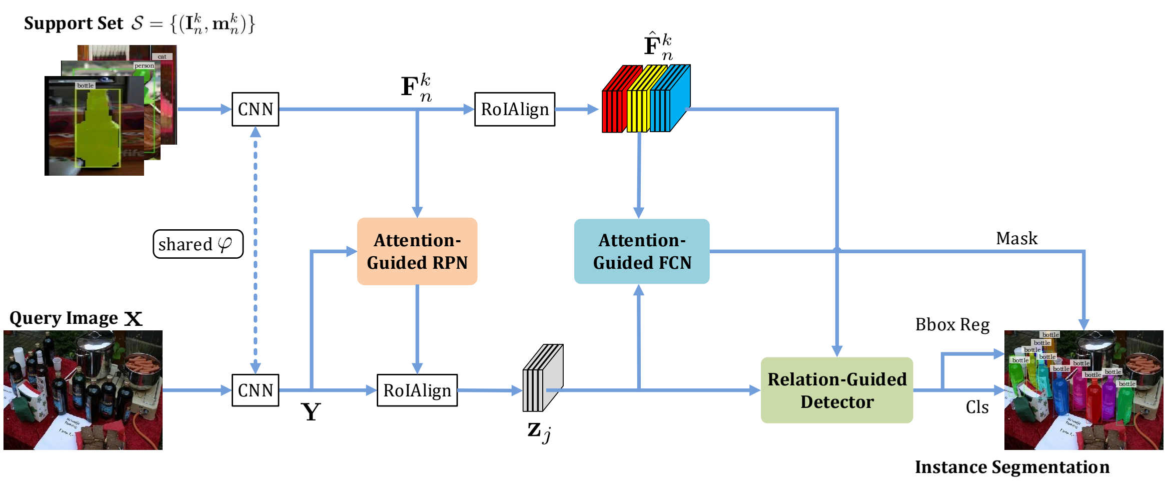

Figure 2. An overview of the proposed Fully Guided Network (FGN). FGN is established upon Mask R-CNN [12], where a support set is encoded and utilized to guide the three key components in Mask R-CNN, through the Attention-Guided RPN (AG-RPN), the Relation-Guided Detector (RG-DET) and the Attention-Guided FCN (AG-FCN), respectively.

ent schemes for encoding the support set or for imposing modulation on the segmentation branch [22, 32, 5].

Few-Shot Object Detection. It is more challenging to adapt FSL to object detection (termed as few-shot object detection) since object detection requires localization. Some works address this problem from the perspectives of self-paced learning [5] or transfer learning [2]. In [25], Schwartz *et al.* proposed to integrate a representative-based metric learning approach with the Faster R-CNN framework. In [15], Kang *et al.* presented a conditioned YOLO framework [23] with reweighted features for few shot object detection. These methods can only yield object bounding boxes, rather than instance masks.

Most closely related to ours, the works in [21, 30] consider FSIS by constructing guided networks upon Mask R-CNN. However, the overall performance is still limited, possibly due to the fact that, guidance driven by the support set cannot fully affect the base network as aforementioned. More effective guidance mechanisms for FSIS largely remain to be explored.

3. Approach

In this section, we start with the problem statement of few-shot instance segmentation. Then we describe the proposed Fully Guided Network, followed by the strategy for model training.

3.1. Problem Statement

Suppose for a set of *base classes* $\mathcal{C}^{\text{base}}$, we have a large set of images annotated with object instances, denoted by $\mathcal{D}^{\text{base}}$. Now let us consider a different set of semantic classes $\mathcal{C}^{\text{novel}}$ (called *novel classes*), which do not overlap with the base classes, *i.e.*, $\mathcal{C}^{\text{base}} \cap \mathcal{C}^{\text{novel}} = \phi$. For these novel classes, we only have a very limited number of annotated instances $\mathcal{D}^{\text{novel}}$, referred to as *support set*. In practice, this is usually due to difficulties in collecting images or acquiring instance-level annotations. The task of *few-shot instance segmentation (FSIS)* is to segment, from any given *query image* $\mathbf{I}^q$, all the object instances belonging to the novel classes. Note that when $|\mathcal{C}^{\text{novel}}| = N$ ($|\cdot|$ represents the cardinality of a set throughout this paper) and there are K annotated instances for each novel class, we call it an N-*way* K*-shot instance segmentation* task.

In this paper, we conjoin the few-shot learning paradigm with general instance segmentation to address the FSIS problem. Following the spirit of few-shot classification [29, 27], we simulate a quantity of N-way K-shot instance segmentation tasks $\mathcal{T} = \{(\mathcal{S}_i, \mathbf{x}_i)\}_{i=1}^{|\mathcal{T}|}$ by randomly sampling support sets and queries from $\mathcal{D}^{\text{base}}$ (of the base classes $\mathcal{C}^{\text{base}}$), where the i-th task is formed by sampling a support set $\mathcal{S}_i$ and a query image $\mathbf{x}_i$. By the use of these simulated tasks $\mathcal{T}$, we learn a conditional instance segmentation model $f_\theta(\mathbf{x}|\mathcal{S})$ parameterized by θ, which performs instance segmentation on the query image $\mathbf{x}$ conditioned on the support set $\mathcal{S}$. The learned model $f_\theta(\mathbf{x}|\mathcal{S})$ can then be applied to the target task, *i.e.*, N-way K-shot instance segmentation over the novel classes $\mathcal{C}^{\text{novel}}$ (simply letting $\mathcal{S} = \mathcal{D}^{\text{novel}}$ and $\mathbf{x} = \mathbf{I}^q$). It is worth pointing out that, instead of straightforwardly learning $f_\theta(\mathbf{x})$, our strategy is to learn a conditional model $f_\theta(\mathbf{x}|\mathcal{S})$, which can be viewed as to utilize the support set $\mathcal{S}$ to guide the instance segmentation of $\mathbf{x}$. The presence of guidance plays a critical role for the model trained on the base classes $\mathcal{C}^{\text{base}}$ to generalize well to the novel classes $\mathcal{C}^{\text{novel}}$.

3.2. Fully Guided Network

Central to any FSIS approach is how to effectively encode and utilize the support set to guide the basic in-

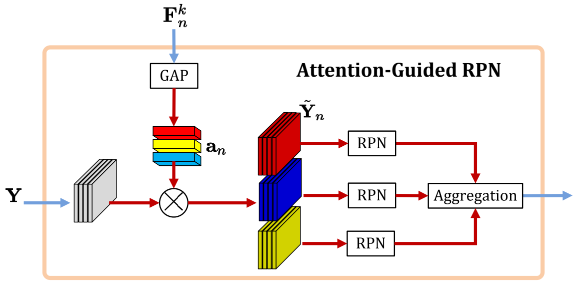

Figure 3. The structure of Attention-Guided RPN (AG-RPN).

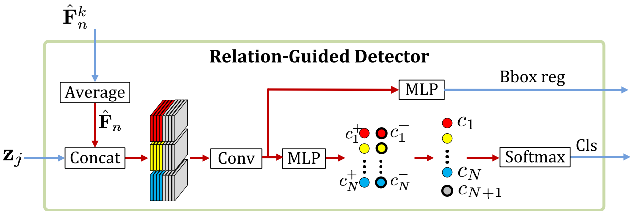

Figure 4. The structure of Relation-Guided Detector (RG-DET).

stance segmentation network (mostly typically Mask R-CNN [12]). Previous works fulfill such guidance by incorporating a single guidance module at a certain location in Mask R-CNN, which may undesirably enforce components for different tasks to share the same guidance [29], or neglect certain components [27]. We present the Fully Guided Network (FGN) in this paper, which is distinct from previous works [29, 27] in that, components for different tasks in Mask R-CNN are guided by the support set differently to achieve full guidance.

An overview of the proposed FGN is demonstrated in Fig. 2. Generally, FGN introduces guidance into Mask R-CNN at three key components, *i.e.*, the RPN, the detection branch (including classification and bbox regression) and the mask branches, leading to the Attention-Guided RPN (AG-RPN), the Relation-Guided Detector (RG-DET) and the Attention-Guided FCN (AG-FCN), respectively. In the proposed FGN, the given support set $\mathcal{S}$ (containing K annotated instances for each of the N classes) and the query image $\mathbf{x}$ are encoded by a shared backbone φ (ResNet101 [13] in our implementation) to give the feature maps $\mathbf{F}_n^k$, $\mathbf{Y} \in \mathbb{R}^{H\times W\times C}$ respectively. $\mathbf{F}_n^k$ encodes the support set, which is used by AG-RPN to guide the proposal generation from $\mathbf{Y}$ in the first stage. Then, in the second stage, for each proposal [also called Region-of-Interest (RoI)] with the aligned feature maps $\mathbf{z}_j \in \mathbb{R}^{h\times w\times C}$, the aligned $\hat{\mathbf{F}}_n^k \in \mathbb{R}^{h\times w\times C}$ is utilized by RG-DET to guide the classification and bbox heads, and by AG-FCN to guide the mask head. Another key contribution of our work is to design novel and effective guidance mechanisms for these modules, which are detailed as below.

Attention-Guided RPN. Mask R-CNN relies on RPN to obtain class-agnostic proposals of potential objects for subsequent processing. Under the problem setting of FSIS, RPN has to be trained on the base classes $\mathcal{C}^{\text{base}}$ and tested on a solely different set of novel classes $\mathcal{C}^{\text{novel}}$. In this case, RPN may generate a lot of undesired proposals but miss the ones of concern, especially when $\mathcal{C}^{\text{novel}}$ departs far from $\mathcal{C}^{\text{base}}$, or the number of novel classes is small, which will largely degrade overall performance. To tackle this issue, our idea is to introduce guidance from the support set into RPN such that it can focus on the classes of concern and generate class-aware proposals, which we call *Attention-Guided RPN (AG-RPN)*.

The structure of AG-RPN is depicted in Fig. 3. The feature maps $\mathbf{F}_n^k \in \mathbb{R}^{H\times W\times C}$ with $n = 1, ..., N, k = 1, ..., K$, which encode the support set, undergo the global average pooling (GAP) and the averaging operation over each individual class, given by

$$\mathbf{a}_n = \frac{1}{K}\sum_{k=1}^{K} \text{GAP}\left(\mathbf{F}_n^k\right), \quad n = 1, ..., N, \qquad (1)$$

with $\{\mathbf{a}_1, ..., \mathbf{a}_N\} \in \mathbb{R}^{C\times 1}$ being the *class-attentive vectors* associated with the N novel classes. Each $\mathbf{a}_n$ is then taken to weight the feature maps of the query image $\mathbf{Y} \in \mathbb{R}^{H\times W\times C}$ as below

$$\tilde{\mathbf{Y}}_n = \mathbf{Y} \otimes \mathbf{a}_n, \quad n = 1, ..., N, \qquad (2)$$

which means taking $\mathbf{a}_n$ to perform element-wise multiplication along the channel dimension at every spatial location in $\mathbf{Y}$. Each $\tilde{\mathbf{Y}}_n$ is fed into the basic RPN for proposal generation independently and the results are then aggregated to yield the final proposals. The aggregation procedure can be described as follows: For each particular anchor, an objectness score can be acquired through the RPN over every $\tilde{\mathbf{Y}}_n$, and the softmax results over the N scores are taken as the class-aware confidence of the anchor. Anchor refinement is conducted by the regression corresponding to the top matching score during inference. The final proposals are picked up from the anchors by thresholding their confidence and performing non-maximal suppression.

Relation-Guided Detector. The guidance on the detector branch in Mask R-CNN (including the classification and bbox regression heads) is imposed in an implicit way in previous works [21, 30], which just simply modulate the feature extraction in the first or second stage by the use of support set. In this paper, we propose a different guidance mechanism for the detector (actually the classification branch), termed as *Relation-Guided Detector (RG-DET)*. RG-DET achieves guidance by explicitly comparing the features extracted from the support set and the RoI, inspired by the Relation Network (RN) [28] originally proposed for few-shot classification. We favor RN mainly because it is characterized by that, both the feature embedding

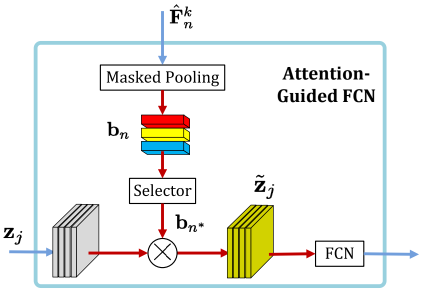

Figure 5. The structure of Attention-Guided FCN.

and the similarity measure are learnable, compared to other competitors like [29, 27, 16].

Unfortunately, RN cannot be directly deployed to our task because there exists an essential difference between our problem here and the general few-shot classification, that is, the rejection of background class. RG-DET operates on individual RoIs output by AG-RPN, which may inevitably contain background RoIs belonging to neither of the novel classes in the support set. By contrast, recall that few-shot classification methods (including RN) always classifies the query to be one of the classes indicated by the support set. Taking into account the background rejection issue, the structure of RG-DET is illustrated in Fig. 4.

For a particular RoI, its aligned feature maps $\mathbf{z}_j \in \mathbb{R}^{h \times w \times C}$ are concatenated with the N aligned feature maps $\hat{\mathbf{F}}_n = \left(\frac{1}{K}\sum_k \hat{\mathbf{F}}_n^k\right) \in \mathbb{R}^{h \times w \times C}$ extracted from the support set (as shown in Fig. 4), followed by a stack of conv and fc layers (termed as MLP), to give the matching scores (the cls branch) and the object box (the bbox reg branch). The matching score between $\mathbf{z}_j$ and the i-th feature maps $\hat{\mathbf{F}}_n$ is represented by a doublet (c_i^+, c_i^-), where c_i^+ and c_i^- stand for the confidence of matching the i-th class and the background respectively. To enable background rejection, we need to derive an $(N+1)$-length matching vector $\mathbf{c} = (c_1, ..., c_N, c_{N+1})$ from the $2N$ original scores, with $c_i, i = 1, ..., N$ reflecting the confidence of the i-th class and c_{N+1} the background. For this purpose, we set $c_i = c_i^+$ and $c_{N+1} = c_{i^*}^-$ with $i^* = \arg\max_i \left\{c_i^+\right\}$, which physically means we depend on the best-matched class (the most reliable one) to estimate the confidence of background c_{N+1}. A softmax operation is then performed over the matching vector $\mathbf{c}$, yielding the final classification score.

The bbox regression branch shares the concatenation and the first conv layer with the classification branch, but has a separate MLP layer as shown in Fig. 4.

Attention-Guided FCN. As illustrated in Fig. 5, the *Attention-Guided FCN (AG-FCN)* introduces guidance into the FCN-based mask head. AG-FCN basically follows the guidance scheme for few-shot semantic segmentation [26], except for two modifications. First, an operation of *masked pooling* [32] is performed on the aligned feature vectors $\hat{\mathbf{F}}_n^k \in \mathbb{R}^{h \times w \times C}$ before computing the class-attentive vectors $\{\mathbf{b}_1, ..., \mathbf{b}_N\} \in \mathbb{R}^{C \times 1}$ as described in Eq. (1). Masked pooling on $\hat{\mathbf{F}}_n^k$ means pooling $\hat{\mathbf{F}}_n^k$ within the binary mask $\hat{\mathbf{m}}_n^k \in \mathbb{R}^{h \times w \times C}$, which is obtained by performing RoIAlign over the original instance mask $\mathbf{m}_n^k \in \mathbb{R}^{H \times W \times C}$. Second, a *selector* is used to pick up the one $\mathbf{b}_{n^*}$ from $\{\mathbf{b}_1, ..., \mathbf{b}_N\}$, where n^* is chosen to be the ground truth class for training, and the one with the highest classfication score for testing. Note that $\tilde{\mathbf{z}}_j = \mathbf{z}_j \otimes \mathbf{b}_{n^*}$ where the operator $\otimes$ is identical to that in Eq. (2).

3.3. Training Strategy

FGN is a two-stage structure since it is based on Mask R-CNN. Hence, our pipeline for training is basically similar to Mask R-CNN (including the loss functions). But differently, following the common practice in [2, 15, 30], our training includes two steps. For the first step, we purely take $\mathcal{D}^{\text{base}}$ of the base classes $\mathcal{C}^{\text{base}}$ as the training data. And for the second step, we take data from both the base classes and the novel classes, *i.e.*, $\mathcal{C}^{\text{base}} \cup \mathcal{C}^{\text{novel}}$, to further fine-tune the model. More precisely, the second-step training data consist of the whole support set $\mathcal{D}^{\text{novel}}$ (containing NK instances) and $3K$ instances for each class in $\mathcal{C}^{\text{base}}$ randomly sampled from $\mathcal{D}^{\text{base}}$, which contain totally $(N+3|\mathcal{C}^{\text{base}}|)K$ instances. Our training requires randomly sampling the training set to simulate the target FSIS tasks (constructing the episodes), which will be detailed in Section 4.1.

4. Experiments and Results

In this section, we present experimental results to evaluate the effectiveness of our method, mainly including: 1) comparison with the state-of-the-art methods; 2) ablation study with several variant baselines. Our method was implemented in TensorFlow and Keras on a workstation with 4 NVIDIA Titan XP GPUs.

4.1. Experimental Settings

We adopt two commonly-used datasets for our experiments, *i.e.*, Microsoft COCO 2017 [18] and PASCAL VOC 2012 [6] (termed as **COCO** and **VOC** respectively). COCO has 80 object classes, consisting of a training set (trainset) with $118,287$ images and a validation set (valset) with $4,952$ images. VOC covers 20 classes that are a subset of COCO's 80 classes, with a trainset of $1,464$ images (annotated with instance masks) and a valset of $1,449$ images.

General Settings. According to the problem definition in Section 3.1, our evaluation requires the following basic settings: *1) Setting the base classes $\mathcal{C}^{base}$ and the novel classes $\mathcal{C}^{novel}$, and accordingly the training set $\mathcal{D}^{base}$ and the query set $\mathcal{D}^{novel}$ (testing set):* As our main setting, we adopt a challenging cross-dataset setting to better compare the generalization ability of various models, inspired by pre-

Methods	Segmentation			Detection		
	1way-1shot	3way-1shot	3way-3shot	1way-1shot	3way-1shot	3way-3shot
MRCNN-FT	0.4	0.5	2.7	6.0	5.2	10.2
Siamese MRCNN [21]	13.8	6.3	6.6	23.9	11.5	13.3
Meta R-CNN [30]	12.5	12.1	15.3	20.1	19.2	23.4
FGN	**16.2**	**13.0**	**17.9**	**30.8**	**23.5**	**32.9**

Table 1. Performance in terms of mAP$_{50}$ obtained by various methods under the **COCO2VOC** setting. Both the segmentation and detection results are reported for comparison.

Methods	Segmentation			Detection		
	1way-1shot	3way-1shot	3way-3shot	1way-1shot	3way-1shot	3way-3shot
MRCNN-FT	25.3	25.0	27.4	27.3	27.1	29.7
Siamese MRCNN [21]	24.2	8.8	9.1	26.4	9.7	10.1
Meta R-CNN [30]	14.9	14.1	15.2	18.5	17.8	19.3
FGN	24.2	13.2	14.3	27.2	16.7	17.3

Table 2. Addition experimental results to demonstrate the challenges of the FSIS problem setting. In this experiment, the settings of $\mathcal{C}^{\text{base}}$ and $\mathcal{D}^{\text{base}}$ are identical to those in COCO2VOC, but $\mathcal{C}^{\text{novel}} \subset \mathcal{C}^{\text{base}}$ and the testing tasks are sampled from COCO's validation set.

vious works [15, 30]. Specifically, we set the 20 classes at the intersection of COCO and VOC to be $\mathcal{C}^{\text{novel}}$ and the rest 60 classes covered by COCO but not VOC to be $\mathcal{C}^{\text{base}}$. Further, we take from COCO's trainset the subset belonging to $\mathcal{C}^{\text{base}}$ as the training set $\mathcal{D}^{\text{base}}$, and take VOC's valset (belonging to the 20 novel classes $\mathcal{C}^{\text{novel}}$) to construct the testing set (see details later). We refer to this main experimental setting as **COCO2VOC**. Additionally, we also consider another setting termed as **VOC2VOC**, which only uses the VOC dataset. More precisely, we randomly sample 15 out of 20 classes covered by VOC to be the base classes $\mathcal{C}^{\text{base}}$ and the rest 5 are taken as $\mathcal{C}^{\text{novel}}$. The training set $\mathcal{D}^{\text{base}}$ and the query set $\mathcal{D}^{\text{novel}}$ are constructed respectively from VOC's trainset and valset. *2) Specifying the numbers of N and K:* We consider three different settings (a) $N = 1, K = 1$ (termed as **1way-1shot**); (b) $N = 3, K = 1$ (termed as **3way-1shot**); (c) $N = 3, K = 3$ (termed as **3way-3shot**).

Methods for Comparison. To our knowledge, there exist only two FSIS methods in the literature so far, *i.e.*, **Siamese MRCNN** [21] and **Meta R-CNN** [30], which are included in our comparison. Similar to our FGN, Siamese MRCNN and Meta R-CNN also achieve FSIS by introducing guidance into Mask R-CNN (but using different guidance mechanisms), for which we use the source codes released by the authors for our experiments. Besides, we also build a baseline for comparison, termed as **MRCNN-FT**, which is basically a Mask R-CNN trained with the strategy detailed in Section 3.3.

Implementation Details. We follow the training strategy in Section 3.3 and the settings of $\{\mathcal{C}^{\text{base}}, \mathcal{D}^{\text{base}}, \mathcal{C}^{\text{novel}}, \mathcal{D}^{\text{novel}}, N, K\}$ above in Section 4.1 to train our FGN model. We use ResNet101 [13] as the backbone for our model. The initial learning rates of SGD for training the first-stage AG-RPN and the second-stage RG-DET and AG-FCN are set to 0.01 and 0.001 respectively. We train for $60,000$ steps and a 10-times learning rate decay is applied to the second-half steps.

To construct the simulated tasks $\mathcal{T} = \{(\mathcal{S}_i, \mathbf{x}_i)\}_{i=1}^{|\mathcal{T}|}$ (typically called "episodes") for training, we basically follow the sampling strategy proposed in [29]. Note that, we crop the local patches extended by 20 pixels around ground truth boxes of instances to form the support set, rather than using holistic images. And for testing, the tasks $\{(\mathcal{D}_i^{\text{novel}}, \mathbf{I}_i^q)\}_i$ are constructed to ensure every novel class in every image in the testing set is tested for once. Specifically, for each image $\mathbf{I}_i^q$, we collect all the classes it covers. Then, for each class we randomly sample other $N - 1$ classes and pick up instances accordingly to form an N-way K-shot episode. We report the average performance over all the testing tasks.

4.2. Results

We present the main results under the settings of COCO2VOC and VOC2VOC and related analysis respectively in the following.

COCO2VOC. The FSIS performance obtained by the various methods under the COCO2VOC setting is comparatively reported in Table 1, where we use mAP$_{50}$ as the quantitative performance measure. As can be observed that, our FGN can generally outperform the two state-of-the-art methods Siamese MRCNN [21] and Meta R-CNN [30] to a large margin for the three settings of N and K. Siamese MRCNN [21] performs comparatively to ours in case of 1way-1shot, but degrades heavily under the other two settings. This is probably because that, the guidance in this approach follows the Siamese Network mechanism which is originally designed for pairwise input. Meta R-CNN [30] does not perform well either, probably because this method relies much on the finetuning procedure in training, which cannot acquire sufficient data for finetuning when N and

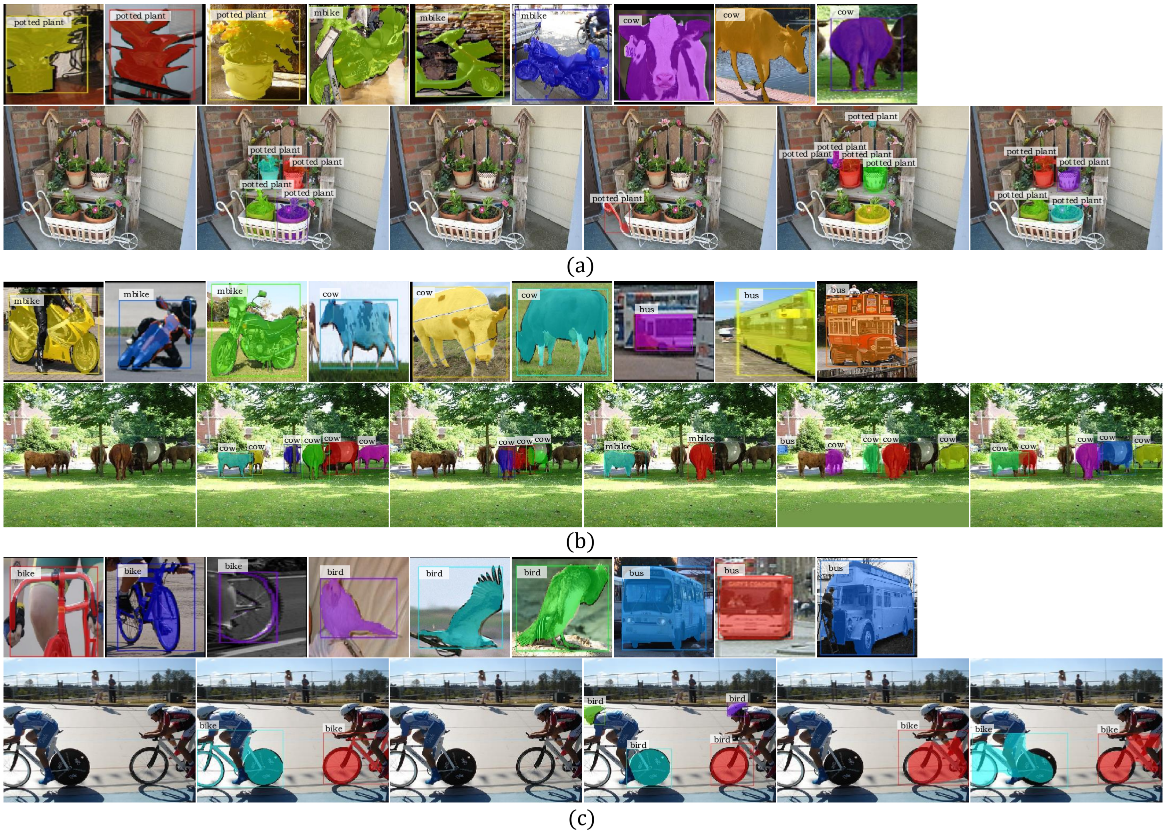

Figure 6. Exemplary results obtained by various results under the **COCO2VOC 3way-3shot** setting. In each group (a) - (c), the images in the top row are the support set. And in the bottom row, from left to right are the query image, the ground truth, and the results obtained by MRCNN-FT, Siamese MRCNN [21], Meta R-CNN [30] and our FGN.

Methods	Segmentation			Detection		
	1way-1shot	3way-1shot	3way-3shot	1way-1shot	3way-1shot	3way-3shot
Siamese MRCNN [21]	8.2	4.4	5.2	**17.9**	8.7	9.0
Meta R-CNN [30]	4.2	3.6	7.3	8.0	7.3	14.4
FGN	**8.4**	**7.3**	**9.6**	15.4	**11.3**	**16.2**

Table 3. Performance in terms of mAP$_{50}$ obtained by various methods under the **VOC2VOC** setting. Both the segmentation and detection results are reported for comparison.

K are small like in our settings. As expected, the baseline MRCNN-FT performs very poorly, which suggests that the strategy of naively finetuning a model pretained from base classes with data from novel classes is inappropriate for FSIS.

In addition to segmentation, we also compare the various methods on the task of few-shot object detection, as shown in Table 1. Our FGN can also outperform the other methods consistently for all the settings. One can further observe that, there is an obvious performance drop from detection to segmentation for all the methods, which may indicate that FSIS cannot be achieved by trivial extension of few-shot object detection methods. We also provide some exemplary results obtained by various methods for visual comparison in Fig. 6.

While the proposed FGN can outperform the state-of-the-art as stated above, one may be concerned with a fact that, the performance of various methods (including ours) generally looks limited, significantly worse than conventional instance segmentation. We argue this is likely due to the intrinsic challenges of the FSIS problem, especially in case of low numbers of ways and shots like ours. To justify this point, we further carry out another experiment where the settings of $\mathcal{C}^{\text{base}}$ and $\mathcal{D}^{\text{base}}$ are identical to those in COCO2VOC, but the novel classes $\mathcal{C}^{\text{novel}} \subset \mathcal{C}^{\text{base}}$ and the testing tasks are sampled from COCO's validation set (the data used for testing are different). Such case where $\mathcal{C}^{\text{novel}} \subset \mathcal{C}^{\text{base}}$ does not coincide with the problem definition of FSIS but general instance segmentation. Also, MRCNN-FT is a Mask R-CNN trained by the common strategy de-

	AG-RPN	RG-DET	AG-FCN	**Segmentation**	**Detection**
FGN-P	✓			13.7	23.8
FGN-DS		✓	✓	15.1	26.8
FGN-PS	✓		✓	15.6	24.8
FGN-PD	✓	✓		15.1	29.1
FGN (Ours)	✓	✓	✓	**17.9**	**32.9**

Table 4. Ablation study on the effectiveness of full guidance. Comparison among the variants of FGN in terms of mAP$_{50}$.

RPN	**AG-RPN-v1**	**AG-RPN**
64.5	74.8	**92.5**

Table 5. Comparison among the variants of AG-RPN in terms of AR$_{50}$.

FCN	**AG-FCN-v1**	**AG-FCN-v2**	**AG-FCN**
15.1	14.5	15.6	**17.9**

Table 6. Comparison among the variants of AG-FCN in terms of mAP$_{50}$.

scribed in Section 3.3, which is shared by all the compared methods (including ours). As shown in Table 2, under the setting of general instance segmentation, even the standard Mask R-CNN trained in the same fashion as commonly required by FSIS approaches can only achieve limited performance. This may reflect that, the FSIS problem setting is inherently challenging, and the training strategy adopted by these FSIS methods (including our FGN) is effective in this sense. It is worth noticing that, it is not meaningful to make comparison among the various methods under this experimental setting.

VOC2VOC. In addition to our main setting of COCO2VOC, we also evaluate under the VOC2VOC setting. The results obtained by various methods in terms of mAP$_{50}$ are listed in Table 3. Although VOC2VOC shares the same validation set as COCO2VOC, it has a far smaller training set ($\sim$ 1.4K in contrast to $\sim$ 118K images). As a result, the performance of VOC2VOC is worse than that of COCO2VOC for all the methods. In this case, our FGN can still achieve the best overall performance among the compared methods for both segmentation and detection.

4.3. Ablation Study

We perform ablation study to further reveal the merits of our FGN. All the following experiments are conducted under the **COCO2VOC 3way-3shot** setting.

Full Guidance. One key reason of FGN's effectiveness is that we carefully design three guidance mechanisms, *i.e.*, AG-RPN (P), RG-DET (D) and AG-FCN (S) to achieve full guidance. To verify the contributions of these modules, we construct several variants by disabling one or more modules from the full FGN model.

The results obtained by these variants in terms of mAP$_{50}$ for segmentation and detection are comparatively reported in Table 4. It can be seen from the degraded performance of these variants that, each module contributes to some extent on both tasks.

AG-RPN. We compare our AG-RPN with the basic RPN in Mask R-CNN and a variant termed as **AG-RPN-v1** by evaluating separately the quality of the proposals generated. AG-RPN-v1 follows the design in [21] to achieve guidance. As can be observed from Table 5 that, AG-RPN (ours) obtains the best performance in terms of AR$_{50}$.

AG-FCN. We construct two variants of AG-FCN (ours) for comparison, termed as **AG-FCN-v1** and **AG-FCN-v2**. AG-FCN-v1 is the FCN guidance mechanism suggested in [32] for the task of semantic segmentation. AG-FCN-v2 tiles the channel attention vectors $\mathbf{b}_{n^*}$ to be of the same size as $\mathbf{z}_j$ and then concatenates them together (see Fig. 5). We also include the basic FCN used by Mask R-CNN (without guidance) for comparison. As can be seen from Table 6, AG-FCN (ours) performs the best among all the variants.

5. Conclusion

In this paper, we have presented the Fully Guided Network (FGN), a novel network to address few-shot instance segmentation. FGN can be viewed as a guided network where a support set is encoded and utilized to guide the base network, *i.e.*, Mask R-CNN. Compared to previous works, FGN is characterized by introducing different guidance mechanisms into the three key components in Mask R-CNN to make full use of the guidance effect of support set. Comparative experiments on public datasets have demonstrated that FGN can outperform state-of-the-art methods. Ablation study has also been conducted to further verify the effectiveness of FGN. Despite the superiority of FGN over previous works, few-shot instance segmentation by nature is a very challenging task and there is still large room for improvement, especially on classification branch where more complicated features and background rejection are engaged. In future work, we will explore new guidance mechanisms to further boost the overall performance.

Acknowledgement

This work was supported by the National Natural Science Foundation of China under Grant 61703166 and Grant 61633010, the Guangdong Natural Science Foundation under Grant 2014A030312005, the Key R&D Program of Guangdong Province under Grant 2018B030339001, the Guangzhou Science and Technology Program under Grant 201904010299, and the Fundamental Research Funds for the Central Universities, SCUT, under Grant 2018MS72.

References

[1] Zhen Liu Irfan Essa Byron Boots, Amirreza Shaban, and Shray Bansal. One-shot learning for semantic segmentation. In *British Machine Vision Conference*, 2017. 4322

[2] Hao Chen, Yali Wang, Guoyou Wang, and Yu Qiao. Lstd: A low-shot transfer detector for object detection. In *AAAI Conference on Artificial Intelligence*, 2018. 4323, 4325

[3] Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, et al. Hybrid task cascade for instance segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4974–4983, 2019. 4321, 4322

[4] Liang-Chieh Chen, Alexander Hermans, George Papandreou, Florian Schroff, Peng Wang, and Hartwig Adam. Masklab: Instance segmentation by refining object detection with semantic and direction features. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4013–4022, 2018. 4321, 4322

[5] Nanqing Dong and Eric Xing. Few-shot semantic segmentation with prototype learning. In *British Machine Vision Conference*, 2018. 4321, 4323

[6] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010. 4325

[7] Cheng-Yang Fu, Mykhailo Shvets, and Alexander C Berg. Retinamask: Learning to predict masks improves state-of-the-art single-shot detection for free. *arXiv preprint arXiv:1901.03353*, 2019. 4322

[8] Victor Garcia and BrunaJoan. Few-shot learning with graph neural networks. In *International Conference on Learning Representations*, 2018. 4321, 4322

[9] Saurabh Gupta, Ross Girshick, Pablo Arbeláez, and Jitendra Malik. Learning rich features from rgb-d images for object detection and segmentation. In *European Conference on Computer Vision*, pages 345–360, 2014. 4321

[10] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Simultaneous detection and segmentation. In *European Conference on Computer Vision*, pages 297–312, 2014. 4321, 4322

[11] Zeeshan Hayder, Xuming He, and Mathieu Salzmann. Boundary-aware instance segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5696–5704, 2017. 4321, 4322

[12] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *IEEE International Conference on Computer Vision*, pages 2961–2969, 2017. 4321, 4322, 4323, 4324

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 4322, 4324, 4326

[14] Zhaojin Huang, Lichao Huang, Yongchao Gong, Chang Huang, and Xinggang Wang. Mask scoring r-cnn. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6409–6418, 2019. 4322

[15] Bingyi Kang, Zhuang Liu, Xin Wang, Fisher Yu, Jiashi Feng, and Trevor Darrell. Few-shot object detection via feature reweighting. In *IEEE International Conference on Computer Vision*, pages 8420–8429, 2019. 4323, 4325, 4326

[16] Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. Siamese neural networks for one-shot image recognition. In *International Conference on Machine Learning Workshops*, volume 2, 2015. 4321, 4322, 4325

[17] Xiaodan Liang, Liang Lin, Yunchao Wei, Xiaohui Shen, Jianchao Yang, and Shuicheng Yan. Proposal-free network for instance-level object segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2978–2991, 2017. 4322

[18] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755, 2014. 4325

[19] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8759–8768, 2018. 4321, 4322

[20] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015. 4322

[21] Claudio Michaelis, Ivan Ustyuzhaninov, Matthias Bethge, and Alexander S Ecker. One-shot instance segmentation. *arXiv preprint arXiv:1811.11507*, 2018. 4322, 4323, 4324, 4326, 4327, 4328

[22] 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*, 2018. 4321, 4323

[23] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 779–788, 2016. 4323

[24] Mengye Ren and Richard S Zemel. End-to-end instance segmentation with recurrent attention. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6656–6664, 2017. 4321, 4322

[25] Eli Schwartz, Leonid Karlinsky, Joseph Shtok, Sivan Harary, Mattias Marder, Sharathchandra Pankanti, Rogerio Feris, Abhishek Kumar, Raja Giries, and Alex M Bronstein. Repmet: Representative-based metric learning for classification and one-shot object detection. *arXiv preprint arXiv:1806.04728*, 2018. 4323

[26] Amirreza Shaban, Shray Bansal, Zhen Liu, Irfan Essa, and Byron Boots. One-shot learning for semantic segmentation. *arXiv preprint arXiv:1709.03410*, 2017. 4321, 4325

[27] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Neural Information Processing Systems*, pages 4077–4087, 2017. 4321, 4322, 4323, 4324, 4325

[28] 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, pages 1199–1208, 2018. 4321, 4322, 4324

[29] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *Neural Information Processing Systems*, pages 3630–3638, 2016. 4321, 4322, 4323, 4324, 4325, 4326

[30] 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*, pages 9577–9586, 2019. 4322, 4323, 4324, 4325, 4326, 4327

[31] Jin-Gang Yu, Yansheng Li, Changxin Gao, Hongxia Gao, Gui-Song Xia, Zhu Liang Yu, and Yuanqing Li. Exemplar-based recursive instance segmentation with application to plant image analysis. *IEEE Transactions on Image Processing*, 29:389–404, 2019. 4321

[32] Xiaolin Zhang, Yunchao Wei, Yi Yang, and Thomas Huang. Sg-one: Similarity guidance network for one-shot semantic segmentation. *arXiv preprint arXiv:1810.09091*, 2018. 4321, 4323, 4325, 4328

[33] 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*, pages 669–677, 2016. 4321