

# Few-Shot Segmentation Without Meta-Learning: A Good Transductive Inference Is All You Need?

Malik Boudiaf \*  
ÉTS Montreal

Hoel Kervadec  
ÉTS Montreal

Ziko Imtiaz Masud  
ÉTS Montreal

Pablo Piantanida  
CentraleSupélec-CNRS  
Université Paris-Saclay

Ismail Ben Ayed  
ÉTS Montreal

Jose Dolz  
ÉTS Montreal

## Abstract

We show that the way inference is performed in few-shot segmentation tasks has a substantial effect on performances—an aspect often overlooked in the literature in favor of the meta-learning paradigm. We introduce a transductive inference for a given query image, leveraging the statistics of its unlabeled pixels, by optimizing a new loss containing three complementary terms: i) the cross-entropy on the labeled support pixels; ii) the Shannon entropy of the posteriors on the unlabeled query-image pixels; and iii) a global KL-divergence regularizer based on the proportion of the predicted foreground. As our inference uses a simple linear classifier of the extracted features, its computational load is comparable to inductive inference and can be used on top of any base training. Foregoing episodic training and using only standard cross-entropy training on the base classes, our inference yields competitive performances on standard benchmarks in the 1-shot scenarios. As the number of available shots increases, the gap in performances widens: on PASCAL-5<sup>i</sup>, our method brings about 5% and 6% improvements over the state-of-the-art, in the 5- and 10-shot scenarios, respectively. Furthermore, we introduce a new setting that includes domain shifts, where the base and novel classes are drawn from different datasets. Our method achieves the best performances in this more realistic setting. Our code is freely available online: <https://github.com/mboudiaf/RePRI-for-Few-Shot-Segmentation>.

## 1. Introduction

Few-shot learning, which aims at classifying instances from unseen classes given only a handful of training ex-

amples, has witnessed a rapid progress in the recent years. To quickly adapt to novel classes, there has been a substantial focus on the meta-learning (or learning-to-learn) paradigm [27, 31, 35]. Meta-learning approaches popularized the need of structuring the training data into *episodes*, thereby simulating the tasks that will be presented at inference. Nevertheless, despite the achieved improvements, several recent image classification works [2, 4, 6, 12, 32, 44] observed that meta-learning might have limited generalization capacity beyond the standard 1- or 5-shot classification benchmarks. For instance, in more realistic settings with domain shifts, simple classification baselines may outperform much more complex meta-learning methods [4, 12].

Deep-learning based semantic segmentation has been generally nurtured from the methodological advances in image classification. Few-shot segmentation, which has gained popularity recently [10, 17, 19, 23, 25, 33, 36, 37, 38, 39, 41, 42], is no exception. In this setting, a deep segmentation model is first pre-trained on *base* classes. Then, model generalization is assessed over few-shot *tasks* and novel classes unseen during base training. Each task includes an unlabeled test image, referred to as the *query*, along with a few labeled images (the *support* set). The recent literature in few-shot segmentation follows the learning-to-learn paradigm, and substantial research efforts focused on the design of specialized architectures and episodic-training schemes for base training. However, i) episodic training itself implicitly assumes that testing tasks have a structure (e.g., the number of support shots) similar to the tasks used at the meta-training stage; and ii) base and novel classes are often assumed to be sampled from the same dataset.

In practice, those assumptions may limit the applicability of the existing few-shot segmentation methods in realistic scenarios [3, 4]. In fact, our experiments proved consistent with findings in few-shot classification when going beyond the standard settings and benchmarks. Particularly, we ob-

\*Corresponding author: malik.boudiaf.1@etsmtl.net

served among state-of-the-art methods a saturation in performances [3] when increasing the number of labeled samples (See Table 3). Also, in line with very recent observations in image classification [4], existing meta-learning methods prove less competitive in cross-domains scenarios (See Table 4). This casts doubts as to the viability of the current few-shot segmentation benchmarks and datasets; and motivates re-considering the relevance of the meta-learning paradigm, which has become the *de facto* choice in the few-shot segmentation literature.

## Contributions

In this work, we forego meta-learning, and re-consider a simple cross-entropy supervision during training on the base classes for feature extraction. Additionally, we propose a *transductive* inference that better leverages the support-set supervision than the existing methods. Our contributions can be summarized as follows:

- We present a new transductive inference—*RePRI* (Region Proportion Regularized Inference)—for a given few-shot segmentation task. RePRI optimizes a loss integrating three complementary terms: *i*) a standard cross-entropy on the labeled pixels of the support images; *ii*) the entropy of the posteriors on the query pixels of the test image; and *iii*) a global KL divergence regularizer based on the proportion of the predicted foreground pixels within the test image. RePRI can be used on top of any trained feature extractor, and uses exactly the same information as standard inductive methods for a given few-shot segmentation task.
- Although we use a basic cross-entropy training on the base classes, without complex meta-learning schemes, RePRI yields highly competitive performances on the standard few-shot segmentation benchmarks, PASCAL-5<sup>i</sup> and COCO-20<sup>i</sup>, with gains around 5% and 6% over the state-of-the-art in the 5- and 10-shot scenarios, respectively.
- We introduce a more realistic setting where, in addition to the usual shift on classes between training and testing data distributions, a shift on the images’ feature distribution is also introduced. Our method achieves the best performances in this scenario.
- We demonstrate that a precise region-proportion information on the query object improves substantially the results, with an average gain of 13% on both datasets. While assuming the availability of such information is not realistic, we show that inexact estimates can still lead to drastic improvements, opening a very promising direction for future research.

## 2. Related Work

**Few-Shot Learning for classification** Meta-learning has become the *de facto* solution to learn novel tasks from a few labeled samples. Even though the idea is not new [28], it has been revived recently by several popular works in few-shot classification [9, 26, 27, 31, 35]. These works can be categorized into gradient- or metric-learning-based methods. Gradient approaches resort to stochastic gradient descent (SGD) to learn the commonalities among different tasks [26, 9]. Metric-learning approaches [35, 31] adopt deep networks as feature-embedding functions, and compare the distances between the embeddings. Furthermore, in a recent line of works, the transductive setting has been investigated for few-shot classification [6, 2, 14, 16, 20, 24, 31, 44], and yielded performance improvements over inductive inference. These results are in line with established facts in classical transductive inference [34, 15, 5], well-known to outperform its inductive counterpart on small training sets. To a large extent, these transductive classification works follow well-known concepts in semi-supervised learning, such as graph-based label propagation [20], entropy minimization [6] or Laplacian regularization [44]. While the entropy is a part of our transductive loss, we show that it is not sufficient for segmentation tasks, typically yielding trivial solutions.

**Few-shot segmentation** Segmentation can be viewed as a classification at the pixel level, and recent efforts mostly went into the design of specialized architectures. Typically, the existing methods use a two-branch comparison framework, inspired from the very popular prototypical networks for few-shot classification [31]. Particularly, the support images are employed to generate class prototypes, which are later used to segment the query images via a prototype-query comparison module. Early frameworks followed a dual-branch architecture, with two independent branches [29, 7, 25], one generating the prototypes from the support images and the other segmenting the query images with the learned prototypes. More recently, the dual-branch setting has been unified into a single-branch, employing the same embedding function for both the support and query sets [42, 30, 37, 38, 21]. These approaches mainly aim at exploiting better guidance for the segmentation of query images [42, 23, 36, 40], by learning better class-specific representations [37, 19, 21, 38, 30] or iteratively refining these [41]. Graph CNNs have also been employed to establish more robust correspondences between the support and query images, enhancing the learned prototypes [36]. Alternative solutions to learn better class representations include: imprinting the weights for novel classes [30], decomposing the holistic class representation into a set of part-aware prototypes [21] or mixing several prototypes, each corresponding to diverse image regions [38].

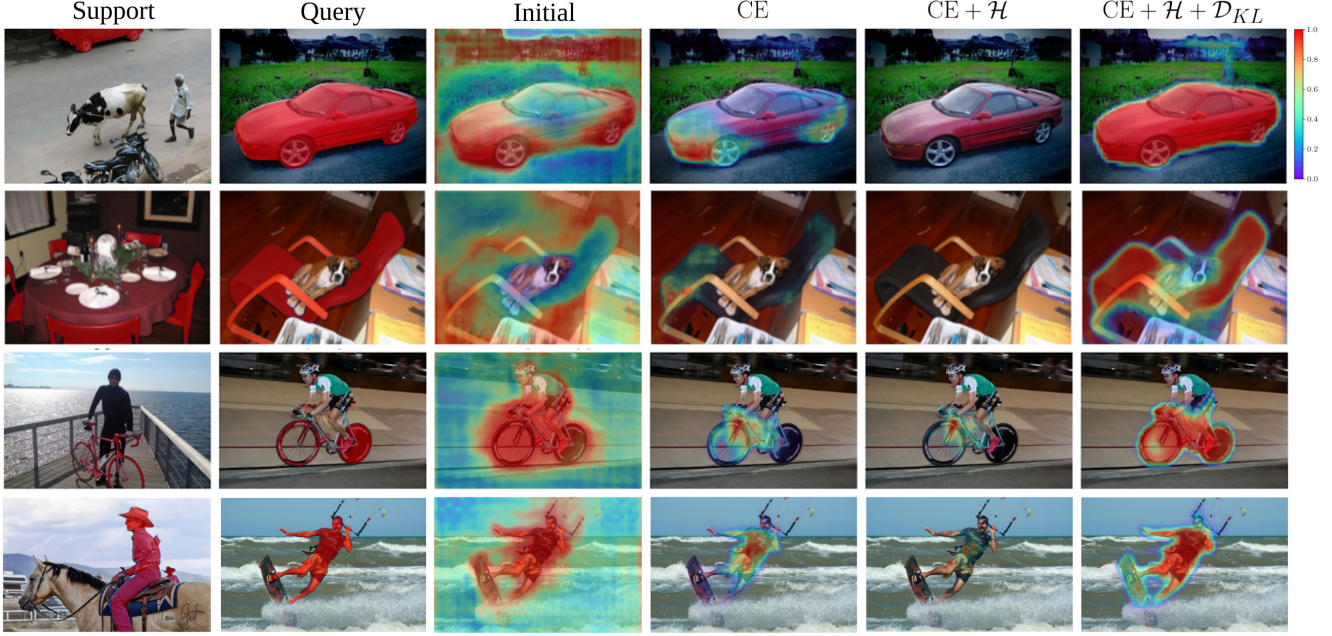


Figure 1: Probability maps for several 1-shot tasks. For each task, the two first columns show the ground truth of support and query. *Initial* column represents the probability map with the initial classifier  $\theta^{(0)}$ , and the last three columns show the final soft predicted segmentation after finetuning with each of the three losses. Best viewed in colors.

### 3. Formulation

#### 3.1. Few-shot Setting

Formally, we define a *base* dataset  $\mathcal{D}_{\text{base}}$  with base semantic classes  $\mathcal{Y}_{\text{base}}$ , employed for training. Specifically,  $\mathcal{D}_{\text{base}} = \{(x_n, y_n)\}_{n=1}^N$ ,  $\Omega \subset \mathbb{R}^2$  an image space,  $x_n : \Omega \rightarrow \mathbb{R}^3$  an input image, and  $y_n : \Omega \rightarrow \{0, 1\}^{|\mathcal{Y}_{\text{base}}|}$  its corresponding pixelwise one-hot annotation. At inference, we test our model through a series of  $K$ -shots tasks. Each  $K$ -shots task consists of a *support* set  $\mathcal{S} = \{(x_k, y_k)\}_{k=1}^K$ , i.e.  $K$  fully annotated images, and one unlabelled query image  $x_Q$ , all from the same novel class. This class is randomly sampled from a set of *novel* classes  $\mathcal{Y}_{\text{novel}}$  such that  $\mathcal{Y}_{\text{base}} \cap \mathcal{Y}_{\text{novel}} = \emptyset$ . The goal is to leverage the supervision provided by the support set in order to properly segment the object of interest in the query image.

#### 3.2. Base training

**Inductive bias in episodic training** There exist different ways of leveraging the base set  $\mathcal{D}_{\text{base}}$ . Meta-learning, or *learning to learn*, is the dominant paradigm in the few-shot literature. It emulates the test-time scenario during training by structuring  $\mathcal{D}_{\text{base}}$  into a series of training tasks. Then, the model is trained on these tasks to learn how to best leverage the supervision from the support set in order to enhance its query segmentation. Recently, Cao et al. [3] formally proved that the number of shots  $K_{\text{train}}$  used in train-

ing episodes in the case of prototypical networks represents a learning bias, and that the testing performance saturates quickly when  $K_{\text{test}}$  differs from  $K_{\text{train}}$ . Empirically, we observed the same trend for current few-shot segmentation methods, with minor improvements from 1-shot to 5-shot performances (Table 1).

**Standard training** In practice, the format of the test tasks may be unknown beforehand. Therefore, we want to take as little assumptions as possible on this. This motivates us to employ a feature extractor  $f_\phi$  trained with standard cross-entropy supervision on the whole  $\mathcal{D}_{\text{base}}$  set instead, **without resorting to episodic training**.

#### 3.3. Inference

**Objective** In what follows, we use  $\cdot$  as a placeholder to denote either a support subscript  $k \in \{1, \dots, K\}$  or the query subscript  $Q$ . At inference, we consider the 1-way segmentation problem:  $y_\cdot : \Omega \rightarrow \{0, 1\}^2$  is the function representing the dense *background/foreground* (B/F) mask in image  $x_\cdot$ . For both support and query images, we extract features  $z_\cdot := f_\phi(x_\cdot)$  and  $z_\cdot : \Psi \rightarrow \mathbb{R}^C$ , where  $C$  is the channel dimension in the feature space  $\Psi$ , with lower pixel resolution  $|\Psi| < |\Omega|$ .

Using features  $z_\cdot$ , our goal is to learn the parameters  $\theta$  of a classifier that properly discriminates foreground from background pixels. Precisely, our classifier  $p_\cdot : \Psi \rightarrow [0, 1]^2$  assigns a (B/F) probability vector to each pixel  $j \in \Psi$  in the



extracted feature space.

For each test task, we find the parameters  $\theta$  of the classifier by optimizing the following transductive objective:

$$\min_{\theta} \text{CE} + \lambda_{\mathcal{H}} \mathcal{H} + \lambda_{\text{KL}} \mathcal{D}_{\text{KL}}, \quad (1)$$

where  $\lambda_{\mathcal{H}}, \lambda_{\text{KL}} \in \mathbb{R}$  are non-negative hyper-parameters balancing the effects of the different terms.

We now describe in details each of the terms in Eq. (1):

$$\text{CE} = -\frac{1}{K|\Psi|} \sum_{k=1}^K \sum_{j \in \Psi} \tilde{y}_k(j)^\top \log(p_k(j))$$

is the cross-entropy between the downsampled labels  $\tilde{y}_k$  from support images and our classifier's soft predictions. Simply minimizing this term will often lead to degenerate solutions, especially in the 1-shot setting, as observed in Figure 1—the classifier  $\theta$  typically overfits the support set  $\mathcal{S}$ , translating into small activated regions on the query image.

$$\mathcal{H} = -\frac{1}{|\Psi|} \sum_{j \in \Psi} p_{\mathcal{Q}}(j)^\top \log(p_{\mathcal{Q}}(j))$$

is the Shannon entropy of the predictions on the query-image pixels. The role of this entropy term is to make the model's predictions more confident on the query image. The use of  $\mathcal{H}$  originates from the semi-supervised literature [11, 22, 1]. Intuitively, it pushes the decision boundary drawn by the linear classifier towards low-density regions of the extracted query feature space. While this term plays a crucial role in conserving object regions that were initially predicted with only medium confidence, its sole addition to CE does not solve the problem of degenerate solutions, and may even worsen it in some cases.

$$\mathcal{D}_{\text{KL}} = \hat{p}_{\mathcal{Q}}^\top \log \left( \frac{\hat{p}_{\mathcal{Q}}}{\pi} \right),$$

with  $\hat{p}_{\mathcal{Q}} = \frac{1}{|\Psi|} \sum_{j \in \Psi} p_{\mathcal{Q}}(j)$ , is a Kullback-Leibler (KL) Divergence term that encourages the B/F proportion predicted by the model to match a parameter  $\pi \in [0, 1]^2$ . Notice that the division inside the log applies element-wise. The joint estimation of parameter  $\pi$  in our context is further discussed in a following paragraph. Here, we argue that this term plays a **key role** in our loss. First, in the case where parameter  $\pi$  does not match the exact B/F proportion of the query image, this term still helps avoiding the degenerate solutions stemming from CE and  $\mathcal{H}$  minimization. Second, should an accurate estimate of the B/F proportion in the query image be available, it could easily be embedded through this term, resulting in a substantial performance boost, as discussed in Section 4.

**Choice of the classifier** As we optimize  $\theta$  for each task at inference, we want our method to add as little computational load as possible. In this regard, we employ a simple linear classifier with learnable parameters  $\theta^{(t)} = \{w^{(t)}, b^{(t)}\}$ , with  $t$  the current step of the optimization procedure,  $w^{(t)} \in \mathbb{R}^C$  the *foreground* prototype and  $b^{(t)} \in \mathbb{R}$  the corresponding bias. Thus, the probabilities  $p_k^{(t)}$  and  $p_{\mathcal{Q}}^{(t)}$  at iteration  $t$ , for pixel  $j \in \Psi$  can be obtained as follow:

$$p_{\cdot}^{(t)}(j) := \begin{pmatrix} 1 - s_{\cdot}^{(t)}(j) \\ s_{\cdot}^{(t)}(j) \end{pmatrix}, \quad (2)$$

where  $s_{\cdot}^{(t)}(j) = \text{sigmoid}(\tau [\cos(z_{\cdot}(j), w^{(t)}) - b^{(t)}])$ ,  $\tau \in \mathbb{R}$  is a temperature hyper-parameter and  $\cos$  the cosine similarity. The same classifier is used to estimate the support set probabilities  $p_k$  and the query predicted probabilities  $p_{\mathcal{Q}}$ . At initialization, we set prototype  $w^{(0)}$  to be the average of the foreground support features, i.e.  $w^{(0)} = \frac{1}{K|\Psi|} \sum_{k=1}^K \sum_{j \in \Psi} \tilde{y}_k(j)_1 z_k(j)$ , with  $\tilde{y}_k(j)_1$  the foreground component of the one-hot label of image  $x_k$  at pixel  $j$ . Initial bias  $b^{(0)}$  is set as the mean of the foreground's soft predictions on the query image:  $b^{(0)} = \frac{1}{|\Psi|} \sum_{j \in \Psi} p_{\mathcal{Q}}(j)_1$ . Then,  $w^{(t)}$  and  $b^{(t)}$  are optimized with gradient descent. The computational footprint of this per-task optimization is discussed in Section 4.

**Joint estimation of B/F proportion  $\pi$**  Without additional information, we leverage the model's label-marginal distribution over the query image  $\hat{p}_{\mathcal{Q}}^{(t)}$  in order to learn  $\pi$  jointly with classifier parameters. Note that minimizing Eq. (1) with respect to  $\pi$  yields  $\pi^{(t)} = \hat{p}_{\mathcal{Q}}^{(t)}$ . Empirically, we found that, after initialization, updating  $\pi$  only once during optimization, at a later iteration,  $t_{\pi}$  was enough:

$$\pi^{(t)} = \begin{cases} \hat{p}_{\mathcal{Q}}^{(0)} & 0 \leq t \leq t_{\pi} \\ \hat{p}_{\mathcal{Q}}^{(t_{\pi})} & t > t_{\pi}. \end{cases} \quad (3)$$

Intuitively, the entropy term  $\mathcal{H}$  helps gradually refine initially blurry soft predictions (third column in Fig. 1), which turns  $\hat{p}_{\mathcal{Q}}^{(t)}$  into an improving estimate of the true B/F proportion. A quantitative study of this phenomenon is provided in Section 4.3. Therefore, our inference can be seen as a joint optimization over  $\theta$  and  $\pi$ , with  $\mathcal{D}_{\text{KL}}$  serving as a *self-regularization* that prevents the model's marginal distribution  $\hat{p}_{\mathcal{Q}}^{(t)}$  from diverging.

**Oracle case with a known  $\pi$**  As an upper bound, we also investigate the *oracle* case, where we have access to the true B/F proportion in  $x_{\mathcal{Q}}$ :

$$\pi^* = \frac{1}{|\Psi|} \sum_{j \in \Psi} \tilde{y}_{\mathcal{Q}}(j). \quad (4)$$

## 4. Experiments

**Datasets** We resort to two public few-shot segmentation benchmarks, PASCAL-5<sup>i</sup> and COCO-20<sup>i</sup>, to evaluate our method. PASCAL-5<sup>i</sup> is built from PASCALVOC 2012 [8], and contains 20 object categories split into 4 folds. For each fold, 15 classes are used for training and the remaining 5 categories for testing. COCO-20<sup>i</sup> is built from MSCOCO [18] and is more challenging, as it contains more samples, more classes and more instances per image. Similar to PASCAL-5<sup>i</sup>, COCO-20<sup>i</sup> dataset is divided into 4-folds with 60 base classes and 20 test classes in each fold.

**Training** We build our model based on PSPNet [43] with Resnet-50 and Resnet-101 [13] as backbones. We train the feature extractor with standard cross-entropy over the base classes during 100 epochs on PASCAL-5<sup>i</sup>, and 20 epochs on COCO-20<sup>i</sup>, with batch size set to 12. We use SGD as optimizer with the initial learning rate set to  $2.5e-3$  and we use cosine decay. Momentum is set to 0.9, and weight decay to  $1e-4$ . Label smoothing is used with smoothing parameter  $\epsilon = 0.1$ . We did not use multi-scaling, nor deep supervision, unlike the original PSPNet paper [43]. As for data augmentations, we only use random mirror flipping.

**Inference** At inference, following previous works [21, 37], all images are resized to a fixed  $417 \times 417$  resolution. For each task, the classifier  $\theta$  is built on top of the features from the penultimate layer of the trained network. For our model with ResNet-50 as backbone, this results in a  $53 \times 53 \times 512$  feature map. SGD optimizer is used to train  $\theta$ , with a learning rate of 0.025. For each task, a total of 50 iterations are performed. The parameter  $t_\pi$  is set to 10. For the main method, the weights  $\lambda_{\mathcal{H}}$  and  $\lambda_{\text{KL}}$  are both initially set to  $1/K$ , such that the CE term plays a more important role as the number of shots  $K$  grows. For  $t \geq t_\pi$ ,  $\lambda_{\text{KL}}$  is increased by 1 to further encourage the predicted proportion close to  $\pi^{(t_\pi)}$ . Finally, the temperature  $\tau$  is set to 20.

**Evaluation** We employ the widely adopted mean Intersection over Union (mIoU). Specifically, for each class, the classwise-IoU is computed as the sum over all samples within the class of the intersection over the sum of all unions. Then, the mIoU is computed as the average over all classes of the classwise-IoU. Following previous works [21], 5 runs of 1000 tasks each are computed for each fold, and the average mIoU over runs is reported.

### 4.1. Benchmark results

**Main method** First, we investigate the performance of the proposed method in the popular 1-shot and 5-shot settings on both PASCAL-5<sup>i</sup> and COCO-20<sup>i</sup>, whose results are reported in Table 1 and 2. Overall, we found that our

method compares competitively with state-of-the-art approaches in the 1-shot setting, and significantly outperforms recent methods in the 5-shot scenario. Additional qualitative results on PASCAL-5<sup>i</sup> are shown in the supplemental material.

**Beyond 5-shots** In the popular learning-to-learn paradigm, the number of shots leveraged during the meta-training stage has a direct impact on the performance at inference [3]. Particularly, to achieve the best performance, meta-learning based methods typically require the numbers of shots used during meta-training to match those employed during meta-testing. To demonstrate that the proposed method is more robust against differences on the number of labeled support samples between the base and test sets, we further investigate the 10-shot scenario. Particularly, we trained the methods in [33, 38] by using one labeled sample per class, i.e., 1-shot task, and test the models on a 10-shots task. Interestingly, we show that the gap between our method and current state-of-the-art becomes larger as the number of support images increases (Table 3), with significant gains of 6% and 4% on PASCAL-5<sup>i</sup> and COCO-20<sup>i</sup>, respectively. These results suggest that our transductive inference leverages more effectively the information conveyed in the labeled support set of a given task.

**Oracle results** We now investigate the ideal scenario where an oracle provides the exact foreground/background proportion in the query image, such that  $\pi^{(t)} = \pi^*, \forall t$ . Reported results in this scenario, referred to as *Oracle* (Table 1 and 2) show impressive improvements over both our current method and all previous works, with a consistent gain across datasets and tasks. Particularly, these values range from 11% and 14 % on both PASCAL-5<sup>i</sup> and COCO-20<sup>i</sup> and in both 1-shot and 5-shot settings. We believe that these findings convey two important messages. First, it proves that there exists a simple linear classifier that can largely outperform state-of-the-art meta-learning models, while being built on top of a feature extractor trained with a standard cross-entropy loss. Second, these results indicate that having a precise size of the query object of interest acts as a strong regularizer. This suggests that more efforts could be directed towards properly constraining the optimization process of  $w$  and  $b$ , and opens a door to promising avenues.

### 4.2. Domain shift

We introduce a more realistic, cross-domain setting (COCO-20<sup>i</sup> to PASCAL-VOC). We argue that such setting is a step towards a more realistic evaluation of these methods, as it can assess the impact on performances caused by a domain shift between the data training distribution and the testing one. We believe that this scenario can be easily found in practice, as even slight alterations in the data collection process might result in a distributional shift. We

Table 1: Results of 1-way 1-shot and 1-way 5-shot segmentation on PASCAL-5<sup>i</sup> using the mean-IoU. Best results in bold.

Method	Backbone	1 shot					5 shot				
		Fold-0	Fold-1	Fold-2	Fold-3	Mean	Fold-0	Fold-1	Fold-2	Fold-3	Mean
OSLSM [29] (BMVC'18)	VGG-16	33.6	55.3	40.9	33.5	40.8	35.9	58.1	42.7	39.1	43.9
co-FCN [25] (ICLRW'18)		36.7	50.6	44.9	32.4	41.1	37.5	50.0	44.1	33.9	41.4
AMP [30] (ICCV'19)		41.9	50.2	46.7	34.7	43.4	41.8	55.5	50.3	39.9	46.9
PANet [37] (ICCV'19)		42.3	58.0	51.1	41.2	48.1	51.8	64.6	59.8	46.5	55.7
FWB [23] (ICCV'19)		47.0	59.6	52.6	48.3	51.9	50.9	62.9	56.5	50.1	55.1
SG-One [42] (TCYB'20)		40.2	58.4	48.4	38.4	46.3	41.9	58.6	48.6	39.4	47.1
CRNet [19] (CVPR'20)		-	-	-	-	55.2	-	-	-	-	58.5
FSS-1000 [17] (CVPR'20)		-	-	-	-	-	37.4	60.9	46.6	42.2	56.8
RPM [21] (ECCV'20)		47.1	65.8	50.6	48.5	53.0	50.0	66.5	51.9	47.6	54.0
CANet [41] (CVPR'19)	ResNet50	52.5	65.9	51.3	51.9	55.4	55.5	67.8	51.9	53.2	57.1
PGNet [40] (ICCV'19)		56.0	66.9	50.6	50.4	56.0	57.7	68.7	52.9	54.6	58.5
CRNet [19] (CVPR'20)		-	-	-	-	55.7	-	-	-	-	58.8
SimPropNet [10] (IJCAI'20)		54.9	67.3	54.5	52.0	57.2	57.2	68.5	58.4	56.1	60.0
LTM [39] (MMMM'20)		52.8	69.6	53.2	52.3	57.0	57.9	69.9	56.9	57.5	60.6
RPM [38] (ECCV'20)		55.2	66.9	52.6	50.7	56.3	56.3	67.3	54.5	51.0	57.3
PPNet [21] (ECCV'20)*		47.8	58.8	53.8	45.6	51.5	58.4	67.8	64.9	56.7	62.0
PFENet [33] (TPAMI'20)		<b>61.7</b>	<b>69.5</b>	55.4	<b>56.3</b>	<b>60.8</b>	63.1	70.7	55.8	57.9	61.9
RePRI (ours)		60.2	67.0	<b>61.7</b>	47.5	59.1	<b>64.5</b>	<b>70.8</b>	<b>71.7</b>	<b>60.3</b>	<b>66.8</b>
Oracle-RePRI	ResNet50	72.4	78.0	77.1	65.8	73.3	75.1	80.8	81.4	74.4	77.9
FWB [23] (ICCV'19)	ResNet101	51.3	64.5	56.7	52.2	56.2	54.9	67.4	62.2	55.3	59.9
DAN [36] (ECCV'20)		54.7	68.6	57.8	51.6	58.2	57.9	69.0	60.1	54.9	60.5
PFENet [33] (TPAMI'20)		<b>60.5</b>	<b>69.4</b>	54.4	<b>55.9</b>	<b>60.1</b>	62.8	70.4	54.9	57.6	61.4
RePRI (ours)		59.6	68.6	<b>62.2</b>	47.2	59.4	<b>66.2</b>	<b>71.4</b>	<b>67.0</b>	<b>57.7</b>	<b>65.6</b>
Oracle-RePRI	ResNet101	73.9	79.7	76.1	65.1	73.7	76.8	81.7	79.5	74.5	78.1

\* We report the results where no additional unlabeled data is employed.

Table 2: Results of 1-way 1-shot and 1-way 5-shot segmentation on COCO-20<sup>i</sup> using mean-IoU metric. Best results in bold.

Method	Backbone	1 shot					5 shot				
		Fold-0	Fold-1	Fold-2	Fold-3	Mean	Fold-0	Fold-1	Fold-2	Fold-3	Mean
PPNet* [21] (ECCV'20)	ResNet50	34.5	25.4	24.3	18.6	25.7	48.3	30.9	35.7	30.2	36.2
RPM [38] (ECCV'20)		29.5	36.8	29.0	27.0	30.6	33.8	42.0	33.0	33.3	35.5
PFENet [33] (TPAMI'20)		<b>36.5</b>	<b>38.6</b>	<b>34.5</b>	<b>33.8</b>	<b>35.8</b>	36.5	43.3	37.8	38.4	39.0
RePRI (ours)		31.2	38.1	33.3	33.0	34.0	<b>38.5</b>	<b>46.2</b>	<b>40.0</b>	<b>43.6</b>	<b>42.1</b>
Oracle-RePRI	ResNet50	49.3	51.4	38.2	41.6	45.1	51.5	60.8	54.7	55.2	55.5

\* We report the results where no additional unlabeled data is employed.

Table 3: Aggregated results for 1-way 1-, 5- and 10-shot tasks with Resnet50 as backbone and averaged over 4 folds. Per fold results are available in the supplementary material.

Method	PASCAL-5 <sup>i</sup>			COCO-20 <sup>i</sup>		
	1-S	5-S	10-S	1-S	5-S	10-S
RPM [38]	56.3	57.3	57.6	30.6	35.5	33.1
PFENet [33]	<b>60.8</b>	61.9	62.1	<b>35.8</b>	39.0	39.7
RePRI (ours)	59.1	<b>66.8</b>	<b>68.2</b>	34.0	<b>42.1</b>	<b>44.4</b>
Oracle-RePRI	73.3	77.9	78.6	45.1	55.5	58.7

reproduce the scenario where a large labeled dataset is available (e.g., COCO-20<sup>i</sup>), but the evaluation is performed on a target dataset with a different feature distribution (e.g., PASCAL-VOC). As per the original work [18], significant differences exist between the two original datasets. For instance, images in MS-COCO have on average 7.7 instances of objects coming from 3.5 distinct categories, while PASCAL-VOC only has an average of 3 instances from 2 distinct categories.

**Evaluation** We reuse models trained on each fold of COCO-20<sup>i</sup> and generate tasks using images from all the classes in PASCAL-VOC that were not used during train-

Table 4: Aggregated domain-shift results, averaged over 4 folds, on COCO-20<sup>i</sup> to PASCAL-VOC. Best results in bold. Per-fold results are available in the supplementary material.

Method	Backbone	COCO → PASCAL	
		1 shot	5 shot
RPM [38]		49.6	53.8
PFENet [33]	ResNet50	61.1	63.4
RePRI (ours)		<b>63.2</b>	<b>67.7</b>
Oracle-RePRI	Resnet-50	76.2	79.7

ing. For instance, fold-0 of this setting means the model was trained on fold-0 of COCO-20<sup>i</sup> and tested on the whole PASCAL-VOC dataset, after removing the classes seen in training. A complete summary of all the folds is available in the Supplemental material.

**Results** We reproduced and compared to the two best performing methods [33, 21] using their respective official GitHub repositories. Table 4 summarizes the results for the 1-shot and 5-shot cross-domain experiments. We observe that in the presence of domain-shift, our method outperforms existing methods in both 1-shot and 5-shot scenarios, with again the improvement jumping from 2% in 1-shot to 4% in 5-shot.

### 4.3. Ablation studies

**Impact of each term in the main objective** While Fig. 1 provides *qualitative* insights on how each term in Eq. (1) affects the final prediction, this section provides a *quantitative* evaluation of their impact, evaluated on PASCAL-5<sup>i</sup> (Table 5). Quantitative results confirm the qualitative insights observed in Fig. 1, as both CE and CE +  $\mathcal{H}$  losses drastically degrade the performance compared to the proportion-regularized loss, i.e., CE +  $\mathcal{H}$  +  $\mathcal{D}_{KL}$ . For example, in the 1-shot scenario, simply minimizing the CE results in more than 20% of difference compared to the proposed model. In this case, the prototype  $w$  tends to overfit the support sample and only activates regions of the query object that strongly correlate with the support object. Such behavior hampers the performance when the support and query objects exhibit slight changes in shape or histogram colors, for example, which may be very common in practice. Adding the entropy term  $\mathcal{H}$  to CE partially alleviates this problem, as it tends to reinforce the model in being confident on positive pixels initially classified with mid or low confidence. Nevertheless, despite improving the naive CE based model, the gap with the proposed model remains considerably large, with 10% difference. One may notice that the differences between CE and CE +  $\mathcal{H}$  +  $\mathcal{D}_{KL}$  decrease in the 5-shot setting, since overfitting 5 support samples si-

multaneously becomes more difficult. The results from this ablation experiment reinforce our initial hypothesis that the proposed KL term based on the size parameter  $\pi$  acts as a strong regularizer.

**Influence of the parameter  $t_\pi$**  In Fig. 2, we plot the averaged mIoU (over 4 folds) as a function of  $t_\pi$  varying over the full range  $t_\pi \in [1, 50]$ . For 5-shot, the performances are stable and remain largely above SOTA for all  $t_\pi$ . As for the 1-shot case, the range  $[5, 15]$  yields roughly similar results. While selection of optimal  $t_\pi$  would lead to performance gains in each setting, in the paper, we used a single value of  $t_\pi = 10$  for all the settings.

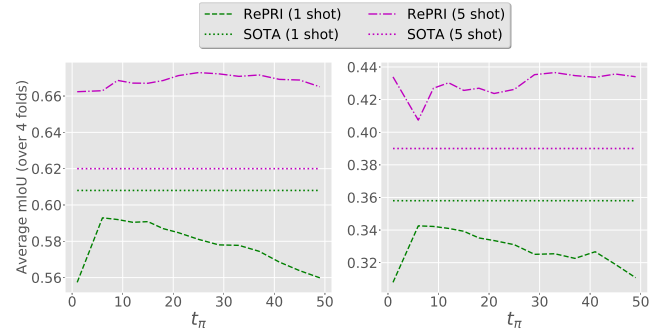


Figure 2: Average mIoU (over 4 folds) as a function of  $t_\pi$  on PASCAL-5<sup>i</sup> (left) and COCO-20<sup>i</sup> (right).

**Influence of parameter  $\pi$  misestimation** Precisely knowing the foreground/background (B/F) proportion of the query object is unrealistic. To quantify the deviation from the exact B/F proportion  $\pi^*$ , we introduce the relative error on the foreground size:

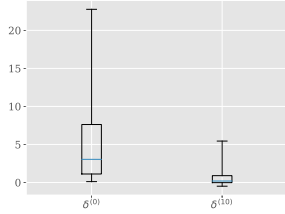
$$\delta^{(t)} = \frac{\pi_1^{(t)}}{\pi_1^*} - 1, \quad (5)$$

where  $\pi_1^*$  represents the exact foreground proportion in the query image, extracted from its corresponding ground truth, and  $\pi_1^{(t)}$  our estimate at iteration  $t$ , which is derived from the soft predicted segmentation. As observed from Fig. 1, the initial prototype often results in a blurred probability map, from which only a very coarse estimate of the query proportion can be inferred and used as  $\pi^{(0)}$ . The distribution of  $\delta$  over 5000 tasks is presented in Fig. 3a. It clearly shows that the initial prediction typically provides an overestimate of the actual query foreground size, while finetuning the classifier  $\theta$  for 10 iterations with our main loss (Eq. 1) already provides a strictly more accurate estimate, as conveyed by the right box plot in Fig. 3a, with an average  $\delta$  around 0.7.

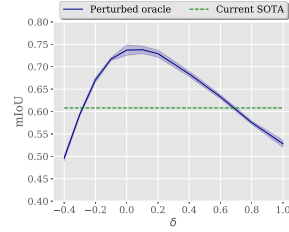
Now, a natural question remains: **how good does the estimate need to be in order to approach the oracle results?** To answer this, we carry out a series of controlled

Table 5: Ablation study on the effect of each term in our loss in Eq. (1), evaluated on PASCAL-5<sup>i</sup>.

Loss	1 shot					5 shot				
	Fold-0	Fold-1	Fold-2	Fold-3	Mean	Fold-0	Fold-1	Fold-2	Fold-3	Mean
CE	39.7	49.3	37.3	27.5	38.5	56.5	66.4	60.1	49.0	58.0
CE + $\mathcal{H}$	45.7	61.7	48.2	36.4	48.0	56.8	68.5	61.3	47.0	58.4
CE + $\mathcal{H}$ + $\mathcal{D}_{\text{KL}}$	<b>60.2</b>	<b>67.0</b>	<b>61.7</b>	<b>47.5</b>	<b>59.1</b>	<b>64.5</b>	<b>70.8</b>	<b>71.7</b>	<b>60.3</b>	<b>66.8</b>



(a) Relative error  $\delta$  distribution of our current method, at initialization  $\delta^{(0)}$  and after 10 gradient iterations  $\delta^{(10)}$ .



(b) Mean-IoU versus enforced relative foreground size error  $\delta$  in the parameter  $\pi^{(0)}$ .

Figure 3: Experiments on  $\pi$  misestimation. Both figures are computed using 5 runs of 1000 1-shot tasks, each on the fold-0 of PASCAL5<sup>i</sup>.

Table 6: Number of tasks performed per second, and the corresponding mIoU performances on PASCAL-5<sup>i</sup>.

Method	1-shot		5-shot	
	FPS	mIoU	FPS	mIoU
RPMMS [38]	18.2	51.5	9.4	57.3
PFENet [33]	15.9	60.8	5.1	61.9
RePRI (ours)	12.8	59.1	4.4	66.8

experiments where, instead of computing  $\pi^{(t)}$  with Eq. (3), we use a  $\delta$ -perturbed oracle at initialization, such that  $\pi_1^{(t)} = \pi_1^*(1 + \delta)$ . Each point in Fig. 3b represents the mIoU obtained over 5000 tasks for a given perturbation  $\delta$ . Fig. 3b reveals that exact B/F proportion is not required to significantly close the gap with the oracle. Specifically, foreground size estimates ranging from -10% to +30% with respect to the oracle proportion are sufficient to achieve 70%+ of mIoU, which represents an improvement of 10% over the current state-of-the-art. This suggests that more refined size estimation methods may significantly increase the performance of the proposed method.

**Computational efficiency** We now inspect the computational cost of the proposed model, and compare to recent existing methods. Unlike prior work, we solve an optimization

problem at inference, which naturally slows down the inference process. However, in our case, only a single prototype vector  $w \in \mathbb{R}^C$ , where we recall  $C$  is the feature channel dimension, and a bias  $b \in \mathbb{R}$  need to be optimized for each task. Furthermore, in our setting  $C = 512$ , and therefore the problem can still be solved relatively efficiently, leading to reasonable inference times. In Table 6, we summarize the FPS rate at inference for our method, as well as for two competing approaches that only require a forward pass. We can observe that, unsurprisingly, our method reports lower FPS rates, without becoming unacceptably slower. The reported values indicate that the differences in inference times are small compared to, for example, the approach in [33]. Particularly, in the 1-shot scenario, our method processes tasks 3 FPS slower than [33], whereas this gap narrows down to 0.7 FPS in the 5-shot setting.

## 5. Conclusion

Without resorting to the popular meta-learning paradigm, our proposed RePRI achieves new state-of-the-art results on standard 5-shot segmentation benchmarks, while being close to best performing approaches in the 1-shot setting. RePRI is modular and can, therefore, be used in conjunction with any feature extractor regardless how the base training was performed. Supported by the findings in this work, we believe that the relevance of the episodic training should be re-considered in the context of few-shot segmentation, and we provide a strong baseline to stimulate future research on this topic. Our results indicate that current state-of-the-art methods may have difficulty with more challenging settings, when dealing with domain shift or conducting inference on tasks whose structures are different from those seen in training—scenarios that have been overlooked in the literature. These findings align with recent observations in few shot classification [4, 3]. Furthermore, embedding more accurate foreground-background proportion estimates appears to be a very promising way of constraining the inference, as demonstrated with the significantly improved results obtained by the oracle. Our implementation is publicly available online: <https://github.com/mboudiaf/RePRI-for-Few-Shot-Segmentation>.



## References

- [1] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 4
- [2] Malik Boudiaf, Ziko Imtiaz Masud, Jérôme Rony, José Dolz, Pablo Piantanida, and Ismail Ben Ayed. Transductive information maximization for few-shot learning. In *Neural Information Processing Systems (NeurIPS)*, 2020. 1, 2
- [3] Tianshi Cao, Marc Law, and Sanja Fidler. A theoretical analysis of the number of shots in few-shot learning. In *International Conference on Learning Representations (ICLR)*, 2020. 1, 2, 3, 5, 8
- [4] 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. 1, 2, 8
- [5] Zhou Dengyong, Olivier Bousquet, Thomas N Lal, Jason Weston, and Bernhard Schölkopf. Learning with local and global consistency. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2004. 2
- [6] Guneet Singh Dhillon, Pratik Chaudhari, Avinash Ravichandran, and Stefano Soatto. A baseline for few-shot image classification. In *International Conference on Learning Representations (ICLR)*, 2019. 1, 2
- [7] Nanqing Dong and Eric Xing. Few-shot semantic segmentation with prototype learning. In *British Machine Vision Conference (BMVC)*, 2018. 2
- [8] 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)*, 2010. 5
- [9] 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. 2
- [10] Siddhartha Gairola, Mayur Hemani, Ayush Chopra, and Balaji Krishnamurthy. SimPropNet: Improved similarity propagation for few-shot image segmentation. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2020. 1, 6
- [11] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. In *Advances in neural information processing systems (NeurIPS)*, 2005. 4
- [12] Yunhui Guo, Noel Codella, Leonid Karlinsky, James V. Codella, John R. Smith, Kate Saenko, Tazjana Rosing, and Rogerio Feris. A broader study of cross-domain few-shot learning. In *European Conference on Computer Vision (ECCV)*, 2020. 1
- [13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 770–778, 2016. 5
- [14] Ruibing Hou, Hong Chang, MA Bingpeng, Shiguang Shan, and Xilin Chen. Cross attention network for few-shot classification. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 2
- [15] Thorsten Joachims. Transductive inference for text classification using support vector machines. In *International Conference on Machine Learning (ICML)*, 1999. 2
- [16] Jongmin Kim, Taesup Kim, Sungwoong Kim, and Chang D Yoo. Edge-labeling graph neural network for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2
- [17] Xiang Li, Tianhan Wei, Yau Pun Chen, Yu-Wing Tai, and Chi-Keung Tang. FSS-1000: A 1000-class dataset for few-shot segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 6
- [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 (ECCV)*, 2014. 5, 6
- [19] Weide Liu, Chi Zhang, Guosheng Lin, and Fayao Liu. CRNet: Cross-reference networks for few-shot segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 2, 6
- [20] Yanbin Liu, Juho Lee, Minseop Park, Saehoon Kim, Eunho Yang, Sung Ju Hwang, and Yi Yang. Learning to propagate labels: Transductive propagation network for few-shot learning. In *International Conference on Learning Representations (ICLR)*, 2019. 2
- [21] Yongfei Liu, Xiangyi Zhang, Songyang Zhang, and Xuming He. Part-aware prototype network for few-shot semantic segmentation. In *European Conference on Computer Vision (ECCV)*, 2020. 2, 5, 6, 7
- [22] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. In *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2018. 4
- [23] Khoi Nguyen and Sinisa Todorovic. Feature weighting and boosting for few-shot segmentation. In *In-*

- ternational Conference on Computer Vision (ICCV), 2019. 1, 2, 6
- [24] Limeng Qiao, Yemin Shi, Jia Li, Yaowei Wang, Tiejun Huang, and Yonghong Tian. Transductive episodic-wise adaptive metric for few-shot learning. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019. 2
- [25] Kate Rakelly, Evan Shelhamer, Trevor Darrell, Alyosha Efros, and Sergey Levine. Conditional networks for few-shot semantic segmentation. In *International Conference on Learning Representations (ICLR) Workshop*, 2018. 1, 2, 6
- [26] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *International Conference on Learning Representations (ICLR)*, 2017. 2
- [27] Mengye Ren, Eleni Triantafillou, Sachin Ravi, Jake Snell, Kevin Swersky, Joshua B Tenenbaum, Hugo Larochelle, and Richard S Zemel. Meta-learning for semi-supervised few-shot classification. In *International Conference on Learning Representations (ICLR)*, 2018. 1, 2
- [28] Jürgen Schmidhuber. *Evolutionary principles in self-referential learning, or on learning how to learn: the meta-meta-... hook*. PhD thesis, Technische Universität München, 1987. 2
- [29] Amirreza Shaban, Shray Bansal, Zhen Liu, Irfan Essa, and Byron Boots. One-shot learning for semantic segmentation. In *British Machine Vision Conference (BMVC)*, 2018. 2, 6
- [30] Mennatullah Siam, Boris N Oreshkin, and Martin Jagersand. AMP: Adaptive masked proxies for few-shot segmentation. In *International Conference on Computer Vision (ICCV)*, 2019. 2, 6
- [31] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in neural information processing systems (NeurIPS)*, 2017. 1, 2
- [32] Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B Tenenbaum, and Phillip Isola. Rethinking few-shot image classification: a good embedding is all you need? *arXiv preprint arXiv:2003.11539*, 2020. 1
- [33] Zhuotao Tian, Hengshuang Zhao, Michelle Shu, Zhicheng Yang, Ruiyu Li, and Jiaya Jia. Prior guided feature enrichment network for few-shot segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2020. 1, 5, 6, 7, 8
- [34] Vladimir N Vapnik. An overview of statistical learning theory. In *IEEE Transactions on Neural Networks (TNN)*, 1999. 2
- [35] 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. 1, 2
- [36] Haochen Wang, Xudong Zhang, Yutao Hu, Yandan Yang, Xianbin Cao, and Xiantong Zhen. Few-shot semantic segmentation with democratic attention networks. In *ECCV*, 2020. 1, 2, 6
- [37] Kaixin Wang, Jun Hao Liew, Yingtian Zou, Daquan Zhou, and Jiashi Feng. PANet: Few-shot image semantic segmentation with prototype alignment. In *International Conference on Computer Vision (ICCV)*, 2019. 1, 2, 5, 6
- [38] Boyu Yang, Chang Liu, Bohao Li, Jianbin Jiao, and Qixiang Ye. Prototype mixture models for few-shot semantic segmentation. In *European Conference on Computer Vision (ECCV)*, 2020. 1, 2, 5, 6, 7, 8
- [39] Yuwei Yang, Fanman Meng, Hongliang Li, Qingbo Wu, Xiaolong Xu, and Shuai Chen. A new local transformation module for few-shot segmentation. In *International Conference on Multimedia Modeling (ICMM)*. Springer. 1, 6
- [40] Chi Zhang, Guosheng Lin, Fayao Liu, Jiushuang Guo, Qingyao Wu, and Rui Yao. Pyramid graph networks with connection attentions for region-based one-shot semantic segmentation. In *International Conference on Computer Vision (ICCV)*, 2019. 2, 6
- [41] 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 *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 2, 6
- [42] Xiaolin Zhang, Yunchao Wei, Yi Yang, and Thomas S Huang. SG-one: Similarity guidance network for one-shot semantic segmentation. *IEEE Transactions on Cybernetics*, 2020. 1, 2, 6
- [43] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2017. 5
- [44] Imtiaz Masud Ziko, Jose Dolz, Eric Granger, and Ismail Ben Ayed. Laplacian regularized few-shot learning. In *International Conference on Machine Learning (ICML)*, 2020. 1, 2