Supplementary Material

In this supplementary material, we first present extensive details on the datasets used in our experiments. Then, we show additional ablation studies and results that support the satisfactory performance of the proposed method.

1. Extended details on the employed datasets

PASCAL-5^i. PASCAL-5^i [26] is the most popular few-shot segmentation benchmark, which inherits from the well-known PASCAL dataset [7]. The images in PASCAL-5^i are split into 4 folds, each having 5 classes, with 3 folds used for training and 1 for evaluation. Following the standard procedure in [26, 21], we employ 1000 support-query pairs randomly sampled in the test split for each class at test time. More details on PASCAL-5^i are provided in [26].

FSS-1000. A limitation of PASCAL-5^i is that it contains relatively few distinct tasks, i.e., 20 excluding background and unknown categories. FSS-1000 dataset [16] alleviates this issue by introducing a more realistic dataset for few-shot semantic segmentation, which emphasizes the number of object classes rather than the number of images. Indeed, FSS-1000 contains a total of 1000 classes, where only 10 images and their corresponding ground truth for each category are provided. Out of the 1000 classes, 240 are dedicated to the test task and the remaining for training. The FSS-1000 dataset [16] only provides pixel-level annotations. Thus, to investigate the effect of using weak annotations in this dataset we generated bounding box annotations. Each bounding box is obtained from one randomly chosen instance mask in each support image. The generated bounding box annotations are provided with the code employed in the experiments.

COCO is a challenging large-scale dataset, which contains 80 object categories. Following [26], we choose 40 classes for training, 20 classes for validation and 20 classes for test.

2. Importance of the pyramidal setting

The integration of DoG in our model is strongly inspired by the seminal work in [20]. Thus, we followed the recommended setting, which suggests that 5 scale levels gives optimal results. To understand why employing a single DoG with a larger difference of σ between the gaussian kernels will not perform at the same level than a pyramid of progressive DoG we need to consider how we recognize images at different distances. When we try to recognize objects that are far away, we might be able to just identify rough details, while fine-grained object details become more clear as the image gets closer. Thus, the level of the scale-space is a key factor when trying to recognize discriminative features in an image. The problem, however, is that the optimal scale-space level to discriminate important features for each object is unknown. By blurring the image with different σ values each image represents a different scale-space level, each of them specializing on features at a given 'distance'. In contrast, if we assume a single DoG with a larger difference between the Gaussian kernel variance, intermediate scale-scape levels will be missed. To demonstrate this empirically, we investigated the setting where a single DoG with σ_0 and σ_4 is integrated into the CNN. Results reported in Table 1 shows that a single DoG obtains a mIoU value of 77.67 on the FSS-1000 dataset, underperforming by 3% the pyramidal setting.

Table 1. Effect of employing a single DoG with σ_0 and σ_4 vs. a pyramidal DoG with progressive σ values. Results for 1-shot on the FSS-1000 class dataset.

	mIoU
Single DoG	77.7
Pyramidal DoG	80.8

3. Ablation study on multi-scale fusion features.

Similarly to [40], we investigated the effect of employing different levels of features, or a combination of those. Particularly, we investigated the three last blocks of VGG-16. In our case, *block5* gives the best performance when a single block is used. If multiple blocks are used instead, we observed that combining the three blocks provides the best performance, even though the contribution of the *block4* is marginal compared to the fused features from *block3* and *block5* (+0.26%). The low performance of shallower layers alone can be explained by the fact that they exploit lower-level cues, which are insufficient to properly find object regions. By integrating these with higher-level features, which correspond to object categories, our model can efficiently identify class-agnostic regions on new images. Furthermore, fusion of features at several levels of abstraction can help to handle larger scale object variations. Thus, the final multi-scale model employed in our experiments corresponds to the architecture combining the three last feature blocks.

Table 2. Effect of combining different level feature maps in the encoder network. Best result is highlighted in bold.

Block 3	Block 4	Block 5	mIoU
✓			76.3
	✓		78.3
		✓	79.5
✓	✓		78.1
✓		✓	80.6
	✓	✓	79.5
✓	✓	✓	**80.8**

4. Model complexity.

The functionality of the proposed method in the demand of computational resources is also investigated in this work. Table 3 shows the model complexity of several methods, as well as their segmentation results on Pascal5^i for 1-shot. In this table, we include the models that either report their number of parameters or provide reproducible code. We observe that the proposed method is ranked among the lightest methods, while typically achieving the best segmentation performance. Compared to similar methods, in terms of complexity (e.g., co-FCN [22], RPMM[18] or SG-One [42]), our model brings between 2 and 17% gain on improvement.

Table 3. Parameter complexity in different approaches and their performance (mIoU) on 1-shot segmentation on PASCAL-5^i. Methods are ordered based on number of learnable parameters.

Method	**1-shot mIoU**	**#params(M)**
OSLSM [26]⋄	40.8	276.7
Meta-Seg [3]⋄	48.6	268.5
AMP [27]⋄	43.4	34.7
co-FCN [22]⋄	41.1	34.2
Proposed$^{\diamond}$	58.0	22.7
RPMM [36]†	56.3	19.6
SG-One [42]⋄	46.3	19.0
CANet [40]†	55.4	19.0
PGNet [39]†	56.0	17.2
Proposed‡	58.7	16.3
PANet [34]⋄	48.1	14.7
PFNet [30]‡	60.1	10.8

*Employed architectures: ⋄, VGG, † ResNet50, ‡ ResNet101

5. Results on COCO

Table 4 reports the results for 1- and 5-shot segmentation on COCO dataset. As the backbone architecture plays an important role on the performance of the whole model, we split the results on methods relying on VGG-16 *(top)* and on ResNet *(top)*. From these results we can see that the proposed method achieves the best performance for 1-shot setting on the VGG-16 group, also outperforming a recent approach with ResNet, i.e., [21]. Regarding the results on 5-shot, our model obtains similar results, but slightly worst, to those obtained by several approaches with ResNet as backbone. This, together with results on FSS-1000 and Pascal5^i, supports our hypothesis that removing the texture bias can be more efficient in scenarios with very limited supervision (e.g., 1-shot), where our method consistently achieves the best results across three different datasets (under the exact same conditions, i.e., same architecture as backbone).

6. Additional visual results

We include additional qualitative results to assess the performance of our method. First, in Fig. 1, visual results on the FSS-1000 class dataset are shown. Similarly to the qualitative examples shown in the main paper, we can observe how our method satisfactorily handles target objects presenting high variability on shape or perspective. This is evident, for example, in the bat images, where our method is able to capture the whole context of a bat flying, while the support image just contained an image of three bats standing in a branch. Then, we also depict failure cases (Fig. 2), where our method does not achieve satisfactory segmentations, or not as good as expected. Typically, these failures come in the form of incomplete segmentations, with small regions of the object not properly identified. The next figure (Fig. 3) depicts the results when a bounding box is employed as supervisory signal in the support sample (depicted in purple). Despite the fact that the support mask is noisy, the results achieved by our method are close to the ground truth masks. This, in addition to the quantitative results reported in Table 4 (main paper), shows that the proposed method, once trained on a base dataset, is robust to noise on the support masks. Last, in Figure 4, we depict few samples from the FSS-1000 class dataset, with their corresponding ground truth and the generated bounding box annotation.

Table 4. Results of 1-way 1-shot and 5-shot segmentation on COCO-20^i data set employing the mean Intersection Over Union (mIoU) metric. Methods are divided according to the backbone used.

		1-shot					5-shot				
Method		$fold^1$	$fold^2$	$fold^3$	$fold^4$	Mean	$fold^1$	$fold^2$	$fold^3$	$fold^4$	Mean
Backbone (VGG-16)											
PANet[34]	ICCV'19	-	-	-	-	20.9	-	-	-	-	**29.7**
Proposed	-	20.2	17.8	21.6	26.8	**21.6**	22.6	22.0	24.2	31.7	25.1
Backbone (ResNet)											
FWB[21] ‡	ICCV'19	18.4	16.7	19.6	25.4	20.0	20.9	19.2	21.9	28.4	22.6
OANet [43] ‡	Arxiv'20	29.6	22.9	20.3	17.5	22.6	36.6	27.1	25.9	21.9	27.9
DAN [33] ‡	ECCV'20	-	-	-	-	24.4	-	-	-	-	29.6
RPMM (Baseline) [36] †	ECCV'20	25.1	30.3	24.5	24.7	26.1	26.0	32.4	26.1	27.0	27.9
RPMM [36] †	ECCV'20	29.5	36.8	29.0	27.0	30.6	33.8	42.0	33.0	33.3	35.5
PPNet* [18] †	ECCV'20	34.5	25.4	24.3	18.6	25.7	48.3	30.9	35.7	30.2	36.2
PFNet [30] ‡	TPAMI'20	36.8	41.8	38.7	36.7	**38.5**	40.4	46.8	43.2	40.5	**42.7**

Employed architectures: † ResNet50, ‡ ResNet101

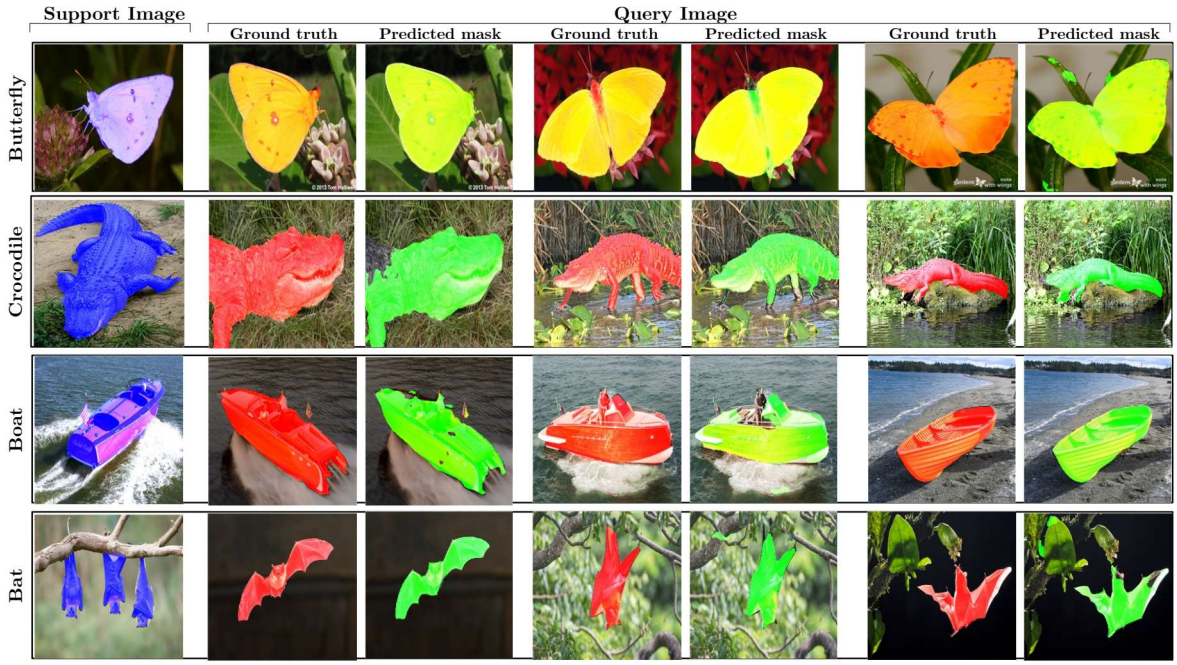

Figure 1. Visual results on FSS-1000 class dataset in 1-way 1-shot setting using the proposed method. The support set, as well as predictions on several query images with corresponding ground truths are shown.

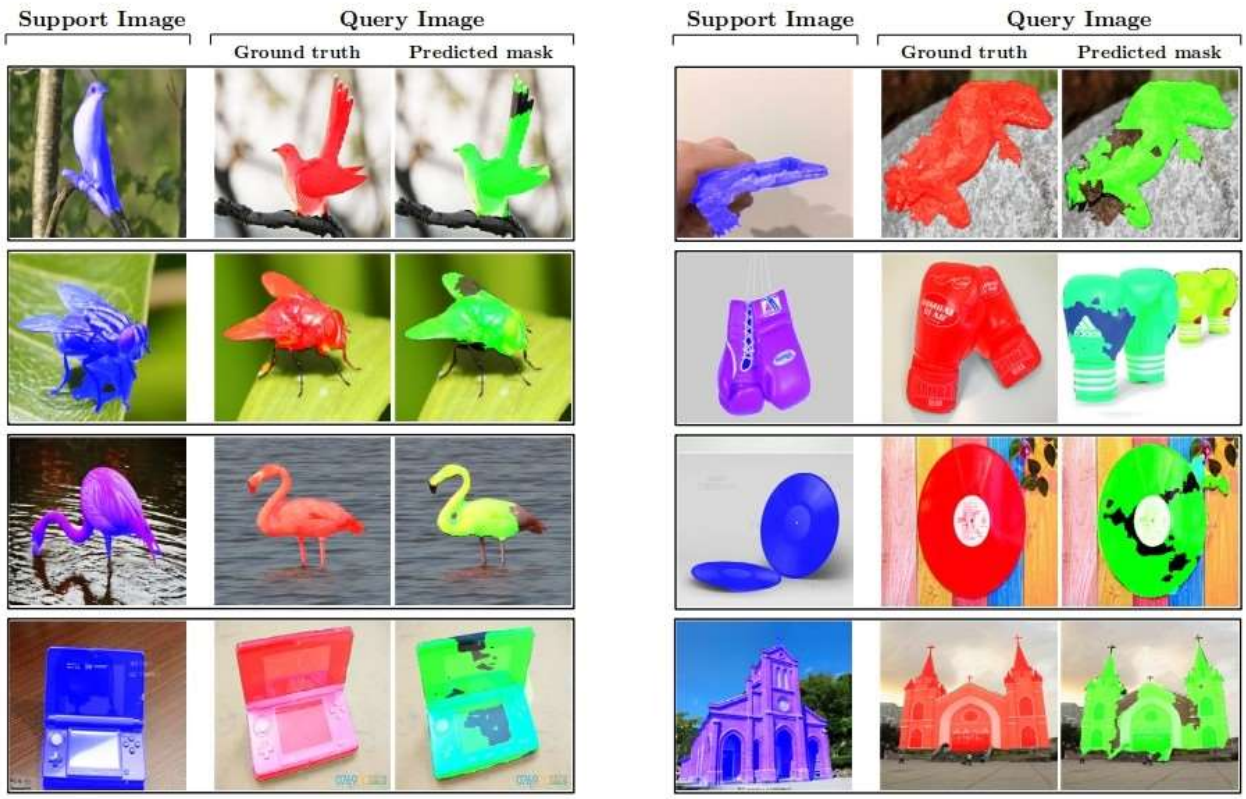

Figure 2. Visual examples of *bad* segmentation results on the FSS-1000 class dataset in 1-way 1-shot setting using the proposed method. The support set, as well as predictions on several query images with corresponding ground truths are shown.

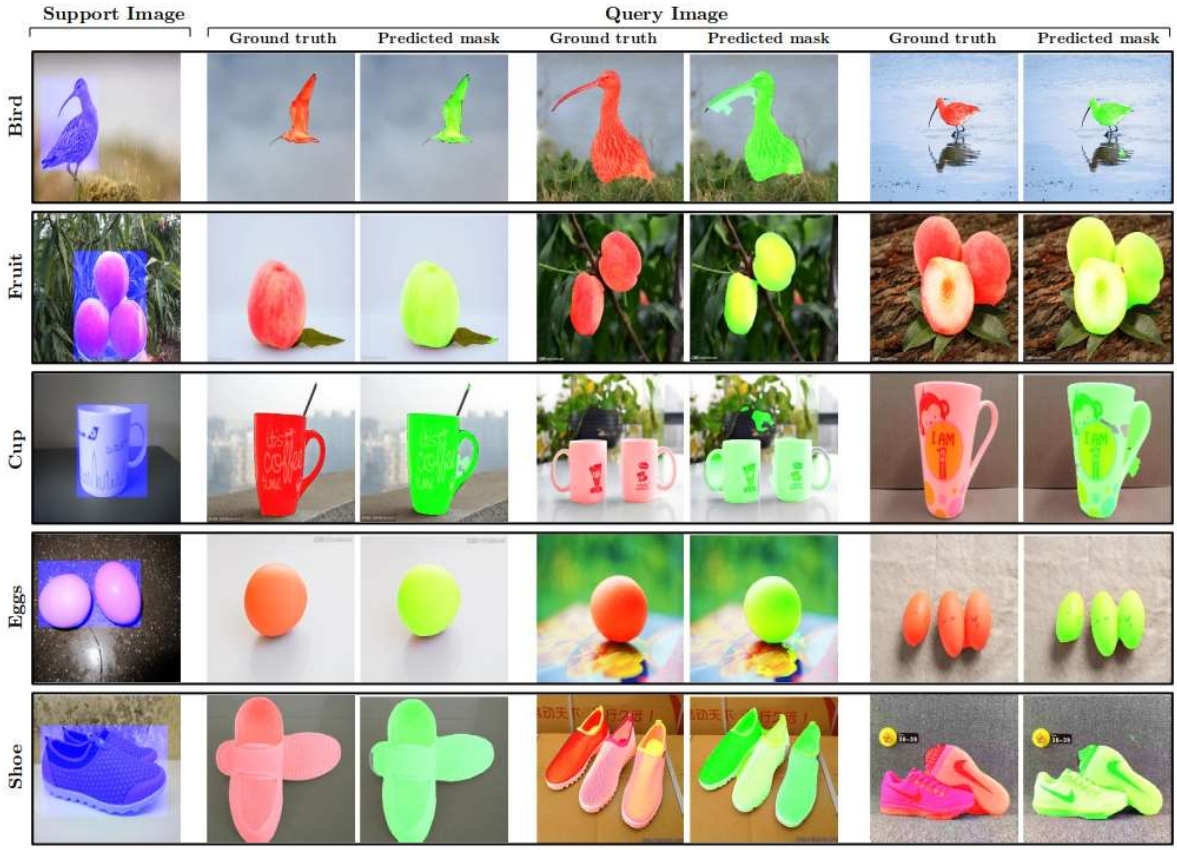

Figure 3. Visual examples of segmentation results on the FSS-1000 class dataset in 1-way 1-shot setting using the proposed method *with bounding box annotations*. The support set (i.e., image and its corresponding bounding box annotation), as well as predictions on several query images with corresponding ground truths are shown.

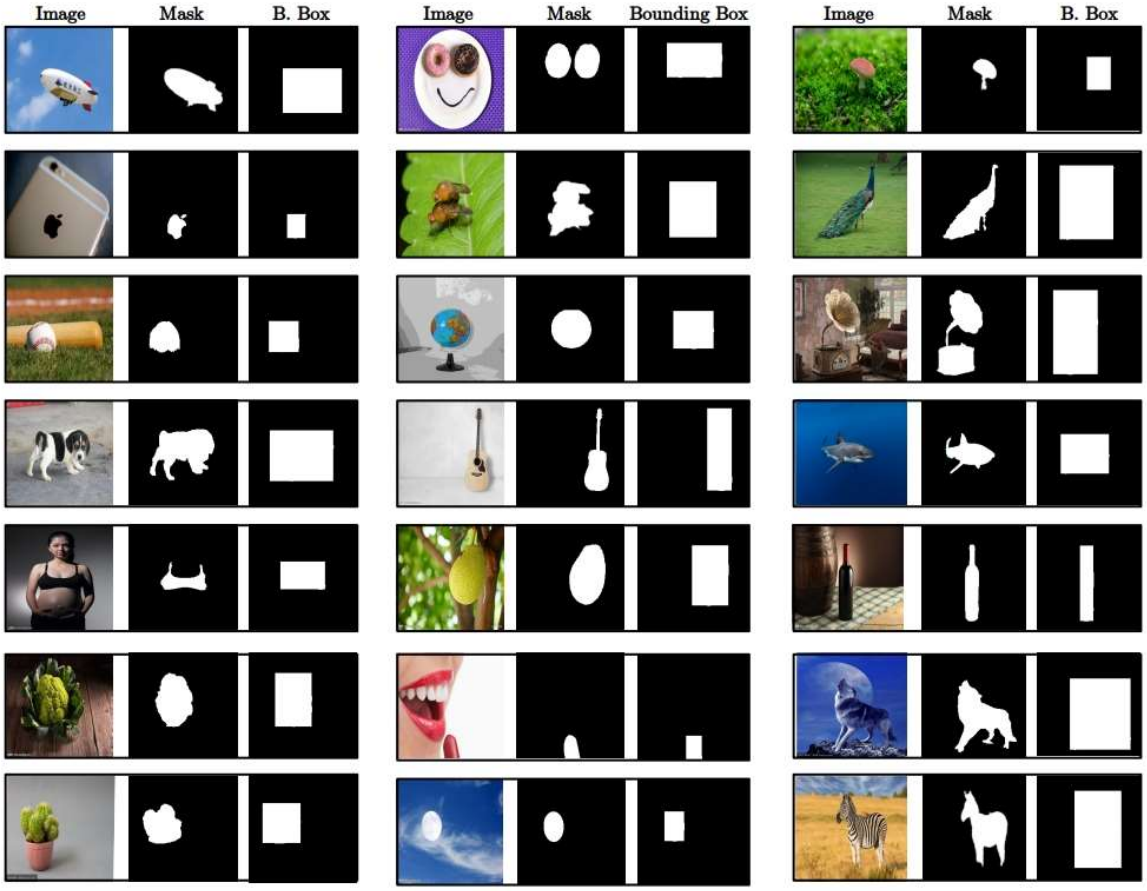

Figure 4. Examples of *bounding box annotations* generated on the FSS-1000 class dataset.