A Strong Baseline for Generalized Few-Shot Semantic Segmentation Appendix

A. Datasets

We evaluated our method on two widely-used few-shot segmentation benchmarks: PASCAL-5^i [5] and COCO-20^i [5]. The former is built based on PASCAL VOC 2012 [1] (containing 20 semantic classes) with additional annotations from SDS [2], while the latter is built from MS-COCO [4] (containing 80 semantic classes). In both datasets, the classes are split into 4 disjoint subsets, and the experiments are done in a cross-validation manner. For each fold, the set of novel classes is extracted from one of these subsets while the union of the remaining subsets will be the set of base classes. Furthermore, as discussed in the main paper, we have introduced a new scenario, where the number of novel classes is increased, referred to as PASCAL-10^i. The semantic classes in each fold of PASCAL-10^i are detailed in Tab. 1.

B. Ablation on the number of iterations

In our empirical validation, the proposed loss function, $\mathcal{L}_{\text{DIaM}}$, is optimized for a fixed number of iterations ($n = 100$), *which was chosen arbitrarily*. As demonstrated in Fig. 1 and Tab. 2 the metrics reach high values using only a few iterations. This finding shows that we can speed up the adaptation further by reducing the number of iterations while keeping the performance relatively intact. Also, continuing the adaptation for longer does not hurt the performance and the metrics stay almost the same. Please note that, as stated, the number of iterations in all the experiments in the main manuscript was set to 100 arbitrarily, disregarding the findings of this section.

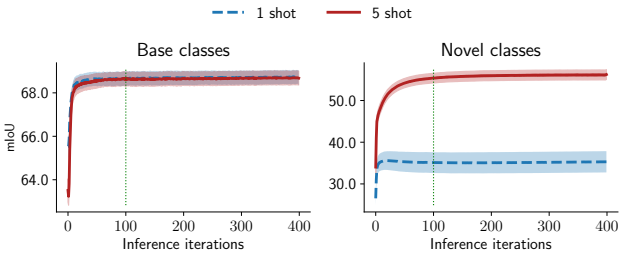

Figure 1. **Performance of the method as the number of iterations in the adaptation phase increases.** Metrics improve rapidly and first and the improvements slow down as the model is further optimized. The dotted green line indicates our choice for the number of iterations in the main manuscript. Results are provided for PASCAL-5^i.

C. Detailed results

The evaluation of our approach is performed in a cross-validation manner. In particular, there exist 4 folds for each of PASCAL-5^i and COCO-20^i benchmarks, and 2 folds for PASCAL-10^i. In the main manuscript, the reported results are obtained by averaging over all the folds in these benchmarks. Table 3 shows the performance of our model on each fold individually.

D. Harmonic mean

Following [8], we provide in Tab. 4 the harmonic mean score, referred to as *H-Mean*, of CAPL [7], BAM [3], and our method for reference. Using this metric increases the overall performance gap between our method and existing approaches.

	Novel classes	Base classes
PASCAL-10^0	aeroplane, bicycle, bird, boat, bottle, bus, car, cat, chair, cow	diningtable, dog, horse, motorbike, person, potted plant, sheep, sofa, train, tv/monitor
PASCAL-10^1	diningtable, dog, horse, motorbike, person, potted plant, sheep, sofa, train, tv/monitor	aeroplane, bicycle, bird, boat, bottle, bus, car, cat, chair, cow

Table 1. **Semantic classes in each fold of PASCAL-10^i.** In this benchmark, each fold contains 10 novel classes and hence introduces more difficulties in the generalized few-shot segmentation scenario.

# iterations	1-Shot			5-Shot		
	Base	Novel	Mean	Base	Novel	Mean
10	70.15	35.38	52.77	69.92	48.51	59.22
50	70.79	35.33	53.06	70.63	54.15	62.39
100	70.89	35.11	53.00	70.85	55.31	63.08
200	70.87	35.10	52.99	70.81	56.00	63.41
300	70.91	35.18	53.05	70.83	56.19	63.51
400	70.88	35.30	53.09	70.85	56.22	63.54

Table 2. **Precise values of the performance metrics, at selected points on the plots in Fig. 1.** The shaded row indicates our choice for the number of iterations reported in the main manuscript. Results are provided for PASCAL-5^i.

Benchmark	Fold	1-Shot			5-Shot		
		Base	Novel	Mean	Base	Novel	Mean
PASCAL-5^i	0	71.33	29.36	50.35	71.06	53.72	62.39
	1	69.54	46.72	58.13	69.63	63.33	66.48
	2	69.10	27.07	48.09	69.12	54.01	61.57
	3	73.60	37.30	55.45	73.60	50.19	61.90
	mean	70.89	35.11	53.00	70.85	55.31	63.08
COCO-20^i	0	49.01	15.89	32.45	48.90	24.86	36.88
	1	46.83	19.50	33.17	47.10	33.94	40.52
	2	48.82	16.93	32.88	49.12	27.15	38.14
	3	48.45	16.57	32.51	48.37	28.95	38.66
	mean	48.28	17.22	32.75	48.37	28.73	38.55
PASCAL-10^i	0	68.69	34.40	51.55	68.49	55.94	62.22
	1	71.83	28.17	50.00	72.00	47.84	59.92
	mean	70.26	31.29	50.77	70.25	51.89	61.07

Table 3. **Detailed results for each fold.** For each of the benchmarks, the performance of our method is presented for all the folds.

	PASCAL-5^i					
	1-Shot			5-Shot		
Method	Base	Novel	H-Mean	Base	Novel	H-Mean
CAPL [7]	64.80	17.46	27.51	65.43	24.43	35.58
BAM [3]	**71.60**	27.49	39.73	**71.60**	28.96	41.24
DIaM	70.89	**35.11**	**46.96**	70.85	**55.31**	**62.12**
	COCO-20^i					
	Base	Novel	H-Mean	Base	Novel	H-Mean
CAPL [7]	43.21	7.21	12.36	43.71	11.00	17.58
BAM [3]	**49.84**	14.16	22.05	**49.85**	16.63	24.94
DIaM	48.28	**17.22**	**25.39**	48.37	**28.73**	**36.05**

Table 4. **Quantitative evaluation on PASCAL-5^i and COCO-20^i compared to GFSS methods, using harmonic mean as the overall score.**

E. Including the background in the base score

CAPL [7] takes into account the background IoU, which is generally higher than the IoU of base classes, when computing the *Base* metric. We, the same as [3], believe that since background does not represent an object of interest, the model's performance on this class should not be considered. Nevertheless, including the background IoU in the metrics leads to marginal performance differences that are consistent across all methods. In Tab. 5, for GFSS methods, the background IoU is included in the *Base* metric, reframing it as *Base w/ bg* to avoid confusion.

Method	1-Shot			5-Shot		
	Base w/ bg	Novel	Mean	Base w/ bg	Novel	Mean
CAPL [7]	66.37	17.46	41.92	66.95	24.43	45.69
BAM [3]	72.00	27.49	49.75	**72.36**	28.96	50.66
DIaM	**72.04**	**35.11**	**53.58**	72.12	**55.31**	**63.72**

Table 5. **Quantitative evaluation on PASCAL-5^i compared to GFSS methods, including the background performance in the metrics.**

F. Practical setting: employing the whole training dataset

As discussed in the main manuscript, CAPL [7] and BAM [3] filter out training images that contain novel classes. This procedure is impractical in real-world scenarios since it needs a training set in which novel classes are labeled, undermining the goal of few-shot learning, *i.e.*, having only a few labeled examples of the novel classes. Recent empirical evidence [6] has shown that such additional step can lead to performance gain on novel classes. In Tab. 6, we have changed the training procedure of CAPL and BAM and avoided removing images containing novel classes from training. More specifically, the potential objects from novel classes are labeled as *background* during training.

Method	1-Shot			5-Shot		
	Base	Novel	Mean	Base	Novel	Mean
CAPL [7]	71.59	12.69	42.14	**71.71**	19.58	45.65
BAM [3]	**71.61**	19.35	45.48	71.66	26.33	49.00
DIaM	70.89	**35.11**	**53.00**	70.85	**55.31**	**63.08**

Table 6. **Quantitative evaluation on PASCAL-5^i compared to GFSS methods, in the experimental setting in which the whole training dataset is employed.** In this setting, images containing novel classes are not removed from the training process. Confirming the findings in [6], this procedure enhances the performance on novel classes. It is also worth noting that although *Novel* score has been decreased for CAPL, its *Base* score has been considerably increased.

G. Adaptation of BAM [3] to multi-class GFSS

The results reported in [3] for the GFSS task are based on an evaluation protocol in which only one novel class can be recognized in a query image. Indeed, the *meta-learner*

from this method can only provide binary (*i.e.*, *background vs foreground*) predictions and is not practical in the setting where multiple novel classes are to be predicted at the same time. To be able to incorporate BAM in our empirical validation, we had to adapt it so that it can predict multiple novel classes simultaneously. These modifications are detailed in what follows. First, instead of selecting K support samples of a novel class c and asking the model to segment class c in a query image, we form $|\mathcal{C}^n|$ different support sets, $\mathbb{S}_i$, one for each $i \in \mathcal{C}^n$. Recall that $\mathcal{C}^n$ is the set of novel classes and that $\mathbb{S}_i$ contains K samples labeled for each novel class i. Second, we run BAM $|\mathcal{C}^n|$ times and, in each inference, we give the same query image alongside $\mathbb{S}_i$, resulting in a foreground probability map for each class i, called $\mathbf{m}_i$. Then, we need to create a single mask containing all the novel class predictions to further use it in BAM's fusion mechanism. To do this, we create an aggregated novel map, $\mathbf{a}$, which is formed based on the resulted $|\mathcal{C}^n|$ maps, in such a way that for each pixel j:

$$\mathbf{a}(j) = \underset{i \in \mathcal{C}^n}{\operatorname{argmax}} \ \mathbf{m}_i(j). \tag{1}$$

We also form $\mathbf{p_a}$ to preserve the probability of the selected indices, which will later be compared to the predefined threshold τ introduced in [3]:

$$\mathbf{p_a}(j) = \max_{i \in \mathcal{C}^n} \ \mathbf{m}_i(j). \tag{2}$$

Then, $\mathbf{a}$ and $\mathbf{p_a}$ alongside the base map predicted by BAM's *base-learner* for the query, $\hat{\mathbf{m}}_b$, are used to perform the fusion procedure following [3]. More specifically, the final prediction is formulated as

$$\hat{\mathbf{m}}_g(j) = \begin{cases} \mathbf{a}(j) & \mathbf{p_a}(j) > \tau, \\ \hat{\mathbf{m}}_b(j) & \mathbf{p_a}(j) \leq \tau \quad \text{and} \quad \hat{\mathbf{m}}_b(j) \neq 0, \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

This change definitely slows the inference by an order of magnitude, but this is inevitable because of the nature of meta-learning few-shot segmentation, which needs to be accommodated to produce multi-class semantic maps, as they are tailored to binary maps.

H. Visual examples

In the main manuscript, we presented qualitative results on PASCAL-5^i using different versions of our loss function. We observed that in the absence of the knowledge distillation term, the model misclassifies some of the previously learned base classes as novel ones. Figure 2 shows similar results on COCO-20^i, where the same trend is observed. For instance, in the first two rows, base classes *cell phone* and *keyboard* are mistakenly classified as the novel class *remote*. Note that this problem is fixed when the knowledge distillation term is added to the loss function.

References

[1] Mark Everingham, SM Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111(1):98–136, 2015. 1

[2] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Simultaneous detection and segmentation. In *European conference on computer vision*, pages 297–312. Springer, 2014. 1

[3] Chunbo Lang, Gong Cheng, Binfei Tu, and Junwei Han. Learning what not to segment: A new perspective on few-shot segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8057–8067, 2022. 1, 2, 3

[4] 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. Springer, 2014. 1

[5] Amirreza Shaban, Shray Bansal, Liu Zhen, Irfan Essa, and Byron Boots. One-shot learning for semantic segmentation. In *Proceedings of the British Machine Vision Conference (BMVC)*, pages 167.1–167.13. BMVA Press, September 2017. 1

[6] Yanpeng Sun, Qiang Chen, Xiangyu He, Jian Wang, Haocheng Feng, Junyu Han, Errui Ding, Jian Cheng, Zechao Li, and Jingdong Wang. Singular value fine-tuning: Few-shot segmentation requires few-parameters fine-tuning. In *NeurIPS*, 2022. 2

[7] Zhuotao Tian, Xin Lai, Li Jiang, Shu Liu, Michelle Shu, Hengshuang Zhao, and Jiaya Jia. Generalized few-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11563–11572, 2022. 1, 2

[8] Han-Jia Ye, Hexiang Hu, and De-Chuan Zhan. Learning adaptive classifiers synthesis for generalized few-shot learning. *International Journal of Computer Vision*, 129:1930–1953, 2021. 1

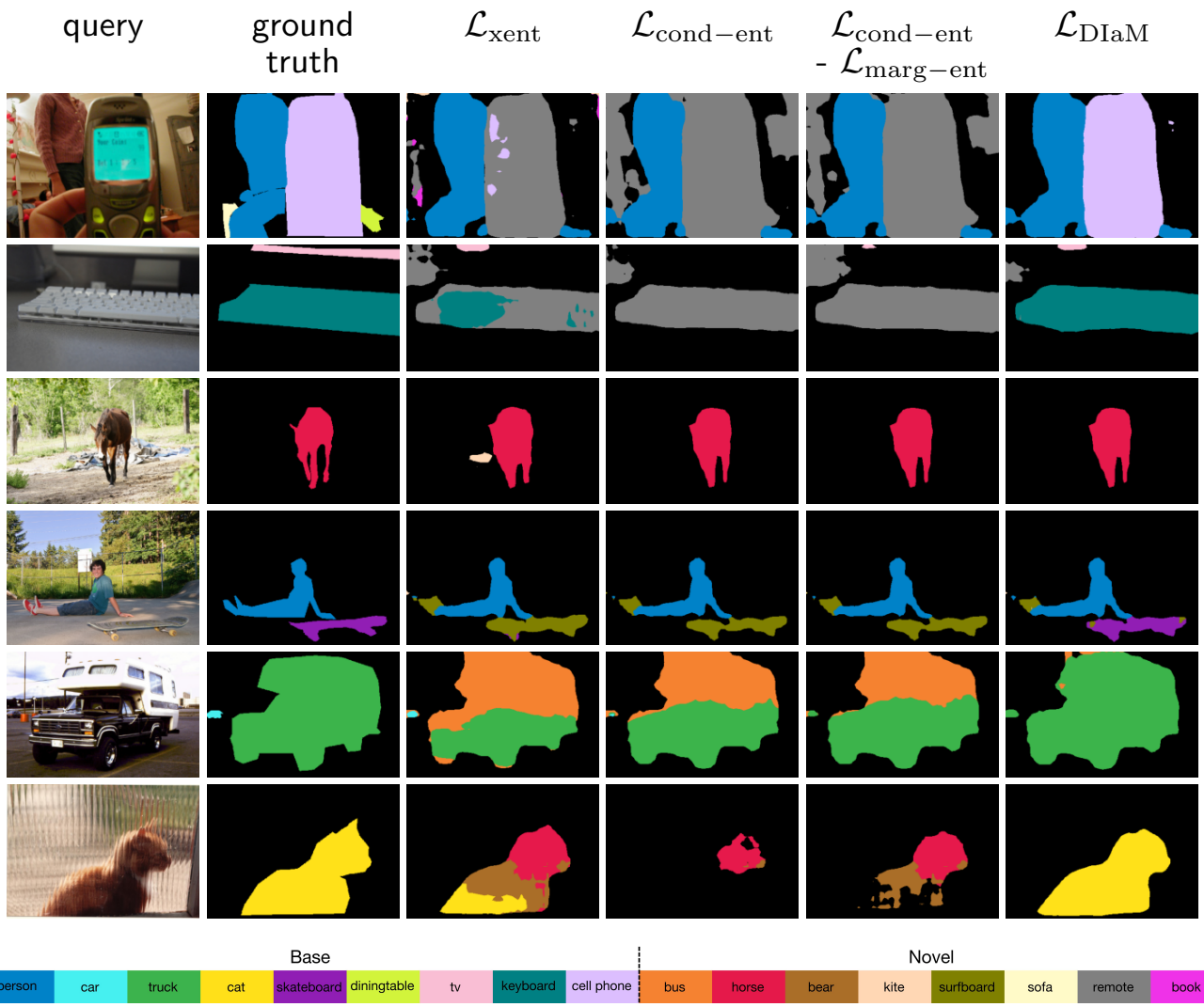

Figure 2. **Qualitative results of different terms of DIaM's loss function on COCO-20^i.** A single support set, containing the following novel classes is used for predicting every query image: *bicycle*, *bus*, *traffic light*, *bench*, *horse*, *bear*, *umbrella*, *frisbee*, *kite*, *surfboard*, *cup*, *bowl*, *orange*, *pizza*, *sofa*, *toilet*, *remote*, *oven*, *book*, *teddy bear*. Query images can contain any classes and every one of them is to be recognized. From the left, the first two columns show the query image and the ground truth, and the following columns display predictions of models using different loss functions. Results are on COCO-20^i under the 5-shot setting.