Supplementary Material: Dense Cross-Query-and-Support Attention Weighted Mask Aggregation for Few-Shot Segmentation

A. More Results

Stability and robustness to trials and hyperparameter. In addition to the 4-fold cross validation in Table 1, we conduct three repeated trials on Fold-0 of PASCAL-5^i and obtain stable IoUs (72.1±0.25). We further experiment with three hyperparamter pairs of batch size and learning rate (bs/lr) and obtain stable IoUs, too: 72.2 (48/0.001), 71.9 (24/0.001), and 71.9 (24/0.0005).

Region-wise over- and under-segmentation measures. We further employ the region-wise over-segmentation measure (ROM) and region-wise under-segmentation measure (RUM) [51] for quantitative evaluation. ROM and RUM are two novel threshold-free metrics assessing region-based over- and under-segmentation, and expected to lend greater explainability to semantic segmentation performance in real-world applications. A smaller ROM (RUM) value indicates less over- (under-) segmentation and is preferred. The 1-shot ROM and RUM values are tabulated in Table S1, where our DCAMA slightly outperforms the competent HSNet in both metrics overall.

Table S1. One-shot evaluation results in terms of ROM and RUM [51]. HSNet†: our reimplementation based on the official codes.

Method	Backbone	PASCAL-5^i		COCO-20^i		FSS-100		Overall	
		ROM	RUM	ROM	RUM	ROM	RUM	ROM	RUM
HSNet† [24]	Swin-B	0.26	**0.06**	0.15	0.07	**0.12**	0.03	0.177	0.053
DCAMA	Swin-B	**0.21**	0.07	**0.13**	**0.06**	0.15	**0.02**	**0.163**	**0.050**

Computational efficiency. In addition to what is reported in the main text, below we provide more metrics regarding computational efficiency in Table S2. Our DCAMA has slightly larger FLOPS but comparable times with HSNet for training an epoch and inference, and converges in substantially fewer epochs thus needing much less time for training.

Table S2. Computational efficiency (1-shot on COCO-20^i). HSNet†: our reimplementation based on the official codes.

Method	Backbone	Epoch to converge	Epoch time	FLOPS	Inference time
HSNet† [24]	Swin-B	355	∼4 min	103.8 G	0.13 s
DCAMA	Swin-B	90	∼4 min	109.4 G	0.13 s

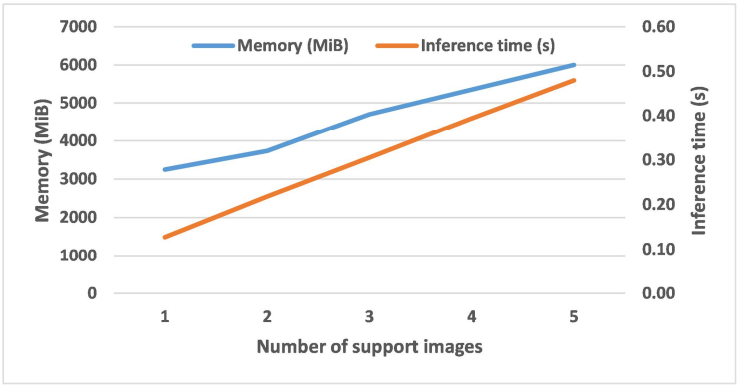

Fig. S1. Memory and latency analysis for n-shot inference.

Memory and latency analysis for n-shot inference. The analysis for $n = 1$ to 5 is presented in Fig. S1. The increases in memory and latency are approximately linear with n.

B. Further Ablation Studies

Configuration of skip connections. In Section 3.2 of the main text, we propose to skip connect (concatenate) extracted features of the input images to the integrated output of the multi-scale multi-layer Dense Cross-query-and-support Attention weighted Mask Aggregation (DCAMA) blocks, following successful experience of previous works [30, 52]. Here, we empirically determine the optimal configuration of the skip connections, by comparing the effects of concatenating (or not) the $\frac{1}{4}$, $\frac{1}{8}$, and $\frac{1}{16}$ scale features individually and jointly. As shown in Table S3, concatenating either $\frac{1}{4}$ or $\frac{1}{8}$ scale features individually brings notable performance improvement upon the baseline of no skip connection, and their joint concatenation brings further improvement to achieve the optimal performance (+3.1% and +1.1% with respect to the baseline in mIoU and FB-IoU, respectively). On the other hand, concatenating the $\frac{1}{16}$ scale features, either individually or jointly with the shallower features, leads to obvious performance

Table S3. Ablation study on feature skip-connection configuration (1-shot on PASCAL-5^i [31] with Swin-B [21] as backbone).

Feature scale			Fold-0	Fold-1	Fold-2	Fold-3	mIoU	FB-IoU
1/4	1/8	1/16						
			70.6	72.6	61.4	64.5	66.2	77.4
✓			71.0	73.3	61.7	67.0	68.2	78.1
	✓		70.7	**74.0**	63.9	65.7	68.6	77.9
✓	✓		**72.2**	73.8	**64.3**	**67.1**	**69.3**	**78.5**
		✓	66.7	62.6	49.6	60.1	59.8	71.9
✓	✓	✓	67.3	59.6	50.6	58.1	58.9	72.2

deterioration. Similar findings were also reported in previous works [38, 48], with the possible explanation that semantic information contained in high-level features is more class-specific and less generalizable. Based on these results, we choose to skip connect the $\frac{1}{4}$ and $\frac{1}{8}$ scale features in our DCAMA framework for comparison with other methods.

Contribution of multi-scale attention. In Section 3.2 of the main text, we implement the multi-layer DCAMA blocks at all scales (i.e., $\frac{1}{8}$, $\frac{1}{16}$, and $\frac{1}{32}$) allowed by our hardware, as the multi-scale strategy has proven effective in various computer vision applications. Here, to empirically evaluate the impact of the multi-scale attention, we conduct experiments to ablate the multi-scale attention—one scale at a time, following [24]. The results are shown in Table S4. As we can see, removing $\frac{1}{8}$ scale attention results in modest and slight decreases in mIoU and FB-IoU by 2.1% and 1.4%, respectively, and consecutively removing 1/16 scale attention leads to further, substantial performance degradation in both metrics by 8.5% and 6.6%, respectively. These results indicate the indispensable role of the multi-scale attention to our proposed framework.

Table S4. Ablation study on the multi-scale attention strategy (1-shot on PASCAL-5^i [31] with Swin-B [21] as backbone).

Attention scale			Fold-0	Fold-1	Fold-2	Fold-3	mIoU	FB-IoU
1/8	1/16	1/32						
		✓	62.0	66.8	49.4	56.7	58.7	70.5
	✓	✓	70.0	73.3	61.5	64.0	67.2	77.1
✓	✓	✓	**72.2**	**73.8**	**64.3**	**67.1**	**69.3**	**78.5**

Impact of skip-connecting support features/the number of support foreground and background pixels. The ablation results in Table S5 show that both excluding support features from skip connection (row a) and normalizing foreground and background region sizes (to 600 pixels following CyCTR; row b) impair the performance. Specifically, the high-level feature maps that are skip connected have relative large receptive fields, hence are helpful even not aligned.

Table S5. Ablation study on skip-connecting support features and the number of support foreground and background pixels (1-shot on PASCAL-5^i [31] with Swin-B [21] as backbone).

Ablation	Supp. feat.	Num pix.	Fold-0	Fold-1	Fold-2	Fold-3	mIoU	FB-IoU
a	✗	All	71.1	73.6	63.5	**67.5**	68.9	77.9
b	✓	600	70.7	66.6	62.9	62.6	65.7	75.9
DCAMA	✓	All	**72.2**	**73.8**	**64.3**	67.1	**69.3**	**78.5**

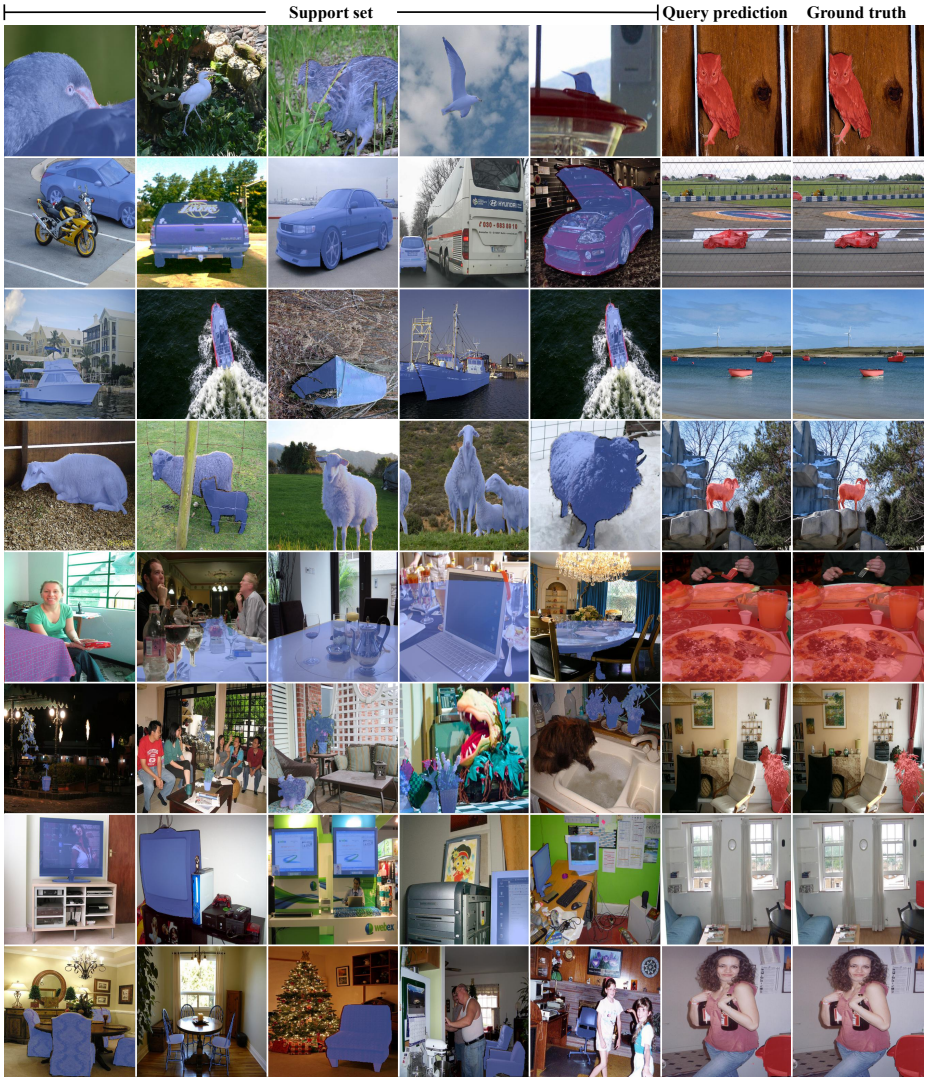

Fig. S2. Example 5-shot segmentation results by the proposed DCAMA framework (with Swin-B [21] as backbone) on PASCAL-5^i, in the presence of intra-class variations, size differences, complex background, and occlusions.

C. More Visual Analysis

Visualization of more segmentation results. Fig. S2 shows example 5-shot segmentation results by our proposed DCAMA framework, complementary to the 1-shot segmentation results shown in the main text.

Qualitative limitation analysis. We empirically explore the limitations of the proposed DCAMA framework, based on qualitative analysis of failure cases in the 1-shot settings. Above all, as shown in Fig. S3, all the failure cases in the 1-

Fig. S3. Left: some representative failure cases on PASCAL-5^i in 1-shot setting. Right: the same cases in 5-shot setting, where the extra support images and masks help our DCAMA framework produce accurate segmentation of the query images in these challenging cases.

shot setting are accurately segmented with the extra support images in the 5-shot setting, in accordance with the findings of Min et al. [24]. As this finding clearly demonstrates the efficacy of 5-shot segmentation and is expected, it is interesting to dig deeper to find out why the 1-shot segmentation fails in these cases and how the extra support images help. Systematically, we roughly categorize the 1-shot failures into three types of limitations: limited representativeness of the support image, intra-class variation, and inter-class similarity, which sometimes occur together, too.

Limited representativeness happens when the target class is under-represented in the support image, e.g., the object is largely occluded (row (a) of Fig. S3), too small (row (b)), or in a quite different perspective (row (c)). In contrast, for intra-class variation, although the object in the support image is complete and in normal size and view, the instance in the query image can look differently despite belonging to the same major class (rows (d) and (e)). The third type of limitation is inter-class similarity, i.e., the similarity between different classes causing difficulty in differentiating them (rows (f)–(h)). We can also observe its concurrence with the other two types (rows (f) and (h)). These limitations are

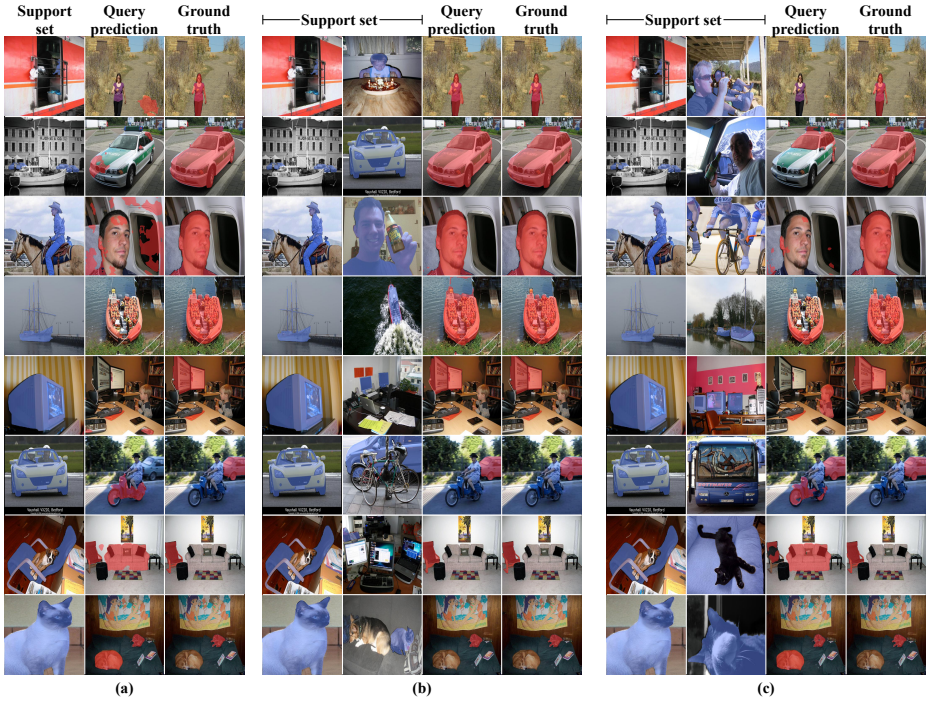

Fig. S4. Control experiment on fixing 1-shot failures by adding one extra support image: (a) the same cases as shown in Fig. S3 (left); (b) the failures are effectively fixed with an informative support image added; and (c) adding a non-informative support image helps little.

inherent in Few-Shot Learning (FSL) and faced by all FSL algorithms. Theocratically, these limitations can be effectively overcome by introducing the missing support information with a few additional, informative support images.

To validate this, we pick only one extra informative support image and use it together with the original 1-shot support image for a 2-shot inference. As shown in Fig. S4(b), by providing more completed information about the target class and extra information about the within-class variance and inter-class differentiation, respectively, the 1-shot failures are fixed, as expected. On the other hand, we also experiment with replacing the added informative support image with a non-informative one as a control group, and the corresponding results in Fig. S4(c) are apparently inferior to those in Fig. S4(b), with little or no improvement upon the original 1-shot results. This suggests that the actual information contained in the support images may matter more than the absolute number of them. To this end, it is desirable to integrate active learning with the few-shot segmentation for practical application.

Visualization of point-wise attention maps. For an intuitive perception of how a specific point in the query image correlates to all pixels of a support image, we visualize the attention weight maps for two query pixels—one foreground

Fig. S5. Point-wise attention weight maps for two pixels (red dot: foreground; and green dot: background) in the query image, to all pixels of a support image. The attention maps are obtained by averaging all the multi-scale multi-layer attention maps (upsampled where applicable).

and one background—in Fig. S5. As we can see, both of the query pixels have the strongest attention weights around their most similar regions in the support image, i.e., the dark cat's eye and below the dark cat's hind legs, respectively, while faint responses can also be observed around similar regions of the orange cats.

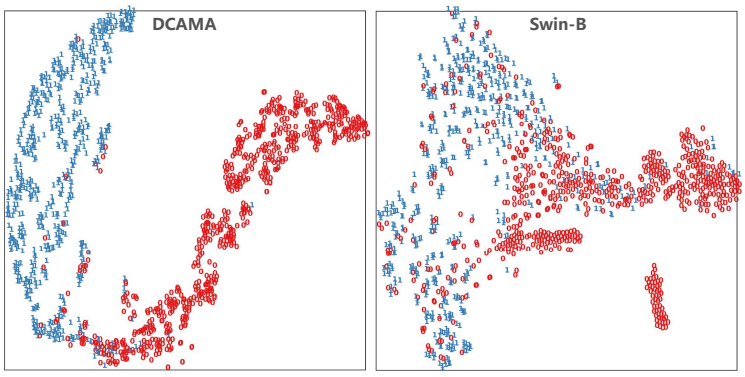

Fig. S6. T-SNE plots for a testing class "plane": blue (red) indicates foreground (background).

T-SNE visualization. For an intuitive perception of how well the model learns the metric space, we employ t-SNE for qualitative analysis (Fig. S6). The penultimate-layer features of DCAMA are more separable than those of the frozen backbone (Swin-B), indicating effective learning of the metric space.