Supplementary Material for
Learning What Not to Segment: A New Perspective on Few-Shot Segmentation

A. Calculation of FLOPs

Floating point operations per second (FLOPs) is utilized to evaluate the computational complexity of our model in the ablation study (refer to Sec. 5.3). Here we introduce the specific calculation process. Given the low-level features $\mathbf{f}^{\mathrm{s}}_{\mathrm{low}}, \mathbf{f}^{\mathrm{q}}_{\mathrm{low}} \in \mathbb{R}^{C\times H\times W}$, we first compute the corresponding Gram matrices $\mathbf{G}^{\mathrm{s}}, \mathbf{G}^{\mathrm{q}}$. Second, we perform subtraction $\mathbf{G}^{\mathrm{s}} - \mathbf{G}^{\mathrm{q}}$ to evaluate the difference. Finally, we calculate the Frobenius norm of the difference to get an overall indicator ψ. The number of FLOPs can be defined as:

$$\begin{aligned} \mathrm{FLOPs} &= 4C^2N + C^2 + 2C^2 \\ &= C^2\left(4N+3\right) \end{aligned}, \tag{S1}$$

where $N = H \times W$, and the three terms in the first row correspond to the three processes above[‡].

B. Implementation Details

As mentioned in Sec. 4.5, we simply fuse the predictions of the base learner and the final predictions after ensemble according to a predefined threshold τ=0.9 to obtain the holistic segmentation results $\hat{\mathbf{m}}_{\mathrm{g}}$, as presented in Eq. (18). There is actually another alternative extension scheme, which can be defined as:

$$\hat{\mathbf{m}}_{\mathrm{g}}^{(x,y)} = \begin{cases} 1 & \mathbf{p}_{\mathrm{f}}^{1;(x,y)} > \tau \text{ and } \hat{\mathbf{m}}_{\mathrm{b}}^{(x,y)} = 0 \\ \hat{\mathbf{m}}_{\mathrm{b}}^{(x,y)} & \hat{\mathbf{m}}_{\mathrm{b}}^{(x,y)} \neq 0 \\ 0 & \text{otherwise} \end{cases}. \tag{S2}$$

The differences between two schemes are as follows: the former is mainly based on the final predictions, leveraging the base learner to determine the base pixels in the background region; the latter, by contrast, is primarily based on the predictions of the base learner, using the final predictions to determine the novel pixels in the background region. In our experiments, BAM achieves similar results using both schemes, with the former slightly better; however, the baseline approach (w/o ensemble module) can only produce tolerable results when the latter scheme is adopted, which further verifies the importance of the ensemble module to correct the coarse predictions of meta learners.

[‡]Taking the third convolutional block B_2 of ResNet50 [16] backbone as an example, the FLOPs is 3.78G with $\mathbf{f}_{\mathrm{low}} \in \mathbb{R}^{512\times 60\times 60}$.