Addressing Background Context Bias in Few-Shot Segmentation through Iterative Modulation

Supplementary Material

The supplementary materials are arranged as follows. In Sec.A, additional details are presented to provide a more comprehensive understanding of our method. In Sec.B, further experiments are included to validate the effectiveness of our method. In Sec.C, we present more visualizations of predictions generated from different iterations of our iterative structure.

A. More Details of Method

A.1. Spire-shape Function for Histograms.

In Sec.4.3 of the main paper, to obtain the structure-wise evolution feature $\mathbf{E}^s$, we introduce histograms $\mathbf{H}_I$ and $\mathbf{H}_O$ to extract fix-shaped and structure-wise features from affinity maps $\mathbf{A}_I$ and $\mathbf{A}_O$ respectively. For example, to obtain $\mathbf{H}^I$, the continuous subtraction range [0,1] is subdivided into L discrete bins, with the l-th one denoting the interval $I_l = [(l-1)/L, l/L]$. We flatten $\mathbf{A}_I$ to the shape $\mathbb{R}^{N_f^2}$, then each dimension $\mathbf{H}_I^l$ on $\mathbf{H}_I$ counts the number of dimensions in $\mathbf{A}_I$ with values falling into the interval I_l. This step is formulated as:

$$\mathbf{H}_I^l = \sum_{i=1}^{N_f^2} \mathbb{1}\left(\frac{l-1}{L} < \mathbf{A}_I^i < \frac{l}{L}\right), \; l \in [0, L-1]. \qquad (1)$$

The ideal $\mathbb{1}$ should be a binary judge function that outputs 1 if $\mathbf{A}_I^i$ falls into I_l, and 0 otherwise. However, such a function is non-differentiable, thus prohibiting the gradients from being back-propagated to the earlier layers. To address the issue, we replace the original binary function with a differentiable spire-shape function. To facilitate expression, we use a to represent $\frac{l-1}{L}$ and b to represent $\frac{l}{L}$. In this way, this spire-shape function can be formulated as:

$$\mathbb{1}\left(a < \mathbf{A}_l^i < b\right) = \begin{cases} 1 - \left|\mathbf{A}_I^i - \dfrac{a+b}{2}\right| & if \quad a < \mathbf{A}_I^i < b \\ 0 \quad else & \end{cases} \qquad (2)$$

By doing so, the gradients can be propagated successfully, allowing the network to be end-to-end trained in optimization.

A.2. Extension to K-shot Setting

In the main paper, we describe our method under the 1-shot setting where only one support image is available ($K = 1$). Our method can be easily extended to K-shot setting by solely modifying the initialization approach of $\mathbf{S}_{\mathrm{QP}}^1$ for the first iteration in the iterative framework. Specifically in the K-shot setting, for each support image $I_s^k, k \in [1, K]$, we first extract a feature f_s^k from the backbone network. Then, to initialize $\mathbf{S}_{\mathrm{QP}}^1$ for the QP step of the first iteration, we concatenate the features of all foreground pixels across $\{f_s^k\}_{k=1}^K$ from all K support images. The subsequent steps remain the same as in the 1-shot setting.

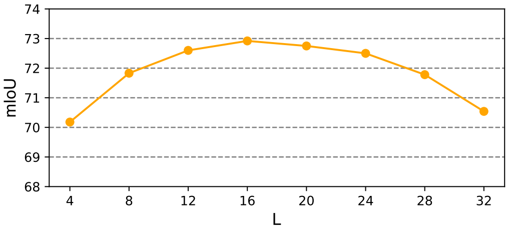

Figure 1. mIoUs when setting L to different values. We choose L=16 as our final setting.

B. More Experimental Results

B.1. Ablation of Bin Number

The generation of histograms used for structure-wise evolution feature $\mathbf{E}^s$ involves a step that equally subdivides the continuous range [0, 1] into L bins, with the l-th one denoting the interval $[(l-1)/L, l/L]$ (refer to Sec 4.3 of the main paper for details). In Fig.1, we present the validation results of using different numbers of bins. The experiments are conducted on PASCAL-5^i fold-0 with the ResNet50 backbone. It is observed that when L is greater than 12 and less than 24, the mIoU remains stable. When L is too small, the histogram is coarse, resulting in less effective structure information extraction and lower validation accuracy. Conversely, when L is too large, overfitting may occur to hinder the model's effectiveness. Based on experimental results, we choose 16 as the setting of L.

B.2. Ablation of Information Cleansing

In the information cleansing (IC) step, we propose a confidence-biased attention to extract the accumulated noisy information, in which the softmax matrix in an attention is summed with the confidence variance V. To validate the effectiveness of this key design, we remove the bias term V in Eq.5 of main paper, transferring the confidence-biased attention into a normal one. The results are shown in Table. 1. This modification decreases mIoU from 72.92% to 68.53%. Compared to the method without the IC step,

Method	mIoU
Ours	72.92
Ours w/o using V in Eq.5 of main paper	68.53
Ours w/o IC step	68.20

Table 1. Ablation study of information cleansing step. Removing V in Eq.5 of main paper significantly decreases performance, demonstrating the importance of the proposed confidence-biased attention.

which yields 68.20% mIoU, using IC step without the bias term V almost brings no improvement. This demonstrates that the improvement of IC step is not from the increase in the number of parameters, but due to the removement of noisy information that is extracted by our proposed novel attention mechanism.

C. visualization from Different Iterations

We present more visualization results of our method in Fig.2, which shows the segmentation predictions generated from different iterations in our framework ($t = 1, 2, 3$). Due to the feature misalignment resulting from the different background contexts, the predictions from the first iteration are suboptimal. Our iterative structure remarkably improves the performance, resulting in gradually refined segmentation results from the first to the final iteration.

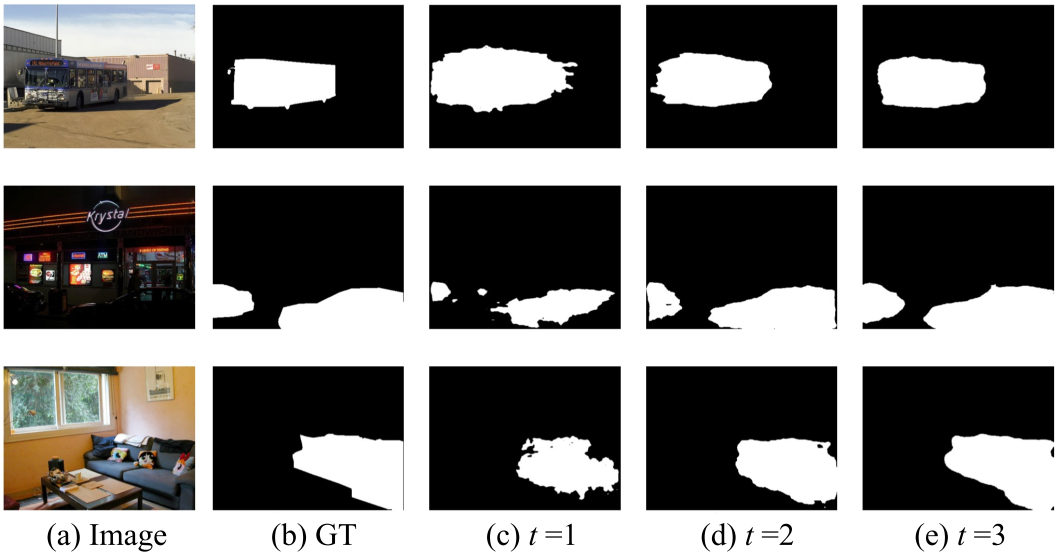

Figure 2. Visualizations of predictions from different iterations t of our iterative structure.