Frequency-Aware Affinity for Weakly Supervised Semantic Segmentation

Supplementary Material

This supplementary material provides more details and results that are not included in the main paper. The contents are organized as follows:

- Additional Formulations in the Methodology
- Analysis of Hyper-parameters
- Analysis of Masks
- CAM Visualizations

1. Additional Formulations in the Methodology

Quality Value In the methodology section, we introduce a quality value α [2] to design the selective mask $\mathbf{M}_d$. This value is computed from the discrepancy between the smoothed affinity $\mathbf{A}_s$ and the low-frequency feature relations $\mathbf{S}_l$ as :

$$\alpha = \sum_{i=1}^{hw}\sum_{j=1}^{hw} |\mathbf{S}_l^{i,j} - \mathbf{A}_s^{i,j}|, \tag{1}$$

where hw is the patch token number. By using this quality value, we successfully filter out the irrelevant relations and only transfer the denser distribution from $\mathbf{S}_l$ to $\mathbf{A}_s$ in the distribution alignment loss $\mathcal{L}_a$.

Frequency Features Moreover, we also utilize the low-frequency features $\mathbf{F}_{ld}$ and high-frequency features $\mathbf{F}_{hd}$ through applying the Fourier Transform $\mathcal{F}$ and inverse Fourier Transform $\mathcal{F}^{-1}$ to the decoder features $\mathbf{F}_d$, the process [1] can be defined as:

$$\mathbf{F}_{ld} = \mathcal{F}^{-1}(\mathbf{B}\mathcal{F}(\mathbf{F}_d)), \tag{2}$$

where $\mathbf{B}$ denotes the binary mask in the frequency domain for selecting the low-frequency components. Then, the high-frequency features $\mathbf{F}_{hd}$ can be obtained by subtracting the $\mathbf{F}_{ld}$ from $\mathbf{F}_d$.

Identification of Unreliable Relations The selection of unreliable relations as follows: we first define the unreliable relations mask $\mathbf{H}$ as:

$$\mathbf{H} = (1 - \mathbf{M}_+) \cdot (1 - \mathbf{M}_-), \tag{3}$$

where 1 denotes an all-ones mask of size $hw \times hw$. For identifying these unreliable relations, we choose to apply a $r \times 1$ window for each patch token pair (p, q) selected (is equal to 1) in $\mathbf{H}$ and measure the number of positive and negative relations around them in this window separately. If the neighboring relations of the patch tokens p and q within the window are strongly positive, the relation between (p, q) is likely to be reinforced, as adjacent patch tokens typically exhibit similar relational properties and structural consistency. The number of their neighboring positive $\mathbf{N}_+$ and negative relations $\mathbf{N}_-$ in the window can be measured as :

$$\begin{aligned}\mathbf{N}_+^{p,q} &= \sum_{u=q-r}^{q+r} \mathbf{M}_+[u,p] + \sum_{v=p-r}^{p+r} \mathbf{M}_+[q,v], \\ \mathbf{N}_-^{p,q} &= \sum_{u=q-r}^{q+r} \mathbf{M}_-[u,p] + \sum_{v=p-r}^{p+r} \mathbf{M}_-[q,v].\end{aligned} \tag{4}$$

where u and v are the neighboring patch tokens of q and p in the window. The final reliable masks can be updated as :

$$\begin{aligned}\mathbf{M}_+^{p,q} &= \begin{cases} 1, & \mathbf{H}^{p,q} = 1 \text{ and } \mathbf{N}_+^{p,q} > \mathbf{N}_-^{p,q}, \\ \mathbf{M}_+^{p,q}, & \text{otherwise}, \end{cases} \\ \mathbf{M}_-^{p,q} &= \begin{cases} 1, & \mathbf{H}^{p,q} = 1 \text{ and } \mathbf{N}_-^{p,q} \geq \mathbf{N}_+^{p,q}, \\ \mathbf{M}_-^{p,q}, & \text{otherwise}, \end{cases}\end{aligned} \tag{5}$$

2. Analysis of Hyper-parameters

In this method, we utilize two hyper-parameters λ_1 and λ_2 in the overall training objective. We further validate the effectiveness of these choices through comprehensive ablation studies, as shown in Table 1, Table 2, respectively.

Table 1. Ablation study of λ_1 selection. 'M' denotes the mIoU (%) of CAM performance.

λ_1	0.05	0.1	0.2	0.5
M	78.9	79.1	78.4	78.2

Table 2. Ablation study of λ_2 selection.

λ_2	0.1	0.2	0.4
M	79.0	79.1	78.4

3. Analysis of Masks

In Table 3, we validate the effectiveness of our designed masks, $\mathbf{M}_d$ and $\mathbf{H}$. Here, $\mathbf{H}$ denotes the updated unreliable positive and negative masks used in $\mathcal{L}_r$. The results show that both masks play crucial roles in their respective losses. Specifically, $\mathbf{M}_d$ successfully selects the dense relations distribution during the optimization of $\mathcal{L}_a$, while $\mathbf{H}$ accurately identifies unreliable positive and negative relations during the optimization of $\mathcal{L}_r$.

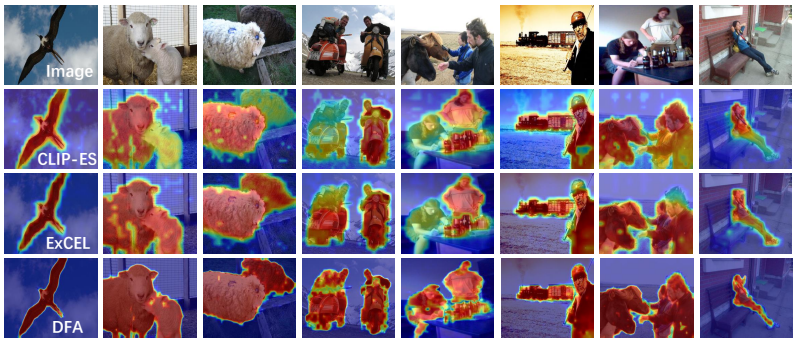

Figure 1. More visualization comparison of CAMs between other modules and our DFA method. The experimental dataset is PASCAL VOC 2012 validation set.

Table 3. Ablation study of our designed masks (*i.e.*, $\mathbf{M}_d$ and $\mathbf{H}$) in DFA losses (*i.e.*, $\mathcal{L}_a$ and $\mathcal{L}_r$) on train set. 'w/ and 'w/o means adding and removing these masks. 'Seg.' denotes the mIoU (%) of semantic segmentation performance.

#	$\mathcal{L}_a$	$\mathcal{L}_r$	M	Seg.
0	w/o $\mathbf{M}_d$	w/o $\mathbf{H}$	77.1	79.5
1	w/ $\mathbf{M}_d$	w/o $\mathbf{H}$	78.1	80.0
2	w/o $\mathbf{M}_d$	w/ $\mathbf{H}$	78.4	80.1
3	w/ $\mathbf{M}_d$	w/ $\mathbf{H}$	79.1	80.6

4. CAM Visualizations

We provide more CAMs visualization comparisons with the other two competitive methods. As illustrated in Figure 1, our DFA produces higher-quality CAMs for both single-object and multiple-object cases, owing to the effectiveness of the proposed low- and high-frequency-aware affinities and the Frequency-Guided (FG) CAM generation.

References

[1] Ziqian Yang, Xinqiao Zhao, Xiaolei Wang, Quan Zhang, and Jimin Xiao. Ffr: Frequency feature rectification for weakly supervised semantic segmentation. In *CVPR*, 2025. 1

[2] Bingfeng Zhang, Siyue Yu, Yunchao Wei, Yao Zhao, and Jimin Xiao. Frozen clip: A strong backbone for weakly supervised semantic segmentation. In *CVPR*, 2024. 1