FFR: Frequency Feature Rectification for Weakly Supervised Semantic Segmentation

Supplementary Material

1. Implementation Details

The learning rate is warmed up to $6e^{-5}$ during the first 2000 iterations, and then decayed with a rate of 0.9 using a polynomial learning rate scheduler for the remaining iterations. Besides, the network is trained for 20,000 iterations with a batch size of 4 on the PASCAL VOC 2012 dataset and for 80,000 iterations with a batch size of 8 on the MS COCO 2014 dataset. We also use the intermediate classifier to leverage the patch token knowledge in the intermediate layer, where high-frequency features remain less attenuated by the multi-head self-attention mechanism. The calculation of its associated loss is the same as multi-label soft margin loss.

2. Multi-class segmentation

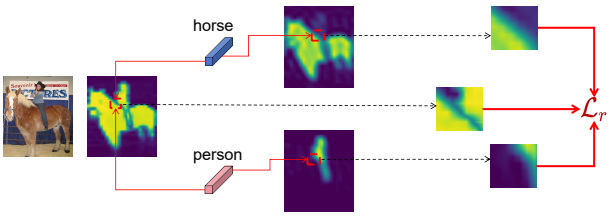

Figure 1. FFR for Multi-class Case.

In this part, we give an additional explanation of how our FFR framework handles multi-class object segmentation. As shown in Figure 1, when one patch contains multiple classes, FFR will generate prototypes for each class. Then, the attenuated high-frequency features will associate with corresponding prototypes and be rectified by their low-frequency features. For example, a patch featuring person and horse will correlate with the prototypes of person and horse respectively, and be rectified by $\mathcal{L}_r$.

3. Motivation for Prototype

FFR employs class-aware prototypes derived from CAMs and image-specific features to compute cosine similarity maps with patch tokens, pinpointing high-frequency patches with significant similarity gaps for attenuated feature rectification. These prototypes simultaneously encode class-discriminative patterns and image-specific feature distributions, enabling precise identification of attenuated class-relevant high-frequency components. Direct similarity computation between frequency-domain features without class guidance risks incorporating noisy background patterns during rectification (*i.e.*, $\mathcal{L}_r$ calculation), which could compromise FFR's performance gain.

	μ	M	Seg.
0	0.1	60.2	57.3
1	0.2	68.2	66.1
2	0.3	73.1	70.9
3	0.4	76.8	74.8

Table 1. Ablation study of μ selection in the Frequency Features Masking Strategy on PASCAL VOC 2012 validation split. 'M' denotes the mIoU (%) of CAM performance and 'Seg.' denotes the mIoU (%) of segmentation performance.

λ_h	1	2	3	4
Seg.	74.2	74.8	74.1	73.5

Table 2. Ablation study of λ_h selection in FFR training objective.

4. Ablation Study

Analysis of μ selection In our FFR framework, μ is used to control the low-frequency regions in the spectrum. The higher the value of ratio μ, the larger the region of the low-frequency part, and vice versa. In Table 1, we analyze the influence of different μ selections on the CAM and Segmentation performance, we select the suitable μ choice in our work based on this experiment. Setting #0 denotes the CAM performance of the original feature map without any reduction in the frequency domain. When the $\mu = 0.4$, we achieve the highest CAM performance 76.8% and segmentation performance 74.8%. If the μ value is less than 0.4, some effective low-frequency features might be masked, which leads to the token rectification of some correct segmented tokens. Once the μ value is larger than 0.4, the noise high-frequency features will not be adequately removed. Therefore, our calculation for μ is optimal, the false segmentations in the boundary regions of segmentation results are obviously reduced and maintain high accuracy when we use $\mu = 0.4$.

Analysis of other hyper-parameters We also conduct ablation studies of other hyper-parameters in the training objective, including the selection of λ_h in Table 2, and the selection of λ_1 in Table 3. After comparing these results, we finally select $\{2, 1\}$ for $\{\lambda_h, \lambda_1\}$.

λ_1	0.5	1	2	4
Seg.	73.9	74.8	74.5	74.0

Table 3. Ablation study of λ_1 selection in FFR training objective.

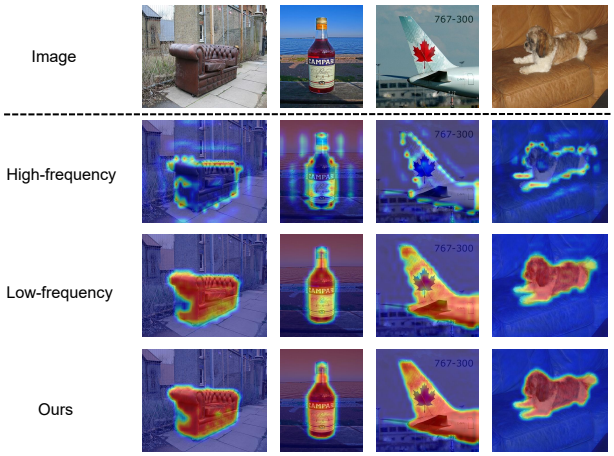

Figure 2. Visualization comparison of Segmentation CAMs between different frequencies in our FFR framework. The experimental dataset is PASCAL VOC 2012 validation split.

5. Visualizations Comparison

Visualizations of high-frequency and low-frequency segmentation CAMs In Figure 2, we visualize the segmentation CAMs from low- and high-frequency in the decoder. As mentioned above, we select the optimal choice $\mu = 0.4$ to decompose the feature map in the frequency domain into the low-frequency and high-frequency feature maps. This allows us to generate the corresponding CAMs in the classifier for a comparison of the activation regions. The visualizations reveal that the segmentation CAMs from high-frequency feature maps tend to incorrectly activate boundary regions and certain parts of the background, which adversely affects the final segmentation accuracy. This observation underscores the attenuation of the high-frequency features we have selected. In contrast, the segmentation CAMs derived from low-frequency feature maps exhibit clear boundary activations. Although some inner regions of the objects remain insufficiently activated, this issue can be addressed through the application of dense energy loss [2] and CRF processing [1]. For instance, the 'bottle' in the high-frequency segmentation CAM has falsely activated in both the boundary and background regions, while its activations in the low-frequency segmentation CAM are inadequate. Within our FFR framework, we rectify these false activations and enhance the segmented boundary regions using reliable low-frequency features, ultimately generating more accurate segmentation CAMs and final segmentation results.

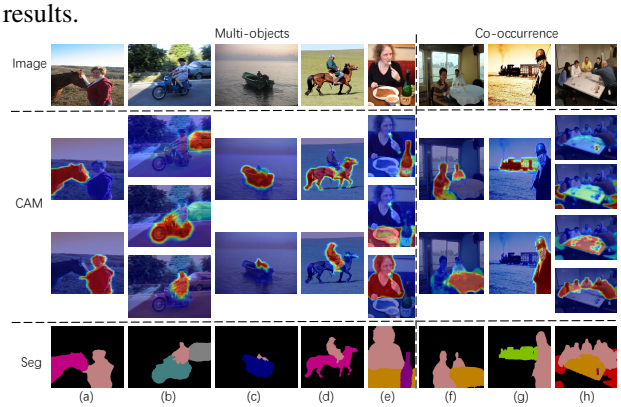

Figure 3. (a)-(e) are visualizations from multi-objects scenarios; (f)-(h) are visualizations from co-occurrence scenarios.

Visualizations of multi-objects co-occurrence scenarios. Moreover, the visualizations of multi-object segmentation are illustrated in Figure 3. All these results verify that our FFR can generate high-quality segmentations with accurate boundaries between multiple objects. Moreover, as can be seen from (f)-(h), our FFR can also effectively solve the co-occurrence issue by rectifying the attenuated high-frequency features.

Visualizations of Motivation. This work is motivated by the observation that attenuated high-frequency features in ViT patches mislead WSSS models, causing inaccurate segmentations along object boundaries. To further validate this motivation, we provide additional experiments and discussions in Figure.4. All this content will be included in our final paper.

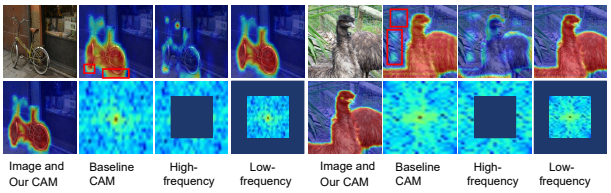

Figure 4. Further experiments of segmentation CAMs.

References

[1] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE TPAMI*, 2017. 2

[2] Bingfeng Zhang, Jimin Xiao, Yunchao Wei, Mingjie Sun, and Kaizhu Huang. Reliability does matter: An end-to-end weakly supervised semantic segmentation approach. In *AAAI*, 2020. 2