Supplementary Materials for Phase Concentration and Shortcut Suppression for Weakly Supervised Semantic Segmentation

Hoyong Kwon, Jaeseok Jeong, Sung-Hoon Yoon, and Kuk-Jin Yoon

KAIST
{kwonhoyong3,jason.jeong,yoon307,kjyoon}@kaist.ac.kr

The supplementary materials contain the following content: Sec. A. Training Details on MS COCO, Sec. B. Discussions regarding FSPM, Sec. C. Importance of Image Pair in MPC, and Sec. D. Visualization Results.

A Training Details: MS COCO

Setting for CAMs Our framework was trained on the MS COCO dataset for 13 epochs using the Adam optimizer with an initial learning rate of 5e-4 and a batch size of 64. Similar to the procedure on the PASCAL VOC 2012 dataset, the same data augmentation techniques as MCTformer [7] were applied, with the image resize scale set to (0.6-1.2). The loss balance hyperparameters λ_1 and λ_2 were both set to 0.4. For the COCO dataset, the classification accuracy in the initial epochs is considerably lower in contrast to the PASCAL dataset. Therefore, the CAM-level losses of MPC and FSS are applied after the accuracy reaches a certain level (after 3 epochs). The Frequency Shortcut Potential Map (FSPM) was updated every three epochs as in the PASCAL dataset.

Setting for Semantic Segmentation For a fair comparison with the prior WSSS works [3–5] and SoTA ViT-based WSSS approach [6], we trained the DeepLab-V2 with the ResNet101 backbone using our high-quality pseudo labels. The training was conducted using an SGD optimizer with a momentum of 0.9, an initial learning rate of 5e-5, and a batch size of 10.

B Discussions regarding FSPM

B.1 Obtaining FSPM

For the Frequency Shortcut Suppression (FSS) in our framework, it is necessary to generate the Frequency Shortcut Potential Map (FSPM) as shown in Fig. A1. The Frequency Influence Measurement (FIM) for creating the FSPM continuously assesses the impact of each frequency component on the classification accuracy by masking them. As for the masking process, since frequency components in similar locations in the frequency domain tend to have similar periods and directions of progression, we measured them by masking on a patch basis. When

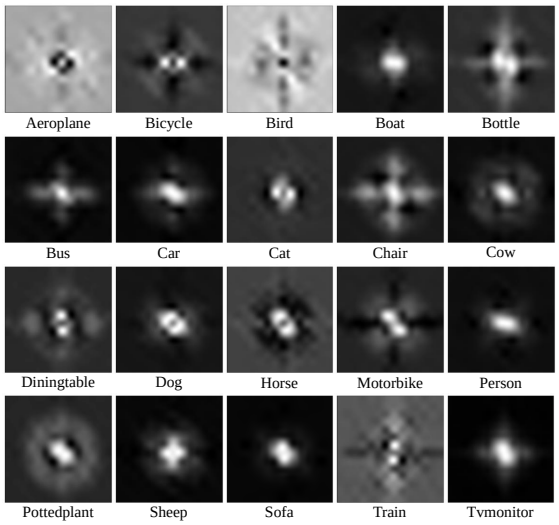

Fig. A1: Visualization of FSPM measured in PASCAL VOC 2012. The frequency components in brighter areas have a higher influence on classification.

Table A1: Ablation study of the FSPM update period. **Bold** number implies the best performance.

Period (epoch)	w/o FSS	1	2	3	4	5	6	7	8
mIoU(%)	67.2	67.7	68.2	**69.5**	68.0	67.3	67.7	67.5	67.5

Discrete Fourier Transform (DFT) is applied to input with real numbers, the output is Hermitian-symmetric, where the negative-frequency components are the complex conjugates of the positive-frequency components at point symmetry positions. To completely mask the frequency component, masking is performed in a point-symmetric fashion to ensure the conjugates are also masked.

Since the FSPM is calculated by masking each patch, which could potentially require considerable resources, a quantitative comparison was conducted to demonstrate consumed resources or times. On the PASCAL VOC 2012 dataset, training of our method takes 140 seconds per epoch, and FIM takes 340 seconds in a single TITAN RTX GPU using a batch size of 64. Considering the FIM is obtained once every three epochs, the process only requires an additional computational time of 0.8 times compared to the training time for each epoch. As the FIM process does not require backward propagation and the need to store gradients, the computation of FIM does not impose much burden on memory. Additionally, as our DFT-based method does not require the saving of weights or parameters, it requires the same number of parameters as our baseline (MCTformer [7]). Lastly, as our MPC module and FSS modules are applied only at training, the running cost is the same as the baseline in the inference phase.

Table A2: Comparison of Magnitude modification methods. $\times N$ means that the magnitude component of the feature is scaled by N.

Method	MPC	(a) Gaussian Noise	(b) $\times$ 0.3	$\times$ 0.5	$\times$ 2.0
mIoU(%)	67.2	62.2	62.7	62.6	59.0

B.2 Update Period of FSPM

An ablation study was conducted to determine the optimal epoch period for updating the FSPM, with the results presented in Tab. A1. It was found that the maximum gain from the FSS occurs when the period is set to 3. When the period is too short, frequent updates to FSPM may prevent the suppression module from optimizing and being fully effective. Likewise, as the period increases, FSS loses its value as the model may attempt to find new shortcuts in a suppressed situation using the updated FSPM. The lack of frequent updates to FSPM can make it challenging to fully disrupt the shortcut learning of the model. Nevertheless, obtaining a gain of 0.8-2.3%p near the optimal period (2-4) supports the effectiveness of the FSS in disrupting the shortcut learning of the model.

C Importance of Image Pair in MPC

To validate the importance of sourcing magnitude information from another image, an ablation study was conducted using an arbitrary source of information composed of random Gaussian values instead of the real image. An image size tensor filled with random Gaussian values was generated and used as input to the ViT with an identical setting in Tab. 1-b of the main paper. This led to a vicious cycle with a tendency for the model to increasingly activate incorrect locations as training progressed. Furthermore, we conducted experiments by replacing the magnitude derived from another image, $|\mathcal{F}(\dot{\mathbf{T}}^{\ell}_{patch})|$, in Eq. 4 of the main paper with Gaussian noise, as shown in Tab. A2-a. The results yielded a 62.2% mIoU, indicating that magnitude modification by Gaussian noise is ineffective. Additionally, instead of using magnitudes derived from another image, we performed magnitude modification using only the magnitude derived from the anchor image, $|\mathcal{F}(\mathbf{T}^{\ell}_{patch})|$, by scaling it by factors of 0.3, 0.5, and 2.0. As illustrated in Tab. A2-b, all cases exhibited poor performance, underscoring the importance of magnitude modification using another image. These outcomes highlight the necessity of providing magnitude with valid statistics from real images.

D Visualization Results

D.1 PASCAL VOC 2012

CAMs In Fig. A2, we present a qualitative comparison of the CAMs $\mathbf{M}_{ref}$ between the baseline and our framework. All samples were selected from the

PASCAL VOC 2012 *train* set. Clear activation around the boundaries of objects in $\mathbf{M}_{ref}$ from our framework implies the effectiveness of the MPC. Particularly, whereas the baseline fails to localize the *Dog* class of the sixth row, ours successfully utilizes the high-level semantic information in the phase to accurately activate along the boundaries of the *Dog*. Furthermore, the effect of the FSS is demonstrated through the reduction of false positives in the *Tv-monitor*, as indicated in the fourth row.

CAMs with FSS To support the effectiveness of FSS, we present a qualitative comparison of the CAMs $\mathbf{M}$ between the baseline and ours with only FSS applied, shown in Fig. A3. The tendency that $\mathbf{M}$ from the baseline exhibits object non-related over-activation throughout the image is observable. Particularly, at *Bird* in the second row, where the classification model has found a frequency shortcut, a strong activation across the entire image is shown. By employing FSS to suppress the frequency shortcut learning, the model focuses on objectness when generating $\mathbf{M}$, resulting in the reduction of false positive activation.

Semantic Segmentation Following the prior WSSS works [6–8], we refined CAMs using PSA [2] to generate high-quality pseudo labels that achieved new state-of-the-art performance. These pseudo labels are used to train a semantic segmentation model, achieving new SoTA mIoU performance on both the *val* and *test* set. Our semantic segmentation results on the PASCAL VOC 2012 *val* set are displayed in Fig. A4. The segmentation performance on the *test* set, validated through the official online server, can be found here.

D.2 MS COCO 2014

CAMs To further validate the superiority of our method, experiments were also conducted on the MS COCO dataset. As the MS COCO dataset contains a large number of classes and complex scenes, obtaining precise CAMs solely with weak signals poses a challenge. Nevertheless, as can be seen in Fig. A5, our method manages to capture the precise boundaries of objects and effectively separates objects from each other as in the fourth and fifth rows.

Semantic Segmentation We applied the IRN [1] to our CAMs to generate pseudo labels and used them to train a semantic segmentation model. We achieved new state-of-the-art performance on the MS COCO *val* set, and the qualitative results are shown in Fig. A6.

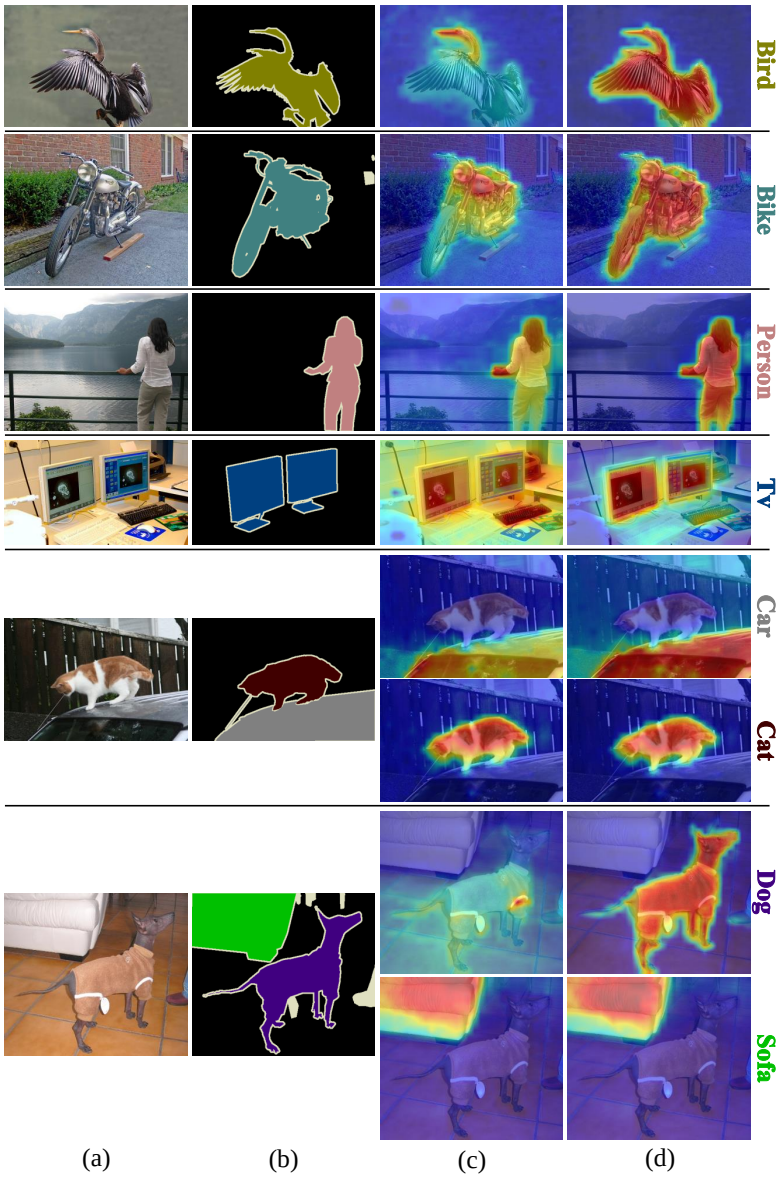

Fig. A2: Comparison of CAMs ($\mathbf{M}_{ref}$) on PASCAL VOC 2012 *train* set. (a) Image, (b) Ground Truth, (c) $\mathbf{M}_{ref}$ of baseline, (d) $\mathbf{M}_{ref}$ of ours.

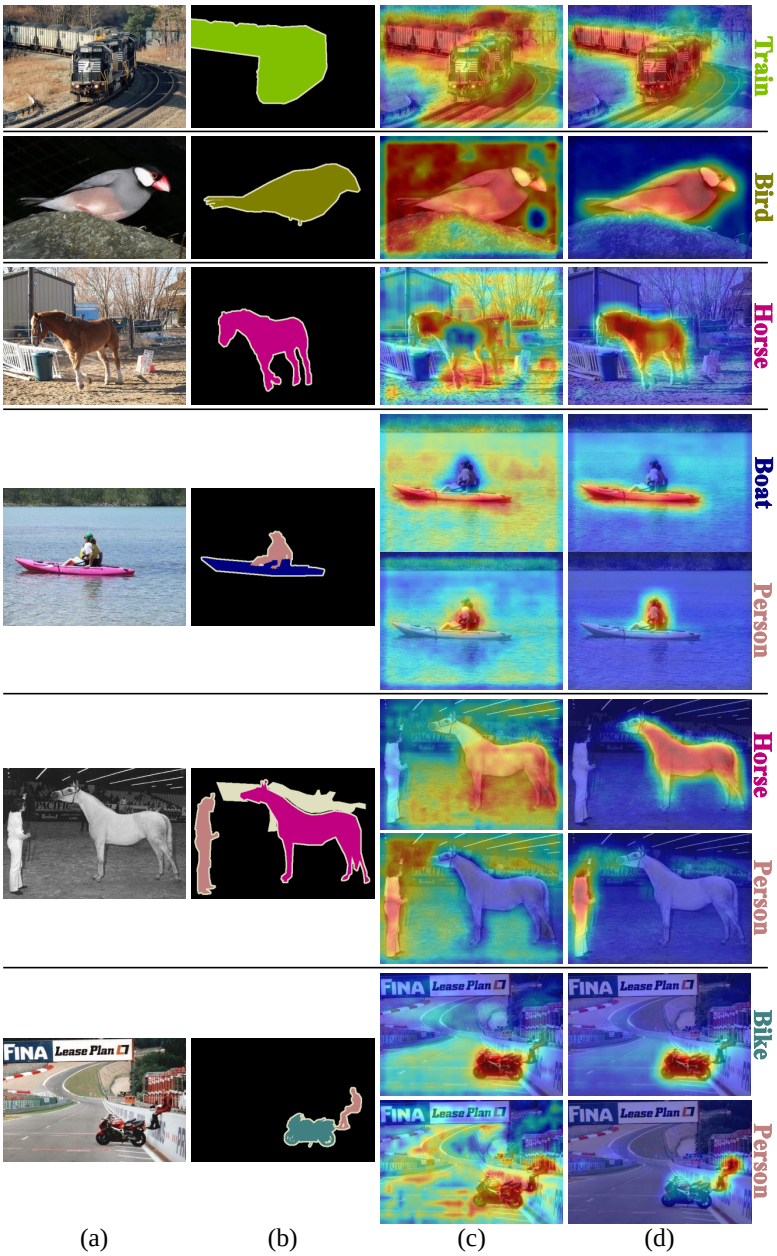

Fig. A3: Comparsion of CAMs (**M**) on PASCAL VOC 2012 *train* set. (a) Image, (b) Ground Truth, (c) **M** of baseline, (d) **M** of ours only with FSS.

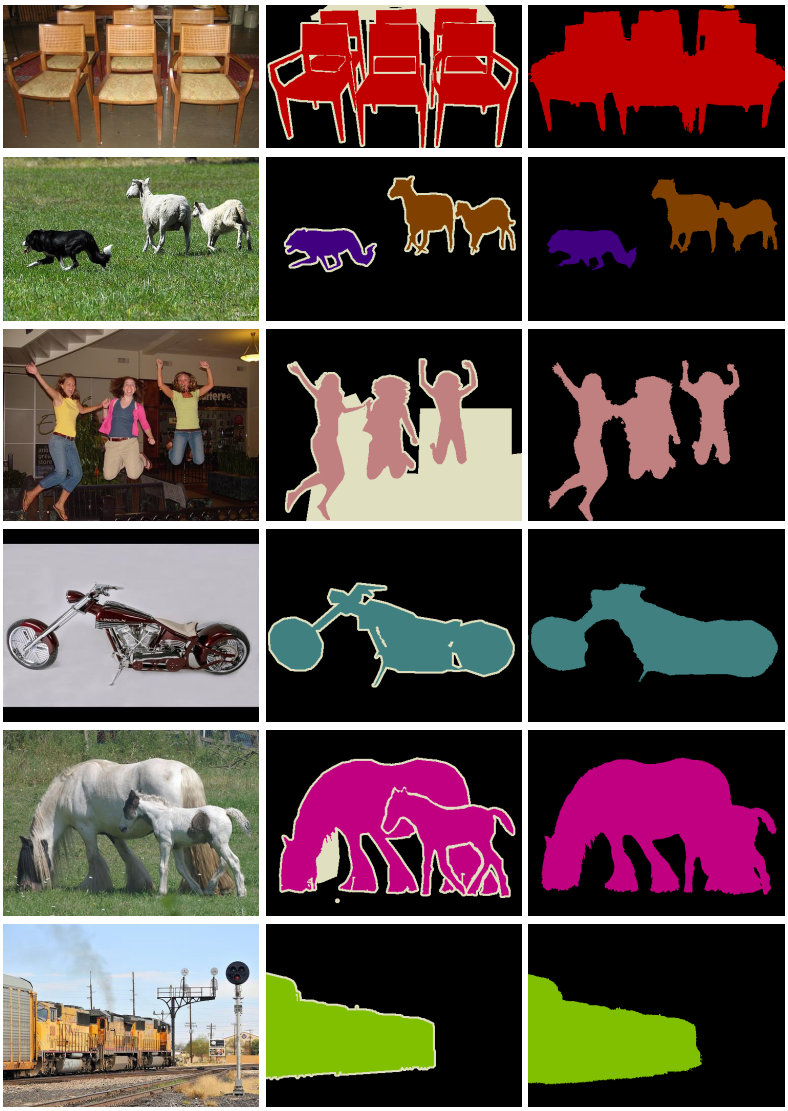

Fig. A4: Visualization of semantic segmentation results on PASCAL VOC 2012 *val* set. (a) Image, (b) Ground Truth, (c) Ours.

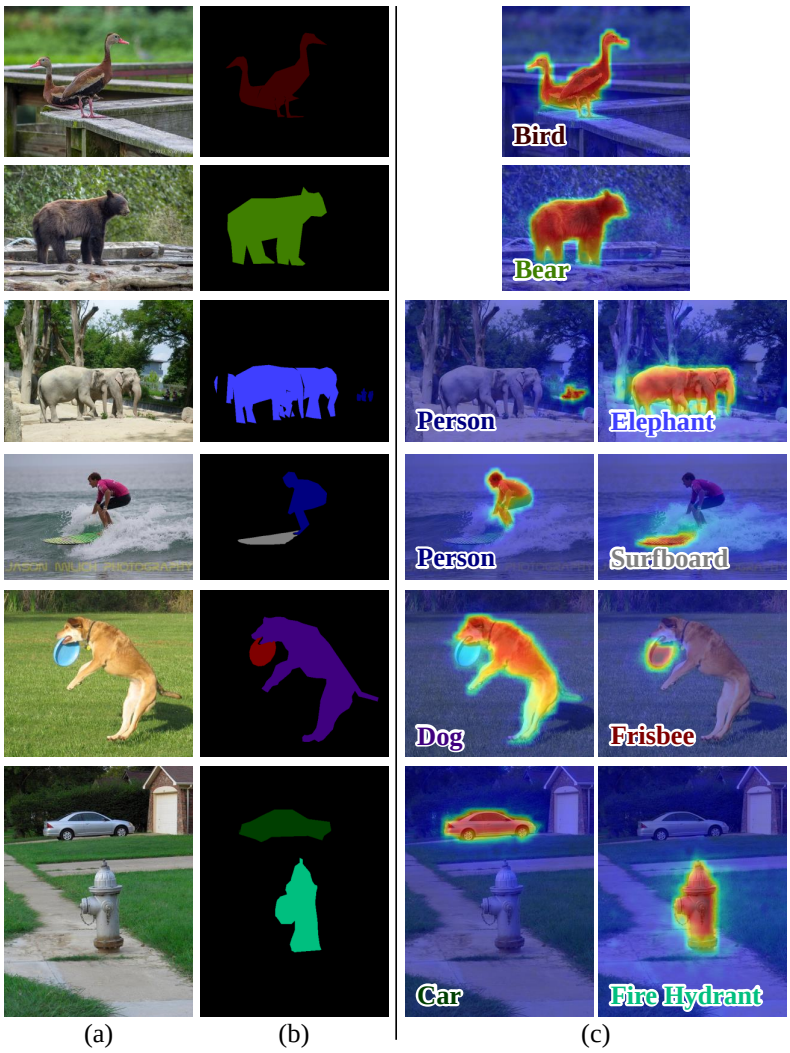

Fig. A5: Visualization of CAMs on MS COCO 2014 *train* set. (a) Image, (b) Ground Truth, (c) Our CAMs.

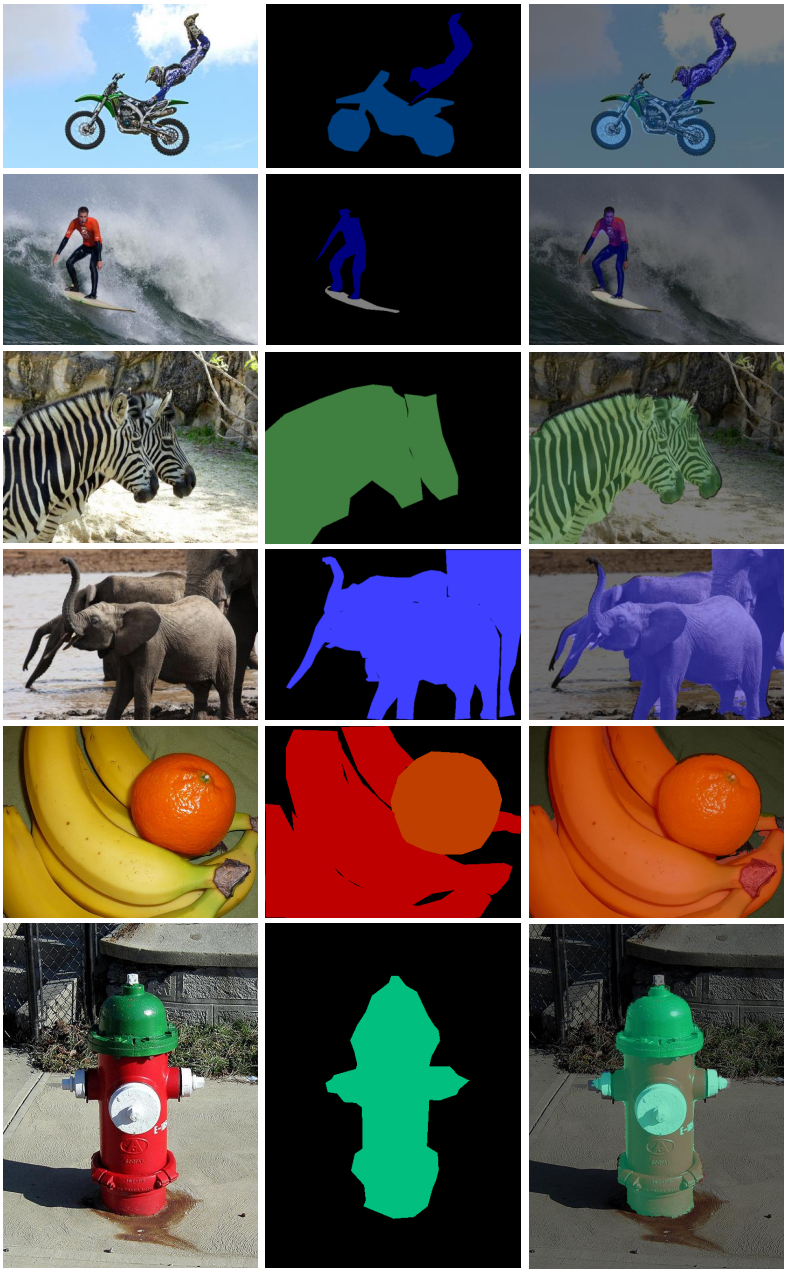

Fig. A6: Visualization of semantic segmentation results on MS COCO 2014 *val* set. (a) Image, (b) Ground Truth, (c) Ours with image.

References

1. Ahn, J., Cho, S., Kwak, S.: Weakly supervised learning of instance segmentation with inter-pixel relations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2209–2218 (2019)
2. Ahn, J., Kwak, S.: Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4981–4990 (2018)
3. Chen, L., Lei, C., Li, R., Li, S., Zhang, Z., Zhang, L.: Fpr: False positive rectification for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1108–1118 (2023)
4. Chen, Z., Wang, T., Wu, X., Hua, X.S., Zhang, H., Sun, Q.: Class re-activation maps for weakly-supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 969–978 (2022)
5. Lee, J., Choi, J., Mok, J., Yoon, S.: Reducing information bottleneck for weakly supervised semantic segmentation. Advances in Neural Information Processing Systems **34**, 27408–27421 (2021)
6. Peng, Z., Wang, G., Xie, L., Jiang, D., Shen, W., Tian, Q.: Usage: A unified seed area generation paradigm for weakly supervised semantic segmentation. arXiv preprint arXiv:2303.07806 (2023)
7. Xu, L., Ouyang, W., Bennamoun, M., Boussaid, F., Xu, D.: Multi-class token transformer for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4310–4319 (2022)
8. Yoon, S.H., Kweon, H., Cho, J., Kim, S., Yoon, K.J.: Adversarial erasing framework via triplet with gated pyramid pooling layer for weakly supervised semantic segmentation. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIX. pp. 326–344. Springer Nature Switzerland Cham (2022)