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

Abstract. Weakly Supervised Semantic Segmentation (WSSS) with image-level supervision typically acquires object localization information from Class Activation Maps (CAMs). While Vision Transformers (ViTs) in WSSS have been increasingly explored for their superior performance in understanding global context, CAMs from ViT still show imprecise localization in boundary areas and false positive activation. This paper proposes a novel WSSS framework that targets these issues based on the information from the frequency domain. In our framework, we introduce the Magnitude-mixing-based Phase Concentration (MPC) module, which guides the classifier to prioritize phase information containing high-level semantic details. By perturbing and mixing the magnitude, MPC guides the classifier to accentuate and concentrate on the shape information in the phase, thereby leading to finer distinctions in CAMs boundary regions. Additionally, inspired by empirical observations that the classification "shortcut" in the frequency domain can induce false positives in CAMs, we introduce a Frequency Shortcut Suppression (FSS) module. This module aims to discourage the formation of such shortcuts, thereby mitigating false positives. The effectiveness of our approach is demonstrated by achieving new state-of-the-art performance on both PASCAL VOC 2012 and MS COCO 2014 datasets. The code is available at `https://github.com/kwonhoyong3/PCSS-WSSS`.

Keywords: Weakly Supervised Semantic Segmentation · Fourier Transform · Shortcut Learning

1 Introduction

Weakly Supervised Semantic Segmentation (WSSS) aims to enhance its utility by reducing the dependency on human-annotated pixel-level labels in semantic segmentation. With this goal, WSSS mainly explores methods to generate accurate pseudo-labels using only easily obtainable weak labels such as image-level labels [1–3,25,26,29,51,60,68,69], scribbles [37,56], and bounding boxes [8,30,43]. Among these methods, leveraging image-level labels has been extensively investigated, primarily due to the availability of large existing datasets.

Since image-level labels do not provide spatial information of objects, image-level based WSSS typically involves multiple stages: 1) extract Class Activation Maps (CAMs) to localize the objects, 2) refine CAMs to generate pseudo

Fig. 1: Experimental results verifying the relationship between CAMs and Frequency Shortcut. The Filtered Image (c) was obtained by removing the Principal Frequency components (b) related to the class of the Image (a). (d) and (e) are CAMs of images (a) and (c), respectively, derived from the same ViT trained on clean images.

ground truth (Pseudo-GT), and 3) train a semantic segmentation model using the Pseudo-GT. As the second and third stages are dependent on the CAMs from the first stage, our work targets to improve the quality of CAMs. In the first stage, unlike Convolutional Neural Networks (CNNs) that focus on discriminative regions, Vision Transformer (ViT) [9] stands out in WSSS for its ability to understand global context through the Multi-Head Self-Attention (MHSA) mechanism [55].

Despite these advantages over CNN-based CAMs, CAMs from ViT still face several issues; 1) CAMs exhibit imprecise object boundaries due to lack of spatial supervision, and 2) CAMs often localize object-unrelated (false positive) regions. In this paper, we propose two methodologies to address the issues with CAMs.

When humans and ViTs visually assess an object, semantics is predominantly influenced by shape over texture [4, 24]. While shape and texture information are intertwined in images, applying Discrete Fourier Transform (DFT) [54] can help extract the two information into magnitude and phase spectral components. Inspired by the fact that the phase spectrum encapsulates shape and high-level semantic information [16, 42], we propose Magnitude-mixing-based Phase Concentration (MPC) to encourage models to focus more on the phase and improve object boundaries in generated CAMs. In MPC, we select a random pair of images consisting of an anchor image and another arbitrary image. Within the intermediate layers of a ViT, we merge the *magnitude* information from the arbitrary image into the magnitude from the anchor image. This results in the creation of two anchor image features with identical phases but different magnitudes. By ensuring that both features are guided by the same weak supervision and that the generated CAMs are consistent regardless of changes in magnitude, we induce the model to focus on the phase. Consequently, MPC enables the model to concentrate on the crucial information within the phase, leading to CAMs with more precise boundaries.

While MPC facilitates precise activation near the boundaries, the issue of activation in object-unrelated regions persists. Interestingly, certain frequency components have been shown to significantly impact classification during the training process of ViT, acting as Frequency Shortcut [58]. A frequency shortcut is a simplicity bias in classification networks, where the network relies on a

specific frequency rather than semantic content for classification. Additionally, we empirically observed that models tend to exhibit false positives in the CAMs of classes attempting to find shortcuts in the frequency domain. To explore the effects of principal frequency and shortcuts in CAMs, we compare the CAMs from a clean image and a corresponding image with the principal frequency removed (*i.e.*, filtered image), as shown in Fig. 1. While the CAMs from clean image (Fig. 1(d)) suffers from false activations, false activations are relieved in the CAMs from filtered image (Fig. 1(e)). Here, we can verify that certain principal frequencies influence the entire image and cause widespread false activation regardless of the semantics; these can be regarded as shortcut frequencies.

To resolve the false positives induced by frequency shortcuts, we propose the Frequency Shortcut Suppression (FSS). Initially, we measured the impact of each frequency on the classification of each class to aggregate a Frequency Shortcut Potential Map (FSPM). Based on FSPM, we removed the frequency shortcut components from the image to generate the filtered image. Through classification on the filtered image and by comparing CAMS generated from the original image and the filtered image, FSS reduces the reliance of the generated CAMs to false positive inducing shortcut frequency components. This removal of shortcut frequency components happens iteratively, with the FSPM constantly adjusted.

The main contributions of this paper are summarized as follows:

- Through Magnitude-mixing-based Phase Concentration (MPC), we guide ViT to focus on the meaningful phase information, resulting in precise activation around boundaries.
- We proposed a Frequency Shortcut Suppression (FSS) module that effectively reduces false positive activations in CAMs by suppressing shortcut learning in the Frequency Domain.

To demonstrate the effectiveness of our methodology, we conducted comparisons with state-of-the-art WSSS methods on the PASCAL VOC 2012 [11] and MS COCO 2014 datasets [38], achieving new SoTA performance.

2 Related Works

2.1 Weakly Supervised Semantic Segmentation

Most existing WSSS approaches utilize Class Activation Maps (CAMs) [72] obtained from Convolutional Neural Networks (CNNs) to localize objects when only image-level labels are provided. However, raw CAMs (*i.e.* seeds) fail to localize the whole object with imprecise boundaries. In pursuit of generating more accurate pixel-level pseudo-labels, research in WSSS aims to propose a method to improve the quality of raw CAMs or further refine/post-processing the CAMs.
CAMs Quality Improvement To guide CAMs to localize the less discriminative regions, various methods are proposed including online CAMs aggregation [21] , sub-categories exploration [3], information bottleneck reduction [27]

and mining cross-image relations [12, 34, 52]. Also, [22, 59, 70] introduce various auxiliary tasks to regularize CAMs training. Along with these methods, Adversarial Erasing (AE) [25,33,53,67,71] methods, which are based on the erase-and-find mechanism, are also actively researched to expand CAMs. Apart from techniques aimed at expanding CAMs, contrastive learning-based WSSS [6, 62, 73] and adversarial learning between classifier-reconstructor [26] are proposed to produce CAMs with precise object boundaries. Though CNN-based WSSS methods show promising performance, with the success of Vision Transformer (ViT) in various vision tasks, several methods [5, 13, 39, 44, 48, 49, 63, 64, 69] are proposed to migrate the ViT for WSSS. MCTformer [63] introduces multi-class token to extract class-specific attention map from ViT by improving the TS-CAM [13], which is designed for weakly-supervised object localization. In the same line, ViT-PCM [47] proposed a method to extract CAMs based on patch-class mapping. Along with these methods, large language model-based WSSS [39, 61, 64] and token-level contrasting [49] are also introduced. Unlike the prior methods, our work is the first to utilize frequency domain information to address the issues with CAMs in ViT-based WSSS.

CAMs Post-processing To obtain high-quality pseudo labels for semantic segmentation, various approaches are introduced in WSSS based on affinity learning [1, 2], removing intra-object edges in contours [32], iterative refinements [35, 36]. Besides these methods, Image-matting based refinements [57] and utilizing unsupervised semantic segmentation model [23] greatly boost the quality of pseudo labels. As the post-processing methods can be synergistically applied with methods that generate precise initial seeds (CAMs), this paper focuses on improving the quality of the CAMs.

2.2 Fourier Transform in Computer Vision

Transforming images from the spatial domain to the frequency domain via the Fourier Transform enables the separation of information into phase and amplitude components. Early studies [16, 42] have shown that while amplitude carries low-level statistics and texture information, phase encapsulates high-level semantic and shape information. Leveraging this advantage, Fourier Transform has been applied to a variety of vision tasks such as Exposure Correction [20], Semantic Segmentation [18], and Learning with noisy label [17]. Along with the successful application of frequency components, the fields of Domain Adaptation [19, 66] and Generalization [15, 31, 65] have emphasized the importance of phase for its high-level semantic and structural information. FDA [66] performed Domain Adaptation by transforming low-level amplitude in the Fourier Domain, maintaining high-level semantics and structure, building on the understanding that low-level amplitude contains image style information. [65] highlighted the significance of the phase component, which carries semantic information, and proposed Fourier-based data augmentation through Image-level amplitude to perform Domain Generalization with an emphasis on phase information.

These studies underscore the significant role of the frequency domain in neural networks. Intriguingly, classification networks engage in shortcut learning

in the frequency domain [58], identifying specific frequencies for classification. Since shortcut learning [14] can disrupt semantic capture, research efforts have been directed towards reducing such effects [40, 46]. SGT [41] analyzed how vanilla ViT activates areas unrelated to the object due to shortcut learning in the background regions during the classification process and proposed rectifying this through saliency maps.

3 Methods

3.1 Preliminary

Vision Transformer for CAMs To propagate the image $\mathbf{I}$ into Vision Transformer (ViT), it is split into $N \times N$ patches and forms the patch tokens $\mathbf{T}^0_{patch} \in \mathbb{R}^{D \times N^2}$, where D is the embedding dimension. Along with C class tokens $\mathbf{T}^0_{cls} \in \mathbb{R}^{D \times C}$, the patch tokens $\mathbf{T}^0_{patch}$ pass through the Vision Transformer (ViT). Consequently, the final class tokens $\mathbf{T}^L_{cls} \in \mathbb{R}^{D \times C}$ and patch tokens $\mathbf{T}^L_{patch} \in \mathbb{R}^{D \times N^2}$ are derived as outputs, where L represents the number of ViT layers. This procedure can be formally defined as follows:

$$\mathbf{T}^i_{cls}, \mathbf{T}^i_{patch} = \mathbf{ViT}_i(\mathbf{T}^{i-1}_{cls}, \mathbf{T}^{i-1}_{patch}), 1 \le i \le L \tag{1}$$

where $\mathbf{ViT}_i$ is i^{th} layer of ViT. The class tokens $\mathbf{T}^L_{cls}$ are processed through average pooling to derive the classification logits $y_c \in \mathbb{R}^C$. Meanwhile, the patch tokens $\mathbf{T}^L_{patch}$, arranged according to their spatial positions, pass through a convolution layer to produce the Class Activation Maps (CAMs) $\mathbf{M} \in \mathbb{R}^{C \times N \times N}$. By applying Global Average Pooling (GAP) to $\mathbf{M}$, we can obtain classification logits $y_p \in \mathbb{R}^C$ in a patch-level. These two classification logits are supervised using a multi-label soft margin loss $\mathcal{L}_{cls}$ with the given classification labels $\mathbf{y}$. Additionally, following [63], we extract the token-to-token attention $\mathbf{A}_{t2t} \in \mathbb{R}^{(C+N^2) \times (C+N^2)}$ based on multi-head self-attention. From this token-to-token attention $\mathbf{A}_{t2t}$, the class-to-patch attention $\mathbf{A}_{c2p} \in \mathbb{R}^{C \times N^2}$ and patch-to-patch attention $\mathbf{A}_{p2p} \in \mathbb{R}^{N^2 \times N^2}$ can be extracted. These two types of attention maps are used to refine the CAMs, forming a refined CAMs $\mathbf{M}_{ref} \in \mathbb{R}^{C \times N \times N}$. The aforementioned process can be noted as follows:

$$\mathbf{M}_{ref} = (\mathbf{M} \odot \mathbf{A}_{c2p}) \times \mathbf{A}_{p2p} \tag{2}$$

where $\odot$ and $\times$ are Hadamard product and matrix multiplication, respectively.

Discrete Fourier Transform Let $\mathcal{F}$ denote the Discrete Fourier Transform (DFT) [54] applied across the spatial dimensions. For a feature map $\mathbf{X} \in \mathbb{R}^{D \times H \times W}$, the result of applying $\mathcal{F}$ to a single channel can be represented as follows:

$$\mathcal{F}(\mathbf{X})(m, n) = \sum_{h,w} \mathbf{X}(h, w) e^{-j2\pi(\frac{h}{H}m + \frac{w}{W}n)}, \; j^2 = -1 \tag{3}$$

Since the absolute value and the angle component of $\mathcal{F}(\mathbf{X})$ respectively represent magnitude and phase components, we denote the magnitude spectrum as $|\mathcal{F}(\mathbf{X})|$ and the phase spectrum as $\angle\mathcal{F}(\mathbf{X})$ throughout the paper.

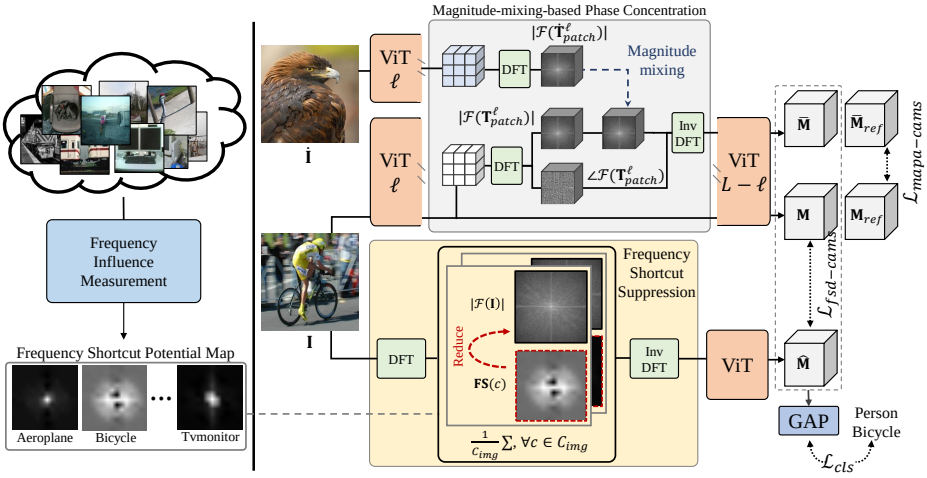

Fig. 2: Visualization of the proposed framework. The Magnitude-mixing-based Phase Concentration (MPC) is executed at an intermediate layer. Here, the *magnitude* of features from the anchor image $\mathbf{I}$ and another arbitrary image $\dot{\mathbf{I}}$ are mixed to make a feature with identical *phase*, but different *magnitude*. The model is guided to perform correct classification and derive consistent CAMs $\mathbf{M}_{ref}$ from a pair of features. In the Frequency Shortcut Suppression (FSS), based on the Frequency Shortcut Potential Map (FSPM), frequency components that have a significant impact on the classification of classes in $\mathbf{I}$ are reduced before being inputted to the ViT. The model is encouraged to classify correctly and minimize the difference between $\mathbf{M}$ and $\hat{\mathbf{M}}$.

3.2 Magnitude-mixing-based Phase Concentration

With the aid of multi-head self-attention, CAMs from ViT are more proficient in capturing the less-discriminative regions than CNN-based methods. Nonetheless, they still show imprecise localization along object boundaries due to a lack of spatial constraint. To improve localization of CAMs, we design the Magnitude-mixing-based Phase Concentration (MPC) module to extract high-level semantic and boundary information shown to be present in the phase component of the frequency domain [16, 42].

We consider two ideas in MPC to guide the model to prioritize the phase component: 1) The model should correctly classify the image even when the magnitude component is altered. Similarly, 2) the CAMs generated after the magnitude change should be similar to those obtained before the change. The key to these two ideas is to change the magnitude while keeping the phase component intact. Since the magnitude needs to be 'feasible' yet different, we designed a magnitude-mixing method. Here, we form a random image pair consisting of an anchor image $\mathbf{I}$ and an arbitrary image $\dot{\mathbf{I}}$ for perturbation. Let the input class and patch tokens of $\mathbf{I}$ be denoted as $[\mathbf{T}^0_{cls}, \mathbf{T}^0_{patch}]$, and those of $\dot{\mathbf{I}}$ as $[\dot{\mathbf{T}}^0_{cls}, \dot{\mathbf{T}}^0_{patch}]$. We then process each set of tokens through the ViT layers, as outlined in Eq. 1, for $1 \le i \le \ell$, to obtain $[\mathbf{T}^\ell_{cls}, \mathbf{T}^\ell_{patch}]$ and $[\dot{\mathbf{T}}^\ell_{cls}, \dot{\mathbf{T}}^\ell_{patch}]$ respectively. After reshaping

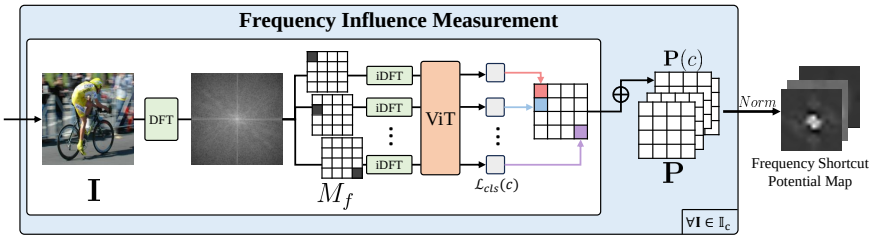

Fig. 3: Visualization of Frequency Influence Measurement. For an image $\mathbf{I}$ containing class c, a set of images that each masked at different frequency components using M_f in the frequency domain of $\mathbf{I}$ are inputted into ViT. The classification loss for class c is calculated as the influence of frequency f on class c. Accumulate the process across the image set $\mathbb{I}_c$ to produce $\mathbf{P}(c)$, and repeat for each c. Finally, $\mathbf{P}$ is normalized to obtain the Frequency Shortcut Potential Map (FSPM) for each class.

and aligning $\mathbf{T}^{\ell}_{patch}$ and $\dot{\mathbf{T}}^{\ell}_{patch}$ according to their spatial positions, the spectral magnitude is obtained and perturbed using the following formula:

$$\bar{\mathbf{T}}^{\ell}_{patch} = \mathcal{F}^{-1}(\langle\!\langle r|\mathcal{F}(\mathbf{T}^{\ell}_{patch})| + (1-r)|\mathcal{F}(\dot{\mathbf{T}}^{\ell}_{patch})|,\ \angle\mathcal{F}(\mathbf{T}^{\ell}_{patch})\rangle\!\rangle) \tag{4}$$

where r is the perturbation ratio between the anchor image and arbitrary image, and $\mathcal{F}^{-1}$ is the inverse DFT. Tokens and maps that have been affected by magnitude-mixing are denoted by a bar on top (*i.e.* $\bar{\mathbf{T}}^{\ell}_{patch}$). $\langle\!\langle A, B\rangle\!\rangle$ is coupling of magnitude A and phase B. As shown in Eq. 4, the phase component of $\bar{\mathbf{T}}^{\ell}_{patch}$ remains unchanged as the magnitude is perturbed. After iDFT, the set of tokens $[\mathbf{T}^{\ell}_{cls}, \bar{\mathbf{T}}^{\ell}_{patch}]$ are reshaped and passed through the remaining layers of ViT, following Eq. 1 from $\ell < i \leq L$, to produce the output tokens $[\bar{\mathbf{T}}^{L}_{cls}, \bar{\mathbf{T}}^{L}_{patch}]$. The class prediction $\bar{y}_c$ is obtained from $\bar{\mathbf{T}}^{L}_{cls}$, while the class prediction $\bar{y}_p$ and refined CAMs $\bar{\mathbf{M}}_{ref}$ are generated from $\bar{\mathbf{T}}^{L}_{patch}$. $\bar{y}_c$ and $\bar{y}_p$ are used as class predictions in multi-label soft margin loss to obtain $\mathcal{L}_{mpc\text{-}cls}$. Following the second idea, the difference between the refined CAMs $\mathbf{M}_{ref}$ and $\bar{\mathbf{M}}_{ref}$ are minimized using the following loss:

$$\mathcal{L}_{mpc\text{-}cams} = |\mathbf{M}_{ref} - \bar{\mathbf{M}}_{ref}|_1 \tag{5}$$

Ultimately, the MPC is supervised through the following loss:

$$\mathcal{L}_{mpc} = \mathcal{L}_{mpc\text{-}cls} + \lambda_1 \mathcal{L}_{mpc\text{-}cams} \tag{6}$$

where λ_1 is a hyperparameter to balance the losses. Additionally, the gradient to $\mathbf{M}_{ref}$ from $\mathcal{L}_{mpc\text{-}cams}$ is detached to guide $\bar{\mathbf{M}}_{ref}$ to follow $\mathbf{M}_{ref}$. This increases the dependency of the model on the phase information, enabling the model to better understand the boundary information in the phase component.

3.3 Frequency Shortcut Suppression

Through the MPC module, we have addressed the issues of activation around the boundary in CAMs. However, the problem of false activations in CAMs

persists. Various factors contribute to false activations in ViT, such as GAP for generating classification logits from CAMs and the over-smoothing problem caused by self-attention of ViT [49]. Interestingly, we experimentally found that false activations can also arise due to Shortcut Learning in the frequency domain [14], which we aim to address. Shortcut Learning during the classification training process in ViT can negatively impact genuine semantic understanding [40], and such shortcuts can lead to wrong activations [41]. Furthermore, it has been revealed that Neural Networks can create shortcuts in the frequency domain [58]. We empirically found a correlation between frequency shortcuts and false positives in CAMs $\mathbf{M}$. Since Frequency Shortcuts are unrelated to objectness, activations caused by frequency shortcuts occur in locations unrelated to objects. We have resolved over-activation by disrupting the creation of frequency shortcuts through the Frequency Shortcut Suppression(FSS) module.

Frequency Shortcut Potential Map To implement the FSS, the potential of specific frequency components to be used as shortcuts in classification needs to be measured. Inspired by Wang *et al*. [58], we utilized a Frequency Influence Measurement (FIM) framework, as depicted in Fig. 3, to assess the impact of each component. If the classification loss of the model increases significantly for an image $\mathbf{I}$ of class c when information about a certain frequency component f is removed, it indicates that f has an important role in classifying class c for the model. Therefore, for each class c, the influence of f on the classification of the image set $\mathbb{I}_c$ corresponding to that class is calculated by accumulating classification loss. To remove a specific frequency component f from an image, a Masking map $M_f \in \mathbb{R}^{H\times W}$ is used, where the value for f is set to 0, and others are set to 1. For an image $\mathbf{I}$, let $\mathbf{I}_f = \mathcal{F}^{-1}(\mathcal{F}(\mathbf{I}) \odot M_f)$ represent the image with the f component masked where $\odot$ is element-wise multiplication. The classification logit derived after passing $\mathbf{I}_f$ through a ViT is denoted by $y_{\mathbf{I}_f}$. When $\mathbf{P} \in \mathbb{R}^{C\times H\times W}$ is frequency influence accumulated map, the influence of the frequency component f on the classification of class c is calculated as follows:

$$\mathbf{P}(c, f) = \sum_{\mathbf{I}\in\mathbb{I}_c} \mathcal{L}_{bce}(y_{\mathbf{I}_f}(c), 1) \tag{7}$$

where $y(c)$ and $\mathcal{L}_{bce}$ are probability of class c and binary cross-entropy loss, respectively. $\mathbf{P}(c) \in \mathbb{R}^{H\times W}$ ultimately holds the classification influence information of each frequency for class c, with higher values indicating a significant impact on classifying class c. $\mathbf{P}$ is normalized for each class to create the Frequency Shortcut Potential Map (FSPM) $\mathbf{FS} \in \mathbb{R}^{C\times H\times W}$:

$$\mathbf{FS}(c) = Norm(\mathbf{P}(c)),\ c \in \{1, 2, ..., C\} \tag{8}$$

where $Norm$ is the normalization function that scales each value to fall between 0 and 1.

Frequency Shortcut Suppression We propose the Frequency Shortcut Suppression (FSS) utilizing the FSPM generated through the FIM process. The FSPM encapsulates information on which frequency components significantly impact classification at the time of FIM execution, indicating frequencies that

the model is likely to use as shortcuts. The image without shortcut frequency can be generated by reducing the magnitude of frequency components based on the FSPM. Subsequently, the model can be prevented from forming shortcuts by encouraging it to classify correctly and generate consistent CAMs from the original and filtered images.

Let an image $\mathbf{I}$ contains classes $c \in \mathbb{C}_\mathbf{I}$. The filtered image $\hat{\mathbf{I}}$ is synthesized by decreasing the frequency components from $\mathbf{I}$ based on the FSPM, making it challenging for the model to classify through shortcuts:

$$\hat{\mathbf{I}} = \frac{1}{N} \sum_{c \in \mathbb{C}_\mathbf{I}} \mathcal{F}^{-1}(\mathcal{F}(\mathbf{I}) \odot (1 - \mathbf{FS}(c))) \tag{9}$$

where N is the number of classes in $\mathbb{C}_\mathbf{I}$, and $\odot$ denotes the Hadamard product. The predictions $\hat{y}_c$ and $\hat{y}_p$ derived from $\hat{\mathbf{I}}$ via the ViT are trained through a multi-label soft margin loss $\mathcal{L}_{fss\text{-}cls}$. Additionally, the CAMs $\hat{\mathbf{M}}$ generated from $\hat{\mathbf{I}}$ and the CAMs $\mathbf{M}$ derived from the original image $\mathbf{I}$ are supervised to minimize their difference:

$$\mathcal{L}_{fss\text{-}cams} = |\mathbf{M} - \hat{\mathbf{M}}|_1 \tag{10}$$

Assuming the model has learned to use a frequency component f as a shortcut for classifying class c, the model cannot use this shortcut f for classification when f is reduced in $\hat{\mathbf{I}}$. In the absence of the frequency shortcut f, the model performs classification based on the semantic understanding of class c, resulting in activation within areas related to the object. Simultaneously, by reducing the difference between the two CAMs $\mathbf{M}$ and $\hat{\mathbf{M}}$, false positives caused by shortcuts can be resolved.

The FSS is ultimately supervised through the following loss:

$$\mathcal{L}_{fss} = \mathcal{L}_{fss\text{-}cls} + \lambda_2 \mathcal{L}_{fss\text{-}cams} \tag{11}$$

where λ_2 is a hyperparameter to balance between the losses.

The final loss of the proposed overall framework is as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{cls} + \mathcal{L}_{mpc} + \mathcal{L}_{fss} \tag{12}$$

4 Experiments

4.1 Experimental Settings

Datasets Our proposed method was evaluated using the PASCAL VOC 2012 and MS-COCO 2014 datasets, widely utilized for Weakly Supervised Semantic Segmentation (WSSS). The PASCAL VOC 2012 includes 20 classes of foreground objects along with a background class, and consists of three subsets (*train*, *val*, *test*) containing 10582, 1449, and 1456 images, respectively. Meanwhile, the MS-COCO 2014 is made up of 80 foreground object classes and a background class and is composed of two subsets (82k *train* set and 40k *val* set).

Table 1: Ablation study of components on the PASCAL VOC 2012 *train* set. P: Precision, R: Recall. Values from **M** are measured under similar R for a fair comparison of P. The best result is represented in **Bold**.

				(x) **M**			(y) $\mathbf{M}_{ref}$		
	$\mathcal{L}_{cls}$	$\mathcal{L}_{fss}$	$\mathcal{L}_{mpc}$	P(%)	R(%)	mIoU(%)	P(%)	R(%)	mIoU(%)
Baseline	✓			53.2	79.6	45.3	74.5	80.9	63.2
(a)	✓	✓		71.3	79.8	59.4	76.5	82.1	65.4
(b)	✓		✓	67.1	80.6	55.9	78.8	82.2	67.2
(c)	✓	✓	✓	76.2	80.7	63.9	79.5	84.8	**69.5**

Evaluation Metric Following prior research [44,63], the mean Intersection over Union (mIoU) metric is used to assess the quality of generated CAMs and the performance of semantic segmentation models trained on CAMs-based pseudo-labels. To measure the quality of CAMs, the mIoU with the most optimal threshold is employed by default. Typically, CAMs are evaluated on the *train* set, whereas the performance of the semantic segmentation model is assessed on the *val* set. While ground-truth pixel-level labels are accessible for both datasets on the *val* set and are analyzed locally, the evaluation for the PASCAL VOC 2012 *test* set is carried out via the official website.

Implementation Details Our framework utilizes DeiT-S pre-trained on ImageNet [50] as the classification backbone for a fair comparison with ViT-based WSSS methods [44,63]. The classification network is trained for 60 epochs with a batch size of 64 and an initial learning rate of 5e-4, using the Adam optimizer. We adopted the same data augmentation technique as in MCTformer [63], but with different image resize scales. We crop images to a size of 224×224 for equal comparison. In the MPC module, $r = 0.3$ was used as the mixing factor, with MPC applied at ViT layer $\ell = 8$. To measure the influence of each frequency for FSPM, 100 images were used per class for all classes. Additionally, the FSPM was updated every three epochs through the FIM process. Min-max normalization was utilized in FIM to normalize $\mathbf{P}$, with the highest measured loss per class serving as the normalization maximum. For balancing each loss relative to $\mathcal{L}_{cls}$, λ_1 and λ_2 were both set to 2. For the semantic segmentation model training on the PASCAL VOC 2012 dataset, Deeplab-V1 with ResNet38 was employed, while Deeplab-V2 with ResNet101 was used for the MS COCO 2014 dataset. Additional training details are in *Supplementary Materials*.

4.2 Ablation Studies

Component Analysis To demonstrate the significance of each module proposed in our framework, we ablated each component as shown in Tab. 1. Observing the performance of refined CAMs in Tab. 1-y, introducing the Frequency Shortcut Suppression (FSS) and Magnitude-mixing-based Phase Concentration (MPC) led to the gain of 2.2%p and 4.0%p respectively over baseline. Our method achieved 69.5% mIoU by adopting both components. FSS still resulted in a gain of 2.3%p in Tab. 1-c compared to Tab. 1-b, implies that FSS and MPC offer distinct advantages.

Table 2: Ablation study of ViT layer index that MPC performed. Peak performance is observed at layer 8, which is represented in **Bold**.

	early	layer	middle	layer	late	layer
ℓ	3	4	7	8	10	11
mIoU(%)	68.0	68.1	68.6	**69.5**	67.8	66.4

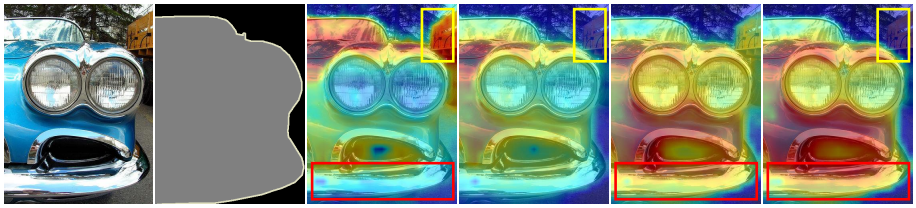

Fig. 4: Comparison of refined CAMs $\mathbf{M}_{ref}$ between baseline and components in ours. (a) Image, (b) Ground Truth, (c) Baseline, (d) only FSS, (e) only MPC, (f) ours. The yellow and red boxes imply effect of FSS and MPC, respectively.

To understand the benefits of FSS in CAMs, we examined the results of ViT CAMs that are not refined by the attention-map in Tab. 1-x. To conduct a comparative analysis considering the Precision-Recall tradeoff, thresholds corresponding to similar Recall levels were used. Low precision of baseline $\mathbf{M}$ indicates a significant number of over-activation. Through FSS, we observed a remarkable improvement in precision(18.1%p) and mIoU(14.1%p) compared to the baseline, confirming that false positives were successfully resolved (Tab. 1-b).

Ablation Studies of MPC The qualitative comparison between the refined CAMs of baseline and MPC in Fig. 4-e demonstrates the efficacy of the MPC by clearly delineating the boundary of the object. Meanwhile, we investigate the impact of the ratio r in MPC, which is used to preserve the magnitude information of the anchor feature. In experiments conducted within a search space below 0.5, there was a minimum gain of 2.3%p when compared to applying FSS alone, with the best performance observed at $r = 0.3$

The influence of the layer index ℓ where the MPC is performed was analyzed by conducting experiments on early, middle, and late layers in Tab. 2. The performance was highest when ℓ was set to 8, with a noticeable decline observed in layers beyond this point. This tendency supports the knowledge that later layers of the ViT play a crucial role in capturing high-level semantic information [9]. Introducing changes to the magnitude while preserving the phase information at these later stages could disrupt the semantic analysis of the model. This disruption may stem from the difference between the coupled information used in previous layers and the re-coupled information with the magnitude alteration. This result shows that MPC applied before when ViT features highly condense to high-level semantics can encourage the model to capture boundary information well in the phase spectrum.

Table 3: (a) True/False Positive Rate (TPR/FPR) of classification where the high-frequency components following FSPM are retained in images. Metrics are measured under the early training stage of the Baseline. **Bold** indicates the shortcut occurred, and Underline represents that model is trying to make a shortcut. (b) Comparative analysis of the Precision of CAMs **M** between baseline and only with FSS. Precision is evaluated under a similar Recall considering the Precision-Recall trade-off.

			bkg	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow
(a)	Baseline	TPR(%)	-	58.96	36.43	**88.47**	53.36	**68.46**	50.83	79.26	68.35	55.48	17.87
		FPR(%)	-	00.23	00.41	**15.78**	00.05	**33.74**	00.00	01.42	00.53	07.04	00.41
(b)	Baseline	P(%)	91.9	39.2	24.8	17.3	18.9	58.1	76.5	42.8	80.9	44.9	59.7
	w/ FSS	P(%)	92.5	52.9	42.2	59.2	33.3	71.1	87.5	62.9	94.1	43.2	90.7

			table	dog	horse	mbk	person	plant	sheep	sofa	train	tv	total
(a)	Baseline	TPR(%)	04.76	61.24	79.40	60.26	75.47	41.88	43.61	45.45	92.51	57.23	-
		FPR(%)	00.00	01.97	00.64	00.16	05.56	00.15	00.38	00.63	00.68	00.13	-
(b)	Baseline	P(%)	68.6	66.0	50.4	68.6	58.0	39.9	56.4	79.0	39.2	37.0	53.2
	w/ FSS	P(%)	89.6	87.5	79.7	78.5	82.7	64.7	92.5	81.2	54.5	56.6	71.3

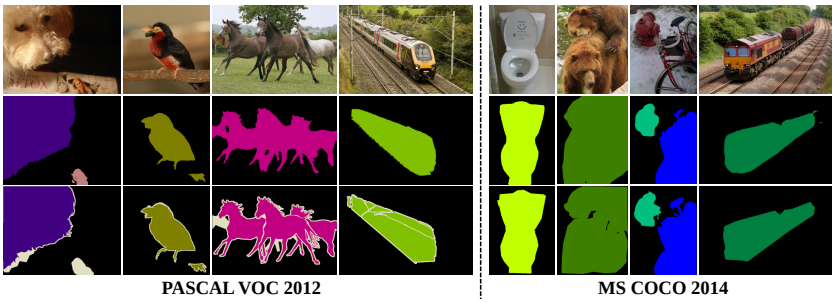

Fig. 5: Semantic segmentation results on PASCAL VOC 2012 (left) and MS COCO 2014 (right) datasets. From top to bottom: Image, Ours, GT.

Frequency Shortcut Analysis and Result of Suppression In Tab. 3, we investigated whether shortcut learning occurs during the model training process in early steps and analyzed the correlation with the improvement in CAMs precision facilitated by the FSS. To conduct this analysis, classification influence according to frequency component for each class was first measured using Eq. 7. Based on this measurement, the classification behavior of the model was analyzed when only some high frequencies of high-influence components were retained.

A high TPR indicates that class c possesses the principal frequency components utilized in the experiment (i.e., *bird* will be correctly classified as *bird* only with the high-influence frequency of *bird*). This suggests that the model has identified a straightforward method for classifying class c and is attempting to find frequency shortcuts. To consider a principal frequency as a shortcut, it should direct images to the respective class regardless of the actual semantic content in the image (*i.e. car* with frequency shortcut of *bird* will falsely classify as *bird*). This is expressed in Tab. 3 through high TPR and FPR, which repre-

Table 4: Comparison between ours and multi-stage WSSS methods. mIoU(%) is evaluated on PASCAL VOC 2012 *train* set in both CAMs (seed) and Pseudo-GT (Mask). The backbone and applied post-processing methods (PSA [2], IRN [1]) are listed for a fair comparison. **Bold** numbers represent the best performance.

Method	Backbone	Seed	Post	Mask
AdvCAM [28] CVPR'21	RN50	55.6	IRN	69.9
OC-CSE [25] ICCV'21	WRN38	56.0	PSA	66.9
ECS [53]ICCV'21	WRN38	56.6	PSA	67.8
CPN [70] ICCV'21	WRN38	57.4	PSA	67.8
AMR [45] AAAI'22	RN50	56.8	PSA	69.7
ReCAM [7] CVPR'22	RN50	54.8	IRN	70.5
SIPE [6] CVPR'22	RN50	58.6	IRN	-
PPC [10] CVPR'22	WRN38	61.5	IRN	70.1
AEFT [67] ECCV'22	WRN38	56.0	PSA	71.0
ACR [26] CVPR'23	WRN38	60.3	IRN	72.3
MCT [63] CVPR'22	DeiT-S	61.7	PSA	69.1
USAGE [44] ICCV'23	DeiT-S	67.7	PSA	72.8
Ours	DeiT-S	**69.5**	PSA	**73.2**

sent the rates in which principal frequencies of images are correctly and falsely classified to the shortcut respective class c. Examining Tab. 3-a, we can observe that class *bird* and *bottle* exhibit high TPR and FPR, indicating the occurrence of shortcuts. Meanwhile, classes such as *car*, *horse*, and *train* show high TPR, suggesting attempts at creating shortcuts.

Tab. 3-b records the class-specific precision of CAMs $\mathbf{M}$ for the baseline model and after the introduction of the FSS. To ensure a fair comparison while considering the Precision-Recall trade-off, precision was measured at similar levels of Recall. The introduction of FSS to disrupt shortcut creation resulted in an average increase of 24.0%p in precision for $\mathbf{M}$ across classes where the baseline model had created or attempted to create shortcuts. This increase confirms that the FSS effectively reduces false positives. Notably, for *bird*, a class where shortcuts had occurred, the precision of CAMs surged by 42.1%p, showcasing the resolve of over-activation in areas unrelated to the object by FSS. Qualitative comparison of refined CAMs $\mathbf{M}_{ref}$ before and after the introduction of FSS in Fig. 4-d further validates the effectiveness of FSS.

4.3 Comparison with State-of-The-Arts

To train the semantic segmentation model, the CAMs generated by our method were refined using PSA [2] as performed in previous WSSS works [44, 63, 67]. In Tab. 4, we compared the performance of CAMs (Seed) and pseudo-ground-truth (Mask) on the PASCAL VOC 2012 *train* set, where our method achieved the best performance in both categories. Utilizing our high-quality pseudo labels for training the semantic segmentation model, we compared the results with other WSSS methods in Tab. 5. Our framework demonstrated superiority by achieving the highest performance on the PASCAL VOC 2012 *val* and *test* sets.

Table 5: Comparison of semantic segmentation performance with multi-stage WSSS methods. Performance (mIoU) is evaluated in PASCAL VOC 2012 and MS COCO 2014 datasets. CNN-based methods are represented above the horizontal line, and ViT-based methods are depicted below the line. **Bold** represents the best performance.

Method	Backbone	VOC *val*	VOC *test*	COCO *val*
AdvCAM [28] CVPR'21	RN101	68.1	68.0	-
OC-CSE [25] ICCV'21	WRN38	68.4	68.2	36.4
ECS [53] ICCV'21	WRN38	66.6	67.6	-
CPN [70] ICCV'21	WRN38	67.8	68.5	-
AMR [45] AAAI'22	RN101	68.8	69.1	-
ReCAM [7] CVPR'22	RN101	68.5	68.4	42.9
SIPE [6] CVPR'22	RN101	68.8	69.7	-
SIPE [6] CVPR'22	WRN38	-	-	43.6
AEFT [67] ECCV'22	WRN38	70.9	71.7	44.8
ACR [26] CVPR'23	WRN38	71.9	71.9	45.3
MCT [63] CVPR'22	WRN38	71.9	71.6	42.0
USAGE [44] ICCV'23	WRN38	71.9	72.8	42.7
USAGE [44] ICCV'23	RN101	-	-	44.3
Ours	WRN38	**73.2**	**73.0**	**-**
Ours	RN101	**-**	**-**	**45.7**

While ViT-based WSSS works tend to outperform CNN-based works on the PASCAL VOC 2012 dataset, they showed lower performance on the MS COCO 2014 dataset due to activation overlapping between classes. Nonetheless, our method achieved the SoTA performance on the MS COCO 2014 dataset by effectively reducing false positives. Additional CAMs and semantic segmentation results are in the *Supplementary Materials*.

5 Conclusion

Weakly Supervised Semantic Segmentation suffers from CAMs that have imprecise activation around boundary areas and object-unrelated activations. In this paper, we aim to address these challenges by applying a novel perspective in the frequency domain, leveraging the Fourier Transform. Firstly, we propose Magnitude-mixing-based Phase Concentration (MPC), which generates features with identical *phase* but different *magnitude* at the intermediate layer, encouraging the model to capture the same semantics from these features. This approach leads the model to pay more attention to shape information in the *phase*, thereby learning to activate boundaries more clearly. Additionally, we identify the occurrence of frequency shortcuts and their association with the over-activation of CAMs. To resolve this, we introduce Frequency Shortcut Suppression (FSS), which discourages the creation of frequency shortcuts, successfully resolving object-unrelated activations. Our experimental results support the utility of the proposed method, demonstrating its effectiveness in addressing the aforementioned issues in WSSS. Furthermore, we have achieved new state-of-the-art performance on the PASCAL VOC and MS COCO datasets.

Acknowledgement This work was supported by the Technology Innovation Program (1415187329,20024355, Development of autonomous driving connectivity technology based on sensor-infrastructure cooperation) funded By the Ministry of Trade, Industry & Energy(MOTIE, Korea) and the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (NRF2022R1A2B5B03002636).

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. Chang, Y.T., Wang, Q., Hung, W.C., Piramuthu, R., Tsai, Y.H., Yang, M.H.: Weakly-supervised semantic segmentation via sub-category exploration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8991–9000 (2020)
4. Chen, G., Peng, P., Ma, L., Li, J., Du, L., Tian, Y.: Amplitude-phase recombination: Rethinking robustness of convolutional neural networks in frequency domain. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 458–467 (October 2021)
5. 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)
6. Chen, Q., Yang, L., Lai, J.H., Xie, X.: Self-supervised image-specific prototype exploration for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4288–4298 (2022)
7. Chen, Z., Wang, T., Wu, X., Hua, X.S., Zhang, H., Sun, Q.: Class reactivation maps for weakly-supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 969–978 (2022)
8. Dai, J., He, K., Sun, J.: Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In: Proceedings of the IEEE international conference on computer vision. pp. 1635–1643 (2015)
9. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
10. Du, Y., Fu, Z., Liu, Q., Wang, Y.: Weakly supervised semantic segmentation by pixel-to-prototype contrast. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4320–4329 (2022)
11. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International journal of computer vision **88**(2), 303–338 (2010)
12. Fan, J., Zhang, Z., Tan, T., Song, C., Xiao, J.: Cian: Cross-image affinity net for weakly supervised semantic segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 10762–10769 (2020)

13. Gao, W., Wan, F., Pan, X., Peng, Z., Tian, Q., Han, Z., Zhou, B., Ye, Q.: Ts-cam: Token semantic coupled attention map for weakly supervised object localization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2886–2895 (2021)
14. Geirhos, R., Jacobsen, J.H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., Wichmann, F.A.: Shortcut learning in deep neural networks. Nature Machine Intelligence **2**(11), 665–673 (2020)
15. Guo, J., Wang, N., Qi, L., Shi, Y.: Aloft: A lightweight mlp-like architecture with dynamic low-frequency transform for domain generalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24132–24141 (2023)
16. Hansen, B.C., Hess, R.F.: Structural sparseness and spatial phase alignment in natural scenes. JOSA A **24**(7), 1873–1885 (2007)
17. Huang, H., Kang, H., Liu, S., Salvado, O., Rakotoarivelo, T., Wang, D., Liu, T.: Paddles: Phase-amplitude spectrum disentangled early stopping for learning with noisy labels. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16719–16730 (2023)
18. Huang, H., Xie, S., Lin, L., Tong, R., Chen, Y.W., Li, Y., Wang, H., Huang, Y., Zheng, Y.: Semicvt: Semi-supervised convolutional vision transformer for semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11340–11349 (2023)
19. Huang, J., Guan, D., Xiao, A., Lu, S.: Rda: Robust domain adaptation via fourier adversarial attacking. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 8988–8999 (2021)
20. Huang, J., Liu, Y., Zhao, F., Yan, K., Zhang, J., Huang, Y., Zhou, M., Xiong, Z.: Deep fourier-based exposure correction network with spatial-frequency interaction. In: European Conference on Computer Vision. pp. 163–180. Springer (2022)
21. Jiang, P.T., Han, L.H., Hou, Q., Cheng, M.M., Wei, Y.: Online attention accumulation for weakly supervised semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(10), 7062–7077 (2021)
22. Jiang, P.T., Yang, Y., Hou, Q., Wei, Y.: L2g: A simple local-to-global knowledge transfer framework for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16886–16896 (2022)
23. Jo, S., Yu, I.J., Kim, K.: Mars: Model-agnostic biased object removal without additional supervision for weakly-supervised semantic segmentation. arXiv preprint arXiv:2304.09913 (2023)
24. Kim, G., Kim, J., Lee, J.S.: Exploring adversarial robustness of vision transformers in the spectral perspective. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 3976–3985 (2024)
25. Kweon, H., Yoon, S.H., Kim, H., Park, D., Yoon, K.J.: Unlocking the potential of ordinary classifier: Class-specific adversarial erasing framework for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6994–7003 (2021)
26. Kweon, H., Yoon, S.H., Yoon, K.J.: Weakly supervised semantic segmentation via adversarial learning of classifier and reconstructor. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11329–11339 (2023)
27. 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)

28. Lee, J., Kim, E., Yoon, S.: Anti-adversarially manipulated attributions for weakly and semi-supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4071–4080 (2021)
29. Lee, J., Oh, S.J., Yun, S., Choe, J., Kim, E., Yoon, S.: Weakly supervised semantic segmentation using out-of-distribution data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16897–16906 (2022)
30. Lee, J., Yi, J., Shin, C., Yoon, S.: Bbam: Bounding box attribution map for weakly supervised semantic and instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2643–2652 (2021)
31. Lee, S., Bae, J., Kim, H.Y.: Decompose, adjust, compose: Effective normalization by playing with frequency for domain generalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11776–11785 (2023)
32. Li, J., Fan, J., Zhang, Z.: Towards noiseless object contours for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16856–16865 (2022)
33. Li, K., Wu, Z., Peng, K.C., Ernst, J., Fu, Y.: Tell me where to look: Guided attention inference network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 9215–9223 (2018)
34. Li, X., Zhou, T., Li, J., Zhou, Y., Zhang, Z.: Group-wise semantic mining for weakly supervised semantic segmentation. arXiv preprint arXiv:2012.05007 (2020)
35. Li, Y., Duan, Y., Kuang, Z., Chen, Y., Zhang, W., Li, X.: Uncertainty estimation via response scaling for pseudo-mask noise mitigation in weakly-supervised semantic segmentation. arXiv preprint arXiv:2112.07431 (2021)
36. Li, Y., Kuang, Z., Liu, L., Chen, Y., Zhang, W.: Pseudo-mask matters in weakly-supervised semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6964–6973 (2021)
37. Lin, D., Dai, J., Jia, J., He, K., Sun, J.: Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3159–3167 (2016)
38. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
39. Lin, Y., Chen, M., Wang, W., Wu, B., Li, K., Lin, B., Liu, H., He, X.: Clip is also an efficient segmenter: A text-driven approach for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15305–15314 (2023)
40. Luo, X., Wei, L., Wen, L., Yang, J., Xie, L., Xu, Z., Tian, Q.: Rectifying the shortcut learning of background for few-shot learning. Advances in Neural Information Processing Systems **34**, 13073–13085 (2021)
41. Ma, C., Zhao, L., Chen, Y., Guo, L., Zhang, T., Hu, X., Shen, D., Jiang, X., Liu, T.: Rectify vit shortcut learning by visual saliency. IEEE Transactions on Neural Networks and Learning Systems (2023)
42. Oppenheim, A.V., Lim, J.S.: The importance of phase in signals. Proceedings of the IEEE **69**(5), 529–541 (1981)
43. Papandreou, G., Chen, L.C., Murphy, K.P., Yuille, A.L.: Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In: Proceedings of the IEEE international conference on computer vision. pp. 1742–1750 (2015)

44. 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)
45. Qin, J., Wu, J., Xiao, X., Li, L., Wang, X.: Activation modulation and recalibration scheme for weakly supervised semantic segmentation. arXiv preprint arXiv:2112.08996 (2021)
46. Robinson, J., Sun, L., Yu, K., Batmanghelich, K., Jegelka, S., Sra, S.: Can contrastive learning avoid shortcut solutions? Advances in neural information processing systems **34**, 4974–4986 (2021)
47. Rossetti, S., Zappia, D., Sanzari, M., Schaerf, M., Pirri, F.: Max pooling with vision transformers reconciles class and shape in weakly supervised semantic segmentation. In: European Conference on Computer Vision. pp. 446–463. Springer (2022)
48. Ru, L., Zhan, Y., Yu, B., Du, B.: Learning affinity from attention: End-to-end weakly-supervised semantic segmentation with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16846–16855 (2022)
49. Ru, L., Zheng, H., Zhan, Y., Du, B.: Token contrast for weakly-supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3093–3102 (2023)
50. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International journal of computer vision **115**(3), 211–252 (2015)
51. Su, Y., Sun, R., Lin, G., Wu, Q.: Context decoupling augmentation for weakly supervised semantic segmentation. arXiv preprint arXiv:2103.01795 (2021)
52. Sun, G., Wang, W., Dai, J., Van Gool, L.: Mining cross-image semantics for weakly supervised semantic segmentation. arXiv preprint arXiv:2007.01947 (2020)
53. Sun, K., Shi, H., Zhang, Z., Huang, Y.: Ecs-net: Improving weakly supervised semantic segmentation by using connections between class activation maps. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7283–7292 (2021)
54. Sundararajan, D.: The discrete Fourier transform: theory, algorithms and applications. World Scientific (2001)
55. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
56. Vernaza, P., Chandraker, M.: Learning random-walk label propagation for weakly-supervised semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7158–7166 (2017)
57. Wang, C., Xu, R., Xu, S., Meng, W., Zhang, X.: Treating pseudo-labels generation as image matting for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 755–765 (2023)
58. Wang, S., Veldhuis, R., Brune, C., Strisciuglio, N.: What do neural networks learn in image classification? a frequency shortcut perspective. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 1433–1442 (October 2023)
59. Wang, Y., Zhang, J., Kan, M., Shan, S., Chen, X.: Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12275–12284 (2020)

60. Wu, T., Huang, J., Gao, G., Wei, X., Wei, X., Luo, X., Liu, C.H.: Embedded discriminative attention mechanism for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16765–16774 (2021)
61. Xie, J., Hou, X., Ye, K., Shen, L.: Clims: Cross language image matching for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4483–4492 (2022)
62. Xie, J., Xiang, J., Chen, J., Hou, X., Zhao, X., Shen, L.: C2am: Contrastive learning of class-agnostic activation map for weakly supervised object localization and semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 989–998 (2022)
63. 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)
64. Xu, L., Ouyang, W., Bennamoun, M., Boussaid, F., Xu, D.: Learning multi-modal class-specific tokens for weakly supervised dense object localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19596–19605 (2023)
65. Xu, Q., Zhang, R., Zhang, Y., Wang, Y., Tian, Q.: A fourier-based framework for domain generalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14383–14392 (2021)
66. Yang, Y., Soatto, S.: Fda: Fourier domain adaptation for semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4085–4095 (2020)
67. 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)
68. Yoon, S.H., Kweon, H., Jeong, J., Kim, H., Kim, S., Yoon, K.J.: Exploring pixel-level self-supervision for weakly supervised semantic segmentation. arXiv preprint arXiv:2112.05351 (2021)
69. Yoon, S.H., Kwon, H., Kim, H., Yoon, K.J.: Class tokens infusion for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3595–3605 (2024)
70. Zhang, F., Gu, C., Zhang, C., Dai, Y.: Complementary patch for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7242–7251 (2021)
71. Zhang, X., Wei, Y., Feng, J., Yang, Y., Huang, T.S.: Adversarial complementary learning for weakly supervised object localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1325–1334 (2018)
72. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2921–2929 (2016)
73. Zhou, T., Zhang, M., Zhao, F., Li, J.: Regional semantic contrast and aggregation for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4299–4309 (2022)