Class Tokens Infusion for Weakly Supervised Semantic Segmentation

Sung-Hoon Yoon, Hoyong Kwon, Hyeonseong Kim, and Kuk-Jin Yoon
KAIST
{yoon307,kwonhoyong3,brian617,kjyoon}@kaist.ac.kr

Abstract

Weakly Supervised Semantic Segmentation (WSSS) relies on Class Activation Maps (CAMs) to extract spatial information from image-level labels. With the success of Vision Transformer (ViT), the migration of ViT is actively conducted in WSSS. This work proposes a novel WSSS framework with Class Token Infusion (CTI). By infusing the class tokens from images, we guide class tokens to possess class-specific distinct characteristics and global-local consistency. For this, we devise two kinds of token infusion: 1) Intra-image Class Token Infusion (I-CTI) and 2) Cross-image Class Token Infusion (C-CTI). In I-CTI, we infuse the class tokens from the same but differently augmented images and thus make CAMs consistent among various deformations (i.e. *view, color). In C-CTI, by infusing the class tokens from the other images and imposing the resulting CAMs to be similar, it learns class-specific distinct characteristics. Besides the CTI, we bring the background (BG) concept into ViT with the BG token to reduce the false positive activation of CAMs. We demonstrate the effectiveness of our method on PASCAL VOC 2012 and MS COCO 2014 datasets, achieving state-of-the-art results in weakly supervised semantic segmentation. The code is available at* https://github.com/yoon307/CTI.

1. Introduction

Fully supervised semantic segmentation shows great improvement in various fields in exchange for expensive and labor-intensive labels. To ease the burden of acquiring labels, Weakly Supervised Semantic Segmentation (WSSS) has emerged while utilizing only weak supervision.

With the weak labels that are relatively easy to acquire with extensive amounts, WSSS researches using image-level labels [1, 2, 4, 12, 19, 22, 28, 44, 52, 54], scribbles [29, 42], and bounding boxes [8, 18, 23, 32] are actively conducted. Among these, our work only utilizes image-level classification labels, the most practical and challenging setting.

Since classification labels only convey the presence or absence of objects with specific classes at an image level, objects are localized using Class Activation Maps (CAMs) [56] in WSSS. Numerous studies have aimed to improve the precision of CAMs to enable the obtained CAMs to serve as pseudo-labels for semantic segmentation.

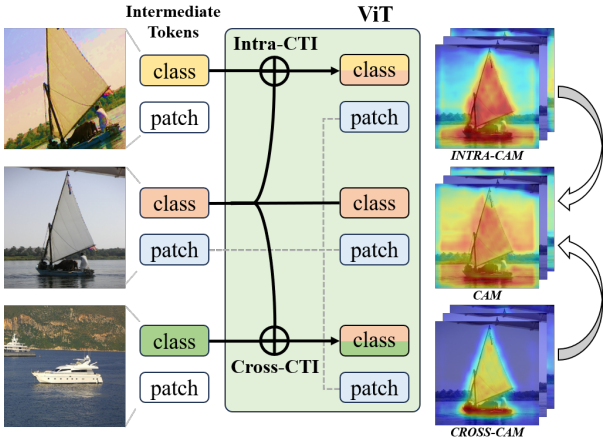

Figure 1. An overview of the proposed Class Token Infusion (CTI). We infuse class tokens in the intermediate layer by I-CTI and C-CTI. CTI guides class tokens and CAMs to be consistent in view differences and possess class-specific distinct representations.

Before Vision Transformer (ViT) emerged, most WSSS research relied on Convolutional Neural Networks (CNNs) to generate CAMs. However, CNNs are trained to localize the objects with limited receptive fields and the classification task itself can not impose spatial constraint, CAMs from CNN often focus only on discriminative object regions (*i.e.* sparseness).

Unlike CNNs that emphasize local features, ViTs capture long-range dependency through self-attention mechanisms; thus, it alleviates the sparseness problem of the activation map. Also, the ViT model trained in a self-supervised manner like DINO[3] or MOCO[15] shows that the localized objects are precise enough to function as a segmentation mask. However, since the original ViT uses single-class tokens to perform classification, the resulting localization map is acquired in a class-agnostic manner.

To bring this high localization capability to WSSS, several works employed multi-class tokens [49, 50] or directly trained the classifier with patch tokens [35] to extract class-specific activation maps.

Though the multi-class token-based ViT shows the best-performing results in WSSS, several problems remain unsolved. 1) While each class token is desirable to be distinct and have low correlation, we experimentally observed that the feature representation from each class token is highly correlated across the images, as shown in Fig. 4 above. 2) Due to the architectural nature of ViT, it is prone to over-expand the CAMs, leading to an increase in false positive regions [33, 35, 37]. The elevated occurrence of false positive regions in ViT can be attributed to two factors. Firstly, ViT's proficiency in capturing long-range dependencies enables the effective extraction of global contextual information. However, the GAP averages all features uniformly, resulting in the activation of irrelevant regions. The second contributing factor is the over-smoothing issue. As elucidated in ToCo [37], the self-attention mechanism functions as a low-pass filter, smoothing the input. Consequently, the mapping of patch tokens to similar latent representations intensifies, contributing to an increase in false positives.

To tackle the issues mentioned above and generate more precise CAMs, this paper introduces two methods: Class Token Infusion (CTI) and Background Token (BGT). Here, each method is designed to resolve the issues in ViT, respectively. In the CTI, as shown in Fig. 1, we propose two types of class token infusion: Intra-image Class Token infusion (I-CTI) and Cross-image Class Token Infusion (C-CTI). The class tokens encapsulate information about patch tokens as they undergo the attention process within the ViT layer to satisfy the goal of classification. However, we empirically find that classification loss is insufficient to make each class token distinct and observe overlaps in the feature space among different class tokens. Thus, with the proposed CTI, we aim to enhance the representation capability of class tokens, ensuring that each token uniquely condenses information relevant to its respective class. In the I-CTI, we first apply two different transformations to the image, producing positive pair images. By infusing intermediate class tokens from this pair and imposing the resulting CAMs to be consistent, we bestow global and local consistency upon CAMs. In the C-CTI, similar to the infusion method employed in I-CTI, we conduct class token infusion from the *other* images with at least one shared class. By ensuring consistency in CAMs before and after infusing class tokens obtained from different images, each class token enhances its ability to condense class-specific information and thus improves CAMs. Besides the CTI, we introduce the concept of background (BG) CAMs to ViT to address the issue of over-expansion. Though many prior CNN-based works utilize BG CAMs [6, 12, 48] in the training pipeline, less research focus is made on ViT-based WSSS to incorporate BG during **training**. By incorporating the BG token into ViT and instructing the network to predict BG CAMs, we can proficiently address false activations.

To show the effectiveness of our method, we conduct comparisons with the other state-of-the-art (SoTA) WSSS methods with two widely used datasets: PASCAL VOC 2012 [11] and MS COCO 2014 [30] datasets. In both datasets, the proposed framework achieves a new SoTA.

The contribution of this paper is twofold:

- We propose two forms of Class Token Infusion (CTI) methods to enhance the class-specific representation capability of class tokens and improve the quality of CAMs in ViT.
- We define a background token and propose a simple yet effective method to utilize the background CAMs in the learning process of ViT. We also demonstrate that utilizing the background token greatly reduces the false activation of CAMs.

2. Related Works

2.1. Weakly Supervised Semantic Segmentation

Improving CAMs Quality. To localize the object with only image-level labels, most WSSS approaches utilize Class Activation Maps [56] from CNNs. However, these CAMs (*i.e.* seeds) tend to focus on the discriminative regions of objects and localize imprecise boundaries. To enable CAMs to localize non-discriminative regions, various approaches have been researched in WSSS. By erasing the most discriminative regions and guiding the classifier to keep searching for object-related regions, Adversarial Erasing (AE) methods [19, 25, 40, 51, 55] effectively expands the CAMs. Other than AE methods, prior works introduced various training protocols such as sub-categories [4], cross-image semantic relations [13, 26, 39], complementary patches [54]. Recently, local-global consistency [16] and local prototype clustering [50] effectively extract non-discriminative features. In addition to expanding CAMs, many attempts [6, 12, 20, 48, 57] have been made to obtain CAMs with precise boundaries. By leveraging the strength of contrastive learning in semantic representation learning, many works with prototype-based contrastive learning [6, 48, 57] were actively conducted. ACR [20] first brought the reconstruction task to WSSS by conducting adversarial learning of the reconstructor and classifier.

Refining CAMs Since the pseudo-pixel-level ground truth generated from CAMs contains noisy information, several works have attempted to refine CAMs to get reliable labels. PSA [1] and IRNet [2] estimate the semantic affinity between pixels to further improve the mask quality. Adv-CAM [22] explored the less-discriminative region by manipulating the image in an anti-adversarial manner in a di-

rection to increase the classification score. Under-fitting strategy [28] to relieve the noisy information of pseudo labels or estimating the uncertainty of CAMs [27] by scaling the CAMs prediction multiple times are also proposed. Mat-Label [43] proposed an image-matting-based pseudo-label generation pipeline, and MARS [17] integrated the unsupervised semantic segmentation to WSSS for the removal of biased region. Though these post-processing methods efficiently improve the quality of CAMs, these methods are dependent on the initial seeds and can be used in conjunction with CAMs improvement methods, our method targets to improve the CAMs.

2.2. Vision Transformer with WSSS

With the powerful attention mechanism, ViT shows great improvements in various vision tasks. In line with this, many works [14, 31, 36, 37, 49, 50] are proposed to bring powerful localization capability to WSSS. After the successful migration of ViT to Weakly Supervised Object Localization (WSOL) in TS-CAM [14], MCTformer [49] leverages multi-class tokens to extract class-specific attention maps in the self-attention mechanism. AFA [36] proposed an end-to-end Transformer-based framework while utilizing the affinity from attention for CAMs refinement. With the advent of the Vision-Language foundation model, *i.e*. CLIP, Xu *et al*. [50] and Lin *et al*. [31] proposed framework to transfer the rich class representation capability of CLIP for WSSS. ToCo [37] points out the over-smoothing issue in ViT and addresses the issue by contrasting the class token from a global view with class tokens from local positive/negative images. ToCo [37] shares similarities with our approach by enhancing the representation capability of the class token. However, ToCo [37] focuses on ensuring representation consistency within a single image, whereas our method aims to achieve consistency for class tokens across both intra- and cross-images. Furthermore, our method optimizes with BG CAMs and BG token, while ToCo [37] requires two additional hyperparameters to distinguish the reliable foreground, background, and uncertain regions.

3. Method

In this section, we propose a ViT-based token infusion framework to overcome the limitations of conventional multi-class token-based WSSS methods. In our work, we introduce a Background Class Token (BGT) to reduce the false positive activation of CAMs. Then, to enhance the representation capability of Class Token, we also propose Intra/Cross-image Class Token Infusion (I/C-CTI).

3.1. Overall Framework

The overall framework of our method is shown in Fig. 2. As shown in the figure, three paired images are utilized for training. For the given RGB image $\mathbf{I}$, we apply various transformations (*e.g*., color jittering, resizing and cropping) and construct a strong-positive image $\dot{\mathbf{I}}$. Also, we sample one image if there are any overlapping class labels and use it as a weak-positive image $\ddot{\mathbf{I}}$. To propagate those images to the network, each image is split into $N \times N$ patches and embedded as patch tokens $\mathbf{T}_{patch} \in \mathbb{R}^{N^2 \times D}$ with embedding dimension D. Though we inherit [49] that uses C foreground (FG) class tokens $\mathbf{T}_{cls-fg} \in \mathbb{R}^{C \times D}$, **one more class token** $\mathbf{T}_{cls-bg} \in \mathbb{R}^{D}$ that represent background is used to form input class tokens $\mathbf{T}_{cls} \in \mathbb{R}^{(C+1) \times D}$ in our framework. The class tokens $\mathbf{T}_{cls}$ and patch tokens $\mathbf{T}_{patch}$ are concatenated to form input token $\mathbf{T}_{input} \in \mathbb{R}^{P \times D}$. Here, P is the sum of the length of the class token and patch token ($P = C + 1 + N^2$). After adding the positional embedding to the input token $\mathbf{T}_{input}$, the token is propagated to L transformer blocks. When $\mathbf{T}^k$ denotes the output token from k^{th} transformer block, we can acquire the CAMs $\mathbf{M} \in \mathbb{R}^{N \times N \times (C+1)}$ from the patch token output $\mathbf{T}^L_{patch} \in \mathbb{R}^{N^2 \times D}$ by applying reshaping and 2D convolutional layer. Note that the channel dimension of CAMs is $C + 1$ instead of C due to BG CAM ($\mathbf{M}_{bg} \in \mathbb{R}^{N \times N}$). To produce class prediction $y_{pred} \in \mathbb{R}^{C}$, we pool the class token output $\mathbf{T}^L_{cls}$ in embedding dimension except for BG token output. With multi-label soft margin loss, $\mathcal{L}_{cls}$ computes the entropy difference between class prediction y_{pred} and classification labels $\mathbf{y}$. Also, the classification loss at the patch level $\mathcal{L}_{cls-patch}$ is calculated by pooling the CAMs $\mathbf{M}$ (w/o BG) in spatial dimensions. Details on how we train the background token and infuse class tokens are explained in the following section.

3.2. Background Class Token

Though the Multi-Head Self-Attention (MHSA) mechanism in ViT effectively enlarges the CAMs to localize non-discriminative regions, it often leads to overly searching object non-related regions. Thus, we introduce a BG class token to ViT-based WSSS to reduce these false activation regions. Methods for improving CAMs using background information have been actively researched in CNN-based WSSS [6, 12, 48]. However, there has been minimal exploration of utilizing background in the context of ViTs while training. In our work, we explicitly define the background class token and guide it to interact with other patch/class tokens. With the self-attention map $\mathbf{A} \in \mathbb{R}^{P \times P}$ that computes Scaled Dot-Product Attention [41] between querys $\mathbf{Q} \in \mathbb{R}^{P \times D}$ and keys $\mathbf{K} \in \mathbb{R}^{P \times D}$, we extract class-patch attention $\mathbf{A}_{cp} \in \mathbb{R}^{N \times N \times (C+1)}$ and patch-patch attention $\mathbf{A}_{pp} \in \mathbb{R}^{N^2 \times N^2}$ similar to [49] by aggregating the self-attention maps along the L layers.

The training objective to get the BG CAM is formulated

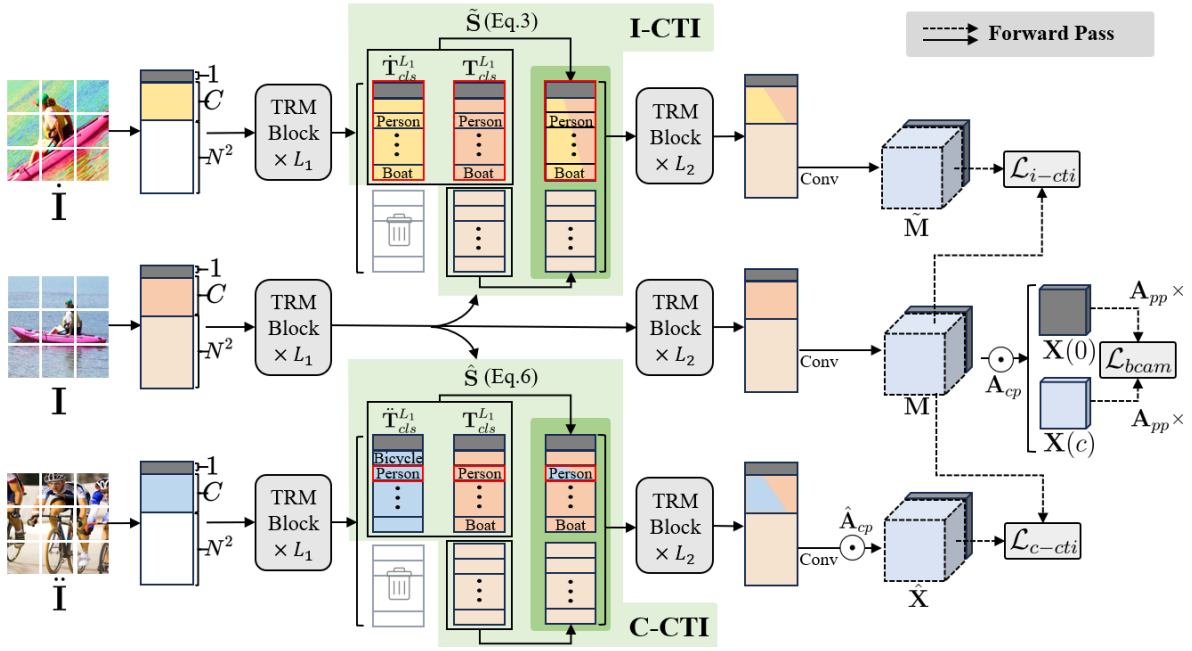

Figure 2. Visualization of the proposed framework. Our framework initiates with three paired images - two images $\mathbf{I}$ and $\dot{\mathbf{I}}$ originated from the **same** image with different transformation, and the $\ddot{\mathbf{I}}$ is sampled from the **other** images that share at least one same class with $\mathbf{I}$. These images are fed into the ViT block with C class tokens and a single background token. At the intermediate layer of the ViT, we perform Intra-image Class Token Infusion (I-CTI) and Cross-image Class Token Infusion (C-CTI). In I-CTI, class tokens corresponding to existing classes of $\dot{\mathbf{I}}$ are infused to class tokens belonging to $\mathbf{I}$. In C-CTI, a single class token from the overlapping classes between $\mathbf{I}$ and $\ddot{\mathbf{I}}$ is infused. The background token and background CAM are colored in black, while classification loss is omitted for simplicity. Note that the networks are weight-shared.

as follows:

$$\mathcal{L}_{bcam} = |\{1 - \{\mathbf{A}_{pp} \times \underset{c\in\mathbb{C}}{max}(\mathbf{X}(c))\}\} - (\mathbf{A}_{pp} \times \mathbf{X}(0))|_1, \tag{1}$$

where $\mathbf{X} = \mathbf{A}_{cp} \odot \mathbf{M}$. Here, $\odot$ and $\times$ respectively represent element-wise and matrix multiplication. $|\cdot|_1$ and $\mathbb{C}$ are L1 loss and set of foreground classes, respectively. Here, $\mathbf{X}(c)$ means the feature map at c^{th} channel. The output $\mathbf{A}_{pp} \times \mathbf{X}$ is normalized to 0-1 via min-max normalization.

3.3. Class Token Infusion

In this section, we propose Class Token Infusion (CTI) to ensure that the class tokens are robust to image view variation (global-local view consistency) and have a class-specific distinct representation. As shown in Fig. 2, we apply infusion method between the class token outputs ($\mathbf{T}^{L_1}_{cls}$, $\dot{\mathbf{T}}^{L_1}_{cls}$, $\ddot{\mathbf{T}}^{L_1}_{cls}$) from three images; original image $\mathbf{I}$, strong-positive image $\dot{\mathbf{I}}$, and weak-positive image $\ddot{\mathbf{I}}$, at the intermediate layer L_1. Here, we propose two types of class token infusion: 1) Intra-image Class Token Infusion (I-CTI) and 2) Cross-image Class Token Infusion (C-CTI).

Intra-image Class Token Infusion To make the ViT produce consistent CAMs for image view change, we propose I-CTI. Though the image $\mathbf{I}$ and strong-positive image $\dot{\mathbf{I}}$ come from the same image, different augmentation techniques such as color jittering and cropping/resizing, are employed. When we formulate the forward process as $\mathbf{T}^i_{cls}, \mathbf{T}^i_{patch} = \mathcal{F}_i(\mathbf{T}^{i-1}_{cls}, \mathbf{T}^{i-1}_{patch})$, where $\mathcal{F}_i$ is the i^{th} transformer block of ViT, I-CTI can be written as follows:

$$\begin{cases} \mathbf{T}^{i+1}_{cls}, \mathbf{T}^{i+1}_{patch} = \mathcal{F}_{i+1}(\mathbf{T}^i_{cls}, \mathbf{T}^i_{patch}), \; i < L_1 \\ \tilde{\mathbf{T}}^{i+1}_{cls}, \tilde{\mathbf{T}}^{i+1}_{patch} = \mathcal{F}_{i+1}(\tilde{\mathcal{S}}(\mathbf{T}^i_{cls}, \dot{\mathbf{T}}^i_{cls}, \dot{\mathbb{C}}), \mathbf{T}^i_{patch}), \; i = L_1 \\ \tilde{\mathbf{T}}^{i+1}_{cls}, \tilde{\mathbf{T}}^{i+1}_{patch} = \mathcal{F}_{i+1}(\tilde{\mathbf{T}}^i_{cls}, \tilde{\mathbf{T}}^i_{patch}), \; i > L_1, \end{cases} \tag{2}$$

where $\tilde{\mathbf{T}}$, $\tilde{\mathcal{S}}(\cdot)$, and $\dot{\mathbb{C}}$ denote the token output as a result of the I-CTI, infusion operation, and set of all classes, respectively. $\tilde{\mathcal{S}}(\cdot)$ is the operation that executes $\tilde{\mathcal{S}}_c(\cdot)$ for each $c \in \dot{\mathbb{C}}$, where the detailed formula for $\tilde{\mathcal{S}}_c(\cdot)$ is as follows:

$$\tilde{\mathcal{S}}_c(\mathbf{T}_1, \mathbf{T}_2, \dot{\mathbb{C}}) = \frac{\mathbf{T}_1(c) + \mathbf{T}_2(c)}{2}. \tag{3}$$

By applying $\tilde{\mathcal{S}}$, class tokens $\dot{\mathbf{T}}_{cls}$ are infused to class tokens of $\mathbf{T}_{cls}$. As shown in the second row of Eq. 2, the infused class token, which is result of $\tilde{\mathcal{S}}(\mathbf{T}^{L_1}_{cls}, \dot{\mathbf{T}}^{L_1}_{cls}, \dot{\mathbb{C}})$, interacts with the patch token $\mathbf{T}^{L_1}_{patch}$ in transformer block $\mathcal{F}$ with self-attention mechanism and produce Token $\tilde{\mathbf{T}}$. After applying $\tilde{\mathbf{T}}$ to transformer block $\mathcal{F}$ several times, we can obtain the $\tilde{\mathbf{T}}^L$; the final result of token infusion. Then, we minimize the difference $\mathbf{M}$ and $\tilde{\mathbf{M}}$ in CAMs-level which

are from $\mathbf{T}^L_{patch}$ and $\tilde{\mathbf{T}}^L_{patch}$, respectively. This constraint can be formulated as follows:

$$\mathcal{L}_{i-cti} = \sum_{c=0}^{C+1} |\mathbf{M}(c) - \tilde{\mathbf{M}}(c)|_1. \quad (4)$$

Cross-image Class Token Infusion To guide each class token to possess class-distinct and specific information, we propose C-CTI. To perform classification using ViT, class tokens interact with patch tokens and are modified to possess image-specific information. Thus, global contextual representation corresponding to the class is stored in the class token. However, we empirically find that classification is insufficient to guide class tokens to learn class-relevant information, and the feature representation from each class token is highly correlated. To resolve this problem, in C-CTI, we infuse class tokens from the other images: $\mathbf{I}$ and $\ddot{\mathbf{I}}$. By guiding the resulting CAMs to be similar before and after infusing class tokens, the class token maintains a meaningful global representation that can express each class. At the same time, the model can learn to focus on the unique class-relevant characteristics by sharing a unique representation obtained from different images. Similar to Eq. 2, the C-CTI can be written as follows:

$$\begin{cases} \mathbf{T}^{i+1}_{cls}, \mathbf{T}^{i+1}_{patch} = \mathcal{F}_{i+1}(\mathbf{T}^i_{cls}, \mathbf{T}^i_{patch}), \; i < L_1 \\ \hat{\mathbf{T}}^{i+1}_{cls}, \hat{\mathbf{T}}^{i+1}_{patch} = \mathcal{F}_{i+1}(\hat{\mathcal{S}}(\mathbf{T}^i_{cls}, \ddot{\mathbf{T}}^i_{cls}, \ddot{\mathbb{C}}), \mathbf{T}^i_{patch}), \; i = L_1 \\ \hat{\mathbf{T}}^{i+1}_{cls}, \hat{\mathbf{T}}^{i+1}_{patch} = \mathcal{F}_{i+1}(\hat{\mathbf{T}}^i_{cls}, \hat{\mathbf{T}}^i_{patch}), \; i > L_1, \end{cases} \quad (5)$$

where $\hat{\mathbf{T}}$ and $\ddot{\mathbb{C}}$ denote the token output as a result of C-CTI and the set which contains selected class, respectively. The infusion operation $\hat{\mathcal{S}}$ for C-CTI can be expressed as follows:

$$\hat{\mathcal{S}}_c(\mathbf{T}_1, \mathbf{T}_2, \ddot{\mathbb{C}}) = \mathbb{1}_{c \notin \ddot{\mathbb{C}}} \mathbf{T}_1(c) + \mathbb{1}_{c \in \ddot{\mathbb{C}}} \frac{\mathbf{T}_1(c) + \mathbf{T}_2(c)}{2}, \quad (6)$$

where $\mathbb{1}_{\mathcal{P}}$ denotes the indicator function that returns 1 if $\mathcal{P}$ is satisfied, and 0 otherwise. Here, by applying $\hat{\mathcal{S}}$, only one shared class between image $\mathbf{I}$ and weak-positive image $\ddot{\mathbf{I}}$ is sampled, and the corresponding class token is infused. Then, we minimize the difference $\mathbf{M}$ and $\hat{\mathbf{X}}$ in CAMs-level which are from $\mathbf{T}^L_{patch}$ and $\hat{\mathbf{T}}^L_{patch}$, respectively. This constraint can be formulated as follows:

$$\mathcal{L}_{c-cti} = \sum_{c=1}^{C+1} |\mathbf{M}(c) - \hat{\mathbf{X}}(c)|_1. \quad (7)$$

Here, to leverage the class-patch attention $\mathbf{A}_{cp}$ information from the other images, we used $\hat{\mathbf{X}}$ instead of $\hat{\mathbf{M}}$.

Training Objective In the proposed CTI, only the class tokens from $\dot{\mathbf{I}}$ and $\ddot{\mathbf{I}}$ are infused to class token from $\mathbf{I}$, while the patch tokens from $\dot{\mathbf{I}}$ and $\ddot{\mathbf{I}}$ are not utilized in the deeper layers. In I-CTI at layer $i = L_1$, only the **class** token $\mathbf{T}^i_{cls}$ (from $\mathbf{I}$) and the **class** token $\dot{\mathbf{T}}^i_{cls}$ (from $\dot{\mathbf{I}}$) are fused. The **patch** token $\dot{\mathbf{T}}^i_{patch}$ is not utilized further in the deeper layer ($i > L_1$). As shown in the second and third row of Eq. 2, only the **patch** token $\mathbf{T}^i_{patch}$ is forwarded with the infused class token $\tilde{\mathcal{S}}(\mathbf{T}^{L_1}_{cls}, \dot{\mathbf{T}}^{L_1}_{cls}, \dot{\mathbb{C}})$ and forms $\tilde{\mathbf{T}}^L$. Since the shape of CAMs is determined by patch tokens (of the last layer) and $\tilde{\mathbf{M}}$ is based on patch tokens obtained from image $\mathbf{I}$, $\tilde{\mathbf{M}}$ thus shares the **same position** with $\mathbf{M}$. Since the C-CTI shares a similar infusion mechanism, the 'shape' of $\mathbf{M}, \tilde{\mathbf{M}}, \hat{\mathbf{M}}$ are the same. The final loss of the proposed framework is formulated as follows:

$$\mathcal{L}_{all} = \mathcal{L}_{cls} + \mathcal{L}_{cls-patch} + \lambda \mathcal{L}_{bcam} + \mathcal{L}_{i-cti} + \mathcal{L}_{c-cti},$$

where λ is hyperparameter that balance the weight with respect to $\mathcal{L}_{cls}$.

4. Experiments

4.1. Experimental Settings

Datasets Following the prior WSSS works, we evaluate our method on the PASCAL VOC 2012 dataset [11] and the MS-COCO 2014 dataset [30], two most widely used benchmarks. The PASCAL VOC 2012 dataset contains 20 foreground object classes and one background class with 10582, 1449, and 1456 images in *train*, *val*, *test* set, respectively. The MS-COCO 2014 dataset, which is a more challenging dataset with 82k *train* set and 40k *val* set, consists of 80 foreground object classes and one background class.

Evaluation Metric For the evaluation of semantic segmentation performance, we use mean Intersection over Union (mIoU) by following the prior works [20, 49–51]. The mIoU of the semantic segmentation model is evaluated on *val* set while the CAMs performance is evaluated on *train* set. The results on the PASCAL VOC 2012 *test* set are evaluated through the online official server.

Implementation Details In our framework, we use DeiT-S pre-trained on ImageNet [9] as a backbone of classifier for the fair comparison with previous ViT-based WSSS [14, 20, 49, 50]. The classification network is trained for 60 epochs on both datasets, employing the Adam optimizer with an initial learning rate of 5e-4 and a batch size of 64. As in MCTformer [49], the same data augmentation methods are applied with different image resize scales, and images are cropped to 224×224. Unlike the MCTformer [49], in which the multiple class tokens are initialized with the same pre-trained class token, our method splits the class token with a fully-connected layer. Using a fully connected (FC) layer for class tokens does not improve performance (mIoU 64.9%) when used solely with a classification loss. However, using class tokens with an FC layer is effective when incorporating with the background owing to the correlation between the background and the foreground. The

Table 1. Ablation study on the PASCAL VOC 2012 *train* set. **Bold** numbers represent the best results. P: Precision, R: Recall.†: We re-implement the baseline [49] for a fair comparison within our setting.

	$\mathcal{L}_{bcam}$	$\mathcal{L}_{i-cti}$	$\mathcal{L}_{c-cti}$	P(%)	R(%)	mIoU (%)
Baseline†				75.2	81.8	64.7
(a)	✓			78.7	82.6	67.6
(b)	✓	✓		79.0	83.7	68.7
(c)	✓		✓	78.9	84.7	69.0
(d)	✓	✓	✓	80.0	84.0	69.5

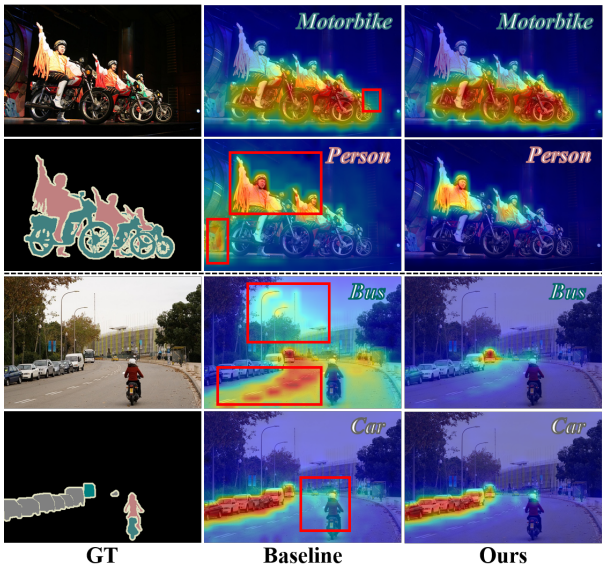

Figure 3. Comparison of CAM between the Baseline and Ours with only background CAM loss, $\mathcal{L}_{bcam}$, applied. The red box indicates the over-activated regions towards the background area.

background token is initialized with zeros. For the C-CTI, to prevent cases where no corresponding classes exist in the mini-batch, we stacked the class tokens into memory and used them. To balance the magnitude of weight with respect to $\mathcal{L}_{cls}$, the λ is set to 0.1. To generate the pseudo labels for the training of the semantic segmentation model, we employed the same post-processing model (IRN [2]) as in prior WSSS works [20, 22, 24, 50, 53]. For the semantic segmentation model, Deeplab-V1 with a ResNet38 backbone is used for the PASCAL VOC 2012 dataset. For the MS COCO 2014 dataset, we used Deeplab-V2 with a ResNet101 backbone. Additional training details are in the *Supp. Materials*.

4.2. Ablation Studies

Component analysis To demonstrate the importance of each method we propose, we ablate the methods as shown in Table 1. By bringing the concept of background CAMs to ViT with a background token, we can obtain a 2.9%p gain compared to the baseline (Table 1-a). CAMs results in Fig. 3 clearly show the effect of background token. BG CAM visualization results are in the *Supp. Materials*. With the proposed intra-class token infusion (I-CTI), we can get additional performance improvement with 1.1%p (Table 1-b). We boost the performance to 69.5% with the proposed cross-class token infusion (C-CTI) as in Table 1-d.

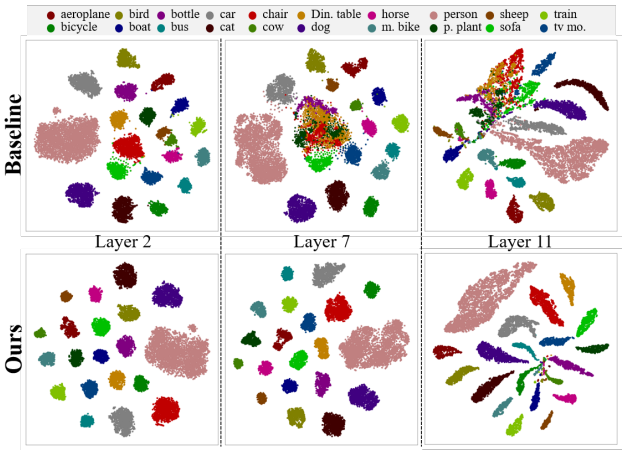

Figure 4. t-SNE comparison result between Baseline [49] (top) and Ours (bottom). Class tokens $\mathbf{T}^i_{cls}$ in layer $i \in \{2, 7, 11\}$ are used for t-SNE and sampled from the PASCAL VOC 2012 *train* set (10,582 images). Only the class token existing in each image is sampled. Each class is represented with distinct colors and the color scheme follows the legend at the top.

Importance of Background Class Token To emphasize the importance of utilizing the BG class token $\mathbf{T}_{cls-bg}$ in our approach, we conducted an experiment by training the baseline model without the BG class token but with BG CAM. To achieve this, we increased the number of class prediction heads for the patch token from C to C+1 and trained it solely with $\mathcal{L}_{bcam}$. Since $\mathbf{X}(0)$ is not defined in the absence of the BG class token, we trained it using $\mathbf{M}(0)$. The results indicated a 4.1%p decrease compared to the baseline, underscoring the significance of the BG class token $\mathbf{T}_{cls-bg}$ in training a BG CAM.

Role of Class Token Infusion The class tokens $\mathbf{T}^i_{cls}$ in layer $i \in 2, 7, 11$ are visualized using t-SNE in Fig. 4 to demonstrate the effectiveness of Token Infusion. Specifically, the class tokens corresponding to the class labels are sampled from the PASCAL VOC 2012 *train* set, which comprises 10,582 images. Referring to the t-SNE results of the Baseline [49] in Fig. 4 above, we observe that class tokens acquired in the early layers (*i.e.*, Layer 2) exhibit some degree of distinctiveness between classes. However, the feature space is not well-separated in the intermediate (*i.e.*, Layer 7) and late layers (*i.e.*, Layer 11). The t-SNE result of Layer 11 suggests a lack of clear separation in the feature space among classes. In contrast, as illustrated in the figure

Table 2. The performance (mIoU,%) variation of the proposed method based on the infusion index L_1. Results consistently outperform the baseline, which is 64.7%. Peak performance, denoted in **Bold**, is observed at layer 3.

L_1	2	3	4	5	6	7	8	9	10
mIoU(%)	68.8	**69.5**	68.6	68.9	69.4	68.7	69.0	68.7	68.8

Table 3. Comparisons between our method and the other WSSS methods. Methods that use the same post-processing methods (PSA [1] or IRN [2]) are listed in the table. The mIoU (%) on the PASCAL VOC 2012 *train* set is reported for CAMs (Seed) and pseudo-ground-truth (Mask), respectively. The backbone for each method is also listed (Backbone). **Bold** represents the best results. Results above the table line are CNN-based, while those below are ViT-based.

Methods	Backbone	Seed	Mask
CDA [38] $_{ICCV21}$	ResNet38	58.4	66.4
OC-CSE [19]$_{ICCV21}$	ResNet38	56.0	66.9
EDAM [45] $_{CVPR21}$	ResNet38	52.8	68.1
AMR [34] $_{AAAI22}$	ResNet50	56.8	69.7
ReCAM [7]$_{CVPR22}$	ResNet50	54.8	70.5
RIB [21] $_{NeurIPS}$	ResNet50	56.5	70.6
CLIMS [47]$_{CVPR22}$	ResNet38	56.6	70.5
PPC [10]$_{CVPR22}$	ResNet38	61.5	70.1
AEFT [51] $_{ECCV22}$	ResNet38	56.0	71.0
ACR [20] $_{CVPR23}$	ResNet38	60.3	72.3
MCT [49] $_{CVPR22}$	DeiT-S	61.7	69.1
FPR [5] $_{ICCV23}$	DeiT-S	63.8	-
ACR+ViT [20] $_{CVPR23}$	DeiT-S	65.5	70.9
USAGE [33] $_{ICCV23}$	DeiT-S	67.7	72.8
Ours	DeiT-S	**69.5**	**73.7**

below, the t-SNE results support the idea that class tokens from our method yield well-distinguished feature spaces not only in the final layer but across all layers. Considering that CAMs from ViT utilize the class-patch attention $\mathbf{A}_{cp}$ that is aggregated from all layers, this clear separation between classes in feature space can help to improve the quality of CAMs. The qualitative comparison results between the baseline [49] and our approach in Fig. 5 support that the CAMs generated by our methods exhibit greater distinctiveness and do not encroach upon the regions of other classes, while the CAMs from the baseline activate the wrong regions (red box). Since the class tokens from the baseline are not well-separated, as illustrated in Fig 5 above, the activation of 'Bus' class intrudes the regions of 'Train' class, while the activation of the 'Person' class extends into the regions of the 'Bus' class with higher confidence.

Effect of infusing index To show the effect of an index in infusion, we conduct an ablation study. As shown in Table 2, the performance in mIoU (%) is calculated by varying the infusion index L_1 from 2 to 10 where the total number of layers L is 12. The performance is highest when the in-

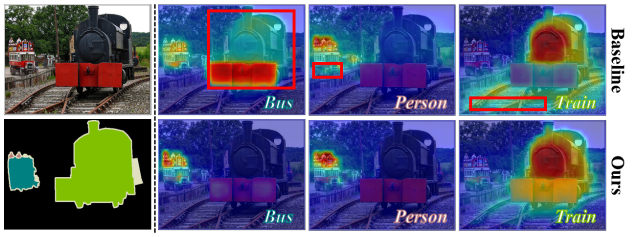

Figure 5. Qualitative comparison results between the Baseline (top) and Ours (bottom). Here, the red box indicates the false positive activation.

Table 4. Semantic segmentation performance comparison in mIoU (%) with the existing WSSS methods. *Sup.* denotes supervision; I: image-level label, S: Saliency maps, L: pre-trained Language model. For a fair comparison, only the methods using the same post-processing methods (PSA [1] or IRN [2]) are listed in this table. **Bold** numbers represent the best results.

Methods	*Backbone*	*Seg.*	*Sup.*	*Val*	*Test*
OC-CSE [19]$_{ICCV21}$	ResNet38	V1	I	68.4	68.2
ReCAM [7]$_{CVPR22}$	ResNet101	V2	I	68.5	68.4
CPN [54]$_{ICCV21}$	ResNet38	V1	I	67.8	68.5
RIB [21]$_{NeurIPS21}$	ResNet101	V2	I	68.3	68.6
CLIMS [47] $_{CVPR22}$	ResNet101	V2	I+L	69.3	68.7
PMM [28]$_{ICCV21}$	ResNet38	V1	I	68.5	69.0
EDAM [45] $_{CVPR21}$	ResNet101	V2	I+S	70.9	70.6
FPR [5] $_{ICCV23}$	ResNet38	V1	I	70.0	70.6
Spatial-BCE [46]$_{ECCV22}$	ResNet101	V2	I	70.0	71.3
AEFT [51]$_{ECCV22}$	ResNet38	V1	I	70.9	71.7
ACR [20] $_{CVPR23}$	ResNet38	V1	I	71.9	71.9
L2G [16] $_{CVPR22}$	ResNet38	V1	I+S	72.0	73.0
MCT [49] $_{CVPR22}$	ResNet38	V1	I	71.9	71.6
ACR+ViT [20] $_{CVPR23}$	ResNet38	V1	I	72.4	72.4
USAGE [33] $_{ICCV23}$	ResNet38	V1	I	71.9	72.8
Ours	ResNet38	V1	I	**74.1**	**73.2**

fusion index L_1 is set to 3. There is a slight performance difference depending on the index performing the infusion, but it consistently demonstrates high performance. Referring to the t-SNE results of baseline [49] in Fig. 4 above, the feature space of class tokens becomes less distinct in the later layers, thus the effectiveness of the proposed CTI is more evident when conducted in the early layers. Yet it still brings a minimal 1.0%p increase when compared to the case without CTI (Table 1-(a)). With these robust results, we set the infusion index to 3.

4.3. Comparisons to State-of-The-Arts

PASCAL VOC As shown in Table 3, we compare the performance of CAMs (seed) and pseudo pixel-level ground-truth (Mask) on *train* set. Our method shows better performance both at the seed and mask levels. Compared with the second best result [33], we achieve +1.8%p and +0.9%p gain at seed and mask performance, respectively. In Table 4, the performance of the semantic segmentation on the PASCAL VOC 2012 dataset is listed. The semantic segmenta-

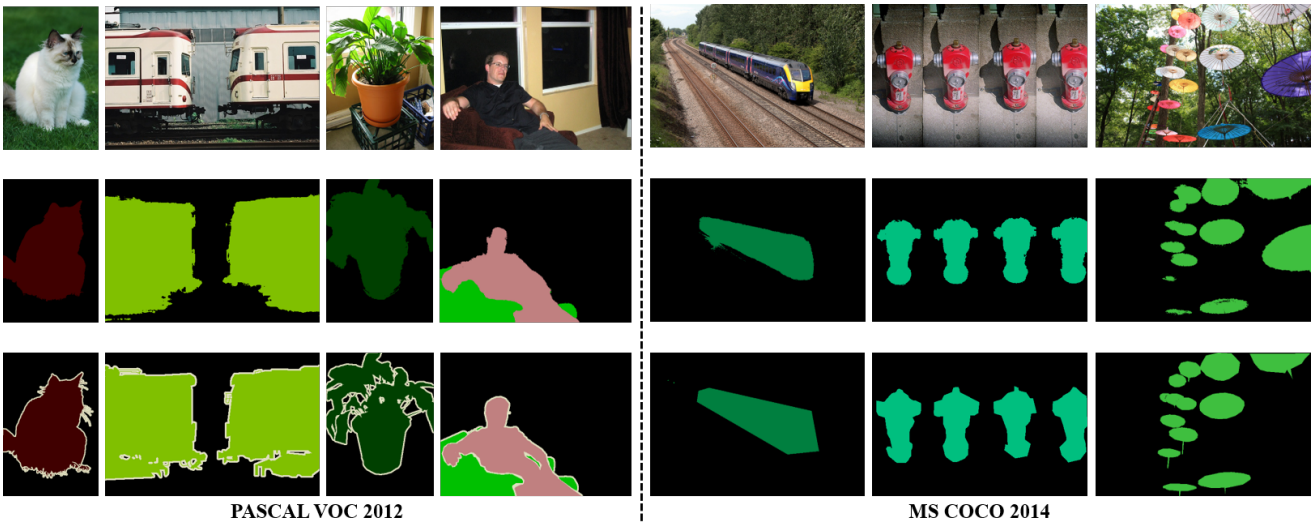

Figure 6. Our semantic segmentation results on VOC 2012 (left) and COCO 2014 (right). From top to bottom: Image, Ours, GT.

Table 5. Comparison in mIoU (%) performance between the proposed method and the existing WSSS methods. Evaluation is conducted on MS-MOCO 2014 *val* set. **Bold** numbers represent the best results.

Methods	*Backbone*	*Seg.*	*Sup.*	*Val*
IRNet [2]$_{CVPR19}$	ResNet50	V2	I	41.4
ReCAM [7]$_{CVPR22}$	ResNet101	V2	I	42.9
SIPE [6]$_{CVPR22}$	ResNet38	V1	I	43.6
RIB [21]$_{NeurIPS21}$	ResNet101	V2	I	43.8
FPR [5] $_{ICCV23}$	ResNet101	V2	I	43.9
L2G [5] $_{ICCV23}$	ResNet101	V2	I+S	44.2
AEFT [51]$_{ECCV22}$	ResNet38	V1	I	44.8
ACR [20] $_{CVPR23}$	ResNet38	V1	I	45.3
MCT [49] $_{CVPR22}$	ResNet38	V1	I	42.0
USAGE [33] $_{ICCV23}$	ResNet101	V2	I	44.3
Ours	ResNet101	V2	I	**45.4**

tion trained with these high-quality labels outperforms the SoTA both on *val* and *test* set with a great margin. Interestingly, our semantic segmentation model shows better performance than the pseudo-labels, considering that other methods show slightly lower performance than the pseudo-labels. Furthermore, though increasing performance becomes more challenging as it saturates, we have more than 1.7%p gain on *val* set compared to the second-best model.

MS COCO As shown in Table 5, we also trained and evaluated our model on MS COCO 2014 dataset. Though the dataset contains more classes with complex scenes, our method shows promising results with 45.4% mIoU and supports the robust generalization ability of our model. ViT-based WSSS methods have exhibited superior performance compared to CNN-based methods on PASCAL VOC 2012. However, on the MS COCO 14 dataset, they have shown lower performance due to false activation and activation overlap between classes. However, with the proposed method, we reduce the gap by effectively reducing the false activation regions while guiding class tokens to be distinct. Additional CAMs and semantic segmentation visualization results are in the *Supp. Materials*.

5. Conclusion

In this work, we aim to enhance the class-specific representation capability of the class tokens in ViT to localize the objects in an image distinctively. For this, we propose two types of Class Token Infusion (CTI): Intra-image Class Token Infusion (I-CTI) and Cross-image Class Token Infusion (C-CTI), which infuses the class tokens from the same or other images. In I-CTI, we infused the class tokens from the same but augmented image to the class token from the original image. By guiding the CAMs before and after the infusion to be the same, we bestow global-local consistency. C-CTI extends this infusion process to the other images with at least one shared class. Through the C-CTI, both class tokens and CAMs possess consistent class-specific knowledge that can be shared across the images. Furthermore, to reduce the false activation of CAMs, we incorporate the Background Token (BGT) into ViT. We also experimentally demonstrate that using BGT effectively addresses over-expansion issues. Extensive experimental results on the VOC and COCO support the validity and generalizability of the proposed method. We also achieved state-of-the-art in both datasets.

Acknowledgements This research was supported by National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (NRF2022R1A2B5B03002636).

References

[1] Jiwoon Ahn and Suha Kwak. 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*, pages 4981–4990, 2018. 1, 2, 7

[2] Jiwoon Ahn, Sunghyun Cho, and Suha Kwak. Weakly supervised learning of instance segmentation with inter-pixel relations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2209–2218, 2019. 1, 2, 6, 7, 8

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 1

[4] Yu-Ting Chang, Qiaosong Wang, Wei-Chih Hung, Robinson Piramuthu, Yi-Hsuan Tsai, and Ming-Hsuan Yang. Weakly-supervised semantic segmentation via sub-category exploration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8991–9000, 2020. 1, 2

[5] Liyi Chen, Chenyang Lei, Ruihuang Li, Shuai Li, Zhaoxiang Zhang, and Lei Zhang. Fpr: False positive rectification for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1108–1118, 2023. 7, 8

[6] Qi Chen, Lingxiao Yang, Jian-Huang Lai, and Xiaohua Xie. Self-supervised image-specific prototype exploration for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4288–4298, 2022. 2, 3, 8

[7] Zhaozheng Chen, Tan Wang, Xiongwei Wu, Xian-Sheng Hua, Hanwang Zhang, and Qianru Sun. Class re-activation maps for weakly-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 969–978, 2022. 7, 8

[8] Jifeng Dai, Kaiming He, and Jian Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1635–1643, 2015. 1

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5

[10] Ye Du, Zehua Fu, Qingjie Liu, and Yunhong Wang. Weakly supervised semantic segmentation by pixel-to-prototype contrast. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4320–4329, 2022. 7

[11] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 2, 5

[12] Junsong Fan, Zhaoxiang Zhang, Chunfeng Song, and Tieniu Tan. Learning integral objects with intra-class discriminator for weakly-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4283–4292, 2020. 1, 2, 3

[13] Junsong Fan, Zhaoxiang Zhang, Tieniu Tan, Chunfeng Song, and Jun Xiao. Cian: Cross-image affinity net for weakly supervised semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10762–10769, 2020. 2

[14] Wei Gao, Fang Wan, Xingjia Pan, Zhiliang Peng, Qi Tian, Zhenjun Han, Bolei Zhou, and Qixiang Ye. Ts-cam: Token semantic coupled attention map for weakly supervised object localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2886–2895, 2021. 3, 5

[15] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 1

[16] Peng-Tao Jiang, Yuqi Yang, Qibin Hou, and Yunchao Wei. 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*, pages 16886–16896, 2022. 2, 7

[17] Sanghyun Jo, In-Jae Yu, and Kyungsu Kim. Mars: Model-agnostic biased object removal without additional supervision for weakly-supervised semantic segmentation. *arXiv preprint arXiv:2304.09913*, 2023. 3

[18] Anna Khoreva, Rodrigo Benenson, Jan Hosang, Matthias Hein, and Bernt Schiele. Simple does it: Weakly supervised instance and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 876–885, 2017. 1

[19] Hyeokjun Kweon, Sung-Hoon Yoon, Hyeonseong Kim, Daehee Park, and Kuk-Jin Yoon. 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*, pages 6994–7003, 2021. 1, 2, 7

[20] Hyeokjun Kweon, Sung-Hoon Yoon, and Kuk-Jin Yoon. Weakly supervised semantic segmentation via adversarial learning of classifier and reconstructor. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11329–11339, 2023. 2, 5, 6, 7, 8

[21] Jungbeom Lee, Jooyoung Choi, Jisoo Mok, and Sungroh Yoon. Reducing information bottleneck for weakly supervised semantic segmentation. *Advances in Neural Information Processing Systems*, 34:27408–27421, 2021. 7, 8

[22] Jungbeom Lee, Eunji Kim, and Sungroh Yoon. Anti-adversarially manipulated attributions for weakly and semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4071–4080, 2021. 1, 2, 6

[23] Jungbeom Lee, Jihun Yi, Chaehun Shin, and Sungroh Yoon. 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*, pages 2643–2652, 2021. 1

[24] Jungbeom Lee, Seong Joon Oh, Sangdoo Yun, Junsuk Choe, Eunji Kim, and Sungroh Yoon. Weakly supervised semantic segmentation using out-of-distribution data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16897–16906, 2022. 6

[25] Kunpeng Li, Ziyan Wu, Kuan-Chuan Peng, Jan Ernst, and Yun Fu. Tell me where to look: Guided attention inference network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9215–9223, 2018. 2

[26] Xueyi Li, Tianfei Zhou, Jianwu Li, Yi Zhou, and Zhaoxiang Zhang. Group-wise semantic mining for weakly supervised semantic segmentation. *arXiv preprint arXiv:2012.05007*, 2020. 2

[27] Yi Li, Yiqun Duan, Zhanghui Kuang, Yimin Chen, Wayne Zhang, and Xiaomeng Li. Uncertainty estimation via response scaling for pseudo-mask noise mitigation in weakly-supervised semantic segmentation. *arXiv preprint arXiv:2112.07431*, 2021. 3

[28] Yi Li, Zhanghui Kuang, Liyang Liu, Yimin Chen, and Wayne Zhang. Pseudo-mask matters in weakly-supervised semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6964–6973, 2021. 1, 3, 7

[29] Di Lin, Jifeng Dai, Jiaya Jia, Kaiming He, and Jian Sun. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3159–3167, 2016. 1

[30] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 2, 5

[31] Yuqi Lin, Minghao Chen, Wenxiao Wang, Boxi Wu, Ke Li, Binbin Lin, Haifeng Liu, and Xiaofei He. 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*, pages 15305–15314, 2023. 3

[32] George Papandreou, Liang-Chieh Chen, Kevin P Murphy, and Alan L Yuille. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1742–1750, 2015. 1

[33] Zelin Peng, Guanchun Wang, Lingxi Xie, Dongsheng Jiang, Wei Shen, and Qi Tian. Usage: A unified seed area generation paradigm for weakly supervised semantic segmentation. *arXiv preprint arXiv:2303.07806*, 2023. 2, 7, 8

[34] Jie Qin, Jie Wu, Xuefeng Xiao, Lujun Li, and Xingang Wang. Activation modulation and recalibration scheme for weakly supervised semantic segmentation. *arXiv preprint arXiv:2112.08996*, 2021. 7

[35] Simone Rossetti, Damiano Zappia, Marta Sanzari, Marco Schaerf, and Fiora Pirri. Max pooling with vision transformers reconciles class and shape in weakly supervised semantic segmentation. In *European Conference on Computer Vision*, pages 446–463. Springer, 2022. 2

[36] Lixiang Ru, Yibing Zhan, Baosheng Yu, and Bo Du. 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*, pages 16846–16855, 2022. 3

[37] Lixiang Ru, Heliang Zheng, Yibing Zhan, and Bo Du. Token contrast for weakly-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3093–3102, 2023. 2, 3

[38] Yukun Su, Ruizhou Sun, Guosheng Lin, and Qingyao Wu. Context decoupling augmentation for weakly supervised semantic segmentation. *arXiv preprint arXiv:2103.01795*, 2021. 7

[39] Guolei Sun, Wenguan Wang, Jifeng Dai, and Luc Van Gool. Mining cross-image semantics for weakly supervised semantic segmentation. *arXiv preprint arXiv:2007.01947*, 2020. 2

[40] Kunyang Sun, Haoqing Shi, Zhengming Zhang, and Yongming Huang. 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*, pages 7283–7292, 2021. 2

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3

[42] Paul Vernaza and Manmohan Chandraker. Learning random-walk label propagation for weakly-supervised semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7158–7166, 2017. 1

[43] Changwei Wang, Rongtao Xu, Shibiao Xu, Weiliang Meng, and Xiaopeng Zhang. Treating pseudo-labels generation as image matting for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 755–765, 2023. 3

[44] Yude Wang, Jie Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12275–12284, 2020. 1

[45] Tong Wu, Junshi Huang, Guangyu Gao, Xiaoming Wei, Xiaolin Wei, Xuan Luo, and Chi Harold Liu. Embedded discriminative attention mechanism for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16765–16774, 2021. 7

[46] Tong Wu, Guangyu Gao, Junshi Huang, Xiaolin Wei, Xiaoming Wei, and Chi Harold Liu. Adaptive spatial-bce loss for weakly supervised semantic segmentation. In *European Conference on Computer Vision*, pages 199–216. Springer, 2022. 7

[47] Jinheng Xie, Xianxu Hou, Kai Ye, and Linlin Shen. Cross language image matching for weakly supervised semantic segmentation. *arXiv preprint arXiv:2203.02668*, 2022. 7

[48] Jinheng Xie, Jianfeng Xiang, Junliang Chen, Xianxu Hou, Xiaodong Zhao, and Linlin Shen. 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*, pages 989–998, 2022. 2, 3

[49] Lian Xu, Wanli Ouyang, Mohammed Bennamoun, Farid Boussaid, and Dan Xu. Multi-class token transformer for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4310–4319, 2022. 2, 3, 5, 6, 7, 8

[50] Lian Xu, Wanli Ouyang, Mohammed Bennamoun, Farid Boussaid, and Dan Xu. 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*, pages 19596–19605, 2023. 2, 3, 5, 6

[51] Sung-Hoon Yoon, Hyeokjun Kweon, Jegyeong Cho, Shinjeong Kim, and Kuk-Jin Yoon. 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*, pages 326–344. Springer Nature Switzerland Cham, 2022. 2, 5, 7, 8

[52] Bingfeng Zhang, Jimin Xiao, Yunchao Wei, Mingjie Sun, and Kaizhu Huang. Reliability does matter: An end-to-end weakly supervised semantic segmentation approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12765–12772, 2020. 1

[53] Dong Zhang, Hanwang Zhang, Jinhui Tang, Xiansheng Hua, and Qianru Sun. Causal intervention for weakly-supervised semantic segmentation. *Advances in Neural Information Processing Systems*, 2020. 6

[54] Fei Zhang, Chaochen Gu, Chenyue Zhang, and Yuchao Dai. Complementary patch for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7242–7251, 2021. 1, 2, 7

[55] Xiaolin Zhang, Yunchao Wei, Jiashi Feng, Yi Yang, and Thomas S Huang. Adversarial complementary learning for weakly supervised object localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1325–1334, 2018. 2

[56] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016. 1, 2

[57] Tianfei Zhou, Meijie Zhang, Fang Zhao, and Jianwu Li. Regional semantic contrast and aggregation for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4299–4309, 2022. 2