From SAM to CAMs: Exploring Segment Anything Model for Weakly Supervised Semantic Segmentation

Hyeokjun Kweon
KAIST
0327june@kaist.ac.kr

Kuk-Jin Yoon
KAIST
kjyoon@kaist.ac.kr

Abstract

Weakly Supervised Semantic Segmentation (WSSS) aims to learn the concept of segmentation using image-level class labels. Recent WSSS works have shown promising results by using the Segment Anything Model (SAM), a foundation model for segmentation, during the inference phase. However, we observe that these methods can still be vulnerable to the noise of class activation maps (CAMs) serving as initial seeds. As a remedy, this paper introduces From-SAM-to-CAMs (S2C), a novel WSSS framework that directly transfers the knowledge of SAM to the classifier during the training process, enhancing the quality of CAMs itself. S2C comprises SAM-segment Contrasting (SSC) and a CAM-based prompting module (CPM), which exploit SAM at the feature and logit levels, respectively. SSC performs prototype-based contrasting using SAM's automatic segmentation results. It constrains each feature to be close to the prototype of its segment and distant from prototypes of the others. Meanwhile, CPM extracts prompts from the CAM of each class and uses them to generate class-specific segmentation masks through SAM. The masks are aggregated into unified self-supervision based on the confidence score, designed to consider the reliability of both SAM and CAMs. S2C achieves a new state-of-the-art performance across all benchmarks, outperforming existing studies by significant margins. The code is available at https://github.com/sangrockEG/S2C.

1. Introduction

Semantic segmentation is a computer vision task, aiming to partition an image into semantically meaningful segments. Over the past decade, learning-based methods have achieved remarkable progress; however, they often rely on a fully supervised approach, demanding labor-intensive and costly pixel-level annotations.

As a remedy, Weakly Supervised Semantic Segmentation (WSSS) has emerged, harnessing weak yet inexpensive labels. Notably, the approach using image-level class labels [1–3, 5, 8, 14, 17, 26, 29, 34, 43, 45, 51, 52, 59, 60, 64, 65, 68, 71] have gained most attention due to their practicality. The conventional works have utilized Class Activation Maps (CAMs) [70], which identify the regions crucial for a classifier's decision-making. While CAMs offer a coarse localization of relevant objects, they usually exhibit a bias towards the most discriminative regions, resulting in incomplete activation. Furthermore, due to the lack of pixel-wise supervision, the CAMs usually have imprecise boundaries.

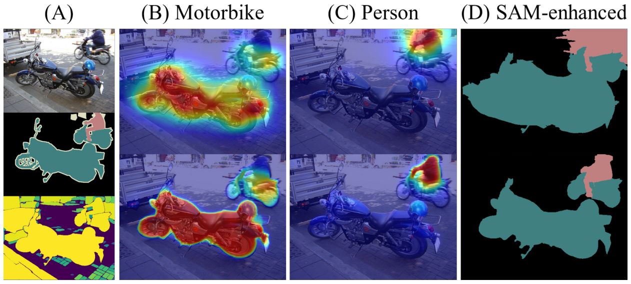

Figure 1. **(A)**: From top to bottom, the RGB image, ground truth (GT), and the segmentation result of SAM. **(B)-(D)**: The top and bottom rows represent the results from vanilla CAMs and the proposed method, respectively. For **(D)**, we apply the SAM-based post-processing method of [9]. Despite SAM's segmentation capability, the initial errors present in CAMs often cannot be rectified and may worsen in some cases. In contrast, the proposed S2C framework effectively harnesses SAM during the training phase, leading to a substantial improvement in the quality of CAMs.

The field of WSSS has endeavored to overcome these challenges, incorporating additional information sources such as off-the-shelf saliency prediction modules [19–21, 28, 49, 53], background-only external data [31], or vision-language models [37, 57, 61]. Leveraging publicly available resources, these works have achieved meaningful improvements while preserving the cost-effectiveness of WSSS.

In line with this philosophy, this paper explores the utilization of another potent information source: the Segment-Anything Model (SAM) [22], a recently introduced foundational model for promptable segmentation. Note that our focus is not merely on using SAM but on **how to use it effec-**

tively to address persistent challenges in WSSS. In this paper, we begin by examining how contemporary works have incorporated SAM for WSSS through post-processing [9] or zero-shot inference [10]. Despite promising results, these approaches primarily employ SAM during the inference phase, leaving the overall pipeline susceptible to noise in the initial seeds, *i.e*., CAMs, as in the top row of Fig. 1.

In response, we introduce a novel WSSS framework named **"from-SAM-to-CAMs (S2C)"**, aiming to enhance the quality of CAMs itself by transferring the knowledge of SAM to the classifier. S2C comprises two innovative methods: SAM-Segment Contrasting (SSC) and CAM-based Prompting Module (CPM), distilling SAM's capabilities at the feature and CAM levels, respectively.

In SSC, we segment an image using SAM's automatic segmentation option (segment-everything), creating prototypes for each segment by averaging the features of our classifier. Through contrastive learning, we enforce each feature to be close to the prototype of its segment and distant from prototypes of other segments. Meanwhile, CPM leverages SAM to directly refine the CAMs into self-supervisory signals. CPM identifies local peaks in each class's CAM, using them as point prompts to generate class-wise masks through SAM. We devise a reliability metric considering both SAM's stability score and CAM's activation score, to aggregate the masks into a unified self-supervision. Note that our novelty stems not merely from the use of SAM itself, but mainly from the way we use it, effectively transferring SAM's knowledge into the domain of WSSS.

Through comprehensive experiments, we analyze the functioning of each component in S2C, presenting extensive qualitative and quantitative results. Furthermore, we compare the proposed framework with the state-of-the-art (SoTA) WSSS methods on PASCAL VOC 2012 [16] and MS COCO [36] datasets. Our S2C significantly outperforms all the other methods on both benchmarks. The substantial semantic segmentation results achieved by S2C also can be shown in the bottom row of Fig. 1.

2. Related Work

2.1. Weakly Supervised Semantic Segmentation

The standard pipeline of WSSS is (1) training a model with image-level class labels for obtaining CAMs of each training image, (2) refining the CAMs into pseudo-labels, and (3) training a semantic segmentation model using the pseudo-labels. Some existing WSSS works aim to improve phase (2), introducing post-processing techniques such as dense conditional random field (denseCRF) [6], AffinityNet [1] or AdvCAM [29]. These strategies have demonstrated promising results and are still widely employed in WSSS. Nevertheless, these offline post-processing methods are susceptible to noisy and imprecise activations, heavily reliant on the initial CAM quality. Hence, the majority of existing WSSS studies primarily concentrate on phase (1) to enhance the quality of CAMs (seeds) themselves. Various approaches have been explored to achieve high-quality CAMs, such as cross-attention across a set of images [33, 49, 58], adversarial erasing [26, 27, 64, 69], consistency enforcement through data augmentation [52, 68], boundary-aware mechanisms [17, 23, 45, 46], adjustment of cross-entropy loss [12, 54], and the integration of Vision Transformer (ViT) architecture [47, 60].

2.2. WSSS with Additional Source of Information

While the abovementioned methods have shown promising results, they still face challenges in learning segmentation concepts without spatial supervision. To address this, numerous studies have explored the utilization of external sources of knowledge. Notably, numerous WSSS methods have incorporated off-the-shelf pre-trained saliency modules [19–21, 28, 49, 53] to distinguish salient regions. Additionally, Lee *et. al*. [31] have proposed using an external dataset to facilitate the discrimination background regions, while Kweon *et. al*. [25] leverages 3D point cloud data in a joint manner. Furthermore, recent studies [37, 57, 61] have attempted to integrate the Contrastive Language-Image Pretraining (CLIP) [44] model into WSSS. These approaches share a core strategy of leveraging the knowledge of publicly available sources to address inherent challenges in WSSS, without harming the cost-effective advantage. In line with this philosophy, this paper proposes S2C, a novel WSSS framework to fully utilize the semantic capability of SAM by directly enhancing the quality of CAMs.

3. Exploring the Use of SAM for WSSS

In this paper, we propose a novel method to integrate the SAM, a foundational model for generic promptable segmentation, into the standard pipeline of WSSS. Before delving into our approach, as a preliminary, we provide a concise overview of SAM and explore several primary methods for employing SAM within the context of WSSS.

SAM [22] is composed of three modules: 1) an image encoder for embedding an input image, 2) a prompt encoder designed to embed various types of input prompts (*e.g*., points, bounding boxes, masks, etc), and 3) a decoder utilized for predicting the mask using these embeddings. The training objective of SAM is to generate a valid mask given an input image and prompts. Here, it is noteworthy that the training process does not involve explicit semantic supervision at all. This enables SAM to attain a generalized capability, focusing on the segmentation aspect rather than being biased toward the semantic meanings of an image. Another notable feature of SAM is its automatic segmentation capability, referred to as "*segment-everything*". This

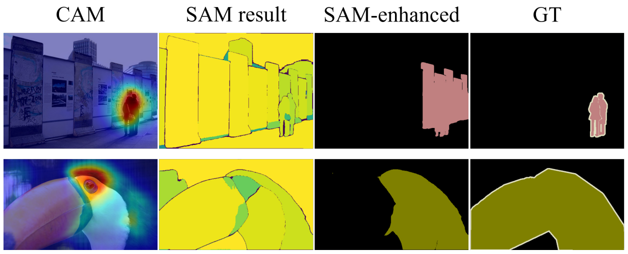

Figure 2. Limitations of inference-only use of SAM in WSSS. SAM-enhanced refers to the outcome achieved through the SAM-based post-processing method [9]. In the top row, the post-processing exacerbates errors present in CAM, leading to an even worse result. In the bottom row, the inherent ambiguity of SAM results in an over-segmentation of the image, and thereby the enhanced result is still incomplete as well as the CAM.

method feeds multiple uniformly distributed points across the image as input prompts for SAM.

Over the past few months, numerous works [66, 67] have reported SAM's ability to produce precise masks from ordinary images and to perform reasonably well, even in specialized domains like medical imagery [40, 41]. It appears that SAM truly possesses *foundational* capability to some extent, and could potentially address the long-standing challenge of WSSS. Then, how can we integrate SAM, this powerful newcomer, into the standard pipeline of WSSS?

One of the most straightforward approaches is utilizing the segment-everything option for refining the pseudo-labels obtained from CAMs [9]. In this method, each segment obtained by segment-everything is assigned to the class of the pseudo label that overlaps with the segment most. This refinement is simple yet effective, showing consistent performance gain when integrated into the existing WSSS methods. Nevertheless, the post-processing approaches are sensitive to the noises stemming from pseudo-labels, propagating the errors unintentionally as in Fig 2. Besides, the masks predicted by SAM do not always cover the entire object, due to the inherent ambiguity of promptable segmentation task [22].

Another actively studied direction is a zero-shot approach [10, 35], collaborating with language-guided models, such as CLIP [44] or Grounding-DINO [39]. In these methods, the image-level class labels in words are fed into the referring object detection model. Then, the predicted regions of each class are subsequently employed as bounding box prompts for SAM to acquire precise masks. The collaboration of two foundational models shows substantial semantic segmentation results. Nevertheless, it is still difficult to recover the wrong initial prediction of the language-based model or SAM-induced errors, similar to the aforementioned SAM-based post-processing approaches.

Specifically, the language-based models may struggle to be generalized, comprehending entirely new semantic concepts not encountered during training, or the purpose of the given task. Considering that WSSS is actively applied to specialized and unique tasks lacking labels, such as medical image segmentation [11, 18, 48, 62], these zero-shot approaches are still insufficient for practical scenarios.

In summary, both approaches have presented the remarkable potential of using SAM for WSSS. However, at the same time, they share **the main limitation of using SAM for the inference phase only**. This makes the system vulnerable to both errors stemming from SAM itself and noisy seeds (*i.e.*, CAMs) serving as input prompts. Against this background, we conclude that the most effective way of integrating SAM into WSSS is directly using it to enhance the quality of CAMs. This leads us to propose S2C, a novel WSSS method that effectively transfers the knowledge from SAM to CAMs during the training process.

4. Method

4.1. Obtaining CAMs

In this paper, we propose a novel WSSS method that effectively leverages SAM's segmentation capability for learning CAMs. We first define a classifier $\mathcal{G}$, primarily as a CAMs generator. Our classifier consists of an encoder $\mathcal{G}_E$ and a classification head $\mathcal{G}_H$. The encoder is responsible for extracting a feature map $\mathbf{F} \in \mathbb{R}^{D\times h\times w}$ from an input image $\mathbf{I} \in \mathbb{R}^{3\times H\times W}$ as follows:

$$\mathbf{F} = \mathcal{G}_E(\mathbf{I}). \tag{1}$$

Then, with the classification head $\mathcal{G}_H$, which is a 1x1 convolutional layer, we generate CAMs $\mathbf{A} \in \mathbb{R}^{C\times h\times w}$ as

$$\mathbf{A} = \mathcal{G}_H(\mathbf{F}), \tag{2}$$

where C is the number of classes and the indices are $\{1,2,\ldots,C\}$. Finally, by applying a Global Average Pooling (GAP) layer to the CAMs along the spatial axes, we obtain an image-level class prediction logit $\mathbf{y} \in \mathbb{R}^C$ as follows:

$$\mathbf{y} = GAP(\mathbf{A}). \tag{3}$$

For multi-label classification, we minimize $\mathcal{L}_{CLS}$, which is a binary cross-entropy loss as

$$\mathcal{L}_{CLS} = \ell_{bce}(\mathbf{y}, \mathbf{t}), \tag{4}$$

where $\mathbf{t}$ is an image-level classification label.

4.2. SAM-Segment Contrasting (SSC)

As we discussed in Section 3, the masks obtained by the segment-everything option of SAM can be reliable segmentation results of the given image. However, the predicted masks lack explicit semantic information since the input prompt is a set of locations without semantics. In addition, the images often include more than one object of the same

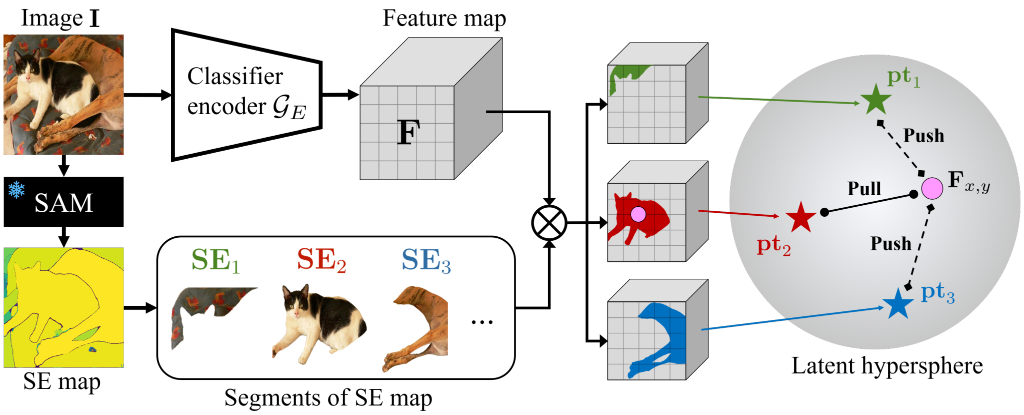

Figure 3. Visualization of the proposed SSC. First, SAM performs segment-everything to obtain a SE map of the given input image. We then group the features of the classifier encoder according to the SE map. Segment-wise prototype features (stars) are computed by averaging the features of each segment. For every feature, we perform regional prototype-based contrastive learning. Specifically, we make the feature (pink circle) close to the prototype that the feature belongs to (red star) and far from all the other prototypes (red and blue stars).

class. Besides, the grid-like distribution of the points leads to over-segmentation of the image even after the merging process. For example, objects composed of various components, such as bicycles or motorbikes, are usually segmented separately, rather than captured as a whole object in one segment. To sum up, the pixels located on the different segments of the segment-everything do not always belong to the different classes.

Therefore, instead of directly applying constraints using the segment-everything at the logit level, we focus on guiding the classifier to learn the concept of segmentation at the feature level. Specifically, we propose **SAM-Segment Contrasting (SSC)** to transfer the segmentation potential of SAM to the classifier at a **feature level**. Our SSC is mainly based on a regional prototype-based contrastive approach [38]. It aims to help the classifier understand which pixels of the image should be grouped into one segment. Specifically, in SSC, we exploit SAM from a clustering perspective, contrasting the classifier's features according to the given segments of the segment-everything.

Figure 3 illustrates the proposed SSC. To begin, we input the image $\mathbf{I}$ into SAM and utilize the segment-everything option to generate segments. Given that the predicted segments may overlap, we sort them based on area and prioritize smaller segments. Specifically, when a pixel belongs to multiple segments, we select the segment with the smallest area. This process results in a single segmentation map, referred to as the SE map throughout this paper. The ith segment of the SE map is formally denoted as $\mathbf{SE}_i$.

Simultaneously, our classifier produces a feature map F as an intermediate outcome of classification, as explained in Equ. (1). Due to the smaller spatial dimension of the feature map, we resize it using bilinear interpolation to match the size of the SE map. Subsequently, we generate a prototype for each segment $\mathbf{SE}_i$ by averaging the features of the pixels located on the segment, as follows:

$$\mathbf{pt}_i = \frac{1}{|\mathbf{SE}_i|} \sum_{(x,y)\in\mathbf{SE}_i} \mathbf{F}_{x,y}, \tag{5}$$

where $\mathbf{pt}_i$ is the prototype of the ith segment. Note that we normalize the features/prototypes along the channel dimension before and after averaging, to ensure they are constrained to lie on the unit hypersphere.

Then, we enforce each feature to be close to the prototype of the segment it belongs to and distant from the prototypes of other segments. This strategy encourages the features of pixels within one segment to form clusters, facilitating the classifiers in distinguishing them from the features of pixels in the other segments. Essentially, the proposed SSC method transfers segmentation knowledge from SAM to the classifier at the feature level.

We formalize this contrasting process with the InfoNCELoss [42] in the following manner:

$$\mathcal{L}_{SSC} = -\sum_{i=1}^{N} \sum_{(x,y)\in\mathbf{SE}_i} \frac{\mathbf{F}_{x,y}\cdot\mathbf{pt}_i/T}{\sum_{j=1}^{N}\mathbf{F}_{x,y}\cdot\mathbf{pt}_j/T}, \tag{6}$$

where N is the total number of segments in the SE map, and T is the temperature. Given the lack of pixel-wise dense GT in WSSS, we guide the feature of every pixel as in [63], instead of sampling hard negative pixels as in ReCo [38].

4.3. CAM-based Prompting Module (CPM)

CPM aims to leverage SAM's promptable segmentation capability to enhance the CAMs dynamically during training. Specifically, we utilize the CAM of each class (existing in the input image) as a prompt for SAM and subsequently obtain the corresponding class-specific mask. This mask can

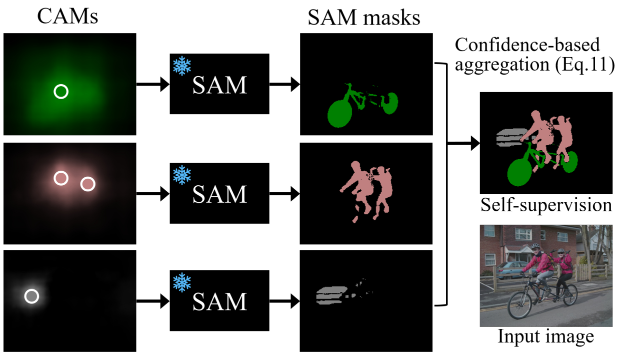

Figure 4. Visualization of the proposed **CAMs-based Prompting Module (CPM)**. From the CAMs obtained by the classifier, we extract multiple peaks (represented as colored circles). Subsequently, SAM predicts class-wise SAM masks using the peaks of each class as point prompts. Using the confidence-based aggregation, the masks are unified into a single semantic segmentation map, which serves as self-supervision for training the CAMs.

be considered a SAM-refined version of the input CAM. The masks of all classes are then aggregated into a single map, serving as self-supervision to guide the input CAMs. However, CAM is inherently a score map, which does not align with the prompts used for training SAM (*i.e.*, mask, bounding box, and point). Therefore, to fully harness the *foundational* capabilities of SAM, it is essential to convert the CAM into the prompt type suitable for SAM.

The most straightforward approach is to threshold the CAM to obtain a binary mask and use it as a mask prompt for SAM. However, we empirically observe that the officially released version of SAM performs poorly when using mask prompts. Meanwhile, utilizing box prompts also presents challenges from several aspects. First, converting the continuous score map into discrete bounding boxes is sensitive to the threshold and requires extensive tuning, which is undesirable in WSSS. Additionally, since we do not know how many objects are in the image, images potentially containing multiple objects bring about another issue. For instance, when the CAM exhibits multiple local maxima (peaks), it becomes problematic to decide whether 1) to use one large bounding box covering all peaks or 2) to use individual bounding boxes for each peak, assuming the presence of multiple objects. Please refer to the *Supplementary Material* for examples of difficulties when converting CAM into the bounding box prompts.

Therefore, we opt for a point format rather than using masks or bounding boxes to efficiently transform CAM into SAM prompts. Specifically, we use a local maximum filter LMF to extract multiple peaks from CAM as follows:

$$\mathbf{P}^c = \mathrm{LMF}(\mathbf{A}^c), \tag{7}$$

where $\mathbf{A}^c$ is the CAM of class c and $\mathbf{P}^c = \{p_1^c, \dots, p_{k_c}^c\}$ is a set of the obtained peaks. We reject the peaks with CAM scores lower than a certain threshold τ since they are often located outside the target objects. In Fig 4, we visualize the peaks as circles. Note that there can be multiple peaks, as the input image may include more than one object of the same class (person, in this case). Refer to *Supplementary Material* for the details about LMF.

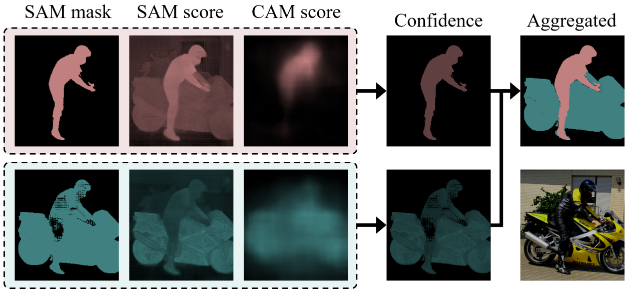

Figure 5. Illustration of the proposed aggregation approach for the CPM. As shown on the leftmost side, the class-specific masks predicted by SAM may exhibit overlaps. To consolidate these masks and assign each pixel to the correct class, we introduce a confidence score that takes into account both the stability of SAM masks and CAM scores. This aggregation method enhances the reliability of self-supervision during the training of CAMs.

We then utilize the peaks as point prompts for the SAM and obtain a refined class-specific mask as follows:

$$\mathbf{M}^c, \alpha^c_{sam} = \mathrm{SAM}(\mathbf{I}; \mathbf{P}^c). \tag{8}$$

In addition to the SAM mask $\mathbf{M}^c$, we also obtain the stability of the SAM mask α^c_{sam} to preserve the reliability of each pixel. Here, the pixels with higher stability scores are highly probable (*i.e.*, more confident) to be segmented along the given prompt. We perform the above process for every class existing in the input image. For efficiency, we run the encoder of SAM only once and iteratively use the obtained embedding for decoding with class-wise point prompts.

Finally, we aggregate the obtained class-specific masks into a single semantic segmentation map. However, as illustrated in Fig. 5, these masks are often not exclusive. The overlaps are mainly due to the erroneous activation of CAMs or SAM ambiguity, both challenging to mitigate directly. As a remedy, we introduce a novel confidence-based aggregation method to accurately assign each pixel to the appropriate class, even when the masks overlap.

In this aggregation method, we consider the reliability of both the SAM masks and CAMs of the existing classes, We use the obtained SAM stability α^c_{sam} as a reliability score map for the SAM mask. On the other hand, for the CAMs, we average the activation of the CAM of each class on the SAM mask of that class as follows:

$$\alpha^c_{cam} = \frac{1}{|\mathbf{M}^c|} \sum_{(x,y)\in\mathbf{M}^c} \mathbf{A}^c_{x,y}, \tag{9}$$

where α^c_{cam} denotes the reliability score map of the CAM. Subsequently, the proposed confidence is defined as

$$\alpha^c = \alpha^c_{sam} \otimes \alpha^c_{cam}, \tag{10}$$

where $\otimes$ denotes element-wise product.

The unified segmentation map $\hat{\mathbf{S}}$ is then acquired by

$$\hat{\mathbf{S}}_{x,y} = \begin{cases} 0 & \text{if } \max_c \alpha^c_{x,y} < \tau \\ \text{argmax}_c \, \alpha^c_{x,y} & \text{otherwise,} \end{cases} \tag{11}$$

where x, y is the pixel coordinate and τ is the threshold for discriminating background regions. Its value is the same as the one used in the local maximum filter.

The obtained $\hat{\mathbf{S}}$ is a self-supervisory signal for training the CAMs. Unlike CAMs defined for C foreground classes only, $\hat{S}$ inherently includes the background class (index 0, in our implementation). To involve the concept of background in training the CAMs, we define the background score map $\mathbf{A}^0$ using CAMs as follows:

$$\mathbf{A}^0_{i,j} = 1 - \max_c \mathbf{A}^c_{i,j}. \tag{12}$$

Then, we concatenate $\mathbf{A}^0$ with the original CAMs to obtain $\mathbf{A}^{+0}$, score maps regarding both background and foreground classes. We define the loss function for CPM as a standard cross-entropy loss between $\mathbf{A}^{+0}$ and $\hat{\mathbf{S}}$ as

$$\mathcal{L}_{CPM} = \ell_{ce}(\mathbf{A}^{+0}, \hat{\mathbf{S}}). \tag{13}$$

Finally, the overall loss function of the proposed S2C framework can be summed up as follows:

$$\mathcal{L}_{S2C} = \mathcal{L}_{CLS} + \mathcal{L}_{SSC} + \mathcal{L}_{CPM}. \tag{14}$$

5. Experimental Results

5.1. Dataset and Evaluation Metric

We evaluate the proposed method on PASCAL VOC 2012 [16] and MS-COCO [36] datasets. The PASCAL VOC dataset comprises 1464/1449/1456 images for the *train*/*val*/*test* sets, respectively, and includes 20 foreground classes and a background class. Following the standard practice of conventional WSSS studies, we adopt the additionally augmented dataset for training our models. It consists of 10582 images along with image-level classification labels. On the other hand, MS-COCO dataset contains around 80k/40k images for the *train*/*val* sets, respectively, with 80 foreground classes and a background class. As an evaluation metric, we use the mean Intersection over Union (mIoU) between the prediction and the GT.

5.2. Implementation Details

We use ResNet38 [56] as the feature encoder of our classifier, followed by a 1×1 convolution layer as the classification head to generate CAMs. For the final semantic segmentation model, we use Deeplab [4] with ResNet38 backbone as in [1, 26, 34, 50, 60, 64, 68]. For SAM [22], we utilize the officially released version with ViT-H model.

As data augmentation, we apply horizontal flipping, random cropping/resizing, and color jittering [24] to the input image. We use a poly learning rate [7], which multiplies $(1 - \frac{iter}{max\ iter})^{0.9}$ to the initial learning rate (0.02). For PASCAL, the classifier is trained for six epochs, which took around 8 hours with two Tesla V100s. As the classification accuracy is low at the initial stage of training, the sampled query from CPM can be erroneous. Therefore, we do not apply $\mathcal{L}_{CPM}$ on the first epoch. We set both τ and T to 0.5 by default. Additional details can be found in the *Supplementary Material*.

Table 1. Results from ablations studies on SSC and CPM. The baseline is a setting with mere image-level classification loss only. SSC (logit-level) and SSC (feature-level) denote that the contrasting performed at the logit-level and feature-level, respectively. **Bold** number represents the best result.

Baseline	SSC (logit-level)	SSC (feature-level)	CPM	mIoU (%)
✓				50.1
✓	✓			57.3
✓		✓		63.2
✓		✓	✓	**76.1**

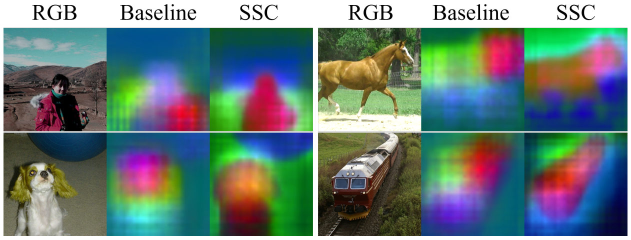

Figure 6. Qualitative comparison between the features of the baseline classifier and those of our model with SSC.

5.3. Ablation Studies

Comparative analysis in Table 1 demonstrates the significant performance improvement by SSC compared to the baseline, which solely relies on a classification loss. The result also confirms that the proposed feature-level contrasting strategy is notably more effective than the naive logit-level contrasting approach. As explained in Section 4.2, when the input image is over-segmented in the SE map, the pixels of the different segments do not always belong to different classes. Therefore, as a more flexible criterion, we perform contrast at the feature level rather than directly utilizing pixel-wise classification logits for contrast. Finally, CPM contributes further to enhancing the quality of CAMs, resulting in a remarkable increase in mIoU.

To provide a more intuitive grasp of SSC's functionality, we present a comparison of classifier features between the baseline classifier (without SSC) and ours (with SSC)

Table 2. Results from ablations studies on the components within CPM. **Multi-peaks**: sampling multiple local peaks from the CAM in Equ. 7. **Aggregation SAM&CAM**: using SAM stability score and CAM activation score for confidence score in Equ. 10, respectively. **Using BG**: incorporating background score map in Equ. 12.

Multi-peaks	Aggregation		Using BG	mIoU
	SAM	CAM		(%)
	✓	✓	✓	73.0
✓			✓	72.7
✓	✓		✓	73.2
✓	✓	✓		74.9
✓	✓	✓	✓	**76.1**

in Fig. 6. We apply Principal Component Analysis (PCA) to reduce the channel dimension of the features from 256 to 3. Subsequently, these channels are used to represent RGB colors for visualization. As shown in Fig. 6, the features with SSC are precisely aligned along objects, while the baseline classifier's features exhibit indistinct and coarse boundaries. The results strongly support the effectiveness of SSC in enabling the classifier to comprehend the concept of segmentation, leading to better CAMs.

In Table 2, we analyze various components within CPM. Initially, we experiment with extracting a single global peak from the CAM of each class in Equ. 7, instead of sampling multiple local peaks originally. However, this configuration results in a performance drop (-3.1%). This implies that it is essential to extract multiple peaks per CAM, particularly for images containing multiple objects.

Secondly, we validate the effectiveness of the proposed score-based aggregation strategy by ablating the use of the SAM stability score and CAM activation score in Equ. 10. The results show that both SAM stability and CAM activation scores meaningfully contribute to the reliable aggregation process, resulting in a high-quality self-supervisory signal within CPM.

Finally, we confirm the advantage of including a background score map in Equ. 12. To verify this, we obtain class-wise binary masks from the unified self-supervision and then minimize the L1 loss between the CAMs and these masks. While this setting yields substantial CAMs, the results indicate that incorporating BG is even more beneficial.

5.4. Comparisons to State-of-The-Arts

In Table 3, we present a comparative analysis of the CAMs obtained by conventional methods and our S2C. Additionally, we conduct a comparison of the masks derived from these CAMs, which serve as pseudo-labels for training a semantic segmentation model. The masks are obtained through post-processing following the standard protocols.

The results demonstrate a significant performance improvement in our CAMs over the CAMs of the conventional methods. This improvement is maintained through the post-processing step, resulting in the masks of substantial quality. Specifically, our S2C achieves a performance gain of more than 5% and 6% at the CAM and mask levels, respectively, surpassing the previous SoTAs.

Table 3. Comparisons between the proposed framework and the conventional WSSS methods. We evaluate mIoU (%) performance on the PASCAL VOC 2012 *train* set at CAMs and Mask (pseudo-label) levels. **Bold** numbers represent the best results.

Methods	Backbone	CAMs	Mask
W-OoD [30]$_{CVPR22}$	RN50	53.3	-
ReCAM [13]$_{CVPR22}$	RN50	54.8	70.5
SIPE [8]$_{CVPR22}$	RN50	58.6	-
MCT [60] $_{CVPR22}$	ViT	61.7	69.1
PPC [15]$_{CVPR22}$	WRN38	61.5	70.1
Spatial-BCE [55]$_{ECCV22}$	WRN38	68.1	70.4
AEFT [64] $_{ECCV22}$	WRN38	56.0	71.0
ACR [27] $_{CVPR23}$	ViT	65.5	70.9
BECO [45] $_{CVPR23}$	RN101	65.5	70.9
USAGE [43] $_{ICCV23}$	WRN38	67.7	72.8
MAT [51]$_{ICCV23}$	WRN38	62.3	72.9
CLIMS [58]$_{CVPR22}$	WRN38	56.6	70.5
Xu *et. al.* [58]$_{CVPR23}$	ViT	66.3	-
CLIP-ES [37]$_{CVPR23}$	RN101	70.8	75.0
S2C (Ours)	WRN38	**76.1**	**81.7**

Furthermore, we evaluate the performance of semantic segmentation models trained using pseudo-labels from each method, as shown in Table 4. The proposed S2C demonstrates superior performance, surpassing the other methods by a significant margin, around 5% across all benchmarks. These results strongly support the effectiveness of our approach in integrating SAM into the WSSS pipeline.

We acknowledge that the comparisons may not be entirely fair, given that the proposed method leverages SAM as a powerful source of information, unlike previous methods. To address this, we conduct additional comparisons between our framework and existing works enhanced by [9]. Table 5 shows that, even after SAM-based post-processing, the masks generated by our method still outperform those of other approaches. These results suggest that SAM's segmentation capability is effectively transferred to our classifier within the S2C framework, proving more advantageous than using SAM solely for inference.

Finally, Fig. 7 presents the qualitative comparisons with the baseline. In contrast to baseline CAMs, the CAMs generated by S2C exhibit precise boundaries while effectively covering the entire object. Notably, our framework excels even in challenging scenarios, such as thin objects (*e.g.*, horse legs) or widely distributed small objects (*e.g.*, cows). The semantic segmentation network trained by our high-quality pseudo-labels yields remarkable results, accurately capturing fine details without semantic confusion. Additional results can be found in *Supplementary Material*.

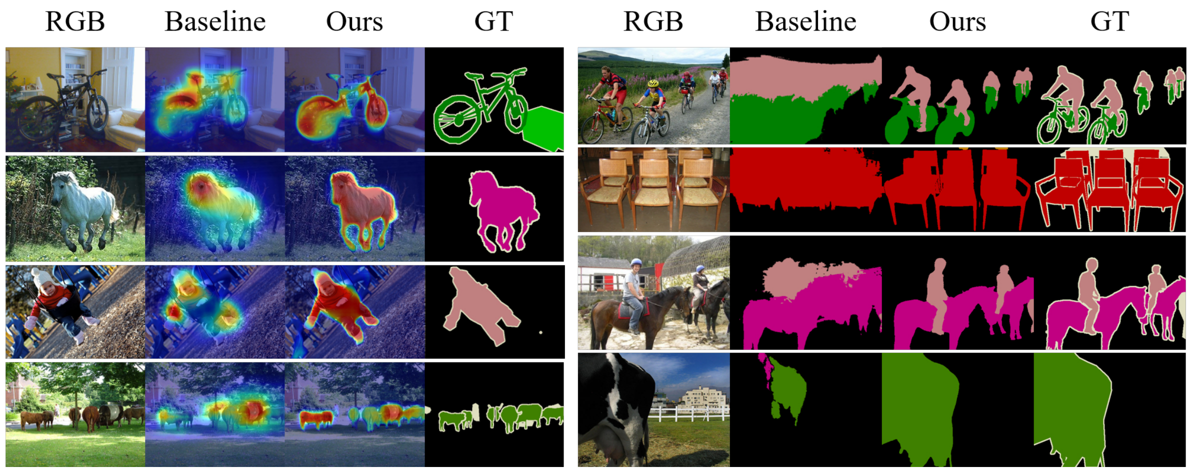

Figure 7. Qualitative comparison between the results of the baseline and the proposed method on PASCAL VOC 2012 dataset. **Left**: CAMs of the vanilla classifier and ours. **Right**: semantic segmentation results of AffinityNet [1] and ours. Our CAMs not only activate the entire regions of each object but also present precise boundaries, ultimately achieving remarkable semantic segmentation results.

Table 4. Comparison in mIoU (%) performance between ours and the existing WSSS methods. We evaluate the methods on both the PASCAL VOC 2012 and MS-COCO 2014 datasets.

Methods	*Backbone*	VOC *val*	VOC *test*	COCO *val*
AffinityNet [1]$_{CVPR18}$	WRN38	61.7	63.7	-
IRNet [2]$_{CVPR19}$	RN50	63.5	64.8	41.4
W-OoD [30]$_{CVPR22}$	WRN38	70.7	70.1	-
MCT [60] $_{CVPR22}$	WRN38	71.9	71.6	42.0
ReCAM [13]$_{CVPR22}$	RN101	68.5	68.4	42.9
AEFT [64]$_{ECCV22}$	WRN38	70.9	71.7	44.8
OCR [14]$_{CVPR23}$	ViT	72.7	72.0	42.5
ACR [27]$_{CVPR23}$	WRN38	72.4	72.4	45.3
BECO [45]$_{CVPR23}$	RN101	72.1	71.8	45.1
USAGE [43] $_{ICCV23}$	WRN38	71.9	72.8	44.3
FPR [5] $_{ICCV23}$	WRN38	70.0	70.6	43.9
MAT [51] $_{ICCV23}$	RN101	73.0	72.7	45.6
CLIMS [58]$_{CVPR22}$	RN50	69.3	68.7	-
CLIP-ES [37]$_{CVPR23}$	RN101	71.1	71.4	45.4
Xu *et. al.* [61]$_{CVPR23}$	WRN38	72.2	72.2	45.9
S2C (Ours)	WRN38	**78.2**	**77.5**	**49.8**

Table 5. Comparisons between our framework and the existing WSSS methods enhanced by SAM-based post-processing [9]. We compare the quality of the pseudo-labels on the PASCAL VOC 2012 *train* set. **VOC *val*** denotes the performance of the final segmentation model trained by the pseudo-labels.

Methods	Backbone	Pseudo-label	VOC *val*
TransCAM [32]	Conf.-S	75.2	69.9
SIPE [8]	RN50	73.8	69.7
L2G [21]	WRN38	77.8	72.4
CLIMS [58]	RN50	75.2	71.1
CLIP-ES [37]	RN101	79.7	73.1
Baseline	WRN38	68.6	65.2
S2C (Ours)	WRN38	**81.7**	**78.2**

6. Conclusion

This paper addresses the challenges in WSSS by leveraging the recently published foundation model for segmentation, SAM, to enhance the quality of CAMs during the training process. Our main novelty lies in how we effectively transfer the knowledge of SAM into WSSS. While recent works have demonstrated promising results using SAM during inference, either as post-processing or in a zero-shot manner, we have identified their potential vulnerabilities to noise in CAMs used as initial seeds. To address this, we propose the S2C framework, facilitating direct knowledge transfer from SAM to the classifier, thereby improving the quality of CAMs. Within S2C, SSC performs prototype-based contrasting with SAM's automatic segmentation results, guiding features to be close to their segment prototype and distant from others. Simultaneously, CPM extracts prompts from CAMs, generating class-specific segmentation masks through SAM. These masks are aggregated into unified self-supervision, guided by a novel reliability score considering both SAM and CAM confidence. Through extensive experiments, we demonstrate the working logic of the proposed S2C in both qualitative and quantitative manners. Furthermore, S2C establishes a new state-of-the-art performance across all benchmarks, surpassing existing studies by significant margins. We believe that our approach pioneers the effective utilization of SAM in the domain of WSSS.

Acknowledgements This research was supported by National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (NRF2022R1A2B5B03002636) and the Challengeable Future Defense Technology Research and Development Program through the Agency For Defense Development (ADD) funded by the Defense Acquisition Program Administration (DAPA) in 2024 (No.912768601).

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, 6, 8

[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. 8

[3] 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

[4] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. In *3rd International Conference on Learning Representations, ICLR*, 2015. 6

[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. 1, 8

[6] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*, 2014. 2

[7] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 6

[8] 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. 1, 7, 8

[9] Tianle Chen, Zheda Mai, Ruiwen Li, and Wei-lun Chao. Segment anything model (sam) enhanced pseudo labels for weakly supervised semantic segmentation. *arXiv preprint arXiv:2305.05803*, 2023. 1, 2, 3, 7, 8

[10] Zhaozheng Chen and Qianru Sun. Weakly-supervised semantic segmentation with image-level labels: from traditional models to foundation models. *arXiv preprint arXiv:2310.13026*, 2023. 2, 3

[11] Zhang Chen, Zhiqiang Tian, Jihua Zhu, Ce Li, and Shaoyi Du. C-cam: Causal cam for weakly supervised semantic segmentation on medical image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11676–11685, 2022. 3

[12] 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. 2

[13] 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

[14] Zesen Cheng, Pengchong Qiao, Kehan Li, Siheng Li, Pengxu Wei, Xiangyang Ji, Li Yuan, Chang Liu, and Jie Chen. Out-of-candidate rectification for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23673–23684, 2023. 1, 8

[15] 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

[16] 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, 6

[17] 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

[18] Zijie Fang, Yang Chen, Yifeng Wang, Zhi Wang, Xiangyang Ji, and Yongbing Zhang. Weakly-supervised semantic segmentation for histopathology images based on dataset synthesis and feature consistency constraint. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 606–613, 2023. 3

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1, 2

[20] Qibin Hou, PengTao Jiang, Yunchao Wei, and Ming-Ming Cheng. Self-erasing network for integral object attention. In *Advances in Neural Information Processing Systems*, pages 549–559, 2018.

[21] 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. 1, 2, 8

[22] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 1, 2, 3, 6

[23] Alexander Kolesnikov and Christoph H Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *European conference on computer vision*, pages 695–711. Springer, 2016. 2

[24] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017. 6

[25] Hyeokjun Kweon and Kuk-Jin Yoon. Joint learning of 2d-3d weakly supervised semantic segmentation. *Advances in Neural Information Processing Systems*, 35:30499–30511, 2022. 2

[26] 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, 6

[27] 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, 7, 8

[28] Jungbeom Lee, Eunji Kim, Sungmin Lee, Jangho Lee, and Sungroh Yoon. Ficklenet: Weakly and semi-supervised semantic image segmentation using stochastic inference. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5267–5276, 2019. 1, 2

[29] 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

[30] 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. 7, 8

[31] Seungho Lee, Minhyun Lee, Jongwuk Lee, and Hyunjung Shim. Railroad is not a train: Saliency as pseudo-pixel supervision for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5495–5505, 2021. 1, 2

[32] Ruiwen Li, Zheda Mai, Zhibo Zhang, Jongseong Jang, and Scott Sanner. Transcam: Transformer attention-based cam refinement for weakly supervised semantic segmentation. *Journal of Visual Communication and Image Representation*, 92:103800, 2023. 8

[33] 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

[34] 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, 6

[35] Yi Li, Hualiang Wang, Yiqun Duan, and Xiaomeng Li. Clip surgery for better explainability with enhancement in open-vocabulary tasks. *arXiv preprint arXiv:2304.05653*, 2023. 3

[36] 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, 6

[37] 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. 1, 2, 7, 8

[38] Shikun Liu, Shuaifeng Zhi, Edward Johns, and Andrew J Davison. Bootstrapping semantic segmentation with regional contrast. *arXiv preprint arXiv:2104.04465*, 2021. 4

[39] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 3

[40] Jun Ma and Bo Wang. Segment anything in medical images. *arXiv preprint arXiv:2304.12306*, 2023. 3

[41] Maciej A Mazurowski, Haoyu Dong, Hanxue Gu, Jichen Yang, Nicholas Konz, and Yixin Zhang. Segment anything model for medical image analysis: an experimental study. *Medical Image Analysis*, 89:102918, 2023. 3

[42] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 4

[43] 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. 1, 7, 8

[44] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 3

[45] Shenghai Rong, Bohai Tu, Zilei Wang, and Junjie Li. Boundary-enhanced co-training for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19574–19584, 2023. 1, 2, 7, 8

[46] Shenghai Rong, Bohai Tu, Zilei Wang, and Junjie Li. Boundary-enhanced co-training for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19574–19584, 2023. 2

[47] 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

[48] Holger R Roth, Dong Yang, Ziyue Xu, Xiaosong Wang, and Daguang Xu. Going to extremes: weakly supervised medical image segmentation. *Machine Learning and Knowledge Extraction*, 3(2):507–524, 2021. 3

[49] 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. 1, 2

[50] 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. 6

[51] 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. 1, 7, 8

[52] 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, 2

[53] 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. 1, 2

[54] 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. 2

[55] 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

[56] Zifeng Wu, Chunhua Shen, and Anton Van Den Hengel. Wider or deeper: Revisiting the resnet model for visual recognition. *Pattern Recognition*, 90:119–133, 2019. 6

[57] Jinheng Xie, Xianxu Hou, Kai Ye, and Linlin Shen. Clims: Cross language image matching for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4483–4492, 2022. 1, 2

[58] Jinheng Xie, Xianxu Hou, Kai Ye, and Linlin Shen. Cross language image matching for weakly supervised semantic segmentation. *arXiv preprint arXiv:2203.02668*, 2022. 2, 7, 8

[59] 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. 1

[60] 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. 1, 2, 6, 7, 8

[61] 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. 1, 2, 8

[62] Guanyu Yang, Chuanxia Wang, Jian Yang, Yang Chen, Lijun Tang, Pengfei Shao, Jean-Louis Dillenseger, Huazhong Shu, and Limin Luo. Weakly-supervised convolutional neural networks of renal tumor segmentation in abdominal cta images. *BMC medical imaging*, 20:1–12, 2020. 3

[63] Sung-Hoon Yoon, Hyeokjun Kweon, Jaeseok Jeong, Hyeonseong Kim, Shinjeong Kim, and Kuk-Jin Yoon. Exploring pixel-level self-supervision for weakly supervised semantic segmentation. *arXiv preprint arXiv:2112.05351*, 2021. 4

[64] 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. 1, 2, 6, 7, 8

[65] 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

[66] Chunhui Zhang, Li Liu, Yawen Cui, Guanjie Huang, Weilin Lin, Yiqian Yang, and Yuehong Hu. A comprehensive survey on segment anything model for vision and beyond. *arXiv preprint arXiv:2305.08196*, 2023. 3

[67] Chaoning Zhang, Sheng Zheng, Chenghao Li, Yu Qiao, Taegoo Kang, Xinru Shan, Chenshuang Zhang, Caiyan Qin, Francois Rameau, Sung-Ho Bae, et al. A survey on segment anything model (sam): Vision foundation model meets prompt engineering. *arXiv preprint arXiv:2306.06211*, 2023. 3

[68] 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, 6

[69] 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

[70] 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

[71] 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. 1