Progressive Feature Self-Reinforcement for Weakly Supervised Semantic Segmentation

Jingxuan He[1], Lechao Cheng[1*], Chaowei Fang[2], Zunlei Feng[3], Tingting Mu[4], Mingli Song[3]

[1]Zhejiang Lab,
[2]Xidian University,
[3]Zhejiang University,
[4]University of Manchester
22021025@zju.edu.cn, chenglc@zhejianglab.com, chaoweifang@outlook.com, zunleifeng@zju.edu.cn,
tingting.mu@manchester.ac.uk. brooksong@ziu.edu.cn

Abstract

Compared to conventional semantic segmentation with pixel-level supervision, weakly supervised semantic segmentation (WSSS) with image-level labels poses the challenge that it commonly focuses on the most discriminative regions, resulting in a disparity between weakly and fully supervision scenarios. A typical manifestation is the diminished precision on object boundaries, leading to deteriorated accuracy of WSSS. To alleviate this issue, we propose to adaptively partition the image content into certain regions (e.g., confident foreground and background) and uncertain regions (e.g., object boundaries and misclassified categories) for separate processing. For uncertain cues, we propose an adaptive masking strategy and seek to recover the local information with self-distilled knowledge. We further assume that confident regions should be robust enough to preserve the global semantics, and introduce a complementary self-distillation method that constrains semantic consistency between confident regions and an augmented view with the same class labels. Extensive experiments conducted on PASCAL VOC 2012 and MS COCO 2014 demonstrate that our proposed single-stage approach for WSSS not only outperforms state-of-the-art counterparts but also surpasses multi-stage methods that trade complexity for accuracy.

Introduction

Weakly supervised semantic segmentation (WSSS) reduces the cost of annotating "strong" pixel-level labels by using "weak" labels such as bounding boxes (Dai, He, and Sun 2015; Song et al. 2019), scribbles (Lin et al. 2016; Vernaza and Chandraker 2017), points (Bearman et al. 2016; Su et al. 2022) and image-level class labels (Araslanov and Roth 2020; Ru et al. 2022; Wu et al. 2023; Ru et al. 2023). Among these, image-level class labels are the most affordable, but challenging to exploit. A commonly used WSSS approach based on image-level class labels typically includes the following steps: (1) to train a neural network for image classification; (2) to use the network to generate class activation maps (CAMs) (Zhou et al. 2016) as seed regions; (3) to refine the CAMs to pseudo segmentation labels that will be used as the ground truth for supervising a segmentation network. These steps can either be implemented as separate stages or as a single collaborative stage, and single-stage frameworks are usually more efficient as they streamline the training pipeline. In general, high-quality pseudo labels lead to superior semantic segmentation performance. In this work, we focus on developing an effective single-stage approach to generate more accurate pseudo labels from image-level class labels.

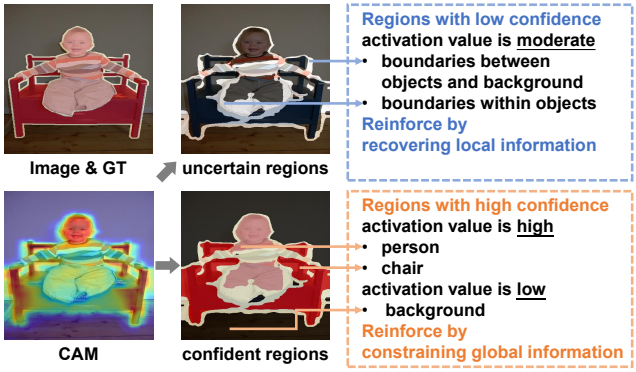

Figure 1: Visualization of our main idea. The flawed CAM only identifies discriminative regions. To address this, we propose to partition the image into uncertain regions (e.g., object boundaries) and confident regions (e.g., the main body of an object) and reinforce features of these regions in a complementary manner.

Unfortunately, CAMs are essentially flawed because they are intended for classification, i.e., they strive to identify the most discriminative regions of an object aiming at improved classification accuracy. To tackle this, one can improve the initial seeds (Lee et al. 2019; Wang et al. 2020) or refine pseudo labels (Ahn, Cho, and Kwak 2019; Chen et al. 2020), by expanding activations or labels to semantically consistent pixels in the neighborhood. Recent studies have found that the restricted receptive field of convolution negatively affects the recognition of integral objects (Ru et al. 2022, 2023) and use vision transformer (Dosovitskiy et al. 2020) to model the global relationships for improve-

*corresponding author.

ment. But this does not resolve the issue of CAM seeds or pseudo labels, and we still observe empirically high uncertainty in (1) boundary regions between foreground objects and background, and (2) misclassified regions within multiple semantically-different objects. In the example of Figure 1, the generated CAM is uncertain about the two arms of the *person* on the *chair*, also the boundary between the foreground (*person* and *chair*) and the background is unclear. These uncertain regions are easily confused by obscure semantic clues due to the absence of pixel-level supervision.

Our goal is to clarify the visual semantics of uncertain regions mentioned above. We emphasize that the local visual patterns should be explicitly modeled and captured. As can be seen from Figure 1, *head* and *upper thighs* are well recognized, while the recognition of *arms* and *lower legs* needs improvement. A better understanding of that *arms and lower legs surely belong to a person* should be established using local visual context. Although some methods can deal with noisy object boundaries by employing off-the-shelf saliency detection models for rich object contours (Lee et al. 2021b; Li, Fan, and Zhang 2022), they overlook uncertain regions caused by low confidence within objects. Alternatively, it has been proposed to attain the training objective using knowledge gathered from the past training iterations, i.e., self-distillation (Caron et al. 2021). Encouraged by the success of self-distillation, we discard saliency detection models, but take advantage of the strategy of self-distillation in our model training.

To this end, to explore and strengthen semantics over uncertain regions, we propose a progressive self-reinforcement method. To distinguish uncertain regions from confident ones, we define those with intermediate CAM scores as uncertain regions, since a very low/high score strongly indicates the background/foreground. Specifically, we propose to mask uncertain features (equivalent to image patch tokens) and learn to recover the original information with the help of an online momentum teacher. This masking strategy aligns with a state-of-the-art pre-training paradigm called masked image modeling (MIM) that brings locality inductive bias to the model (Xie et al. 2023). We upgrade its random masking strategy with semantic uncertainty so that the network can focus on uncertain features controlled by the masking ratio. This design is beneficial to facilitate features in both object boundaries and misclassified regions. Assuming that confident features should be robust enough to present global semantics, we also introduce a complementary method that constrains semantic consistency between two augmented views with the same class labels. Our proposal can be seamlessly integrated into a vision transformer based single-stage WSSS framework. We summarize our main contributions as follows :

- We propose a novel WSSS approach, *progressive feature self-reinforcement*, to effectively enhance the semantics of uncertain regions. The investigation of uncertain regions, including both object boundaries and misclassified categories, significantly improves WSSS performance.
- We design an adaptive masking strategy to identify uncertain regions. Unlike most previous works that adopt additional saliency detection models, we locate uncertain regions with the guidance of semantic-aware CAMs.
- Exhaustive experiments on PASCAL VOC 2012 (Everingham et al. 2010) and MS COCO 2014 (Lin et al. 2014) show that our method outperforms SOTA single-stage competitors, even better than existing sophisticated multi-stage methods.

Related Work

Weakly Supervised Semantic Segmentation

Multi-stage WSSS methods adopt a classification model to generate CAMs as pseudo labels, then train a segmentation model for evaluating the final performance. To overcome the commonly acknowledged weakness that CAMs can only focus on discriminative regions, several works aim at improving the training dynamic by erasing and seeking (Hou et al. 2018) or adversarial learning (Yoon et al. 2022). Some recent approaches also adopt vision transformer (Dosovitskiy et al. 2020) for the WSSS task, considering its favorable long-range modeling capability. TS-CAM (Gao et al. 2021) proposes to couple class-agnostic attention maps with semantic-aware patch tokens to promote object localization. MCTformer (Xu et al. 2022) introduces multiple class tokens so that class-specific attention maps can be generated. Other approaches incorporate extra data into training or post-processing, e.g., saliency maps (Lee et al. 2021b) or contrastive language-image pre-training (CLIP) models (Lin et al. 2023). Our solution aims at improving pseudo labels as well, but it is integrated into a single-stage framework for simplicity, and it requires neither extra data nor off-the-shelf saliency detection models.

Single-stage WSSS methods treat multiple stages such as classification, pseudo label refinement, segmentation as a whole and perform joint training. 1Stage (Araslanov and Roth 2020) achieves comparable performance with dominant multi-stage approaches by ensuring local consistency, semantic fidelity and mask completeness. AFA (Ru et al. 2022) explores the intrinsic architecture of ViT and derives reliable semantic affinity from multi-head self-attention for pseudo label refinement. ToCo (Ru et al. 2023) tackles the issue of over-smoothing observed in ViT by supervising the final patch tokens with intermediate knowledge. Despite the simplified and streamlined training procedure, single-stage methods often suffer from inferior performance compared with multi-stage ones. In this work, we achieve superior semantic segmentation results using a single-stage framework by discovering and reinforcing underlying semantic layouts.

Self-Distillation

Self-distillation associates self-supervised learning (He et al. 2020) with knowledge distillation (Hinton, Vinyals, and Dean 2015), where knowledge is transferred and learned without resorting to any labels. It is primarily designed to compress large networks, and is hoping to promote performance on downstream tasks via mimicking the output of a frozen teacher (Noroozi et al. 2018; Zhang et al. 2023; Wang et al. 2023). Recently, some approaches (Caron et al. 2021;

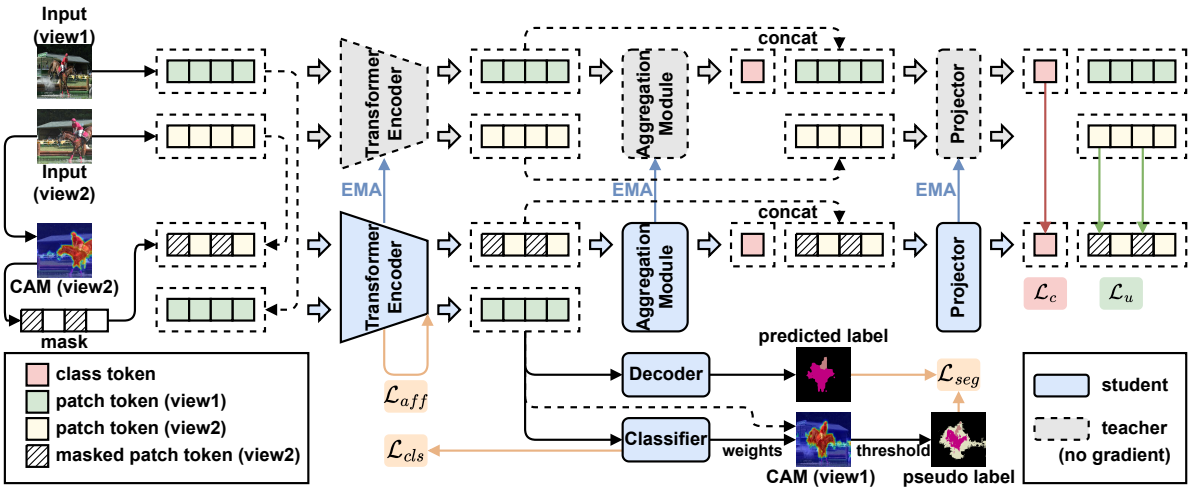

Figure 2: Overview of our framework. For the student pipeline, we forward one view through the encoder, and the encoded patch tokens are fed into the classifier for classification and the decoder for semantic segmentation, separately. The other view is masked and sequentially forwarded through the encoder, the aggregation module, and the projector. For the teacher pipeline, both views are propagated through the encoder, the aggregation module, and the projector. The teacher network requires no gradient and is an exponential moving average (EMA) of the student network.

Zhou et al. 2021) build the teacher dynamically during training, where the teacher adopts the same architecture as that of the student and is updated with the knowledge of past iterations. The resulting framework simplifies the training process and achieves compelling results compared with other self-training frameworks. This motivates us to adapt the core idea of self-distillation to the WSSS task for the purpose of rectifying inaccurate object boundaries as well as improving discriminative object features.

Methodology

A Single-Stage Framework for WSSS

The proposed single-stage framework for WSSS is illustrated in Figure 2. We use an encoder-decoder architecture to accomplish semantic segmentation with image-level supervision. The encoder is a vision transformer supervised by image-level class labels. We adopt patch token contrast (PTC) (Ru et al. 2023) for affinity learning as it is crucial to constrain affinities between patch tokens of the last layer against over-smoothing (Gong et al. 2021). As for semantic segmentation, we borrow a lightweight convolutional decoder from DeepLab (Chen et al. 2017), which is supervised by pseudo segmentation labels that are generated from CAMs. An aggregation module is used to summarize patch tokens into one class token and an MLP-based projector to transform all tokens into an appropriate feature space for feature learning. To improve model training, we enable a student and a teacher pipeline to achieve self-distillation.

Formally, let $\mathcal{F}$ be the transformer encoder with its output embedding dimension denoted by D, $\mathcal{P}$ the projector, $\mathcal{M}$ the masking operator, and $\mathcal{A}$ the aggregating operator. We start from explaining the student pipeline. An input image is randomly augmented to two distorted views: x_1 and x_2. Each view is subsequently divided into HW non-overlapping patch tokens, denoted as $T_1 = \left\{t_1^{(i)}\right\}_{i=1}^{HW}$ and $T_2 = \left\{t_2^{(i)}\right\}_{i=1}^{HW}$, respectively. We forward T_1 into the encoder to obtain the logits $Z_1 = \mathcal{F}(T_1) \in \mathbb{R}^{HW \times D}$, which are then fed into the classifier for classification, and also the decoder for segmentation, following the standard image classification and segmentation setup. To reinforce features, we divide T_2 into uncertain and confident tokens and mask the uncertain ones with learnable parameters, for which the uncertain token selection and masking approaches will be explained later. The resulting masked view, denoted as $\hat{T}_2 = \mathcal{M}(T_2)$, is also fed into the encoder to obtain $\hat{Z}_2 = \mathcal{F}\left(\hat{T}_2\right)$. Embeddings of the unmasked confident tokens in $\hat{Z}_2$ are summarized into a class token by an aggregation module, denoted by $\mathcal{A}\left(\hat{Z}_2\right) \in \mathbb{R}^{1 \times D}$. This class token is concatenated with $\hat{Z}_2$, and further projected and normalized to resemble probabilities distributions in $\hat{P}_2 \in \mathbb{R}^{(1+HW) \times D}$, as

$$\hat{P}_2 = \sigma\left(\mathcal{P}\left(\left[\mathcal{A}\left(\hat{Z}_2\right); \hat{Z}_2\right]\right)\right), \tag{1}$$

where σ is the row-wise softmax function, and $[;]$ the concatenation. We will explain the aggregation design later.

The teacher shares the same architecture as the student's encoder and projector, and has a similar pipeline described by Eq. (1), except it takes the unmasked inputs T_1 and T_2, and returns two distributions P_1 and P_2 for the two views, respectively. The student output $\hat{P}_2$ and the teacher outputs P_1 and P_2 are used for feature reinforcement training.

Uncertain Patch Token Selection

We select uncertain patch tokens under the guidance of semantic-aware CAMs, generated using the logits computed

earlier with the first view, i.e., $Z_1 = \mathcal{F}(T_1)$. We linearly project Z_1 using the weights $W \in \mathbb{R}^{C\times D}$ of the classifier for image classification, where C is the class number, and then normalize it by the ReLU function and $\min$-$\max$ normalization. The normalized CAM, denoted as $M_c \in \mathbb{R}^{HW\times C}(0 \leq M_C \leq 1)$, is defined by

$$M_c := \min\text{-}\max\left(\mathrm{ReLU}\left(ZW^\top\right)\right). \tag{2}$$

It encodes the semantic uncertainty for each patch driven by CAM scores $ZW^\top$.

Next, we identify the uncertain regions based on the normalized CAM and mask the uncertain patches, following an adaptive masking strategy. Features in non-reliable regions are considered as uncertain features. However, some reliable regions can be wrongly labeled, and their corresponding features can also be uncertain. To remedy this, we propose an adaptive masking strategy, resulting in a soft masking vector $M_s \in \mathbb{R}^{HW}$ with each element given as

$$M_s^{(i)} = \begin{cases} u_i + 1, & \text{if } \beta_l < \max\left(M_c^{(i,:)}\right) < \beta_h, \\ u_i, & \text{otherwise,} \end{cases} \tag{3}$$

where $u_i \sim \mathrm{U}(0,1)$ draws from a standard uniform distribution and enables a stochastic selection process. The above use of two background thresholds $0 < \beta_l < \beta_h < 1$ for dividing patches into reliable and non-reliable ones is inspired by Zhang et al. (2020) and Ru et al. (2022), which suggests an intermediate score to be a sign of uncertainty. As a result, elements in M_s with larger values suggest uncertain patches.

We use a masking ratio $0 < r < 1$ to control the amount of selected uncertain patches, and defines the following binary selection mask $M_b \in \mathbb{R}^{HW}$ with each element as

$$M_b^{(i)} = \begin{cases} 1, & \text{if } i \in \arg\max_{\mathrm{top(k)}}(M_s), k := \lfloor HW * r \rfloor, \\ 0, & \text{otherwise,} \end{cases} \tag{4}$$

where $\lfloor\cdot\rfloor$ denotes the floor function. The selected uncertain patches, flagged by 1 in M_b, correspond to those top-k large-valued elements in M_s. Our masking strategy is designed to relax the hard foreground-background thresholds by the masking ratio r. When more patches are flagged as uncertain by $\beta_l < \max\left(M_c^{(i,:)}\right) < \beta_h$, the selection is randomly conducted within them through u_i. When less uncertain patches are flagged, part of confident patches are also selected. The original features of the selected tokens to mask are replaced by learnable parameters with the same feature dimension.

Certain Feature Aggregation

We design an attentive aggregation module to compress the embeddings of a sequence of $N = HW$ patch tokens, stored in $\hat{Z} \in \mathbb{R}^{N\times D}$, into one class token embedding $\bar{Z} \in \mathbb{R}^{1\times D}$. As shown in Figure 3, this module contains several aggregation blocks, where each block contains a masked cross-attention (MCA) layer and a feed-forward (FF) layer, given as

$$\begin{aligned} \bar{Z}_{(o)}^{(l)} &= \bar{Z}^{(l)} + \mathrm{MCA}\left(\eta\left(\left[\bar{Z}^{(l)}; \hat{Z}^{(l)}\right]\right)\right), \\ \bar{Z}^{(l+1)} &= \bar{Z}_{(o)}^{(l)} + \mathrm{FF}\left(\eta\left(\bar{Z}_{(o)}^{(l)}\right)\right), \end{aligned} \tag{5}$$

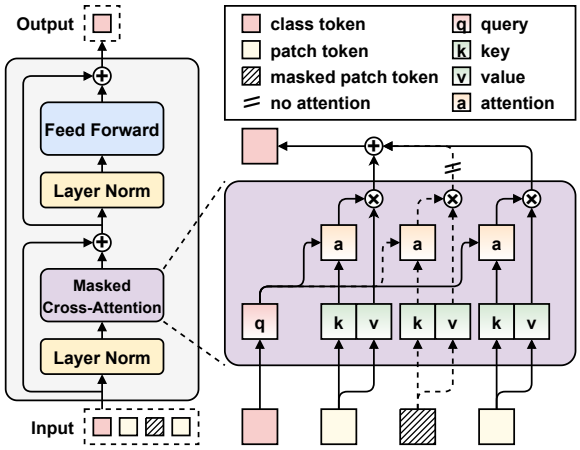

Figure 3: Illustration of the aggregation module. This module is composed of several aggregation blocks, where each block alternates in turn a cross-attention layer and a feed-forward layer. The cross-attention layer computes attention between a class token and a sequence of unmasked patch tokens.

where l denotes the layer index and η is the LayerNorm (Ba, Kiros, and Hinton 2016).

MCA is analogous to self-attention (Vaswani et al. 2017), except that it computes attention between the class token and a set of unmasked patch tokens. We parameterize MCA with projection weights $W_Q, W_K, W_V, W_O \in \mathbb{R}^{D\times D}$, and calculate the queries $Q \in \mathbb{R}^{1\times D}$, keys $K \in \mathbb{R}^{N\times D}$ and values $V \in \mathbb{R}^{N\times D}$ by projection, so that

$$Q = \eta\left(\bar{Z}\right)W_Q^\top, K = \eta\left(\hat{Z}\right)W_K^\top, V = \eta\left(\hat{Z}\right)W_V^\top. \tag{6}$$

Note that queries are derived from the class token, while keys and values are calculated on patch tokens. The masked cross-attention $A \in \mathbb{R}^{1\times N}$ is then formulated as

$$A = \sigma\left(\frac{\left(1 - M_b\right)\left(QK^\top\right)}{\sqrt{D}}\right). \tag{7}$$

The output of MCA is computed as a weighted sum of values, i.e., $(AV)\,W_O^\top$.

Feature Self-reinforcement

We adopt self-distillation (Caron et al. 2021; Zhou et al. 2021; Oquab et al. 2023) to improve the model training for feature reinforcement. As explained earlier, given two distorted views of the same image, we compute one student output $\hat{P}_2$ and two teacher outputs P_1 and P_2, where their first element stores the aggregated token information, while the remaining the individual token content. We propose a self-reinforcement loss $\mathcal{L}_u$ for the uncertain tokens, as the cross-entropy loss between each student's patch token and its corresponding teacher's patch token:

$$\mathcal{L}_u = -\sum_{i=2}^{1+N} M_b^{(i)} P_2^{(i)} \log \hat{P}_2^{(i)}, \tag{8}$$

Method	Sup.	Net.	Val	Test
Multi-stage WSSS methods				
RIB (Lee et al. 2021a)	$\mathcal{I}+\mathcal{S}$	RN101	70.2	70.0
EDAM (Wu et al. 2021)	$\mathcal{I}+\mathcal{S}$	RN101	70.9	70.6
EPS (Lee et al. 2021b)	$\mathcal{I}+\mathcal{S}$	RN101	71.0	71.8
SANCE (Li, Fan, and Zhang 2022)	$\mathcal{I}+\mathcal{S}$	RN101	72.0	72.9
L2G (Jiang et al. 2022)	$\mathcal{I}+\mathcal{S}$	RN101	72.1	71.7
RCA (Zhou et al. 2022)	$\mathcal{I}+\mathcal{S}$	RN38	72.2	72.8
SEAM (Wang et al. 2020)	$\mathcal{I}$	RN38	64.5	65.7
BES (Chen et al. 2020)	$\mathcal{I}$	RN101	65.7	66.6
CPN (Zhang et al. 2021)	$\mathcal{I}$	RN38	67.8	68.5
CDA (Su et al. 2021)	$\mathcal{I}$	RN38	66.1	66.8
ReCAM (Chen et al. 2022)	$\mathcal{I}$	RN101	68.5	68.4
URN (Li et al. 2022b)	$\mathcal{I}$	RN101	69.5	69.7
ESOL (Li et al. 2022a)	$\mathcal{I}$	RN101	69.9	69.3
†ViT-PCM (Rossetti et al. 2022)	$\mathcal{I}$	RN101	70.3	70.9
†MCTformer (Xu et al. 2022)	$\mathcal{I}$	RN38	71.9	71.6
†OCR (Cheng et al. 2023)	$\mathcal{I}$	RN38	72.7	72.0
†BECO (Rong et al. 2023)	$\mathcal{I}$	MiT-B2	73.7	73.5
†MCTformer+ (Xu et al. 2023)	$\mathcal{I}$	RN38	74.0	73.6
Single-stage WSSS methods				
RRM (Zhang et al. 2020)	$\mathcal{I}$	RN38	62.6	62.9
1Stage (Araslanov and Roth 2020)	$\mathcal{I}$	RN38	62.7	64.3
†AFA (Ru et al. 2022)	$\mathcal{I}$	MiT-B1	66.0	66.3
†ToCo (Ru et al. 2023)	$\mathcal{I}$	ViT-B	71.1	72.2
†**Ours**	$\mathcal{I}$	ViT-B	**75.7**	**75.0**

Table 1: Performance comparison of semantic segmentation on PASCAL VOC 2012 in terms of mIoU(%). Sup. denotes the supervision type. $\mathcal{I}$: image-level class labels. $\mathcal{S}$: off-the-shelf saliency maps. Net. denotes the segmentation network for multi-stage methods or the backbone for single-stage methods. RN38: Wide ResNet38 (Wu, Shen, and Van Den Hengel 2019), RN101: ResNet101 (He et al. 2016), MiT: Mix Transformer (Xie et al. 2021). † flags transformer based classification network or backbone.

where M_b is the mask in Eq. (4) to help select masked patch tokens. We also conduct self-reinforcement for the confident tokens, formulated as the cross-entropy loss on the two aggregated class tokens of the two views, given as

$$\mathcal{L}_c = -P_1^{(1)} \log \hat{P}_2^{(1)}. \tag{9}$$

Following a common practice, we adopt the multi-label soft margin loss $\mathcal{L}_{cls}$ for classification, the pixel-wise cross-entropy loss $\mathcal{L}_{seg}$ for segmentation, and the cosine similarity loss $\mathcal{L}_{aff}$ for affinity regularization. Denote the weighting factors as $\{\lambda_i\}_{i=1}^5$, the overall training objective is

$$\mathcal{L} = \lambda_1\mathcal{L}_{cls} + \lambda_2\mathcal{L}_{aff} + \lambda_3\mathcal{L}_{seg} + \lambda_4\mathcal{L}_u + \lambda_5\mathcal{L}_c. \tag{10}$$

It consolidates classification, segmentation and feature self-reinforcement within a single-stage framework.

Experiments

Experimental Settings

Datasets We evaluate our method on two benchmarks: PASCAL VOC 2012 (Everingham et al. 2010) and MS COCO 2014 (Lin et al. 2014). PASCAL VOC contains 20 object classes and one background class. Following the common practice of previous works (Zhang et al. 2020; Araslanov and Roth 2020; Ru et al. 2022, 2023), it is augmented with data from the SBD dataset (Hariharan et al. 2011), resulting in $10,582$, $1,449$ and $1,456$ images for training, validation and testing, respectively. MS COCO contains 80 object classes and one background class. It has $82,081$ images for training and $40,137$ images for validation. Note that we only adopt image-level labels during the training phase. We report mean Intersection-over-Union (mIoU) as the evaluation metric.

Method	Sup.	Net.	Val
Multi-stage WSSS methods			
EPS (Lee et al. 2021b)	$\mathcal{I}+\mathcal{S}$	RN101	35.7
RIB (Lee et al. 2021a)	$\mathcal{I}+\mathcal{S}$	RN101	43.8
L2G (Jiang et al. 2022)	$\mathcal{I}+\mathcal{S}$	RN101	44.2
CDA (Su et al. 2021)	$\mathcal{I}$	RN38	33.2
URN (Li et al. 2022b)	$\mathcal{I}$	RN101	40.7
ESOL (Li et al. 2022a)	$\mathcal{I}$	RN101	42.6
†MCTformer (Xu et al. 2022)	$\mathcal{I}$	RN38	42.0
†ViT-PCM (Rossetti et al. 2022)	$\mathcal{I}$	RN101	45.0
†OCR (Cheng et al. 2023)	$\mathcal{I}$	RN38	42.5
BECO (Rong et al. 2023)	$\mathcal{I}$	RN101	45.1
†MCTformer+ (Xu et al. 2023)	$\mathcal{I}$	RN38	45.2
Single-stage WSSS methods			
†AFA (Ru et al. 2022)	$\mathcal{I}$	MiT-B1	38.9
†ToCo (Ru et al. 2023)	$\mathcal{I}$	ViT-B	42.3
†**Ours**	$\mathcal{I}$	ViT-B	**45.4**

Table 2: Performance comparison of semantic segmentation on MS COCO 2014 in terms of mIoU(%). We use the same notations as in Table 1.

Implementation Details We adopt ViT-B (Dosovitskiy et al. 2020) pretrained on ImageNet (Deng et al. 2009) as the transformer encoder. The convolutional decoder refers to DeepLab-LargeFOV (Chen et al. 2017). We use two aggregation blocks in the aggregation module. The projector comprises a 3-layer perceptron and a weight-normalized fully connected layer (Caron et al. 2021). Parameters in the aggregation module and the projector are randomly initialized. We use a light data augmentation: random resized cropping to 448×448, random horizontal flipping, and random color jittering. The student is optimized with AdamW (Loshchilov and Hutter 2017). The base learning rate is warmed up to $6e-5$ and decayed with a cosine schedule. The weighting factors $(\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5)$ are $(1.0, 0.2, 0.1, 0.1, 0.1)$. The teacher requires no gradient and is updated using exponential moving average (EMA). The momentum for EMA is 0.9995 and increases to 1.0 with a cosine schedule during training. We embrace centering and sharpening techniques suggested in (Caron et al. 2021) to avoid collapsed solutions. The masking ratio r is 0.4 for adaptive uncertain feature selection. The background scores (β_l, β_h) introduced to determine uncertain regions are $(0.2, 0.7)$. Training iterations are 20,000 for PASCAL VOC 2012 and 80,000 for MS COCO 2014. We use multi-scale testing and dense CRF (Chen et al. 2014) at test time following (Ru et al. 2022, 2023).

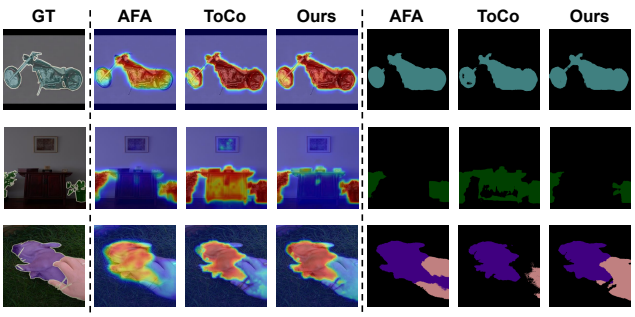

Figure 4: Visualization results of CAMs and predicted segmentation labels with SOTA single-stage frameworks (i.e., AFA and ToCo). (left) Ground truth. (middle) Comparison results of CAMs. (right) Comparison results of predicted segmentation labels.

Comparison with State-of-the-arts

PASCAL VOC 2012 Table 1 shows comparison results of our proposed Feature Self-Reinforcement (FSR) with other state-of-the-art methods on PASCAL VOC 2012. As can be seen from this table, FSR significantly outperforms other single-stage approaches, achieving 75.7% and 75.0% mIoU on the validation and test sets, respectively. It is noticeable that our method achieves even higher mIoU than several sophisticated multi-stage methods, e.g., FSR surpasses BECO (Rong et al. 2023) by margins of 2.0% and 1.5%. Compared with multi-stage methods using both image-level labels and off-the-shelf saliency maps, e.g., L2G (Jiang et al. 2022) and RCA (Zhou et al. 2022), our method still achieves superior performance. We assume although saliency maps are effective in providing additional background clues, our method can strengthen both confident regions (mostly the main body of objects or the background) and uncertain regions (mostly object boundaries), so that semantically distinct objects can be better differentiated. Moreover, it shows that recent methods with transformer-based networks (with †) generally outperform those with convolutional networks (without †). Nevertheless, due to the difficulty of end-to-end optimization, single-stage transformer-based methods (e.g., ToCo reports 71.1% and 72.2%) can only achieve comparable performance with multi-stage ones (e.g., BECO reports 73.7% and 73.5%). Our method proves the efficacy of transformer-based single-stage training by attaining even better results.

MS COCO 2014 Table 2 gives comparison results of semantic segmentation on a more challenging benchmark MS COCO 2014. We achieve 45.5% mIoU on the validation set, which outperforms previous single-stage solutions and is slightly better than multi-stage MCTformer+ (Xu et al. 2023) by 0.2%. This further demonstrates the superiority of our proposed method.

Visualization Results In Figure 4, we visualize CAMs derived from the classifier and semantic segmentation labels predicted by the decoder of three single-stage methods, i.e., AFA (Ru et al. 2022), ToCo (Ru et al. 2023) and our proposed FSR. Compared with AFA, ToCo and FSR can generate more integral and deterministic CAMs. For instance, the wheels of "motorbike" are mildly activated by AFA while strongly confirmed by ToCo and FSR. This proves the effectiveness of FSR for uncertain features. However, AFA only focuses on boosting uncertain features, whereas our method enhances both uncertain and certain ones. For instance, AFA mistakes "drawer" as "chair", while FSR successfully recognizes the different semantics. This shows the importance of FSR for seemingly certain features.

mask ratio	Edge (strict)	CAM (strict)	CAM 0.1	CAM 0.2	CAM 0.3	CAM 0.4	CAM 0.5
Pseu. label results (%)							
random	-	-	73.1	73.6	**74.1**	74.2	73.2
uncertain	73.3	74.0	**74.1**	**74.2**	73.9	**74.4**	**73.7**
Pred. label results (%)							
random	-	-	71.7	72.3	71.3	72.3	71.2
uncertain	71.6	71.8	**72.2**	**72.3**	**72.0**	**72.5**	**72.1**

Table 3: Ablation results of uncertain feature selection methods. "random" means random masking, "uncertain" means our adaptive masking strategy that gives priority to masking uncertain regions.

Masking	unc.FSR	cer.FSR	Pseu. (%)	Pred. (%)
-	-	-	71.1	67.9
CAM	✓		74.4	72.5
		✓(GAP)	72.3	70.9
		✓(GMP)	71.8	70.0
		✓(MCA)	75.2	73.3
	✓	✓(MCA)	**75.7**	**73.6**

Table 4: Ablation results of unc.FSR and cer.FSR. "GAP", "GMP", and "MCA" are aggregation methods adopted for cer.FSR.

Ablation Studies

In this section, we present extensive ablation studies to verify the effectiveness of our proposed FSR. We report segmentation performance of pseudo labels (Pseu.) derived from CAMs as well as predicted labels (Pred.) generated by the decoder. All results are evaluated on PASCAL VOC 2012 val set. Dense CRF is not applied with ablations.

Analysis of Uncertain Feature Selection In Table 3, we compare two *strict* selection methods for uncertain features: edge-based selection and CAM-based selection. For edge-based selection, we choose the conventional Canny edge detector to extract edges in an image and generate exact masks of these edges. Activation thresholds for CAM-based selection are $(0.2, 0.7)$. CAM-based selection is marginally better than edge-based selection; the improvement continues when CAM-based selection is relaxed, i.e., uncertain features are not strictly but preferentially masked. Empirically, we find that $r = 0.4$ gives the best result. In addition, uncertain feature masking achieves higher performance than random

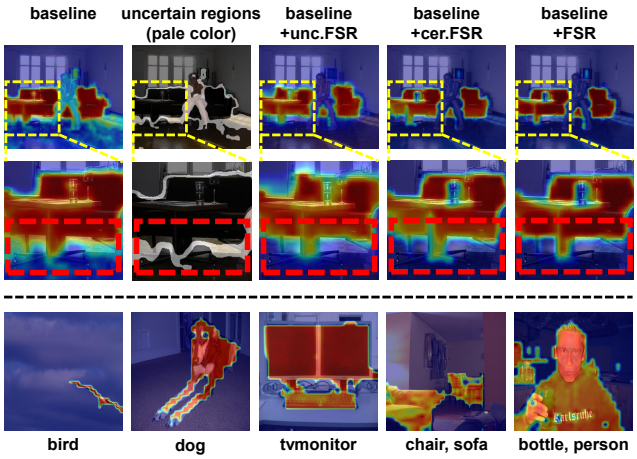

Figure 5: (top) Visualization of FSR optimizing boundary regions (e.g., boundaries of *sofa* in dashed red boxes) via uncertain and confident feature self-reinforcement. (bottom) Class-to-patch attention maps derived from the aggregation module. Class labels are displayed under each example.

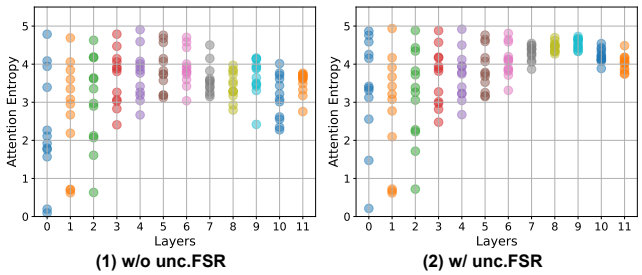

Figure 6: Average attention entropy of different attention heads (dots) across different layers.

	Ours	+GaussianBlur	+Solarization	AutoAugment
Pseu. (%)	**75.7**	**75.9** $\pm$ 0.05	75.3 $\pm$ 0.12	74.8 $\pm$ 0.09
Pred. (%)	**73.6**	**73.6** $\pm$ 0.02	73.2 $\pm$ 0.06	72.8 $\pm$ 0.04

Table 5: Ten-trial experimental results of data augmentations. "Ours" is our default data augmentation setting.

feature masking in most cases, showing it is important to reinforce uncertain features for semantics clarification.

Analysis of Feature Self-reinforcement Table 4 shows the ablation results of FSR on uncertain regions (unc.FSR) and on certain regions (cer.FSR). The masking ratio is set to 0.4 for comparison. It demonstrates the advancement of unc.FSR by achieving 74.4% (compared to 71.1%) on pseudo labels and 72.5% (compared to 67.9%) on predicted labels. This proves that reinforcing uncertain features, which mainly contain ambiguous object boundaries and misclassified categories, is fairly effective. When combining unc.FSR with cer.FSR, the quality of pseudo labels can be further improved, from 74.4% to 75.7%; predicted labels directly supervised by pseudo labels are promoted as well, from 72.5% to 73.6%. This indicates that reinforcing confident features is complementary to unc.FSR with enhanced global understanding. We showcase examples of our FSR method and its effect on object boundaries in Figure 5 (top).

(a) Analysis of unc.FSR To gain a deep understanding of unc.FSR, we investigate the training process by analyzing the attention mechanism. Specifically, we compute average attention entropy (Attanasio et al. 2022) for each attention head across transformer layers. As shown in Figure 6, the entropy at shallow layers (e.g., layer 0 to 6) holds similar without unc.FSR; however, it becomes higher and tighter at deep layers (e.g., layer 7 to 11) when unc.FSR is applied. A large entropy for a specific token indicates that a broad context contributes to this token, while a small entropy tells the opposite. From this point of view, we assume that unc.FSR benefits semantic segmentation by improving the degree of contextualization at deep layers.

(b) Analysis of cer.FSR We compare our attentive aggregation of certain features (MCA) with two conventional methods: Global Average Pooling (GAP) and Global Maximum Pooling (GMP). GAP assigns an equal weight to each unmasked patch token, while GMP picks up the dominant one along each dimension. Table 4 shows that GAP performs better than GMP, as GMP tends to intensify the most discriminative features, which may have an adverse effect in recognizing an integral object. It is noticeable that MCA outperforms GAP by a large margin, indicating an attentive weighting mechanism is superior to average weighting. We visualize class-to-patch attention maps in Figure 5 (bottom), which illustrates that the class token can adaptively learn to pay attention to object regions. Note that the class token is not directly supervised by classification in our design; it interacts with unmasked patch tokens and learns to summarize effective information from them.

Data Augmentation We present comparison results with other data augmentations in Table 5, which reveals that data augmentations have limited impacts on the performance. For example, the performances display variations within the anticipated range when incorporating GaussianBlur or Solarization. Even when we substitute the data augmentation with the robust AutoAugment (Cubuk et al. 2018), the results witness a slight decline as a strong augmentation may interfere with the segmentation objective.

Conclusion

In this work, we propose to estimate boundaries with the guidance of semantic uncertainty identified by CAM. To achieve this, we design an activation-based masking strategy and seek to recover local information with self-distilled knowledge. We further introduce a self-distillation method to reinforce semantic consistency with another augmented view. We integrate our method into the single-stage WSSS framework and validate its effectiveness on PASCAL VOC 2012 and MS COCO 2014 benchmarks.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant No. (62106235, 62202015, 62376206, 62003256), and in part by the Exploratory Research Project of Zhejiang Lab under Grant 2022PG0AN01.

References

Ahn, J.; Cho, S.; and Kwak, S. 2019. Weakly supervised learning of instance segmentation with inter-pixel relations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2209–2218.

Araslanov, N.; and Roth, S. 2020. Single-stage semantic segmentation from image labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4253–4262.

Attanasio, G.; Nozza, D.; Hovy, D.; and Baralis, E. 2022. Entropy-based attention regularization frees unintended bias mitigation from lists. *arXiv preprint arXiv:2203.09192*.

Ba, J. L.; Kiros, J. R.; and Hinton, G. E. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Bearman, A.; Russakovsky, O.; Ferrari, V.; and Fei-Fei, L. 2016. What's the point: Semantic segmentation with point supervision. In *European conference on computer vision*, 549–565. Springer.

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9650–9660.

Chen, L.; Wu, W.; Fu, C.; Han, X.; and Zhang, Y. 2020. Weakly supervised semantic segmentation with boundary exploration. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVI 16*, 347–362. Springer.

Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. L. 2014. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*.

Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. L. 2017. 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.

Chen, Z.; Wang, T.; Wu, X.; Hua, X.-S.; Zhang, H.; and Sun, Q. 2022. Class re-activation maps for weakly-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 969–978.

Cheng, Z.; Qiao, P.; Li, K.; Li, S.; Wei, P.; Ji, X.; Yuan, L.; Liu, C.; and Chen, J. 2023. Out-of-candidate rectification for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23673–23684.

Cubuk, E. D.; Zoph, B.; Mane, D.; Vasudevan, V.; and Le, Q. V. 2018. Autoaugment: Learning augmentation policies from data. *arXiv preprint arXiv:1805.09501*.

Dai, J.; He, K.; and Sun, J. 2015. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, 1635–1643.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Everingham, M.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2010. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88: 303–338.

Gao, W.; Wan, F.; Pan, X.; Peng, Z.; Tian, Q.; Han, Z.; Zhou, B.; and Ye, Q. 2021. Ts-cam: Token semantic coupled attention map for weakly supervised object localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2886–2895.

Gong, C.; Wang, D.; Li, M.; Chandra, V.; and Liu, Q. 2021. Vision transformers with patch diversification. *arXiv preprint arXiv:2104.12753*.

Hariharan, B.; Arbeláez, P.; Bourdev, L.; Maji, S.; and Malik, J. 2011. Semantic contours from inverse detectors. In *2011 international conference on computer vision*, 991–998. IEEE.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9729–9738.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Hou, Q.; Jiang, P.; Wei, Y.; and Cheng, M.-M. 2018. Self-erasing network for integral object attention. *Advances in Neural Information Processing Systems*, 31.

Jiang, P.-T.; Yang, Y.; Hou, Q.; and Wei, Y. 2022. 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*, 16886–16896.

Lee, J.; Choi, J.; Mok, J.; and Yoon, S. 2021a. Reducing information bottleneck for weakly supervised semantic segmentation. *Advances in Neural Information Processing Systems*, 34: 27408–27421.

Lee, J.; Kim, E.; Lee, S.; Lee, J.; and Yoon, S. 2019. Ficklenet: Weakly and semi-supervised semantic image segmentation using stochastic inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5267–5276.

Lee, S.; Lee, M.; Lee, J.; and Shim, H. 2021b. 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*, 5495–5505.

Li, J.; Fan, J.; and Zhang, Z. 2022. Towards noiseless object contours for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16856–16865.

Li, J.; Jie, Z.; Wang, X.; Wei, X.; and Ma, L. 2022a. Expansion and shrinkage of localization for weakly-supervised semantic segmentation. *Advances in Neural Information Processing Systems*, 35: 16037–16051.

Li, Y.; Duan, Y.; Kuang, Z.; Chen, Y.; Zhang, W.; and Li, X. 2022b. Uncertainty estimation via response scaling for pseudo-mask noise mitigation in weakly-supervised semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 1447–1455.

Lin, D.; Dai, J.; Jia, J.; He, K.; and Sun, J. 2016. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3159–3167.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755. Springer.

Lin, Y.; Chen, M.; Wang, W.; Wu, B.; Li, K.; Lin, B.; Liu, H.; and He, X. 2023. 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*, 15305–15314.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Noroozi, M.; Vinjimoor, A.; Favaro, P.; and Pirsiavash, H. 2018. Boosting self-supervised learning via knowledge transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 9359–9367.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Rong, S.; Tu, B.; Wang, Z.; and Li, J. 2023. Boundary-Enhanced Co-Training for Weakly Supervised Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19574–19584.

Rossetti, S.; Zappia, D.; Sanzari, M.; Schaerf, M.; and Pirri, F. 2022. Max pooling with vision transformers reconciles class and shape in weakly supervised semantic segmentation. In *European Conference on Computer Vision*, 446–463. Springer.

Ru, L.; Zhan, Y.; Yu, B.; and Du, B. 2022. 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*, 16846–16855.

Ru, L.; Zheng, H.; Zhan, Y.; and Du, B. 2023. Token contrast for weakly-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3093–3102.

Song, C.; Huang, Y.; Ouyang, W.; and Wang, L. 2019. Box-driven class-wise region masking and filling rate guided loss for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3136–3145.

Su, H.; Ye, Y.; Hua, W.; Cheng*, L.; and Song, M. 2022. SAS-Former: Transformers for Sparsely Annotated Semantic Segmentation. In *IEEE International Conference on Multimedia and Expo (ICME), 2023*.

Su, Y.; Sun, R.; Lin, G.; and Wu, Q. 2021. Context decoupling augmentation for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 7004–7014.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Vernaza, P.; and Chandraker, M. 2017. Learning random-walk label propagation for weakly-supervised semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7158–7166.

Wang, Y.; Cheng*, L.; Duan, M.; Wang, Y.; Feng, Z.; and Kong, S. 2023. Improving Knowledge Distillation via Regularizing Feature Norm and Direction. *arXiv preprint arXiv:2305.17007*.

Wang, Y.; Zhang, J.; Kan, M.; Shan, S.; and Chen, X. 2020. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12275–12284.

Wu, F.; He, J.; Cheng*, L.; Yin, Y.; Hao, Y.; and Huang, G. 2023. Masked Collaborative Contrast for Weakly Supervised Semantic Segmentation. *arXiv preprint arXiv:2305.08491*.

Wu, T.; Huang, J.; Gao, G.; Wei, X.; Wei, X.; Luo, X.; and Liu, C. H. 2021. Embedded discriminative attention mechanism for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16765–16774.

Wu, Z.; Shen, C.; and Van Den Hengel, A. 2019. Wider or deeper: Revisiting the resnet model for visual recognition. *Pattern Recognition*, 90: 119–133.

Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J. M.; and Luo, P. 2021. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34: 12077–12090.

Xie, Z.; Geng, Z.; Hu, J.; Zhang, Z.; Hu, H.; and Cao, Y. 2023. Revealing the dark secrets of masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14475–14485.

Xu, L.; Bennamoun, M.; Boussaid, F.; Laga, H.; Ouyang, W.; and Xu, D. 2023. MCTformer+: Multi-Class Token Transformer for Weakly Supervised Semantic Segmentation. *arXiv preprint arXiv:2308.03005*.

Xu, L.; Ouyang, W.; Bennamoun, M.; Boussaid, F.; and Xu, D. 2022. Multi-class token transformer for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4310–4319.

Yoon, S.-H.; Kweon, H.; Cho, J.; Kim, S.; and Yoon, K.-J. 2022. Adversarial erasing framework via triplet with gated pyramid pooling layer for weakly supervised semantic segmentation. In *European Conference on Computer Vision*, 326–344. Springer.

Zhang, B.; Xiao, J.; Wei, Y.; Sun, M.; and Huang, K. 2020. Reliability does matter: An end-to-end weakly supervised semantic segmentation approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 12765–12772.

Zhang, F.; Gu, C.; Zhang, C.; and Dai, Y. 2021. Complementary patch for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 7242–7251.

Zhang, T.; Xue, M.; Zhang, J.; Zhang, H.; Wang, Y.; Cheng, L.; Song, J.; and Song, M. 2023. Generalization Matters: Loss Minima Flattening via Parameter Hybridization for Efficient Online Knowledge Distillation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2023*.

Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; and Torralba, A. 2016. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2921–2929.

Zhou, J.; Wei, C.; Wang, H.; Shen, W.; Xie, C.; Yuille, A.; and Kong, T. 2021. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*.

Zhou, T.; Zhang, M.; Zhao, F.; and Li, J. 2022. Regional semantic contrast and aggregation for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4299–4309.