applied sciences

MDPI

Article

Enhanced Pseudo-Labels for End-to-End Weakly Supervised Semantic Segmentation with Foundation Models

Xuesheng Zhou [1] and Zhenzhou Tang [2,*]

[1] School of Computer and Information Engineering, Fuyang Normal University, Fuyang 236037, China; zxs@fynu.edu.cn
[2] Wenzhou Key Laboratory for Intelligent Networking, Wenzhou University, Wenzhou 325035, China
* Correspondence: mr.tangzz@gmail.com

Abstract

Weakly supervised semantic segmentation (WSSS) aims to learn pixel-level semantic concepts from image-level class labels. Due to the simplicity and efficiency of the training pipeline, end-to-end WSSS has received significant attention from the research community. However, the coarse nature of pseudo-label regions remains one of the primary bottlenecks limiting the performance of such methods. To address this issue, we propose class-guided enhanced pseudo-labeling (CEP), a method designed to generate high-quality pseudo-labels for end-to-end WSSS frameworks. Our approach leverages pretrained foundation models, such as contrastive language-image pre-training (CLIP) and segment anything model (SAM), to enhance pseudo-label quality. Specifically, following the pseudo-label generation pipeline, we introduce two key components: a Skip-CAM module and a pseudo-label refinement module. The Skip-CAM module enriches feature representations by introducing skip connections from multiple blocks of CLIP, thereby improving the quality of localization maps. The refinement module then utilizes SAM to refine and correct the pseudo-labels based on the initial class-specific regions derived from the localization maps. Experimental results demonstrate that our method surpasses the state-of-the-art end-to-end approaches as well as many multi-stage competitors.

Keywords: weakly supervised semantic segmentation; pseudo-label; foundation model; CLIP; SAM

check for updates

Received: 12 August 2025
Revised: 6 September 2025
Accepted: 9 September 2025
Published: 12 September 2025

Citation: Zhou, X.; Tang, Z. Enhanced Pseudo-Labels for End-to-End Weakly Supervised Semantic Segmentation with Foundation Models. *Appl. Sci.* **2025**, *15*, 10013. https://doi.org/10.3390/app151810013

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Fully supervised semantic segmentation (FSSS) relies heavily on pixel-level annotations for model training, which are typically expensive and labor-intensive to acquire. Alternatively, weakly supervised semantic segmentation (WSSS) attempts to learn semantic segmentation models using weaker annotations, such as image-level labels [1,2], bounding boxes [3,4], or scribbles [5,6]. Based on different training paradigms, WSSS methods can be categorized into multi-stage and single-stage (end-to-end) approaches. Recently, end-to-end WSSS [7,8] has attracted significant attention due to its concise and straightforward training pipeline. End-to-end WSSS methods commonly employ pseudo-labels as supervision signals to train semantic segmentation networks. End-to-end approaches, such as AFA [7] and ToCo [8], also focus on improving the quality of pseudo-labels. For instance, ToCo [8] and WeCLIP [9] replace pseudo-labels with ground-truth labels (which can be regarded as the highest-quality pseudo-labels) as supervision signals for training segmentation networks, and experimental results demonstrate significant improvements in

segmentation performance. Therefore, the quality of pseudo-labels providing supervision signals is one of the key factors in enhancing segmentation performance.

Both multi-stage and end-to-end WSSS approaches typically employ class activation maps (CAM) [10] or its variants to generate initial localization maps. However, CAM usually focuses only on highly discriminative regions. Recently, several studies have attempted to incorporate contrastive language-image pre-training (CLIP) [11] into WSSS frameworks. CLIP, trained on approximately 400 million image–text pairs, exhibits strong image–text alignment capabilities and robust image representation power. Previous studies [9,12] have demonstrated that CLIP possesses remarkable object localization abilities and has been employed to enhance CAM performance. Nevertheless, directly utilizing CAM derived from CLIP for constructing pseudo-labels still encounters issues such as coarse localization and ambiguous boundaries, leaving considerable room for further improvement in pseudo-label quality.

Furthermore, another foundational model, the segment anything model (SAM) [13], has recently been introduced into WSSS frameworks. SAM demonstrates remarkable zero-shot generalization capability, enabling accurate segmentation of previously unseen objects without additional training, thus providing high-quality masks for downstream tasks. However, masks generated by SAM typically lack class discriminability, and their quality greatly depends on the provided input prompts. Existing studies [14,15] generally employ SAM either for pseudo-label refinement or as a supervisory signal for CAM generation networks, both of which fall under the category of multi-stage WSSS methods. Consequently, exploring how to further leverage SAM to refine pseudo-labels while improving the quality of CLIP-generated CAMs, particularly within an end-to-end WSSS framework, remains an open and promising research direction.

To address these challenges, we propose a novel class-guided enhanced pseudo-labeling (CEP) framework for end-to-end WSSS. Specifically, we first introduce a skip-connected CAM (Skip-CAM) module, which integrates features from multiple intermediate layers via skip connections into the final block, extracting richer feature information to construct higher-quality localization maps. Building upon this, CEP incorporates a SAM-based pseudo-label refinement module to further optimize the initial pseudo-labels. In detail, this module employs coarse class regions derived from initial pseudo-labels as guidance to generate point prompts. SAM then produces multiple candidate class-related semantic regions based on these prompts. By combining these regions with the original class regions, our method selectively fuses them to generate refined pseudo-labels with enhanced precision and clear boundaries.

The main contributions of this paper are summarized as follows:

- We propose the CEP method, enabling high-quality online pseudo-label generation for end-to-end WSSS.
- We introduce the Skip-CAM and pseudo-label refinement modules to improve the quality of localization maps and the accuracy of pseudo labels, respectively, based on the pseudo-label generation pipeline.
- Experimental results on PASCAL VOC 2012 and COCO 2014 demonstrate that our approach significantly outperforms existing state-of-the-art methods.

2. Related Work

2.1. Fully Supervised Semantic Segmentation

Traditional image segmentation methods typically rely on feature or classification approaches. Classical feature techniques include histogram of oriented gradients [16] and scale-invariant feature transform [17], while classification techniques include support vector machines [18] and random forests [19]. Compared to deep learning-based segmentation

methods, traditional approaches generally require more domain knowledge for handcrafted feature design, but they still hold broad utility in certain scenarios.

Early deep learning segmentation methods were commonly built upon convolutional neural network (CNN) frameworks, such as FCN [20] and DeepLab [21]. In recent years, Vision Transformer (ViT)-based approaches have gained increasing popularity. Owing to the ability of the self-attention mechanism to effectively capture global semantic information, ViT [22] has been widely adopted as a backbone network for segmentation models to enhance performance. For instance, PVT [23] introduces a pyramid vision transformer for semantic segmentation, achieving competitive performance while reducing computational cost. Mask2Former [24] further proposes a segmentation framework that integrates a Transformer decoder with masked attention and a pixel decoder, thereby attaining stronger general-purpose segmentation capability. However, fully supervised deep learning methods for semantic segmentation depend heavily on large-scale pixel-level annotations, the acquisition of which is typically resource-intensive.

2.2. Weakly Supervised Semantic Segmentation

Weakly supervised semantic segmentation approaches generally fall into two solutions: multi-stage methods and end-to-end (i.e., single-stage) methods.

Multi-stage methods divide pseudo-label generation and semantic segmentation network training into separate stages. Typically, such methods include three stages: (1) localization map generation, where initial localization maps are produced using techniques like CAM for coarse localization; (2) pseudo-label generation, which involves refining localization maps into pseudo-labels; and (3) semantic segmentation network training, where the pseudo-labels supervise segmentation network training. AffinityNet [25] and IRN [1] are typical multi-stage solutions. For example, IRN initially trains a ResNet-based [26] multi-label classifier to generate localization maps, subsequently trains an IRNet to refine these maps into pseudo-labels, and finally uses these labels to supervise a DeepLab segmentation model [21]. Later methods, such as ReCAM [27] and AMN [28], focus on improving localization map quality. Furthermore, some studies (e.g., MCTformer [29] and CTI [2]) leveraged Transformer models for superior localization. Additionally, BBAM [4] and CDL [30] utilize bounding boxes and scribbles, respectively, as extra supervision to enhance WSSS performance. The training process of multi-stage WSSS methods is relatively cumbersome, as different models need to be trained for different target stages. Therefore, recent research has increasingly favored end-to-end WSSS.

Early end-to-end methods typically employed CNNs as backbone networks, jointly training for image classification and semantic segmentation. These methods most focus on extracting high-quality supervision from CNNs. For instance, RRM [31] employs WideResNet38 [32] as a backbone network, using a classification branch to generate CAM and further produce pseudo-labels. However, due to the limited local receptive fields of CNNs, recent methods have adopted Transformers [33] as backbone networks. The self-attention mechanism in Transformers better captures global contexts, addressing the limitations of CNNs and more effectively identifying entire object regions. For example, AFA [7] extracts semantic-level affinity via multi-head self-attention in Mix Transformer [34], refining initial coarse pseudo-labels. Additionally, DuPL [35] introduced a dual-student framework and progressive learning to enhance pseudo-label quality. However, traditional models such as CNNs and ViTs, when used as backbones, struggle to provide more precise localization information and higher-quality semantic features. This limitation hinders the further performance improvement of end-to-end methods.

2.3. CLIP and SAM

Recently, a new trend in WSSS involves utilizing large-scale pre-trained foundational models such as CLIP [11] and SAM [13] to further boost performance.

The CLIP model, consisting of an image encoder and a text encoder, is pre-trained on datasets comprising hundreds of millions of image–text pairs, demonstrating exceptional cross-modal generalization capabilities and superior performance on numerous downstream zero-shot tasks. Consequently, recent multi-stage WSSS approaches have begun integrating CLIP into their frameworks. For example, CLIMS [36] introduces CLIP in classifier training stage to activate more complete object regions and suppress background noise. WeakCLIP [37] leverages CLIP's pre-trained knowledge to build text-to-pixel matching, refining initial CAMs for more accurate pseudo-labels. CLIP-ES [12] incorporates softmax into GradCAM [38] and uses attention-based affinity mechanisms to construct more precise CAMs. Regarding end-to-end methods, WeCLIP [9] utilizes CLIP as the image encoder backbone, employing CLIP-ES to create high-quality supervision and achieving impressive performance.

SAM [13] is pre-trained on an extremely large-scale dataset containing 11 million images and over 1 billion masks. The model, consisting of an image encoder, prompt encoder, and mask decoder, exhibits strong segmentation generalization and flexible applicability across various downstream vision tasks. Unlike CLIP, which emphasizes image–text semantic matching, SAM focuses specifically on fine-grained instance segmentation and boundary recognition. Recently, SAM has been introduced into the WSSS domain, demonstrating significant potential in weakly supervised scenarios. Specifically, FMA [14] employs CLIP as a backbone for pseudo-label generation, learning classification and semantic prompts for image classification and seed segmentation tasks, subsequently generating seed maps with SAM. S2C [15] directly transfers SAM knowledge into the classifier training stage to produce enhanced localization maps.

CLIP and SAM exhibit complementary advantages in functionality. CLIP is capable of extracting high-quality semantic features from images and generating highly accurate initial localization information, while SAM possesses strong fine-grained segmentation capability, which can be utilized to refine the pseudo-labels produced by CLIP. Therefore, this work aims to simultaneously improve localization map quality and pseudo-label accuracy by leveraging CLIP and refining pseudo-labels through SAM, ultimately generating high-precision pseudo-labels suitable for end-to-end WSSS.

3. Method

The proposed method comprises three main modules: a feature extraction module, which extracts feature information from the input image; a Skip-CAM module, which utilizes skip connections from multiple blocks to obtain multi-level features for generating high-quality localization maps; and a pseudo-label refinement module, which employs rough class regions from initial pseudo-labels as guidance for SAM to generate refined pseudo-labels. Figure 1 illustrates the framework of our method.

3.1. Feature Extraction Module

CLIP's image encoder has two architectures, ResNet [26] and ViT [22]; we adopt the ViT architecture in this work. Given an input image $I \in \mathbb{R}^{3\times H\times W}$, where H and W denote image height and width, respectively, the image I is input to SAM's image encoder to extract features F_s. Concurrently, the image is input to CLIP's image encoder to obtain token features T^l from L block layers, with l indexing the block layer and L representing the total number of blocks. Additionally, the image class label c is used to construct a textual

prompt input into CLIP's text encoder, resulting in text features F_{txt}^{c}. The background classes are likewise encoded into text features F_{txt}^{bg}.

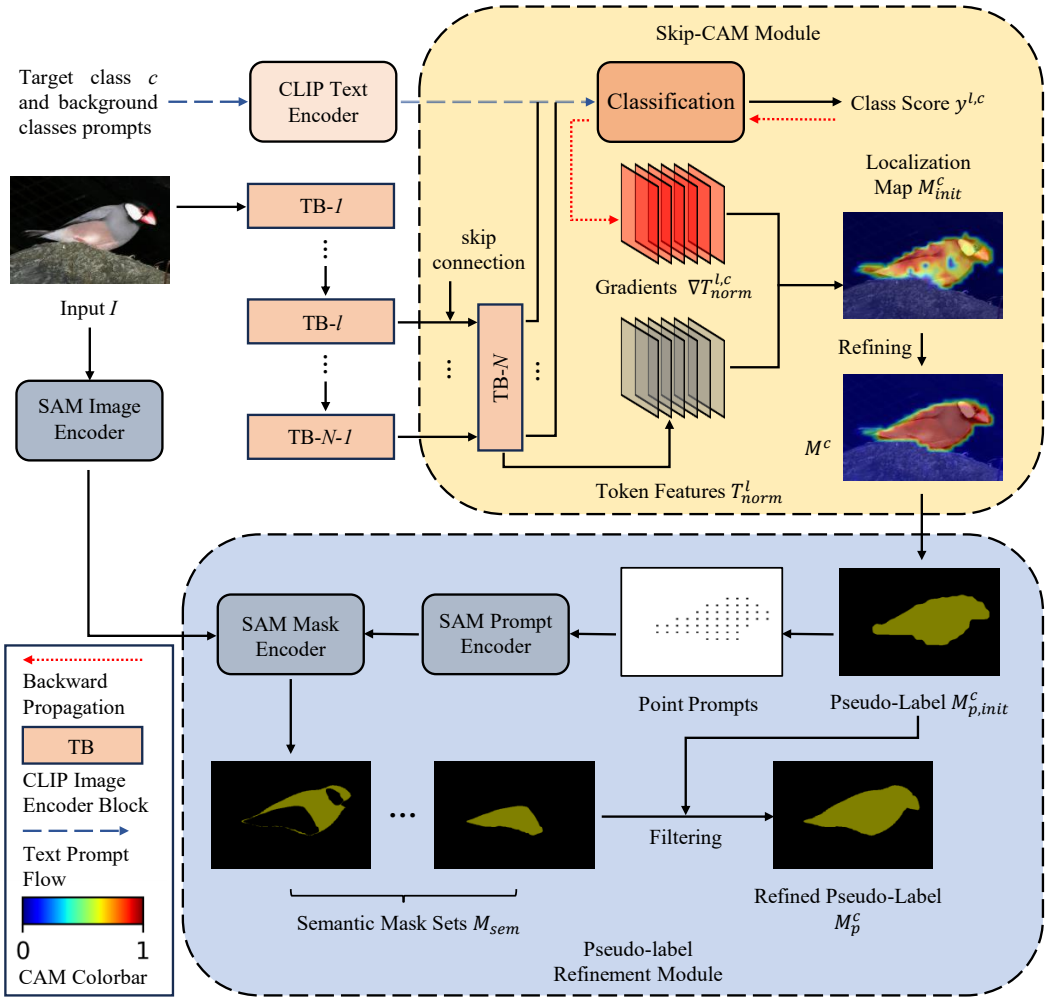


Figure 1. An overview of the proposed CEP framework.

3.2. Skip-CAM Module

Skip connections, widely adopted in deep neural networks, allow cross-layer information flow. This Skip-CAM module utilizes skip connections to aggregate rich feature information for generating localization maps. Specifically, for the token features T^{l} from the l-th block ($l < N$) , we first pass them through the last block of CLIP's image encoder. The output is compared to the target class and background classes text features F_{txt}^{c} and F_{txt}^{bg} to compute feature similarity scores, and the softmax function is applied to produce the target class scores $y^{l,c}$. Algorithm 1 illustrates the process of Skip-CAM.

In this process, we also obtain the normalized features T_{norm}^{l}, corresponding to T^{l}, from the first layer normalization in the N-th block, excluding the class token feature. Then, we compute the gradients of the target class score $y^{l,c}$ with respect to T_{norm}^{l}:

$$\nabla T_{norm}^{l,c} = \frac{\partial y^{l,c}}{\partial T_{norm}^{l}}. \tag{1}$$

Algorithm 1: Skip-CAM: Generating Init Localization Maps

Input: Input image $I \in \mathbb{R}^{3\times H\times W}$;
As an illustrative example, the input image of size 320 × 320 with a single foreground class.
CLIP image encoder with N blocks, where token features from the l-th block are denoted by T^l; # T^l: $401 \times 1 \times 768$
Text features: target class F^c_{txt} and background class F^{bg}_{txt};
F^c_{txt}: 1×512. F^{bg}_{txt}: 25×512
Number of skip-connected blocks N_{skip} (default: $\{N-N_{\text{skip}}, \ldots, N-1\}$).
Output: Initial localization map $M^c_{\text{init}} \in \mathbb{R}^{h\times w}$

Initialization:
Define the skip connection index set $\mathcal{L} = \{\, l \mid l = N - N_{\text{skip}}, \ldots, N-1 \,\}$;
Initialize $\mathcal{F} \leftarrow [\,]$ and $\mathcal{G} \leftarrow [\,]$.
for $l \in \mathcal{L}$ **do**
Pass T^l through the final block of CLIP's image encoder to obtain $\hat{T}^l$;
$y^{l,c} = \text{Classifier}\big(\hat{T}^l, [F^c_{\text{txt}}, F^{bg}_{\text{txt}}]\big)$
Extract normalized token T^l_{norm} from the first LayerNorm of the N-th block;
Remove Class token and reshape:

$$T^{c,orig}_{\text{norm}} = \text{reshape}\Big(T^{c,\text{orig}}_{\text{norm}}[1{:},\,],\ 1,\ 768,\ 20,\ 20\Big)$$

Processed $\nabla T^{l,c}_{\text{norm}}$ shape: $1 \times 768 \times 20 \times 20$.
Compute the gradient $\nabla T^{l,c}_{\text{norm}} = \dfrac{\partial y^{l,c}}{\partial T^l_{\text{norm}}}$;
Append T^l_{norm} to $\mathcal{F}$ and $\nabla T^{l,c}_{\text{norm}}$ to $\mathcal{G}$.

Feature and Gradient Fusion:
Compute the average normalized features and gradients: $\overline{T_{\text{norm}}} = \dfrac{1}{|\mathcal{L}|}\sum_{l\in\mathcal{L}} T^l_{\text{norm}}$,
$\overline{\nabla T^{l,c}_{norm}} = \dfrac{1}{|\mathcal{L}|}\sum_{l\in\mathcal{L}} \nabla T^{l,c}_{\text{norm}}$.
Linear Combination Weights:
$T_{norm,wgt}$: $1 \times 768 \times 1 \times 1$.
Average the gradients along spatial dimensions to obtain channel weights:
$T_{norm,wgt} = \frac{1}{hw}\sum_i \sum_j \overline{\nabla T^{l,c}_{norm,i,j}}$.
Localization Map Generation:
M^c_{init} : $1 \times 20 \times 20$.
Perform channel-wise linear combination followed by ReLU activation:
$M^c_{\text{init}} = \text{ReLU}\big(\overline{T_{\text{norm}}} \odot T_{\text{norm,wgt}}\big)$, where $\odot$ denotes element-wise multiplication and aggregation across channels.
Min-Max Normalization:
Normalize the localization map to $[0,1]$: $M^c_{\text{init}} \leftarrow \dfrac{M^c_{\text{init}} - \min(M^c_{\text{init}})}{\max(M^c_{\text{init}}) - \min(M^c_{\text{init}})}$.
return M^c_{init}.

Next, we fuse the feature information from different skip connections by averaging. The fused features $\overline{T_{norm}}$, and their gradients $\overline{\nabla T^{l,c}_{norm}}$ are computed as follows:

$$\overline{T_{norm}} = \frac{1}{N_{skip}} \sum_{l=N-N_{skip}-1}^{N-1} T_{norm}^{l}, \quad \overline{\nabla T_{norm}^{l,c}} = \frac{1}{N_{skip}} \sum_{l=N-N_{skip}-1}^{N-1} \nabla T_{norm}^{l,c}, \tag{2}$$

where N_{skip} denotes the number of blocks involved in the skip connections. Notably, for $l = N - 1$, it is a normal connection.

We compute linear combination weights $T_{norm,wgt}$ by averaging $\overline{\nabla T_{norm}^{l,c}}$ along the height and width dimensions. The linear combination weights of the features $\overline{T_{norm}}$, $T_{norm,wgt}$, can be calculated as

$$T_{norm,wgt} = \frac{1}{hw} \sum_{i} \sum_{j} \overline{\nabla T_{norm,i,j}^{l,c}}, \tag{3}$$

where h and w represent the height and width of the gradients. The initial localization map M_{init}^{c} is computed by

$$M_{init}^{c} = \mathrm{ReLU}(\overline{T_{norm}} \odot T_{norm,wgt}), \tag{4}$$

where $\odot$ denotes element-wise multiplication, and ReLU is used to remove negative weights. Next, we apply min–max normalization to M_{init}^{c}:

$$M_{init}^{c} = \frac{M_{init}^{c} - \min(M_{init}^{c})}{\max(M_{init}^{c}) - \min(M_{init}^{c})}. \tag{5}$$

Each element of the normalized map M_{init}^{c} represents confidence for class c. To mitigate noise, the confidence map M_{conf}^{c} is defined as

$$M_{conf}^{c}(i,j) = \begin{cases} 1, & \text{if } M_{init}^{c}(i,j) > \tau \\ 0, & \text{otherwise} \end{cases}, \tag{6}$$

where τ is the confidence threshold. Following CLIP-ES [12], we refine M_{conf}^{c} using semantic affinity operations to yield the final refined localization map M^{c}.

3.3. Pseudo-Label Refinement Module

First, we generate the initial pseudo-label $M_{p,init}$ by selecting the maximum index along the class dimension from localization maps of all foreground classes.

To avoid blindly generating prompt points, we selectively construct point prompts based on class regions in the initial pseudo-label $M_{p,init}^{c}$ for the target class c. The input image I is divided into an $H_p \times W_p$ regular grid, and the center points of these grids form a candidate set P_{cand}. We filter candidate points within the target class c:

$$P^{c} = p \in P_{cand} \mid M_{p,init}^{c}(p) = True, \tag{7}$$

where P^{c} represents the selected points for the target class c.

These foreground points P^{c} and features F_s are input into the SAM model, producing semantic mask sets M_{sem}^{c}. We filter masks $M_{sem,i}^{c}$ based on intersection-over-union (IoU) with the following target class pseudo-label mask:

$$M_{new}^{c} = M_{sem,i}^{c} \mid \mathrm{IoU}(M_{sem,i}, M_{p,init}^{c}) > \delta, \tag{8}$$

where δ denotes the selection threshold. Finally, the selected semantic masks M_{new}^{c} are merged to obtain the final refined pseudo-label M_{p}^{c} for the target class c.

4. Experiment

4.1. Datasets

We evaluate the proposed method on two widely used benchmarks for WSSS: PASCAL VOC 2012 [39] (VOC 2012) and MS COCO 2014 [40] (COCO 2014). The PASCAL VOC 2012 contains 20 foreground categories and one background class and is split into training, validation, and test sets with 1464, 1449, and 1456 images, respectively. Following prior works [27,28], we train on the augmented training set provided by SBD [41], which consists of 10,582 images. The MS COCO 2014 dataset comprises 81 categories (including one background class), with approximately 80 k training images and 40 k validation images. Many images in this dataset exhibit complex scenes with multiple objects and categories, making it a challenging benchmark for WSSS.

To quantitatively assess segmentation performance, we adopt mIoU as the evaluation metric. mIoU captures the overall segmentation accuracy between foreground classes and background and is the standard metric for measuring the effectiveness of WSSS methods.

4.2. Implementation Details

For localization map generation, we adopt the pre-trained ViT-B/16 CLIP model [11]. The number of transformer blocks N is 12, and the number of skip connections N_{skip} is 3. The confidence threshold τ is set to 0.4 on PASCAL VOC 2012 and to 0.7 on MS COCO 2014. For pseudo-label refinement, we use the pre-trained SAM ViT-B model [13] with an input resolution of 512×512. In our method, both CLIP and SAM are kept frozen. For candidate points, both H_p and W_p are set to 16, and the IoU filtering threshold δ is 0.5.

We integrate the proposed CEP into the end-to-end WeCLIP framework [9] for online pseudo-label generation. Therefore, the WeCLIP is used as the comparative baseline in this paper. Table 1 reports the relevant settings. For training, following the WeCLIP setup, we use batch sizes of 4 and 8 for VOC 2012 and COCO 2014, respectively. Optimization is performed using the AdamW optimizer with an initial learning rate of 2×10^{-3}. Training is conducted for 30,000 iterations on VOC 2012 and 80,000 iterations on COCO 2014, with an input scale of 320×320. During the testing phase, the use of multi-scale inputs and CRF is a common practice in WSSS; for example, ToCo [8], DuPL [35], and WeCLIP [9] all employ these strategies. WeCLIP is used as the baseline in this paper. Specifically, for a fair comparison, we follow the testing settings of WeCLIP, which adopt multi-scale factors of 0.75 and 1.0, along with flipping and DenseCRF [42] post-processing.

Table 1. Implementation configuration details. LR: learning rate. BS: batch size.

Dataset	CLIP	SAM	Optimizer	LR	BS	Seed	Iterations
VOC 2012	ViT-B/16	ViT-B	AdamW	2×10^{-3}	4	1	3×10^4
COCO 2014	ViT-B/16	ViT-B	AdamW	2×10^{-3}	8	1	8×10^4

In addition, it should be noted that the software configuration for the experiments in this study consisted of Python 3.10, PyTorch 2.1, Torchvision 0.16, and CUDA 12.1. The training is conducted on a single NVIDIA RTX 3090Ti.

4.3. Comparison with State-of-the-Art Methods

We conduct comprehensive comparisons between the proposed WeCLIP+CEP method and existing multi-stage as well as end-to-end WSSS approaches on the PASCAL VOC 2012 and MS COCO 2014 datasets. The methods that adopt CLIP belong to the category of language-supervised approaches and are denoted by L. In addition, we also indicate which methods utilize the SAM model.

Table 2 presents the performance comparison between our method and several representative state-of-the-art techniques on the PASCAL VOC 2012 dataset. As shown, the proposed WeCLIP+CEP achieves a 3.1% mIoU improvement over the WeCLIP baseline on the validation set, reaching 79.5% mIoU. Among end-to-end WSSS methods, WeCLIP+CEP sets a new performance record, outperforming recent methods such as MoRe [43], FFR [44], and SeCo [45] by 3.1%, 3.5%, and 5.5% mIoU, respectively. Even when compared with multi-stage methods, WeCLIP+CEP remains highly competitive. In particular, it surpasses the recent SAM-based S2C [15] method by 1.3% mIoU and outperforms CLIP-ES+POT [46] by 3.4% mIoU on the VOC 2012 validation set.

Table 2. Comparison of state-of-the-art approaches on the PASCAL VOC 2012 val set. mIoU (%) as the evaluation metric. Sup. denotes the supervision type. I: image-level supervision. L: language supervision.

Method	Backbone	Sup.	Val
multi-stage weakly supervised methods			
BECO$_{CVPR'23}$ [47]	MIT-B2	I	73.7
ACR$_{CVPR'23}$ [48]	ResNet38	I	72.4
MCTformer$_{CVPR'22}$ [29]	ResNet38	I	70.4
MCTformer + OCR$_{CVPR'23}$ [49]	ResNet38	I	72.7
USAGE$_{ICCV'23}$ [50]	ResNet38	I	71.9
MCTformer + LPCAM$_{CVPR'23}$ [51]	ResNet38	I	72.6
DiG$_{ECCV'24}$ [52]	ResNet38	I	73.9
CTI$_{CVPR'24}$ [2]	ResNet38	I	74.1
MuP-VSS$_{CVPR'25}$ [53]	ResNet38	I	73.6
S2C$_{CVPR'24}$ [15]	ResNet38	I + SAM	**78.2**
ReCAM$_{CVPR'22}$ [27]	ResNet101	I	68.5
AMN$_{CVPR'22}$ [28]	ResNet101	I	69.5
CDL$_{IJCV'23}$ [30]	ResNet101	I	72.4
CLIMS$_{CVPR'22}$ [36]	ResNet101	I + L	70.4
CLIP-ES$_{CVPR'23}$ [12]	ResNet101	I + L	73.8
CLIP-ES + PSDPM$_{CVPR'24}$ [54]	ResNet101	I + L	74.1
CLIP-ES + CPAL$_{CVPR'24}$ [55]	ResNet101	I + L	74.5
CLIP-ES + POT$_{CVPR'25}$ [46]	ResNet101	I + L	76.1
end-to-end weakly supervised methods			
AFA$_{CVPR'22}$ [7]	MIT-B1	I	66.0
TSCD$_{AAAI'23}$ [56]	MIT-B1	I	67.3
1Stage$_{CVPR'20}$ [57]	ResNet38	I	62.7
RRM$_{AAAI'20}$ [31]	ResNet38	I	62.6
SLRNet$_{IJCV'22}$ [58]	ResNet38	I	67.2
ToCo$_{CVPR'23}$ [8]	ViT-B	I	71.1
DuPL$_{CVPR'24}$ [35]	ViT-B	I	73.3
SeCo$_{CVPR'24}$ [45]	ViT-B	I	74.0
MoRe$_{AAAI'25}$ [43]	ViT-B	I	76.4
FFR$_{CVPR'25}$ [44]	ViT-B	I	76.0
WeCLIP$_{CVPR'24}$ [9]	ViT-B	I + L	76.4
WeCLIP + CEP (single-scale)	ViT-B	I + L + SAM	76.3
WeCLIP + CEP (multi-scale)	ViT-B	I + L + SAM	77.3
WeCLIP + CEP (multi-scale+flip)	ViT-B	I + L + SAM	77.8
WeCLIP + CEP (multi-scale+flip+crf)	ViT-B	I + L + SAM	**79.5**

Table 3 shows the results on the more challenging MS COCO 2014 dataset. WeCLIP+CEP achieves 48.6% mIoU on this dataset, yielding a 1.5% improvement over the WeCLIP baseline. This gain is consistent with the improvement trend observed on VOC, highlighting the cross-dataset generalization ability of the proposed approach. Compared

with other end-to-end methods, WeCLIP+CEP also attains the highest accuracy, surpassing FFR and MoRe by 1.8% and 1.2% mIoU, respectively. Furthermore, our method outperforms many multi-stage approaches, achieving competitive performance.

Table 3. Comparison with other state-of-the-art methods on MS COCO 2014 val set. mIoU (%) as the evaluation metric. Sup. denotes the supervision type. I: image-level supervision. L: language supervision.

Method	Backbone	Sup.	Val
multi-stage weakly supervised methods			
MCTformer$_{CVPR'22}$ [29]	ResNet38	I	42.0
MCTformer + OCR$_{CVPR'23}$ [49]	ResNet38	I	42.5
MCTformer + LPCAM$_{CVPR'23}$ [51]	ResNet38	I	42.8
DiG$_{ECCV'24}$ [52]	ResNet38	I	45.5
MuP-VSS$_{CVPR'25}$ [53]	ResNet38	I	46.6
S2C$_{CVPR'24}$ [15]	ResNet38	I + SAM	**49.8**
ReCAM$_{CVPR'22}$ [27]	ResNet101	I	42.9
AMN$_{CVPR'22}$ [28]	ResNet101	I	44.7
USAGE$_{ICCV'23}$ [50]	ResNet101	I	44.3
CDL$_{IJCV'23}$ [30]	ResNet101	I	45.5
BECO$_{CVPR'23}$ [47]	ResNet101	I	45.1
CLIP-ES$_{CVPR'23}$ [12]	ResNet101	I + L	45.4
CTI$_{CVPR'24}$ [2]	ResNet101	I + L	45.4
CLIP-ES + PSDPM$_{CVPR'24}$ [54]	ResNet101	I + L	47.2
CLIP-ES + CPAL$_{CVPR'24}$ [55]	ResNet101	I + L	46.8
CLIP-ES+POT$_{CVPR'25}$ [46]	ResNet101	I + L	47.9
end-to-end weakly supervised methods			
AFA$_{CVPR'22}$ [7]	MIT-B1	I	38.9
TSCD$_{AAAI'23}$ [56]	MIT-B1	I	40.1
SLRNet$_{IJCV'22}$ [58]	ResNet38	I	35.0
ToCo$_{CVPR'23}$ [8]	ViT-B	I	42.3
DuPL$_{CVPR'24}$ [35]	ViT-B	I	44.6
SeCo$_{CVPR'24}$ [45]	ViT-B	I	46.7
MoRe$_{AAAI'25}$ [43]	ViT-B	I	47.4
FFR$_{CVPR'25}$ [44]	ViT-B	I	46.8
WeCLIP$_{CVPR'24}$ [9]	ViT-B	I + L	47.1
WeCLIP + CEP (single-scale)	ViT-B	I + L + SAM	46.4
WeCLIP + CEP (multi-scale)	ViT-B	I + L + SAM	47.3
WeCLIP + CEP (multi-scale+flip)	ViT-B	I + L + SAM	47.7
WeCLIP + CEP (multi-scale+flip+crf)	ViT-B	I + L + SAM	**48.6**

Overall, across both datasets, WeCLIP+CEP consistently surpasses the WeCLIP baseline and pushes the performance boundary of end-to-end WSSS. These results demonstrate that integrating WeCLIP with the proposed CEP module yields substantial performance gains.

Table 4 presents the confidence intervals of the segmentation results for our method on the VOC 2012 and COCO 2014 validation sets. It can be observed that the 95% confidence intervals of our method on both datasets are entirely above those of the baseline, thereby confirming that our method achieves superior segmentation performance compared to the baseline.

Table 4. Confidence intervals of segmentation results on the validation sets of VOC 2012 and COCO 2014 datasets. CI: confidence interval.

Dataset	95% CI
VOC 2012	[78.3%, 80.8%]
COCO 2014	[48.1%, 49.0%]

Furthermore, we employ the bootstrap method to perform a statistical significance test to determine whether the performance improvement of our method over the baseline model is statistically significant. The results on the VOC 2012 validation set show that $p < 0.05$, further verifying that the segmentation performance of our method is significantly better than that of the baseline.

4.4. Visualizations

To validate the effectiveness of the proposed method, Figure 2 presents a qualitative comparison of semantic segmentation results between WeCLIP+CEP and the baseline WeCLIP. As can be observed, compared with the baseline, the proposed approach produces more accurate segmentation maps with richer target details. For example, in the *horse* image of the first column in Figure 2, our method successfully segments the horse's legs with higher accuracy, whereas the baseline only captures part of them. In the *bird* image of the second column, the proposed method also yields more precise predictions for the target's head. These observations demonstrate that our method can effectively enhance the quality of semantic segmentation results.

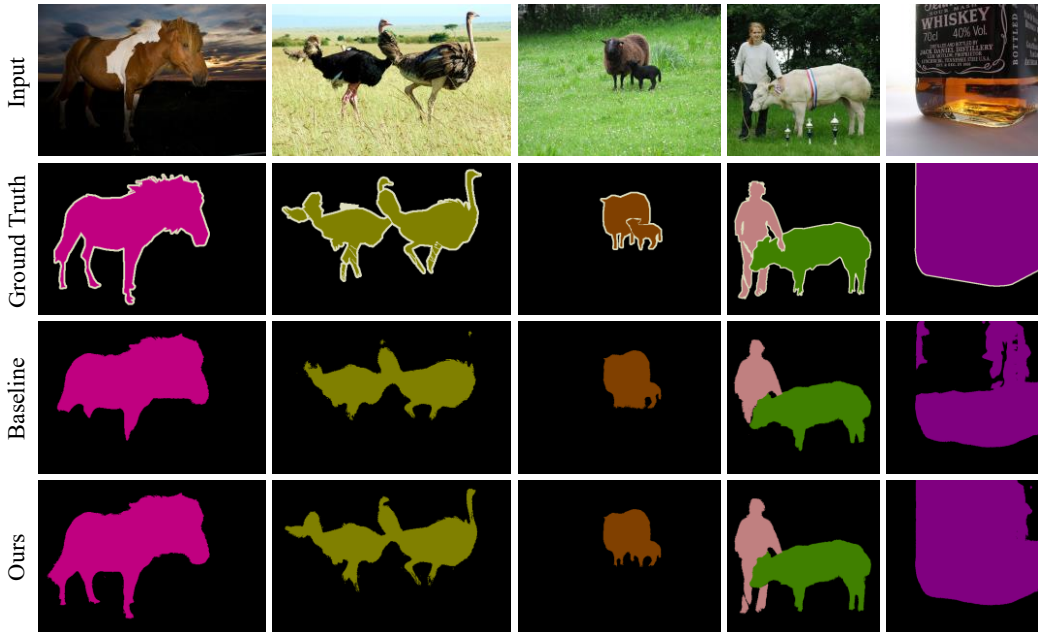


Figure 2. Visualization comparison of semantic segmentation results between our CEP and the baseline.

4.5. Ablation Studies

Effect of Components. We conduct ablation studies on the VOC 2012 training set to evaluate the quality of pseudo-labels and quantitatively validate the effectiveness of each component in our proposed method. Unless otherwise specified, all experiments adopt the pre-trained weights of WeCLIP to eliminate the influence of model performance on the quality of pseudo-labels.

Table 5 demonstrates the effects of the pseudo-label refinement and Skip-CAM modules. Since VOC 2012 does not provide ready-made boundary annotations, for mean

Boundary IoU (mBIoU) evaluation, we apply a morphological erosion operation to the given class segmentation masks. The boundary of each class is then obtained by computing the difference between the original segmentation region and its eroded version, where the kernel size of erosion is set to 23. As shown in the table, the pseudo-label refinement module based on SAM brings improvements of 2.5% in mIoU and 10.1% in mBIoU. This demonstrates that the SAM refinement module not only enhances the accuracy of pseudo-labels but also substantially improves their quality, thereby achieving the purpose of refinement. Furthermore, the Skip-CAM module further contributes an additional improvement of 1.4% in mIoU and 0.5% in mBIoU, as it enriches the semantic information of features through skip connections, leading to more precise localization maps and further improving the quality of pseudo-labels.

Table 5. Ablation study of each component in our CEP on PASCAL VOC 2012 training set. PRM: pseudo-label refinement module. BIoU: boundary IoU.

PRM	Skip-CAM	mIoU (%)	mBIoU (%)
		75.4	32.6
✓		77.9	42.7
✓	✓	79.3	43.2

To qualitatively validate the effectiveness of our CEP method, we conduct a visual comparison of pseudo-labels generated by CEP and WeCLIP [9]. To ensure a fair comparison of pseudo-label quality, we adopt the original training weights of WeCLIP, thereby eliminating the influence of model performance differences on the pseudo-labels. The experiments are conducted on the VOC 2012 training set.

Figure 3 presents the visualization results. Compared with the baseline method (i.e., WeCLIP), pseudo-labels produced by CEP exhibit higher quality and more precise object boundaries. For instance, in the *sheep* image of the second column, the pseudo-label generated by CEP delineates the head contour with greater clarity. This demonstrates that our approach can effectively improve pseudo-label quality. Furthermore, in the *cat* image of the first column, the pseudo-label generated by CEP not only captures more accurate object boundaries but also covers a more complete target region.

Effect of SAM Variants. Table 6 presents a comparison of the pseudo-label quality generated by the proposed CEP combined with different versions of SAM on the VOC 2012 training set. The experimental results show that as the scale of the SAM model increases, the mIoU score of the generated pseudo labels gradually improves. However, larger-scale SAM models also significantly increase the number of parameters, leading to higher computational overhead. Considering both performance and efficiency, we adopt the ViT-B version of SAM as the default configuration in this work.

Table 6. Effect of different SAM variants on the quality of pseudo labels on the VOC 2012 training set.

Variants	ViT-B	ViT-L	ViT-H
mIoU (%)	79.3	81.8	82.2

Effect of N_{skip}. Furthermore, we compare the impact of different numbers of skip connections N_{skip} on the quality of localization maps. Table 7 reports the effect of the number of skip connections N_{skip} in the Skip-CAM module, as defined in Equation (2). Experimental results show that increasing the number of skip connections to 3 leads to a significant performance boost, owing to the fusion of richer semantic information. However, when the number of skip connections exceeds 3, the improvement becomes marginal or even

degrades. This may be due to the large distribution gap between features from shallow and deep blocks, which negatively impacts classification performance and introduces noisy gradients, thus affecting the quality of localization maps. Additionally, increasing the number of skip connections incurs higher computational costs. Therefore, in this work, we set N_{skip} to 3.

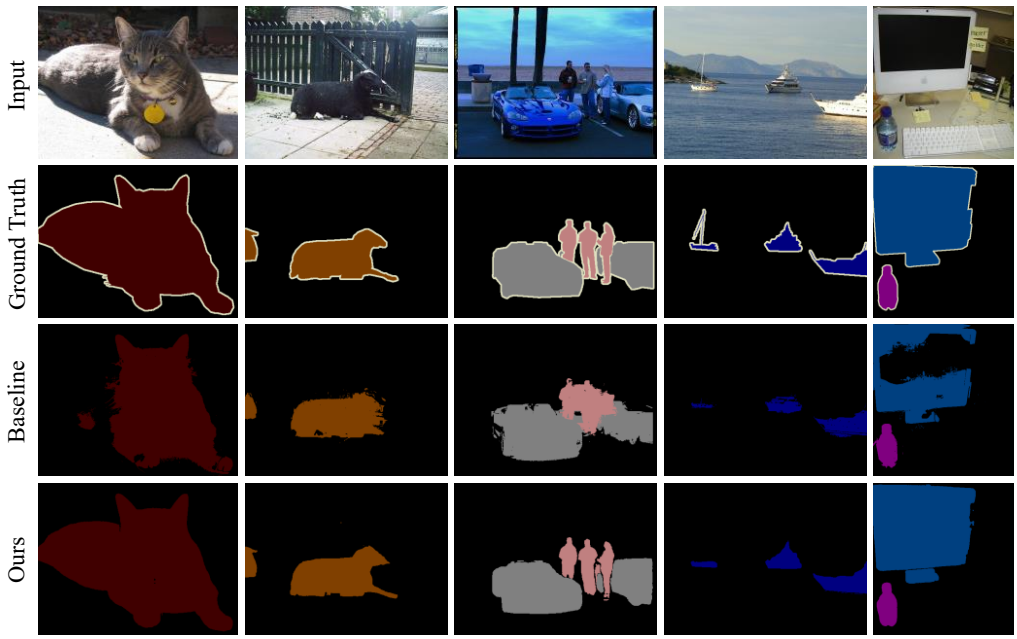


Figure 3. Visualization comparison of pseudo labels between our CEP and the baseline.

Table 7. Ablation study about skip connection numbers in N_{skip} of Equation (2) on PASCAL VOC 2012 training set.

N_{skip}	1	2	3	4	5	6
mIoU (%)	73.53	74.10	74.24	74.25	74.25	73.95

Effect of Candidate Point Settings. In the pseudo-label refinement module, the image region is first divided into $H_p \times W_p$ grids, with the center of each grid taken as a candidate point. Subsequently, the target-class regions in the initial pseudo labels are used as guidance to filter these candidate points. Table 8 presents the effect of different candidate point settings on pseudo-label generation performance on the VOC 2012 training set. The results show that when $H_p \times W_p$ increases from 8×8 to 32×32, the mIoU of the pseudo labels improves from 76.6% to 80.0%, but the average computation time per image also increases from 0.07 s to 0.19 s. In other words, within a certain range, denser candidate point settings lead to higher mIoU but also significantly higher computational costs. Considering both performance and efficiency, $H_p \times W_p$ is set to 16×16 in this work.

Table 8. Effect of different candidate point settings on pseudo-label generation performance on the VOC 2012 training set.

Setting	8×8	16×16	32×32
mIoU (%)	76.6	79.3	80.0
Time (s)	0.07	0.09	0.19

Effect of Prompt Settings. Both our method and the baseline (i.e., WeCLIP [9]) follow CLIP-ES [12] to design text prompts for CLIP. Taking VOC 2012 as an example, the background

class set includes [*ground*, *land*, *grass*, *tree*, *building*, *wall*, *sky*, *lake*, *water*, *river*, *sea*, *railway*, *railroad*, *keyboard*, *helmet*, *cloud*, *house*, *mountain*, *ocean*, *road*, *rock*, *street*, *valley*, *bridge*, *sign*], while the foreground class set includes [*aeroplane*, *bicycle*, *bird avian*, *boat*, *bottle*, *bus*, *car*, *cat*, *chair seat*, *cow*, *diningtable*, *dog*, *horse*, *motorbike*, {*person with clothes*, *people*, *human*}, *pottedplant*, *sheep*, *sofa*, *train*, *tvmonitor screen*]. The prompt prefix set is [*a clean origami {}.*], and we denote the prompt settings from CLIP-ES as Prompt1.

Following the text prompt templates adopted in CLIP, we replace the prompt prefix set in Prompt1 with [*itap of a {}.*, *a bad photo of the {}.*, *a origami {}.*, *a photo of the large {}.*, *a {} in a video game.*, *art of the {}.*, *a photo of the small {}.*], denoted as Prompt2. On the basis of Prompt1, we further add synonyms or attributes to foreground classes, e.g., adding the synonym *airplane* to *aeroplane* and the attribute *wings* to *bird*, denoted as Prompt3. Finally, we combine the settings of Prompt2 and Prompt3 to obtain Prompt4.

Table 9 reports the pseudo-label generation performance under different prompt settings. As observed, the text prompt setting from CLIP-ES (Prompt1) achieves the best performance. The use of synonyms or attributes for class labels in Prompt3 yields slightly worse performance than Prompt1. Introducing diverse prompt prefixes in Prompt2 leads to a noticeable performance drop, and the performance of Prompt4 is almost identical to that of Prompt2. Therefore, both our work and the baseline adopt the text prompt setting from CLIP-ES.

Table 9. Effect of different prompt settings on pseudo-label generation performance on the VOC 2012 training set.

Setting	**Prompt1**	**Prompt2**	**Prompt3**	**Prompt4**
mIoU (%)	79.30	78.68	79.15	78.69

Effect of τ. Table 10 presents the impact of the hyperparameter τ in Equation (6). The value of τ increases from 0.1 to 0.9 with a step size of 0.1. The mIoU of the pseudo-labels first increases and then decreases as τ becomes larger. A larger τ implies that fewer pixels are retained in the initial localization maps. When $\tau = 0.6$, the proposed method achieves the highest pseudo-label accuracy on the VOC 2012 training set, reaching 79.4% mIoU. For a fair comparison, following the settings of WeCLIP, we set $\tau = 0.4$ for VOC 2012 and $\tau = 0.7$ for COCO 2014.

Table 10. Ablation study of the hyperparameter τ with mIoU.

τ	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**
mIoU (%)	77.1	77.9	78.6	79.3	79.3	79.4	78.0	76.1	74.5

Effect of δ. Table 11 shows the impact of the hyperparameter δ in Equation (8). We incrementally increased δ from 0.1 to 0.9 with a step size of 0.1 and evaluated the variation in mIoU of the pseudo-labels generated by the proposed method on the VOC 2012 training set. The experimental results indicate that as δ increases, the mIoU first rises and then falls. When $\delta = 0.4$ and $\delta = 0.5$, the difference in mIoU is only 0.1%, which is almost negligible. Although $\delta = 0.4$ shows slightly better performance, considering the generality and robustness of the method, we finally set $\delta = 0.5$ for both the VOC 2012 and COCO 2014 datasets.

Table 11. Ablation study of the hyperparameter δ with mIoU.

δ	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**
mIoU (%)	73.2	77.4	78.8	79.4	79.3	78.9	77.9	76.4	71.8

5. Training Cost and Performance

In Table 12, we compare the training cost and performance of different variants of our method with baseline methods on the VOC 2012 dataset. Our method that employs both the pseudo-label refinement and Skip-CAM modules is denoted as WeCLIP+CEP-Normal. Our method that uses only the pseudo-label refinement module is denoted as WeCLIP+CEP-SAM. For a fairer comparison, both the baseline and WeCLIP+CEP-SAM adopt the modified CAM construction code when evaluating training cost.

Table 12. Training cost and performance on the PASCAL VOC 2012 dataset. All methods are run on a single NVIDIA RTX 3090Ti GPU. CEP-SAM: CEP with the pseudo-label refinement module. CEP-Normal: CEP with both the pseudo-label refinement and Skip-CAM modules. Time: training time. Mem: GPU memory usage. TPIter: average time per iteration. TPImg: average time per image.

Method	Params	FLOPs	Time	Mem	TPIter	TPImg	mIoU
WeCLIP (Baseline)	155.6 M	0.3 T	1.1 h	4.9 G	0.13 s	0.03 s	76.4%
WeCLIP + CEP-SAM	247.0 M	1.6 T	2.6 h	13.3 G	0.31 s	0.08 s	79.0%
WeCLIP + CEP-Normal	247.3 M	1.6 T	3.1 h	13.7 G	0.38 s	0.09 s	79.5%

It can be observed that WeCLIP+CEP-Normal achieves the highest mIoU score, with parameters, FLOPs, training time, and GPU memory usage of 247.3 M, 1.6 T, 3.1 h, and 13.7 G, respectively. The corresponding average time per iteration is 0.37 s, the average processing time per image is 0.09 s. The high training cost of WeCLIP + CEP-Normal is due to the use of the pseudo-label refinement module based on SAM and the Skip-CAM module with multiple skip connections. WeCLIP + CEP-SAM, which only employs the pseudo-label refinement module based on SAM, exhibits lower performance and training cost than WeCLIP + CEP-Normal but higher than the baseline. The baseline, which lacks both the pseudo-label refinement and Skip-CAM modules, has the lowest training cost, with parameters, FLOPs, training time, and GPU memory usage of 155.6 M, 0.3 T, 1.1 h, and 4.9 G, respectively, while also achieving the lowest performance. The corresponding average time per iteration is 0.13 s, the average processing time per image is 0.03 s.

6. Limitations and Failure Modes

The main limitation of our method lies in its significantly higher training cost compared to the baseline. In Section 5, we discuss the training cost of our method and the baseline in detail. Our method requires 247.3 M parameters, 1.6 T FLOPs, 3.1 h of training time, and 13.7G GPU memory. This substantial cost primarily stems from the pseudo-label refinement module based on SAM and the Skip-CAM module with multi-hop connections. Next, we discuss the failure modes of the proposed method, such as co-occurring scenes.

We visualized several failure modes of our method (including thin structures, objects with strong co-occurrence, and small objects) on samples from the VOC 2012 training set and further analyzed the potential causes. As shown in Figure 4, in the *chair* scene, our method failed to effectively capture its slender structure. In the *person* and *motorbike* co-occurrence scene, pseudo-labels incorrectly classified most of the *person* region as *motorbike*. In the small object scene, the method mistakenly segmented part of the background as foreground.

For errors involving thin structures, we consider one possible reason to be the insufficient segmentation capability of the SAM model for fine structures. For errors in co-occurrence scenes, the primary reason lies in the network's difficulty in encoding highly discriminative feature information for different object categories, making accurate localization challenging. For small object scenes, the errors stem from the network's limited ability to resolve fine details, which easily leads to confusion between foreground and background; this bias is further amplified when combined with the SAM model.

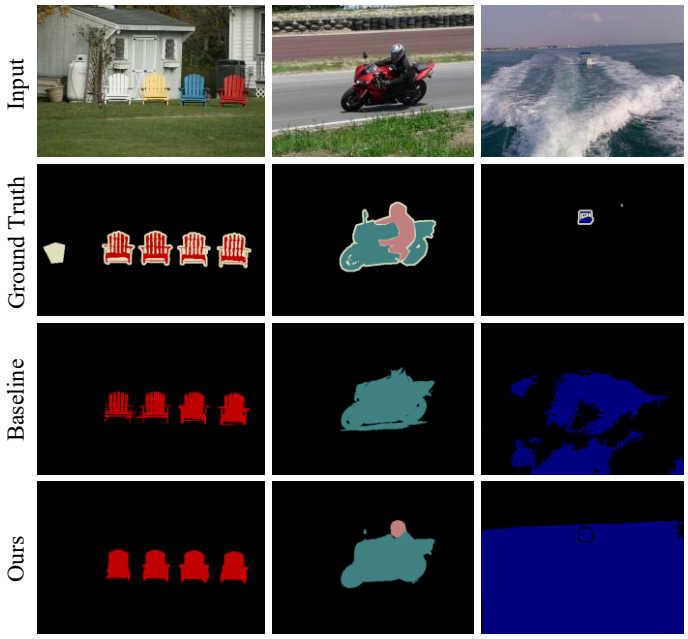


Figure 4. Failure modes of pseudo labels generated by the proposed method.

7. Future Work

Diffusion models, which learn to restore images perturbed by successive Gaussian noise during training, have been widely applied in text-to-image generation tasks. Compared to the weaker local localization capability of ViTs, diffusion models are able to generate higher-quality high-level semantic information. Leveraging the strong performance of foundation models to advance WSSS, we plan to introduce diffusion models into WSSS in future work to further improve segmentation performance. However, efficiently integrating diffusion models with ViT features remains an urgent problem to be addressed.

In addition, we plan to investigate the applicability of CEP for refining pseudo labels in other domains, such as WSSS of medical images. Medical image segmentation plays a crucial role in quantitative studies of diseases and clinical diagnosis. However, pixel-level annotations for medical images also demand extensive human labor. WSSS methods demonstrate certain advantages in medical image segmentation, as they can alleviate the reliance on pixel-level annotations. Nevertheless, compared with natural images, medical image segmentation faces more severe co-occurrence issues. For instance, certain organs almost always appear together with others.

8. Conclusions

In this paper, we propose a pseudo-label enhancement method, CEP, which can online generate high-quality pseudo labels for end-to-end WSSS. The method leverages a foundation model to improve pseudo labels. Specifically, CEP extracts richer feature representations from CLIP via a skip-connection strategy to produce high-quality localization maps. Then, based on the target-class regions provided by the initial pseudo labels generated from the localization maps, SAM is guided to refine the pseudo labels, yielding more accurate labels for model training. Ablation studies validate the effectiveness of each core component in CEP. Experimental results show that, on the VOC 2012 and COCO 2014 datasets, our method outperforms the latest state-of-the-art methods in the end-to-end WSSS. Specifically, on VOC 2012, our method outperforms SeCo, MoRe, and FFR by 5.5%, 3.1%, and 3.5% mIoU, respectively. On COCO 2014, our method surpasses SeCo, MoRe, and FFR by 1.9%, 1.2%, and 1.8% mIoU, respectively.

Author Contributions: Conceptualization, X.Z.; methodology, X.Z.; investigation, X.Z.; writing—original draft preparation, X.Z.; writing—review and editing, Z.T.; supervision, Z.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Zhejiang Provincial Natural Science Foundation of China under Grant LZ20F010008.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Conflicts of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Ahn, J.; Cho, S.; Kwak, S. Weakly Supervised Learning of Instance Segmentation With Inter-Pixel Relations. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 2204–2213. [CrossRef]
2. Yoon, S.H.; Kwon, H.; Kim, H.; Yoon, K.J. Class Tokens Infusion for Weakly Supervised Semantic Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 17–21 June 2024; pp. 3595–3605. [CrossRef]
3. Oh, Y.; Kim, B.; Ham, B. Background-Aware Pooling and Noise-Aware Loss for Weakly-Supervised Semantic Segmentation. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Virtual, 16–20 June 2021; pp. 6909–6918. [CrossRef]
4. Lee, J.; Yi, J.; Shin, C.; Yoon, S. BBAM: Bounding Box Attribution Map for Weakly Supervised Semantic and Instance Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Virtual, 16–20 June 2021; pp. 2643–2652. [CrossRef]
5. Xu, J.; Zhou, C.; Cui, Z.; Xu, C.; Huang, Y.; Shen, P.; Li, S.; Yang, J. Scribble-Supervised Semantic Segmentation Inference. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Virtual, 11–17 October 2021; pp. 15354–15363. [CrossRef]
6. Pan, Z.; Sun, H.; Jiang, P.; Li, G.; Tu, C.; Ling, H. CC4S: Encouraging Certainty and Consistency in Scribble-Supervised Semantic Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 8918–8935. [CrossRef] [PubMed]
7. Ru, L.; Zhan, Y.; Yu, B.; Du, B. Learning Affinity from Attention: End-to-End Weakly-Supervised Semantic Segmentation with Transformers. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 19–24 June 2022; pp. 16825–16834. [CrossRef]
8. Ru, L.; Zheng, H.; Zhan, Y.; Du, B. Token Contrast for Weakly-Supervised Semantic Segmentation. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023; pp. 3093–3102. [CrossRef]
9. Zhang, B.; Yu, S.; Wei, Y.; Zhao, Y.; Xiao, J. Frozen CLIP: A Strong Backbone for Weakly Supervised Semantic Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 17–21 June 2024; pp. 3796–3806.
10. Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; Torralba, A. Learning Deep Features for Discriminative Localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016.
11. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning Transferable Visual Models From Natural Language Supervision. In Proceedings of the 38th International Conference on Machine Learning PMLR, Virtual, 18–24 July 2021; Meila, M., Zhang, T., Eds.; PMLR: New York, NY, USA, 2021; Volume 139, pp. 8748–8763.
12. Lin, Y.; Chen, M.; Wang, W.; Wu, B.; Li, K.; Lin, B.; Liu, H.; He, X. CLIP Is Also an Efficient Segmenter: A Text-Driven Approach for Weakly Supervised Semantic Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023; pp. 15305–15314.
13. Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A.C.; Lo, W.Y.; et al. Segment Anything. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2–6 October 2023; pp. 4015–4026.
14. Yang, X.; Gong, X. Foundation Model Assisted Weakly Supervised Semantic Segmentation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 4–8 January 2024; pp. 523–532.
15. Kweon, H.; Yoon, K.J. From SAM to CAMs: Exploring Segment Anything Model for Weakly Supervised Semantic Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 17–21 June 2024; pp. 19499–19509.

16. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–25 June 2005; Volume 1, pp. 886–893. [CrossRef]
17. Lowe, D.G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [CrossRef]
18. Fulkerson, B.; Vedaldi, A.; Soatto, S. Class segmentation and object localization with superpixel neighborhoods. In Proceedings of the 2009 IEEE 12th International Conference on Computer Vision, Miami, FL, USA, 20–25 June 2009; pp. 670–677. [CrossRef]
19. Shotton, J.; Johnson, M.; Cipolla, R. Semantic texton forests for image categorization and segmentation. In Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, USA, 23–28 June 2008; pp. 1–8. [CrossRef]
20. Long, J.; Shelhamer, E.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015.
21. Liang-Chieh, C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A. Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs. In Proceedings of the International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015.
22. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. International Conference on Learning Representations, Virtual, 3–7 May 2021.
23. Wang, W.; Xie, E.; Li, X.; Fan, D.P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; Shao, L. Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction Without Convolutions. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Virtual, 11–17 October 2021; pp. 568–578.
24. Cheng, B.; Misra, I.; Schwing, A.G.; Kirillov, A.; Girdhar, R. Masked-Attention Mask Transformer for Universal Image Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 19–24 June 2022; pp. 1290–1299.
25. Ahn, J.; Kwak, S. Learning Pixel-Level Semantic Affinity with Image-Level Supervision for Weakly Supervised Semantic Segmentation. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 4981–4990. [CrossRef]
26. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016.
27. Chen, Z.; Wang, T.; Wu, X.; Hua, X.S.; Zhang, H.; Sun, Q. Class Re-Activation Maps for Weakly-Supervised Semantic Segmentation. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 19–24 June 2022; pp. 959–968. [CrossRef]
28. Lee, M.; Kim, D.; Shim, H. Threshold Matters in WSSS: Manipulating the Activation for the Robust and Accurate Segmentation Model Against Thresholds. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 19–24 June 2022; pp. 4320–4329. [CrossRef]
29. Xu, L.; Ouyang, W.; Bennamoun, M.; Boussaid, F.; Xu, D. Multi-class Token Transformer for Weakly Supervised Semantic Segmentation. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 19–24 June 2022; pp. 4300–4309. [CrossRef]
30. Zhang, B.; Xiao, J.; Wei, Y.; Zhao, Y. Credible Dual-Expert Learning for Weakly Supervised Semantic Segmentation. *Int. J. Comput. Vis.* **2023**, *131*, 1892–1908. [CrossRef]
31. Zhang, B.; Xiao, J.; Wei, Y.; Sun, M.; Huang, K. Reliability Does Matter: An End-to-End Weakly Supervised Semantic Segmentation Approach. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020, Volume 34, pp. 12765–12772. [CrossRef]
32. Wu, Z.; Shen, C.; van den Hengel, A. Wider or Deeper: Revisiting the ResNet Model for Visual Recognition. *Pattern Recognit.* **2019**, *90*, 119–133. [CrossRef]
33. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.u.; Polosukhin, I. Attention is All you Need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30.
34. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. In Proceedings of the Advances in Neural Information Processing Systems, Virtual, 6–14 December 2021; Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2021; Volume 34, pp. 12077–12090.
35. Wu, Y.; Ye, X.; Yang, K.; Li, J.; Li, X. DuPL: Dual Student with Trustworthy Progressive Learning for Robust Weakly Supervised Semantic Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 17–21 June 2024; pp. 3534–3543.

36. Xie, J.; Hou, X.; Ye, K.; Shen, L. CLIMS: Cross Language Image Matching for Weakly Supervised Semantic Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 19–24 June 2022; pp. 4483–4492.
37. Zhu, L.; Wang, X.; Feng, J.; Cheng, T.; Li, Y.; Jiang, B.; Zhang, D.; Han, J. WeakCLIP: Adapting CLIP for Weakly-Supervised Semantic Segmentation. *Int. J. Comput. Vis.* **2024**, *133*, 1085–1105. [CrossRef]
38. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017.
39. Everingham, M.; Gool, L.; Williams, C.K.; Winn, J.; Zisserman, A. The Pascal Visual Object Classes (VOC) Challenge. *Int. J. Comput. Vis.* **2010**, *88*, 303–338. [CrossRef]
40. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. In Proceedings of the European Conference on Computer Vision (ECCV), Zurich, Switzerland, 6–12 September 2014; pp. 740–755.
41. Hariharan, B.; Arbeláez, P.; Bourdev, L.; Maji, S.; Malik, J. Semantic contours from inverse detectors. In Proceedings of the 2011 International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; pp. 991–998. [CrossRef]
42. Kraehenbuehl, P.; Koltun, V. Parameter Learning and Convergent Inference for Dense Random Fields. In Proceedings of the 30th International Conference on Machine Learning, Atlanta, GA, USA, 17–19 June 2013; Dasgupta, S., McAllester, D., Eds.; PMLR: New York, NY, USA, 2013; Volume 28, pp. 513–521.
43. Yang, Z.; Meng, Y.; Fu, K.; Wang, S.; Song, Z. MoRe: Class Patch Attention Needs Regularization for Weakly Supervised Semantic Segmentation. Proceedings of the AAAI Conference on Artificial Intelligence, Philadelphia, PA, USA, 25 February–4 March 2025; Volume 39, pp. 9400–9408. [CrossRef]
44. Yang, Z.; Zhao, X.; Wang, X.; Zhang, Q.; Xiao, J. FFR: Frequency Feature Rectification for Weakly Supervised Semantic Segmentation. In Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), Nashville, TN, USA, 11–15 June 2025; pp. 30261–30270.
45. Yang, Z.; Fu, K.; Duan, M.; Qu, L.; Wang, S.; Song, Z. Separate and Conquer: Decoupling Co-occurrence via Decomposition and Representation for Weakly Supervised Semantic Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 17–21 June 2024; pp. 3606–3615.
46. Wang, J.; Dai, T.; Zhang, B.; Yu, S.; Lim, E.G.; Xiao, J. POT: Prototypical Optimal Transport for Weakly Supervised Semantic Segmentation. In Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), Nashville, TN, USA, 11–15 June 2025; pp. 15055–15064.
47. Rong, S.; Tu, B.; Wang, Z.; Li, J. Boundary-Enhanced Co-Training for Weakly Supervised Semantic Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023; pp. 19574–19584.
48. Kweon, H.; Yoon, S.H.; Yoon, K.J. Weakly Supervised Semantic Segmentation via Adversarial Learning of Classifier and Reconstructor. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023; pp. 11329–11339.
49. Cheng, Z.; Qiao, P.; Li, K.; Li, S.; Wei, P.; Ji, X.; Yuan, L.; Liu, C.; Chen, J. Out-of-Candidate Rectification for Weakly Supervised Semantic Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023; pp. 23673–23684.
50. Peng, Z.; Wang, G.; Xie, L.; Jiang, D.; Shen, W.; Tian, Q. USAGE: A Unified Seed Area Generation Paradigm for Weakly Supervised Semantic Segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2–6 October 2023; pp. 624–634.
51. Chen, Z.; Sun, Q. Extracting Class Activation Maps From Non-Discriminative Features As Well. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023; pp. 3135–3144.
52. Yoon, S.H.; Kwon, H.; Jeong, J.; Park, D.; Yoon, K.J. Diffusion-Guided Weakly Supervised Semantic Segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Milan, Italy, 29 September–4 October 2024; pp. 393–411.
53. Duan, S.; Yang, X.; Wang, N. Multi-Label Prototype Visual Spatial Search for Weakly Supervised Semantic Segmentation. In Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), Nashville, TN, USA, 11–15 June 2025; pp. 30241–30250.
54. Zhao, X.; Yang, Z.; Dai, T.; Zhang, B.; Xiao, J. PSDPM: Prototype-based Secondary Discriminative Pixels Mining for Weakly Supervised Semantic Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 17–21 June 2024; pp. 3437–3446.
55. Tang, F.; Xu, Z.; Qu, Z.; Feng, W.; Jiang, X.; Ge, Z. Hunting Attributes: Context Prototype-Aware Learning for Weakly Supervised Semantic Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 17–21 June 2024; pp. 3324–3334.

56. Xu, R.; Wang, C.; Sun, J.; Xu, S.; Meng, W.; Zhang, X. Self Correspondence Distillation for End-to-End Weakly-Supervised Semantic Segmentation. Proceedings of the AAAI Conference on Artificial Intelligence, Washington, DC, USA, 7–14 February 2023, Volume 37, pp. 3045-3053. [CrossRef]
57. Araslanov, N.; Roth, S. Single-Stage Semantic Segmentation From Image Labels. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020.
58. Pan, J.; Zhu, P.; Zhang, K.; Cao, B.; Wang, Y.; Zhang, D.; Han, J.; Hu, Q. Learning Self-supervised Low-Rank Network for Single-Stage Weakly and Semi-supervised Semantic Segmentation. *Int. J. Comput. Vis.* **2022**, *130*, 1181–1195. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.