Segment Anything Model (SAM) Enhances Pseudo-Labels for Weakly Supervised Semantic Segmentation

Tianle Chen [1*] **Zheda Mai** [1*] **Ruiwen Li**[2] **Wei-Lun Chao** [1]
[1]The Ohio State University
[2]EAIGLE Inc
(mai.145, chen.9471)@osu.edu, ruiwen@eaigle.com, chao.209@osu.edu

Abstract

Weakly supervised semantic segmentation (WSSS) aims to bypass the need for laborious pixel-level annotation by using only image-level annotation. Most existing methods rely on Class Activation Maps (CAM) to derive pixel-level pseudo-labels and use them to train a fully supervised semantic segmentation model. Although these pseudo-labels are class-aware, indicating the coarse regions for particular classes, they are not object-aware and fail to delineate accurate object boundaries. To address this, we introduce a simple yet effective method harnessing the Segment Anything Model (SAM), a class-agnostic foundation model capable of producing fine-grained instance masks of objects, parts, and subparts. We use CAM pseudo-labels as cues to select and combine SAM masks, resulting in high-quality pseudo-labels that are both class-aware and object-aware. Our approach is highly versatile and can be easily integrated into existing WSSS methods without any modification. Despite its simplicity, our approach shows consistent gain over the state-of-the-art WSSS methods on both PASCAL VOC and MS-COCO datasets.

1 Introduction

Semantic segmentation, a task aiming to assign a semantic label to each image pixel [38], has found wide applications in various fields, such as medical imaging [5], remote sensing [53] and autonomous driving [17]. The success of deep learning techniques and the availability of large-scale pixel-level annotations have greatly boosted the performance of semantic segmentation in recent years [27]. However, acquiring pixel-level annotations is daunting due to its laborious and costly nature. As an alternative, weakly supervised semantic segmentation (WSSS) seeks to train a segmentation model with cheaper yet weaker annotations such as bounding boxes [39, 29, 44, 45], scribbles [34], points [24, 6], and image-level class labels [28, 55, 19, 48]. Among existing approaches, image-level WSSS has gained widespread popularity due to the abundance of image-level annotations online or in various vision datasets [41, 36] and the availability of strong pre-trained classifiers [18, 42].

As image-level labels do not provide location information for each object class, most of the existing WSSS methods leverage Class Activation Maps (CAM) [56] to derive location cues. These approaches typically follow a four-stage learning process. First, they train a classification model with image-level labels. Then, based on the intermediate feature maps and their weights to a class, CAMs are generated as the coarse estimate of the class location. Subsequently, the initial CAMs are refined with post-processing techniques, such as pixel affinity-based methods [2, 33] or saliency guidance [30, 7, 52], to create pixel-level pseudo-labels. Finally, a semantic segmentation model [10, 8] is trained using the pseudo-labels as pixel-level supervision. The efficacy of WSSS greatly relies on the accuracy of pseudo-labels. However, it is widely recognized that the CAM-derived pseudo-labels often suffer

*Equal contributions.

37th Conference on Neural Information Processing Systems Workshop (NeurIPS 2023).

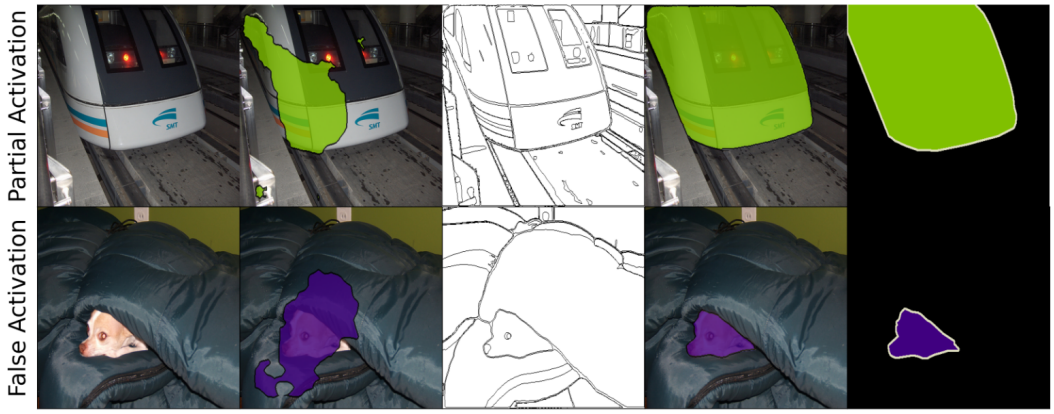


Figure 1: **Illustration of how SAM addresses partial and false activation** on PASCAL VOC 2012 train set: (A) original images; (B) pseudo-labels generated by a SOTA image-level WSSS method, CLIMS [50]; (C) masks from SAM; (D) SAM enhanced pseudo-labels; (E) ground-truth labels.

from *partial activation* [3, 47], activating the most discriminative region instead of the entire object area, and *false activation* [50, 21], wrongly activating the background around the object. (Figure 1 shows partial and false activation examples.) *In other words, CAM-derived pseudo-labels lack the awareness of objects, resulting in poor contours that drastically deviate from object boundaries.*

In this paper, we investigate a novel approach to addressing this issue, directly incorporating object boundary information into pseudo-label generation. Concretely, we leverage the advent of segmentation foundation models [46, 25, 59], in particular, the Segment Anything Model (SAM) [25], which is capable of producing fine-grained, class-agnostic masks of objects, parts, or subparts. We hypothesize that the quality of the resulting pseudo-labels can be greatly enhanced by appropriately integrating the coarse class location from CAM and the object boundary information from SAM.

To this end, we propose **SAM Enhanced pseudo-labels** (**SEPL**). **SEPL** uses CAM-derived pseudo-labels for a particular class as the seed signals to select the most relevant masks from SAM. The union of these masks, which encompass both class and object information, is then treated as the enhanced pseudo-labels for training semantic segmentation models. More specifically, **SEPL** consists of two stages: mask assignment and mask selection (see Figure 2). During **mask assignment**, each SAM mask is assigned to the class (of the image-level annotation) whose CAM-derived pseudo-labels have the largest intersection with the mask. During **mask selection**, SAM masks with substantial overlap with the CAM-derived pseudo-labels are chosen to address *false activation*, given that background masks typically manifest minimal overlap. Meanwhile, we also select SAM masks that substantially encompass the CAM-derived pseudo-labels, targeting the challenge of *partial activation*. Given the precise alignment of SAM masks to object boundaries, we find substantial enhancements in mitigating partial and false activations in the existing pseudo-labels, as depicted in Figure 1.

SEPL is remarkably versatile as it can be seamlessly integrated into existing WSSS methods without modifying the original methods. Despite its simplicity, SEPL achieves a notable improvement in the mean Intersection over Union (mIoU) of pseudo-labels and ground-truth labels compared to eleven state-of-the-art WSSS methods, with an average gain of 5.33% and 3.12% on the train set of PASCAL VOC 2012 dataset [15] and MS COCO 2014 [36], respectively. As far as we know, this is the first study to investigate the potential of SAM in the context of WSSS. We hope this work will pave the way for applying segmentation foundation models in diverse computer vision applications.

Related work. Due to the page limit, we leave it in Appendix A.

2 SAM Enhanced Pseudo Labels

2.1 Preliminary

Following the standard setup, each training image $\boldsymbol{X} \in \mathbb{R}^{H \times W \times C}$ is associated with only an image-level label vector $\boldsymbol{y} = [y_1, y_2, \ldots, y_K]^\top \in \{0, 1\}^K$ for K classes, where $y_k = 1$ indicates the

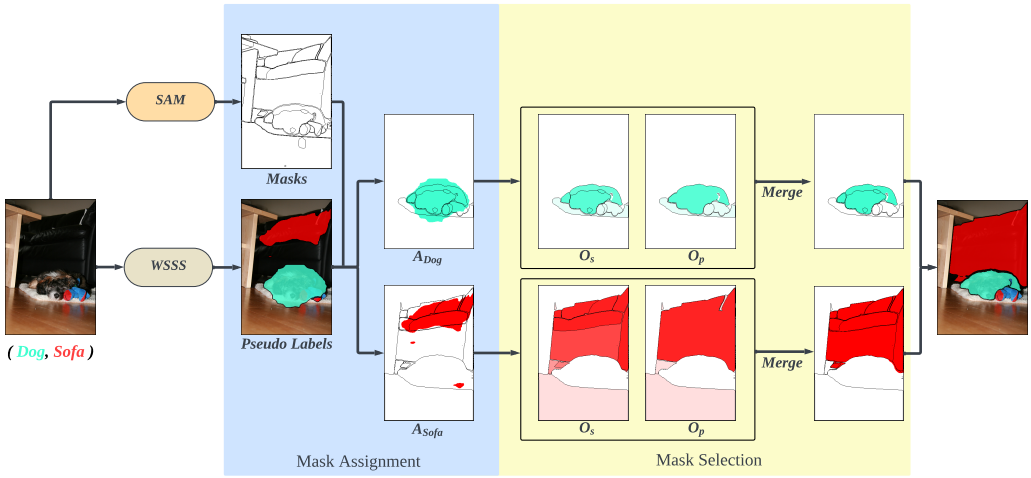


Figure 2: **Illustration of the SEPL pipeline.** SEPL comprises two stages, mask assignment and mask selection. Based on the intersection between each SAM mask and pseudo-labels, a mask is assigned to the class with the largest intersection. For each mask, two metrics are computed: o_s, the fraction of the mask overlapped by pseudo-labels, and o_p, the fraction of pseudo-labels overlapped by the mask. A mask is retained as an enhanced pseudo-label if either metric surpasses the designated threshold.

presence of class k in $\boldsymbol{X}$ and 0 otherwise. Upon training a classifier f with this dataset, WSSS methods feed an image to f and obtain the Class Activation Maps (CAM) $\mathcal{M} = [\boldsymbol{M}_1, \cdots \boldsymbol{M}_K]$, where $\boldsymbol{M}_k \in R^{H \times W}$ highlights the discriminative image regions utilized by f to identify class k. Post-processing techniques, such as AffinityNet [3] and IRNet [1], further refine $\boldsymbol{M}_k$ to produce pseudo-labels $\boldsymbol{P} \in \{0, 1, \cdots, K\}^{H \times W}$, where pixel is mapped to either a class label in $\{1, \cdots, K\}$ or 0 for background regions. $\boldsymbol{P}_k \in \{0, k\}^{H \times W}$ represents the pseudo-labels for class k. Subsequently, a fully supervised semantic segmentation network (e.g., [9, 8]) is trained with $\boldsymbol{P}$. SAM takes an image $\boldsymbol{X} \in \mathbb{R}^{H \times W \times C}$ as input and returns a list of masks capturing either a subpart, a part or an entire object: $\mathcal{S} = [\boldsymbol{S}_0, \boldsymbol{S}_1, \cdots, \boldsymbol{S}_L]$ where $\boldsymbol{S}_l \in \{0, 1\}^{H \times W}$ and L is the number of masks. It is noteworthy that different images may receive various mask quantities from SAM.

2.2 Approach

While CAM-derived pseudo-labels are class-aware and identify discriminative regions for individual classes, they often fail to delineate accurate object boundaries. In contrast, SAM is able to precisely segment most parts or objects in a class-agnostic manner. Bridging their capabilities, we propose SAM Enhanced Pseudo Labels (SEPL) to harness the potential of SAM for pseudo-label enhancement.

Our approach takes pseudo-labels $[\boldsymbol{P}_1, \cdots, \boldsymbol{P}_K]$ and SAM masks $[\boldsymbol{S}_0, \cdots, \boldsymbol{S}_L]$ as input and returns a list of enhanced pseudo-labels $[\hat{\boldsymbol{P}}_1, \cdots, \hat{\boldsymbol{P}}_K]$. Specifically, it consists of two stages, mask assignment and mask selection as elucidated in Figure 2. During the mask assignment phase, we compute the intersection between each SAM mask $\boldsymbol{S}_l$ and pseudo-labels $\boldsymbol{P}_k$ for every class $k \in \{1, \cdots, K\}$. Each $\boldsymbol{S}_l$ is assigned to the class with the largest intersection area and masks without any overlap with existing pseudo-labels are disregarded. After processing all SAM masks, we obtain a mask assignment list $A = [A_1, \cdots, A_K]$ where A_k contains the masks assigned to class k.

It is widely known that the CAM-derived pseudo-labels often suffer from false activation [50, 21] and partial activation [3, 47]. Our mask selection strategy aims to address them by selecting the most relevant masks based on the overlaps between SAM masks and pseudo-labels. False activation arises when pseudo-labels encapsulate the target object along with a marginal section of the surrounding background. Thus, the masks for the entire target object (or its parts) should predominantly align with the pseudo-labels, whereas masks for the background should exhibit minimal overlap with the pseudo-labels. To mitigate false activation, we select masks demonstrating extensive coverage by the pseudo-labels. Conversely, partial activation arises when pseudo-labels only cover the most discriminative part instead of the entire object. Therefore, if a mask covers the majority of the pseudo-labels, it is indicative that this mask likely represents the complete object and should be retained for enhanced pseudo-labels to address partial activation.

Based on the intuitions mentioned above, we iterate through every mask assigned to class k. For each mask $\boldsymbol{S}$, we compute o_s, the fraction of mask $\boldsymbol{S}$ overlapped by pseudo-labels $\boldsymbol{P}_k$, and o_p, the fraction of pseudo-labels $\boldsymbol{P}_k$ overlapped by mask $\boldsymbol{S}$. A mask $\boldsymbol{S}$ is preserved as an enhanced pseudo-label if:

1. $o_s > t_1$ where $t_1 = 0.5$ indicates at least 50% of the mask is covered by the pseudo-labels
2. $o_p > t_2$ where $t_2 = 0.85$ indicates at least 85% of the pseudo-labels is covered by the mask

If the initial pseudo-labels are not covered by any SAM masks, we will keep them unchanged in the enhanced pseudo-labels. The overall algorithm is summarized in Algorithm 1.

Algorithm 1 SAM Enhanced Pseudo-Labels (SEPL) for One Image

Input: Pseudo labels $[\boldsymbol{P}_1, \cdots, \boldsymbol{P}_K]$, Masks $[\boldsymbol{S}_0, \cdots, \boldsymbol{S}_L]$,
threshold $t_1 = 0.5$, threshold $t_2 = 0.85$
Output Enhanced pseudo-labels $[\hat{\boldsymbol{P}}_1, \cdots, \hat{\boldsymbol{P}}_K]$

procedure SEPL($[\boldsymbol{P}_1, \cdots, \boldsymbol{P}_K]$, $[\boldsymbol{S}_0, \cdots, \boldsymbol{S}_L]$)
  $A = [A_1, \cdots, A_K]$ where $A_k = \{\}$ ▷ A_k stores the masks assigned to class k
  for l from 0 to L **do** ▷ Mask assignment
    $k^\star = \arg\max_k \text{Intersect}(\boldsymbol{S}_l, \boldsymbol{P}_k)$
    $A_{k^\star} \leftarrow A_{k^\star} \cup \{\boldsymbol{S}_l\}$
  for k from 1 to K **do** ▷ Mask selection
    if $\boldsymbol{P}_k == \{0\}^{H \times W}$ **then**
      continue
    $tmp = \{\}$ ▷ tmp stores enhanced pseudo-labels for class k
    for each mask S in A_k **do**
      $o_s = \frac{\text{Intersect}(S, \boldsymbol{P}_k)}{\text{nonzero\_area}(S)}$ ▷ fraction of mask S covered by pseudo-labels $\boldsymbol{P}_k$
      $o_p = \frac{\text{Intersect}(S, \boldsymbol{P}_k)}{\text{nonzero\_area}(\boldsymbol{P}_k)}$ ▷ fraction of pseudo-labels $\boldsymbol{P}_k$ covered by mask S
      if $o_s > t_1$ or $o_p > t_2$ **then**
        $tmp \leftarrow tmp \cup \{S\}$
    if $tmp == \{\}$ **then**
      $tmp \leftarrow tmp \cup \{\boldsymbol{P}_k\}$
    $\hat{\boldsymbol{P}}_k \leftarrow$ Merge masks in tmp with element-wise OR and assign k to nonzero elements

3 Experiment

3.1 Experimental setup

Datasets and Evaluation Metric We evaluate our proposed framework on the PASCAL VOC 2012 [15] and MS COCO 2014 dataset [35]. The dataset details can be found in Appendix B.1. We only used image-level ground-truth labels during pseudo-labels generation. The mean Intersection over Union (mIoU) is adopted as the evaluation metric for all experiments. To demonstrate the quality of the pseudo-labels, we evaluate them on the VOC and COCO training set.

Pseudo Labels In our experiments, we generate pseudo-labels using several SOTA WSSS methods, including: Recurseed [21], L2G [20], CLIPES [37], RCA [58], EPS [31], CLIMS [50], TransCAM [33], PPC [14], SIPE [11], and PuzzleCAM [22]. The detailed introductions of them can be found in Appendix B.2.

Implementation Details SAM masks are generated with the official code [25]. t_1 and t_2 by default are set to 0.5 and 0.85 respectively. We use Deeplab V2 (ResNet-101) [9] as the fully supervised semantic segmentation model. More implementation details can be found in Appendix B.3 including SAM inference hyperparameters and training details for Deeplab. The ablation study of t_1 and t_2 can be found in Appendix C.

3.2 Quantitative Evaluation and Comparison

Figures Figure 3 and Figure 4 illustrate the enhanced pseudo-label quality achieved on the PASCAL VOC and MS COCO with our SEPL algorithm. SEPL consistently and significantly elevates the

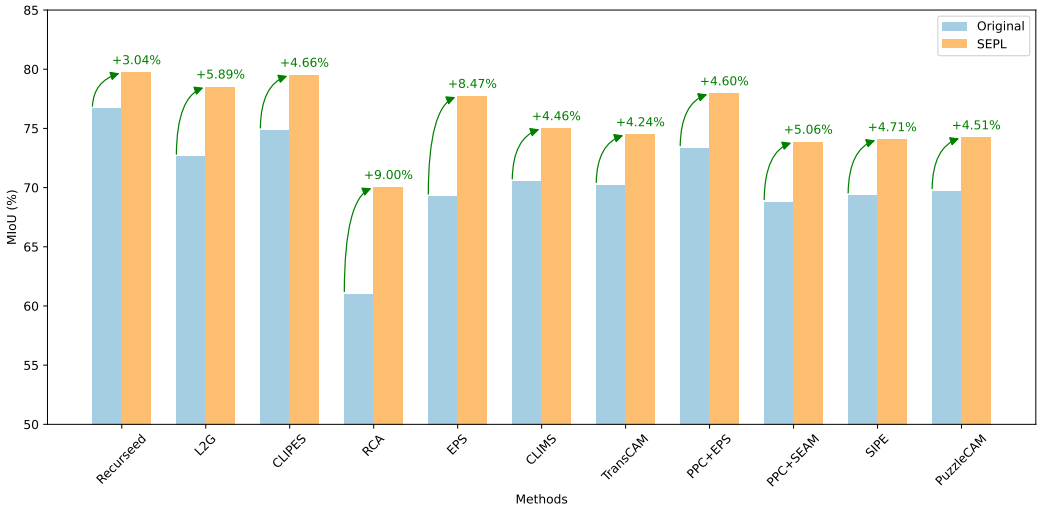


Figure 3: Pseudo labels quality on PASCAL VOC 2012. Original: pseudo-labels from original SOTA WSSS methods. SEPL: pseudo-labels enhanced by SEPL. The improvement after enhancement is indicated in green.

Method	Recurseed [21]	L2G [20]	CLIPES [37]	RCA [58]	EPS [31]	CLIMS [50]	TransCAM [33]	PPC+EPS [14]	PPC+SEAM [14]	SIPE [11]	PuzzleCAM [22]	FS [8]
Origin	71.38	69.38	70.48	69.48	68.16	69.33	68.10	70.30	65.01	67.14	65.84	76.48
SEPL	**72.89**	**72.41**	**73.06**	**69.70**	**72.13**	**71.11**	**69.93**	**71.93**	**68.25**	**69.67**	**68.89**	

Table 1: Performance of Deeplab V2 (ResNet-101) trained on pseudo-labels without post-processing: Original method vs. SAM-enhanced SEPL. Evaluated on the VOC val set. FS: full supervision

quality of pseudo-labels across various original WSSS methods. Moreover, the impact of this enhancement extends beyond the pseudo-label quality. Utilizing the enhanced pseudo-labels for training supervised semantic segmentation models(DeepLab V2) yields notable improvements in performance. The models, when trained on the enhanced pseudo-labels, consistently outperform those trained on original pseudo-labels. Table 1 and Figure 2 provide a detailed quantitative comparison of these performances across both PASCAL VOC and MS COCO datasets. More detailed results can be found in Appendix C.

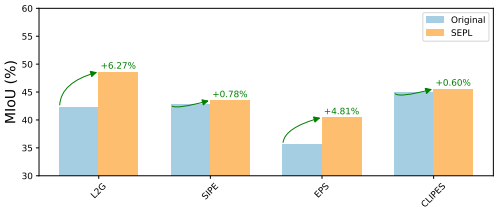


Figure 4: Pseudo-label quality on COCO $train$ set: Original from WSSS methods vs. SEPL-enhanced. Green indicates enhancements.

Method	L2G [20]	SIPE [11]	EPS [31]	CLIPES [37]	FS [8]
Origin	43.06	41.53	39.06	46.29	55.04
SEPL	**46.39**	**45.19**	**41.55**	**47.90**	

Table 2: Reult of Deeplab V2 (ResNet101) without CRF on COCO val set: Original vs. SEPL-enhanced pseudo-labels. FSS: full supervision

4 When does SAM not help?

Upon analyzing instances where SEPL was ineffective, we attributed the shortcomings primarily to three sources: the initial pseudo-labels, the SAM masks, and our enhancement algorithm.

Initial pseudo-labels Since we leverage the initial pseudo-labels as the anchors to find relevant masks, if they activate on incorrect objects or fail to activate on the target objects, the SAM masks won't offer any enhancement. In fact, they may detrimentally affect the quality of the pseudo-labels, as shown in Figure 5. To address this, we turn to better WSSS methods for more precise pseudo-labels.

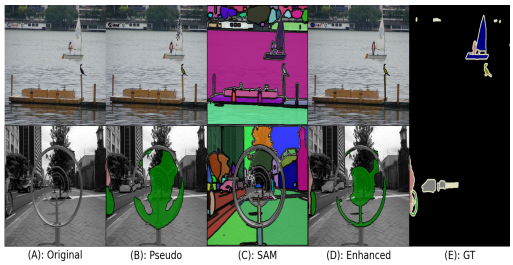


Figure 5: Examples of pseudo-labels activate on incorrect objects or fail to activate on the target objects. Adding SAM mask may detrimentally affect the quality of the pseudo-labels

Figure 6: Example of SAM's failures. The first row shows that SAM overlooks certain parts in an image and the second row shows SAM erroneously groups several objects as a single mask

Figure 7: Certain SAM masks within an image envelop smaller masks. The SAM mask in column (g) entirely covers the masks in columns (e) and (f), which poses a challenge as we cannot ascertain if a mask represents multiple objects or just one.

SAM masks While SAM effectively processes most images in VOC and COCO, it occasionally falters. As depicted in Figure 6, SAM might sometimes overlook certain parts of images or erroneously group several objects as a single segment. Fine-tuning SAM's inference hyperparameters could be a potential remedy to enhance segmentation outcomes.

Enhancement Algorithm While our current algorithm is proficient in many scenarios, it falters in specific situations. As depicted in Figure 7, certain SAM masks within an image overshadow and envelop smaller masks. An example is column (g) entirely covers the masks in columns (e) and (f). Since pseudo-labels often suffer from partial activation, our existing algorithm inclines towards selecting larger masks to produce a more complete pseudo-label. Yet, this approach poses a dilemma: the absence of clarity on whether these masks encapsulate multiple objects or just a singular entity. As illustrated in the first row of Figure 7, the mask in (g) spans both the boat and the sea. Blindly opting for the larger mask risks deteriorating the quality of pseudo-labels. A promising resolution could lie in treating the SAM masks as nodes in a tree, leveraging their inherent hierarchical structure. This tree-based approach might facilitate a more discerning selection of the appropriate masks.

5 Conclusion

This paper presents a pioneer investigation into the application of SAM as a foundation model in WSSS. By leveraging SAM's class-agnostic capability of producing fine-grained instance masks, we use CAM pseudo-labels as cues to select and combine SAM masks, resulting in high-quality pseudo-labels that are both class-aware and object-aware. Our approach is highly versatile and can be easily integrated into existing WSSS methods without any modification. Despite its simplicity, our approach shows consistent improvement over the SOTA WSSS methods on both PASCAL VOC and MS-COCO datasets. We anticipate that this study will catalyze the adoption of segmentation foundational models across a broad spectrum of computer vision tasks.

Acknowledgments

This research is supported in part by NSF (IIS-2107077, OAC2118240, and OAC-2112606) and Cisco Research. We are thankful for the computational resources of the Ohio Supercomputer Center

References

[1] Jiwoon Ahn, Sunghyun Cho, and Suha Kwak. "Weakly supervised learning of instance segmentation with inter-pixel relations". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 2209–2218.

[2] Jiwoon Ahn and Suha Kwak. "Learning Pixel-Level Semantic Affinity With Image-Level Supervision for Weakly Supervised Semantic Segmentation". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2018.

[3] Jiwoon Ahn and Suha Kwak. "Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 4981–4990.

[4] Nikita Araslanov and Stefan Roth. "Single-stage semantic segmentation from image labels". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 4253–4262.

[5] Saeid Asgari Taghanaki, Kumar Abhishek, Joseph Paul Cohen, Julien Cohen-Adad, and Ghassan Hamarneh. "Deep semantic segmentation of natural and medical images: a review". In: *Artificial Intelligence Review* 54 (2021), pp. 137–178.

[6] Amy Bearman, Olga Russakovsky, Vittorio Ferrari, and Li Fei-Fei. "What's the point: Semantic segmentation with point supervision". In: 2016, pp. 549–565.

[7] Arslan Chaudhry, Puneet K. Dokania, and Philip H. S. Torr. *Discovering Class-Specific Pixels for Weakly-Supervised Semantic Segmentation*. 2017. arXiv: 1707.05821 [cs.CV].

[8] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs". In: *TPAMI* (2017).

[9] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs". In: *IEEE transactions on pattern analysis and machine intelligence* 40.4 (2017), pp. 834–848.

[10] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. "Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs". In: *ICLR*. 2015.

[11] Qi Chen, Lingxiao Yang, Jian-Huang Lai, and Xiaohua Xie. "Self-Supervised Image-Specific Prototype Exploration for Weakly Supervised Semantic Segmentation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2022, pp. 4288–4298.

[12] Junsuk Choe, Seungho Lee, and Hyunjung Shim. "Attention-based dropout layer for weakly supervised single object localization and semantic segmentation". In: *IEEE transactions on pattern analysis and machine intelligence* 43.12 (2020), pp. 4256–4271.

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. "An image is worth 16x16 words: Transformers for image recognition at scale". In: *arXiv preprint arXiv:2010.11929* (2020).

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

[15] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. "The pascal visual object classes (voc) challenge". In: 88.2 (2010), pp. 303–338.

[16] Junsong Fan, Zhaoxiang Zhang, Tieniu Tan, Chunfeng Song, and Jun Xiao. "Cian: Cross-image affinity net for weakly supervised semantic segmentation". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 07. 2020, pp. 10762–10769.

[17] Di Feng, Christian Haase-Schütz, Lars Rosenbaum, Heinz Hertlein, Claudius Glaeser, Fabian Timm, Werner Wiesbeck, and Klaus Dietmayer. "Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges". In: *IEEE Transactions on Intelligent Transportation Systems* 22.3 (2020), pp. 1341–1360.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. *Deep Residual Learning for Image Recognition*. 2015. arXiv: 1512.03385 [cs.CV].

[19] Peng-Tao Jiang, Qibin Hou, Yang Cao, Ming-Ming Cheng, Yunchao Wei, and Hong-Kai Xiong. "Integral object mining via online attention accumulation". In: 2019, pp. 2070–2079.

[20] Peng-Tao Jiang, Yuqi Yang, Qibin Hou, and Yunchao Wei. "L2g: A simple local-to-global knowledge transfer framework for weakly supervised semantic segmentation". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 16886–16896.

[21] Sang Hyun Jo, In Jae Yu, and Kyung-Su Kim. "RecurSeed and CertainMix for weakly supervised semantic segmentation". In: *arXiv preprint arXiv:2204.06754* (2022).

[22] Sanghyun Jo and In-Jae Yu. "Puzzle-cam: Improved localization via matching partial and full features". In: *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE. 2021, pp. 639–643.

[23] Tsung-Wei Ke, Jyh-Jing Hwang, and Stella X Yu. "Universal weakly supervised segmentation by pixel-to-segment contrastive learning". In: *arXiv preprint arXiv:2105.00957* (2021).

[24] Anna Khoreva, Rodrigo Benenson, Jan Hosang, Matthias Hein, and Bernt Schiele. "Simple does it: Weakly supervised instance and semantic segmentation". In: Computer Vision Foundation / IEEE, 2017, pp. 876–885.

[25] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. "Segment anything". In: *arXiv preprint arXiv:2304.02643* (2023).

[26] Alexander Kolesnikov and Christoph H Lampert. "Seed, expand and constrain: Three principles for weakly-supervised image segmentation". In: *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*. Springer. 2016, pp. 695–711.

[27] Fahad Lateef and Yassine Ruichek. "Survey on semantic segmentation using deep learning techniques". In: *Neurocomputing* 338 (2019), pp. 321–348.

[28] Jungbeom Lee, Eunji Kim, and Sungroh Yoon. "Anti-Adversarially Manipulated Attributions for Weakly and Semi-Supervised Semantic Segmentation". In: Computer Vision Foundation / IEEE, 2021, pp. 4071–4080.

[29] Jungbeom Lee, Jihun Yi, Chaehun Shin, and Sungroh Yoon. *BBAM: Bounding Box Attribution Map for Weakly Supervised Semantic and Instance Segmentation*. 2021. arXiv: 2103.08907 [cs.CV].

[30] Seungho Lee, Minhyun Lee, Jongwuk Lee, and Hyunjung Shim. "Railroad is not a Train: Saliency as Pseudo-pixel Supervision for Weakly Supervised Semantic Segmentation". In: *CVPR*. 2021.

[31] Seungho Lee, Minhyun Lee, Jongwuk Lee, and Hyunjung Shim. "Railroad is not a train: Saliency as pseudo-pixel supervision for weakly supervised semantic segmentation". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 5495–5505.

[32] Kunpeng Li, Ziyan Wu, Kuan-Chuan Peng, Jan Ernst, and Yun Fu. "Tell me where to look: Guided attention inference network". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 9215–9223.

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

[34] Di Lin, Jifeng Dai, Jiaya Jia, Kaiming He, and Jian Sun. "Scribblesup: Scribble-supervised convolutional networks for semantic segmentation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 3159–3167.

[35] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. "Microsoft coco: Common objects in context". In: *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer. 2014, pp. 740–755.

[36] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. "Microsoft Coco: Common Objects in Context". In: *Computer Vision – ECCV 2014* (2014), pp. 740–755. DOI: 10.1007/978-3-319-10602-1_48.

[37] Yuqi Lin, Minghao Chen, Wenxiao Wang, Boxi Wu, Ke Li, Binbin Lin, Haifeng Liu, and Xiaofei He. "CLIP Is Also an Efficient Segmenter: A Text-Driven Approach for Weakly Supervised Semantic Segmentation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2023, pp. 15305–15314.

[38] Shervin Minaee, Yuri Boykov, Fatih Porikli, Antonio Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. "Image Segmentation Using Deep Learning: A Survey". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44.7 (2022), pp. 3523–3542. DOI: 10.1109/TPAMI.2021.3059968.

[39] Youngmin Oh, Beomjun Kim, and Bumsub Ham. "Background-Aware Pooling and Noise-Aware Loss for Weakly-Supervised Semantic Segmentation". In: *CVPR*. 2021.

[40] Lixiang Ru, Yibing Zhan, Baosheng Yu, and Bo Du. "Learning affinity from attention: end-to-end weakly-supervised semantic segmentation with transformers". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 16846–16855.

[41] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. *ImageNet Large Scale Visual Recognition Challenge*. 2015. arXiv: 1409.0575 [cs.CV].

[42] Karen Simonyan and Andrew Zisserman. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. 2015. arXiv: 1409.1556 [cs.CV].

[43] Krishna Kumar Singh and Yong Jae Lee. "Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization". In: *2017 IEEE international conference on computer vision (ICCV)*. IEEE. 2017, pp. 3544–3553.

[44] Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang. "Box-driven class-wise region masking and filling rate guided loss for weakly supervised semantic segmentation". In: Computer Vision Foundation / IEEE, 2019, pp. 3136–3145.

[45] Weixuan Sun, Jing Zhang, and Nick Barnes. "3d guided weakly supervised semantic segmentation". In: *Proceedings of the Asian Conference on Computer Vision*. 2020.

[46] Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, and Tiejun Huang. "Seggpt: Segmenting everything in context". In: *arXiv preprint arXiv:2304.03284* (2023).

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

[48] Yunchao Wei, Jiashi Feng, Xiaodan Liang, Ming-Ming Cheng, Yao Zhao, and Shuicheng Yan. "Object region mining with adversarial erasing: A simple classification to semantic segmentation approach". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 1568–1576.

[49] Yunchao Wei, Jiashi Feng, Xiaodan Liang, Ming-Ming Cheng, Yao Zhao, and Shuicheng Yan. "Object region mining with adversarial erasing: A simple classification to semantic segmentation approach". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 1568–1576.

[50] Jinheng Xie, Xianxu Hou, Kai Ye, and Linlin Shen. "CLIMS: cross language image matching for weakly supervised semantic segmentation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 4483–4492.

[51] Lian Xu, Wanli Ouyang, Mohammed Bennamoun, Farid Boussaid, and Dan Xu. "Multi-class token transformer for weakly supervised semantic segmentation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 4310–4319.

[52] Qi Yao and Xiaojin Gong. "Saliency guided self-attention network for weakly and semi-supervised semantic segmentation". In: *IEEE Access* (2020).

[53] Xiaohui Yuan, Jianfang Shi, and Lichuan Gu. "A review of deep learning methods for semantic segmentation of remote sensing imagery". In: *Expert Systems with Applications* 169 (2021), p. 114417.

[54] Bingfeng Zhang, Jimin Xiao, Yunchao Wei, Mingjie Sun, and Kaizhu Huang. "Reliability does matter: An end-to-end weakly supervised semantic segmentation approach". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 07. 2020, pp. 12765–12772.

[55] Tianyi Zhang, Guosheng Lin, Weide Liu, Jianfei Cai, and Alex Kot. "Splitting vs. merging: Mining object regions with discrepancy and intersection loss for weakly supervised semantic segmentation". In: 2020.

[56] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. "Learning Deep Features for Discriminative Localization". In: *Computer Vision and Pattern Recognition*. 2016.

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

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

[59] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Gao, and Yong Jae Lee. "Segment everything everywhere all at once". In: *arXiv preprint arXiv:2304.06718* (2023).

Appendix

A Related Work

A.1 Weakly Supervised Semantic Segmentation (WSSS)

Recent approaches in weakly supervised semantic segmentation (WSSS) often rely on Class Activation Maps (CAM) [57] to generate pixel-level pseudo-labels. These pseudo-labels are then used to train the segmentation model in a fully supervised manner. However, CAM often exhibits a bias towards the most discriminative regions of the target object which limits the quality of the pseudo-labels. To overcome this challenge, recent works mainly focus on generating high-quality CAMs with integral activation on the entire object regions.

Early-stage works [43, 49, 32] encourage the network to discover less activated object parts via adversarial erasing. In addition to the classification loss typically used in the WSSS framework, specific loss functions such as SEC loss [26], equivariance regularization [47], and contrastive loss [23, 14] have been exploited in narrowing the gap between the pixel-level and image-level supervisions. Some works also introduce network modules to address the partial activation problem of CAM: SEAM [47] leverages pixel-level semantic affinities with a pixel correlation module; CIAN [16] exploits the additional information from related images with a cross-image affinity module. Recent methods based on Vision Transformer [13] including [33, 51] aim to uncover more comprehensive object regions by exploring the global information from the attention of the transformer network. Most of these works follow the multi-stage framework, where a post-processing step is necessary for refining and improving the initial pseudo-labels generated from CAM.

A.2 Post-Processing in WSSS

Although there are end-to-end WSSS solutions [40, 4, 54] available, most of the recent works still rely on some post-processing techniques to enhance the initial pseudo-labels to achieve superior performance. Among these techniques, two widely utilized methods for refining pseudo-labels are AffinityNet [3] and IRNet [1]. AffinityNet trains a network from CAM to predict the semantic affinities and uses it for propagating local activations, whereas IRNet [1] learns and predicts semantic affinities more effectively by leveraging class boundary maps. Despite the substantial improvement in the pseudo-labels, these methods require training a separate network. The computational cost involved can be a significant barrier, especially when working with large-scale datasets. Additionally, the careful tuning of hyperparameters to obtain accurate foreground and background pixels for training these networks can slow down the entire training pipeline. This requirement for meticulous parameter tuning not only adds complexity to the process but also limits the applicability and scalability of these post-processing techniques.

B Experiment Details

B.1 Dataset Details

Experiments are conducted on two publicly available datasets, PASCAL VOC 2012 [15]and MS COCO 2014 [36]. The PASCAL VOC 2012 dataset contains 20 semantic categories and the background. It is split into three sets, the training, validation, and test sets, each containing 1464, 1449, and 1456 images, respectively. Following the standard setting, we also use the augmented training set, yielding a total of 10582 training images. The MS COCO 2014 dataset has 80 semantic categories. Following [12], the images without target categories are excluded from the dataset, remaining 82081 training images and 40137 validation images. We report the mean Intersection-over-Union (mIoU), precision, and recall for evaluation. To demonstrate the quality116 of the pseudo-labels, we evaluate them on the VOC and COCO training set.

B.2 Baselines

- **CLIMS:** [50] A Cross-Language Image Matching framework leveraging natural language supervision to activate complete object regions and suppress related open background regions for improved CAM quality in WSSS.

- **SIPE:** [11] Self-supervised Image-specific Prototqualitativeype Exploration, which tailors prototypes for each image to capture complete regions, optimizing feature representation and enabling self-correction for improved WSSS performance.
- **PPC:** [14] A weakly-supervised pixel-to-prototype contrast method providing pixel-level supervisory signals, executed across and within different views of an image to enhance the quality of pseudo masks for WSSS.
- **TransCAM:** [33] A Conformer-based solution that refines CAM by leveraging attention weights from the transformer branch of the Conformer, capturing both local features and global representations for WSSS.
- **RecurSeed:** [21] An approach that alternately reduces non-detections and false-detections through recursive iterations, implicitly finding an optimal junction and leveraging a novel data augmentation method, EdgePredictMix, for improved WSSS performance.
- **L2G** [20] A simple online local-to-global knowledge transfer framework for high quality object attention mining. It first leverages a local classification network to extract attentions from multiple local patches randomly cropped from the input image. Then, it utilizes a global network to learn complementary attention knowledge across multiple local attention maps online.
- **CLIPES** [37] An approach that leverages CLIP to improve pseudo-label generation, refinement and final segmentation model training.
- **RCA** [58] RCA is equipped with a regional memory bank to store massive, diverse object patterns appearing in training data, which acts as strong support for exploration of dataset-level semantic structure.
- **EPS** [31] EPS learns from pixel-level feedback by combining two weak supervisions; the image-level label provides the object identity via the localization map and the saliency map from the off-the-shelf saliency detection model offers rich boundaries
- **PuzzleCAM** [22] PuzzleCAM minimizes differences between the features from separate patches and the whole image. It consists of a puzzle module and two regularization terms to discover the most integrated region in an object.

B.3 Implementation Details

B.3.1 SAM Inference Hyperparameters

For our experiments, we adopted the standard settings of the SAM model as provided in their official repository. However, we made modifications to two specific hyperparameters to tailor the model's behavior to our needs:

- pred-iou-thresh was set from None to 0.86.
- stability-score-thresh was set from None to 0.92.

By adjusting these thresholds, our objective was to enable SAM to produce a wider and more diverse range of masks for selection via our algorithm. Importantly, these modifications did not have a detrimental effect on the inference speed, ensuring efficiency was maintained throughout the process. For inference, we employed the default pretrained ViT-H SAM model.

B.3.2 DeepLab Model Training

PASCAL VOC 2012 Dataset For the PASCAL VOC 2012 dataset, our training process of DeepLabV2 adheres to the guidelines provided in the GitHub repository `https://github.com/kazuto1011/deeplab-pytorch`. We utilize the training hyperparameters as configured in the `voc12.yaml` file from this repository. Key hyperparameters are specified as follows:

- **Batch Size:**
  - `TRAIN`: 5
- **Iterations:**
  - `ITER_MAX`: 20000

  - ITER_SIZE: 2
- **Learning Rate (LR)**: 2.5×10^{-4}
- **Momentum**: 0.9
- **Number of Blocks (N_BLOCKS)**: [3, 4, 23, 3]
- **Atrous Rates**: [6, 12, 18, 24]

In our experiments, we did not further refine the results of the DeepLab model with Conditional Random Fields (CRF) or any other post-processing techniques.

COCO2014 Dataset For the COCO2014 dataset, we followed a similar training process to that of the PASCAL VOC 2012 dataset. However, the configuration settings were adopted from a different GitHub repository, available at `https://github.com/PengtaoJiang/L2G`. The specific hyperparameters used in our training are as follows:

- **Batch Size:**
  - TRAIN: 20
- **Iterations:**
  - ITER_MAX: 50000
  - ITER_SIZE: 1
- **Learning Rate (LR)**: 2.5×10^{-4}
- **Momentum**: 0.9
- **Number of Blocks (N_BLOCKS)**: [3, 4, 23, 3]
- **Atrous Rates**: [6, 12, 18, 24]

Similar to our approach with the PASCAL VOC 2012 dataset, we did not apply CRF or other post-processing techniques to the results obtained from the DeepLab model.

C Extra Experiment Results

C.1 More Results for SEPL

Figure 8 illustrates the qualitative improvements of the pseudo-labels enhanced by SEPL in recall, precision, and mIoU, respectively, which are calculated based on the average of all samples. Table 4 shows the quantitative improvement of the pseudo-labels enhanced by SEPL for recall and precision.

C.2 Ablation Study of t_1 and t_2

As mentioned in subsection 3.1, t_1 and t_2 by default are set to 0.5 and 0.85 respectively. To better understand the impact of these hyperparameters, we conducted an ablation study on them. Table 3 shows the pseudo labels quality (mIoU) on PASCAL VOC 2012 for SEPL using Recurseed [21] as the base method with different values of t_1 and t_2. SEPL's performance is robust to t_1 and t_2.

C.3 Apply SAM on CAM without post-processing

As mentioned in Section subsection A.2, most recent works still rely on some post-processing techniques to enhance the initial CAM, aiming to procure more precise pseudo-labels. However, these enhancement procedures often demand substantial computational overhead and extended training durations. Such constraints can potentially hinder the broad-scale deployment of WSSS on extensive datasets. Since the initial CAM also provides an estimation of object localization, our proposed SEPL can be directly applied to the initial CAM. This approach circumvents the need for additional post-processing steps, leading to appreciable reductions in both training duration and computational demands in the WSSS pipeline.

As shown in Figure 9, the initial CAM consistently benefits from SAM masks. More interestingly, CAM+SAM can reach and even surpass the quality of pseudo-labels obtained after conventional

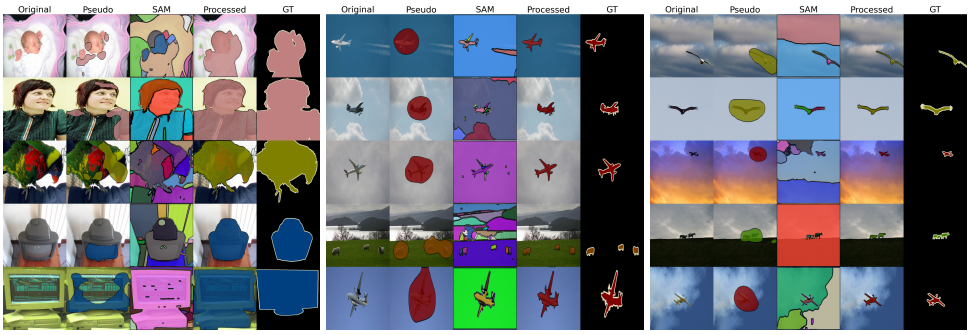
Table 3: Ablation study of t_1 and t_2. Pseudo labels quality (mIoU) on PASCAL VOC 2012 for SEPL using Recurseed [21] as the base method with different values of t_1 and t_2. SEPL's performance is robust to t_1 and t_2.

t1 \ t2	0.2	0.3	0.4	0.5	0.6	0.7	0.8
0.6	79.87	80.88	81.20	81.46	81.39	81.40	79.50
0.65	79.84	80.85	81.10	81.37	81.26	81.25	79.15
0.7	79.84	80.88	81.10	81.28	81.16	81.06	78.64
0.75	79.84	80.84	81.04	81.23	81.10	81.00	78.33
0.8	79.85	80.83	81.04	81.22	81.03	80.87	77.78
0.85	79.83	80.79	81.00	81.12	80.87	80.69	77.13
0.9	79.86	80.82	80.99	81.01	80.67	80.42	76.03

Method	Recurseed	L2G	CLIPES	RCA	EPS	CLIMS	TransCAM	PPC+EPS	PPC+SEAM	SIPE	PuzzleCAM
Pseudo Label Precision	85.10	78.86	83.81	81.53	78.10	80.97	80.46	82.73	78.93	76.75	83.29
SEPL Precision	**85.73**	**84.57**	**85.96**	**82.13**	**84.85**	**83.29**	**81.89**	**84.32**	**81.76**	**79.86**	**83.93**
Precision Delta	0.63	5.71	2.15	0.60	6.75	2.32	1.43	1.59	2.83	3.11	6.39
Pseudo Label Recall	85.85	87.35	85.36	71.04	84.93	83.19	84.83	86.91	83.45	87.68	76.77
SEPL Recall	**91.75**	**92.05**	**91.58**	**79.37**	**90.94**	**89.12**	**90.07**	**92.02**	**90.36**	**91.12**	**86.22**
Recall Delta	5.90	4.70	6.22	8.33	6.01	5.93	5.24	5.11	6.91	3.44	9.45

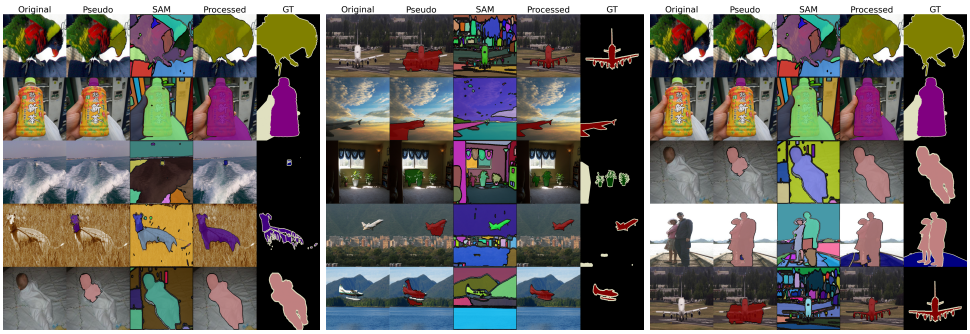
Table 4: Pseudo label precision and recall improvements by incorporating SEPL

post-processing. This finding suggests that SAM has the potential to replace time-consuming post-processing steps, offering a more efficient solution to WSSS tasks. Figure 10 illustrates the qualitative improvements of the initial CAM by SEPL in recall, precision, and mIoU, respectively.



(a) Recall (b) Precision (c) mIoU

Figure 10: Improvements for CAM without post-processing: (a) Recall, (b) Precision, (c) mIoU.



(a) Recall (b) Precision (c) mIoU

Figure 8: Improvements for pseudo-labels: (a) Recall, (b) Precision, (c) mIoU.

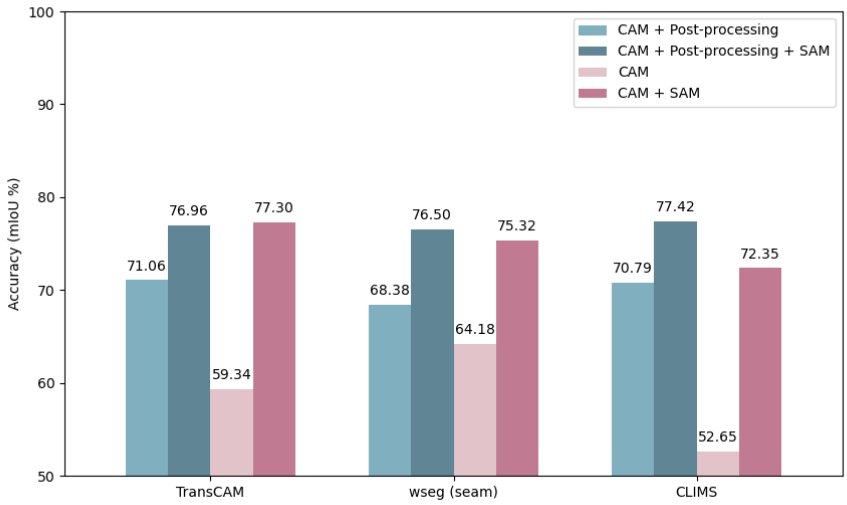


Figure 9: By directly utilizing iinitial CAM with SAM, we achieved comparable performance to that of post-processed pseudo-labels enhanced by SAM. This finding suggests that SAM can be used as a substitute for post-processing modules, resulting in a marked acceleration of the entire WSSS training pipeline