DHR: Dual Features-Driven Hierarchical Rebalancing in Inter- and Intra-Class Regions for Weakly-Supervised Semantic Segmentation

Sanghyun Jo[1], Fei Pan[2], In-Jae Yu[3], and Kyungsu Kim*[4]

[1] OGQ, Seoul, Korea
[2] University of Michigan, Ann Arbor, MI, USA
[3] Samsung Electronics, Suwon, Korea
[4] Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA
{shjo.april, ijyu.phd, kskim.doc}@gmail.com, feipan@umich.edu

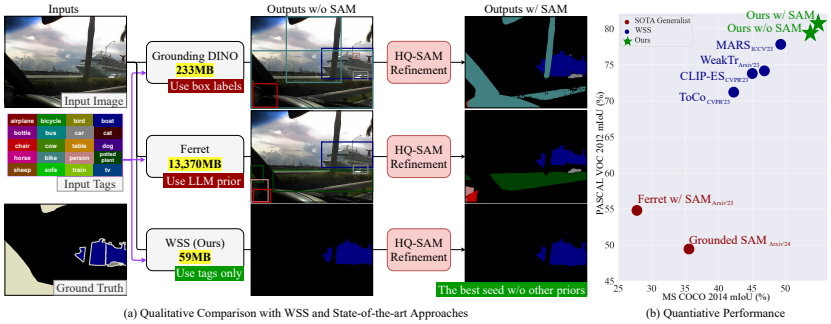
Abstract. Weakly-supervised semantic segmentation (WSS) ensures high-quality segmentation with limited data and excels when employed as input seed masks for large-scale vision models such as Segment Anything. However, WSS faces challenges related to minor classes since those are overlooked in images with adjacent multiple classes, a limitation originating from the overfitting of traditional expansion methods like Random Walk. We first address this by employing unsupervised and weakly-supervised feature maps instead of conventional methodologies, allowing for hierarchical mask enhancement. This method distinctly categorizes higher-level classes and subsequently separates their associated lower-level classes, ensuring all classes are correctly restored in the mask without losing minor ones. Our approach, validated through extensive experimentation, significantly improves WSS across five benchmarks (VOC: 79.8%, COCO: 53.9%, Context: 49.0%, ADE: 32.9%, Stuff: 37.4%), reducing the gap with fully supervised methods by over 84% on the VOC validation set. Code is available at https://github.com/shjo-april/DHR.

Keywords: semantic segmentation · weakly supervised learning

1 Introduction

Semantic segmentation, the process of grouping each pixel in an image into semantic classes, heavily relies on pixel-wise annotations crafted by humans. This labor-intensive requirement has been a significant bottleneck in scaling segmentation models. Recently, large-scale visual models (LSVMs), such as Grounded SAM [48], have emerged as compelling alternatives in segmentation tasks, employing advanced box, point, or scribble supervision with image-text pairs. By contrast, weakly-supervised semantic segmentation (WSS) utilizes image-level class labels with images to produce segmentation outcomes. As shown in Figure 1, recent WSS models [27, 51] outperform existing LSVMs by at least 7% accuracy on segmentation datasets, such as VOC 2012 [17] and COCO 2014 [39].

* Correspondence to



(a) Qualitative Comparison with WSS and State-of-the-art Approaches (b) Quantiative Performance

Fig. 1: Importance of WSS. (a): Our WSS approach (DHR) outperforms large-scale vision models [48, 61] with only image-level supervision and 25% of the parameters, bypassing the need for extensive human annotations (image, text, and box pairs). **(b):** Our DHR significantly exceeds Grounded SAM [48], Ferret [61], and recent WSS models [27, 40, 51, 66] in standard benchmark performances [17, 39].

WSS research efforts [25, 36, 55] primarily focus on generating pseudo masks from image-level tags to mimic pixel-level annotations through the propagation of initial class activation maps (CAMs). Our analysis of discrepancies between WSS predictions and ground-truth masks, as shown in Figure 2, uncovers a significant challenge: adjacent pixels from distinct classes (*e.g.*, person and motorbike) often merge, leading to the disappearance of spatially minor classes. This problem, a result of overlooking class ratios during the propagation process, is especially pronounced in areas of inter-class regions. Notably, on the VOC dataset [17], adjacent regions constitute 35% of the total area, with 79% being inter-class regions, while the COCO dataset [39] has 75% adjacent regions, 55% of which are inter-class. This analysis underscores the importance of addressing the vanishing problem in neighboring classes to enhance WSS performance.

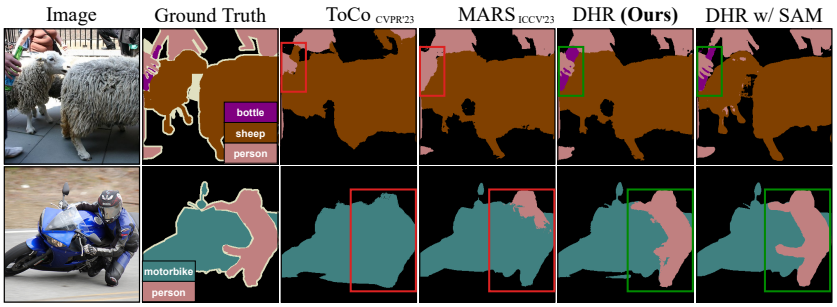


Fig. 2: Vanishing problem of adjacent minor classes in WSS outputs. Red boxes illustrate the false prediction of minor classes in pseudo labels generated from WSS, *e.g.*, bottle, person, and backpack. Green boxes highlight our DHR outperforming state-of-the-art baselines [27, 51] in adjacent class regions.

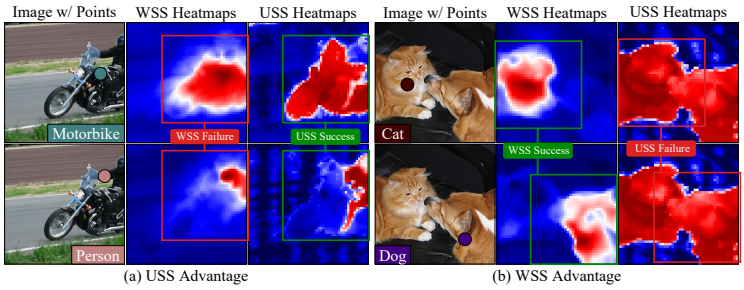


Fig. 3: Visualization of heatmaps with target points. (a): USS features can precisely separate between inter-class regions (*e.g.*, animals vs. vehicles) unlike WSS. **(b):** Thanks to training image-level class labels, WSS features can discern specific classes (*e.g.*, dog vs. cat) in the same inter-class region (*e.g.*, animal).

We observe that unsupervised semantic segmentation (USS) features, as highlighted in studies [19, 29], are adept at distinguishing between inter-class regions (*e.g.*, animal and furniture). Figure 3 showcases pixel-wise cosine similarity maps (*i.e.*, heatmaps) comparing WSS and USS features with class points. Specifically, the inherent difference between USS and WSS features originates from their distinct learning objectives. In Figure 3(a), USS methods learn visual similarities (*e.g.*, color and shape) between images without tags, enabling them to distinguish inter-class regions with dissimilar appearances (*e.g.*, person vs. motorbike). Conversely, WSS methods learn class-specific differences with human-annotated tags, allowing them to discern visually similar classes (*e.g.*, dog vs. cat) within the same inter-class region (*e.g.*, animals), as shown in Figure 3(b).

To harness the strengths of both USS and WSS features, we introduce a pioneering seed propagation method termed **D**ual Features-Driven **H**ierarchical **R**ebalancing (DHR). DHR encompasses three pivotal steps: 1) the seed recovery for disappeared classes in WSS masks, 2) the utilization of USS features for inter-class segregation, and 3) the fine-grained separation between intra-class regions using WSS features. In particular, we automatically group inter- and intra-class regions based on the USS feature correlation matrix to apply subsequent WSS rebalancing in USS-based rebalancing outputs, as visualized in Figure 4(a). Consequently, our novel propagation using dual features achieves class separation across all adjacent classes, as conceptually illustrated in Figure 4(b). Our key contributions are summarized as follows:

- We identify and tackle the issue of minor classes disappearing in WSS, a problem that previously needed to be addressed by existing techniques.
- We introduce DHR, a novel method that 1) enhances class distinction through seed initialization based on optimal transport and 2) leverages USS and WSS features to separate inter- and intra-class regions.
- DHR achieves a state-of-the-art mIoU of 79.8% on the PASCAL VOC 2012 test set, significantly closing the gap with FSS (84%) and showing versatility across multiple USS and WSS models.

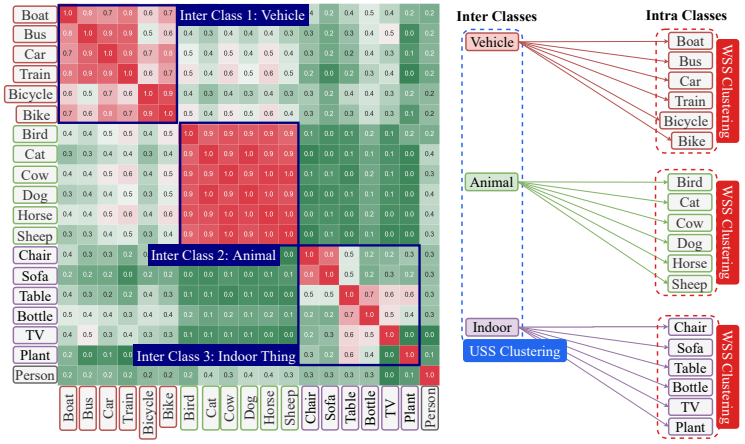


Fig. 4: Conceptual illustration of our hierarchical clustering. *Left.* USS feature correlation automatically groups inter classes. *Right.* Using USS features, we categorize all inter classes (*e.g.*, vehicle and animal) and then separate these intra classes (*e.g.*, car and bus) per each inter class (*e.g.*, vehicle) with WSS features.

– Our method also proves effective as a seeding technique for SAM, outperforming leading approaches like Grounding DINO [42] and establishing the potential of WSS in advanced segmentation tasks.
– We first evaluate our DHR and recent WSS methods [27, 51] across three additional benchmarks [4, 43, 63]. This showcases their adaptability and sets new standards for future explorations in the field.

2 Related Work

2.1 Weakly-supervised Semantic Segmentation

Weakly-supervised semantic segmentation (WSS) aims to minimize the quality gap between pixel-wise annotations and pseudo masks generated with image-level class labels. Unlike previous WSS studies [9, 11, 15, 18, 27, 30, 33, 40, 51, 54, 60,66], our research pioneers in addressing the problem of vanishing minor classes by rebalancing class information in pseudo masks in Table 1.

Table 1: Conceptual comparison of our method with recent approaches.

Properties	ACR	ToCo	WeakTr	CLIP-ES	QA-CLIMS	FMA	MARS	**Ours**
Solving Vanishing Issue of Adjacent Minor Classes	✗	✗	✗	✗	✗	✗	✗	✓
Utilizing WSS features to improve the quality of WSS masks	✓	✓	✓	✓	✓	✓	✗	✓
Utilizing USS features to enhance the quality of WSS masks	✗	✗	✗	✗	✗	✗	✓	✓
No external datasets and models	✓	✓	✓	✗	✗	✗	✓	✓
Following model-agnostic manner	✗	✗	✗	✗	✗	✗	✓	✓

Evolution of WSS Pipelines Traditionally, WSS pipelines are composed of three stages: 1) CAM generation for initial seeds, 2) applying propagation methods like Random Walk (RW) [1, 2] to generate pseudo masks from CAMs, and 3) training the final segmentation network (*e.g.*, DeepLabv2 [7]). Most studies have focused on improving CAM performance using WSS feature correlation [10, 16, 26, 55], patch-based principles [24, 25, 35, 62], and enhancing cross-image information [53, 56, 57, 65]. Recent studies [50, 58] employ self-attention mechanisms from transformer architectures rather than the conventional RW for propagating initial CAMs. However, the absence of unsupervised features in these models leads to the disappearance of minor classes during propagation.

Exploiting Unsupervised Features A few methods [27, 30] utilize unsupervised features, primarily DINO [5], for improving WSS performance. MARS [27] leverages advanced USS features from STEGO [19] for binary separation of foreground and background classes to remove biased objects in pseudo labels. However, relying solely on unsupervised features without WSS features fails to distinguish between intra-class regions. Our approach is the first solution to mitigate the vanishing problem of adjacent minor classes in WSS outputs by combining WSS and USS features hierarchically, using each of the strengths to separate intra- and inter-class regions simultaneously for the first time.

Integration with External Models Current state-of-the-art approaches [9, 11, 15, 40, 54, 60] rely on CLIP [47], large-language models [61], Grounding DINO [42], or SAM [31] to fine-tune WSS models. However, the dependency on pre-trained knowledge limits applicability to novel tasks, such as in medical fields, constraining WSS's scalability. Nevertheless, our approach using both unsupervised and weakly-supervised features outperforms recent WSS methods depending on advanced supervision and datasets, demonstrating the potential of integrating WSS with USS in overcoming limitations presented by external models.

2.2 Unsupervised Semantic Segmentation

Unsupervised semantic segmentation (USS) is dedicated to developing semantically rich features across a collection of images without relying on any annotations. Depending on whether self-supervised vision transformers (*e.g.*, [5, 44]) are used, they can be classified into two types. First, without initializing pre-trained vision transformers, conventional approaches [14, 22, 45] enhance the mutual information across differing perspectives of the same image. Meanwhile, Leopart [67], STEGO [19], HP [52], and CAUSE [29] employ self-supervised vision transformers for initializing spatially coherent representations of images and train a simple feed-forward network to enhance pixel-level representation. Our technique is designed to be compatible with any USS approach by leveraging pixel-level embedding vectors. Thus, our method operates independently from existing USS frameworks. Based on recent USS models [19, 29], we demonstrate our method's flexibility to consistently improve the performance of existing WSS models in Section 4.

2.3 Open-vocabulary Detection and Segmentation

Contrastive Language-Image Pre-training (CLIP) [47], trained on 400 million image-text pairs, set a foundation for generating segmentation from free-form text prompts. Despite advancements with models like MaskCLIP [64] and TCL [6], they fall short against specialized segmentation models, such as DeepLabv3+ [64] and Mask2Former [12], which utilize precise mask annotations for training. Recently, Grounding DINO [42] and Ferret [61] have pushed the envelope by integrating box annotations and leveraging language models to enhance detection performance. The introduction of SAM [28, 31], a zero-shot segmentation model using conditional prompts like points, marks a significant stride in employing open-vocabulary tasks, with Grounded SAM [48] leading the charge by merging two capabilities of Grounding DINO [42] and SAM [31]. Still, these models face challenges in generating accurate seeds from multi-tag inputs due to their reliance on limited tag ranges in training captions. Our analysis reveals that advanced WSS models, including our DHR approach, outperform these open-vocabulary models in multi-class predictions (see Figure 1 and Table 2). Incorporating DHR with SAM for the final refinement also boosts segmentation accuracy, showcasing WSS's indispensable role in navigating multi-tag segmentation scenarios.

3 Method

In this section, we introduce our method, DHR (**D**ual Features-Driven **H**ierarchical **R**ebalancing), which aims to address the challenges of vanishing classes in existing WSS methods. We outline an overview of DHR in Figure 5 for a comprehensive understanding of our framework. Section 3.1 delves into the background of conventional seed propagation techniques and our setup, highlighting the limitations that our method seeks to overcome. The core of our contribution (DHR) is presented in Section 3.2, where we detail how we tackle the disappearance of inter- and inter-class regions through a novel rebalancing strategy. We round off with Section 3.3, discussing DHR's training objectives.

3.1 Background: Conventional Seed Propagation

To train segmentation models, previous WSS studies [2, 26] generate pseudo masks by expanding initial seeds obtained from class activation maps (CAMs). These CAMs are extracted from weakly-supervised feature maps $F^{ws} = E_{\theta^{ws}}(I)$, produced by an image encoder within the WSS network, using the target image I as input. Mathematically, the process of propagating CAMs is described as

$$M^{base} = \mathcal{R}_C(\mathcal{P}^{base}(\mathcal{A}_{F^{ws}})), \tag{1}$$

where $\mathcal{A}_{F^{ws}}$ represents the CAMs, $\mathcal{P}^{base}(\cdot)$ indicates conventional seed propagation techniques [1, 21, 34] such as Random Walk [2]. The process further

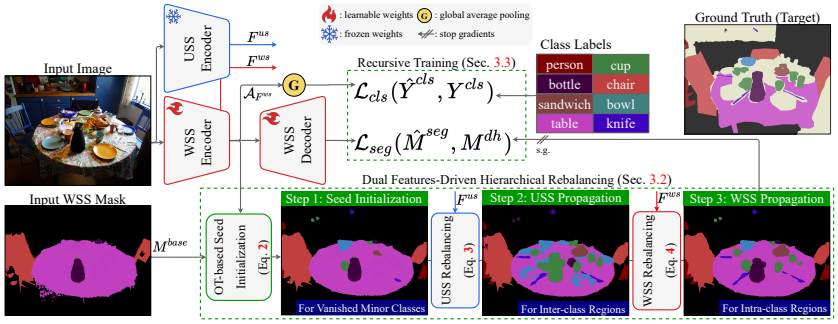


Fig. 5: Overview of DHR. Our framework unfolds in three steps, recovering vanished classes by replacing pixels of input WSS mask with OT-based CAMs for seed initialization. We then employ a hierarchical approach to propagate restored seeds, using unsupervised feature maps for the inter-class segregation (*e.g.*, kitchenware) and weakly-supervised features for the intra-class differentiation (*e.g.*, bottle and cup). Finally, our balanced masks are used to train the segmentation model recursively.

incorporates $\mathcal{R}_C(\cdot)$, boundary correction tools such as CRF [32] and PAMR [3], to produce a segmentation output with C class channels, resulting in the final WSS mask $M^{base} \in \mathbb{R}^{C\times H\times W}$. Although this propagation approach covers a sufficient foreground region compared to CAMs, it leads to the disappearance of adjacent minor classes in Figure 2.

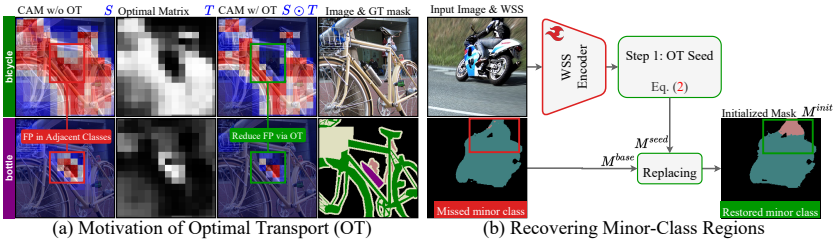
3.2 Dual Features-Driven Hierarchical Rebalancing (DHR)

For our DHR to easily integrate with other WSS methods, we aim to refine pre-propagated WSS masks M^{base} during the segmentation learning phase rather than replacing existing propagation tools. Our DHR approach comprises three steps; this presents a new propagation mechanism $\mathcal{P}^{ours}(\cdot)$ with reconstructing vanished-class regions.

Step 1: Optimal Transport-based Seed Initialization To restore minor-class regions that vanish during the traditional propagation process, we revisit CAMs $\mathcal{A}_{F^{ws}}$ utilized before the propagation $\mathcal{P}^{base}$. As shown in Figure 6(a), we first observe that Optimal Transport (OT) [46] effectively minimizes the occurrence of false positives (FP) in regions adjacent to the target classes:

$$M^{seed} = \mathcal{R}_C(f_{OT}(\mathcal{A}_{F^{ws}}) \odot \mathcal{A}_{F^{ws}}), \tag{2}$$

where $f_{OT}(S) := \arg\min_{T} \sum_{i=1}^{HW} \sum_{j=1}^{C} T_{ij}(1 - S_{ij}) - \lambda\mathcal{H}(T)$ represents the optimal transport matrix T for performing OT based on the input heatmaps S (*e.g.*, CAM). Here, λ is a regularization parameter set to 0.1 and $\mathcal{H}(\cdot)$ denotes the entropy term. Finally, the recovered minor-class regions from M^{seed} are integrated with M^{base} to initialize the WSS mask $M^{init} \in \mathbb{R}^{C\times H\times W}$ in Figure 6(b).



(a) Motivation of Optimal Transport (OT) (b) Recovering Minor-Class Regions

Fig. 6: Illustration of recovering minor-class regions from CAMs with OT. (a): OT reduces false positives in adjacent class regions (*e.g.*, bottle vs. bicycle). **(b):** The input WSS mask often misses most minor-class areas (*i.e.*, person) in the red box. By contrast, refining CAM seeds with OT eliminates overlapping areas, restoring minor-class regions that have previously vanished, as indicated in the green box.

Step 2: USS Feature-based Rebalancing For Inter-class Regions The bottom part (blue) in Figure 7 shows the second step of DHR. We obtain unsupervised feature maps F^{us} from the target image I, defined as $F^{us} = E_{\theta^{us}}(I)$ with dimensions $H \times W \times D_{us}$. We apply class-level average pooling (CAP) to F^{us} using the initial mask M^{init}, producing USS centroids for each class, denoted as $V^{us} = CAP(F^{us}, M^{init})$ with dimensions $C \times D_{us}$. This leads to the creation of the USS-based mask $\hat{S}^{us}$:

$$\hat{S}^{us} := f_{OT}(S^{us}) \odot S^{us}, \tag{3}$$

where $S^{us}_{ij} := ReLU(sim(F^{us}_{ij}, V^{us}))$ in $\mathbb{R}^C$ is the result of updating the initial mask M^{init} by clustering around USS centroids V^{us}, aiding in the distinct categorization of inter-class regions (*e.g.*, kitchenware and furniture). $sim(\cdot)$ is the cosine similarity. The refined outcome, $\hat{S}^{us}$, is the result of applying OT-based optimization to this updated mask. We describe CAP in Appendix A.1.

Step 3: WSS Feature-based Rebalancing For Intra-class Regions In the last stage of DHR, we refine the mask $\hat{S}^{us}$ with weakly-supervised feature maps $F^{ws} = E_{\theta^{ws}}(I)$, sized $\mathbb{R}^{H \times W \times D_{ws}}$, for the precise intra-class segregation, as shown in Figure 7. Following the second step, we generate WSS centroids $V^{ws} = CAP(F^{ws}, M^{init})$ in $\mathbb{R}^C$ by applying CAP to F^{ws} with M^{init}. The final refined mask $\hat{S}^{dh}$ is then formed:

$$\hat{S}^{dh} := f_{OT}(S^{ws}) \odot \hat{S}^{us} \odot \mathbb{1}_{[sim(V^{us}, V^{us}) > \tau]} \tag{4}$$

where $S^{ws}_{ij} := ReLU(sim(F^{ws}_{ij}, V^{ws}))$ in $\mathbb{R}^C$ creates a WSS-based clustered mask for the intra-class distinction. Combined with $\hat{S}^{us}$ through $f_{OT}(S^{ws}) \odot \hat{S}^{us}$, it could enhance the further differentiation across all class regions. Consequently, the final mask $\hat{S}^{dh}$ balances both intra- and inter-class distinctions. The pruning operator $\mathbb{1}_{[sim(V^{us}, V^{us}) > \tau]}$ applies WSS rebalancing for high-correlated USS centroids, thereby leveraging the strengths of WSS and USS features. Lastly, we

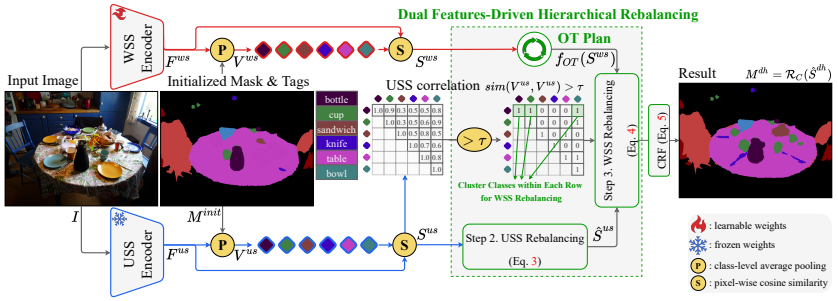


Fig. 7: Visualization of hierarchical rebalancing based on dual features. From initialized seeds M^{init} with restored minor-class regions obtained in the first step of Section 3.2, USS features are used for inter-class grouping, while WSS features are sequentially applied for intra-class separation within each inter-class region. Here, WSS is conditionally applied to cases with high USS correlation (*i.e.*, classes that USS cannot distinguish), successfully differentiating both inter-class and intra-class pixels.

employ the same boundary correction tool $\mathcal{R}_C(\cdot)$ in (1) used in existing WSS studies to generate the final segmentation mask M^{dh}:

$$M^{dh} := \mathcal{R}_C(\hat{S}^{dh}) := \mathcal{R}_C(\mathcal{P}^{ours}(\mathcal{A}_{F^{ws}})) \tag{5}$$

Contrary to former approaches like $\mathcal{R}_C(\mathcal{P}^{base}(\mathcal{A}_{F^{ws}}))$ in (1)) which apply flawed propagation $\mathcal{P}^{base}$, our DHR initiates a new propagation, $\mathcal{P}^{ours}$, incorporating dual feature maps from USS (Step 2) and WSS (Step 3) to discern inter- and intra-class regions for all class rebalancing. We add the specific example related to the pruning operator $\mathbb{1}_{[sim(V^{us},V^{us})>\tau]}$ in Appendix A.2.

3.3 Recursive Learning

The last section presents our training strategy using refined masks M^{dh}. To train both WSS encoder and decoder parameters, we combine two different losses following previous methods [3, 26, 50]:

$$\mathcal{L}_{total} = \mathcal{L}_{cls}(\hat{Y}^{cls}, Y^{cls}) + \mathcal{L}_{seg}(\hat{M}^{seg}, M^{dh}) \tag{6}$$

where $\mathcal{L}_{cls}$ and $\mathcal{L}_{seg}$ denote the multi-label soft margin loss and the per-pixel cross-entropy loss, respectively. We apply global average pooling (GAP) and sigmoid σ to predict class labels, denoted as $\hat{Y}^{cls} = \sigma(GAP(\mathcal{A}_{F^{ws}}))$. The final segmentation output $\hat{M}^{seg} = D_{\theta^{ws}}(F^{ws})$ is obtained from the WSS decoder. As a result, our DHR iteratively refines initial WSS masks, utilizing WSS features to separate inter- and intra-class regions. This process has continuously enhanced the quality of WSS outputs through recursive updates by the WSS network.

4 Experiments

4.1 Experimental Setup

Datasets We conduct all experiments on five segmentation benchmarks, such as PASCAL VOC 2012 (VOC) [17], MS COCO 2014 (COCO) [39], Pascal Context (Context) [43], ADE 2016 (ADE) [63], and COCO-Stuff (Stuff) [4], beyond limited WSS benchmarks. All datasets include image-level class labels and pixel-wise annotations to quantify the performance gap between FSS and WSS. PASCAL VOC 2012 [17], MS COCO 2014 [39], Pascal Context [43], ADE 2016 [63], and COCO-Stuff [4] datasets have 21, 81, 59, 150, and 171 classes, respectively. For a fair comparison, we reproduce previous state-of-the-art WSS models [27, 51] on five segmentation benchmarks.

Reproducibility For our experiments, all unsupervised semantic segmentation (USS) models [19, 29] are trained from the ground up on individual datasets without incorporating additional data. To assess our method's adaptability, we apply our method (DHR) across various weakly-supervised semantic segmentation (WSS) studies [26,36,55] on the PASCAL VOC 2012 dataset [17]. We adhere to the original training protocols of both USS and WSS methods to ensure an equitable comparison. Consequently, our approach has the same evaluation runtime as competing methods. For preliminary results, we generate initial WSS masks from RSEPM [26] with MARS [27] and set τ with 0.8 in (4). In line with standard evaluation practices for segmentation tasks, we employ multi-scale inference and Conditional Random Fields (CRF) [32] for producing segmentation outcomes. All experiments are conducted on a single RTX A100 GPU (80GB) using PyTorch for WSS and USS model implementations.

Evaluation Metrics Our method is evaluated through the mean Intersection over Union (mIoU) metric, aligning with the standard evaluation criterion utilized in previous WSS research [1, 26, 27, 36, 55]. We acquire results for the PASCAL VOC 2012 validation and test datasets directly from the official PASCAL VOC online evaluation platform.

4.2 Comparison with State-of-the-art Approaches

We compare our method with other WSS methods based on CNN and transformer architectures for the quantitative analysis. While recent state-of-the-art methods exploit various supervisions, ranging from LLM [15] to SAM [60], our approach achieves significant results solely by depending on image-level supervision. This outcome underscores the effectiveness of leveraging both USS and WSS features to address vanishing adjacent minor classes. Notably, our method exhibits significant performance gains, surpassing the previous state-of-the-art approach [27] by a margin of at least 1.9% for the PASCAL VOC 2012 validation and test sets utilizing the same configuration. Surprisingly, the proposed method

Table 2: Performance comparison of WSS methods across five benchmarks.

Method	Backbone	Supervision	VOC val	VOC test	COCO val	Context val	ADE val	Stuff val
Weakly-supervised Segmentation Models:								
Feng *et al.* Pattern Recognition'23 [18]	ResNet-101	$\mathcal{I}$	70.5	71.8	-	-	-	
SEPL arXiv'23 [11]	ResNet-101	$\mathcal{I}$+$\mathcal{S}$+$\mathcal{C}$	71.1	-	-	-	-	
SAS AAAI'23 [30]	ResNet-101	$\mathcal{I}$	69.5	70.1	44.8	-	-	-
ToCo CVPR'23 [51]	ViT-B	$\mathcal{I}$	71.1	72.2	42.3	25.0*	10.5*	14.2*
Jiang et al arXiv'23 [23]	ResNet-101	$\mathcal{I}$+$\mathcal{S}$	71.1	72.2	-	-	-	-
ACR CVPR'23 [33]	Wide-ResNet-38	$\mathcal{I}$	71.9	71.9	45.3	-	-	-
BECO CVPR'23 [49]	ResNet-101	$\mathcal{I}$	72.1	71.8	-	-	-	-
MMSCT CVPR'23 [59]	Wide-ResNet-38	$\mathcal{I}$+$\mathcal{C}$	72.2	72.2	45.9	-	-	-
QA-CLIMS MM'23 [15]	ResNet-101	$\mathcal{I}$+$\mathcal{L}$	72.4	72.3	43.2	-	-	-
OCR CVPR'23 [13]	Wide-ResNet-38	$\mathcal{I}$	72.7	72.0	42.5	-	-	-
CLIP-ES CVPR'23 [40]	ResNet-101	$\mathcal{I}$+$\mathcal{C}$	73.8	73.9	45.4	-	-	-
BECO CVPR'23 [49]	MiT-B2	$\mathcal{I}$	73.7	73.5	45.1	-	-	-
WeakTr arXiv'23 [66]	DeiT-S	$\mathcal{I}$	74.0	74.1	46.9	-	-	-
ROSE Information Fusion'24 [9]	ResNet-101	$\mathcal{I}$+$\mathcal{S}$+$\mathcal{C}$	75.4	76.6	48.3	-	-	-
Sun *et al.* arXiv'23 [54]	ResNet-101	$\mathcal{I}$+$\mathcal{S}$+$\mathcal{D}$	77.2	77.1	55.6	-	-	-
FMA-WSSS WACV'24 [60]	ResNet-101	$\mathcal{I}$+$\mathcal{S}$+$\mathcal{C}$	77.3	76.7	48.6	-	-	-
MARS ICCV'23 [27]	ResNet-101	$\mathcal{I}$	77.7	77.2	49.4	39.8*	22.0*	35.7*
DHR (Ours, DeepLabv3+)	ResNet-101	$\mathcal{I}$	**79.6**	**79.8**	**53.9**	**49.0**	**32.9**	**37.4**
DHR (Ours, Mask2Former)	Swin-L	$\mathcal{I}$	**82.3**	**82.3**	**56.8**	**53.6**	**36.9**	**41.1**
Upper Bound (DeepLabv3+) CVPR'18 [8]	ResNet-101	$\mathcal{M}$	80.6*	81.0*	61.8*	54.6*	45.3*	44.2*
Upper Bound (Mask2Former) CVPR'22 [12]	Swin-L	$\mathcal{M}$	86.0	86.1	66.7	64.3*	55.5*	50.6*
Open-vocabulary Segmentation Models:								
MaskCLIP ECCV'22 [64]	ViT-B	$\mathcal{C}$+$\mathcal{T}$	29.3	-	15.5	21.1	10.8	14.7
TCL CVPR'23 [6]	ViT-B	$\mathcal{C}$+$\mathcal{T}$	55.0	-	33.2	33.8	15.6	22.4
Ferret arXiv'23 [61] w/ SAM [28]	ViT-H	$\mathcal{B}$+$\mathcal{T}$+$\mathcal{S}$+$\mathcal{L}$	54.7	-	27.7	22.4	7.6	12.6
Grounded SAM arXiv'24 [48]	Swin-B	$\mathcal{B}$+$\mathcal{T}$+$\mathcal{S}$	46.3	-	35.7	28.1	4.8	18.8

*: we reproduce all results for a fair comparison

$\mathcal{I}$: image-level supervision; $\mathcal{T}$: text supervision (image-text pairs); $\mathcal{A}$: saliency [20]; $\mathcal{L}$: language supervision (*e.g.*, LLM [61]); $\mathcal{M}$: mask supervision; $\mathcal{S}$: SAM [28]; $\mathcal{C}$: CLIP [47]; $\mathcal{B}$: box supervision; $\mathcal{D}$: using the external dataset [37]

consistently demonstrates high performance across five benchmarks, validating the efficacy of WSS techniques compared to open-vocabulary models [48]. Extended results, *e.g.*, per-class comparisons, are provided in Appendix B.

4.3 Discussion

Flexibility We demonstrate the flexibility of our method by comparing it to various WSS and USS methods on the PASCAL VOC 2012 validation dataset. Table 3 presents the results of incorporating two USS methods, such as STEGO [19] and CAUSE [29], into the WSS method (*i.e.*, RSEPM [26]). Both techniques demonstrate a comparable performance improvement of approximately 5%. This observation suggests that if USS methods can sufficiently ensure performance in distinguishing inter-class regions rather than detailed class distinctions, ample efficacy exists in separating intra-class regions through the proposed WSS rebalancing. We visualize qualitative improvements in Appendix C.1.

In Table 4, we compare our method to other model-agnostic WSS approaches [27, 37, 41] based on four WSS methods [1, 26, 36, 55]. We employ CAUSE [29] for our USS method as it shows the best performance in Table 3. The results demonstrate that, regardless of the specific WSS method employed, our approach consistently outperformed others while significantly narrowing the performance gap between WSS baselines and its subsequent FSS models.

Table 3: Comparison with two USS methods in terms of mIoU (%) on the PASCAL VOC 2012 validation set.

WSS	USS	Backbone	mIoU
RSEPM [26]	✗	ResNet-101	74.4
+ MARS [27]	STEGO ICLR'22 [19]	ResNet-101	77.7 (+3.3)
+ DHR (Ours)	STEGO ICLR'22 [19]	ResNet-101	**79.2 (+4.8)**
+ DHR (Ours)	CAUSE Arxiv'23 [29]	ResNet-101	**79.6 (+5.2)**

Table 4: Comparison with four WSS methods on the PASCAL VOC 2012 validation set. Δ means the percentage improvement in the gap between WSS and FSS.

WSS	Backbone	Segmentation	mIoU	Δ
IRNet [1]	ResNet-50	DeepLabv2	63.5	0%
+ MARS [27]	ResNet-50	DeepLabv2	69.8	46%
+ DHR (Ours)	ResNet-50	DeepLabv2	**72.7**	**72%**
Upper Bound (FSS)	ResNet-50	DeepLabv2	76.3	-
SEAM [55]	Wide-ResNet-38	DeepLabv1	64.5	0%
+ ADELE [41]	Wide-ResNet-38	DeepLabv1	69.3	35%
+ MARS [27]	Wide-ResNet-38	DeepLabv1	70.8	46%
+ DHR (Ours)	Wide-ResNet-38	DeepLabv1	**73.6**	**67%**
Upper Bound (FSS)	Wide-ResNet-38	DeepLabv1	78.1	-
AdvCAM [36]	ResNet-101	DeepLabv2	68.1	0%
+ W-OoD [37]	ResNet-101	DeepLabv2	69.8	17%
+ MARS [27]	ResNet-101	DeepLabv2	70.3	22%
+ DHR (Ours)	ResNet-101	DeepLabv2	**73.6**	**56%**
Upper Bound (FSS)	ResNet-101	DeepLabv2	78.0	-
RSEPM [26]	ResNet-101	DeepLabv3+	74.4	0%
+ MARS [27]	ResNet-101	DeepLabv3+	77.7	53%
+ DHR (Ours)	ResNet-101	DeepLabv3+	**79.6**	**84%**
Upper Bound (FSS)	ResNet-101	DeepLabv3+	80.6	-

Effect of DHR To verify the effectiveness of the main components of the proposed method, we evaluate the quality of pseudo masks when applying OT-based seed initialization and dual features-based hierarchical rebalancing (DHR) (see Section 3.2) to different segmentation datasets in Table 5. First, the application of OT-based seed initialization alone enhances the performance by recovering vanished classes in pseudo masks in the second row of Table 5. Second, the addition of DHR substantiated a notably more efficacious impact. Significantly, in the Context and ADE datasets, composed solely of adjacent classes, our approach demonstrated a minimum performance improvement of over 13.6%.

Table 6 presents a more detailed ablation study of the key components of the proposed method on the COCO training dataset, as the COCO dataset has the largest number of classes (*i.e.*, 81) among the official WSS benchmarks [17, 39]. We describe a detailed analysis of our components for other training datasets in Appendix B.3. Applying OT without boundary correction tools like CRF [32] results in a performance increase of +1.7% (the second row), while the introduction of CRF [32] for seed initialization contributed an additional +0.8% (the third row). The joint application of OT and CRF achieved an overall improvement of +2.5% (the fourth row) in mIoU compared to our baseline (*i.e.*, MARS [27]).

Table 5: Effect of key components. We evaluate the quality of pseudo masks (mIoU) on all training datasets.

OT-based Seed Init.	DHR	VOC	COCO	Context	ADE	Stuff
✗	✗	81.8	52.6	51.3	30.2	50.0
✓	✗	82.6 (+0.8)	55.1 (+2.5)	53.6 (+2.3)	32.4 (+2.2)	50.6 (+0.6)
✓	✓	**83.9 (+2.1)**	**58.6 (+6.0)**	**64.9 (+13.6)**	**48.1 (+17.9)**	**53.7 (+3.7)**

Table 6: Detailed analysis of each component on the MS COCO 2014 training dataset.

OT-based Seed Initialization		DHR		mIoU
Optimal Transport	Refinement	USS Balancing	WSS Balancing	
✗	✗	✗	✗	52.6
✓	✗	✗	✗	54.3 (+1.7)
✗	✓	✗	✗	53.4 (+0.8)
✓	✓	✗	✗	55.1 (+2.5)
✓	✓	✓	✗	56.8 (+4.2)
✓	✓	✗	✓	55.3 (+2.7)
✓	✓	✓	✓	**58.6 (+6.0)**

For DHR, employing USS balancing to refine WSS masks with OT-based seed initialization yields a modest improvement of +1.7% (the fifth row). However, the combined application of USS and WSS demonstrated a synergistic effect, resulting in a substantial improvement of +6.0% in the last row, emphasizing the efficacy of their complementary integration. Notably, applying WSS balancing only to WSS masks yields a marginal improvement of 0.2% (the sixth row), underscoring the correctness of the sequential application of USS followed by WSS. Qualitative improvements are elaborated upon in Appendix C.1.

Hyperparameter of DHR In our study, we evaluate the impact of the hyperparameter τ in Eq. (4) for WSS balancing, specifically after segmenting inter-class regions using USS features, as detailed in Eq. (3). As illustrated in Figure 8, setting τ to 1.0, which bypasses WSS balancing, results in only a slight improvement. This result is attributed to increased in incorrect predictions within inter-class areas when WSS features are omitted. Adjusting τ between 0.5 and 0.9 demonstrates minimal impact on segmentation accuracy, highlighting our DHR resilience to variations in τ.

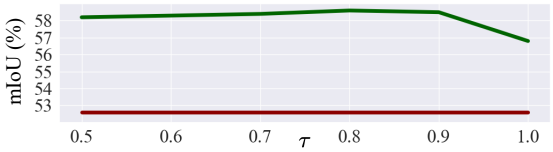


Fig. 8: Sensitivity of a hyperparameter τ. All mIoU values are calculated using the COCO training dataset. The red line is our baseline [6, 27, 48, 51].

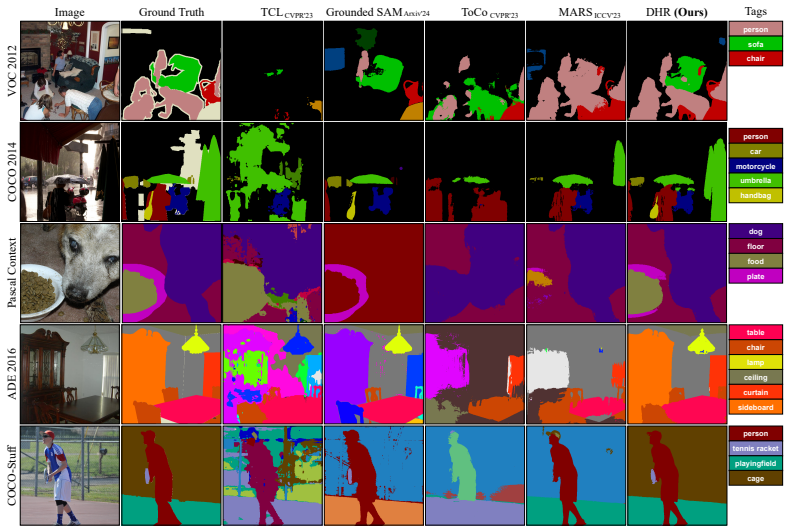


Fig. 9: Qualitative comparison with ours and recent models [6, 27, 48, 51].

4.4 Qualitative Results

Figure 9 compares our DHR's segmentation performance against leading WSS methodologies across five benchmark datasets. The results underscore DHR's enhanced capability in delineating semantic boundaries between adjacent classes, outperforming existing state-of-the-art approaches [6,27,48,51] using tag inputs. For further examples, refer to Appendix C.2.

5 Conclusion

In this study, we present DHR, a novel propagation strategy employing a hierarchical integration of unsupervised and weakly-supervised features to address the issue of adjacent minor classes disappearing. This method has significantly enhanced the performance of leading WSS models, narrowing the gap with FSS by over 50% across five segmentation benchmarks, promising significant impacts on WSS research and extending its use across various fields (*e.g.*, robotics, scene understanding, and autonomous driving) where propagation strategies are essential, showcasing our method's broad utility. DHR distinguishes itself from major vision models like CLIP [38] and Grounded SAM [48] by producing accurate segmentation masks with fewer samples for specific classes. It has become a valuable resource in the industrial and medical fields, where labeling is costly. When used as a starting point for the latest interactive segmentation tools, such as SAM, DHR outperforms other technologies for providing initial seeds, underscoring its potential in WSS. Its efficiency with minimal annotations and limited datasets showcases its considerable promise in segmentation tasks, underlining the importance of WSS in scenarios with significant annotation obstacles.

References

1. Ahn, J., Cho, S., Kwak, S.: Weakly supervised learning of instance segmentation with inter-pixel relations. In: IEEE CVPR. pp. 2209–2218 (2019)
2. Ahn, J., Kwak, S.: Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In: IEEE CVPR. pp. 4981–4990 (2018)
3. Araslanov, N., Roth, S.: Single-stage semantic segmentation from image labels. In: IEEE CVPR. pp. 4253–4262 (2020)
4. Caesar, H., Uijlings, J., Ferrari, V.: COCO-Stuff: Thing and stuff classes in context. In: IEEE CVPR. pp. 1209–1218 (2018)
5. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: IEEE ICCV. pp. 9650–9660 (2021)
6. Cha, J., Mun, J., Roh, B.: Learning to generate text-grounded mask for open-world semantic segmentation from only image-text pairs. In: IEEE CVPR. pp. 11165–11174 (2023)
7. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE TPAMI **40**(4), 834–848 (2017)
8. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: ECCV. pp. 801–818 (2018)
9. Chen, Q., Chen, Y., Huang, Y., Xie, X., Yang, L.: Region-based online selective examination for weakly supervised semantic segmentation. Information Fusion p. 102311 (2024)
10. Chen, Q., Yang, L., Lai, J.H., Xie, X.: Self-supervised image-specific prototype exploration for weakly supervised semantic segmentation. In: IEEE CVPR. pp. 4288–4298 (2022)
11. Chen, T., Mai, Z., Li, R., Chao, W.l.: Segment anything model (sam) enhanced pseudo labels for weakly supervised semantic segmentation. arXiv preprint arXiv:2305.05803 (2023)
12. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: IEEE CVPR. pp. 1290–1299 (2022)
13. 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: IEEE CVPR. pp. 23673–23684 (2023)
14. Cho, J.H., Mall, U., Bala, K., Hariharan, B.: Picie: Unsupervised semantic segmentation using invariance and equivariance in clustering. In: IEEE CVPR. pp. 16794–16804 (2021)
15. Deng, S., Zhuo, W., Xie, J., Shen, L.: Qa-clims: Question-answer cross language image matching for weakly supervised semantic segmentation. In: ACM MM. pp. 5572–5583 (2023)
16. Du, Y., Fu, Z., Liu, Q., Wang, Y.: Weakly supervised semantic segmentation by pixel-to-prototype contrast. In: IEEE CVPR. pp. 4320–4329 (2022)
17. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (VOC) challenge. IJCV **88**(2), 303–338 (2010)
18. Feng, J., Wang, X., Li, T., Ji, S., Liu, W.: Weakly-supervised semantic segmentation via online pseudo-mask correcting. Pattern Recognition Letters **165**, 33–38 (2023)

19. Hamilton, M., Zhang, Z., Hariharan, B., Snavely, N., Freeman, W.T.: Unsupervised semantic segmentation by distilling feature correspondences. In: ICLR (2022), https://openreview.net/forum?id=SaKO6z6Hl0c
20. Hou, Q., Cheng, M.M., Hu, X., Borji, A., Tu, Z., Torr, P.H.: Deeply supervised salient object detection with short connections. In: IEEE CVPR. pp. 3203–3212 (2017)
21. Huang, Z., Wang, X., Wang, J., Liu, W., Wang, J.: Weakly-supervised semantic segmentation network with deep seeded region growing. In: IEEE CVPR. pp. 7014–7023 (2018)
22. Ji, X., Henriques, J.F., Vedaldi, A.: Invariant information clustering for unsupervised image classification and segmentation. In: IEEE ICCV. pp. 9865–9874 (2019)
23. Jiang, P.T., Yang, Y.: Segment anything is a good pseudo-label generator for weakly supervised semantic segmentation. arXiv preprint arXiv:2305.01275 (2023)
24. Jiang, P.T., Yang, Y., Hou, Q., Wei, Y.: L2g: A simple local-to-global knowledge transfer framework for weakly supervised semantic segmentation. In: IEEE CVPR. pp. 16886–16896 (2022)
25. Jo, S., Yu, I.J.: Puzzle-cam: Improved localization via matching partial and full features. In: IEEE ICIP. pp. 639–643. IEEE (2021)
26. Jo, S., Yu, I.J., Kim, K.: Recurseed and edgepredictmix: Single-stage learning is sufficient for weakly-supervised semantic segmentation. arXiv preprint arXiv:2204.06754 (2022)
27. Jo, S., Yu, I.J., Kim, K.: Mars: Model-agnostic biased object removal without additional supervision for weakly-supervised semantic segmentation. In: IEEE ICCV. pp. 614–623 (October 2023)
28. Ke, L., Ye, M., Danelljan, M., Tai, Y.W., Tang, C.K., Yu, F., et al.: Segment anything in high quality. Advances in Neural Information Processing Systems **36** (2024)
29. Kim, J., Lee, B.K., Ro, Y.M.: Causal unsupervised semantic segmentation. arXiv preprint arXiv:2310.07379 (2023)
30. Kim, S., Park, D., Shim, B.: Semantic-aware superpixel for weakly supervised semantic segmentation. In: AAAI. pp. 1142–1150 (2023)
31. 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. arXiv preprint arXiv:2304.02643 (2023)
32. Krähenbühl, P., Koltun, V.: Efficient inference in fully connected CRFs with gaussian edge potentials. NeurIPS **24**, 109–117 (2011)
33. Kweon, H., Yoon, S.H., Yoon, K.J.: Weakly supervised semantic segmentation via adversarial learning of classifier and reconstructor. In: IEEE CVPR. pp. 11329–11339 (2023)
34. Lee, J., Choi, J., Mok, J., Yoon, S.: Reducing information bottleneck for weakly supervised semantic segmentation. NeurIPS **34** (2021)
35. Lee, J., Kim, E., Lee, S., Lee, J., Yoon, S.: Ficklenet: Weakly and semi-supervised semantic image segmentation using stochastic inference. In: IEEE CVPR. pp. 5267–5276 (2019)
36. Lee, J., Kim, E., Yoon, S.: Anti-adversarially manipulated attributions for weakly and semi-supervised semantic segmentation. In: IEEE CVPR. pp. 4071–4080 (2021)
37. Lee, J., Oh, S.J., Yun, S., Choe, J., Kim, E., Yoon, S.: Weakly supervised semantic segmentation using out-of-distribution data. In: IEEE CVPR. pp. 16897–16906 (2022)

38. Li, Y., Hu, P., Liu, Z., Peng, D., Zhou, J.T., Peng, X.: Contrastive clustering. In: AAAI. pp. 8547–8555 (2021)
39. 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: ECCV. pp. 740–755. Springer (2014)
40. 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: IEEE CVPR. pp. 15305–15314 (June 2023)
41. Liu, S., Liu, K., Zhu, W., Shen, Y., Fernandez-Granda, C.: Adaptive early-learning correction for segmentation from noisy annotations. In: IEEE CVPR. pp. 2606–2616 (2022)
42. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. arXiv preprint arXiv:2303.05499 (2023)
43. Mottaghi, R., Chen, X., Liu, X., Cho, N.G., Lee, S.W., Fidler, S., Urtasun, R., Yuille, A.: The role of context for object detection and semantic segmentation in the wild. In: IEEE CVPR. pp. 891–898 (2014)
44. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Howes, R., Huang, P.Y., Xu, H., Sharma, V., Li, S.W., Galuba, W., Rabbat, M., Assran, M., Ballas, N., Synnaeve, G., Misra, I., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision (2023)
45. Ouali, Y., Hudelot, C., Tami, M.: Autoregressive unsupervised image segmentation. In: ECCV. pp. 142–158. Springer (2020)
46. Rachev, S.T.: The monge–kantorovich mass transference problem and its stochastic applications. Theory of Probability & Its Applications **29**(4), 647–676 (1985)
47. 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: ICML. pp. 8748–8763. PMLR (2021)
48. Ren, T., Liu, S., Zeng, A., Lin, J., Li, K., Cao, H., Chen, J., Huang, X., Chen, Y., Yan, F., et al.: Grounded sam: Assembling open-world models for diverse visual tasks. arXiv preprint arXiv:2401.14159 (2024)
49. Rong, S., Tu, B., Wang, Z., Li, J.: Boundary-enhanced co-training for weakly supervised semantic segmentation. In: IEEE CVPR. pp. 19574–19584 (June 2023)
50. Ru, L., Zhan, Y., Yu, B., Du, B.: Learning affinity from attention: end-to-end weakly-supervised semantic segmentation with transformers. In: IEEE CVPR. pp. 16846–16855 (2022)
51. Ru, L., Zheng, H., Zhan, Y., Du, B.: Token contrast for weakly-supervised semantic segmentation. In: IEEE CVPR. pp. 3093–3102 (2023)
52. Seong, H.S., Moon, W., Lee, S., Heo, J.P.: Leveraging hidden positives for unsupervised semantic segmentation. In: IEEE CVPR. pp. 19540–19549 (2023)
53. Sun, G., Wang, W., Dai, J., Van Gool, L.: Mining cross-image semantics for weakly supervised semantic segmentation. In: ECCV. pp. 347–365. Springer (2020)
54. Sun, W., Liu, Z., Zhang, Y., Zhong, Y., Barnes, N.: An alternative to wsss? an empirical study of the segment anything model (sam) on weakly-supervised semantic segmentation problems. arXiv:2305.01586 (2023)
55. Wang, Y., Zhang, J., Kan, M., Shan, S., Chen, X.: Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In: IEEE CVPR. pp. 12275–12284 (2020)

56. Wu, T., Huang, J., Gao, G., Wei, X., Wei, X., Luo, X., Liu, C.H.: Embedded discriminative attention mechanism for weakly supervised semantic segmentation. In: IEEE CVPR. pp. 16765–16774 (2021)
57. Xie, J., Xiang, J., Chen, J., Hou, X., Zhao, X., Shen, L.: C2am: Contrastive learning of class-agnostic activation map for weakly supervised object localization and semantic segmentation. In: IEEE CVPR. pp. 989–998 (2022)
58. Xu, L., Ouyang, W., Bennamoun, M., Boussaid, F., Xu, D.: Multi-class token transformer for weakly supervised semantic segmentation. In: IEEE CVPR. pp. 4310–4319 (2022)
59. Xu, L., Ouyang, W., Bennamoun, M., Boussaid, F., Xu, D.: Learning multi-modal class-specific tokens for weakly supervised dense object localization. In: IEEE CVPR. pp. 19596–19605 (2023)
60. Yang, X., Gong, X.: Foundation model assisted weakly supervised semantic segmentation. In: IEEE WACV. pp. 523–532 (2024)
61. You, H., Zhang, H., Gan, Z., Du, X., Zhang, B., Wang, Z., Cao, L., Chang, S.F., Yang, Y.: Ferret: Refer and ground anything anywhere at any granularity. arXiv preprint arXiv:2310.07704 (2023)
62. Zhang, F., Gu, C., Zhang, C., Dai, Y.: Complementary patch for weakly supervised semantic segmentation. In: IEEE ICCV. pp. 7242–7251 (2021)
63. Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., Torralba, A.: Semantic understanding of scenes through the ade20k dataset. IJCV **127**, 302–321 (2019)
64. Zhou, C., Loy, C.C., Dai, B.: Extract free dense labels from clip. In: ECCV. pp. 696–712. Springer (2022)
65. Zhou, T., Zhang, M., Zhao, F., Li, J.: Regional semantic contrast and aggregation for weakly supervised semantic segmentation. In: IEEE CVPR. pp. 4299–4309 (2022)
66. Zhu, L., Li, Y., Fang, J., Liu, Y., Xin, H., Liu, W., Wang, X.: Weaktr: Exploring plain vision transformer for weakly-supervised semantic segmentation. arXiv preprint arXiv:2304.01184 (2023)
67. Ziegler, A., Asano, Y.M.: Self-supervised learning of object parts for semantic segmentation. In: IEEE CVPR. pp. 14502–14511 (2022)