DuPL: Dual Student with Trustworthy Progressive Learning for Robust Weakly Supervised Semantic Segmentation

Yuanchen Wu[1], Xichen Ye[1], Kequan Yang[1], Jide Li[1], Xiaoqiang Li[1]*
[1] School of Computer Engineering and Science, Shanghai University, China.
{yuanchenwu,yexichen0930,kqyang,iavtvai,xqli}@shu.edu.cn

Abstract

*Recently, One-stage Weakly Supervised Semantic Segmentation (WSSS) with image-level labels has gained increasing interest due to simplification over its cumbersome multi-stage counterpart. Limited by the inherent ambiguity of Class Activation Map (CAM), we observe that one-stage pipelines often encounter confirmation bias caused by incorrect CAM pseudo-labels, impairing their final segmentation performance. Although recent works discard many unreliable pseudo-labels to implicitly alleviate this issue, they fail to exploit sufficient supervision for their models. To this end, we propose a dual student framework with trustworthy progressive learning (**DuPL**). Specifically, we propose a dual student network with a discrepancy loss to yield diverse CAMs for each sub-net. The two sub-nets generate supervision for each other, mitigating the confirmation bias caused by learning their own incorrect pseudo-labels. In this process, we progressively introduce more trustworthy pseudo-labels to be involved in the supervision through dynamic threshold adjustment with an adaptive noise filtering strategy. Moreover, we believe that every pixel, even discarded from supervision due to its unreliability, is important for WSSS. Thus, we develop consistency regularization on these discarded regions, providing supervision of every pixel. Experiment results demonstrate the superiority of the proposed DuPL over the recent state-of-the-art alternatives on PASCAL VOC 2012 and MS COCO datasets. Code is available at* https://github.com/Wu0409/DuPL.

1. Introduction

Weakly supervised semantic segmentation (WSSS) aims at using weak supervision, such as image-level labels [13, 42], scribbles [28, 41], and bounding boxes [22, 32], to alleviate the reliance on pixel-level annotations for segmentation. Among these annotation forms, using image-level labels is the most rewarding yet challenging way, as it only provides the presence of certain classes without offering any localization information. In this paper, we also focus on semantic segmentation using image-level labels.

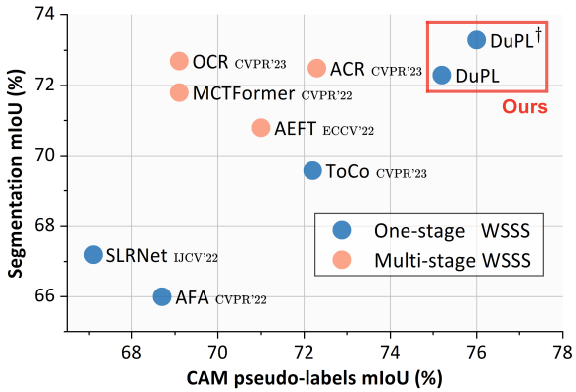


Figure 1. **CAM pseudo-labels (*train*) *vs.* segmentation performance (*val*) on PASCAL VOC 2012**. DuPL outperforms state-of-the-art one-stage competitors and achieves comparable performance with multi-stage methods in terms of CAM pseudo-labels and final segmentation performance. † denotes using ImageNet-21k pretrained parameters.

Prevalent works typically follow a multi-stage pipeline [18], *i.e.*, pseudo-label generation, refinement, and segmentation training. First, the pixel-level pseudo-labels are derived from Class Activation Map (CAM) through classification [46]. Since CAM tends to identify the discriminative semantic regions and fails to distinguish co-occurring objects, the pseudo-labels often suffer from the CAM ambiguity. Thus, they are then refined by training a refinement network [1, 2]. Finally, the refined pseudo-labels are used to train a segmentation model in a fully supervised manner. Recently, to simplify the multi-stage process, many studies proposed one-stage solutions that simultaneously produce pseudo-labels and learn a segmentation head [3, 39, 40]. Despite their enhanced training efficiency, the performance still lags behind their multi-stage counterparts.

One important yet overlooked reason is the confirmation bias of CAM, stemming from the concurrent process of CAM pseudo-label generation and segmentation supervision. For the one-stage pipeline, the segmentation train-

*Corresponding author.

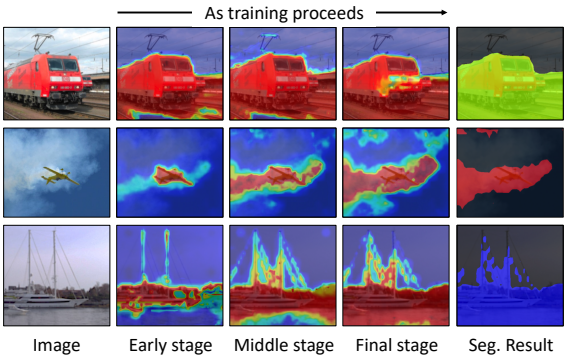


Figure 2. **Confirmation bias of CAM.** As training proceeds, the bias will be consistently reinforced, impairing the final segmentation performance. Here, we use the ViT-B [12] baseline and introduce more unreliable pseudo-labels to amplify this phenomenon.

ing enforces the backbone features to align with the CAM pseudo-labels. Since the backbone features are shared for the segmentation head and the CAM generation, these inaccurate CAM pseudo-labels not only hinder the learning process of segmentation but, more critically, reinforce the CAM's incorrect judgments. As illustrated in Figure 2, this issue consistently deteriorates throughout the training phase and eventually degrades the segmentation performance. Recent one-stage approaches [39, 40, 44] commonly set a fixed and high threshold to filter unreliable pseudo-labels, which prioritizes high-quality supervision to implicitly alleviate this issue. However, this strategy fails to exploit sufficient supervision for their models. Employing a fixed and high threshold inevitably discards many pixels that actually have correct CAM pseudo-labels. Furthermore, these unreliable regions discarded from supervision often exist in semantically ambiguous regions. Excluding them directly from supervision makes the model rarely learn the segmentation in these regions, leading to insufficient training. From this perspective, we believe that every pixel matters for segmentation and should be properly utilized.

To address the above limitations, this work proposes a *dual student framework with trustworthy progressive learning*, dubbed DuPL. Inspired by the co-training [35] paradigm, we equip two student sub-networks that engage in mutual learning. They infer diverse CAMs from different views, and transfer the knowledge learned from one view to the other. To avoid homogenous students, we impose a representation-level discrepancy constraint on the two subnets. This architecture effectively mitigates the confirmation bias resulting from their own incorrect pseudo-labels, thus producing high-fidelity CAMs. Based on our dual student framework, we propose trustworthy progressive learning for sufficient segmentation supervision. We set up a dynamic threshold adjustment strategy to involve more pixels in the segmentation supervision. To overcome the noise in CAM pseudo-labels, we propose an adaptive noise filtering strategy based on the Gaussian Mixture Model. Finally, for the regions where pseudo-labels are excluded from supervision due to their unreliability, we employ an additional strong perturbation branch for each sub-net and develop consistency regularization on these regions. Overall, our main contributions are summarized as follows:

- We explore the CAM confirmation bias in one-stage WSSS. To address this limitation, we propose a dual student architecture. Our experiment proves its effectiveness of reducing the over-activation rate caused by this issue and promotes the quality of CAM pseudo-labels.
- We propose progressive learning with adaptive noise filtering, which allows more trustworthy pixels to participate in supervision. For the regions with filtered pseudo-labels, we develop consistency regularization for sufficient training. This strategy highlights the importance of fully exploiting pseudo-supervision for WSSS.
- Experiments on the PASCAL VOC and MS COCO datasets show that DuPL surpasses state-of-the-art one-stage WSSS competitors and achieves comparable performance with multi-stage solutions (Figure 1). Through visualizing the segmentation results, we observe that DuPL shows much better segmentation robustness, thanks to our dual student and trust-worthy progressive learning.

2. Related work

One-stage Weakly Supervised Semantic Segmentation. Due to the complex process of multi-stage solutions [1, 2], many recent efforts mainly focused on one-stage solutions [3, 39, 40, 44]. A common one-stage pipeline is generating CAM and using online refinement modules to obtain final pseudo-labels [3]. These pseudo-labels are then directly used as the supervision for the segmentation head. Typically, recent works mainly proposed additional modules or training objectives to achieve better segmentation. For instance, Zhang *et al*. [45] introduce a feature-to-prototype alignment loss with an adaptive affinity field, Ru *et al*. [39] leverage pseudo-labels to guide the affinity learning of self-attention, and Xu *et al*. [44] utilize feature correspondence to achieve self-distillation. One common practice of them is that they all set a high and fixed threshold to filter out unreliable pseudo-labels to ensure the quality of supervision. In contrast, we propose a progressive learning strategy fully exploit the potential of every pseudo-label.

Confirmation Bias. This phenomenon commonly occurs in the self-training paradigm of semi-supervised learning (SSL) [21], where the model overfits the unlabeled images assigned with incorrect pseudo-labels. In the above process, this incorrect information is constantly reinforced, causing the unstable training process [4]. Co-training offers an effective solution to this issue [35]. It uses two diverse sub-nets to provide mutual supervision to ensure more

stable and accurate predictions while mitigating the confirmation bias [8, 33]. Motivated by this, we propose a dual student architecture with a representation-level discrepancy loss to generate diverse CAMs. The two sub-nets learn from each other through the other's pseudo-labels, countering the CAM confirmation bias and achieving better object activation. To the best of our knowledge, DuPL is the first work exploring the CAM confirmation bias in one-stage WSSS.

Noise Label Learning in WSSS. In addition to better CAM pseudo-label generation, several recent works aim at learning a robust segmentation model using existing pseudo-labels [10, 27, 31]. URN [27] introduces the uncertainty estimation by the pixel-wise variance between different views to filter noisy labels. Based on the early learning and memorization phenomenon [30], ADELE [31] adaptively calibrates noise labels based on prior outputs in the early learning stage. Different from these works relying on the existing CAM pseudo-labels by other works, the pseudo-labels in one-stage methods continuously update in training. To alleviate the noise pseudo-labels for our progressive learning, we design an online adaptive noise filtering strategy based on the loss feedback from the segmentation head.

3. Method

3.1. Preliminary

We begin with a brief review of how to generate CAM [46] and its pseudo-labels. Given an image, its feature maps $\mathbf{F} \in \mathbb{R}^{D\times H\times W}$ are extracted by a backbone network, where D and $H \times W$ is the channel and spatial dimension, respectively. Then, $\mathbf{F}$ is fed to a global average pooling and a classification layer to output the final classification score. In the above process, we can retrieve the classification weight of every class $\mathbf{W} \in \mathbb{R}^{C\times D}$ and use it to weight and sum the feature maps to generate the CAM:

$$\mathbf{M}(c) = \texttt{ReLU}\left(\sum_{i=1}^{D} \mathbf{W}_{c,i} \cdot \mathbf{F}_i\right), \tag{1}$$

where c is the c-th class and `ReLU` is used to eliminate negative activations. Finally, we apply `max-min` normalization to rescale $\mathbf{M} \in \mathbb{R}^{C\times H\times W}$ to $[0, 1]$. To generate the CAM pseudo-labels, one-stage WSSS methods commonly use two background thresholds τ_l and τ_h to separate the background ($\mathbf{M} \leq \tau_l$), uncertain region ($\tau_l < \mathbf{M} < \tau_h$), and foreground ($\mathbf{M} \geq \tau_h$) [39, 40]. The uncertain part is regarded as unreliable regions with noise, and will not be involved in the supervision of the segmentation head.

3.2. Dual Student Framework

To overcome the confirmation bias of CAM, we propose a co-training-based dual student network where two sub-nets (*i.e.*, ψ_1 and ψ_2) have the same network architecture, and their parameters are independently updated and non-shared. As presented in Figure 3, for the i-th sub-net, it comprises a backbone network ψ_i^f, a classifier ψ_i^c, and a segmentation head ψ_i^s. To ensure that the two sub-nets activate more diverse regions in CAMs, we enforce sufficient diversity to their representations extracted from ψ_i^f, preventing two sub-nets from being homogeneous such that one subnet can learn the knowledge from the other to alleviate the confirmation bias of CAM. Therefore, we set a discrepancy constraint to minimize the cosine similarity between the feature maps of two sub-nets. Formally, denoting the input image as $\mathbf{X}$ and the features from the sub-nets as $\boldsymbol{f}_1 = \psi_1^f(\mathbf{X})$ and $\boldsymbol{f}_2 = \psi_2^f(\mathbf{X})$, we minimize their similarity by:

$$\mathcal{D}(\boldsymbol{f}_1, \boldsymbol{f}_2) = -log\left(1 - \frac{\boldsymbol{f}_1 \cdot \boldsymbol{f}_2}{\|\boldsymbol{f}_1\|_2 \times \|\boldsymbol{f}_2\|_2}\right), \tag{2}$$

where $\|\cdot\|_2$ is $l2$-normalization. Following [7, 14], we define a symmetrized discrepancy loss as:

$$\mathcal{L}_{dis} = \mathcal{D}(\boldsymbol{f}_1, \Delta(\boldsymbol{f}_2)) + \mathcal{D}(\boldsymbol{f}_2, \Delta(\boldsymbol{f}_1)), \tag{3}$$

where Δ is the stop-gradient operation to avoid the model from collapse. This loss is computed for each image, with the total loss being the average across all images.

The segmentation supervision of dual student is bidirectional. One is from $\mathbf{M}_1$ to ψ_2 and the other one is $\mathbf{M}_2$ to ψ_1, where $\mathbf{M}_1$, $\mathbf{M}_2$ are the CAM from the sub-nets ψ_1, ψ_2, respectively. The CAM pseudo-labels $\mathbf{Y}_1$ from $\mathbf{M}_1$ are used to supervise the prediction maps $\mathbf{P}_2$ from the other sub-net's segmentation head ψ_2^s, and vice versa. The segmentation loss of our framework is computed as:

$$\mathcal{L}_{seg} = \texttt{CE}(\mathbf{P}_1, \mathbf{Y}_2) + \texttt{CE}(\mathbf{P}_2, \mathbf{Y}_1), \tag{4}$$

where `CE` is the standard cross-entropy loss function.

3.3. Trustworthy Progressive Learning

Dynamic Threshold Adjustment. As mentioned in Section 3.1, one-stage methods [39, 40, 44] set background thresholds, τ_l and τ_h, to generate pseudo-labels, where τ_h is usually set to a very high value to ensure that only reliable foreground pseudo-labels can participate in the supervision. In contrast, during the training of dual student framework, the CAMs tend to be more reliable gradually. Based on this intuition, to fully utilize more foreground pseudo-labels for sufficient training, we adjust the background threshold τ_h with the cosine descent strategy in every iteration:

$$\tau_h(\mathrm{t}) = \tau_h(0) - \frac{1}{2}\left(\tau_h(0) - \tau_h(\mathrm{T})\right)\left(1 - \cos(\frac{\mathrm{t}\pi}{\mathrm{T}})\right), \tag{5}$$

where t is the current number of iteration and T is the total number of training iterations.

Adaptive Noise Filtering. To further reduce the noise in the produced pseudo-labels that impacts the segmentation

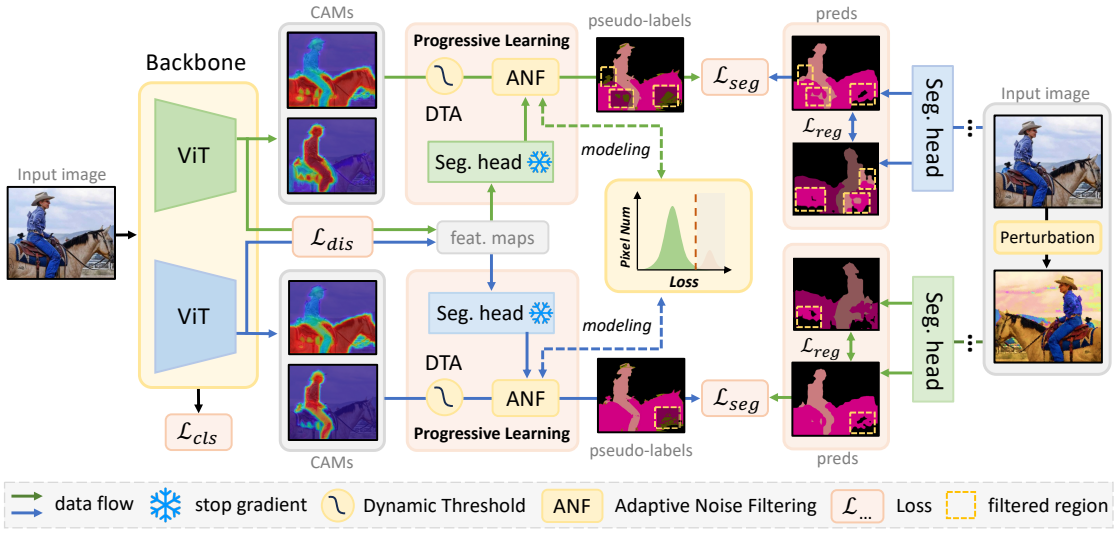


Figure 3. **The overall framework of DuPL.** We use a discrepancy loss $\mathcal{L}_{dis}$ to constrain the two sub-nets to generate diverse CAMs. Their CAM pseudo-labels are utilized for segmentation cross-supervision $\mathcal{L}_{seg}$, which mitigates the CAM confirmation bias. In this process, we set a dynamic threshold to progressively introduce more pixels to segmentation supervision. Adaptive Noise Filtering strategy is equipped to minimize the noise in pseudo-labels via the segmentation loss distribution. To utilize every pixel, the filtered regions are implemented consistency regularization $\mathcal{L}_{reg}$ with their perturbed counterparts. The classifier is simplified for the clear illustration.

generalizability and reinforces the CAM confirmation bias, we develop an adaptive noise filtering strategy to implement *trust-worthy* progressive learning. Previous studies suggest that deep networks tend to fit clean labels faster than noisy ones [5, 15, 37]. This implies that the samples with smaller losses are more likely to be considered as the clean ones before the model overfits the noisy labels. A simple idea is to use a predefined threshold to divide the clean and noisy pseudo-labels based on their training losses. However, it fails to consider that the model's loss distribution is different across various samples, even those within the same class.

To this end, we develop an Adaptive Noise Filtering strategy to distinguish noisy and clean pseudo-labels via the loss distribution, as depicted in Figure 4. Specifically, for the input image $\mathbf{X}$ with its segmentation map $\mathbf{P}$ and CAM pseudo-label $\mathbf{Y}$, we hypothesize the loss of each pixel $x \in \mathbf{X}$, defined as $l^x = \mathrm{CE}\left(\mathbf{P}\left(x\right), \mathbf{Y}\left(x\right)\right)$, is sampled from a Gaussian mixture model (GMM) $\mathcal{P}(x)$ on all pixels with two components, *i.e.*, clean c and noisy n:

$$\mathcal{P}(l^x) = w_c \mathcal{N}(l^x|\mu_c, (\sigma_c)^2) + w_n \mathcal{N}(l^x|\mu_n, (\sigma_n)^2), \quad (6)$$

where $\mathcal{N}(\mu, \sigma^2)$ represents one Gaussian distribution, w_c, μ_c, σ_c and w_n, μ_n, σ_n correspond to the weight, mean, and variance of two components. Thereinto, the component with high loss values corresponds to the noise component. Through Expectation Maximization algorithm [25], we can infer the noise probability $\varrho_n(l^x)$, which is equivalent to the posterior probability of $\mathcal{P}(noise \mid l^x, \mu_n, (\sigma_n)^2)$. If $\varrho_n(l^x) > \gamma$, the corresponding pixel will be classified as noise. Note that not all pseudo-labels $\mathbf{Y}$ are composed of noise, and thus the loss distributions may not have two clear Gaussian distributions. Therefore, we additionally measure the distance between μ_c and μ_n. If $(\mu_n - \mu_c) \leq \eta$, all the pixel pseudo-labels will be regarded as clean ones. Finally, the set of noisy pixel pseudo-labels are determined as

$$\mathcal{X}_n = \{x \mid \varrho_n(l^x) > \gamma, \mu_c - \mu_n > \eta\}, \quad (7)$$

and they are excluded in the segmentation supervision. In DuPL, each sub-net's pseudo-labels are conducted adaptive noise filtering strategy independently.

Every Pixel Matters. In one-stage WSSS, discarding unreliable pseudo-labels that probably contain noises is a common practice to ensure the quality of the segmentation or other auxiliary supervision [39, 40, 44]. Although we gradually introduce more pixels to the segmentation training, there are still many unreliable pseudo-labels being discarded due to the semantic ambiguity of CAM. Typically, throughout the training phase, unreliable regions often exist in non-discriminative regions, boundaries, and background regions. Such an operation may cause the segmentation head to lack sufficient supervision in these regions.

To address this limitation, we treat the regions with unreliable pseudo-labels as *unlabeled* samples. Despite no clear pseudo-labels to supervise the segmentation in these regions, we can regularize the segmentation head to output consistent predictions when fed perturbed versions of the same image. The consistency regularization implicitly imposes the model to comply with the smoothness assumption [6, 20], which provides additional supervision for these regions. Specifically, we first apply strong augmentation ϕ

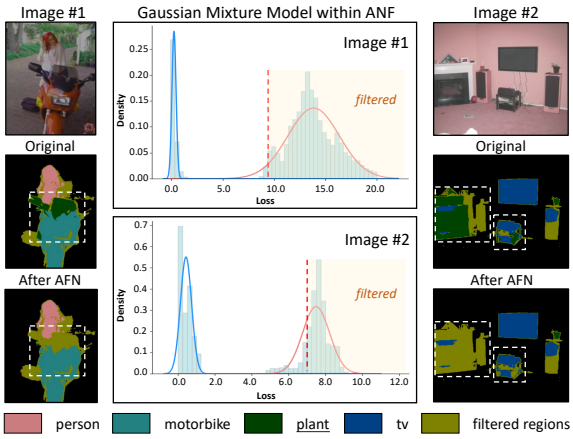


Figure 4. **The loss distribution of images with noisy pseudo-labels.** The model produces incorrect pseudo-labels of `plant`. Two peaks appear in the loss distribution on the two pseudo-labels, and the red peak with anomalous losses is mainly caused by noises. The distribution of normal losses is rescaled for visualization.

to perturb the input image $\phi(\mathbf{X}) \rightarrow \widetilde{\mathbf{X}}$, and then send it to the sub-nets to get the segmentation prediction $\widetilde{\mathbf{P}}_i$ from ψ_i^s. Using the pseudo-label $\phi'(\mathbf{Y}_i)$ taking the same affine transformation in ϕ as the supervision, the consistency regularization of the i-th sub-net is formulated as:

$$\mathcal{L}_{reg\_i} = \frac{1}{\mid \mathcal{M}_i \mid} \sum_{x \in \mathbf{X}} \mathrm{CE}\left[\widetilde{\mathbf{P}}_i(\phi(x)), \phi'(\mathbf{Y}_i(x))\right] \cdot \mathcal{M}_i, \tag{8}$$

where $\mathcal{M}_i$ is the mask indicating the filtered pixels with unreliable pseudo-labels of the i-th sub-net. The filtered pixel is masked as 1, and otherwise it is 0. The total regularization loss of our dual student framework is $\mathcal{L}_{reg} = \mathcal{L}_{reg\_1} + \mathcal{L}_{reg\_2}$. This loss is computed for each image, with the total loss being the average across all images.

3.4. Training objective of DuPL

As illustrated in Figure 3, DuPL consists four training objectives, that are, the classification loss $\mathcal{L}_{cls}$, the discrepancy loss $\mathcal{L}_{dis}$, the segmentation loss $\mathcal{L}_{seg}$, and consistency regularization loss $\mathcal{L}_{reg}$. Following the common practice in WSSS, we use the multi-label soft margin loss for classification. The total optimization objective of DuPL is the linear combination of the above loss terms:

$$\mathcal{L} = \mathcal{L}_{cls} + \lambda_1 \mathcal{L}_{dis} + \lambda_2 \mathcal{L}_{seg} + \lambda_3 \mathcal{L}_{reg}, \tag{9}$$

where λ_i is the weight to rescale the loss terms.

4. Experiments

4.1. Experimental Settings

Datasets. We evaluate the proposed DuPL on the two standard WSSS datasets, *i.e.*, PASCAL VOC 2012 and MS COCO 2014 datasets. For the VOC 2012 dataset, it is extended with the SBD dataset [16] following common practice. The train, val, and test set are composed of 10582, 1449, and 1456 images, respectively. The test performance of DuPL is evaluated on the official evolution server. For the COCO 2014 dataset, its train and val set involve 82k and 40k images, respectively. The mean Intersection-over-Union (mIoU) is reported for performance evaluation.

Method	Backbone	`train`	`val`
Multi-stage WSSS Methods			
PPC [13] CVPR'2022 + PSA [1]	WR38	73.3	–
ACR [19] CVPR'2023 + IRN [2]	WR38	72.3	–
One-stage WSSS Methods			
1Stage [3] CVPR'2020	WR38	66.9	65.3
ViT-PCM [38] ECCV'2022	ViT-B†	67.7	66.0
AFA [39] CVPR'2022	MiT-B1	68.7	66.5
ToCo [40] CVPR'2023	ViT-B	72.2	70.5
DuPL	ViT-B	75.1	73.5
DuPL†	ViT-B†	**76.0**	**74.1**

Table 1. **Evaluation of CAM pseudo labels.** The results are evaluated on the VOC `train` and `val` set and reported in mIoU (%). † denotes using ImageNet-21k pretrained parameters.

Network Architectures of DuPL. We use the ViT-B [12] with a lightweight classifier and a segmentation head, and the patch token contrast loss [40] as our baseline network. The classifier is a fully connected layer. The segmentation head consists of two 3×3 convolutional layers (with a dilation rate of 5) and one 1×1 prediction layer. The patch token contrast loss is applied to alleviate the over-smoothness issue of CAM in ViT-like architectures. DuPL is composed of two subnets with the baseline settings, where the backbones are initialized with ImageNet pretrained weights.

Implement Details. We adopt the AdamW optimizer with an initial learning rate set to $6e^{-5}$ and a weight decay factor 0.01. The input images are augmented using the strategy in [40], and cropped to 448×448. For the strong perturbations, we adopt Random Augmentation strategy [11] on color and apply additional scaling and horizontal flipping. In the inference stage, following the common practice in WSSS, we use multi-scale testing and dense CRF processing.

For experiments on the VOC 2012 dataset, the batch size is set as 4. The total iteration is set as 20k with 2k iterations warmed up for the classifiers and 6k iterations warmed up for the segmentation heads before conducting Adaptive Noise Filtering. The background thresholds $(\tau_l, \tau_h(0), \tau_h(\mathrm{T}))$ are set as $(0.25, 0.7, 0.55)$. The thresholds (γ, η) of Adaptive Noise Filtering are set as $(0.9, 1.0)$. The weight factors $(\lambda_1, \lambda_2, \lambda_3)$ of the loss terms in Section 3.4 are set as $(0.1, 0.2, 0.05)$. For the COCO dataset, the batch size is set as 8. The network is trained for 80k iter-

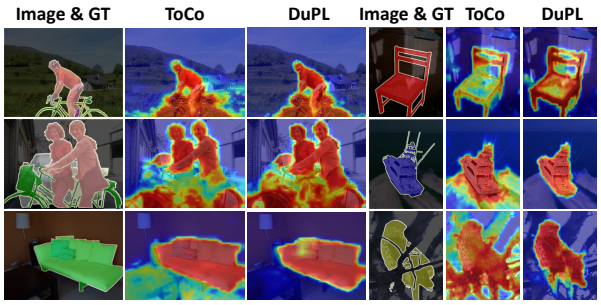


Figure 5. **Visual comparison of CAMs.** We compare the state-of-the-art one-stage approach, ToCo [40], with our proposed DuPL. DuPL not only suppresses over-activations but also achieves more complete object activation coverage.

ations with 5k iterations warmed up for the classifier, and 20k iterations warmed up for the segmentation head. The other settings are remained the same.

4.2. Experimental Results

CAM and Pseudo-labels. We begin by visualizing the CAM of DuPL in Figure 5. We can find that, using the same ViT-B backbone with ImageNet-1k pretrained weights, our method can generate more complete and accurate CAMs when compared to current state-of-the-art one-stage work, *i.e.*, ToCo [40]. Then, we evaluate the CAM pseudo-labels on the `train` and `val` set of the VOC dataset and compare them with recent state-of-the-art WSSS methods. In one-stage methods, the pseudo-labels are directly generated using CAMs, while those of multi-stage methods are produced by the initial seed generation and refinement processes. The results are presented in Table 1. As can be seen, DuPL significantly outperforms the recent one-stage competitors and even surpasses the multi-stage methods. Compared with other methods with ViT-B baseline, our methods can produce higher quality pseudo-labels than the competitors with both the ImageNet-1k and ImageNet-21k pretrained weights. Using ViT-B with ImageNet-21k pretrained weights, we boost the pseudo-label performance to 76.0% (**+3.8%**) and 74.1% (**+3.6%**) on the `train` and `val` set, respectively.

Final Segmentation Results. Table 2 reports the final segmentation performance of DuPL. To show the superiority of the proposed method, we compare our performance with both one-stage and multi-stage prior arts. Notably, the proposed DuPL achieves 73.3% (**+3.5%**), 72.8% (**+2.3%**) and 44.6% (**+3.3%**) mIoU on the VOC `val`, `test` and COCO `val` set, respectively, which significantly surpasses recent one-stage methods. The performance of DuPL strongly supports that *fully exploiting the trustworthy pseudo-labels*

[1] `http://host.robots.ox.ac.uk:8080/anonymous/103D8M.html`

[2] `http://host.robots.ox.ac.uk:8080/anonymous/R7RLMS.html`

	Sup.	Net.	VOC		COCO
			val	test	val
Multi-stage WSSS Methods.					
EPS [24] CVPR'2021	$\mathcal{I}+\mathcal{S}$	DL-V2	71.0	71.8	–
L2G [17] CVPR'2022	$\mathcal{I}+\mathcal{S}$	DL-V2	72.1	71.7	44.2
PPC [13] CVPR'2022	$\mathcal{I}+\mathcal{S}$	DL-V2	72.6	73.6	–
Lin *et al.* [29] CVPR'2023	$\mathcal{I}+\mathcal{T}$	DL-V2	71.1	71.4	45.4
ReCAM [9] CVPR'2022	$\mathcal{I}$	DL-V2	68.4	68.2	45.0
W-OoD [23] CVPR'2022	$\mathcal{I}$	WR-38	70.7	70.1	–
ESOL [26] NeurIPS'2022	$\mathcal{I}$	DL-V2	69.9	69.3	42.6
MCTformer [43] CVPR'2022	$\mathcal{I}$	WR-38	71.9	71.6	42.0
OCR [10] CVPR'2023	$\mathcal{I}$	WR-38	72.7	72.0	42.5
ACR [19] CVPR'2023	$\mathcal{I}$	DL-V2	71.9	71.9	45.3
One-stage WSSS Methods.					
RRM [45] AAAI'2020	$\mathcal{I}$	WR-38	62.6	62.9	–
1Stage [3] CVPR'2020	$\mathcal{I}$	WR-38	62.7	64.3	–
AFA [39] CVPR'2022	$\mathcal{I}$	MiT-B1	66.0	66.3	38.9
SLRNet [34] IJCV'2022	$\mathcal{I}$	WR-38	67.2	67.6	35.0
TSCD [44] AAAI'2023	$\mathcal{I}$	MiT-B1	67.3	67.5	40.1
ToCo [40] CVPR'2023	$\mathcal{I}$	ViT-B	69.8	70.5	41.3
DuPL	$\mathcal{I}$	ViT-B	72.2	71.6[1]	43.5
DuPL†	$\mathcal{I}$	ViT-B†	**73.3**	**72.8**[2]	**44.6**

Table 2. **Semantic Segmentation Results**. "Sup." denotes the supervision type. $\mathcal{I}$: Image-level labels; $\mathcal{S}$: Saliency maps. $\mathcal{T}$: text-driven supervision from CLIP [36]. "Net." denotes the backbone in one-stage methods and the segmentation network in multi-stage methods. † denotes using ImageNet-21k pretrained weights.

is very important to single-stage methods. Also, DuPL proves that using the one-stage pipeline is *strong enough* to achieve competitive WSSS performance with multi-stage approaches. Along with the quantitative comparison results, we visualize and compare the segmentation masks of DuPL, ToCo [40], and ground-truths in Figure 6. We can see that DuPL predicts more accurate objects in challenging scenes, which are close to their ground truths.

Fully-Supervised Counterparts. As presented in Table 3, the one-stage competitors adopt various backbones, *e.g.*, Wide ResNet38 (WR-38), MixFormer-Base1 (MiT-B1), and ViT-Base (ViT-B). To eliminate the impact of backbone on segmentation results for fair comparison, we compared the performance gap between the methods and their fully supervised counterpart. Notably, when using the ImageNet-1k pre-trained weight, DuPL achieves 72.2% mIoU and **90.1%** of its upper bound performance, significantly ahead of recent one-stage one-stage methods (**+3.4%**).

4.3. Ablation studies and Analysis

Effectiveness of Components. The proposed DuPL consists of a dual student (DS) architecture and trust-worthy progressive learning. Within the progressive learning, we have dynamic threshold adjustment (DTA) and Adaptive

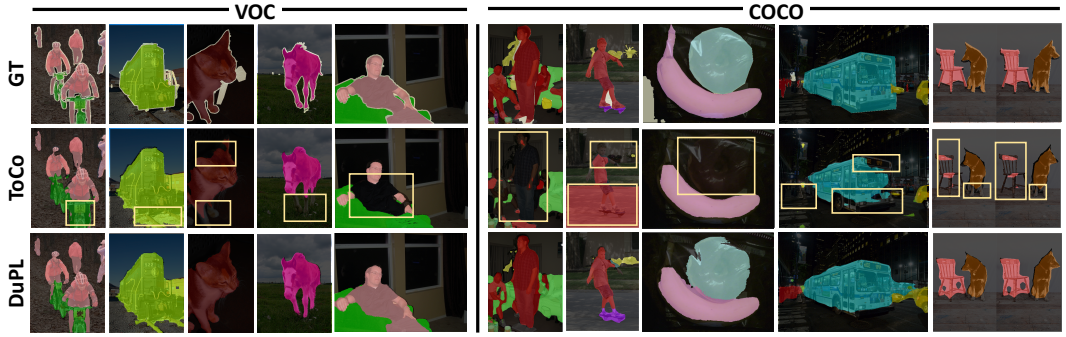


Figure 6. **Visualization of segmentation results on PSCAL VOC 2012 and MS COCO datasets**. We compare the results of DuPL with those of ToCo [40]. Both of them use ViT-B with ImageNet-1k as the backbone for fair comparison.

Noise Filtering (ANF). In addition to the basic classification and segmentation loss, DuPL also incorporates two training losses, *i.e.*, discrepancy loss $\mathcal{L}_{dis}$ and consistency regularization loss $\mathcal{L}_{reg}$. We now investigate the contributions of each module and loss in DuPL.

The experiment results are presented in Table 4. We can observe that employing solely dual student architecture brings a slight improvement of nearly 2% mIoU for CAM pseudo-labels, resulting in 63.8% mIoU of segmentation performance. In this setting, the CAM diversity arises merely from the randomly initialized segmentation heads, thus the CAMs from the two sub-nets are still highly identical, leaving a huge space for improvement. When incorporating $\mathcal{L}_{dis}$, the performance of CAM pseudo-label is improved to 67.3% mIoU, indicating that it can further benefit the effectiveness of dual student architecture. As CAM becomes increasingly reliable, DTA progressively introduces more pixels into the segmentation supervision and improves the segmentation performance by 2.6%. The ANF suppresses noise pseudo-labels and improves segmentation performance by 1.5%. It's noted that high-quality supervision of segmentation benefits the CAM quality, and DTA with ANF significantly improves the pseudo labels by 4.3%. With the motivation of "every pixel matters", $\mathcal{L}_{dis}$ ultimately boosts the segmentation performance to 69.9% mIoU, leading to the state-of-the-art.

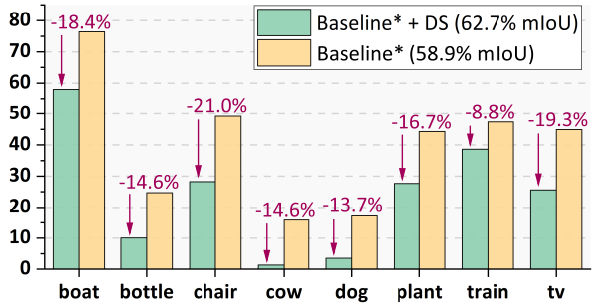
Analysis of Dual Student. DuPL adopts the mutual supervision of two student sub-nets to alleviate the confirmation bias introduced by their own incorrect pseudo-labels. The confirmation bias issue can be reflected by the over-activation (OA) rate. A higher OA rate means the model activates more incorrect pixels for the target classes, causing a more severe CAM confirmation bias. Here, we count the number of the false positive (FP) and true positive (TP) pixel pseudo-labels for each class, and calculate the OA rate (*i.e.*, `FP` / (`TP` + `FP`)). We first compare the baseline and the ablated variant with Dual Student (*i.e.*, baseline + DS + $\mathcal{L}_{dis}$) under a low background threshold setting ($\tau_h = 0.5$). From Figure 7a, we can see due to the confirmation bias, the baseline over-activates lots of incorrect regions, resulting in subpar segmentation outcomes (only 58.9% mIoU). With dual student, the ablated version significantly reduces the OA rate by over 15% in many classes, and even reduces the OA rate to below 5% for some categories (such as `cow` and `dog`). Further, we evaluate the OA rate of ToCo [40] and DuPL. From Figure 7b, we can see that ToCo also suffers from the confirmation bias problem, with OA rate exceeding 30% in several categories. In contrast, the proposed DuPL significantly overcomes this problem in these severely over-activated classes, which reflects the effectiveness of our architecture.

Method	**BB.**	`val` ($\mathcal{F}$)	`val` ($\mathcal{I}$)	*ratio (%)*
1Stage [3]	WR38	80.8	62.7	77.6
SLRNet [34]	WR38	80.8	67.2	83.2
AFA [39]	MiT-B1	78.7	66.0	83.9
ToCo [40]	ViT-B	80.5	69.8	86.7
DuPL	**ViT-B**	**80.5**	**72.2**	**90.1**
DuPL†	**ViT-B**†	**82.3**	**73.3**	**89.1**

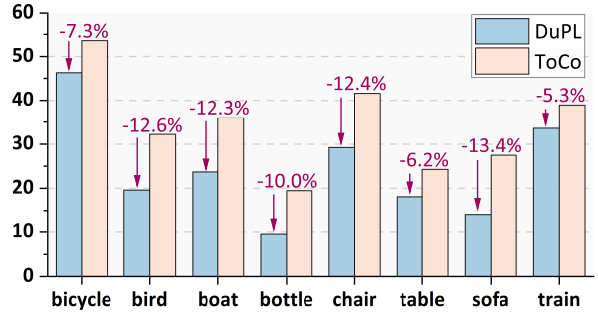
Table 3. **The performance comparison with fully supervised counterparts on the VOC dataset.** The pixel pseudo labels are used to supervise the segment head. $\mathcal{F}$: fully-supervised supervision. $\mathcal{I}$: image-level supervision (WSSS). *ratio* = `val` ($\mathcal{I}$) / `val` ($\mathcal{F}$). † denotes using ImageNet-21k pretrained weights.

Baseline	**DS**	$\mathcal{L}_{\text{dis}}$	**DTA**	**ANF**	$\mathcal{L}_{\text{reg}}$	**M**	**Seg.**
✓						63.2	62.3
✓	✓					65.4	63.8
✓	✓	✓				67.3	64.1
✓	✓	✓	✓			69.2	66.7
✓	✓	✓	✓	✓		71.6	68.2
✓	✓	✓	✓	✓	✓	**73.5**	**69.9**

Table 4. **Ablation Study.** "M" denotes the CAM performance and "Seg." denotes the segmentation performance. CRF post-processing is not conducted in the ablation study.



(a) Comparison of the baseline and the baseline with dual student.



(b) Comparison of ToCo [40] and the proposed DuPL.

Figure 7. **Effectiveness evaluation of our proposed method.** The OA rate (%) are evaluated on the VOC `val` set. "*" denotes the baseline is trained under a low background threshold ($\tau_h = 0.5$) to aggregate the CAM conformation bias. The per-class results can be viewed in *Supplementary Material.*

Dynamic Threshold Adjustment. In DuPL, $\tau_h(\mathrm{t})$ is a dynamic background threshold that progressively decreases to $\tau_h(\mathrm{T})$ with training, aiming at involving more pseudo-labels into the segmentation supervision. Table 5a shows the impact of different $\tau_h(\mathrm{T})$ on the CAM and segmentation performance. We observe that when $\tau_h(\mathrm{T})$ ranges from 0.65 to 0.55, the model's performance exhibits steady improvement. However, when $\tau_h(\mathrm{T})$ is smaller than 0.55, the excessive introduction of noises becomes challenging to suppress, thus yielding a negative impact on the model performance. Nevertheless, the model continues to improve in comparison to the case with a relatively higher $\tau_h(\mathrm{T})$.

Warm-up Stage for The Segmentation Head. Motivated by the Early-learning nature of deep networks, ANF uses the feedback from the segmentation head to filter the noise pseudo-labels. This requires the segmentation head to fit the CAM pseudo-labels properly. Incorporating ANF too early may risk filtering out correct pseudo-labels due to under-fitting, while introducing ANF too late may lead to the model having already memorized noisy pseudo-labels, making it challenging to discriminate them. In Table 5b, we report the impact on the warm-up stage for the segmentation head. We show that warming up the segmentation head using 8000 iterations can achieve the best performance.

Discrepancy strategy in Dual Student. We apply the discrepancy constraint on the representation level to make each sub-nets generate more diverse CAMs. In Table 6, we compare the impact of different discrepancy strategies. It shows that only introducing $\mathcal{L}_{dis}$ on the representation level is more beneficial for two sub-nets to transfer the knowledge learned from one view to the other through CAM pseudo-labels, thus yielding favorable performance.

$\tau_h(\mathrm{T})$	M	Seg.
0.65	69.4	68.1
0.60	71.8	70.9
0.55	**73.5**	**72.2**
0.50	72.3	71.5

(a) Background threshold τ_h.

Iter	M	Seg.
6000	72.4	70.9
8000	**73.5**	**72.2**
10000	72.6	71.7
12000	71.1	69.4

(b) Warm-up stage.

Table 5. **Impact of hyper-parameters.** The results are evaluated on the VOC `val` set. The default settings are marked in color.

	None	Diff. Aug	$\mathcal{L}_{\mathbf{dis}}$	Diff. Aug + $\mathcal{L}_{\mathbf{dis}}$
M	69.6	70.7	**73.5**	70.9
Seg.	68.9	69.8	**72.2**	69.4

Table 6. **Different discrepancy strategies in Dual student.** The results are evaluated on the VOC `val` set. "Diff. Aug" denotes that the input images of two-subnets are augmented differently, and the CAM pseudo-labels will be re-transformed to fit the inputs for the other sub-net.

5. Conclusion

This work aims to address the problem of CAM confirmation bias and fully utilize the CAM pseudo-labels for better WSSS. Specifically, we develop a dual student architecture with two sub-nets that mutually provide the pseudo-labels for the other, which is empirically proved to counter the CAM confirmation bias well. With better CAM activations during the training process, we gradually introduce more pixels into the supervision for sufficient segmentation training. We overcome the excessive noisy pseudo-labels brought by the above operation by proposing an adaptive noise filter strategy. Such a trustworthy progressive learning paradigm significantly boosts the WSSS performance. Motivated by the idea that "every pixel matters", instead of discarding unreliable labels, we fully leverage them through consistency regularizations. The experiment results demonstrate that DuPL significantly outperforms other one-stage competitors and archives competitive performance with multi-stage solutions.

Acknowledgements. This work is supported in part by Shanghai science and technology committee under grant No. 22511106005. We appreciate the High Performance Computing Center of Shanghai University, and Shanghai Engineering Research Center of Intelligent Computing System for the computing resources and technical support.

References

[1] Jiwoon Ahn and Suha Kwak. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4981–4990, 2018. 1, 2, 5

[2] 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*, pages 2209–2218, 2019. 1, 2, 5

[3] 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*, pages 4253–4262, 2020. 1, 2, 5, 6, 7

[4] Eric Arazo, Diego Ortego, Paul Albert, Noel E O'Connor, and Kevin McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2020. 2

[5] Devansh Arpit, Stanislaw Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. In *International conference on machine learning*, pages 233–242. PMLR, 2017. 4

[6] Philip Bachman, Ouais Alsharif, and Doina Precup. Learning with pseudo-ensembles. *Advances in neural information processing systems*, 27, 2014. 4

[7] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758, 2021. 3

[8] Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2613–2622, 2021. 3

[9] Zhaozheng Chen, Tan Wang, Xiongwei Wu, Xian-Sheng Hua, Hanwang Zhang, and Qianru Sun. Class re-activation maps for weakly-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 969–978, 2022. 6

[10] Zesen Cheng, Pengchong Qiao, Kehan Li, Siheng Li, Pengxu Wei, Xiangyang Ji, Li Yuan, Chang Liu, and Jie Chen. Out-of-candidate rectification for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23673–23684, 2023. 3, 6

[11] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703, 2020. 5

[12] 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. *arXiv preprint arXiv:2010.11929*, 2020. 2, 5

[13] Ye Du, Zehua Fu, Qingjie Liu, and Yunhong Wang. Weakly supervised semantic segmentation by pixel-to-prototype contrast. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4320–4329, 2022. 1, 5, 6

[14] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020. 3

[15] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *Advances in neural information processing systems*, 31, 2018. 4

[16] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *2011 international conference on computer vision*, pages 991–998. IEEE, 2011. 5

[17] Peng-Tao Jiang, Yuqi Yang, Qibin Hou, and Yunchao Wei. L2g: A simple local-to-global knowledge transfer framework for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16886–16896, 2022. 6

[18] 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*, pages 695–711. Springer, 2016. 1

[19] Hyeokjun Kweon, Sung-Hoon Yoon, and Kuk-Jin Yoon. Weakly supervised semantic segmentation via adversarial learning of classifier and reconstructor. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11329–11339, 2023. 5, 6

[20] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016. 4

[21] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, page 896. Atlanta, 2013. 2

[22] Jungbeom Lee, Jihun Yi, Chaehun Shin, and Sungroh Yoon. Bbam: Bounding box attribution map for weakly supervised semantic and instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2643–2652, 2021. 1

[23] Jungbeom Lee, Seong Joon Oh, Sangdoo Yun, Junsuk Choe, Eunji Kim, and Sungroh Yoon. Weakly supervised semantic segmentation using out-of-distribution data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16897–16906, 2022. 6

[24] Seungho Lee, Minhyun Lee, Jongwuk Lee, and Hyunjung Shim. Railroad is not a train: Saliency as pseudo-pixel supervision for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5495–5505, 2021. 6

[25] Junnan Li, Richard Socher, and Steven CH Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. *arXiv preprint arXiv:2002.07394*, 2020. 4

[26] Jinlong Li, Zequn Jie, Xu Wang, Xiaolin Wei, and Lin Ma. Expansion and shrinkage of localization for weakly-supervised semantic segmentation. *arXiv preprint arXiv:2209.07761*, 2022. 6

[27] Yi Li, Yiqun Duan, Zhanghui Kuang, Yimin Chen, Wayne Zhang, and Xiaomeng Li. Uncertainty estimation via response scaling for pseudo-mask noise mitigation in weakly-supervised semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1447–1455, 2022. 3

[28] 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*, pages 3159–3167, 2016. 1

[29] Yuqi Lin, Minghao Chen, Wenxiao Wang, Boxi Wu, Ke Li, Binbin Lin, Haifeng Liu, and Xiaofei He. Clip is also an efficient segmenter: A text-driven approach for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15305–15314, 2023. 6

[30] Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Carlos Fernandez-Granda. Early-learning regularization prevents memorization of noisy labels. *Advances in neural information processing systems*, 33:20331–20342, 2020. 3

[31] Sheng Liu, Kangning Liu, Weicheng Zhu, Yiqiu Shen, and Carlos Fernandez-Granda. Adaptive early-learning correction for segmentation from noisy annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2606–2616, 2022. 3

[32] Youngmin Oh, Beomjun Kim, and Bumsub Ham. Background-aware pooling and noise-aware loss for weakly-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6913–6922, 2021. 1

[33] Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12674–12684, 2020. 3

[34] Junwen Pan, Pengfei Zhu, Kaihua Zhang, Bing Cao, Yu Wang, Dingwen Zhang, Junwei Han, and Qinghua Hu. Learning self-supervised low-rank network for single-stage weakly and semi-supervised semantic segmentation. *International Journal of Computer Vision*, 130(5):1181–1195, 2022. 6, 7

[35] Siyuan Qiao, Wei Shen, Zhishuai Zhang, Bo Wang, and Alan Yuille. Deep co-training for semi-supervised image recognition. In *Proceedings of the european conference on computer vision (eccv)*, pages 135–152, 2018. 2

[36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 6

[37] Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *International conference on machine learning*, pages 4334–4343. PMLR, 2018. 4

[38] Simone Rossetti, Damiano Zappia, Marta Sanzari, Marco Schaerf, and Fiora Pirri. Max pooling with vision transformers reconciles class and shape in weakly supervised semantic segmentation. In *European Conference on Computer Vision*, pages 446–463. Springer, 2022. 5

[39] 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*, pages 16846–16855, 2022. 1, 2, 3, 4, 5, 6, 7

[40] Lixiang Ru, Heliang Zheng, Yibing Zhan, and Bo Du. Token contrast for weakly-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3093–3102, 2023. 1, 2, 3, 4, 5, 6, 7, 8

[41] Paul Vernaza and Manmohan Chandraker. Learning random-walk label propagation for weakly-supervised semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7158–7166, 2017. 1

[42] Yuanchen Wu, Xiaoqiang Li, Songmin Dai, Jide Li, Tong Liu, and Shaorong Xie. Hierarchical semantic contrast for weakly supervised semantic segmentation. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 1542–1550, 2023. 1

[43] Lian Xu, Wanli Ouyang, Mohammed Bennamoun, Farid Boussaid, and Dan Xu. Multi-class token transformer for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4310–4319, 2022. 6

[44] Rongtao Xu, Changwei Wang, Jiaxi Sun, Shibiao Xu, Weiliang Meng, and Xiaopeng Zhang. Self correspondence distillation for end-to-end weakly-supervised semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3045–3053, 2023. 2, 3, 4, 6

[45] Bingfeng Zhang, Jimin Xiao, Yunchao Wei, Mingjie Sun, and Kaizhu Huang. Reliability does matter: An end-to-end weakly supervised semantic segmentation approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12765–12772, 2020. 2, 6

[46] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016. 1, 3