Weakly Supervised Semantic Segmentation for Driving Scenes

Dongseob Kim[*1], **Seungho Lee**[*1], **Junsuk Choe**[2], **Hyunjung Shim**[†3]

[1] Yonsei University, South Korea
[2] Sogang University, South Korea
[3] Korea Advanced Institute of Science & Technology, South Korea
{kou.k, seungholee}@yonsei.ac.kr, jschoe@sogang.ac.kr, kateshim@kaist.ac.kr

Abstract

State-of-the-art techniques in weakly-supervised semantic segmentation (WSSS) using image-level labels exhibit severe performance degradation on driving scene datasets such as Cityscapes. To address this challenge, we develop a new WSSS framework tailored to driving scene datasets. Based on extensive analysis of dataset characteristics, we employ Contrastive Language-Image Pre-training (CLIP) as our baseline to obtain pseudo-masks. However, CLIP introduces two key challenges: (1) pseudo-masks from CLIP lack in representing small object classes, and (2) these masks contain notable noise. We propose solutions for each issue as follows. (1) We devise Global-Local View Training that seamlessly incorporates small-scale patches during model training, thereby enhancing the model's capability to handle small-sized yet critical objects in driving scenes (e.g., *traffic light*). (2) We introduce Consistency-Aware Region Balancing (CARB), a novel technique that discerns reliable and noisy regions through evaluating the consistency between CLIP masks and segmentation predictions. It prioritizes reliable pixels over noisy pixels via adaptive loss weighting. Notably, the proposed method achieves 51.8% mIoU on the Cityscapes test dataset, showcasing its potential as a strong WSSS baseline on driving scene datasets. Experimental results on CamVid and WildDash2 demonstrate the effectiveness of our method across diverse datasets, even with small-scale datasets or visually challenging conditions. The code is available at https://github.com/k0u-id/CARB.

Introduction

Recent advancements in weakly supervised semantic segmentation (WSSS) using image-level labels have demonstrated impressive results, achieving performance levels of over 90% compared to full supervised models on the PASCAL VOC dataset (Lee, Kim, and Shim 2022; Yoon et al. 2022). Given this success, it is crucial to transfer the WSSS framework to driving scenes, which are a significant scenario in semantic segmentation. Obtaining pixel-level labels in driving scenes is prohibitively expensive, making label-efficient training methods imperative in this context. For instance, Cityscapes required 1.5 hours per image (Cordts et al. 2016), while PASCAL VOC required 239.7 seconds per image (Bearman et al. 2016).

However, when applied to driving scene datasets like Cityscapes, WSSS models exhibit significant performance degradation. Akiva and Dana attributed this issue to the specific characteristics of the dataset, such as small object size, a high number of objects in each image, and limited diversity in object appearance (Akiva and Dana 2023). However, they only reported this tendency implicitly. In our study, we explicitly compare the driving scene datasets to the existing benchmark datasets (i.e., PASCAL VOC and MS COCO). As a result, we find that the driving scenes datasets lack negative samples and exhibit a remarkably high level of co-occurrence among classes. This poses a challenge in identifying individual objects through image classification, which hinders the effectiveness of common WSSS baselines, such as class activation mapping (CAM).

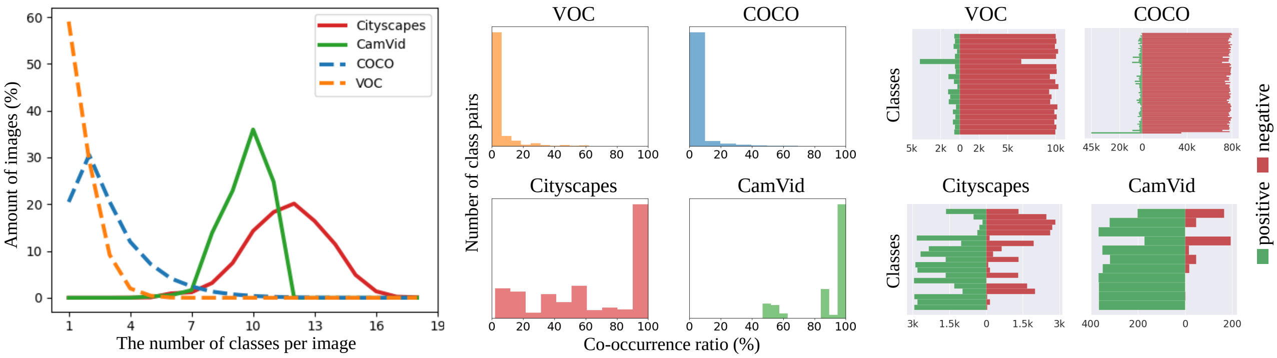
Recently, Contrastive Language Image Pre-training (CLIP), a model trained on a massive set of 400 million image-text pairs, has remarkably performed in open vocabulary classification. Using the open vocabulary classification ability of CLIP, we can avoid the characteristic of the dataset degrading the classifier's performance. As a result, as opposed to CAM, the seed mask generated by CLIP better distinguishes the object regions on the driving dataset like Cityscapes. Despite the potential, it often fails to identify small objects and produces noisy masks (Fig. 3 (a)).

In this paper, we present a novel WSSS framework for driving scene datasets, to address the above two challenges inherent in CLIP. Considering CLIP as a baseline mask generator, we propose (1) global-local view training to handle small-sized objects and (2) *Consistency-Aware Region Balancing (CARB)* to mitigate the negative effects of noisy pseudo-masks. Firstly, we found a unique property of CLIP: it offers considerably different pseudo-masks across input scales. Based on this observation, we use both a local view (i.e., a small-sized patch) and a global view (i.e., an original-sized image) during model training for accurately detecting small but critical objects in driving scenes (e.g., *traffic light*).

Next, we propose CARB, which suppresses the erroneous region of pseudo-masks to train the segmentation model. Specifically, we divide the noisy pseudo-mask into consistent and inconsistent regions according to prediction consistency between the segmentation model and CLIP. The incon-

[*] These authors contributed equally.

[†] Hyunjung Shim is a corresponding author.



(a) Amount of images per number of classes (b) Histogram of co-occurrence ratio (c) Number of positive/negative sample

Figure 1: Dataset statistics for Cityscapes, CamVid, MS COCO, and PASCAL VOC. (a) Counting the number of images given by the number of classes in a single image. (b) Histogram of co-occurrence ratio between classes. (c) The number of positive and negative images for each class.

sistent region contains more false predictions than the consistent region, resulting in a higher loss. This discrepancy in the magnitude of loss values leads to a negative impact on the overall training process. To mitigate this, we propose a strategy to balance the losses from both regions, thereby suppressing the high loss of the inconsistent region.

In summary, we examine the distinct characteristics of driving scenes over the commonly evaluated datasets and highlight the issue of ineffective CAM-based approaches in these scenes. We introduce a new WSSS framework utilizing pseudo-masks generated from CLIP, suggesting global-local view training to handle small-sized objects and CARB to mitigate the negative effects of noisy pseudo-masks.

We demonstrate that the proposed method achieved 51.8% mIoU on the Cityscapes dataset, showcasing the potential as a strong WSSS baseline for driving scenes. The effectiveness of the proposed method was confirmed on CamVid representing a small-scale dataset, and on WildDash2 containing more visually challenging scenes (e.g., diverse weather and lighting conditions). Owing to its advantage in performance and simple training, our method can serve as a valuable baseline for future research to address the challenges of WSSS in the driving scenes.

Related Work

Earlier Works in WSSS. Most WSSS techniques using image-level labels utilize CAM (Zhou et al. 2016). Due to its sparse coverage, recent studies have focused on expanding discriminative regions (Jiang et al. 2019; Wei et al. 2018; Choe and Shim 2019). In terms of using global-local view, L2G (Jiang et al. 2022) strengthened classifier learning by using local attention. This method was also used to widen the discriminative region. Recently, some approaches have attempted to solve co-occurrence problem by incorporating additional information (Lee et al. 2021, 2022; Xie et al. 2022). However, most of existing methods were only evaluated on PASCAL VOC (Everingham et al. 2015) or MS COCO (Lin et al. 2014). Akiva and Dana. 2023 conducted evaluations on more complex datasets like Cityscapes (Cordts et al. 2016) and ADE20k (Zhou et al. 2019), but only revealed the performance limitations of existing WSSS studies. Wang, Ma, and You. 2020 introduced a clustering-based approach in driving scene datasets, while they only achieved a marginal improvement. Unlike most WSSS studies, we analyze distinct characteristics of the driving scene datasets compared to existing benchmark datasets and suggest a new direction for WSSS in driving scene scenarios.

CLIP-based Segmentation. CLIP (Radford et al. 2021) is a framework trained on a large amount of image-text pairs. Several attempts have been made to utilize the characteristics of the multimodal embedding space in the field of segmentation (Ding et al. 2022; Wang et al. 2022b). In WSSS, CLIMS (Xie et al. 2022) employed the embedding spaces by optimizing the mask based on the similarity between masked image and text embedding. CLIP-ES (Lin et al. 2023) generates seed masks in Grad-CAM manner (Selvaraju et al. 2017). Then, it refines masks with class-wise attention-based affinity of the CLIP image encoder.

Existing studies (Li et al. 2022a; Xu et al. 2022) have also shown a significant improvement in zero-shot and few-shot segmentation by leveraging CLIP's zero-shot ability. Recently, MaskCLIP (Zhou, Loy, and Dai 2022) has been proposed to create dense masks from CLIP with category information at the dataset level rather than the image level. We employ MaskCLIP to extract dense labels from images and further propose a training strategy for handling the noise present in its pseudo-masks.

Uncertainty Estimation. Estimating the uncertainty (Kendall and Gal 2017) has been discussed in deep learning since deep neural networks learn to approximate probabilistic models. Focusing on semantic segmentation, the Feng et al. utilizes an ensemble of models that are initialized differently to separate the uncertainty region. In a similar vein, several methods (Oh, Kim, and Ham 2021; Zhang et al. 2020) utilized a combination of confidence thresholding and consistency between the CRF-refined mask and the original mask to define a reliable

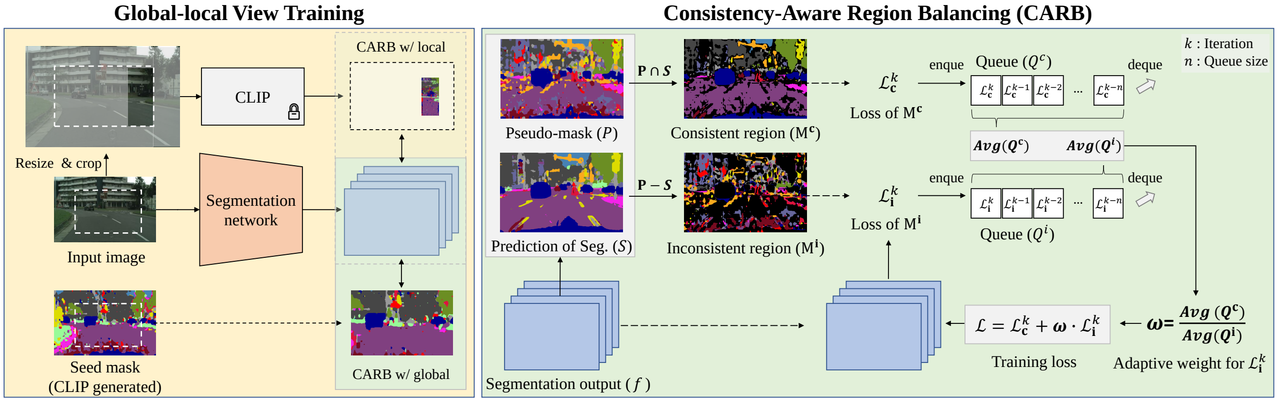

Global-local View Training
Consistency-Aware Region Balancing (CARB)
CLIP
Resize & crop
Input image
Segmentation network
CARB w/ local
CARB w/ global
Seed mask (CLIP generated)
Pseudo-mask (P)
Prediction of Seg. (S)
$P \cap S$
$P - S$
Consistent region (M^c)
Inconsistent region (M^i)
$\mathcal{L}_c^k$
Loss of M^c
$\mathcal{L}_i^k$
Loss of M^i
enque
deque
Queue (Q^c)
Queue (Q^i)
$\mathcal{L}_c^k$ $\mathcal{L}_c^{k-1}$ $\mathcal{L}_c^{k-2}$ ... $\mathcal{L}_c^{k-n}$
$\mathcal{L}_i^k$ $\mathcal{L}_i^{k-1}$ $\mathcal{L}_i^{k-2}$ ... $\mathcal{L}_i^{k-n}$
$Avg(Q^c)$ $Avg(Q^i)$
k : Iteration
n : Queue size
$\mathcal{L} = \mathcal{L}_c^k + \omega \cdot \mathcal{L}_i^k$
Training loss
$\omega = \frac{Avg(Q^c)}{Avg(Q^i)}$
Adaptive weight for $\mathcal{L}_i^k$
Segmentation output (f)

Figure 2: Overall framework of proposed method. (Global-local View Training) CLIP gives different pseudo masks for cropping and resizing. (CARB) The pseudo-mask is divided into the consistent / inconsistent regions and the high loss of inconsistent regions is suppressed via adaptive region balancing.

region. ST++ (Yang et al. 2022) identifies reliable images by utilizing the results of previous checkpoints. Recently, several methods suggest pixel-level entropy to measure the pixel-level uncertainty (Hu, Sclaroff, and Saenko 2020; Huynh et al. 2022; Li et al. 2022b; Wang et al. 2022a).

Statistics of Datasets

In this section, to identify the cause of the poor performance of existing WSSS methods on driving scenes, we compare the characteristics of two types of datasets: standard benchmark datasets (e.g., PASCAL VOC and MS COCO) and driving scene datasets (e.g., Cityscapes and CamVid). Specifically, we investigate the histograms of 1) the number of classes per image, 2) the co-occurrence ratio between classes, and 3) the number of positive/negative samples per class (Fig. 1).

The distinct difference between the two types of datasets is the number of classes in a single image. Although existing benchmark datasets have only one or two classes in most images, driving scene datasets typically contain eight or more classes in a single image, as in Fig. 1 (a). Next, we calculate the frequency ratio of co-occurrence between every pair of classes and plot a histogram of those ratios in Fig. 1 (b). Even worse, these classes in driving scenes often appear together, causing contextual bias. It is clearly different from PASCAL VOC and MS COCO.

Another critical point is the scarcity of negative samples in driving scene datasets. Negative samples are important learning signals for training the image classifier. As shown in Fig. 1 (c), existing datasets have a sufficient number of negative samples, but some classes in driving scene datasets have extremely few or zero negative samples. Most seriously, *road* and *car* in CamVid always appear in all training images and cannot be distinguished using only image-level labels. Understanding these characteristics is essential for developing productive approaches for WSSS using image-level labels in driving scene applications.

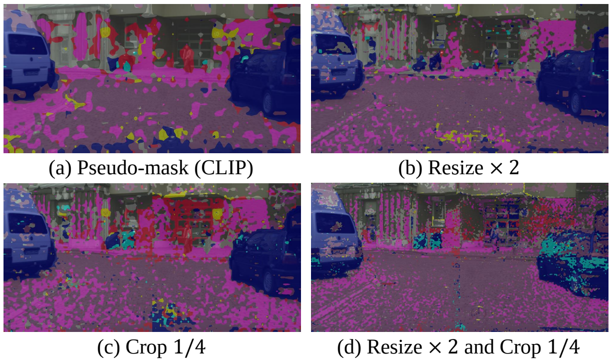

(a) Pseudo-mask (CLIP)
(b) Resize × 2
(c) Crop 1/4
(d) Resize × 2 and Crop 1/4

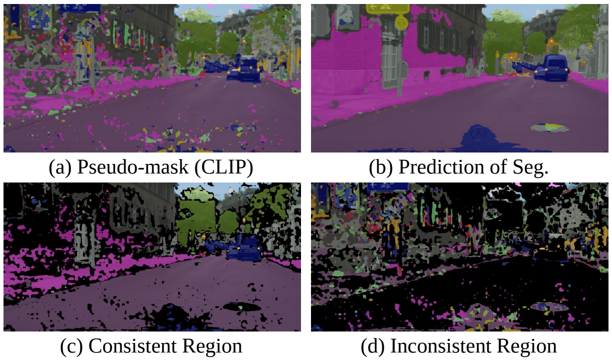
Figure 3: Pseudo-masks after resizing and cropping. (a) The original CLIP mask. (b) CLIP mask with resize ratio 2. (c) The concatenation of quarter-size cropped CLIP masks (d) The mask applying both operations. For visual clarity, we modify color palette of motorcycle to *cyan* in this figure.

Method

Global-local View Training

Owing to the specific nature of driving scenes, certain classes such as *roads* are consistently large, and others like *traffic light* remain small in size. Since the driving scenes capture road environments with a wide range of depth in each image, object sizes vary significantly with distance even within the same class, such as *car*. In Fig. 3 (a), we observe that the pseudo-masks generated by the CLIP model exhibit notably high quality for relatively large objects but poor quality for small objects. We conjecture that this performance degradation may occur from the training mechanism of CLIP (i.e., it mainly concentrate on salient objects corresponding to text prompt rather than small objects).

Building upon this observation, we manipulate the relative object size within the input by adjusting the *image scale* and *field-of-view* (FOV). We then analyze the resultant changes in the pseudo-mask obtained from CLIP. In Fig. 3 (b), when the input is scaled to twice its original size,



(a) Pseudo-mask (CLIP)

(b) Prediction of Seg.

(c) Consistent Region

(d) Inconsistent Region

Figure 4: The characteristics of two different masks. (a) The mask from CLIP contains small and blob-like noisy regions. (b) The output mask from the segmentation network is more systematic. We identify reliable regions (c) based on prediction consistency between (a) and (b).

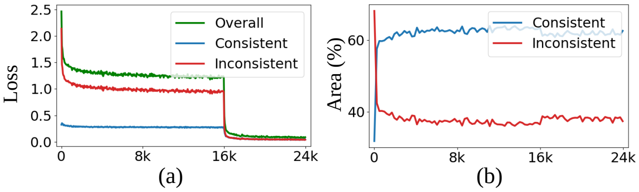


Figure 5: Changes in (a) loss and (b) area of consistent/inconsistent regions during training. Adaptive region balancing is applied from 16K iteration, affecting the training dynamics.

the pseudo-mask exhibit more accurate and fine-grained results along the object boundaries. Additionally, reducing the FOV by half (i.e., the network only observes one-quarter of the input at a time) leads to noticeable changes in the pseudo-mask, particularly pronounced for smaller objects such as *motorcycle* (c.f., Fig 3 (c)). This simple case study unveils distinctive characteristics associated with each adjustment.

Summarizing, we observed distinct responses of CLIP to cropping (changing FOV) and resizing (changing scales). (1) Resizing improves the localization at fine-grained areas such as edges. (2) Cropping enhances the classification of small objects. Capitalizing on the distinctive effects of cropping and resizing, we synergistically incorporate both augmentations into our approach. By jointly leveraging these functions, we enhance pseudo-mask performance, particularly for small objects.

Inspired by this observation, we have developed a new method called *Local View Sampling*. This technique leverages the conventional augmented input known as the global view, commonly used for training segmentation networks. We extract a patch of a specific size (typically small) from an arbitrary position inside the global view. Then, the patch is resized randomly before passing it through CLIP. We obtain the local pseudo-mask by calculating the similarity of local features from the image encoder and text embedding as follows:

$$\mathbf{M}^{\mathbf{l}} = \arg\max(\frac{\boldsymbol{F}^{\mathbf{l}} \cdot \boldsymbol{t}}{\|\boldsymbol{F}^{\mathbf{l}}\| \cdot \|\boldsymbol{t}\|}), \tag{1}$$

where $\boldsymbol{F}^{\mathbf{l}}$ is the feature from CLIP with a local view image, $\boldsymbol{t}$ is text embedding of CLIP. The local view only contains semantic information in a small, confined, area, so it can fully exploit locality from CLIP. This empowers the pseudo-mask of the local view to better focus on small objects. We leverage the masks of both views to train the segmentation model. The loss for the global-local view training is computed by cross-entropy loss for each region as follows:

$$\mathcal{L}_{\mathbf{l}} = -\frac{1}{|\mathbf{M}^{\mathbf{l}}|} \sum_{i,j \in local} y^{\mathbf{l}} \log f_{i,j}, \tag{2}$$

$$\mathcal{L}_{\mathbf{g}} = -\frac{1}{|\mathbf{M}^{\mathbf{g}}|} \sum_{i,j} y^{\mathbf{g}} \log f_{i,j}, \tag{3}$$

where $y^{\mathbf{l}}$ and $y^{\mathbf{g}}$ are one-hot labels of $\mathbf{M}^{\mathbf{l}}$ and $\mathbf{M}^{\mathbf{g}}$, respectively. $\mathbf{M}^{\mathbf{g}}$ and $\mathbf{M}^{\mathbf{l}}$ are the global pseudo-mask and the local pseudo-mask, respectively. $f \in \mathbb{R}^{C \times H \times W}$ is the probability of segmentation network. The total loss for the global-local view training is $\mathcal{L} = \mathcal{L}_{\mathbf{g}} + \mathcal{L}_{\mathbf{l}}$.

Consistency-aware Region Balancing

We identify noisy regions in the pseudo-mask created by CLIP. Fig. 4 showcases an example of a pseudo-mask containing small and blob-like noisy regions, randomly scattered on the image.

Conversely, training a segmentation network with pseudo-masks removes the randomly scattered noise of CLIP's pseudo-mask in the output, resulting in systematic predictions. (e.g., *road* in Fig. 4 (b)) However, the segmentation prediction has misclassified pixels that were originally correct in the pseudo-mask. In particular, we observe that the trained segmentation model produces an indistinct boundary of the object even worse than the pseudo-mask generated by CLIP (e.g., *sidewalk* in Fig. 4 (b)).

Owing to the unique properties of the trained segmentation model and CLIP, we leverage both models to benefit from their respective advantages. However, since the prediction of the segmentation model is already incorporated in the model, directly computing the loss using the segmentation prediction does not provide new evidence for training. To address this, we indirectly employ the segmentation prediction to distinguish the pixels of the pseudo-mask from CLIP. Specifically, utilizing prediction consistency, we regard the pixel as reliable if they are consistent and noisy if the predictions from the two models are inconsistent:

$$\mathbf{M}^{\mathbf{c}} = \{P_{i,j} | P_{i,j} = S_{i,j}\}, \tag{4}$$

$$\mathbf{M}^{\mathbf{i}} = \{P_{i,j} | P_{i,j} \neq S_{i,j}\}, \tag{5}$$

where $\mathbf{M}^{\mathbf{c}}$ and $\mathbf{M}^{\mathbf{i}}$ correspond to consistent and inconsistent regions, respectively. $P \in C^{H \times W}$ and $S \in C^{H \times W}$ are the pseudo-mask from CLIP and the prediction of the segmentation network, where C is a set of classes. Furthermore, we apply label filtering when generating S to prevent misprediction with non-existent classes in the image.

The consistent and inconsistent regions are recalculated in each iteration to update the segmentation model. As training progresses, we notice that the size of the consistent region changes, resulting in performance improvement of the segmentation model (Fig. 5 (b)).

To understand the effects of consistent and inconsistent regions, we separately calculate the cross-entropy loss of each side:

$$\mathcal{L}_{\mathbf{c}} = -\frac{1}{|\mathbf{M}^{\mathbf{c}}|}\sum_{i,j} y^{\mathbf{c}} \log f_{i,j}, \tag{6}$$

$$\mathcal{L}_{\mathbf{i}} = -\frac{1}{|\mathbf{M}^{\mathbf{i}}|}\sum_{i,j} y^{\mathbf{i}} \log f_{i,j}, \tag{7}$$

where $y^{\mathbf{c}}$ and $y^{\mathbf{i}}$ are one-hot labels of $\mathbf{M}^{\mathbf{c}}$ and $\mathbf{M}^{\mathbf{i}}$, respectively. $f \in \mathbb{R}^{C\times H\times W}$ is the probability of segmentation network. We observe that inconsistent regions have much higher loss values than consistent regions in Fig. 5 (a). If we treat the training loss equally across all regions, the network is overly influenced by the high loss produced from the inconsistent regions. Therefore, we suggest assigning different weights to the losses of consistent and inconsistent regions while taking into account the noise level of the data. It helps prevent the conventional cross-entropy loss from being vulnerable to noise in the training data.

To this end, we devise an adaptive region balancing method that dynamically adjusts the loss of the inconsistent region by monitoring the loss profiles in both the consistent and inconsistent regions during training. Specifically, we introduce two fixed-size queues which track the losses of the two regions, denoted as $\mathcal{L}_{\mathbf{c}}$ and $\mathcal{L}_{\mathbf{i}}$, respectively. We then compute the average loss from each queue. We use the ratio of two average losses as the weight for the cross-entropy loss of the inconsistent region, denoted as w, which is multiplied by the loss of the inconsistent region. The CARB training loss is $\mathcal{L} = \mathcal{L}_{\mathbf{c}} + w \cdot \mathcal{L}_{\mathbf{i}}$. This balancing ensures that the training is less influenced by the inconsistent region.

While one might consider that the loss from inconsistent regions can be simply neglected, our observations reveal that the inconsistent regions still possess useful learning signals. Notably, we observe that labels of highly correlated object classes (i.e., the classes sharing visual properties like *bus* and *car*) exist within the inconsistent region. Overlooking those pixels impedes the label imbalance problem. The Cityscapes dataset, for instance, includes classes like *rider* (a subset of *person*), *bus*, and *truck* (subsets of *car*) that are susceptible to such confusion. Considering these challenges, we present a region-balancing method designed to harness meaningful information even from inconsistent regions.

Overall Training. The proposed method consists of two stages. In the first stage, we warm up the baseline segmentation model with global and local views generated from CLIP masks. This step ensures that the segmentation network sufficiently learns the regular patterns of the target dataset. In the second stage, we refine the segmentation network utilizing CARB. We apply CARB for both global and local views.

Method	mIoU
Base	40.1
+ CARB	45.7
+ Local	45.1
+ Local + CARB	50.6
+ Dual	45.8
+ Dual + CARB	**52.1**

Table 1: Ablation study of the proposed modules. The accuracy (mIoU) is evaluated on the Cityscapes validation set. The best score is in **bold** throughout all experiments.

Experiments

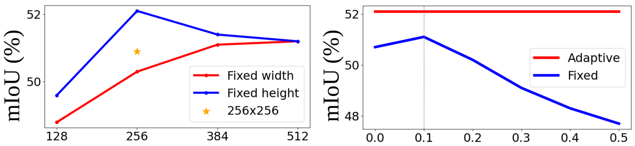
Experimental Setup

Dataset & Evaluation Metric. For performance evaluation, we utilized the well-known Cityscapes (Cordts et al. 2016), CamVid (Brostow, Fauqueur, and Cipolla 2009), and WildDash2 (Zendel et al. 2022), which are autonomous driving datasets. The Cityscapes dataset consists of 2,975 training, 500 validation, and 1,525 test images with fine annotation. It contains a total of 30 classes, and 19 classes are evaluated for public assessment while the rest are void. The CamVid dataset consists of 367 training, 101 validation, and 233 test images, containing a total of 32 classes. In our experiments, only 11 classes are evaluated by following the convention of previous research (Wang, Ma, and You 2020). The WildDash2 dataset consist of 3,618 training, 638 validation, and 812 test images, containing a total of 25 classes. In all experiments, we solely utilized image-level labels for training. The image-level labels are acquired from pixel-level labels of each dataset. Mean Intersection over Union (mIoU) was used as the evaluation criterion, a popular and standard metric for semantic segmentation.

Implementation Detail. We employed ViT-B/16 (Dosovitskiy et al. 2021) as the image encoder for the CLIP, and ResNet50 (He et al. 2016)-based DeepLab-ASPP (Chen et al. 2017) as the segmentation network. The last convolutional layer of ASPP was replaced with text embedding of CLIP. The segmentation network was initialized with an ImageNet pre-trained model provided by MMSeg (Contributors 2020). Considering class definition and object words connoting actual objects, we replaced the *vegetation* and *terrain* class names with *tree* and *grass*, respectively. Furthermore, we changed the *person* class to *pedestrian*, since it is a superset of the *rider*. For generating pseudo-masks from CLIP, we adopt MaskCLIP (Zhou, Loy, and Dai 2022).

Ablation Study

Effects of Each Module. We evaluate the effectiveness of each component of our method in Tab. 1. When we train the segmentation model with additional local view sampling (*Local*), it shows a remarkable improvement of 5.0%p. This implies that additional information from local patches through cropping and resizing provides rich learning signals. CARB alone contributed an impressive improvement of 5.6%p, indicating that adaptively re-weighting the loss



(a) Crop size (b) Weight

Figure 6: Segmentation results (mIoU) on Cityscapes validation set depending on (a) the crop size (b) the weight. We set the length of one side to 512 and varied the length of the other side between 128 and 512. Yellow star indicates the experiment using 256×256 patch.

Method	val	test
DeepLab-ASPP (Full supervision)	78.3	75.8
AffinityNet	8.2	-
SEAM	17.3	-
1-Stage	11.8	-
Wang, Ma, and You	24.2	24.9
CAM	33.0	32.2
AMN	17.5	17.8
CLIMS	18.1	18.0
CLIP-ES	35.4	35.0
Ours	**52.1**	**51.8**

Table 2: Segmentation results (mIoU) on Cityscapes.

according to its reliability plays a critical role in learning with noisy pseudo-masks. By combining local view sampling and CARB ($Local + CARB$), a substantial improvement of 10.5%p was achieved. Instead of $Local$, we added a slight modification named $Dual$, by re-creating the mask of global views depending on the augmentation using CLIP for each iteration. This modification yielded a 0.7%p gain over $Local$. Interestingly, our $Dual + CARB$ method shows a 1.5%p improvement compared to $Local + CARB$, indicating synergy between various-sized mask creations and our noise-handling strategy.

Effects of the Crop Size and Resize Ratio. In our empirical investigation, it was consistently observed that vertically long rectangular patches exhibited superior performance in terms of crop sizes compared to patches of other sizes. This finding is supported by Fig. 6. We conjecture this tendency in the driving scene dataset comes from vertically long structures such as *pole* and *traffic light*. Also, the 512×512 patches for local views are more effective than the 256×256. These experiments suggest that, while the local view represents a smaller portion of the overall scene, excessively small sizes may not benefit from the attention layers equipped in CLIP.

We evaluated our local view sampling under variable resize ratios. Under the fixed ratio from 0.5 to 2.0, we observe the best performance of 52.1% at the ratio of 1.0 while a significant drop with other ratios. However, when focusing on classwise performances under various resize ratios, we confirmed that a large resize rate benefits the performance of small classes, such as *traffic light* and *rider*. Meanwhile, the performance of the large classes, such as *sidewalk* and *wall*, is decreased. Due to the performance trade-off across different classes, we set a random value between 1.0 and 2.0 as the resize ratio. Our choice leads to the overall best performance in both small and large classes.

Effects of Adaptive Region Balancing. We compare our adaptive region balancing strategy with a fixed weighting strategy, where the loss weight for the inconsistent region (w) is set to a specific value (c.f., Fig. 6 (b)). When changing the fixed weight w gradually from 0 to 0.5, we observe the best score at the weight of 0.1 and a significant drop with other values. Although the highest performance of the fixed weight strategy is similar to that of our method (51.06% for fixed strategy and 52.1% for our method), it requires a hyperparameter search for the optimal weight using the validation dataset. In contrast, our method does not require such a search, making it more suitable for WSSS scenario.

Quantitative Comparisons

Remarks on Comparisons. Existing WSSS methods focus on handling object-centric datasets such as PASCAL VOC 2012. Therefore, their methods are primarily designed to distinguish relatively simple object shapes with similar scales, which is still a valuable research direction. Given this dataset mismatch, direct comparisons between our method and existing WSSS approaches might not be entirely fair, as they cater to distinct dataset characteristics. Nevertheless, by adapting established WSSS methods to driving scenes, we intend to show that the existing framework is ineffective for our application scenario.

Existing WSSS Methods. Existing methods can be partitioned into CAM-based methods (i.e., an image classifier for generating the pseudo-masks) and CLIP-based methods (i.e., CLIP for generating the pseudo-masks). Among them, we choose several representative methods such as (1) AffinityNet (Ahn and Kwak 2018), (2) SEAM (Wang et al. 2020), (3) 1-Stage (Araslanov and Roth 2020), (4) SEC (Kolesnikov and Lampert 2016), (5) Wang et al. (Wang, Ma, and You 2020), (6) CAM (Zhou et al. 2016), and (7) AMN (Lee, Kim, and Shim 2022). To compare with CLIP-based WSSS, we reproduce the driving scene results of (8) CLIMS (Xie et al. 2022) and (9) CLIP-ES (Lin et al. 2023), both of which use the same level of information, CLIP, as our method.

Cityscapes. Tab. 2 presents the performance of our proposed CARB compared to other methods in driving scenes. Specifically, our approach achieves 51.8% on the Cityscapes test set, which outperforms the Wang, Ma, and You by 26.9%p and previous CLIP-based WSSS technique by 16.8%p. Additionally, we observed that our method consistently performs better than CLIP-ES in every class. Fig. 7 showcases qualitative examples of segmentation results on the Cityscapes. Notably, CARB successfully eliminates misclassified *sidewalk* regions on the *sky* class (see the first and second rows). These results visually confirm that our method correctly captures each class and successfully reduces the prediction errors.

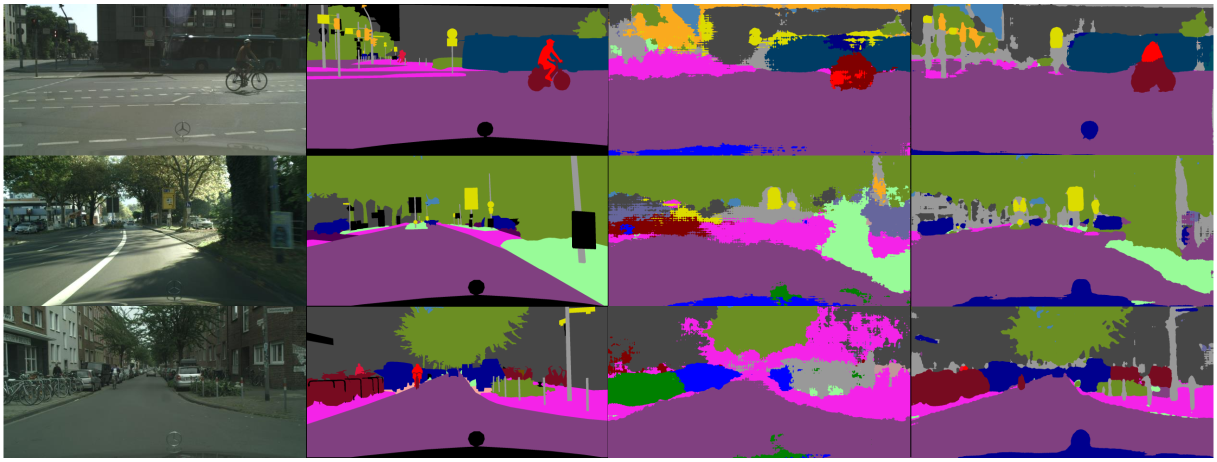


Figure 7: Qualitative results on Cityscapes validation set. (a) Input image, (b) Ground-truth, (c) CLIP-ES, and (d) Our method.

Method	val	test
DeepLab-ASPP (Full supervision)	81.6	74.9
SEC	-	2.5
AffinityNet	11.0	15.5
Wang, Ma, and You	23.5	30.4
CAM	9.6	6.6
AMN	10.7	7.6
CLIMS	2.7	4.3
CLIP-ES	41.7	39.6
Ours	**55.7**	**50.5**

Table 3: Segmentation results (mIoU) on CamVid.

CamVid. The CamVid dataset has a much smaller number of training images compared to Cityscapes, with only 367 images. Additionally, it is not possible to differentiate between the *car* and *road* classes using only image-level labels, as they appear in all images. However, our method can distinguish them by utilizing the pre-trained image-text information from the CLIP model. Tab. 3 shows that the performance of CAM-based methods (e.g., SEC, AffinityNet, Wang, Ma, and You, and CLIMS) is considerably low while our method achieves significantly higher performance. This demonstrates that our proposed approach can address the problem even when the scale of the dataset is small and has severe contextual bias.

WildDash2. Since the WildDash2 dataset possesses extremely high diversity, it is generally challenging even for the fully supervised model. A classifier-based WSSS method such as CLIMS performs poorly, 1% in mIoU, which is worse than a random guess. This poor performance is caused by difficulties in training the classifier due to class imbalance and complex class distribution. Since CLIP-ES and our method are built upon CLIP for generating the pseudo masks, both methods provide relatively reasonable performances. Our method achieves considerably high performance compared to CLIP-ES, with the performance gain primarily observed in small classes such as *billboard*, *rider*, *bicycle*, and *road marking*.

Method	Result
DeepLab-ASPP (Full supervision)	54.0
CAM	15.2
AMN	18.8
CLIMS	1.0
CLIP-ES	24.7
Ours	**32.2**

Table 4: Segmentation results (mIoU) on WildDash2 *val* set.

Conclusion

This paper addressed the limitations of conventional CAM-based, weakly-supervised semantic segmentation (WSSS) methods when handling the driving scene datasets. To break the performance bottleneck of the CAM-based methods, we utilized CLIP as the pseudo-mask generator. Then, we proposed global-local view training, which exploits the characteristics of CLIP generating diverse masks depending on the relative object sizes. We also propose a novel training strategy, namely consistency-aware region balancing (CARB). It distinguishes between reliable and noisy regions utilizing prediction consistency and then suppresses the latter regions during training. By incorporating these two components, our method successfully (1) learns to segment small objects and (2) heavily relies on reliable regions while effectively handling challenging objects from noisy regions. Extensive experiments demonstrate that each component of our method contributes to achieving new state-of-the-art performances on the Cityscapes, CamVid, and WildDash2 datasets in WSSS. Our study introduces a new approach addressing the challenges posed by driving datasets and suggests a promising direction for future research in WSSS.

Acknowledgments

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the MSIP (NRF-2022R1A2C3011154, RS-2023-00219019, RS-2023-00240135) and MOE (NRF-2022R1A6A3A13073319), the IITP grant funded by the Korea government(MSIT) (No. 2019-0-00075, Artificial Intelligence Graduate School Program(KAIST)), KEIT grant funded by the Korea government(MOTIE) (No. 2022-0-00680, 2022-0-01045, 2021-0-02068, Artificial Intelligence Innovation Hub) and Samsung Electronics Co., Ltd (IO230508-06190-01).

References

Ahn, J.; and Kwak, S. 2018. 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*.

Akiva, P.; and Dana, K. 2023. Single Stage Weakly Supervised Semantic Segmentation of Complex Scenes. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*.

Araslanov, N.; and Roth, S. 2020. Single-Stage Semantic Segmentation from Image Labels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Bearman, A.; Russakovsky, O.; Ferrari, V.; and Fei-Fei, L. 2016. What's the point: Semantic segmentation with point supervision. In *Proceedings of European Conference on Computer Vision (ECCV)*. Springer.

Brostow, G. J.; Fauqueur, J.; and Cipolla, R. 2009. Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters*.

Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. L. 2017. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Choe, J.; and Shim, H. 2019. Attention-based dropout layer for weakly supervised object localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Contributors, M. 2020. MMSegmentation: OpenMMLab Semantic Segmentation Toolbox and Benchmark. https://github.com/open-mmlab/mmsegmentation.

Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; and Schiele, B. 2016. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Ding, J.; Xue, N.; Xia, G.-S.; and Dai, D. 2022. Decoupling zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

Everingham, M.; Eslami, S. A.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2015. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*.

Feng, Z.; Zhou, Q.; Gu, Q.; Tan, X.; Cheng, G.; Lu, X.; Shi, J.; and Ma, L. 2022. Dmt: Dynamic mutual training for semi-supervised learning. *Pattern Recognition*.

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

Hu, P.; Sclaroff, S.; and Saenko, K. 2020. Uncertainty-Aware Learning for Zero-Shot Semantic Segmentation. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*. Curran Associates, Inc.

Huynh, D.; Kuen, J.; Lin, Z.; Gu, J.; and Elhamifar, E. 2022. Open-Vocabulary Instance Segmentation via Robust Cross-Modal Pseudo-Labeling. *IEEE Conference on Computer Vision and Pattern Recognition*.

Jiang, P.-T.; Hou, Q.; Cao, Y.; Cheng, M.-M.; Wei, Y.; and Xiong, H.-K. 2019. Integral object mining via online attention accumulation. In *Proceedings of the IEEE International Conference on Computer Vision*.

Jiang, P.-T.; Yang, Y.; Hou, Q.; and Wei, Y. 2022. L2G: A Simple Local-to-Global Knowledge Transfer Framework for Weakly Supervised Semantic Segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

Kendall, A.; and Gal, Y. 2017. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? In Guyon, I.; von Luxburg, U.; Bengio, S.; Wallach, H. M.; Fergus, R.; Vishwanathan, S. V. N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*.

Kolesnikov, A.; and Lampert, C. H. 2016. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *Proceedings of the European Conference on Computer Vision*. Springer.

Lee, J.; Oh, S. J.; Yun, S.; Choe, J.; Kim, E.; and Yoon, S. 2022. Weakly supervised semantic segmentation using out-of-distribution data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Lee, M.; Kim, D.; and Shim, H. 2022. Threshold matters in WSSS: manipulating the activation for the robust and accurate segmentation model against thresholds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Lee, S.; Lee, M.; Lee, J.; and Shim, H. 2021. 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*.

Li, B.; Weinberger, K. Q.; Belongie, S.; Koltun, V.; and Ranftl, R. 2022a. Language-driven Semantic Segmentation. In *International Conference on Learning Representations*.

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

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision*. Springer.

Lin, Y.; Chen, M.; Wang, W.; Wu, B.; Li, K.; Lin, B.; Liu, H.; and He, X. 2023. CLIP Is Also an Efficient Segmenter: A Text-Driven Approach for Weakly Supervised Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15305–15314.

Oh, Y.; Kim, B.; and Ham, B. 2021. 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 (CVPR)*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, 618–626.

Wang, X.; Ma, H.; and You, S. 2020. Deep clustering for weakly-supervised semantic segmentation in autonomous driving scenes. *Neurocomputing*, 381.

Wang, Y.; Wang, H.; Shen, Y.; Fei, J.; Li, W.; Jin, G.; Wu, L.; Zhao, R.; and Le, X. 2022a. Semi-Supervised Semantic Segmentation Using Unreliable Pseudo Labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wang, Y.; Zhang, J.; Kan, M.; Shan, S.; and Chen, X. 2020. Self-supervised Equivariant Attention Mechanism for Weakly Supervised Semantic Segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Wang, Z.; Lu, Y.; Li, Q.; Tao, X.; Guo, Y.; Gong, M.; and Liu, T. 2022b. CRIS: CLIP-Driven Referring Image Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wei, Y.; Xiao, H.; Shi, H.; Jie, Z.; Feng, J.; and Huang, T. S. 2018. Revisiting dilated convolution: A simple approach for weakly-and semi-supervised semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Xie, J.; Hou, X.; Ye, K.; and Shen, L. 2022. CLIMS: Cross Language Image Matching for Weakly Supervised Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Xu, M.; Zhang, Z.; Wei, F.; Lin, Y.; Cao, Y.; Hu, H.; and Bai, X. 2022. A Simple Baseline for Open Vocabulary Semantic Segmentation with Pre-trained Vision-language Model. *Proceedings of the IEEE/CVF European Conference on Computer Vision (ECCV)*.

Yang, L.; Zhuo, W.; Qi, L.; Shi, Y.; and Gao, Y. 2022. St++: Make self-training work better for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Yoon, S.-H.; Kweon, H.; Cho, J.; Kim, S.; and Yoon, K.-J. 2022. Adversarial Erasing Framework via Triplet with Gated Pyramid Pooling Layer for Weakly Supervised Semantic Segmentation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIX*. Springer.

Zendel, O.; Schörghuber, M.; Rainer, B.; Murschitz, M.; and Beleznai, C. 2022. Unifying Panoptic Segmentation for Autonomous Driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 21351–21360.

Zhang, B.; Xiao, J.; Wei, Y.; Sun, M.; and Huang, K. 2020. Reliability Does Matter: An End-to-End Weakly Supervised Semantic Segmentation Approach. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*. AAAI Press.

Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; and Torralba, A. 2016. Learning Deep Features for Discriminative Localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Zhou, B.; Zhao, H.; Puig, X.; Xiao, T.; Fidler, S.; Barriuso, A.; and Torralba, A. 2019. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*.

Zhou, C.; Loy, C. C.; and Dai, B. 2022. Extract free dense labels from clip. In *Proceedings of the European Conference on Computer Vision*. Springer.