

Precision matters: Precision-aware ensemble for weakly supervised semantic segmentation

Junsung Park, Hyunjung Shim

¹Kim Jaechul Graduate School of AI
Korea Advanced Institute of Science and Technology
85, Hoegi-ro, Dongdaemun-gu, Seoul, Korea
{jshackist, kateshim}@kaist.ac.kr

Abstract

Weakly Supervised Semantic Segmentation (WSS) employs weak supervision, such as image-level labels, to train the segmentation model. Despite the impressive achievement in recent WSS methods, we identify that introducing weak labels with high mean Intersection of Union (mIoU) does not guarantee high segmentation performance. Existing studies have emphasized the importance of prioritizing precision and reducing noise to improve overall performance. In the same vein, we propose ORANDNet, an advanced ensemble approach tailored for WSS. ORANDNet combines Class Activation Maps (CAMs) from two different classifiers to increase the precision of pseudo-masks (PMs). To further mitigate small noise in the PMs, we incorporate curriculum learning. This involves training the segmentation model initially with pairs of smaller-sized images and corresponding PMs, gradually transitioning to the original-sized pairs. By combining the original CAMs of ResNet-50 and ViT, we significantly improve the segmentation performance over the single-best model and the naive ensemble model, respectively. We further extend our ensemble method to CAMs from AMN (ResNet-like) and MCTformer (ViT-like) models, achieving performance benefits in advanced WSS models. It highlights the potential of our ORANDNet as a final add-on module for WSS models.

Introduction

Weakly supervised semantic segmentation (WSS) addresses the high annotation cost of fully supervised semantic segmentation (FSS) by utilizing weak supervision such as scribble (Lin et al. 2016), bounding box (Khoreva et al. 2017), point (Chen et al. 2021) or image-level labels (Pinheiro and Collobert 2015). In this paper, we focus on WSS methods using image-level labels since image-level labels are commonly accessible in diverse benchmarks and widely employed in practical applications.

Existing WSS methods using image-level labels typically follow a two-stage process. Firstly, an image classifier is trained using image-level labels, and pseudo-masks are generated through Class Activation Mapping (CAM). Subsequently, a segmentation model is trained using pairs of training images and their corresponding CAM-generated

Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

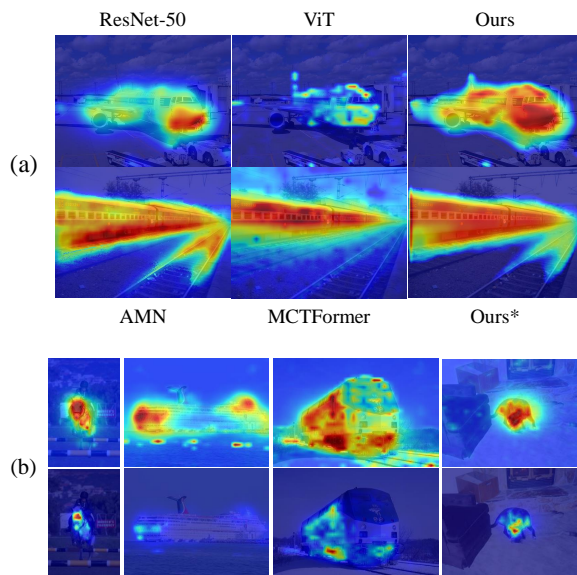


Figure 1: (a) Difference between ResNet-based CAM and ViT-based CAM. “Ours” and “Ours*” indicate the ORANDNet ensemble of ResNet-50-ViT and AMN-MCTformer, respectively. (b) Visualization of OR CAM (1st row) and AND CAM (2nd row).

pseudo-masks. Existing methods have identified three limitations of the CAM-generated pseudo-masks: 1) the sparse coverage by capturing the most discriminative parts, 2) inaccurate boundaries due to lack of pixel-level supervision, and 3) falsely capturing the co-occurring objects. The sparse coverage issue has been resolved by erasing the most discriminative regions of training images (Wei et al. 2017; Kweon et al. 2021) or consistency regularization (Wang et al. 2020) or contrastive learning (Du et al. 2022). Predicting affinity between pixels (Ahn and Kwak 2018; Ahn, Cho, and Kwak 2019) resolved the inaccurate boundary problem. Employing the saliency map (Lee et al. 2021) and separating of object and context via data augmentation (Su et al. 2021) resolved the co-occurrence problem.

Despite the impressive achievement in recent WSS methods, we have identified that solely improving pseudo-

masks in terms of mean Intersection of Union (mIoU) in the first stage does not always result in the high performance of a final model. This counter-intuitive result is due to the fact that mIoU alone may not be the best metric to evaluate the pseudo-mask quality. Recent studies on learning with noisy labels highlight the importance of precision in the early training stage via memorization effects (Rong et al. 2023; Liu et al. 2022). Since the pseudo-mask is inevitably noisy, the importance of precision can be applied to the WSSS framework. However, the mIoU is affected by both recall and precision, and thus, the pseudo-masks with high mIoU can be achieved by increasing the recall. Since the recall does not penalize false positive predictions, the high recall inevitably allows noises in pseudo-masks, degrading the segmentation training in the second stage. This explains the discrepancy between the pseudo-masks’ performance and segmentation performance, as also pointed out by (Rong et al. 2023).

To address the above-mentioned issue, we introduce ORANDNet, an ensemble method for WSSS that focuses on achieving high precision in pseudo-masks. Our approach leverages CAMs of two distinct classifiers to improve the precision and handle label noise within pseudo-masks. We design ORANDNet to take a probabilistic OR of two CAMs as an input and to produce a probabilistic AND of two CAMs as an output. This simple-yet-effective approach leads to higher mIoU as well as precision in the pseudo-mask compared to using any single model or the ensemble average of two CAMs. To fully enjoy the ensemble effects of ORANDNet, we choose two distinct classifiers having clearly different characteristics, such as ResNet (He et al. 2016) and ViT (Dosovitskiy et al. 2020). Furthermore, we confirm that incorporating more advanced WSSS methods such as AMN (Lee, Kim, and Shim 2022) and MCTformer (Xu et al. 2022) can yield even higher performance, showcasing the potential of our approach to be add-on module across different WSSS models. Additionally, We employ ‘scale scheduling’ early in training to downscale image-pseudo-mask pairs, removing small false activations in ORANDNet pseudo masks. This downsizing reduces noisy label impacts, thereby improving pseudo-mask precision, essential in early stages and advantageous in subsequent stages.

We believe that our work suggests the first task-specific ensemble method in WSSS. It can be an add-on module to various WSSS models, serving as a final step for future WSSS studies.

Related Works

Recent studies of WSSS often generate pseudo-labels using CAMs, derived from image-level labels, to form pixel-level pseudo-labels. Techniques like denseCRF (Krähenbühl and Koltun 2011), AffinityNet (Ahn and Kwak 2018), IRN (Ahn, Cho, and Kwak 2019), and PAMR (Araslanov and Roth 2020) are used to improve boundary performance. These pseudo-labels are then fed into a Fully Supervised Semantic Segmentation Network for final training.

AdvErasing (Wei et al. 2017) iteratively obscures and re-learns from a class activation map to uncover alternate discriminative regions, while CGN (Kweon et al. 2021)

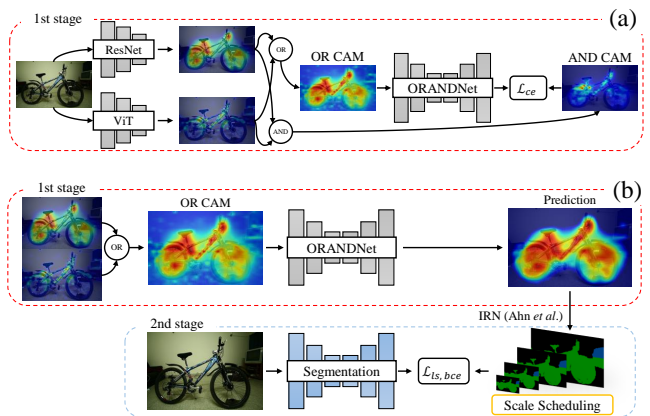


Figure 2: The overall pipeline of our method includes: (a) ORANDNet during training, and (b) ORANDNet in the test time with scale scheduling. \mathcal{L}_{ce} and $\mathcal{L}_{ls, bce}$ mean cross entropy and class-balanced cross entropy with label smoothing, respectively.

masks images with CAMs for iterative reclassification, promoting whole-object region consideration. PPC (Du et al. 2022) boosts WSSS performance by using contrastive learning and consistency regularization for pixel-level supervision. SEAM (Wang et al. 2020) enhances CAM through an equivariance constraint, ensuring pixel class consistency after affine transformations. CDA (Su et al. 2021) addresses WSSS co-occurrence by differentiating objects from their context. Fan et al. (Fan, Zhang, and Tan 2020) propose using multiple seeds with adaptively weighted pseudo-masks per pixel. ADELE (Liu et al. 2022) corrects labels by optimally weighting them during early, less noise-prone learning epochs. BECO (Rong et al. 2023) enhances noise resistance in segmentation networks using a Siamese network for uncertain regions and creating mixed image-label-boundary pairs. WeakTr (Zhu et al. 2023) introduces an adaptive attention fusion module for weighting across attention heads, deriving class activation maps. RCA (Zhou et al. 2022) uses contrastive learning to understand regional priors across datasets, surpassing individual image prior dependency.

Methods

This section provides a detailed procedure for ORANDNet. Unlike existing methods, ORANDNet takes two CAMs from two different classifiers and performs an effective ensemble aimed at boosting precision. The comprehensive process is exhibited in Figure 2. In the following, we first discuss the rationale behind ORANDNet, outline the overall process, and demonstrate ORANDNet with its training and architecture.

Motivation

To fully exploit the benefits of ensemble methods, it is important to use two distinct Class Activation Maps (CAMs), where each possesses distinguishable characteristics compared to another. With this objective in mind, we employ

ResNet and Vision Transformer (ViT) for their unique CAM characteristics: ResNet CAMs show concentrated activations at key object points, while ViT’s are denser but weaker, covering more of the object (see Figure 1 (a)). Both exhibit false activations beyond object boundaries. Activation map inconsistencies presented between ResNet and ViT extend to WSSS models with these backbones. Figure 1 (a) illustrates this, contrasting AMN (based on ResNet-38) and MCTformer (a ViT variant).

Our goal is to develop a task-specific ensemble method that enhances pseudo-mask precision. Leveraging the contrasting distributions in activation maps of ResNet and ViT, we identify consistent activations as reliable pseudo-labels. Thus, a probabilistic AND operation on these CAMs results in a map with reduced false activations, exemplified in Figure 1 (b) using the ResNet-ViT pair.

While consistent activations contribute to improved precision in pseudo masks, achieving sufficient mIoU scores also requires reasonable recall. Training a segmentation model with high precision but low recall is problematic due to data scarcity. Consequently, the result of the AND operation, as shown in Figure 1 (b), is not conducive to effective segmentation training. To achieve high mIoU and precision simultaneously, we propose ORANDNet.

Proposed method

Following the convention of existing WSSS methods, the pipeline of the proposed method consists of two stages. The first stage involves obtaining the pseudo-mask. The second stage uses pairs of images and pseudo-masks to train the semantic segmentation network. The detailed procedures are illustrated in Figure 2.

ORANDNet As mentioned in the previous section, our goal is to enhance pseudo-mask precision and reduce noise, without compromising on the quality of the pseudo-masks. First, we execute pixel-wise probabilistic OR and AND operations on the two CAMs from distinct models. OR and AND operations are illustrated as below respectively:

$$\frac{c_1 + c_2 - c_1 \odot c_2}{\max(c_1 + c_2 - c_1 \odot c_2)}, \frac{c_1 \odot c_2}{\max(c_1 \odot c_2)}. \quad (1)$$

where \odot is a pixel-wise multiplication, and c_1 and c_2 indicate two distinctive CAMs from different models. By applying Equation 1, we can derive the OR CAM and AND CAM.

As previously discussed, training a segmentation model using pseudo-masks generated from the AND CAM is hindered by data scarcity, cause by low recall of pseudo-mask. To address this challenge and produce high-precision pseudo-masks, we propose a simple yet effective approach called ORANDNet. In this method, we design an FCN (Simonyan and Zisserman 2014) to predict the AND CAM from the OR CAM, which is derived from two distinct activation maps. ORANDNet improves pseudo-mask precision and mIoU by refining dense, wide-coverage activation maps into consistent regions while retaining sufficient recall, with Figure 1 (a) demonstrating these enhancements.

After achieving OR CAM and AND CAM through Eqn 1, ORANDNet is then trained with the AND CAM as the

Method	mIoU	Precision	Recall
ResNet-50	48.3	66.9	64.8
ViT	53.4	67.7	72.4
Naive ensemble	49.0	61.5	72.2
Ours	54.3 (+5.0)	71.0 (+9.5)	70.9 (-1.3)
Ours w/IRN	70.9	82.9	82.6
AMN	62.8	74.0	80.5
MCTformer	62.1	74.7	78.8
Ours*	64.3 (+1.5, +2.2)	78.9 (+4.9, +4.2)	77.9 (-2.6, -0.9)
Ours* w/IRN	74.3	85.5	84.3

Table 1: The results of pseudo-masks for ORANDNet and naive ensembles, and single models. Classifier training was done on the PASCAL VOC 2012 *augmented train set*, evaluated on the *train set*. Increments were based on the naive ensemble for ‘Ours’, AMN/MCTFormer for ‘Ours*’.

pseudo-mask and the OR CAM as the input. In this way, ORANDNet learns to reduce noise in the pseudo-masks and enhance precision while preserving the object coverage of activations. During the testing phase, the OR CAM is utilized as the input for ORANDNet. Subsequently, we achieve the pseudo-mask by applying the IRN to the predictions of ORANDNet.

Scale scheduling To reduce noise in pseudo-masks, we adopt a curriculum learning strategy in second stage, progressively downsampling training data (image and pseudo-mask pairs) by factors of 8, 4, and 2 for the initial 2, 4, and 6 epochs, respectively. This method effectively discards smaller noise early in training by introducing variably downsampled image-pseudo mask pairs at different epochs.

Experiments

Datasets and evaluation metrics. The PASCAL VOC 2012 dataset (Everingham et al. 2015) was used for these experiments. The training set of PASCAL VOC 2012 comprises 10,582 images, the validation set contains 1,449 images, and the test set contains 1,456 images. We evaluated pseudo-mask generation using mean Intersection over Union (mIoU) and precision metrics, while segmentation results were assessed solely based on mIoU.

Implementation details. When training ResNet-50 with classification, we used a learning rate of 0.01, a polynomial scheduler, and a weight decay of $5e-4$. For ViT, the learning rate was set to 0.001, with a cosine annealing scheduler and a weight decay of $1e-5$. ORANDNet used a learning rate of 0.001, a polynomial scheduler, and no weight decay. ORANDNet utilizes the VGG16 backbone without batch normalization (Simonyan and Zisserman 2014). For the final segmentation model, the Deeplab-v1 model (Chen et al. 2014) with a ResNet-38 backbone (Wu, Shen, and Van Den Hengel 2019) was selected.

Ablation study

ORANDNet with basic classification models. Firstly, we investigate the effect of ORANDNet using two classifiers, ResNet-50 and ViT-base-patch16.

Table 1 summarizes the mIoU, precision, and recall of ORANDNet compared to those of CAM with ResNet and ViT, respectively. The results clearly show that ORANDNet improves both mIoU and precision better than the individual model. Specifically, by applying ORANDNet, we gain +6 in mIoU and +4.1 in precision relative to ResNet-50. Similarly, with the ViT backbone, we achieve +0.9 in mIoU and +3.3 in precision. These results support that ORANDNet is clearly effective in enhancing precision and mitigating noise. Finally, employing IRN further improves both mIoU and precision.

Add-on on other models. Beyond ResNet and ViT, we explored ORANDNet’s performance enhancement with other high-performing WSSS models, specifically AMN (a ResNet-50 variant) and MCTFormer (a ViT variant). We generated pseudo-masks using both AMN and MCTFormer, integrating them with ORANDNet, akin to our earlier experiments. As shown in row 8 of Tabel 1, ORANDNet gain +1.5 mIoU/+4.9 precision and +2.2 mIoU/+4.2 precision respect to AMN and MCTFormer. These results emphasize the effectiveness of the CAM/pseudo-mask ensemble approach of ORANDNet on advanced WSSS methods. It shows its potential as a final add-on module for future WSSS models.

Comparison with a naïve ensemble. Effectiveness of ORANDNet was compared against a naïve ensemble, which simply averages CAMs from ResNet and ViT. Results in row 4 of Table 1 show +5.0 mIoU and +9.5 precision gain in respect to naïve ensemble. These lead to two conclusions: 1) Ensemble approaches improve pseudo mask quality, and 2) ORANDNet outperforms the naïve ensemble.

Comparing segmentation results with recent WSSS

In these experiments, we utilized pseudo-masks for training a segmentation model and evaluated the performance of recent WSSS methods on the PASCAL VOC 2012 *validation set* and *test set*.

Quantitative results. As in Table 2, we found that precision improved pseudo-mask from ORANDNet can yield comparable performance to state-of-the-art WSSS methods with ResNet-50 and ViT. This indicates that enhancing the precision of pseudo-masks contributes significantly to the enhancement of segmentation quality

In contrast to the methodological complexity of existing WSSS methods and their limited gains, our simpler method with ResNet-50 and ViT gained +6.8/+7.3 mIoU respect to IRN(with ResNet-50) on validation/test set. Considering that +1.0 mIoU gain was often seen in prior WSSS studies, +6.8/+7.3 mIoU gain using the simple method we propose is a significant achievement. To highlight this achievement, our method with AMN and MCTFormer achieved an increase of +2.3 and +1.3 in mIoU on the test set, respectively, compared to the performance of each model when used individually.

Note that the second stage methods (denoted as ‘2nd’ in Table 2) suggest the segmentation training strategy thus they can be easily combined with the first stage methods (denoted as ‘1st’). That means, our method can be used in conjunction with BECO and further improve their performance. Additionally, for future WSSS methods that produce more accu-

Method	Stg	Sup.	Val	Test
EDAM (Wu et al. 2021)	1st	I+S	70.9	70.6
EPS (Lee et al. 2021)	1st	I+S	71.0	71.8
SANCE (Li, Fan, and Zhang 2022)	1st	I+S	72.0	72.9
OC-CSE (Kweon et al. 2021)	1st	I	68.4	68.2
CDA (Su et al. 2021)	1st	I	66.1	66.8
PPC (Du et al. 2022)	1st	I	72.6	73.6
URN (Li et al. 2022)	2nd	I	69.5	69.7
ADELE (Liu et al. 2022)	2nd	I	69.3	68.9
BECO (Rong et al. 2023)	2nd	I	73.7	73.5
IRN (Ahn, Cho, and Kwak 2019)	1st	I	63.5	64.8
Ours	1st	I	70.3	72.1
Relative to IRN			+6.8	+7.3
AMN (Lee, Kim, and Shim 2022)	1st	I	70.7	70.6
MCTformer (Xu et al. 2022)	1st	I	71.9	71.6
Ours*	1st	I	72.2	72.9
Relative to AMN, MCTFormer			+1.5, +0.3	+2.3, +1.3

Table 2: mIoU of segmentation results on PASCAL VOC 2012 validation/test set. Ours/Ours* are ORANDNet ensemble of ResNet-50/ViT and AMN/MCTformer, respectively. Sup. denotes the weak supervision type, I is image-level labels and S is saliency supervision. Stg. denotes stages developed in each method.

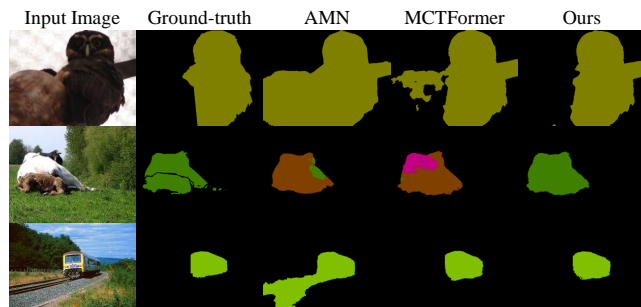


Figure 3: Segmentation results of AMN and MCTformer and their ORANDNet ensemble.

rate pseudo-masks, incorporating ORANDNet into those approach has the potential to further enhance its performance.

Qualitative results. Figure 3 visualizes the improved segmentation results from ORANDNet’s pseudo-masks. The first row highlights improved precision by our method. The second row highlights ORANDNet’s capability to correct false predictions within single objects, which is not observed with existing methods. In the third row, we observe that ORANDNet retains the advantage of existing methods in preventing co-occurrence, inherited from MCTformer. Collectively, these three results demonstrate ORANDNet’s simplicity-yet-powerful ability in enhancing WSSS performance.

Conclusion

We present ORANDNet, a novel ensemble approach in WSSS focusing on pseudo-mask precision. It markedly improves mIoU and precision, achieving performance on par with state-of-the-art even with basic classifiers like ResNet-50 and ViT. As the first WSSS-specific ensemble method, ORANDNet can be employed as add-on to future WSSS

methodologies.

Acknowledgement

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) and IITP grant funded by the Korea government(MSIT) and KEIT grant funded by the Korea government(MOTIE) (No. 2022-0-01045, No. 2022-0-00680).

References

- Ahn, J.; Cho, S.; and Kwak, S. 2019. Weakly supervised learning of instance segmentation with inter-pixel relations. In *CVPR*.
- Ahn, J.; and Kwak, S. 2018. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *CVPR*.
- Araslanov, N.; and Roth, S. 2020. Single-stage semantic segmentation from image labels. In *CVPR*.
- Chen, H.; Wang, J.; Chen, H. C.; Zhen, X.; Zheng, F.; Ji, R.; and Shao, L. 2021. Seminar learning for click-level weakly supervised semantic segmentation. In *ICCV*.
- Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. L. 2014. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*.
- Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.
- Du, Y.; Fu, Z.; Liu, Q.; and Wang, Y. 2022. Weakly supervised semantic segmentation by pixel-to-prototype contrast. In *CVPR*.
- 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. *IJCV*.
- Fan, J.; Zhang, Z.; and Tan, T. 2020. Employing multi-estimations for weakly-supervised semantic segmentation. In *ECCV*. Springer.
- He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*.
- Khoreva, A.; Benenson, R.; Hosang, J.; Hein, M.; and Schiele, B. 2017. Simple does it: Weakly supervised instance and semantic segmentation. In *CVPR*.
- Krähenbühl, P.; and Koltun, V. 2011. Efficient inference in fully connected crfs with gaussian edge potentials. *Advances in neural information processing systems*, 24.
- Kweon, H.; Yoon, S.-H.; Kim, H.; Park, D.; and Yoon, K.-J. 2021. Unlocking the potential of ordinary classifier: Class-specific adversarial erasing framework for weakly supervised semantic segmentation. In *ICCV*.
- 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 *CVPR*.
- 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 *CVPR*.
- Li, J.; Fan, J.; and Zhang, Z. 2022. Towards noiseless object contours for weakly supervised semantic segmentation. In *CVPR*.
- Li, Y.; Duan, Y.; Kuang, Z.; Chen, Y.; Zhang, W.; and Li, X. 2022. Uncertainty estimation via response scaling for pseudo-mask noise mitigation in weakly-supervised semantic segmentation. In *AAAI*.
- Lin, D.; Dai, J.; Jia, J.; He, K.; and Sun, J. 2016. Scribble-sup: Scribble-supervised convolutional networks for semantic segmentation. In *CVPR*.
- Liu, S.; Liu, K.; Zhu, W.; Shen, Y.; and Fernandez-Granda, C. 2022. Adaptive early-learning correction for segmentation from noisy annotations. In *CVPR*.
- Pinheiro, P. O.; and Collobert, R. 2015. From image-level to pixel-level labeling with convolutional networks. In *CVPR*.
- Rong, S.; Tu, B.; Wang, Z.; and Li, J. 2023. Boundary-Enhanced Co-Training for Weakly Supervised Semantic Segmentation. In *CVPR*.
- Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.
- Su, Y.; Sun, R.; Lin, G.; and Wu, Q. 2021. Context decoupling augmentation for weakly supervised semantic segmentation. In *ICCV*.
- Wang, Y.; Zhang, J.; Kan, M.; Shan, S.; and Chen, X. 2020. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In *CVPR*.
- Wei, Y.; Feng, J.; Liang, X.; Cheng, M.-M.; Zhao, Y.; and Yan, S. 2017. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In *CVPR*.
- Wu, T.; Huang, J.; Gao, G.; Wei, X.; Wei, X.; Luo, X.; and Liu, C. H. 2021. Embedded discriminative attention mechanism for weakly supervised semantic segmentation. In *CVPR*.
- Wu, Z.; Shen, C.; and Van Den Hengel, A. 2019. Wider or deeper: Revisiting the resnet model for visual recognition. *Pattern Recognition*.
- Xu, L.; Ouyang, W.; Bennamoun, M.; Boussaid, F.; and Xu, D. 2022. Multi-class token transformer for weakly supervised semantic segmentation. In *CVPR*.
- Zhou, T.; Zhang, M.; Zhao, F.; and Li, J. 2022. Regional semantic contrast and aggregation for weakly supervised semantic segmentation. In *CVPR*.
- Zhu, L.; Li, Y.; Fang, J.; Liu, Y.; Xin, H.; Liu, W.; and Wang, X. 2023. WeakTr: Exploring Plain Vision Transformer for Weakly-supervised Semantic Segmentation. *arXiv preprint arXiv:2304.01184*.