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

Supplementary Material

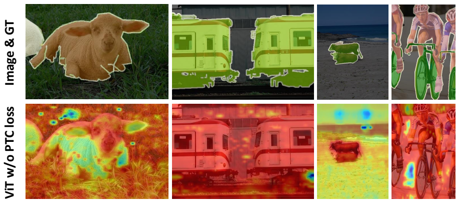

Figure 1. **CAM visualization from vanilla ViT backbone.** The results are evaluated on the VOC dataset. The vanilla ViT is severely suffered from the over-smoothing problem.

Strategy	M	Seg.
Linear Descent	72.5	68.7
Cosine Descent	73.5	69.9

Table 1. **The impact of descent strategy in progressive learning.** The results are evaluated on the VOC `val` set. "M" denotes the CAM performance and **"Seg."** denotes the segmentation performance. CRF post-processing is not implemented for evaluation.

1. ViT backbone with Token Contrast Loss

Following previous works [4, 6, 7] , this paper uses the ViT-B [2] as the backbone. Additionally, we integrate the Patch Token Contrast (PTC) module [6] into the ViT backbone as our experiment baseline. This is because the previous study [6] demonstrates that ViT-like architectures encounter the over-smoothing issue as it tends to smooth the patch tokens progressively (as shown in Figure 1). This phenomenon severely impairs the CAM and segmentation performance. Due to the observation that the learned representations in intermediate layers can still preserve the semantic diversity, the PTC module set an additional classifier in an intermediate layer to extract the auxiliary CAM. Then, the auxiliary CAM is used to generate corresponding pseudo pairwise token relations to supervise the pairwise cosine similarities of final patch tokens. More details can be found in [6].

2. Additional Hyper-parameter Analysis

We report the impact of other hyper-parameters of DuPL in this section.

γ	M	Seg.
0.5	73.1	69.5
0.7	72.7	69.1
0.9	73.5	69.9
0.95	72.9	69.2

Table 2. **The impact of γ in Adaptive Noise Filtering strategy.** The results are evaluated on the VOC `val` set. "M" denotes the CAM performance and **"Seg."** denotes the segmentation performance. CRF post-processing is not implemented for evaluation.

Descent Strategy adopted in Progressive Learning. We develop a dynamic threshold adjustment (DTA) strategy in DuPL, which uses the cosine descent strategy to introduce more pixels to the supervision progressively. In Table 1, we conduct the performance comparison between DTA with the cosine descent strategy and DTA with the linear descent strategy. We show the cosine descent strategy in dynamic threshold adjustment can yield the better performance.

γ in Adaptive Noise Filtering. In the Adaptive Noise Filtering (ANF) strategy, γ is used to control the partition of noise pseudo-labels. The pixel pseudo-labels with the noise probability larger than γ will be considered as noises. Table 2 reports the performance under different settings of γ. We can observe that the setting of $\gamma = 0.9$ can produce the best CAM and segmentation results, while other settings can also achieve favorable performance.

η in Adaptive Noise Filtering. In the Adaptive Noise Filtering (ANF) strategy, η is used to discriminate whether the loss distribution has two distinct peaks. Table 3a reports the performance under different settings of η. We find that $\eta = 1.0$ can achieve the best performance, while $\eta > 1.0$ can also achieve favorable performance.

Loss Weights of the discrepancy loss λ_1. Table 3b reports the analysis of the weights λ_1 of discrepancy loss. A larger λ_1 means that there is a larger proportion of the overall loss, and the model will pay more attention to optimizing this training objective. The results show that $\lambda_1 = 0.1$ can achieve the highest accuracy in most semantic classes. When $\lambda_1 = 0.05$, the penalty of representations from two

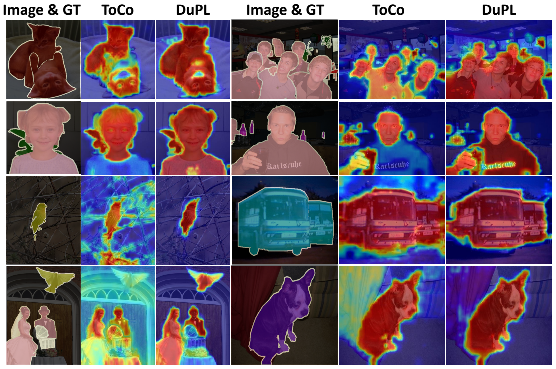

Figure 2. **Visual comparison of CAMs.** The images are from the VOC `val` set. We compare the state-of-the-art one-stage approach, ToCo [6], with our proposed DuPL.

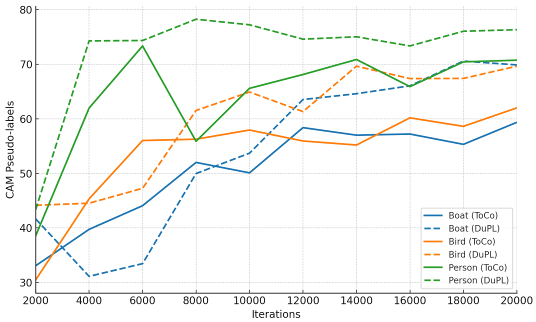

Figure 3. **Tracking of per-class CAM Pseudo-Label quality.** The results are from the VOC `val` set. We compare the state-of-the-art one-stage approach, ToCo [6], with our proposed DuPL.

sub-nets is too little, making the CAMs from two sub-nets have not sufficient diversity. This problem leads to the case with λ_1 has inferior performance.

Loss Weights of the consistency regularization loss λ_3. Table 3c reports the analysis of the weights λ_3 of consistency regularization loss on the predictions of filtered regions. We can observe that $\underline{\lambda_3 = 0.05}$ can achieve the best performance, and the performance decreases after λ_3 is larger than 0.05.

3. Additional Experiment Results

Per-class Segmentation Results. We report the per-class semantic segmentation results on the VOC 2012 `val` set in Table 4. We can see DuPL achieves remarkable performance improvement in most semantic classes (**16 of 21**). We also compare the DuPL performance between using ImageNet-1k and ImageNet-21k pretrained weights, we find that the performance of both versions has its own pros and cons on specific classes, and it is noted that the model pre-trained with larger scale data (*e.g.*, ImageNet-21k) is not necessarily well-performed in every class.

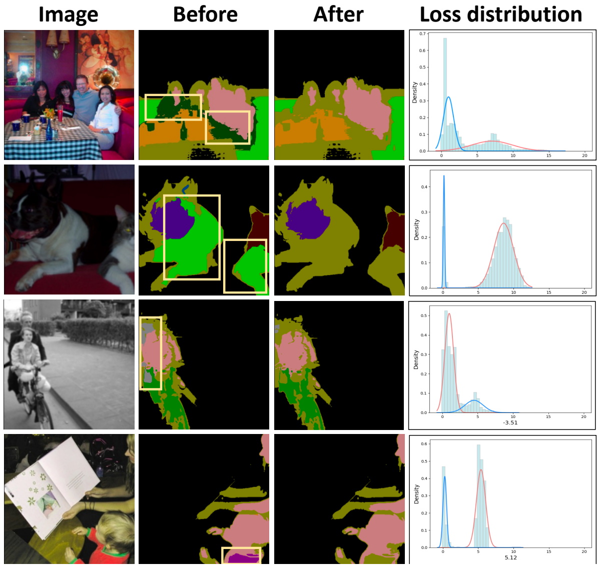

Figure 4. **Visual comparison of ANF.** The images are from the VOC `val` set. The component with larger losses is the noise component in GMM model. The distribution of normal losses is rescaled for visualization.

CAM Results. In Figure 2, we provide more qualitative results of CAM produced by ToCo [6] and our DuPL. We can observe that DuPL can outperform ToCo and provide better segmentation supervision for the segmentation head. Additionally, we present tracking of the CAM pseudo-labels quality on a per-class basis through line graphs in Figure 3. These graphs showcase the evolution of CAM quality across different iterations, further illustrating the superiority of DuPL in generating more accurate pseudo-labels for each class.

ANF Results. In Figure 4, we provide more qualitative results of the pseudo-labels and those adopted adaptive noise filtering strategy (ANF). We can see that the loss distribution of the noise pseudo-labels have clear difference with that of the clean ones. With our ANF strategy, it can effectively discard the noisy pseudo-labels and improve the quality of the segmentation supervision.

Per-class OA Rate Results. We report the per-class semantic segmentation results on the VOC 2012 `val` set in Figure 5. It shows that, compare with the recent state-of-the-art works, *i.e*., ToCo [6], the proposed DuPL can significantly overcome the over-activation problem caused by CAM confirmation bias in most of semantic classes.

Semantic Segmentation Results. In Figure 6, we provide more qualitative results of semantic segmentation predicted by ToCo [6] and the proposed DuPL. We can see DuPL can achieve better object coverage and get more closer predictions to the ground-truths.

η	M	Seg.
0.5	71.5	67.8
1.0	73.5	69.9
1.5	73.4	69.6
2.0	73.2	69.1

(a) The impact of $\boldsymbol{\eta}$ in ANF strategy.

λ_1	M	Seg.
0.05	69.3	66.1
0.1	73.5	69.9
0.2	71.1	67.8

(b) The impact of weight of discrepancy loss.

λ_3	M	Seg.
0.05	69.3	66.1
0.1	73.5	69.9
0.2	71.1	67.8

(c) The impact of weight of regularization loss.

Table 3. **Results of Hyper-parameter Analysis.** The results are evaluated on the VOC `val` set. The default settings are marked in color. "M" denotes the CAM performance and **"Seg."** denotes the segmentation performance. CRF post-processing is not implemented for evaluation.

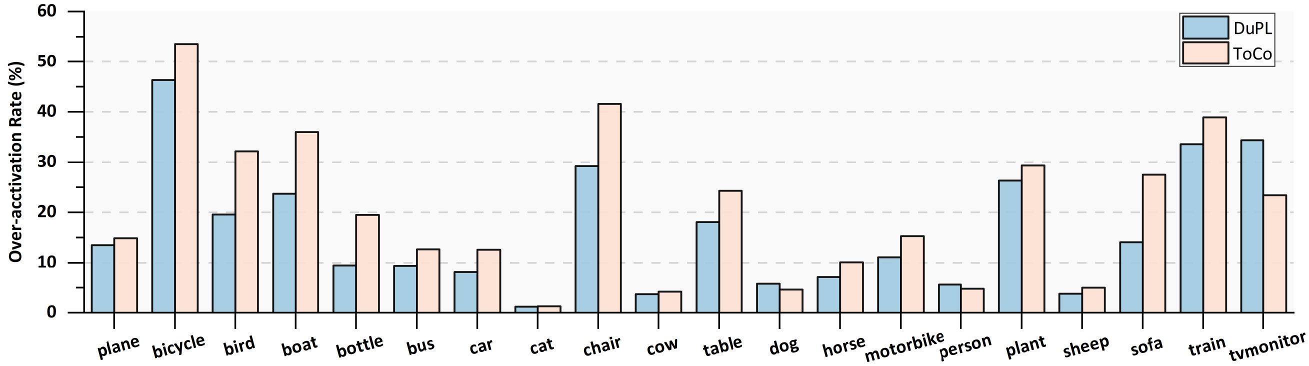

Figure 5. **Comparison of ToCo [6] and the proposed DuPL in OA rate.** The results are evaluated on the VOC 2012 `val` dataset.

	bkg	*aero*	*bicycle*	*bird*	*boat*	*bottle*	*bus*	*car*	*cat*	*chair*	*cow*	*table*	*dog*	*horse*	*motor*	*person*	*plant*	*sheep*	*sofa*	*train*	*tv*	*mIoU*
1Stage [1]	88.7	70.4	35.1	75.7	51.9	65.8	71.9	64.2	81.1	30.8	73.3	28.1	81.6	69.1	62.6	74.8	48.6	71.0	40.1	68.5	64.3	62.7
AFA [5]	89.9	79.5	31.2	80.7	67.2	61.9	81.4	65.4	82.3	28.7	83.4	41.6	82.2	75.9	70.2	69.4	53.0	85.9	44.1	64.2	50.9	66.0
ToCo	89.9	81.8	35.4	68.1	62.0	76.6	83.6	80.4	87.7	24.5	88.1	54.9	87.0	84.0	76.0	68.2	**65.6**	85.8	42.4	57.7	**65.6**	69.8
ToCo†	91.1	80.6	**48.7**	68.6	45.4	**79.6**	87.4	**83.3**	89.9	35.8	84.7	60.5	83.7	83.2	76.8	83.0	56.6	87.9	43.5	60.5	63.1	71.1
DuPL	**91.9**	**82.7**	41.3	74.7	**65.8**	77.2	88.2	82.3	89.9	25.6	88.3	52.4	87.7	86.7	**80.6**	82.0	63.7	90.6	49.7	62.7	52.9	72.2
DuPL †	91.8	77.8	47.1	**81.7**	58.9	78.6	**88.8**	77.6	**91.9**	**38.2**	**91.5**	55.5	**88.0**	**90.0**	77.7	**85.9**	60.7	**92.7**	**54.0**	**66.1**	45.5	**73.3**

Table 4. **Evaluation and comparison of the semantic segmentation results in mIoU on the VOC `val` set.** † denotes using ImageNet-21k [3] pretrained weights.

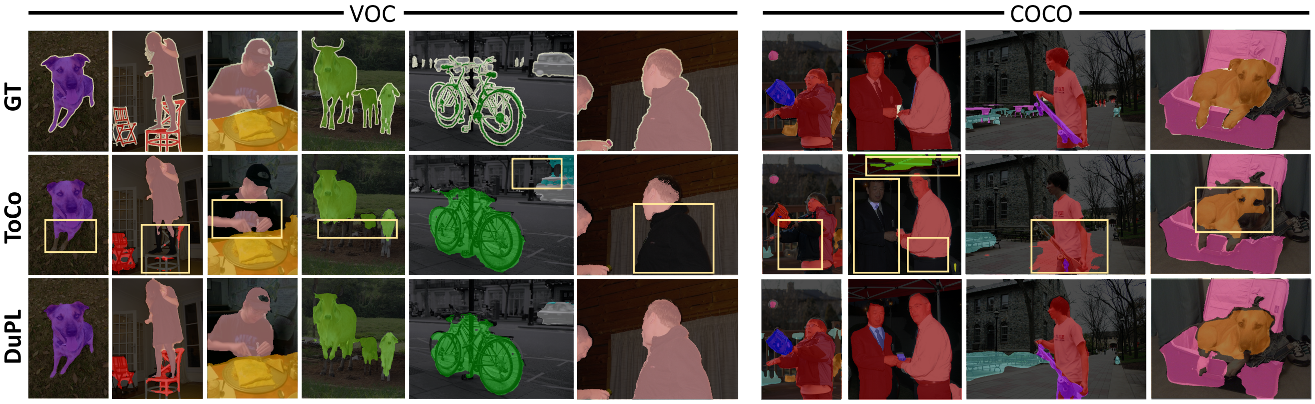

Figure 6. **Visual comparison of segmentation performance.** We compare the state-of-the-art one-stage approach, ToCo [6], with our proposed DuPL. Both of them use ViT-B with ImageNet-1k for fair comparison.

References

[1] 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. 3

[2] 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. 1

[3] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. *arXiv preprint arXiv:2104.10972*, 2021. 3

[4] 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. 1

[5] 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. 3

[6] 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

[7] 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. 1