Supplementary of Token Contrast for Weakly-Supervised Semantic Segmentation

1. Additional Results

1.1. Backbone with ViT Variants

In the previous experiments, we mainly conducted experiments with ViT-B as the backbone. In Figure 1 and Figure 2, we report the evaluation of the generated CAM and semantic results with ViT using other configurations (ViT-S, ViT-L† [3]). ViT-S and ViT-L consist of 12 and 24 Transformer blocks, respectively. We show that other backbones also encounter the over-smoothing issue and the proposed ToCo can finely address it. Specifically, without the proposed ToCo, the generated CAM typically activates all image regions, and the semantic segmentation results also perform badly. In a contrast, the proposed ToCo finely addresses the over-smoothing issue and promotes the semantic segmentation performance to 65.2% and 71.2% mIoU with ViT-S and ViT-L as the backbone, respectively.

1.2. Hyper-parameters

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

Background Thresholds. In Table 2a, we report the impact of background thresholds to differentiate the foreground, background, and uncertain regions. We show the combination of $\beta_h = 0.7$ and $\beta_l = 0.25$ can achieve the best performance.

Temperature Factors. Table 2b presents the performance w.r.t. the temperature factor τ in Equation (4). τ control the sharpness of the logits. In Table 2b, we observe that τ_g=0.5 yields the best performance, while other values can also achieve favorable performance.

Loss Weights. Table 2c reports the analysis of the weights of loss terms. The combination of (λ_1 = 0.2, λ_2 = 0.5, λ_3 = 0.1) can produce the best semantic segmentation results.

1.3. Setting of $\mathcal{L}_{ptc}$

In the PTC module, due to the observation that the cosine similarities of patch tokens are usually positive values, as indicated in [2, 4], we use the absolute cosine similarity instead of the origin cosine similarity in $\mathcal{L}_{ptc}$. In Table 1, we report the evaluation of pseudo labels M and semantic segmentation results Seg.

Table 1 shows that directly minimizing the cosine similarity (`CosSim`) cannot produce satisfactory results. A possible reason is that two patch tokens with negative similarities are still correlated. When ignoring the negative parts (`ReLu(CosSim)`), the results are remarkably promoted. Finally, using the absolute cosine similarity (`Abs(CosSim)`) can finely optimize the positive and negative parts and yield the best results.

1.4. Additional Quantitative/Qualitative Results

Per-Class Results. We report the per-class semantic segmentation results on the VOC val set in Table 3. Table 3 shows that the proposed ToCo can achieve the highest accuracy in most semantic classes.

Pseudo Labels. We present the generated CAM in Figure 3. Figure 3 demonstrates that the proposed PTC and CTC can address the over-smoothing issue and further distinguish the uncertain regions, respectively. Besides, ToCo can generate better pseudo labels than the recent state-of-the-art single-stage method, AFA [6].

Semantic Segmentation Results. The qualitative semantic segmentation in Figure 4 shows that ToCo can surpass AFA [6] and achieve close results with the ground-truth.

Attention Maps of Class Token. In Figure 5, we visualize more attention maps of class token w.r.t. other patch tokens. Figure 5 shows the global view can discover most object regions but ignore some uncertain local regions, which can be activated in the local view. By contrasting the class tokens of global and local views in the CTC module, the representation of the integral regions can be more consistent.

	M	Seg.
`CosSim`	36.5	29.6
`ReLu(CosSim)`	63.0	60.1
`Abs(CosSim)`	70.5	68.1

Table 1. **The impact of CosSim in $\mathcal{L}_{ptc}$.**

ViT-S		CAM	*Seg.*
`train`	*w/o* ToCo	27.3	–
	w/ ToCo	**69.1**	**68.1**
`val`	*w/o* ToCo	27.6	–
	w/ ToCo	**68.2**	**65.2**

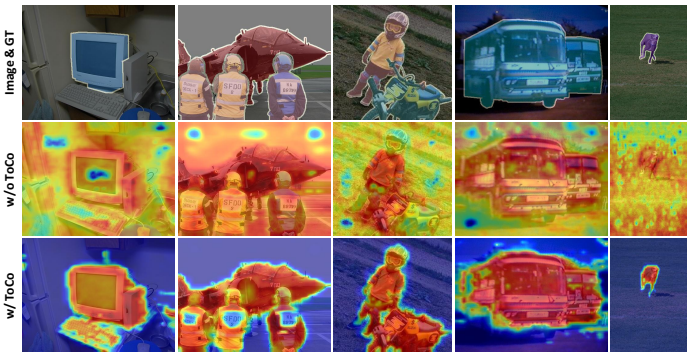

Figure 1. **Evaluation of the generated CAM and semantic segmentation results with ViT-S**. The results are evaluated on the VOC dataset.

ViT-L†		CAM	*Seg.*
`train`	*w/o* ToCo	25.3	–
	w/ ToCo	**73.8**	**74.2**
`val`	*w/o* ToCo	25.6	–
	w/ ToCo	**72.6**	**71.2**

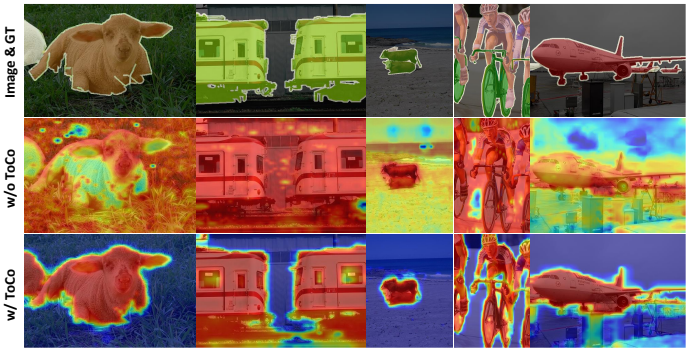

Figure 2. **Evaluation of the generated CAM and semantic segmentation results with ViT-L†**. The results are evaluated on the VOC dataset.

		β_l			
		0.15	0.2	0.25	0.3
β_h	0.6	–	55.1	62.9	66.7
	0.65	51.2	61.5	67.2	66.3
	0.7	56.2	65.5	**68.1**	67.1
	0.75	63.0	65.6	66.3	64.3

(a) **Background thresholds**.

τ	0.1	0.2	0.5	0.8
Seg.	66.0	67.2	**68.1**	67.3

(b) **Temperatures**.

λ_1	0.05	0.1	0.2	0.5
Seg.	64.2	67.0	**68.1**	65.4
λ_2	0.1	0.2	0.5	0.8
Seg.	65.1	65.9	**68.1**	67.1
λ_3	0.05	0.1	0.2	0.5
Seg.	67.8	**68.1**	67.6	66.0

(c) **Loss Weights**.

Table 2. **Impact of hyper-parameters.** The performance is evaluated on the VOC `val` set. The default settings are marked in gray.

	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 [6]	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**

Table 3. Evaluation and comparison of the semantic segmentation results in mIoU on the *val* set. † denotes using ImageNet-21k [5] pretrained weights.

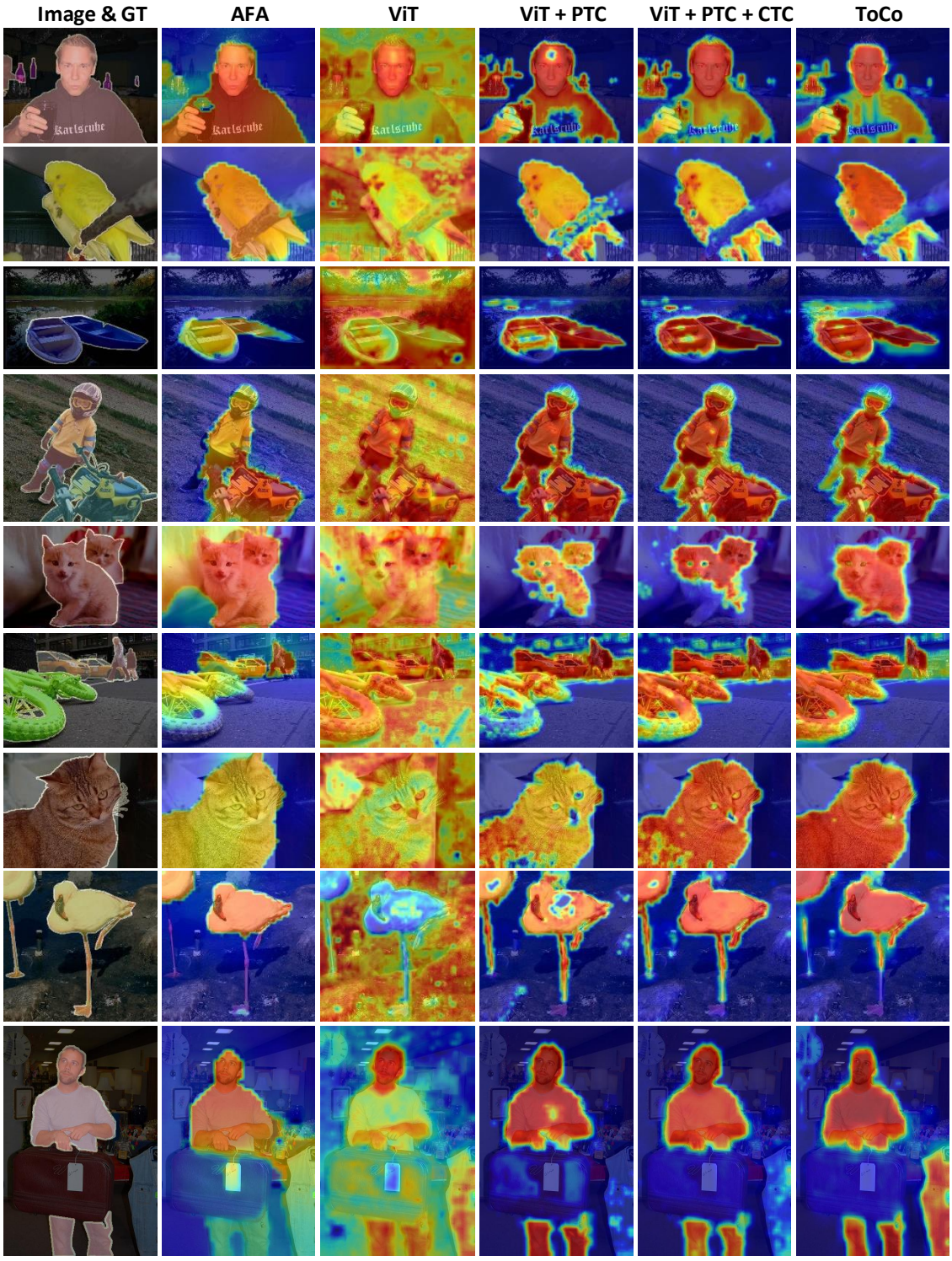

Figure 3. **Visualization of CAM.** From left to right, the CAM is generated with AFA [6], ViT baseline, ViT with PTC, ViT with PTC and CTC, and the proposed ToCo.

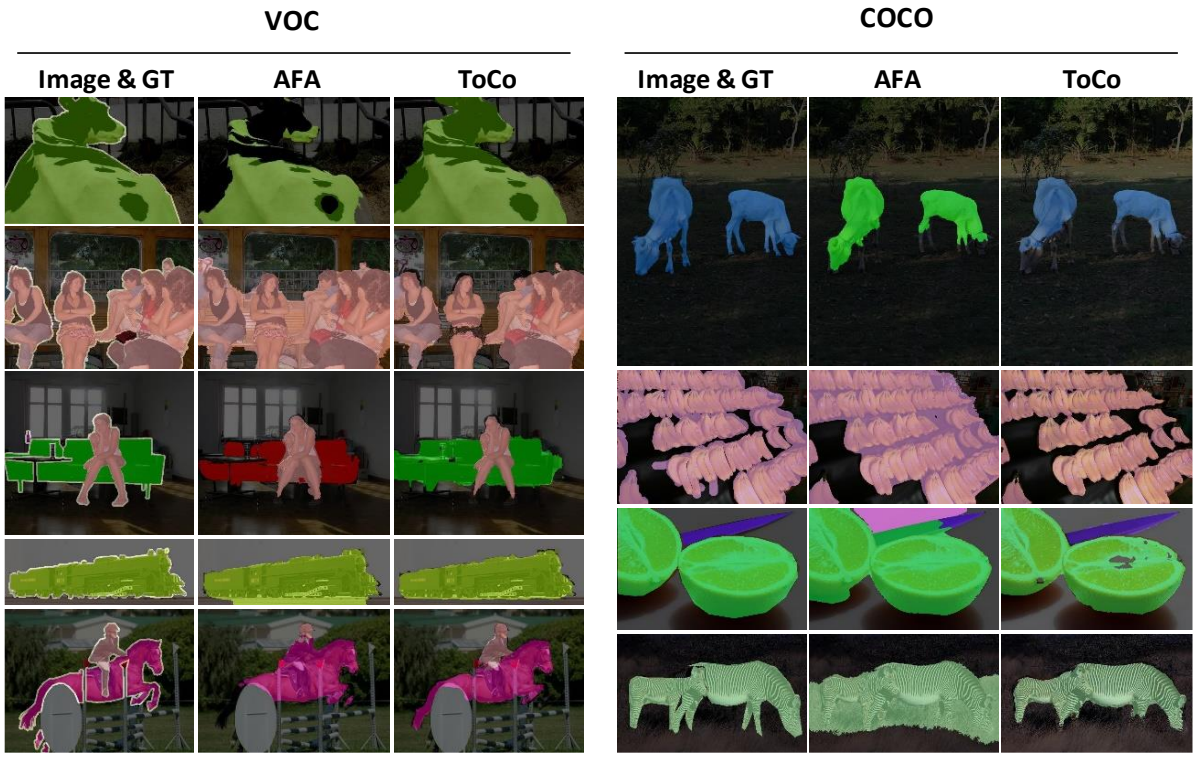

Figure 4. **Semantic segmentation results on the VOC and COCO dataset**.

Figure 5. **Visualization of the attention map of class token w.r.t. patch tokens**. The brighter region indicates a larger attention value. *Left*: the global view image in CTC; *Right*: the local view image randomly cropped from the global view in CTC.

References

[1] Nikita Araslanov and Stefan Roth. Single-stage semantic segmentation from image labels. In *CVPR*, pages 4253–4262, 2020. 2

[2] Tianlong Chen, Zhenyu Zhang, Yu Cheng, Ahmed Awadallah, and Zhangyang Wang. The principle of diversity: Training stronger vision transformers calls for reducing all levels of redundancy. In *CVPR*, pages 12020–12030, 2022. 1

[3] 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. In *ICLR*, 2020. 1

[4] Chengyue Gong, Dilin Wang, Meng Li, Vikas Chandra, and Qiang Liu. Vision transformers with patch diversification. *arXiv preprint arXiv:2104.12753*, 2021. 1

[5] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. *arXiv preprint arXiv:2104.10972*, 2021. 2

[6] Lixiang Ru, Yibing Zhan, Baosheng Yu, and Bo Du. Learning affinity from attention: End-to-end weakly-supervised semantic segmentation with transformers. In *CVPR*, 2022. 1, 2, 3