

DINO is Also a Semantic Guider: Exploiting Class-aware Affinity for Weakly Supervised Semantic Segmentation

– Supplementary Materials

Anonymous Author(s)
Submission Id: 853

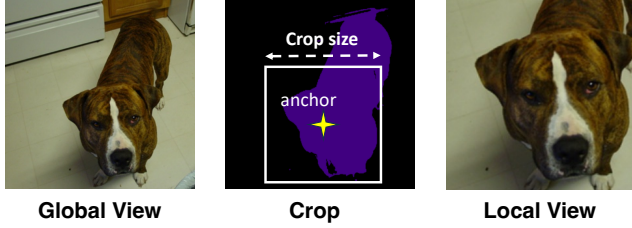


Figure 1: The modified cropping strategy. We use this strategy to generate the local views guided by the produced pseudo-labels.

1 DETAILS OF MIT BACKBONE IN ECA

Following previous single-stage prior arts [24, 33], we adopt MiT-B1, proposed in SegFormer [31], as the backbone for our ECA framework. Compared to the vanilla ViT [11], SegFormer is specifically designed for semantic segmentation, making it a more suitable architecture for WSSS tasks. SegFormer incorporates overlapped patch merging layers with different strides to compute multi-head self-attention and generate multi-scale feature maps. The standard SegFormer with MiT-B1 has a stride of [4, 8, 16, 32]. In our work, we modify the stride of the last patch merging layer to 16 to increase the resolution of the feature maps. Early experiments shows that this modification can effectively benefit the CAM performance, which can produce better CAM pseudo-labels.

2 DETAILS OF CROPPING STRATEGY IN CCE

In the CAM Correspondence Enhancement (CAE) module, we modify the random cropping technique to generate the local views guided by the produced pseudo-labels. This is because utilizing a basic random cropping strategy may result in some local views lacking target objects, deviating from our intended design objective. Early experiments also show that performing CCE on the local views without objects would lead to performance degradation. To this end, we modify the random cropping strategy to guarantee that the local views can contain partial objects. Specifically, we identify pixels corresponding to the objects indicated by the pseudo-labels as “anchors”. Then, we randomly select an anchor pixel as the center and crop the image with crop size 256 (default). This process is illustrated in Figure 1.

3 MORE EXPERIMENTAL RESULTS

Threshold γ in the SAE module. In the Semantic Affinity Exploitation module (SAE), we set γ to generate the binary adjacency vector A . Table 1 reports the performance under different settings

γ	CAM	val (msc)
-0.15	54.6	54.4
0.00	69.2	68.2
0.15	69.1	69.4
0.30	67.5	67.3

Table 1: Impact of the threshold γ in the SAE module. The results are evaluated on the VOC 2012 val set. “msc”: the multi-scale inference test. The segmentation results are not implemented dense CRF post-processing.

τ	Seg	Seg (msc)
0.10	66.8	68.7
0.50	66.9	69.0
0.90	67.1	69.4
0.99	67.6	69.4

Table 2: Impact of the momentum τ on prototype update for CARM calibration. The results are evaluated on the VOC 2012 val set. “msc”: the multi-scale inference test. The segmentation results are not implemented dense CRF post-processing.

of γ . We can observe that the setting of $\gamma \geq 0$ can produce favorable CAM performance, while the setting of $\gamma = 0.15$ can produce the best segmentation results, and we use this setting as default.

Momentum τ in the CARM Calibration. After the production of CARM, we utilize momentum prototypes to calibrate them effectively. Higher values of τ result in more stable prototypes for each class. To investigate the impact of different momentums, we conduct experiments within the range of [0.1, 0.99]. From Table 2, we can observe that ECA reveals a notable improvement in both single-scale and multi-scale segmentation performance as the momentum increases. This observation highlights the significance of maintaining stable prototypes in CARM calibration, ultimately enhancing the overall segmentation performance. In our ECA framework, $\tau = 0.99$ is adopted as default.

Weight of Loss Terms. We present the segmentation results on the VOC 2012 val set with different weight factors of loss terms. From Table 3, we can observe that $\lambda_1 = 0.1, \lambda_2 = 0.2, \lambda_3 = 0.01$ is a preferred choice for our ECA framework.

Semantic Segmentation Results. In Figure 3, we provide more qualitative results of semantic segmentation predicted by AFA [24] (with the same MiT-B1 backbone) and the proposed ECA. We can

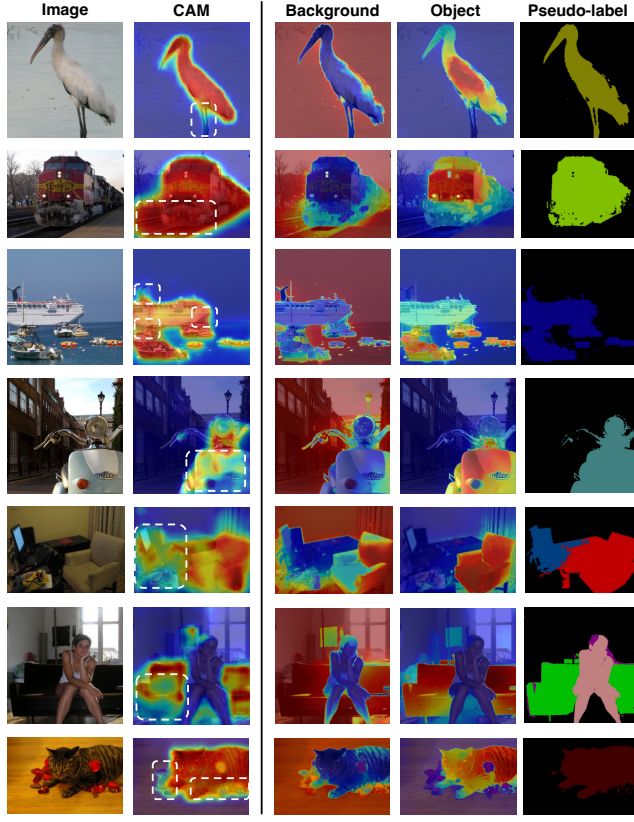


Figure 2: Visualization of CARM and its generated pseudo-labels. Some images have multiple object classes, we only visualize the CAM and CARM of one target class for clear demonstration. The produced pseudo-labels contain multiple classes' annotations.

see ECA can achieve better object coverage and get more closer predictions to the ground-truths.

4 MORE QUALITATIVE RESULTS

CARM and its pseudo-labels. We present some cases of CARM pseudo-label generation in Figure 2. Although some CARMs of target objects are incomplete, we can still obtain satisfactory pseudo-labels with the help of background CARM. Compared to CAM pseudo-labels, it can provide complementary semantic guidance to overcome the over-activation and under-activation issues for the segment decoder, thus boosting the segmentation performance.

Semantic Segmentation Results. In Figure 3, we provide more qualitative results of semantic segmentation predicted by AFA [24] (with the same MiT-B1 backbone) and the proposed ECA. We can see ECA can achieve better object coverage and get more closer predictions to the ground-truths.



Figure 3: Visualization of segmentation results on PASCAL VOC val set and MS COCO val set. We compared the results of the proposed ECA with AFA [24] (with the same MiT-B1 backbone). Our methods outperforms AFA and are more closer to the ground-truths.

	λ_1	λ_2	λ_3	val
Default	0.1	0.1	0.1	69.4
	0.05			68.3
	0.2			66.9
		0.05		68.1
		0.2		67.5
			0.05	68.8
			0.02	67.3

Table 3: Impact of the weights of loss terms. The results are evaluated on the val set of PASCAL VOC 2012. The segmentation results are not implemented dense CRF post-processing.

REFERENCES

- [1] Jiwoon Ahn, Sunghyun Cho, and Suha Kwak. 2019. Weakly supervised learning of instance segmentation with inter-pixel relations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2209–2218.
- [2] Jiwoon Ahn and Suha Kwak. 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*. 4981–4990.
- [3] Nikita Araslanov and Stefan Roth. 2020. Single-stage semantic segmentation from image labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4253–4262.
- [4] Amy Bearman, Olga Russakovsky, Vittorio Ferrari, and Li Fei-Fei. 2016. What’s the point: Semantic segmentation with point supervision. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14*. Springer, 549–565.
- [5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*. 9650–9660.
- [6] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. 2017. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence* 40, 4 (2017), 834–848.
- [7] Xinlei Chen, Saining Xie, and Kaiming He. 2021. An Empirical Study of Training Self-Supervised Vision Transformers. *arXiv preprint arXiv:2104.02057* (2021).
- [8] Zhaozheng Chen, Tan Wang, Xiongwei Wu, Xian-Sheng Hua, Hanwang Zhang, and Qianru Sun. 2022. Class re-activation maps for weakly-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 969–978.
- [9] Zesen Cheng, Pengchong Qiao, Kehan Li, Siheng Li, Pengxu Wei, Xiangyang Ji, Li Yuan, Chang Liu, and Jie Chen. 2023. Out-of-candidate rectification for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 23673–23684.
- [10] Jifeng Dai, Kaiming He, and Jian Sun. 2015. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*. 1635–1643.
- [11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiuhua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).
- [12] Ye Du, Zehua Fu, Qingjie Liu, and Yunhong Wang. 2022. Weakly supervised semantic segmentation by pixel-to-prototype contrast. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4320–4329.
- [13] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. 2010. The pascal visual object classes (voc) challenge. *International journal of computer vision* 88 (2010), 303–338.
- [14] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhansu Maji, and Jitendra Malik. 2011. Semantic contours from inverse detectors. In *2011 international conference on computer vision*. IEEE, 991–998.
- [15] Longlong Jing and Yingli Tian. 2020. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence* 43, 11 (2020), 4037–4058.
- [16] Hyeokjun Kwon, Sung-Hoon Yoon, and Kuk-Jin Yoon. 2023. Weakly Supervised Semantic Segmentation via Adversarial Learning of Classifier and Reconstructor. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11329–11339.
- [17] Jungbeom Lee, Seong Joon Oh, Sangdoo Yun, Junsuk Choe, Eunji Kim, and Sungroh Yoon. 2022. Weakly supervised semantic segmentation using out-of-distribution data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16897–16906.
- [18] Jinlong Li, Zequn Jie, Xu Wang, Xiaolin Wei, and Lin Ma. 2022. Expansion and Shrinkage of Localization for Weakly-Supervised Semantic Segmentation. *arXiv preprint arXiv:2209.07761* (2022).
- [19] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*. Springer, 740–755.
- [20] Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).
- [21] Youngmin Oh, Beomjun Kim, and Bumsub Ham. 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*. 6913–6922.
- [22] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193* (2023).
- [23] Junwen Pan, Pengfei Zhu, Kaihua Zhang, Bing Cao, Yu Wang, Dingwen Zhang, Junwei Han, and Qinghua Hu. 2022. Learning self-supervised low-rank network for single-stage weakly and semi-supervised semantic segmentation. *International Journal of Computer Vision* 130, 5 (2022), 1181–1195.
- [24] Lixiang Ru, Yibing Zhan, Baosheng Yu, and Bo Du. 2022. 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*. 16846–16855.
- [25] Lixiang Ru, Heliang Zheng, Yibing Zhan, and Bo Du. 2023. Token Contrast for Weakly-Supervised Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3093–3102.
- [26] Oriane Siméoni, Gilles Puy, Huy V Vo, Simon Roburin, Spyros Gidaris, Andrei Bursuc, Patrick Pérez, Renaud Marlet, and Jean Ponce. 2021. Localizing Objects with Self-Supervised Transformers and no Labels. In *BMVC-British Machine Vision Conference*.
- [27] Xudong Wang, Rohit Girdhar, Stella X Yu, and Ishan Misra. 2023. Cut and learn for unsupervised object detection and instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3124–3134.
- [28] Yangtao Wang, Xi Shen, Yuan Yuan, Yuming Du, Maomao Li, Shell Xu Hu, James L Crowley, and Dominique Vaufreydaz. 2022. TokenCut: Segmenting objects in images and videos with self-supervised transformer and normalized cut. *arXiv preprint arXiv:2209.00383* (2022).
- [29] Yude Wang, Jie Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. 2020. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12275–12284.
- [30] Yuanchen Wu, Xiaoqiang Li, Songmin Dai, Jide Li, Tong Liu, and Shaorong Xie. 2023. Hierarchical Semantic Contrast for Weakly Supervised Semantic Segmentation. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*. 1542–1550.
- [31] Enze Xie, Wenhui Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. 2021. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems* 34 (2021), 12077–12090.
- [32] Lian Xu, Wanli Ouyang, Mohammed Bennamoun, Farid Boussaid, and Dan Xu. 2022. Multi-class token transformer for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4310–4319.
- [33] Rongtao Xu, Changwei Wang, Jiaxi Sun, Shibao Xu, Weiliang Meng, and Xiaopeng Zhang. 2023. Self Correspondence Distillation for End-to-End Weakly-Supervised Semantic Segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 3045–3053.
- [34] Bingfeng Zhang, Jimin Xiao, Jianbo Jiao, Yunchao Wei, and Yao Zhao. 2021. Affinity attention graph neural network for weakly supervised semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 11 (2021), 8082–8096.
- [35] Bingfeng Zhang, Jimin Xiao, Yunchao Wei, Mingjie Sun, and Kaizhu Huang. 2020. Reliability does matter: An end-to-end weakly supervised semantic segmentation approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 12765–12772.
- [36] Xiangrong Zhang, Zelin Peng, Peng Zhu, Tianyang Zhang, Chen Li, Huiyu Zhou, and Licheng Jiao. 2021. Adaptive affinity loss and erroneous pseudo-label refinement for weakly supervised semantic segmentation. In *Proceedings of the 29th ACM International Conference on Multimedia*. 5463–5472.
- [37] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. 2021. Re-thinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 6881–6890.
- [38] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. 2016. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2921–2929.
- [39] Tianfei Zhou, Meijie Zhang, Fang Zhao, and Jianwu Li. 2022. Regional semantic contrast and aggregation for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4299–4309.
- [40] Adrian Ziegler and Yuki M Asano. 2022. Self-supervised learning of object parts for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14502–14511.