S^{4}M: Boosting Semi-Supervised Instance Segmentation with SAM

Supplementary Material

A. More Qualitative Results

Extended qualitative results. We present extended qualitative results of **S^{4}M** on Cityscapes [11] in A.1 and for COCO [35] in A.2. The results demonstrate that our approach consistently achieved improvements over the supervised teacher network across all experimental settings.

Qualitative comparison of the improved teacher with structural distillation. In addition to the quantitative results presented in Tab. 1 of the main paper, we provide qualitative evidence in A.3 to illustrate the improvements achieved by the teacher model trained with structural distillation (SD). A.3 demonstrates that the supervised model with the additional SD loss $\mathcal{L}_{\text{SD}}$ detects objects more effectively and reduces instances where multiple instance masks are merged into a single pseudo-label compared to the baseline model without $\mathcal{L}_{\text{SD}}$.

B. Examples of augmented images with refined pseudo-labels

In Fig. A.4, we present sample outputs of our proposed Refined Instance Mixing (**ARP**). Pseudo-label masks are initially generated from teacher predictions and then refined using SAM, yielding higher-quality pseudo-labels. Building upon these enhanced labels, ARP craft synthetic data by blending instances from paired images, thereby introducing diverse spatial and contextual variations such as novel backgrounds and potential occlusions. This augmentation strategy encourages consistent model performance under challenging conditions and fosters improved robustness and generalization to a wide range of transformations.

C. Additional Analysis

Analysis on pseudo-label quality. Analysis of pseudo-label quality for the original teacher prediction, as provided in Fig. 1 of the main paper, was conducted on the Cityscapes validation set. The segmentation quality (SQ) was quantified using the mean IoU of true positive labels, where a prediction was considered a positive label if it shared the same class with the ground truth and had an IoU exceeding 0.5. Class accuracy (CA) was computed as the ratio of correctly matched predictions (true positives) to the total number of predictions, with matched pairs defined as predictions exceeding an IoU threshold of 0.5. Notably, we did not utilize the region quality metric commonly employed in panoptic quality (PQ) for evaluating class accuracy, as its computation considers false negatives, leading to the inclusion of undetected pseudo-labels and making accurate assessment difficult. Building on this analysis, we evaluated teacher predictions refined through structural distillation, confirming its effectiveness in improving pseudo-label quality metrics and addressing the identified challenges. Tab A.1 indicates that structural distillation effectively enhances both metrics used to evaluate pseudo-label quality, thereby substantively addressing the challenges we discussed.

	Baseline	Baseline+SD
CA	96.9	**97.3**
SQ	47.7	**49.4**

Table A.1. **Comparison of pseudo-label quality analysis.**

Analysis on prompt types for pseudo-label refinement. In Tab A.2, we present the results of applying different SAM prompt types (bounding box, mask, and point) to our method. The results show that multiple point prompts offer the highest performance, aligning with our proposed approach. While bounding boxes follow closely, they can introduce ambiguity when multiple objects appear within a single box. Single-point prompts can lead to degraded performance due to SAM's over-segmentation tendencies. Furthermore, as discussed in prior work [13, 60], relying on mask prompts may lower mask quality and thus negatively affect training.

Prompt Type	AP
Bounding Box	32.1
Mask	24.3
Single point	30.4
K-sampled points (Ours)	**32.8**

Table A.2. **Performance comparison across different prompt type configurations.**

Figure A.1. **Qualitative results on Cityscapes under different labeled data settings.** Predictions from supervised training (top) and our semi-supervised approach (bottom) across different labeled data settings. "Supervised" refers to the pretrained teacher network, while "semi-supervised" denotes the student model trained jointly on both labeled and unlabeled data.

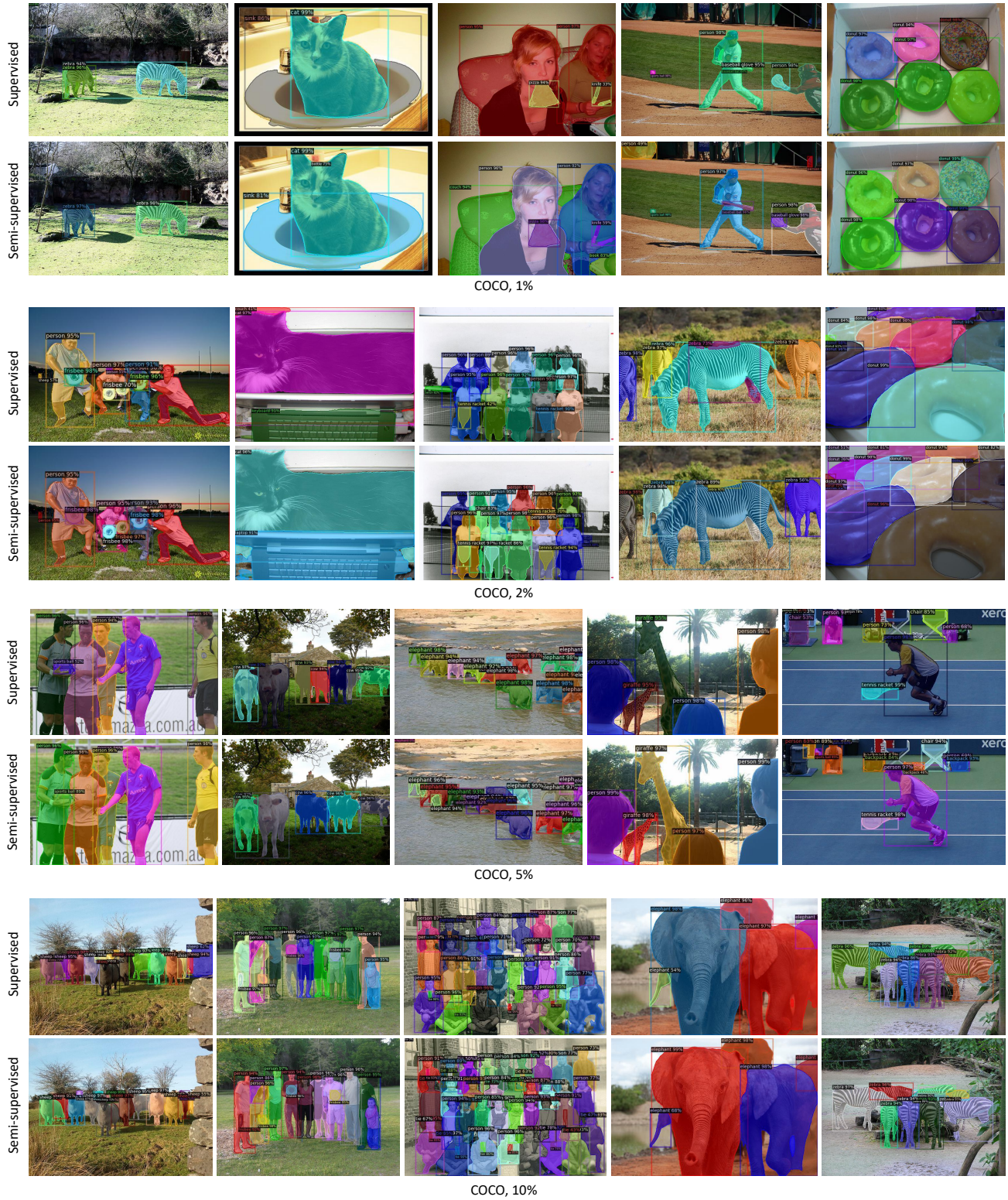

Figure A.2. **Qualitative results on COCO under different labeled data settings.** Predictions from supervised training (top) and our semi-supervised approach (bottom) across different labeled data settings.

Figure A.3. **Qualitative comparison between the improved teacher model, enhanced by structural distillation and trained on 20% of the labeled data, and the baseline model on Cityscapes.**

Figure A.4. **Visualization of augmented samples with refined pseudo-labels**

Acknowledgement

This research was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (RS-2019-II190075, RS-2024-00509279, RS-2025-II212068, RS-2023-00227592, RS-2025-02214479, RS-2024-00457882) and the Culture, Sports, and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism (RS-2024-00345025, RS-2024-00333068), and National Research Foundation of Korea (RS-2024-00346597).

References

[1] Shir Amir, Yossi Gandelsman, Shai Bagon, and Tali Dekel. Deep vit features as dense visual descriptors. *arXiv preprint arXiv:2112.05814*, 2(3):4, 2021. 2

[2] Honggyu An, Jinhyeon Kim, Seonghoon Park, Jaewoo Jung, Jisang Han, Sunghwan Hong, and Seungryong Kim. Cross-view completion models are zero-shot correspondence estimators. *arXiv preprint arXiv:2412.09072*, 2024. 2

[3] Tariq Berrada, Camille Couprie, Karteek Alahari, and Jakob Verbeek. Guided distillation for semi-supervised instance segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 475–483, 2024. 1, 2, 3, 4, 6, 7, 8

[4] Yunkang Cao, Xiaohao Xu, Chen Sun, Yuqi Cheng, Zongwei Du, Liang Gao, and Weiming Shen. Segment any anomaly without training via hybrid prompt regularization. *arXiv preprint arXiv:2305.10724*, 2023. 2

[5] Xin Chen, Jie Hu, Xiawu Zheng, Jianghang Lin, Liujuan Cao, and Rongrong Ji. Depth-guided semi-supervised instance segmentation. *arXiv preprint arXiv:2406.17413*, 2024. 1, 2, 3, 6

[6] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in neural information processing systems*, 34:17864–17875, 2021. 1

[7] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022. 1, 3, 6, 7

[8] Seokju Cho, Sunghwan Hong, Sangryul Jeon, Yunsung Lee, Kwanghoon Sohn, and Seungryong Kim. Cats: Cost aggregation transformers for visual correspondence. *Advances in Neural Information Processing Systems*, 34:9011–9023, 2021. 4

[9] Seokju Cho, Sunghwan Hong, and Seungryong Kim. Cats++: Boosting cost aggregation with convolutions and transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):7174–7194, 2022.

[10] Seokju Cho, Heeseong Shin, Sunghwan Hong, Anurag Arnab, Paul Hongsuck Seo, and Seungryong Kim. Cat-seg: Cost aggregation for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4113–4123, 2024. 4

[11] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 5, 6, 1

[12] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703, 2020. 5

[13] Haixing Dai, Chong Ma, Zhiling Yan, Zhengliang Liu, Enze Shi, Yiwei Li, Peng Shu, Xiaozheng Wei, Lin Zhao, Zihao Wu, et al. Samaug: Point prompt augmentation for segment anything model. *arXiv preprint arXiv:2307.01187*, 2023. 1

[14] Terrance DeVries. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017. 5

[15] 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. 4

[16] Dominik Filipiak, Andrzej Zapała, Piotr Tempczyk, Anna Fensel, and Marek Cygan. Polite teacher: Semi-supervised instance segmentation with mutual learning and pseudo-label thresholding. *IEEE Access*, 12:37744–37756, 2024. 1, 2, 4

[17] Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D Cubuk, Quoc V Le, and Barret Zoph. Simple copy-paste is a strong data augmentation method for instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2918–2928, 2021. 5

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[19] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 1, 6

[20] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 1

[21] Tong He, Chunhua Shen, Zhi Tian, Dong Gong, Changming Sun, and Youliang Yan. Knowledge adaptation for efficient semantic segmentation. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 578–587, 2019. 3

[22] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015. 3

[23] Sunghwan Hong and Seungryong Kim. Deep matching prior: Test-time optimization for dense correspondence. In

Proceedings of the IEEE/CVF international conference on computer vision, pages 9907–9917, 2021. 4

[24] Sunghwan Hong, Seokju Cho, Jisu Nam, Stephen Lin, and Seungryong Kim. Cost aggregation with 4d convolutional swin transformer for few-shot segmentation. In *European Conference on Computer Vision*, pages 108–126. Springer, 2022. 1

[25] Sunghwan Hong, Jisu Nam, Seokju Cho, Susung Hong, Sangryul Jeon, Dongbo Min, and Seungryong Kim. Neural matching fields: Implicit representation of matching fields for visual correspondence. *Advances in Neural Information Processing Systems*, 35:13512–13526, 2022.

[26] Sunghwan Hong, Seokju Cho, Seungryong Kim, and Stephen Lin. Unifying feature and cost aggregation with transformers for semantic and visual correspondence. *arXiv preprint arXiv:2403.11120*, 2024.

[27] Sunghwan Hong, Jaewoo Jung, Heeseong Shin, Jiaolong Yang, Seungryong Kim, and Chong Luo. Unifying correspondence pose and nerf for generalized pose-free novel view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20196–20206, 2024. 4

[28] Jie Hu, Chen Chen, Liujuan Cao, Shengchuan Zhang, Annan Shu, Guannan Jiang, and Rongrong Ji. Pseudo-label alignment for semi-supervised instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16337–16347, 2023. 1, 2, 6

[29] Yuhao Huang, Xin Yang, Lian Liu, Han Zhou, Ao Chang, Xinrui Zhou, Rusi Chen, Junxuan Yu, Jiongquan Chen, Chaoyu Chen, et al. Segment anything model for medical images? *Medical Image Analysis*, 92:103061, 2024. 2

[30] Peter J Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics: Methodology and distribution*, pages 492–518. Springer, 1992. 4

[31] Deyi Ji, Feng Zhao, Hongtao Lu, Feng Wu, and Jieping Ye. Structural and Statistical Texture Knowledge Distillation and Learning for Segmentation . *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 5555. 3

[32] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9404–9413, 2019. 1

[33] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 2, 5

[34] Hyeokjun Kweon and Kuk-Jin Yoon. From sam to cams: Exploring segment anything model for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19499–19509, 2024. 2

[35] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. 5, 7, 1

[36] Yifan Liu, Ke Chen, Chris Liu, Zengchang Qin, Zhenbo Luo, and Jingdong Wang. Structured knowledge distillation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[37] Juzheng Miao, Cheng Chen, Keli Zhang, Jie Chuai, Quanzheng Li, and Pheng-Ann Heng. Cross prompting consistency with segment anything model for semi-supervised medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 167–177. Springer, 2024. 2

[38] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. Ieee, 2016. 3

[39] Viktor Olsson, Wilhelm Tranheden, Juliano Pinto, and Lennart Svensson. Classmix: Segmentation-based data augmentation for semi-supervised learning. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1369–1378, 2021. 5

[40] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 1

[41] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3967–3976, 2019. 3

[42] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 1

[43] Ilija Radosavovic, Piotr Dollár, Ross Girshick, Georgia Gkioxari, and Kaiming He. Data distillation: Towards omni-supervised learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4119–4128, 2018. 6

[44] Frano Rajič, Lei Ke, Yu-Wing Tai, Chi-Keung Tang, Martin Danelljan, and Fisher Yu. Segment anything meets point tracking. *arXiv preprint arXiv:2307.01197*, 2023. 2

[45] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 2, 4, 6

[46] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024. 2

[47] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014. 4, 8

[48] Chaitanya Ryali, Yuan-Ting Hu, Daniel Bolya, Chen Wei, Haoqi Fan, Po-Yao Huang, Vaibhav Aggarwal, Arkabandhu

Chowdhury, Omid Poursaeed, Judy Hoffman, et al. Hiera: A hierarchical vision transformer without the bells-and-whistles. In *International Conference on Machine Learning*, pages 29441–29454. PMLR, 2023. 4, 6

[49] Heeseong Shin, Chaehyun Kim, Sunghwan Hong, Seokju Cho, Anurag Arnab, Paul Hongsuck Seo, and Seungryong Kim. Towards open-vocabulary semantic segmentation without semantic labels. *Advances in Neural Information Processing Systems*, 37:9153–9177, 2025. 2

[50] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020. 2, 5

[51] Vibashan VS, Shubhankar Borse, Hyojin Park, Debasmit Das, Vishal Patel, Munawar Hayat, and Fatih Porikli. Possam: Panoptic open-vocabulary segment anything. *arXiv preprint arXiv:2403.09620*, 2024. 4

[52] Haoxiang Wang, Pavan Kumar Anasosalu Vasu, Fartash Faghri, Raviteja Vemulapalli, Mehrdad Farajtabar, Sachin Mehta, Mohammad Rastegari, Oncel Tuzel, and Hadi Pouransari. Sam-clip: Merging vision foundation models towards semantic and spatial understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3635–3647, 2024. 2, 4

[53] Zhenyu Wang, Yali Li, and Shengjin Wang. Noisy boundaries: Lemon or lemonade for semi-supervised instance segmentation? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16826–16835, 2022. 2, 6, 7

[54] Zhenyu Wang, Yali Li, and Shengjin Wang. Noisy boundaries: Lemon or lemonade for semi-supervised instance segmentation? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16826–16835, 2022. 1, 6

[55] Philippe Weinzaepfel, Vincent Leroy, Thomas Lucas, Romain Brégier, Yohann Cabon, Vaibhav Arora, Leonid Antsfeld, Boris Chidlovskii, Gabriela Csurka, and Jérôme Revaud. Croco: Self-supervised pre-training for 3d vision tasks by cross-view completion. *Advances in Neural Information Processing Systems*, 35:3502–3516, 2022. 1

[56] Guoping Xu, Xiaoxue Qian, Hua-Chieh Shao, Jax Luo, Weiguo Lu, and You Zhang. A segment anything model-guided and match-based semi-supervised segmentation framework for medical imaging. *Medical physics*, 2025. 2

[57] Jinyu Yang, Mingqi Gao, Zhe Li, Shang Gao, Fangjing Wang, and Feng Zheng. Track anything: Segment anything meets videos. *arXiv preprint arXiv:2304.11968*, 2023. 2

[58] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019. 5

[59] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 5

[60] Renrui Zhang, Zhengkai Jiang, Ziyu Guo, Shilin Yan, Junting Pan, Xianzheng Ma, Hao Dong, Peng Gao, and Hongsheng Li. Personalize segment anything model with one shot. *arXiv preprint arXiv:2305.03048*, 2023. 1

[61] Yichi Zhang, Jin Yang, Yuchen Liu, Yuan Cheng, and Yuan Qi. Semisam: Enhancing semi-supervised medical image segmentation via sam-assisted consistency regularization. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 3982–3986. IEEE, 2024. 2

[62] Ziyu Zhang, Sanja Fidler, and Raquel Urtasun. Instance-level segmentation for autonomous driving with deep densely connected mrfs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1

[63] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *European Conference on Computer Vision*, pages 696–712. Springer, 2022. 2

[64] Juan Carlos Ángeles Cerón, Gilberto Ochoa Ruiz, Leonardo Chang, and Sharib Ali. Real-time instance segmentation of surgical instruments using attention and multi-scale feature fusion. *Medical Image Analysis*, 81:102569, 2022. 1