Segment Anything, Even Occluded

Supplementary Material

Appendix

This supplementary document provides additional experimental results and visualizations supporting our main paper.

- Section A presents visual examples from our collected amodal datasets.
- Section B illustrates qualitative comparisons between SAMEO and AISFormer [9].
- Section C shows the adaptation from modal to amodal segmentation compared to EfficientSAM [11].
- Section D extends our quantitative evaluation with class-specific metrics.
- Section E highlights the limitations of SAMEO and suggests potential directions for future research.

A. Amodal Dataset Visualization

Our collected amodal datasets, shown in Figure B, serve as essential training data for zero-shot amodal instance segmentation. Across ten diverse examples (COCOA [13], COCOA-cls [2], DYCE [1], KINS [7], MUVA [3], D2SA [2], KITTI-360-APS [5], MP3D-amodal [12], WALT [8], and pix2gestalt [6]), we display modal and amodal mask pairs. Our proposed Amodal-LVIS dataset features dual annotations of both occluded and unoccluded versions of each instance. This curated collection provides rich training signals that enable our model to learn generalizable amodal segmentation capabilities across different domains and object categories.

B. Qualitative Comparison

We compare SAMEO's amodal instance segmentation capabilities with state-of-the-art AISFormer on COCOA-cls (Figure C) and MUVA (Figure D) datasets. Using AISFormer's box predictions as prompts, SAMEO generates amodal masks for detected instances. The qualitative results demonstrate SAMEO's superior performance in mask boundary precision and occlusion estimation, particularly for complex shapes and instances with multiple overlaps. Our method significantly outperforms AISFormer in terms of overall mask quality.

C. Amodal Mask Adaptation

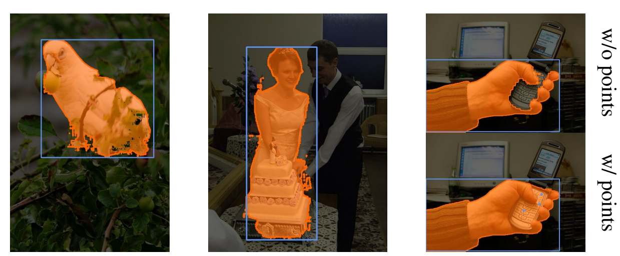
We demonstrate SAMEO's adaptation from modal to amodal segmentation through visualization experiments on the pix2gestalt dataset (Figure E). Comparing the modal mask predictions from the original EfficientSAM with SAMEO's amodal predictions and ground truth masks reveals successful adaptation to amodal segmentation. Our specialized training enables SAMEO to effectively estimate occluded regions while preserving the high-quality mask prediction and zero-shot capabilities inherent to the original model.

D. Class-specific Results

Table A and Table B present the class-specific AP/AR evaluations as a complement to class-agnostic results, following identical experimental settings from **??**. In both standard and zero-shot settings, SAMEO consistently improves the baseline models' performance. In standard evaluation, using RTMDet [4] as the front-end detector with SAMEO achieves the best performance on COCOA-cls, while using ConvNeXt-V2 [10] as the front-end detector with SAMEO leads on D2SA. For zero-shot settings, using CO-DETR [14] as the front-end detector with SAMEO shows strong results on both COCOA-cls and D2SA, indicating SAMEO's effectiveness generalizes well across both class-specific and class-agnostic scenarios.

E. Limitation and Future Work

Although SAMEO notably outperforms SOTA methods in both scores and quality, it still faces challenges with difficult cases, as shown in Figure A: incomplete amodal masks (a), rough edges (b), and unexpected modal outputs (top of (c)). When multiple objects overlap, using box prompts alone can cause model confusion. Additional experiments with using both box and point prompts show promising results in enhancing target region predictions (bottom of (c)). We believe exploring different prompt types is a direction for future work.



(a) (b) (c)

Figure A. Failure cases of SAMEO: (a) incomplete amodal masks, (b) rough edges, and (c) unexpected modal output.

References

[1] Kiana Ehsani, Roozbeh Mottaghi, and Ali Farhadi. Segan: Segmenting and generating the invisible. In *2018 IEEE Con-*

Figure B. Visualization of collected amodal datasets. For Amodal-LVIS, each instance has unoccluded (left) and occluded (right) versions.

Model	COCOA-cls				D2SA			
	AP	**AP$_{50}$**	**AP$_{75}$**	**AR**	**AP**	**AP$_{50}$**	**AP$_{75}$**	**AR**
AISFormer [9]	35.5	58.0	37.6	49.6	62.9	83.4	68.3	72.0
RTMDet* [4]	50.4	68.1	55.4	69.9	53.9	71.9	57.3	75.8
ConvNeXt-V2* [10]	46.9	64.1	51.3	70.7	60.7	81.3	63.5	74.3
AISFormer+SAMEO	46.4	62.1	50.5	62.8	72.2	84.3	**76.6**	79.2
RTMDet*+SAMEO	**54.4**	**71.4**	**59.6**	**73.5**	62.1	72.7	65.8	75.7
ConvNeXt-V2*+SAMEO	53.7	69.8	58.8	72.8	**72.9**	**84.9**	76.5	**83.3**

Table A. Class-specific performance on COCOA-cls and D2SA datasets. * denotes modal object detectors that provide modal bounding boxes as prompts. Bold numbers indicate the best performance.

Model	COCOA-cls				D2SA			
	AP	**AP$_{50}$**	**AP$_{75}$**	**AR**	**AP**	**AP$_{50}$**	**AP$_{75}$**	**AR**
AISFormer	35.5	58.0	37.6	49.6	62.9	83.4	68.3	72.0
AISFormer+EfficientSAM†	42.0	59.2	45.3	59.3	62.7	80.5	64.9	72.4
RTMDet*+EfficientSAM†	48.7	67.5	53.3	65.8	55.9	72.4	57.4	77.3
AISFormer+SAMEO†	45.2	61.5	49.8	61.4	66.9	81.7	70.5	74.6
RTMDet*+SAMEO†	53.3	70.6	59.2	72.5	60.2	74.1	62.6	**81.0**
CO-DETR* [14]+SAMEO†	**53.6**	**70.6**	**59.5**	**73.3**	**72.2**	**87.7**	**74.6**	79.4

Table B. Zero-shot class-specific performance on COCOA-cls and D2SA datasets. † indicates zero-shot evaluation without training on the test dataset. * denotes modal object detectors that provide modal bounding boxes as prompts. Bold numbers indicate the best performance.

ference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, pages 6144–6153. Computer Vision Foundation / IEEE Computer Society, 2018. I

[2] Patrick Follmann, Rebecca König, Philipp Härtinger, Michael Klostermann, and Tobias Böttger. Learning to see the invisible: End-to-end trainable amodal instance segmentation. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2019, Waikoloa Village, HI, USA, January 7-11, 2019*, pages 1328–1336. IEEE, 2019. I

[3] Zhixuan Li, Weining Ye, Juan Terven, Zachary Bennett, Ying Zheng, Tingting Jiang, and Tiejun Huang. MUVA: A new large-scale benchmark for multi-view amodal instance segmentation in the shopping scenario. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 23447–23456. IEEE, 2023. I

[4] Chengqi Lyu, Wenwei Zhang, Haian Huang, Yue Zhou, Yudong Wang, Yanyi Liu, Shilong Zhang, and Kai Chen. Rtmdet: An empirical study of designing real-time object detectors. *CoRR*, abs/2212.07784, 2022. I, III

[5] Rohit Mohan and Abhinav Valada. Amodal panoptic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 20991–21000. IEEE, 2022. I

[6] Ege Ozguroglu, Ruoshi Liu, Dídac Surís, Dian Chen, Achal Dave, Pavel Tokmakov, and Carl Vondrick. pix2gestalt: Amodal segmentation by synthesizing wholes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 3931–3940. IEEE, 2024. I

[7] Lu Qi, Li Jiang, Shu Liu, Xiaoyong Shen, and Jiaya Jia. Amodal instance segmentation with KINS dataset. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 3014–3023. Computer Vision Foundation / IEEE, 2019. I

[8] N. Dinesh Reddy, Robert Tamburo, and Srinivasa G. Narasimhan. WALT: watch and learn 2d amodal representation from time-lapse imagery. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 9346–9356. IEEE, 2022. I

[9] Minh Q. Tran, Khoa Vo, Kashu Yamazaki, Arthur A. F. Fernandes, Michael Kidd, and Ngan Le. Aisformer: Amodal instance segmentation with transformer. In *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022*, page 712. BMVA Press, 2022. I, III

[10] Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. Convnext V2: co-designing and scaling convnets with masked autoencoders. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 16133–16142. IEEE, 2023. I, III

[11] Yunyang Xiong, Bala Varadarajan, Lemeng Wu, Xiaoyu Xiang, Fanyi Xiao, Chenchen Zhu, Xiaoliang Dai, Dilin Wang, Fei Sun, Forrest N. Iandola, Raghuraman Krishnamoorthi, and Vikas Chandra. Efficientsam: Leveraged masked im-

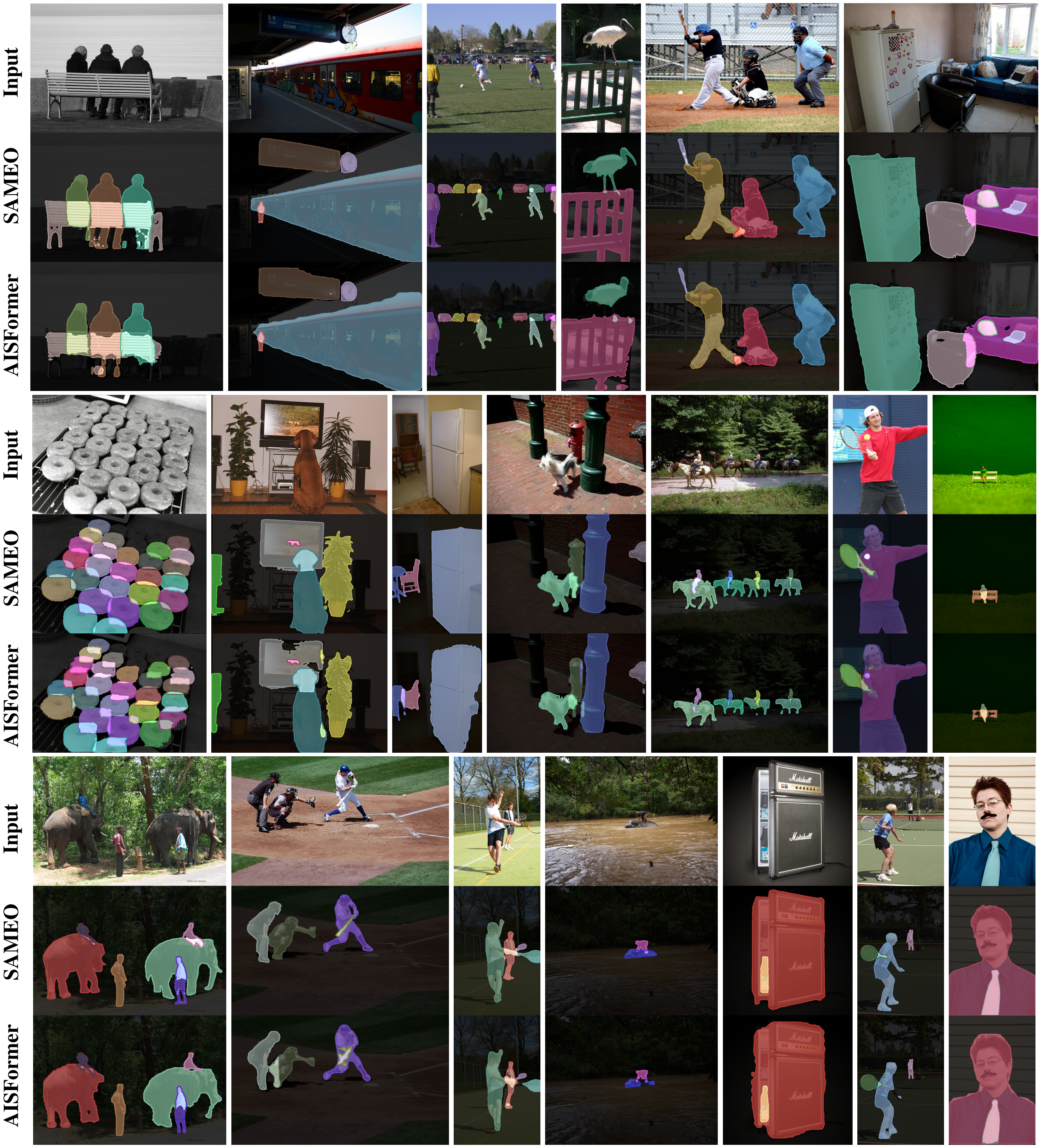


Figure C. Qualitative comparison of amodal instance segmentation on COCOA-cls dataset. Each row shows: *i*) input RGB image, *ii*) SAMEO's amodal prediction using AISFormer boxes as prompts, and *iii*) AISFormer's prediction. SAMEO demonstrates superior mask boundary delineation and more accurate occluded region estimation compared to the baseline.

age pretraining for efficient segment anything. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 16111–16121. IEEE, 2024. I

[12] Guanqi Zhan, Chuanxia Zheng, Weidi Xie, and Andrew Zisserman. Amodal ground truth and completion in the wild.

Figure D. Qualitative comparison of amodal instance segmentation on MUVA dataset. Each row displays: *i*) input RGB image, *ii*) amodal masks predicted by SAMEO with AISFormer box prompts, and *iii*) AISFormer predictions. Our approach yields more precise boundaries and better handles occlusion estimation.

In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 28003–28013. IEEE, 2024. I

[13] Yan Zhu, Yuandong Tian, Dimitris N. Metaxas, and Piotr Dollár. Semantic amodal segmentation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 3001–3009. IEEE Computer Society, 2017. I

[14] Zhuofan Zong, Guanglu Song, and Yu Liu. Detrs with collaborative hybrid assignments training. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 6725–6735. IEEE, 2023. I, III

Figure E. From modal to amodal segmentation on pix2gestalt dataset. Each row demonstrates: *i*) input RGB image, *ii*) modal mask prediction from the original EfficientSAM, *iii*) amodal mask prediction from our SAMEO, *iv*) ground truth amodal mask. The results showcase SAMEO's successful adaptation to amodal segmentation while maintaining zero-shot capabilities.