Supplementary Material for ZIM: Zero-Shot Image Matting for Anything

Beomyoung Kim Chanyong Shin Joonhyun Jeong Hyungsik Jung
Se-Yun Lee Sewhan Chun Dong-Hyun Hwang Joonsang Yu
NAVER Cloud, ImageVision

Figure S1. **Visualization of samples** from the MicroMat-3K test set, providing a high-quality benchmark for zero-shot matting models.

A. MicroMat-3K Details

Inspired by the data engine procedure in the SAM [23], we constructed the MicroMat-3K test set. The dataset construction involved four key steps: (1) We collected high-resolution images from the DIV2K dataset [1], which is originally intended for super-resolution tasks. (2) We generated pseudo-segmentation labels using automatic mask generation of SAM, powered by a large backbone model to ensure high-quality segmentation. (3) We transformed these segmentation labels into matte labels using our label converter, which is also powered by a large backbone network. This step provided an initial set of pseudo-matte labels for each image. (4) Our annotators inspected the pseudo-matte labels. High-quality labels were directly retained as ground-truth annotations, while any low-quality matte labels were manually revised to ensure high-fidelity ground-truth labels. Additionally, we categorized the final matte labels into two classes: fine-grained masks (750 samples) and coarse-grained masks (2250 samples). Figure S1 showcases visualization examples from MicroMat-3K, which provides diverse and high-quality micro-level matte labels.

B. Downstream Task

B.1. Zero-Shot Image Matting

To showcase the generalizability of ZIM, we evaluate it across 23 diverse datasets, including ADE20K [56], BBBC038v1 [6], Cityspcaes [12], DOORS [35], EgoHOS [55], DRAM [11], GTEA [17, 28], Hypersim [40], IBD [8], iShape [48], COCO [29], NDD20 [46], NDISPark [9, 10], OVIS [36], PIDRay [47], Plittersdorf [20], PPDLS [33], STREETS [42], TimberSeg [18], TrashCan [21], VISOR [13, 14], WoodScape [52], and Zero Waste-f [4]. Using the `Automatic Mask Generation` strategy introduced by SAM [23], we apply a regular grid of point prompts to each image and perform post-processing with thresholding and non-maximum suppression (NMS) to generate the final matting masks. Figure S2 and Figure S3 show that ZIM produces high-quality *matte anything* results for all datasets with considerably detailed matte quality and powerful generalization capability. Although SAM shows powerful generalization capability, the output mask is the coarse quality. In addition, existing interactive matting methods ($i.e.$, Matte-Any [49], Matting-Any [27], and SMat [50]) often fail to generalize the unseen data.

B.2. Image Inpainting

Image inpainting is an important application in generative AI, where precise mask generation plays a critical role in removing or reconstructing parts of an image. Following the Inpaint Anything framework [54], we guide SAM and ZIM masks to the inpainting model [45]. As shown in Figure S4, ZIM produces more accurate object masks than SAM, leading to significantly better inpainting results. For complex objects like flowerpots and hair, SAM's coarse masks fail to remove the object cleanly, leaving noticeable artifacts. In contrast, our precise matting masks enable the inpainting model to smoothly remove objects without artifacts. To provide quantitative analysis, we use CLIP Distance and CLIP Accuracy [16, 51] as evaluation metrics. CLIP Distance measures the similarity between the source and inpainted regions, where a larger distance indicates better removal. CLIP Accuracy evaluates the change in class predictions after removal, considering the task successful if the original class is absent from Top-1/3/5 predictions. Following [51], we use the text prompt `a photo of a {category name}`. Table S1 shows that ZIM outperforms SAM on both metrics by better preserving surrounding context after inpainting through enhanced mask quality.

B.3. 3D Object Segmentation with NeRF

The quality of the segmentation mask is crucial when converting a 2D mask into a 3D representation. In this work, we adopt the SA3D framework [7], which utilizes SAM to segment 3D objects from 2D masks by manually prompting the target object in a single view. By replacing SAM with ZIM in the 2D mask segmentation process, we significantly improve the quality of the resulting 3D objects. Figure S5 presents the qualitative results of segmented 3D objects guided by the SAM and ZIM models on the LLFF-trex and LLFF-horns [32] dataset. Compared to SAM, which often misses finer details due to its coarse-level mask generation, ZIM captures more intricate object features. These findings demonstrate that the precise matting capabilities of ZIM extend beyond 2D tasks, significantly enhancing the quality of 3D object segmentation. For quantitative evaluation, we employ the NVOS [39] dataset, which includes finely annotated 2D masks. Since the SA3D framework relies on binary masks for projecting 2D masks into 3D space, we binarized the ZIM results using a threshold of 0.3. We fixed the number of self-prompting points at 10 and followed the experimental setup of SA3D [7] for pre-trained NeRFs and manual prompts. As shown in Table S2, ZIM outperforms SAM in quantitative 3D segmentation results on the target view, reporting higher mask IoU scores. These findings demonstrate that the precise matting capabilities of ZIM extend beyond 2D tasks, significantly enhancing the quality of 3D object segmentation.

Mask	**CLIP Dist ↑**	**CLIP Acc ↑**		
		Top-1	Top-3	Top-5
COCO GT [29]	67.07	0.7767	0.6236	0.5415
SAM [23]	68.17	0.7940	0.6521	0.5753
ZIM (ours)	**73.11**	**0.8616**	**0.7543**	**0.6855**

Table S1. **Quantitative results of image inpainting** using the Inpainting Anything framework [54]. The inpainting model takes three types of input masks (COCO ground-truth [29], SAM [23], and ZIM) and we evaluate the corresponding inpainting results using CLIP distance and accuracy metrics [16, 51].

Scenes	**Mask IoU (%) ↑**	
	SAM	**ZIM**
Fern	84.9	**86.6**
Flower	94.0	**95.7**
Fortress	96.5	**98.1**
Horns-center	94.0	**97.4**
Horns-left	92.9	**94.7**
Leaves	92.2	**92.9**
Orchids	89.9	**91.8**
Trex	83.7	**85.4**

Table S2. **Quantitative results** of mask IoU scores on the target view for the NVOS dataset [39].

Figure S2. **Qualitative samples of automatic mask generation results** on (1) ADE20K [56], (2) BBBC038v1 [6], (3) Cityscapes [12], (4) DOORS [35], (5) EgoHOS [55], (6) DRAM [11], (7) GTEA [17, 28], (8) Hypersim [40], (9) IBD [8], (10) iShape [48], and (11) COCO [29] datasets.

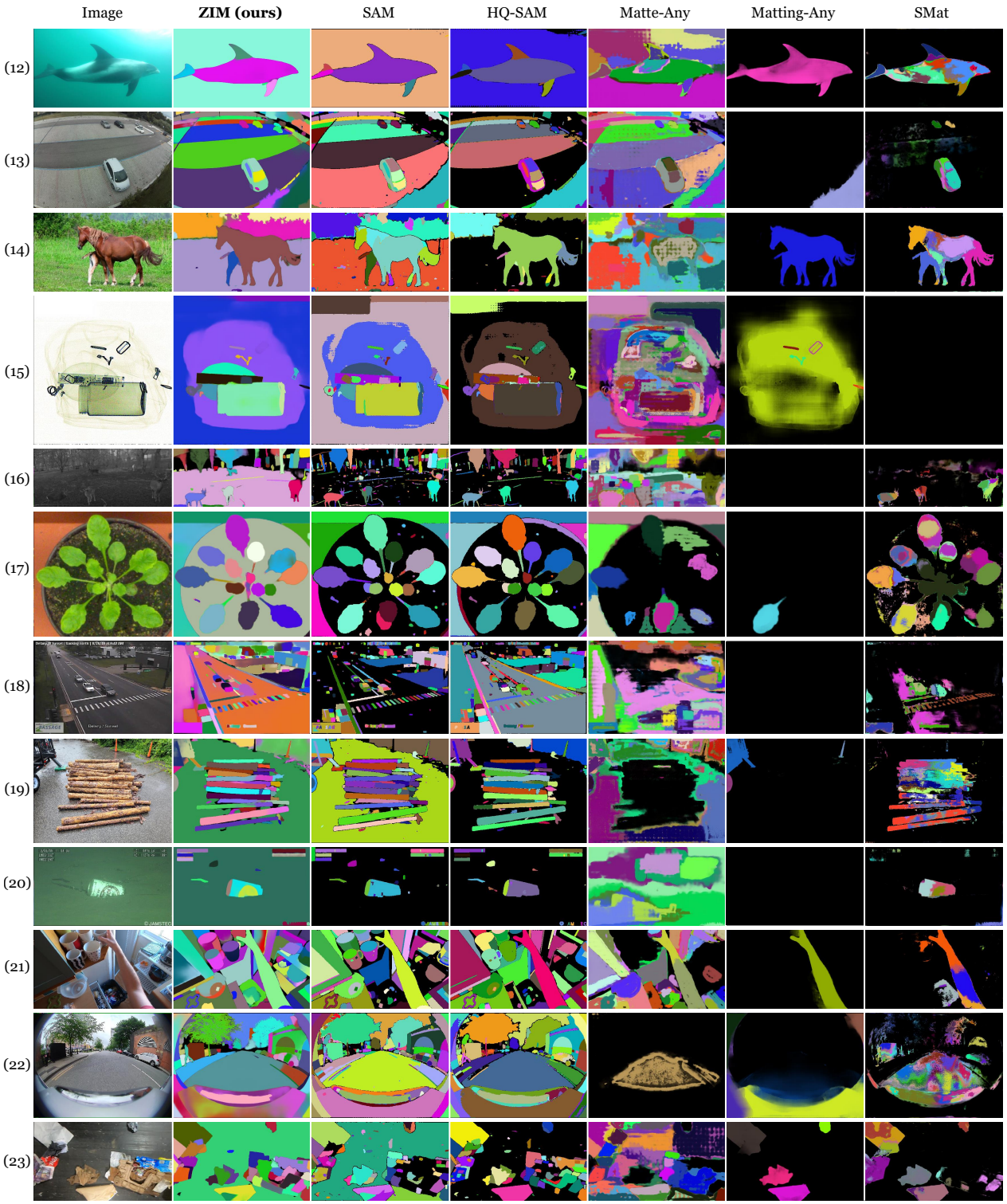


Figure S3. (continue) **Qualitative samples of automatic mask generation results** on (12) NDD20 [46], (13) NDISPark [9, 10], (14) OVIS [36], (15) PIDRay [47], (16) Plittersdorf [20], (17) PPDLS [33], (18) STREETS [42], (19) TimberSeg [18], (20) TrashCan [21], (21) VISOR [13, 14], (22) WoodScape [52], and (23) ZeroWaste-f [4] datasets.

Figure S4. **Qualitative results** of three kinds of input masks (*i.e.*, COCO ground-truth [29], SAM [23], and ZIM) along with their corresponding image inpainting results using the Inpainting Anything framework [54].

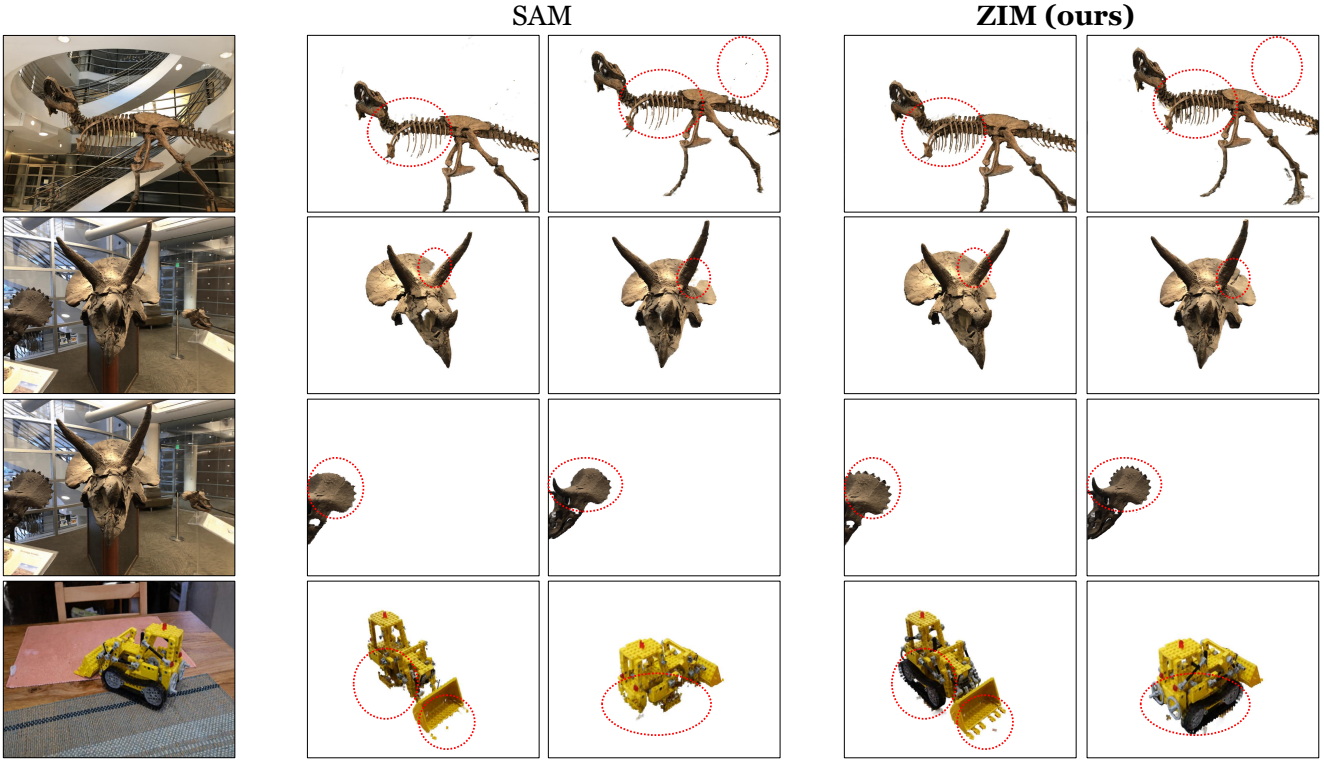


Figure S5. **Qualitative samples** of 3D object segmentation results guided by SAM [23] and ZIM models within the SA3D framework [7] for the LLFF-trex, LLFF-horns [32], and 360°-kitchen (Lego) [3] datasets.

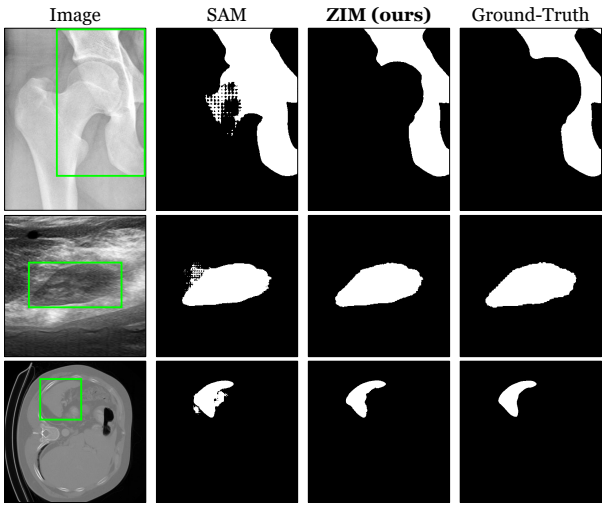


Figure S6. **Qualitative samples** of SAM and ZIM output masks on the medical image datasets [2, 19, 34, 44] using the box prompt.

B.4. Medical Image Segmentation

Segmentation models are essential in medical image analysis, where they assist in identifying key anatomical structures and abnormalities. Building on the recent evaluation of SAM's performance in medical imaging by [31], we explore the applicability of ZIM for zero-shot medical image segmentation. Given that neither SAM nor ZIM has been trained on medical image datasets, this experiment focuses on evaluating their zero-shot segmentation capabilities. Both SAM and ZIM, using the ViT-B backbone, are evaluated across five medical imaging datasets: the hippocampus and spleen datasets from the Medical Image Decathlon [2], an ultrasonic kidney dataset [44], an ultrasonic nerve dataset [34], and an X-ray hip dataset [19]. Since these datasets comprise binary ground-truth masks, we apply a threshold of 0.3 to the matte output of ZIM. Following the evaluation protocol from [31], we employ five prompt modes: (1) a single point at the center of the largest contiguous region, (2) multiple points centered on up to three regions, (3) a box surrounding the largest region, (4) multiple boxes around up to three regions, and (5) a box encompassing the entire object.

Figure S7 illustrates the distribution of IoU scores across the five datasets for each prompt mode. These results show that ZIM consistently outperforms SAM, particularly in point-based prompts (modes 1 and 2). In addition, as shown in Figure S6, SAM frequently exhibits checkerboard artifacts when dealing with the indistinct visual details of medical images. In contrast, ZIM produces more robust and precise segmentation masks due to our advanced pixel decoder. This demonstrates ZIM's strong generalization on unseen and complex medical image data, highlighting its superior zero-shot capabilities compared to SAM.

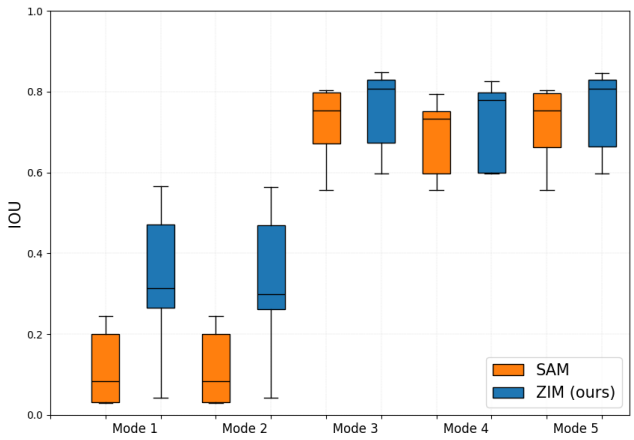


Figure S7. **Mask IoU distribution** across the five medical image analysis datasets [2, 19, 34, 44] for the five prompt modes.

C. Additional Experiemental Results

C.1. Detailed Experiment Settings

We evaluate interactive matting performance using point and box prompts derived from ground-truth masks. For point sampling, we adopt RITM [43] with a maximum of 12 points. Box prompts are extracted from the min-max foreground coordinates, with up to 10% random perturbation to assess robustness to noisy box inputs. This sampling strategy is applied to MicroMat3K and public matting datasets such as AIM-500 [25] and P3M-500-NP [24].

ZIM is trained on 1% of SA1B-Matte, containing 2.2M matte labels. Since SAM pre-trained weights are utilized, increasing the training data to 10% did not yield noticeable improvements, indicating that 1% suffices for enhancing fine-grained representations.

All experiments are conducted using 8 V100 GPUs. Training the label converter and ZIM takes approximately 7 and 5 days, respectively.

C.2. Detailed Quantitative Comparison

Table S3 provides an in-depth evaluation of zero-shot matting models, with additional experiments utilizing various backbone networks: ViT-L and ViT-H [15] for SAM [23], HQ-SAM [22], and Matte-Any [49], and Hiera-L [41] for SAM2 [37]. Furthermore, we investigate model behavior based on the size of the target object by defining three object size groups, according to the ratio of the foreground region in the image: small (ratio $< 1\%$), medium ($1\% \leq$ ratio $< 10\%$), and large (ratio $\geq 10\%$). The MSE error is reported for each object size group, offering a detailed understanding of model performance across varying object sizes. Additionally, we measure the model throughput using an NVIDIA V100 GPU to assess computational efficiency.

Method	Backbone	Latency (ms)↓	Prompt	Fine-grained↓ MSE	MSE$_S$	MSE$_M$	MSE$_L$	Coarse-grained↓ MSE	MSE$_S$	MSE$_M$	MSE$_L$
SAM [23]	ViT-B [15]	172.5	point	21.651	2.717	7.145	70.502	5.569	4.092	5.405	77.761
			box	11.057	0.329	2.983	38.501	1.044	0.181	2.456	24.519
	ViT-L [15]	361.6	point	15.663	3.312	5.165	49.202	4.293	3.640	3.389	46.278
			box	7.989	0.320	2.423	27.276	0.534	**0.145**	1.606	5.575
	ViT-H [15]	605.2	point	14.534	3.619	5.753	43.371	2.100	**0.653**	3.278	56.086
			box	6.188	0.281	1.687	21.389	0.468	0.152	1.334	4.610
SAM2 [37]	Hiera-B+ [41]	147.4	point	25.296	15.657	15.970	53.346	14.794	14.786	12.128	49.058
			box	12.024	0.322	2.155	43.670	0.613	0.152	1.600	10.194
	Hiera-L [41]	195.0	point	16.937	0.871	5.613	56.702	2.572	0.954	4.308	57.804
			box	10.616	0.263	1.888	38.636	0.704	0.149	1.424	18.019
HQ-SAM [22]	ViT-B [15]	177.2	point	36.674	5.094	9.356	123.199	6.457	3.602	9.596	102.834
			box	42.457	0.392	4.369	160.456	2.733	0.230	2.521	124.372
	ViT-L [15]	368.1	point	20.481	1.928	6.496	67.979	3.046	1.200	4.987	66.373
			box	19.881	0.326	2.683	73.913	0.762	0.163	1.408	21.129
	ViT-H [15]	608.1	point	22.547	5.308	6.601	71.446	3.599	1.659	5.034	77.643
			box	23.743	0.263	2.538	89.518	0.789	0.164	1.266	24.469
Matte-Any [49]	ViT-B [15]	668.5	point	20.844	3.381	7.953	65.108	6.053	4.739	5.897	70.443
			box	9.746	0.918	3.856	31.109	1.983	1.235	3.039	24.426
	ViT-L [15]	814.1	point	15.230	3.985	6.243	44.842	5.116	4.570	3.996	44.606
			box	7.323	0.975	3.692	21.711	1.597	1.222	2.418	9.119
	ViT-H [15]	1036.9	point	14.119	4.270	7.085	38.704	3.022	1.669	3.913	56.044
			box	6.048	0.917	2.992	17.872	1.571	1.228	2.294	8.836
Matting-Any [27]	ViT-B [15]	200.3	point	77.335	27.256	41.349	202.692	36.187	31.549	40.993	197.340
			box	68.372	18.286	33.111	192.566	23.780	17.344	36.194	175.418
SMat [50]	ViT-B [15]	263.4	point	123.664	31.549	40.993	197.340	59.113	52.305	68.709	250.587
			box	133.515	17.344	36.194	175.418	61.157	53.280	74.099	261.544
ZIM (ours)	ViT-B [15]	187.8	point	8.213	0.870	3.962	24.934	1.788	1.444	1.731	**18.983**
			box	1.893	0.205	2.228	3.617	0.448	0.200	**1.382**	**0.632**
	ViT-L [15]	373.4	point	**5.825**	**0.563**	**2.724**	**17.898**	**1.719**	**1.041**	**1.574**	35.687
			box	**1.589**	**0.191**	**1.888**	**2.982**	**0.446**	**0.175**	1.458	0.686

Table S3. **Detailed Quantitative comparison** of our ZIM model and six existing methods on the MicroMat-3K dataset. Results are presented for different backbone networks, model throughput, and MSE scores across object sizes (small, medium, and large). The latency is measured on the NVIDIA V100 GPU.

The results in Table S3 provide some meaningful insights: (1) ZIM consistently outperforms SAM, especially for larger objects, as indicated by the MSE_L metric. This improvement is likely due to the reduction of checkerboard artifacts, a known issue in SAM's pixel decoder, which our advanced decoder addresses effectively, as evidenced in Figures 1, S2, and S3. (2) ZIM demonstrates highly competitive results even with the smaller ViT-B backbone, outperforming models like SAM and HQ-SAM with larger backbones such as ViT-H. Additionally, the performance of ZIM with the ViT-L backbone suggests that further improvements could be achieved with more powerful architectures. (3) Despite our advanced decoder, ZIM introduces only a marginal increase in latency (just 10ms more than SAM) making it a lightweight and efficient option for zero-shot matting tasks. (4) Compared to existing matting models (*e.g.*, Matte-Any, Matting-Any, and SMat), ZIM delivers superior performance while maintaining efficiency.

C.3. Discussion on Transparency Prediction

In the image matting task, there is a distinct scenario, that is, predicting the transparency of objects, such as glasses and fire. Due to its inherent difficulties, existing matting methods [5, 30, 53] typically rely on curated transparency datasets [5] containing close-up transparent object images with clear backgrounds, making them effective in constrained environments but less adaptable to open-world scenarios. ZIM, on the other hand, is designed for fine-grained mask representation in general object segmentation across diverse and complex scenes. However, predicting

Model	Input	Transparent-460 [5] MSE ↓	SAD ↓
IndexNet [30]	Trimap	112.53	573.09
MGMatting [53]	Trimap	6.33	111.92
TransMatting [5]	Trimap	4.02	88.34
ZIM (ours)	Box	15.55	298.78

Table S4. **Quantitative results on the transparent object matting dataset, Transprent-460 [5].**

transparency within ZIM remains a challenge due to the lack of large-scale open-world transparency datasets and the inherent complexity of identifying transparent regions in natural images.

To investigate ZIM's adaptability to transparency prediction, we fine-tune it on the Transparent-460 dataset [5] and compare its performance against specialized transparency matting methods [5, 30, 53], as shown in Table 4. While these methods leverage trimaps to provide explicit spatial guidance for transparency estimation, ZIM relies only on sparse box prompts, which offer less precise object boundary information. Despite this limitation, ZIM achieves reasonable results, demonstrating its strong transferability.

C.4. Expanding Prompt Sources

Interactive models, such as SAM, commonly support only point and box prompts. Here, we demonstrate the potential ZIM offers a more flexible approach by expanding the variety of prompt types, including text and scribble prompts.

Text Prompt. To enable text prompts, we integrate ZIM with the Grounded-SAM framework [38]. Grounded-SAM uses a grounding object detection model that processes an image-text pair and returns bounding boxes for objects mentioned in the text. ZIM then uses these bounding boxes as prompts to produce detailed matte outputs. We refer to this combined model as Grounded-ZIM. As shown in Figure S8, Grounded-ZIM provides high-quality outputs with a simple text prompting pipeline, offering more precise and robust mask generation than Grounded-SAM.

Scribble Prompt. In addition, ZIM can support scribble prompts, which provide users with an intuitive way to mark regions of interest. We implement this functionality by sampling points along the scribble path. To ensure comprehensive coverage of the scribble region, we employ uniform sampling with setting the maximum number of sampled points to 24. It allows ZIM to effectively handle the scribble input and generate high-quality matte outputs. Figure S9 shows an example of how the scribble prompt leads to accurate and stable results. These expansions highlight the versatility of ZIM in accommodating diverse input prompts.

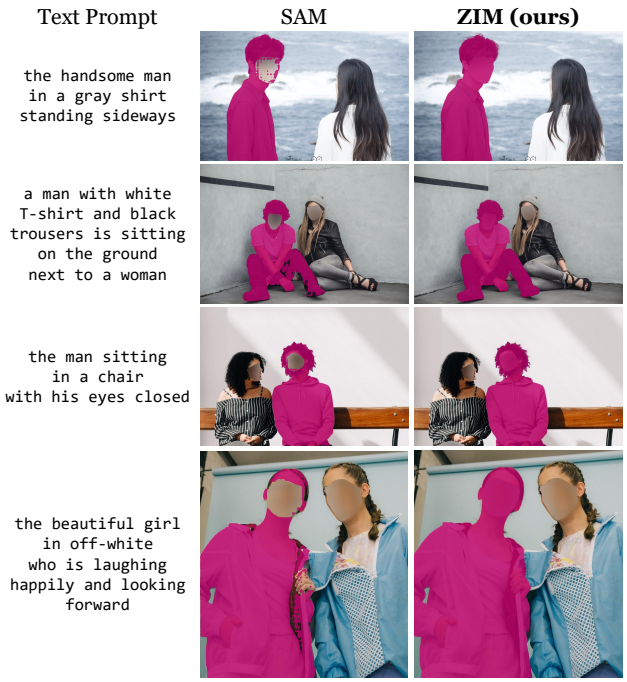


Figure S8. **Qualitative samples** of text prompting results on the RefMatte-RW100 [26] dataset.

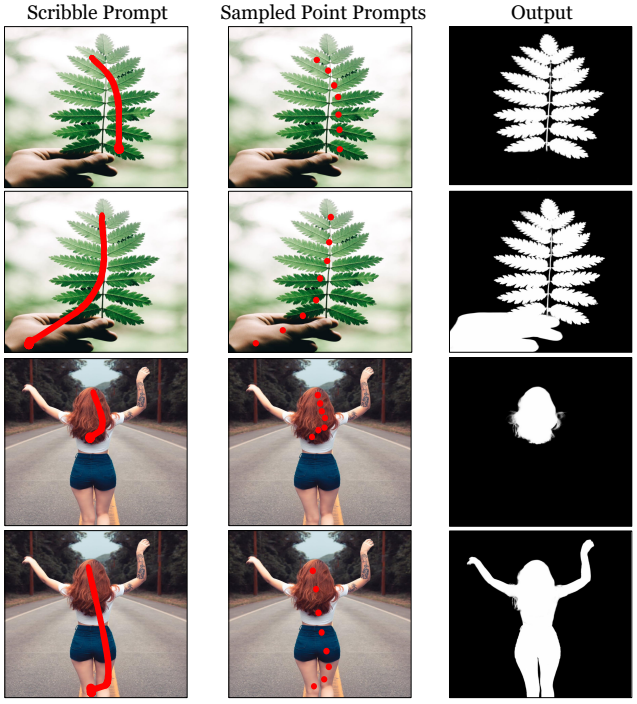


Figure S9. **Qualitative examples** of scribble prompting results.

Model	MSE↓
SAM [23]	11.05
SAM [23] + Converter	7.34
SAM [23] + SA1B-Matte	2.71
HQSAM [22]	42.45
HQSAM [22] + SA1B-Matte	13.35
Matting-Any [27]	68.37
Matting-Any [27] + SA1B-Matte	12.26
ZIM (ours)	**1.89**

Table S5. **Analysis** of our proposed components (*i.e.*, Label Converter and SA1B-Matte dataset) when applied to existing models.

C.5. Additional Ablation Study

Impact of our proposed components on existing segmentation models. Table S5 demonstrates that both the Label Converter and SA1B-Matte training data provide substantial improvements to baseline performance on the MicroMat3K fine-grained dataset. The Label Converter, when applied to SAM's coarse output masks, reduces MSE from 11.05 to 7.34, representing a significant improvement in fine-grained mask quality. However, the converter cannot address inherent limitations in SAM's base predictions, such as checkerboard artifacts, resulting in performance that remains below ZIM's 1.89 MSE.

In addition, training existing models (*i.e.*, SAM, HQ-SAM, and Matting-Any) on our SA1B-Matte dataset yields consistent improvements across all models. SAM trained with SA1B-Matte achieves 2.71 MSE, while HQSAM and Matting-Any reach 13.35 and 12.26 MSE, respectively. Despite these improvements, all existing models underperform compared to ZIM, highlighting the importance of our architectural innovations, including the Hierarchical Pixel Decoder and Prompt-Aware Masked Attention. The performance gap is particularly notable for HQ-SAM and Matting-Any, which rely on the frozen SAM model that limits their capacity to learn fine-grained representations during training.

Effect of Hyperparameter σ. The hyperparameter σ controls the standard deviation of the 2D Gaussian map used to create the soft attention mask for point prompts. A larger σ results in a wider spread of the Gaussian, covering a broader region around the point. Table S6a presents the performance of the ZIM model using point prompts on the MicroMat-3K dataset for varying values of σ. Through experimentation, we found that setting σ to 21 strikes a balance by generating an appropriately sized soft attention mask that effectively captures relevant features while minimizing unnecessary coverage.

σ	Fine-grained			Coarse-grained		
	SAD↓	MSE↓	Grad↓	SAD↓	MSE↓	Grad↓
11	31.476	8.381	5.505	6.741	1.816	1.538
21	**31.286**	**8.213**	**5.324**	**6.645**	**1.788**	**1.469**
41	31.341	8.298	5.476	6.686	1.807	1.493

(a)

λ	Fine-grained			Coarse-grained		
	SAD↓	MSE↓	Grad↓	SAD↓	MSE↓	Grad↓
5	33.086	9.056	7.392	6.966	1.970	1.817
10	**31.286**	**8.213**	**5.324**	**6.645**	**1.788**	**1.469**
20	31.402	8.295	5.356	6.712	1.838	1.475

(b)

Table S6. **Analysis** of ZIM using point prompt evaluations: (a) Effect of the hyperpameter σ. (b) Effect of the hyperpameter λ.

Effect of Hyperparameter λ. The hyperparameter λ, as defined in Eq (1), controls the weight assigned to the Gradient loss, influencing the emphasis on edge detail during training. We evaluate how different values of λ affect the performance of the ZIM on the MicroMat-3K dataset. The results in Table S6b indicate that a λ value of 10 achieves optimal performance, providing a balanced trade-off between smoothness and edge accuracy.

C.6. Additional Qualitative Samples

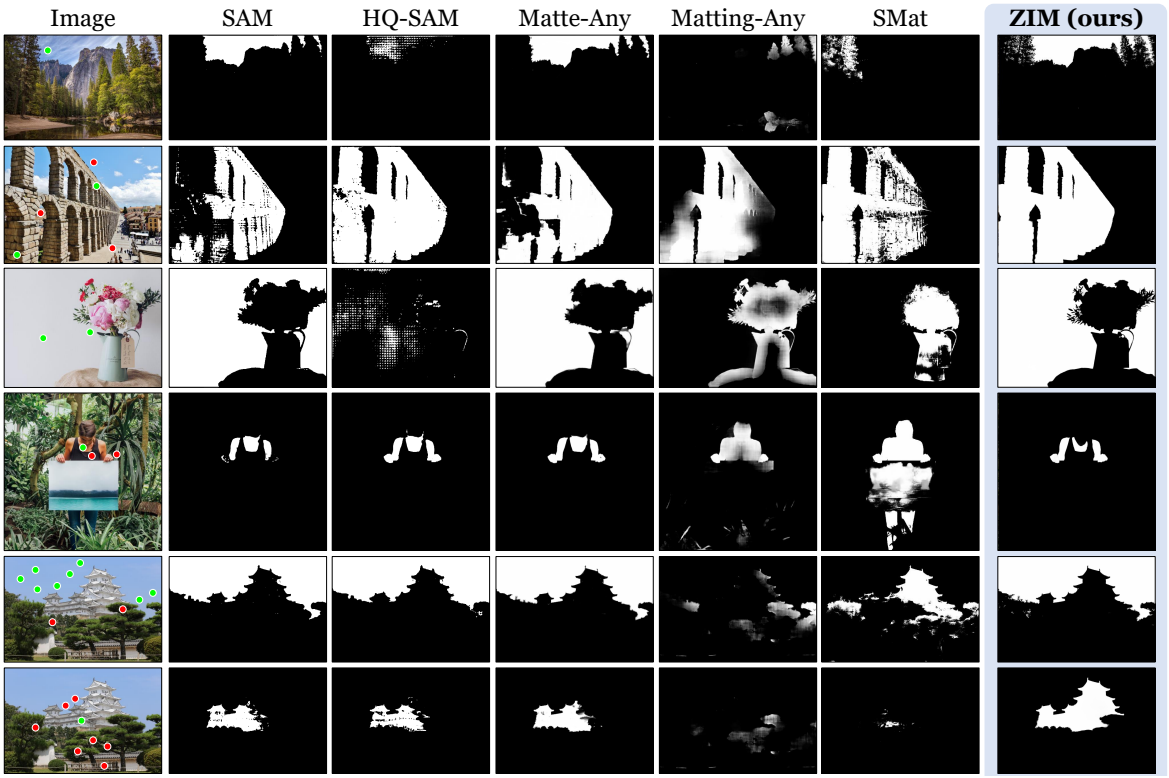
We provide more qualitative samples for the SA1B-Matte dataset (Figure S10) and ZIM output mattes (Figure S11).

References

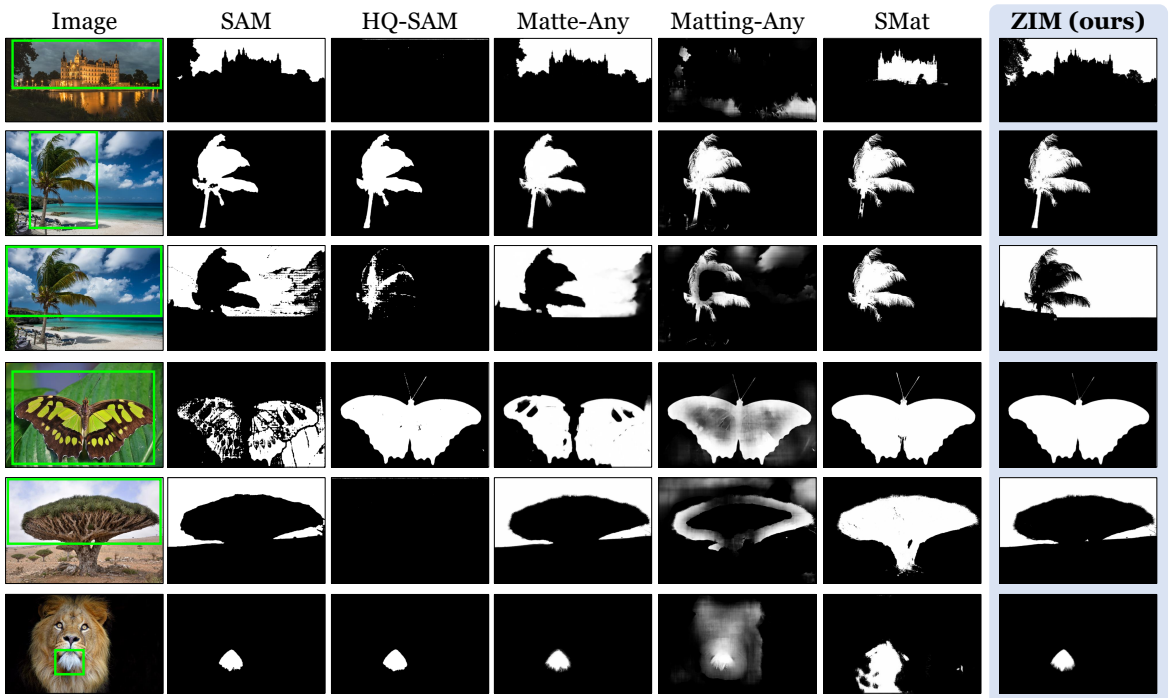
[1] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 126–135, 2017. 1

[2] Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A. Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M. Summers, Bram van Ginneken, Michel Bilello, Patrick Bilic, Patrick F. Christ, Richard K. G. Do, Marc J. Gollub, Stephan H. Heckers, Henkjan Huisman, William R. Jarnagin, Maureen K. McHugo, Sandy Napel, Jennifer S. Golia Pernicka, Kawal Rhode, Catalina Tobon-Gomez, Eugene Vorontsov, James A. Meakin, Sebastien Ourselin, Manuel Wiesenfarth, Pablo Arbeláez, Byeonguk Bae, Sihong Chen, Laura Daza, Jianjiang Feng, Baochun He, Fabian Isensee, Yuanfeng Ji, Fucang Jia, Ildoo Kim, Klaus Maier-Hein, Dorit Merhof, Akshay Pai, Beomhee Park, Mathias Perslev, Ramin Rezaiifar, Oliver Rippel, Ignacio Sarasua, Wei Shen, Jaemin Son, Christian Wachinger, Liansheng Wang, Yan Wang, Yingda Xia, Daguang Xu, Zhanwei Xu, Yefeng Zheng, Amber L. Simpson, Lena Maier-Hein, and M. Jorge Cardoso. The medical segmentation decathlon. *Nature Communications*, 13(1), 2022. 6

Figure S10. **Additional qualitative samples** of the SA1B dataset [23] with micro-level coarse labels and our SA1B-Matte dataset with micro-level fine labels.



(a)



(b)

Figure S11. **Additional qualitative samples** of ZIM with five existing zero-shot models (SAM [23], HQ-SAM [22], Matte-Any [49], Matting-Any [27], and SMat [50]) based on (a) point prompts and (b) box prompts.

[3] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. *CVPR*, 2022. 5

[4] Dina Bashkirova, Mohamed Abdelfattah, Ziliang Zhu, James Akl, Fadi Alladkani, Ping Hu, Vitaly Ablavsky, Berk Calli, Sarah Adel Bargal, and Kate Saenko. Zerowaste dataset: Towards deformable object segmentation in cluttered scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21147–21157, 2022. 2, 4

[5] Huanqia Cai, Fanglei Xue, Lele Xu, and Lili Guo. Transmatting: Enhancing transparent objects matting with transformers. In *European conference on computer vision*, pages 253–269. Springer, 2022. 7, 8

[6] Juan C Caicedo, Allen Goodman, Kyle W Karhohs, Beth A Cimini, Jeanelle Ackerman, Marzieh Haghighi, CherKeng Heng, Tim Becker, Minh Doan, Claire McQuin, et al. Nucleus segmentation across imaging experiments: the 2018 data science bowl. *Nature methods*, 16(12):1247–1253, 2019. 2, 3

[7] Jiazhong Cen, Zanwei Zhou, Jiemin Fang, Wei Shen, Lingxi Xie, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, et al. Segment anything in 3d with nerfs. *Advances in Neural Information Processing Systems*, 36:25971–25990, 2023. 2, 5

[8] Jiazhou Chen, Yanghui Xu, Shufang Lu, Ronghua Liang, and Liangliang Nan. 3-d instance segmentation of mvs buildings. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–14, 2022. 2, 3

[9] Luca Ciampi, Carlos Santiago, Joao Paulo Costeira, Claudio Gennaro, and Giuseppe Amato. Domain adaptation for traffic density estimation. In *VISIGRAPP (5: VISAPP)*, pages 185–195, 2021. 2, 4

[10] Luca Ciampi, Carlos Santiago, Joao Costeira, Claudio Gennaro, and Giuseppe Amato. Night and Day Instance Segmented Park (NDISPark) Dataset: a Collection of Images taken by Day and by Night for Vehicle Detection, Segmentation and Counting in Parking Areas, 2022. 2, 4

[11] Nadav Cohen, Yael Newman, and Ariel Shamir. Semantic segmentation in art paintings. In *Computer graphics forum*, pages 261–275. Wiley Online Library, 2022. 2, 3

[12] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 2, 3

[13] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision*, pages 1–23, 2022. 2, 4

[14] Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma, Amlan Kar, Richard Higgins, Sanja Fidler, David Fouhey, and Dima Damen. Epic-kitchens visor benchmark: Video segmentations and object relations. *Advances in Neural Information Processing Systems*, 35:13745–13758, 2022. 2, 4

[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. 6, 7

[16] Yigit Ekin, Ahmet Burak Yildirim, Erdem Eren Caglar, Aykut Erdem, Erkut Erdem, and Aysegul Dundar. Clipaway: Harmonizing focused embeddings for removing objects via diffusion models, 2024. 2

[17] Alireza Fathi, Xiaofeng Ren, and James M Rehg. Learning to recognize objects in egocentric activities. In *CVPR 2011*, pages 3281–3288. IEEE, 2011. 2, 3

[18] Jean-Michel Fortin, Olivier Gamache, Vincent Grondin, François Pomerleau, and Philippe Giguère. Instance segmentation for autonomous log grasping in forestry operations. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6064–6071. IEEE, 2022. 2, 4

[19] Daniel Gut. X-ray images of the hip joints, 2021. , Mendeley Data, V1, doi: 10.17632/zm6bxzhmfz.1. 6

[20] Timm Haucke, Hjalmar S Kühl, and Volker Steinhage. Socrates: Introducing depth in visual wildlife monitoring using stereo vision. *Sensors*, 22(23):9082, 2022. 2, 4

[21] Jungseok Hong, Michael Fulton, and Junaed Sattar. Trashcan: A semantically-segmented dataset towards visual detection of marine debris. *arXiv preprint arXiv:2007.08097*, 2020. 2, 4

[22] Lei Ke, Mingqiao Ye, Martin Danelljan, Yu-Wing Tai, Chi-Keung Tang, Fisher Yu, et al. Segment anything in high quality. *Advances in Neural Information Processing Systems*, 36, 2024. 6, 7, 9, 11

[23] 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. 1, 2, 5, 6, 7, 9, 10, 11

[24] Jizhizi Li, Sihan Ma, Jing Zhang, and Dacheng Tao. Privacy-preserving portrait matting. In *Proceedings of the 29th ACM international conference on multimedia*, pages 3501–3509, 2021. 6

[25] Jizhizi Li, Jing Zhang, and Dacheng Tao. Deep automatic natural image matting. *arXiv preprint arXiv:2107.07235*, 2021. 6

[26] Jizhizi Li, Jing Zhang, and Dacheng Tao. Referring image matting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22448–22457, 2023. 8

[27] Jiachen Li, Jitesh Jain, and Humphrey Shi. Matting anything. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1775–1785, 2024. 2, 7, 9, 11

[28] Yin Li, Zhefan Ye, and James M Rehg. Delving into egocentric actions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 287–295, 2015. 2, 3

[29] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 2, 3, 5

[30] Hao Lu, Yutong Dai, Chunhua Shen, and Songcen Xu. Indices matter: Learning to index for deep image matting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3266–3275, 2019. 7, 8

[31] Maciej A Mazurowski, Haoyu Dong, Hanxue Gu, Jichen Yang, Nicholas Konz, and Yixin Zhang. Segment anything model for medical image analysis: an experimental study. *Medical Image Analysis*, 89:102918, 2023. 6

[32] Ben Mildenhall, Pratul P. Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics (TOG)*, 2019. 2, 5

[33] Massimo Minervini, Andreas Fischbach, Hanno Scharr, and Sotirios A Tsaftaris. Finely-grained annotated datasets for image-based plant phenotyping. *Pattern recognition letters*, 81:80–89, 2016. 2, 4

[34] Anna Montoya, Hasnin, kaggle446, shirzad, Will Cukierski, and yffud. Ultrasound nerve segmentation. https://kaggle.com/competitions/ultrasound-nerve-segmentation, 2016. Kaggle. 6

[35] Mattia Pugliatti and Francesco Topputo. Doors: Dataset for boulders segmentation. *Zenodo*, 9(20):6, 2022. 2, 3

[36] Jiyang Qi, Yan Gao, Yao Hu, Xinggang Wang, Xiaoyu Liu, Xiang Bai, Serge Belongie, Alan Yuille, Philip HS Torr, and Song Bai. Occluded video instance segmentation: A benchmark. *International Journal of Computer Vision*, 130(8): 2022–2039, 2022. 2, 4

[37] 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. 6, 7

[38] 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. 8

[39] Zhongzheng Ren, Aseem Agarwala†, Bryan Russell†, Alexander G. Schwing†, and Oliver Wang†. Neural volumetric object selection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. († alphabetic ordering). 2

[40] Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10912–10922, 2021. 2, 3

[41] 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. 6, 7

[42] Corey Snyder and Minh Do. Streets: A novel camera network dataset for traffic flow. *Advances in Neural Information Processing Systems*, 32, 2019. 2, 4

[43] Konstantin Sofiiuk, Ilya A Petrov, and Anton Konushin. Reviving iterative training with mask guidance for interactive segmentation. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 3141–3145. IEEE, 2022. 6

[44] Yuxin Song, Jing Zheng, Long Lei, Zhipeng Ni, Baoliang Zhao, and Ying Hu. Ct2us: Cross-modal transfer learning for kidney segmentation in ultrasound images with synthesized data. *Ultrasonics*, 122:106706, 2022. 6

[45] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2149–2159, 2022. 2

[46] Cameron Trotter, Georgia Atkinson, Matt Sharpe, Kirsten Richardson, A Stephen McGough, Nick Wright, Ben Burville, and Per Berggren. Ndd20: A large-scale few-shot dolphin dataset for coarse and fine-grained categorisation. *arXiv preprint arXiv:2005.13359*, 2020. 2, 4

[47] Boying Wang, Libo Zhang, Longyin Wen, Xianglong Liu, and Yanjun Wu. Towards real-world prohibited item detection: A large-scale x-ray benchmark. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5412–5421, 2021. 2, 4

[48] Lei Yang, Yan Zi Wei, Yisheng He, Wei Sun, Zhenhang Huang, Haibin Huang, and Haoqiang Fan. ishape: A first step towards irregular shape instance segmentation. *arXiv preprint arXiv:2109.15068*, 2021. 2, 3

[49] Jingfeng Yao, Xinggang Wang, Lang Ye, and Wenyu Liu. Matte anything: Interactive natural image matting with segment anything model. *Image and Vision Computing*, 147: 105067, 2024. 2, 6, 7, 11

[50] Zixuan Ye, Wenze Liu, He Guo, Yujia Liang, Chaoyi Hong, Hao Lu, and Zhiguo Cao. Unifying automatic and interactive matting with pretrained vits. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25585–25594, 2024. 2, 7, 11

[51] Ahmet Burak Yildirim, Vedat Baday, Erkut Erdem, Aykut Erdem, and Aysegul Dundar. Inst-inpaint: Instructing to remove objects with diffusion models, 2023. 2

[52] Senthil Yogamani, Ciarán Hughes, Jonathan Horgan, Ganesh Sistu, Padraig Varley, Derek O'Dea, Michal Uricár, Stefan Milz, Martin Simon, Karl Amende, et al. Woodscape: A multi-task, multi-camera fisheye dataset for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9308–9318, 2019. 2, 4

[53] Qihang Yu, Jianming Zhang, He Zhang, Yilin Wang, Zhe Lin, Ning Xu, Yutong Bai, and Alan Yuille. Mask guided matting via progressive refinement network. In *Proceedings*

of the IEEE/CVF conference on computer vision and pattern recognition, pages 1154–1163, 2021. 7, 8

[54] Tao Yu, Runseng Feng, Ruoyu Feng, Jinming Liu, Xin Jin, Wenjun Zeng, and Zhibo Chen. Inpaint anything: Segment anything meets image inpainting. *arXiv preprint arXiv:2304.06790*, 2023. 2, 5

[55] Lingzhi Zhang, Shenghao Zhou, Simon Stent, and Jianbo Shi. Fine-grained egocentric hand-object segmentation: Dataset, model, and applications. In *European Conference on Computer Vision*, pages 127–145. Springer, 2022. 2, 3

[56] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017. 2, 3