Supplementary Material - Detect Any Mirrors: Boosting Learning Reliability on Large-Scale Unlabeled Data with an Iterative Data Engine

Zhaohu Xing[1] Lihao Liu[2] Yijun Yang[1] Hongqiu Wang[1]
Tian Ye[1] Sixiang Chen[1] Wenxue Li[1] Guang Liu[3] Lei Zhu[1,4*]

[1]The Hong Kong University of Science and Technology (Guangzhou) [2]Amazon
[3]Beijing Academy of Artificial Intelligence [4]The Hong Kong University of Science and Technology

1. Data Collection

We collect large-scale unlabeled data from five data sources: **unsplash**, **pexels**, **flickr**, **google**, **baidu**. We use 'mirror' as the keyword to search for and download data in batches. To ensure that each collected image contains mirrors, we use the MiniCPM-V 2.6, a multi-modal large language model, to determine whether there is a mirror in the image. The text prompt is *'Give you an image:<image>./</image>. Does this image contain mirrors? Answer me Yes or No and do not explain.'* To ensure the accuracy of this approach, we randomly checked 15% of the collected mirror images for which the MiniCPM-V 2.6 model identified as 'Yes,' resulting in an accuracy of 99.8%.

2. More Details in Our Experiments

Compared Methods. We compare our method with seven supervised networks and four semi-supervised methods. Seven supervised methods are SegFormer [12], Mask2Former [2], MirrorNet [15], PMDNet [5], SANet [3], VCNet [10], and HetNet [4]. Four semi-supervised methods are: Mean Teacher [11], UAMT [16], DepthAnything [14], and UniMatch [13]. The Mean Teacher method was originally designed for image classification tasks, while UAMT was developed for 3D medical image segmentation. Therefore, we re-implement these two methods using a more advanced segmentation model, specifically SegFormer [12], in this paper. For the DepthAnything method, we employ the DINOv2-small encoder [8] for feature extraction and use the DPT [9] decoder for mirror detection. Consistent with the original paper, we adopt DeepLab-V3 [1] with ResNet-101 as the feature encoder for mirror detection in the UniMatch [13] method. For all semi-supervised methods, we utilize the same training set with collected unlabeled data as our method for fair comparisons.

More Implementation Details. We test two input sizes, 518×518 and 336×336, during training and inference to observe the performance effects of different image sizes.

Datasets	Methods	Image Size	IoU ↑	Accuracy ↑	F_β ↑	MAE ↓	BER ↓
MSD [15]	DAM (Ours)	336	0.928	0.965	0.960	0.024	2.55
	DAM (Ours)	518	**0.934**	**0.980**	**0.968**	**0.019**	**2.32**
PMD [5]	DAM (Ours)	336	0.723	0.839	0.841	0.027	9.14
	DAM (Ours)	518	**0.734**	**0.851**	**0.852**	**0.019**	**8.08**
RGBD* [7]	DAM (Ours)	336	0.801	0.876	0.852	0.039	7.51
	DAM (Ours)	518	**0.808**	**0.885**	**0.870**	**0.030**	**6.42**

Table 1. The performance effects of different input sizes are evaluated. RGBD* denotes the RGBD-Mirror* dataset. Decreasing the input sizes has a slight effect on various evaluation metrics, but our method still demonstrates strong performance across the three test datasets.

These two input sizes are multiples of 14, as the pre-defined patch size of DINOv2 encoders [8] is 14. During training, we use 8 NVIDIA RTX 4090 GPUs, with a batch size of 4 on each GPU, and the loss function is cross-entropy loss [6]. We adopt horizontal flip, color jitter, random grayscale, and random blur for unlabeled images with probabilities of 0.5, 1.0, 0.2, and 0.5, respectively. Moreover, for the SegFormer backbone in Mean Teacher and UAMT, we use an input size of 512×512. For the DINOv2 backbone in the DepthAnything method, we utilize an input size of 518×518. These input sizes are consistent with those used in the original papers [12, 14].

3. More Ablation Study

Ablation of the Smaller Input Size. Table 1 shows the quantitative comparisons for the different input sizes (i.e., 518×518 and 336×336). Our method demonstrates robust performance across the three test datasets.

Ablation on the Iteration Number of the Data Engine. As shown in Figure 1, we conduct additional experiments where we increase the iteration number of our data engine. In this figure, we observe that the performance gains diminish when the iteration number exceeds 3. Therefore, to improve training efficiency, we choose an iteration number of 3 as our default setting in this paper.

4. Future Work and Limitation

Currently, the model size is DINOv2-small [8]. In the future, we plan to increase the model size to further enhance the generalization capability of our model. Furthermore, in scenes with smaller mirrors, we also need to support a larger input size than the current 518×518 or 336×336.

5. More Visualization

Please refer to the following figures (i.e., Figure 2, 3, 4) for comprehensive qualitative results in more challenging scenes. Our method demonstrates impressive generalization capability across various challenging scenarios. Additionally, we present the corresponding output feature maps from the segmentation head in our network. Our method not only accurately detects mirror areas but also outputs distinctive values for different objects in the feature maps.

References

[1] Liang-Chieh Chen. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 1

[2] Bowen Cheng, Anwesa Choudhuri, Ishan Misra, Alexander Kirillov, Rohit Girdhar, and Alexander G Schwing. Mask2former for video instance segmentation. *arXiv preprint arXiv:2112.10764*, 2021. 1

[3] Huankang Guan, Jiaying Lin, and Rynson WH Lau. Learning semantic associations for mirror detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5941–5950, 2022. 1

[4] Ruozhen He, Jiaying Lin, and Rynson WH Lau. Efficient mirror detection via multi-level heterogeneous learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 790–798, 2023. 1

[5] Jiaying Lin, Guodong Wang, and Rynson WH Lau. Progressive mirror detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3697–3705, 2020. 1

[6] Anqi Mao, Mehryar Mohri, and Yutao Zhong. Cross-entropy loss functions: Theoretical analysis and applications. In *International conference on Machine learning*, pages 23803–23828. PMLR, 2023. 1

[7] Haiyang Mei, Bo Dong, Wen Dong, Pieter Peers, Xin Yang, Qiang Zhang, and Xiaopeng Wei. Depth-aware mirror segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3044–3053, 2021. 1

[8] 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, 2

[9] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12179–12188, 2021. 1

[10] Xin Tan, Jiaying Lin, Ke Xu, Pan Chen, Lizhuang Ma, and Rynson WH Lau. Mirror detection with the visual chirality cue. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3492–3504, 2022. 1

[11] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017. 1

[12] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021. 1

[13] Lihe Yang, Lei Qi, Litong Feng, Wayne Zhang, and Yinghuan Shi. Revisiting weak-to-strong consistency in semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7236–7246, 2023. 1

[14] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10371–10381, 2024. 1

[15] Xin Yang, Haiyang Mei, Ke Xu, Xiaopeng Wei, Baocai Yin, and Rynson WH Lau. Where is my mirror? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8809–8818, 2019. 1

[16] Lequan Yu, Shujun Wang, Xiaomeng Li, Chi-Wing Fu, and Pheng-Ann Heng. Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation. In *Medical image computing and computer assisted intervention–MICCAI 2019: 22nd international conference, Shenzhen, China, October 13–17, 2019, proceedings, part II 22*, pages 605–613. Springer, 2019. 1

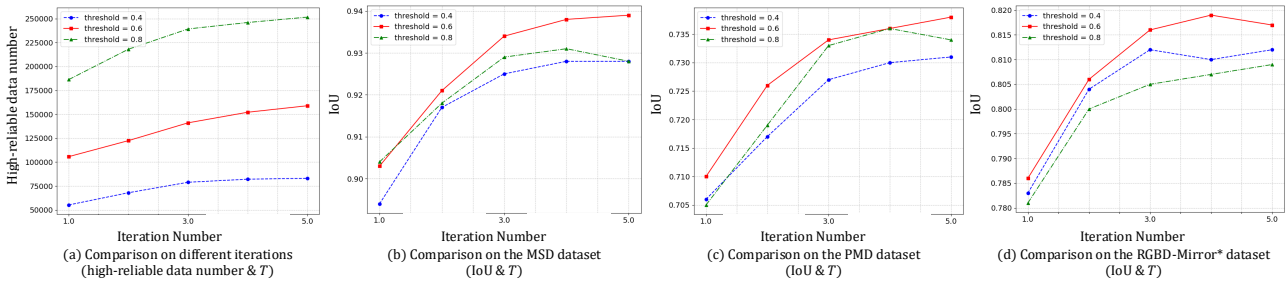

Figure 1. Ablation on the iteration number of the data engine. We observe that the performance gains diminish when the iteration number exceeds 3. Therefore, to reduce training time, we choose an iteration number of 3 as our default setting in this paper.

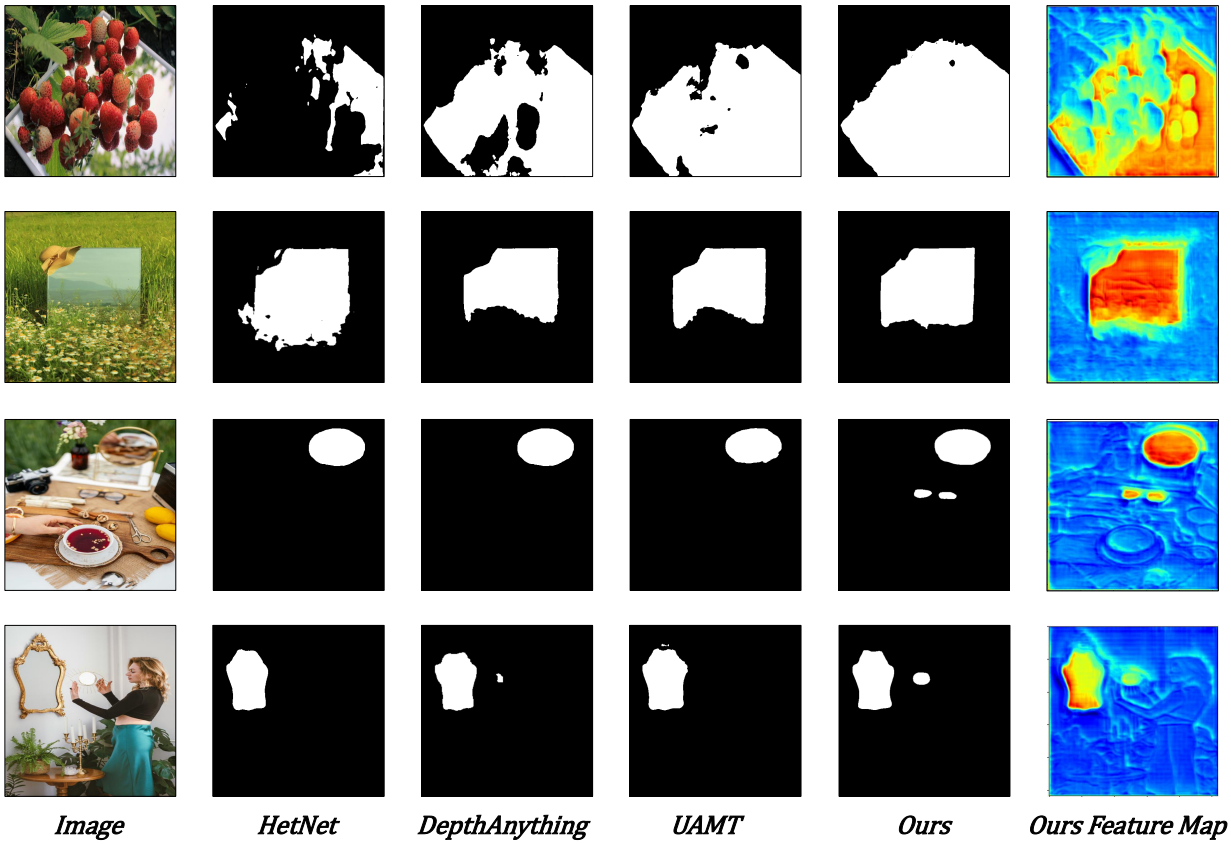

Figure 2. Additional visualizations of challenging scenes using our DAM and other state-of-the-art methods, along with the feature maps from our method.

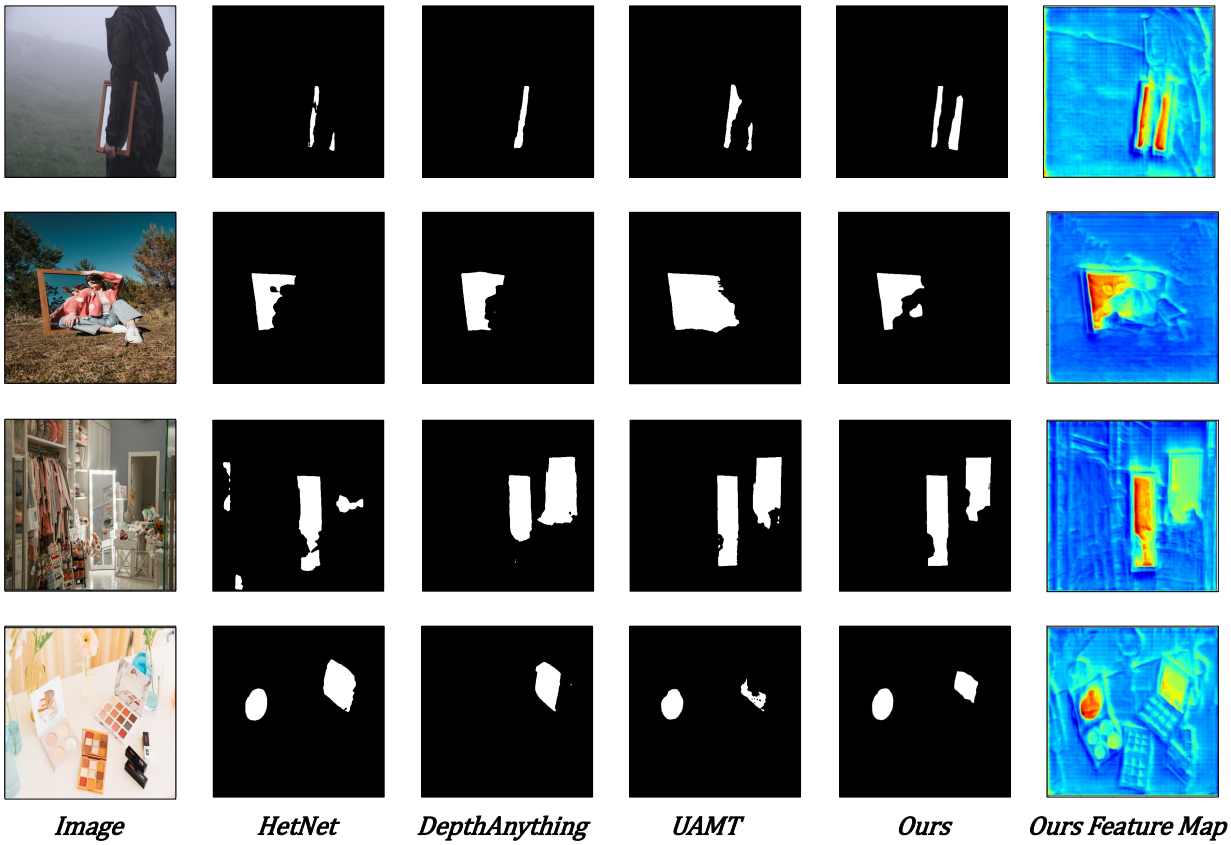

Figure 3. Additional visualizations of challenging scenes using our DAM and other state-of-the-art methods, along with the feature maps from our method.

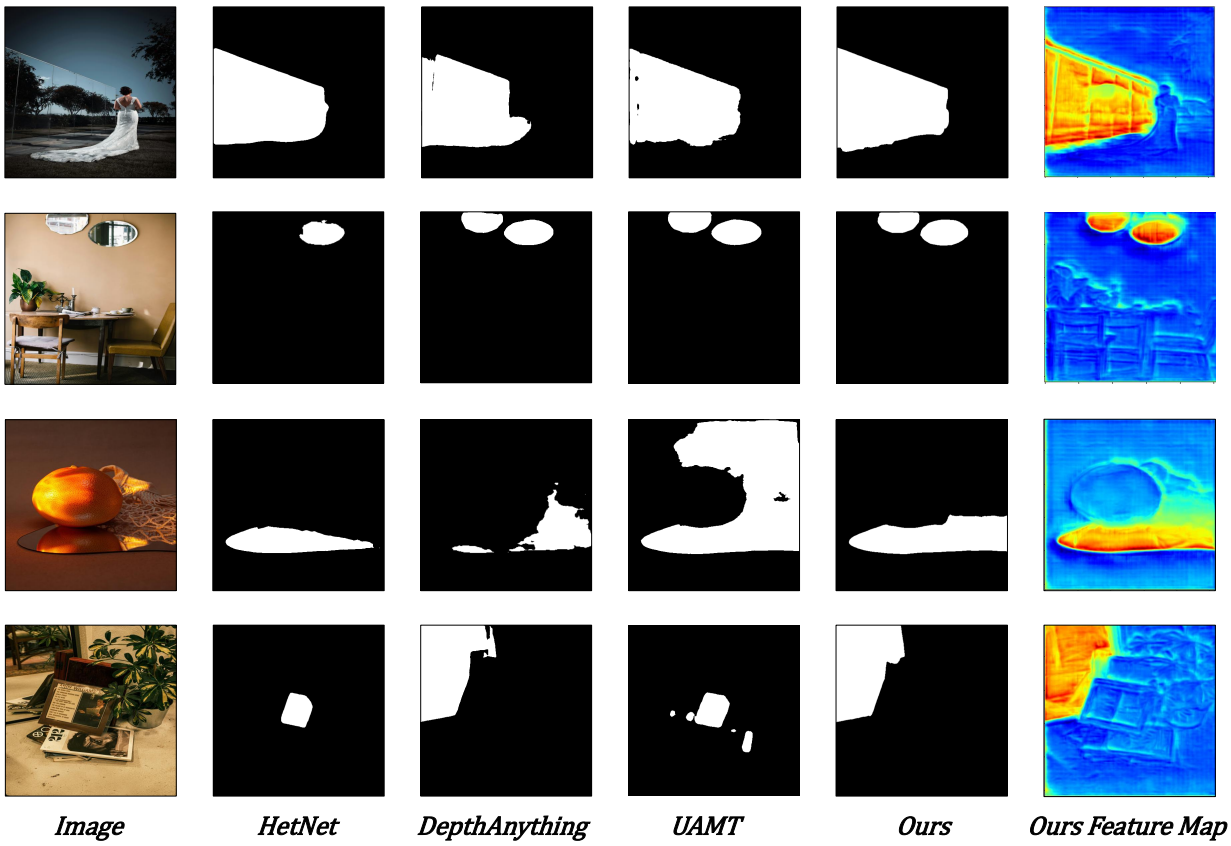

Figure 4. Additional visualizations of challenging scenes using our DAM and other state-of-the-art methods, along with the feature maps from our method.