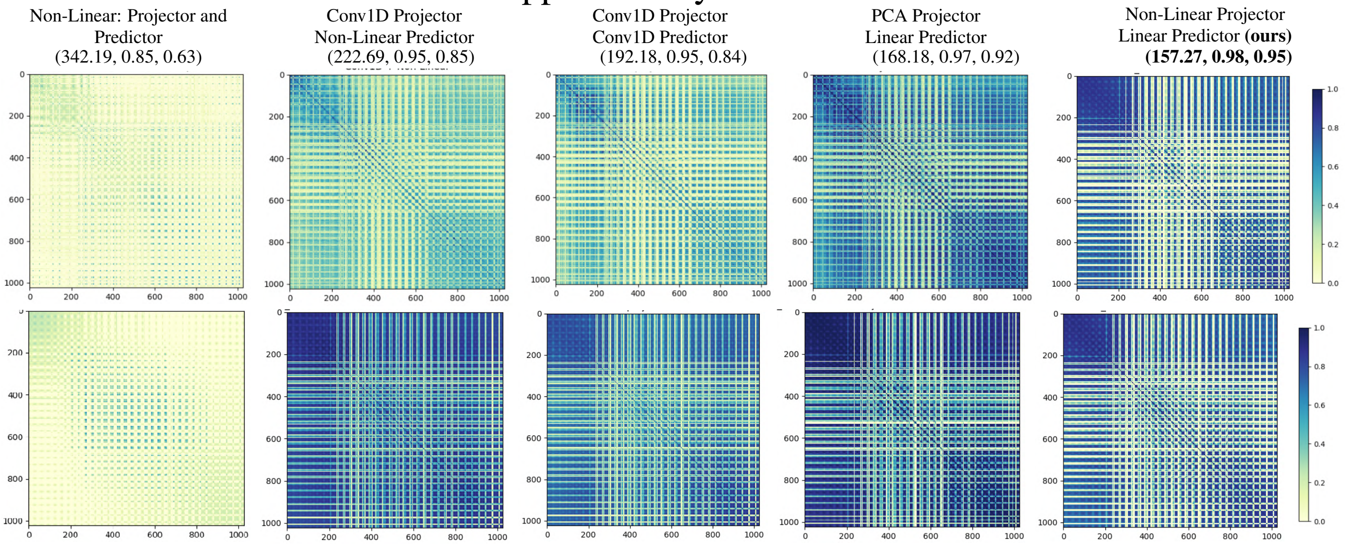
SIMSAM: SIMPLE SIAMESE REPRESENTATIONS BASED SEMANTIC AFFINITY MATRIX FOR UNSUPERVISED IMAGE SEGMENTATION

Chanda Grover Kamra[⋆] *Indra Deep Mastan*[†] *Nitin Kumar*[^] *Debayan Gupta*[⋆]

[⋆] Ashoka University, India. [†] The LNM Institute of Information Technology (LNMIIT), India.
[^] Shiv Nadar University, New Delhi.

Supplementary Material


Non-Linear: Projector and Predictor (342.19, 0.85, 0.63)

Conv1D Projector Non-Linear Predictor (222.69, 0.95, 0.85)

Conv1D Projector Conv1D Predictor (192.18, 0.95, 0.84)

PCA Projector Linear Predictor (168.18, 0.97, 0.92)

Non-Linear Projector Linear Predictor **(ours)** **(157.27, 0.98, 0.95)**

Fig. 1: The top row represents the affinity matrices of the segmentation masks obtained in different configurations of projectors and predictors of a particular image. The bottom row represents the affinity matrices of the ground truth mask. The values reported on top of the figures are *Frobenius Norm, Accuracy and mIOU* scores.

1. IMPLEMENTATION DETAILS

Broadly, we present two modules: One is the *SimSAM framework*, which is used for obtaining a semantically consistent, improved semantic matrix (leading to better image segmentation). We trained our Siamese-based SimSAM framework for 10 iterations for each image.

Projector. We experimented with Principal Component Analysis (PCA) as a projector with 64, 128, and 256 components, as shown in Table 1. Among these configurations, 64 components gave the best accuracy and IOU scores for the segmentation mask. We also conducted studies for multiple configurations of the Projector as shown in Fig. 1. We found that using a non-linear layer as Projector reported the best Frobenius Norm, Accuracy and mIoU scores.

Predictor. We also conducted studies for multiple configurations of predictors on a single image. Fig. 1 shows the affinity matrix obtained in each configuration with respect to its ground truth affinity. As we can see, the linear layer as a predictor reported the best accuracy scores, mIoU and Frobenius norm.

Components	**32**	**64**	**128**
Accuracy	0.77	**0.89**	0.88
mIoU	0.54	**0.75**	0.72

Table 1: Number of PCA components. We tested 32, 64 and 128 PCA components. n=64 gave the best scores.

2. EXPERIMENTAL RESULTS

Object Segmentation masks Outputs. Fig. 2-4 are extended outputs of Fig. 3 of the main manuscript. As shown, the predicted mask obtained with SimSAM (ours) is closest to the ground truth mask as compared to DSM [1] and Deep Cut [2]. We reproduced the segmentation masks of baseline methods with the GitHub codes available in DSM [3] and DeepCut [4].
Semantic Segmentation Outputs. Fig. 5 and Fig. 6 shows the extended results of semantic segmentation masks of Fig. 5 and Fig. 1 of the main manuscript.

Distinguishable Dense Representations. Fig. 7-8 are detailed scores of individual images whose corresponding average values were reported in Table 3, Ablation Study-(I) of the main manuscript. Semantic Affinity Matrix, Frobenius Norm, mIoU and accuracy scores of 10 randomly sampled images from the ECSSD dataset are presented. The difference between DSM [1] and SimSAM is visible in the visualization of affinity matrices with respect to ground truth affinity matrices and with quantitative scores.

3. DATASETS

Object Segmentation. We considered ECSSD [5], DUTS [6], DUTS-OMRON [7] and CUB[8] dataset for training and evaluating the performance of segmentation masks obtained with our method. During training on our SimSAM framework, we considered a batch size of two. During inference, we passed entire image on the trained network to obtain the projected DINO-ViT features for computing a semantically consistent better affinity matrix.

Semantic Segmentation. We performed qualitative and quantitative experiments on the PASCAL VOC dataset for the semantic segmentation task.

4. EVALUATION METRICS

Image Segmentation. We presented mIOU scores for evaluating the quality of the segmentation mask, as mentioned in Section 4 of the main paper. Our method outperformed the DSM method on ECSSD, DUTS and OMRON datasets. We randomly sampled ten images from ECSSD data and computed their Frobenius Norm, Accuracy and mIoU scores for this. We found that our method performed better on those individual images. See the Fig. 7 and 8.

5. ABLATION STUDIES

Ablation Study-(I). We performed more ablation studies to explore the effect of normalization on the Vanilla affinity matrix $W_A = (F_L^K)(F_L^K)^T - \mathcal{M}\big((F_L^K)(F_L^K)^T$. Let $W_{A'} = (F_L^K)(F_L^K)^T$ be the matrix without subtracting mean as in W_A. Table 2 shows that subtracting the mean from the matrix improves the accuracy and mIoU scores.

Affinity Matrix	**Accuracy**	**mIoU**
$W_{A'}$	0.859	0.678
W_A	**0.893**	**0.757**

Table 2: **Ablation Study-(I)**: Deep Spectral Affinity Matrix on ECSSD dataset. Effect of subtracting mean of feature values from correlation affinity matrix.

Abalation Study-(II). We considered different for fine-tuning the values of κ and found that $\kappa = 0.1$ works best for obtaining the best scores on the object segmentation task on ECSSD dataset.

	κ=0.1	κ=0.3	κ=0.5	κ=0.7	κ=0.9
mIoU	**0.762**	0.742	0.723	0.733	0.752
accuracy	**0.896**	0.888	0.842	0.850	0.874

Table 3: Parametric tuning of κ on ECSSD dataset

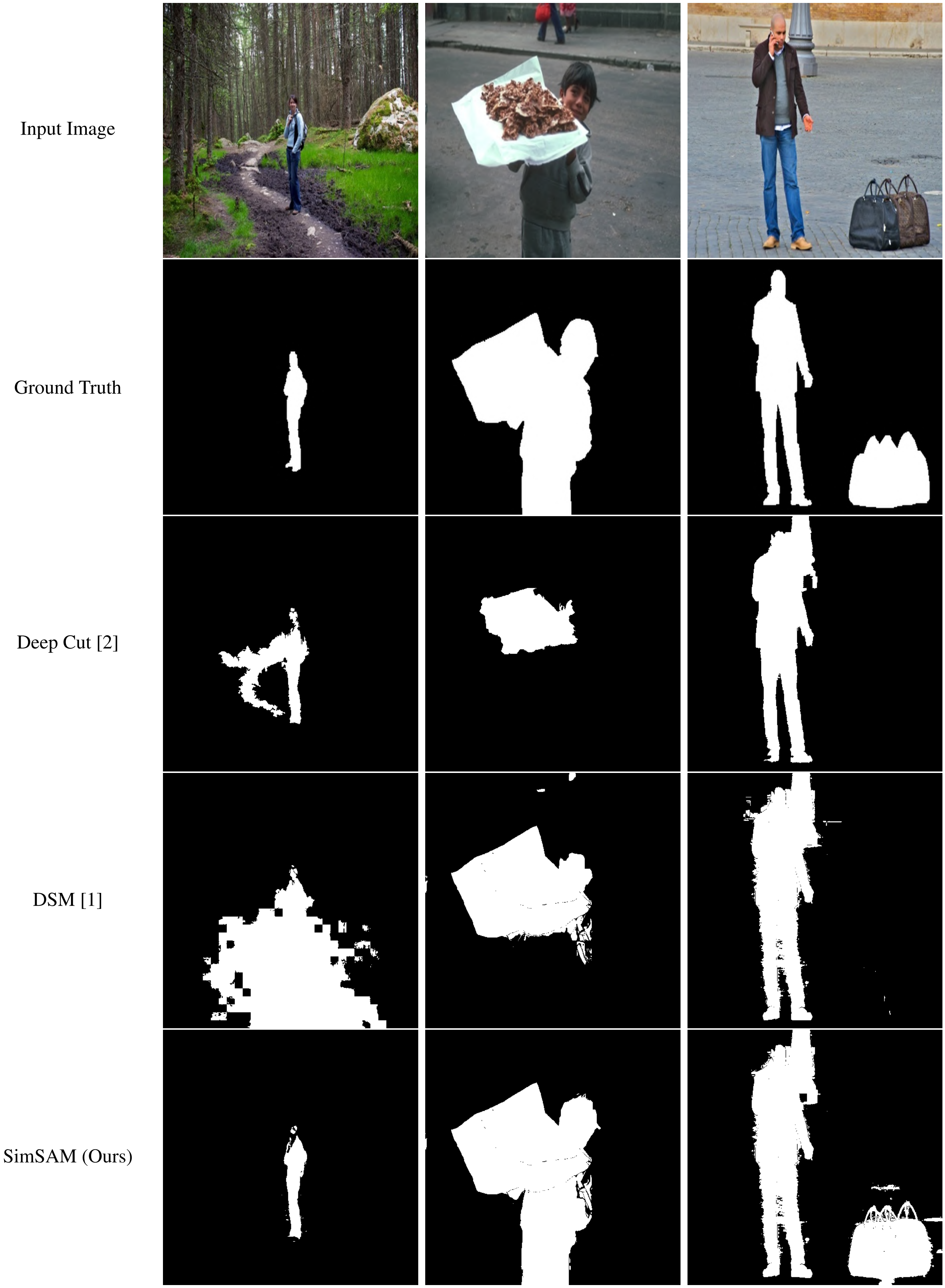


Fig. 2: Object Segmentation Outputs. SimSAM (ours) is closer to Ground Truth in comparison to baseline methods.

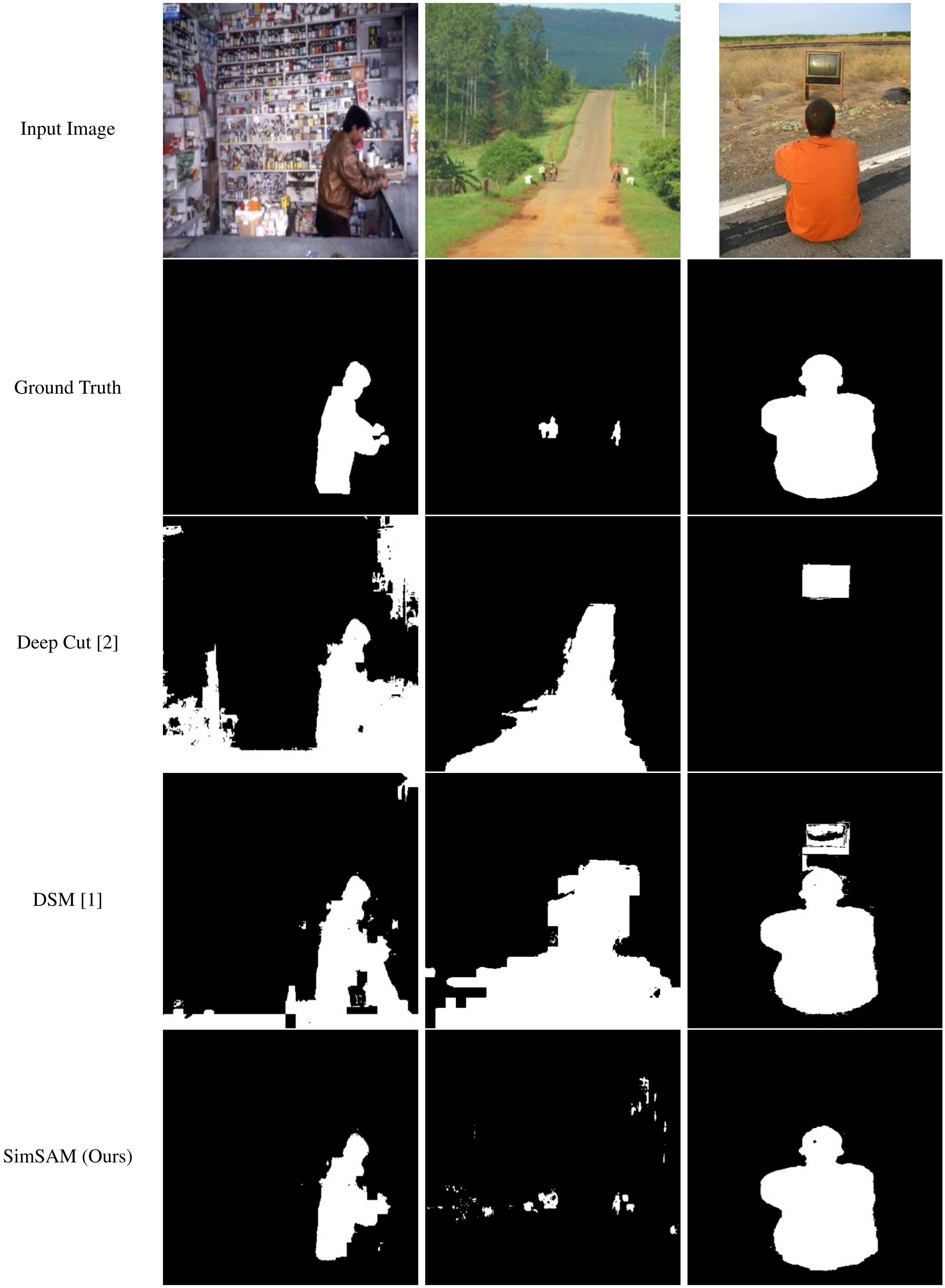


Fig. 3: Object Segmentation Outputs. SimSAM (ours) is closer to Ground Truth in comparison to baseline methods.

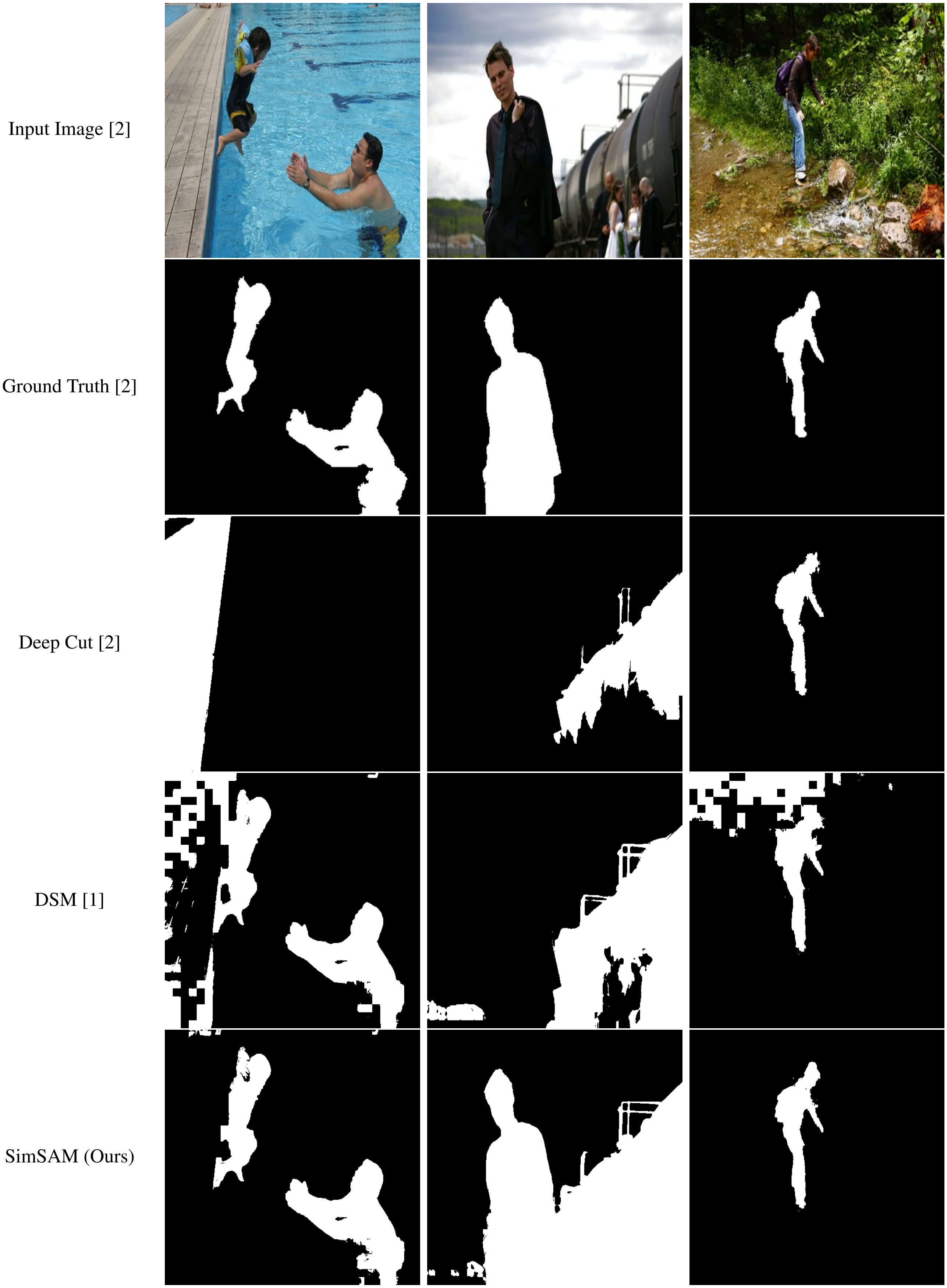


Fig. 4: Object Segmentation Outputs. SimSAM (ours) is closer to Ground Truth in comparison to baseline methods.

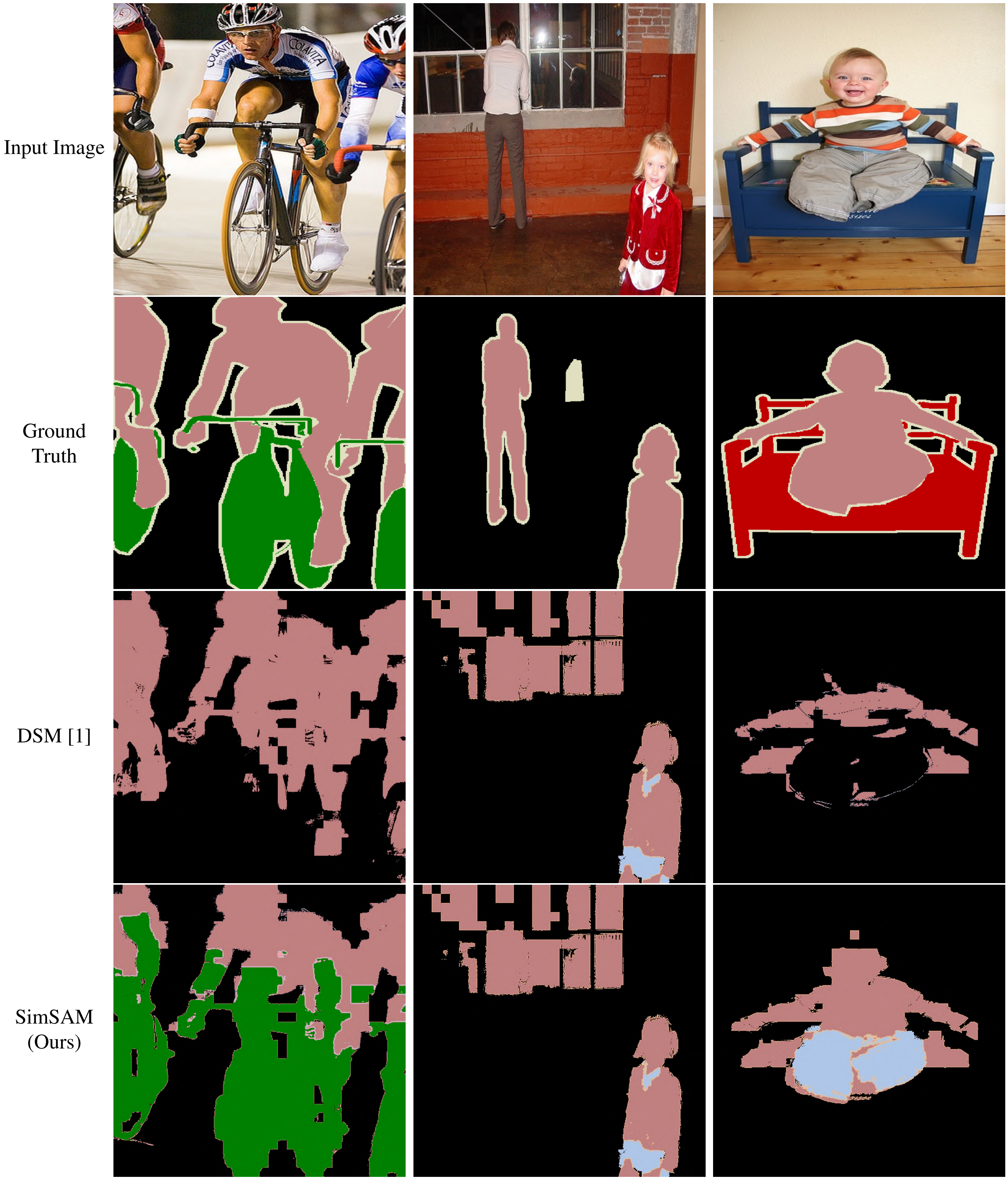


Fig. 5: Semantic Segmentation Outputs.

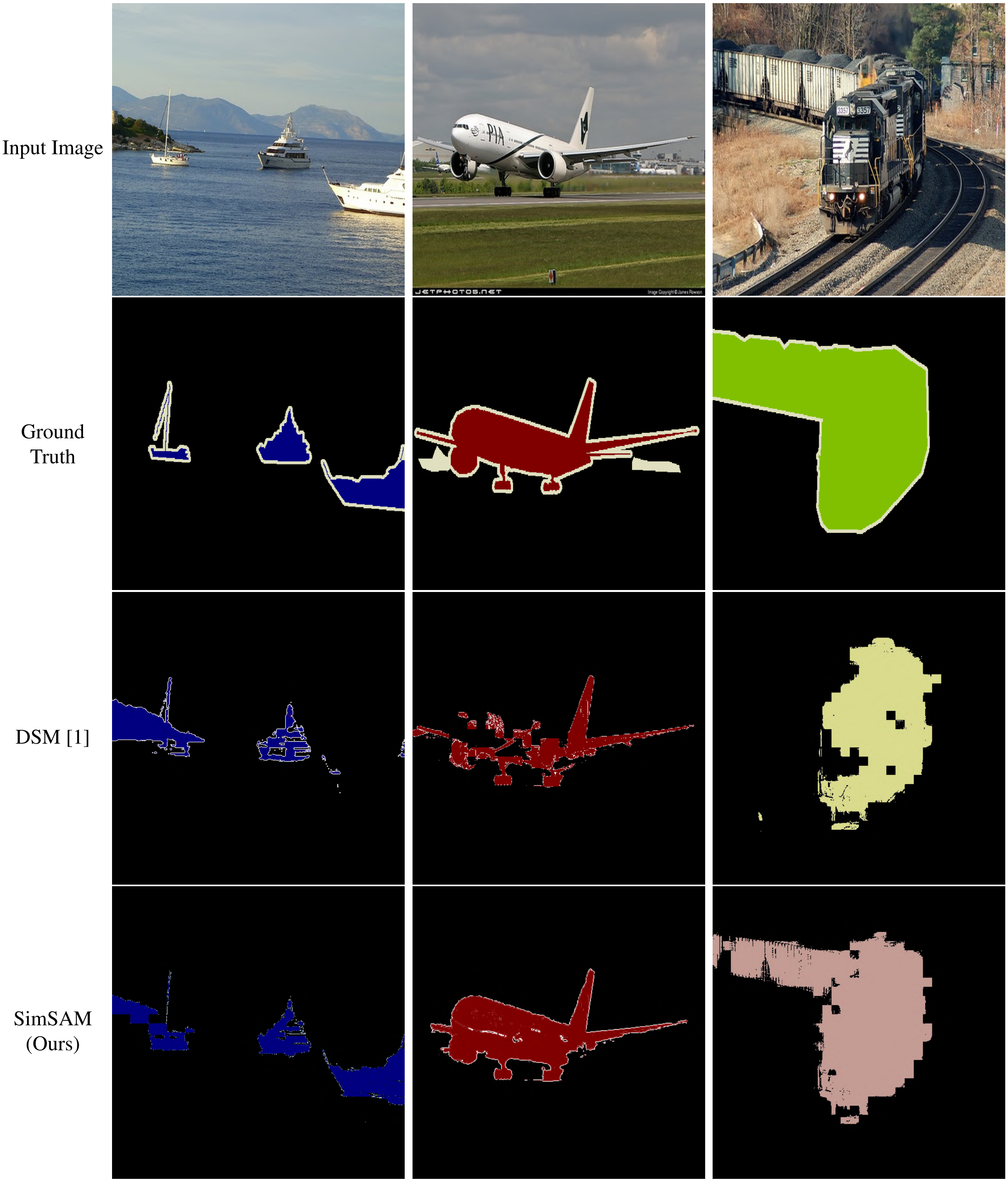


Fig. 6: Semantic Segmentation Outputs.

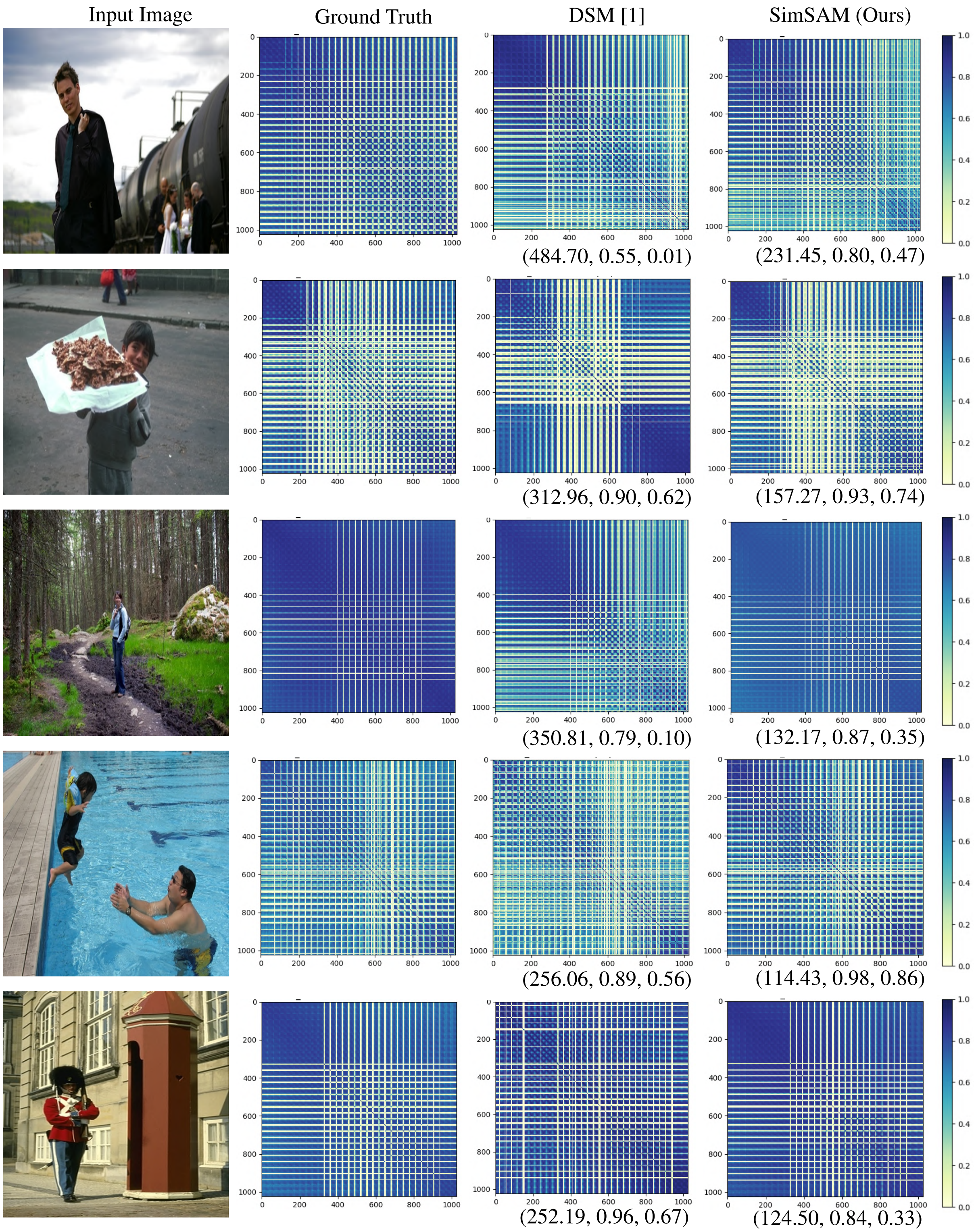


Fig. 7: Values reported at bottom of DSM [1] and SimSAM (ours) method is Frobenius Norm, Accuracy and mIoU scores of randomly sampled image from ECSSD dataset. **Ablation Study-(I)** of main manuscript.

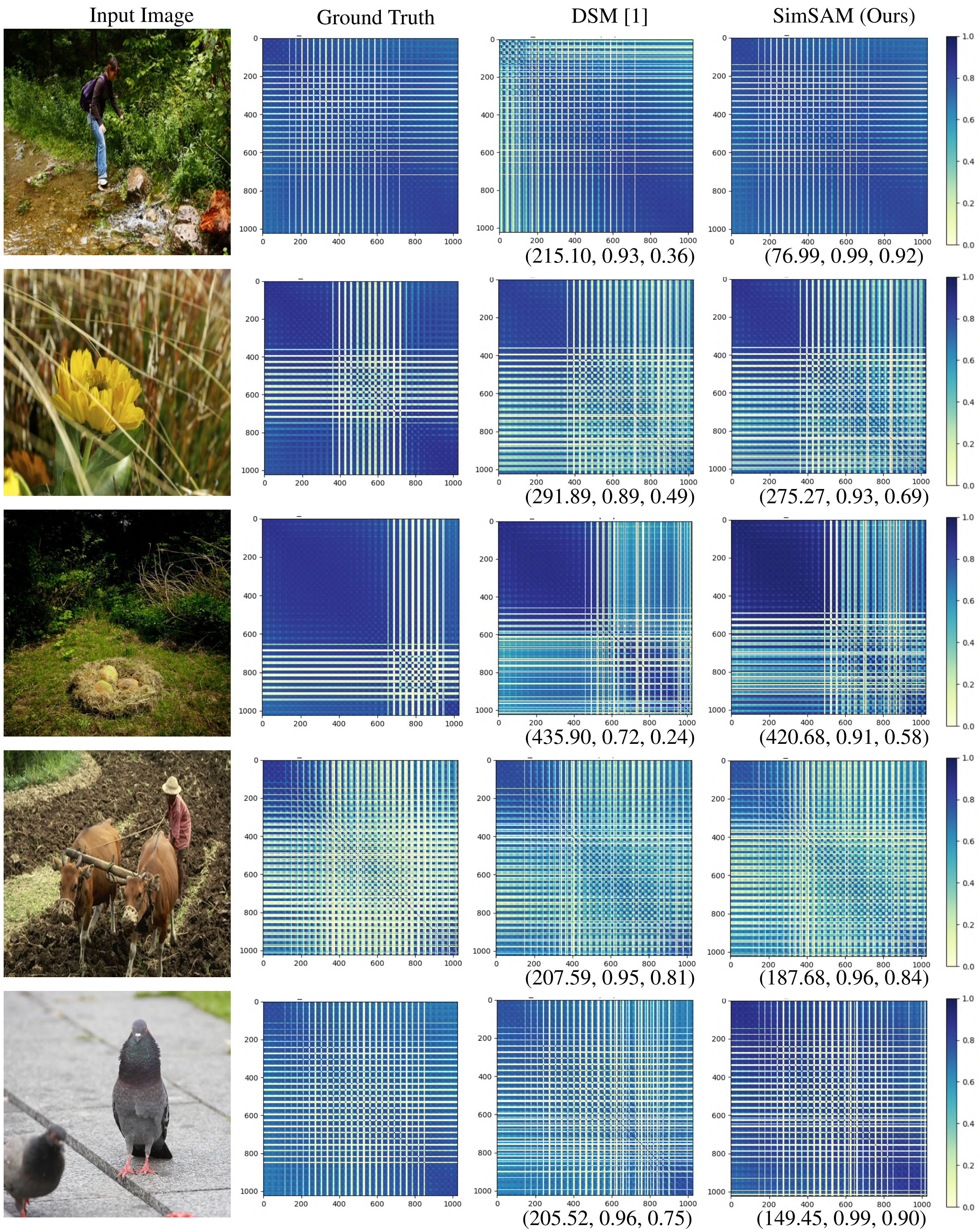


Fig. 8: Values reported at bottom of DSM [1] and SimSAM (ours) method is Frobenius Norm, Accuracy and mIoU scores of randomly sampled image from ECSSD dataset. **Ablation Study-(I)** of main manuscript.

6. REFERENCES

[1] Luke Melas-Kyriazi, Christian Rupprecht, Iro Laina, and Andrea Vedaldi, "Deep spectral methods: A surprisingly strong baseline for unsupervised semantic segmentation and localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 8364–8375.

[2] Amit Aflalo, Shai Bagon, Tamar Kashti, and Yonina Eldar, "Deepcut: Unsupervised segmentation using graph neural networks clustering," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, October 2023, pp. 32–41.

[3] Luke Melas-Kyriazi, "deep-spectral-segmentation," `https://github.com/lukemelas/deep-spectral-segmentation`, 2022.

[4] Amit Aflalo, "Deepcut," `https://github.com/SAMPL-Weizmann/DeepCut`, 2022.

[5] Jianping Shi, Qiong Yan, Li Xu, and Jiaya Jia, "Hierarchical image saliency detection on extended cssd," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 4, pp. 717–729, 2015.

[6] Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan, "Learning to detect salient objects with image-level supervision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 136–145.

[7] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang, "Saliency detection via graph-based manifold ranking," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 3166–3173.

[8] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "Cub-200-2012," Tech. Rep. CNS-TR-2011-001, California Institute of Technology, 2011.