Indices Matter: Learning to Index for Deep Image Matting Supplementary Material

Hao Lu[†] Yutong Dai[†] Chunhua Shen[†*] Songcen Xu[‡]
[†]The University of Adelaide, Australia [‡]Noah's Ark Lab, Huawei Technologies
e-mail: {hao.lu, yutong.dai, chunhua.shen}@adelaide.edu.au

1. Introduction

This supplementary material includes the following contents:

- Extension of IndexNet to image classification. We evaluate IndexNet on the CIFAR-10 and CIFAR-100 datasets [8];
- Extention of IndexNet to monocular depth prediction. The NYUv2 dataset [13] is used;
- Extention of IndexNet to scene understanding. We evaluate IndexNet on the SUN RGB-D dataset [15];
- Further visualizations of learned index maps;
- Further qualitative results on the Composition-1k testing set and the `alphamatting.com` online benchmark;
- Further ablation study supporting the normalization design in IndexNet.
- Some failure cases of image matting.

In following experiments, 'Context' is always included in IndexNet.

2. Image Classification

Here we extend IndexNet to the task of image classification. LeNet [9], MobileNet [6] and VGG-16 [14] are chosen as our backbones. We conduct experiments on these architectures with/without IndexNet on the CIFAR-10 and CIFAR-100 datasets. Note that, since models for image classification do not have upsampling stages, we only apply the $\mathcal{IP}$ operator. All networks are trained from scratch for 160 epochs. We set the batch size to 128 and use the standard SGD for optimization. The initial learning rate is set to 0.1, and reduced by $\times 10$ at the 80-th epoch and the 120-th epoch. As what can be seen in Table 1, IndexNet can generally bring $1\% \sim 2\%$ improvements when the complexity of the problem increases or the model capacity is low (LeNet and MobileNet). However, the improvement becomes marginal with increased model capacity (VGG-16) or reduced problem complexity (CIFAR-10). A possible explanation is that the training loss converges rapidly so that no informative gradient signal is backpropagated to optimize IndexNet.

Method	IndexNet	CIFAR-10	CIFAR-100
LeNet	–	75.47	39.34
LeNet + $\mathcal{IP}$	Nonlinear HIN	77.65	41.65
LeNet + $\mathcal{IP}$	O2O Linear DIN	77.73	40.34
LeNet + $\mathcal{IP}$	M2O Nonlinear DIN	77.54	41.74
MobileNet	–	90.72	67.93
MobileNet + $\mathcal{IP}$	Nonlinear HIN	91.25	70.18
MobileNet + $\mathcal{IP}$	O2O Linear DIN	90.68	69.32
MobileNet + $\mathcal{IP}$	M2O Nonlinear DIN	90.49	70.19
VGG-16	–	93.76	72.93
VGG-16 + $\mathcal{IP}$	Nonlinear HIN	93.96	73.37
VGG-16 + $\mathcal{IP}$	O2O Linear DIN	94.09	73.02
VGG-16 + $\mathcal{IP}$	M2O Nonlinear DIN	94.00	73.01

Table 1: Accuracy (%) on the CIFAR-10 and CIFAR-100 image datasets.

3. Monocular Depth Prediction

Here we demonstrate the effectiveness of IndexNet on the task of monocular depth prediction (metric depth). A ResNet-50 based architecture proposed recently by [7] is regarded as our baseline. We compare the performance of architectures with/without IndexNet on the NYUDv2 dataset. The original architecture and our modified architecture are shown in Fig. 1. Note that, the only modification is to replace all the bilinear upsampling in the decoder (D) and multi-scale feature fusion (MFF) modules with IndexNet and the indexed upsampling ($\mathcal{IU}$) operator. These models are trained with a batch size of 10 for 20 epochs. We start with a learning rate of 0.0001 and divide it by $\times 10$ every 5 epochs. We use the following measures to quantify the performance:

*Corresponding author.

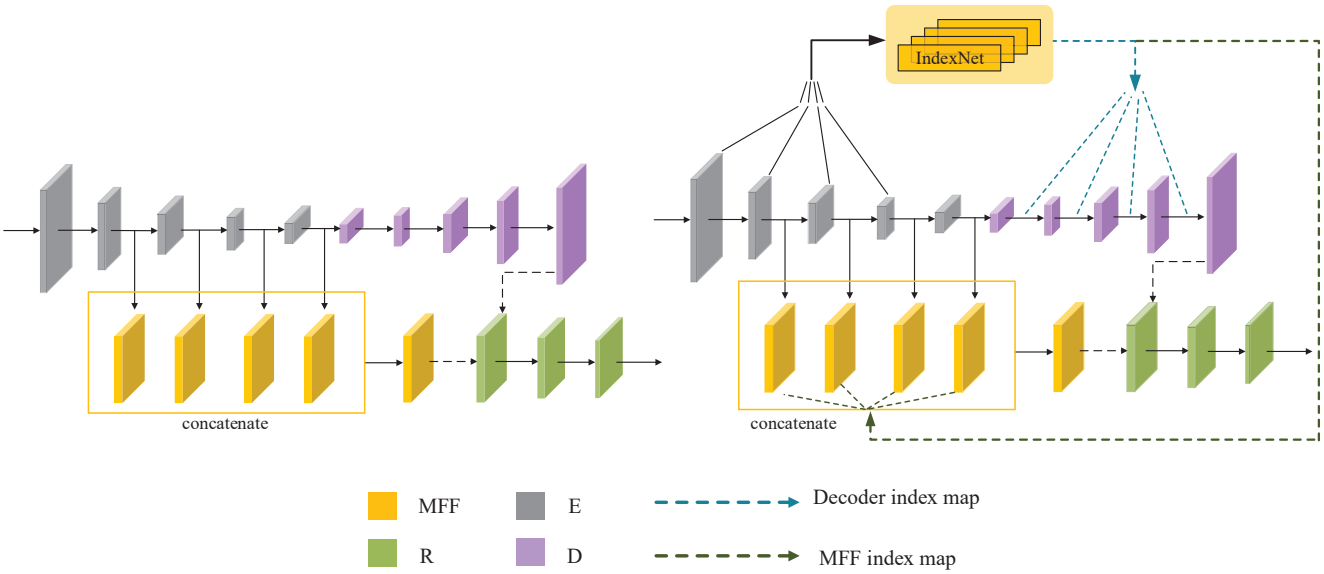

Figure 1: The schematic diagram of the depth estimation network proposed by Hu *et al*. [7] (left) and its modified version with IndexNet (right).

Method	IndexNet	RMS	REL	log10	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Re-implementation of Hu *et al*. [7]	—	0.558	0.129	0.055	0.837	0.968	**0.992**
Hu *et al*. [7] + $\mathcal{IU}$	Nonlinear HIN	**0.553**	**0.128**	**0.055**	**0.841**	**0.968**	0.991

Table 2: Results on the NYUDv2 dataset. For RMS, REL and log10, lower is better. For $\delta < 1.25$, $\delta < 1.25^2$ and $\delta < 1.25^3$, higher is better. The best performance is boldfaced.

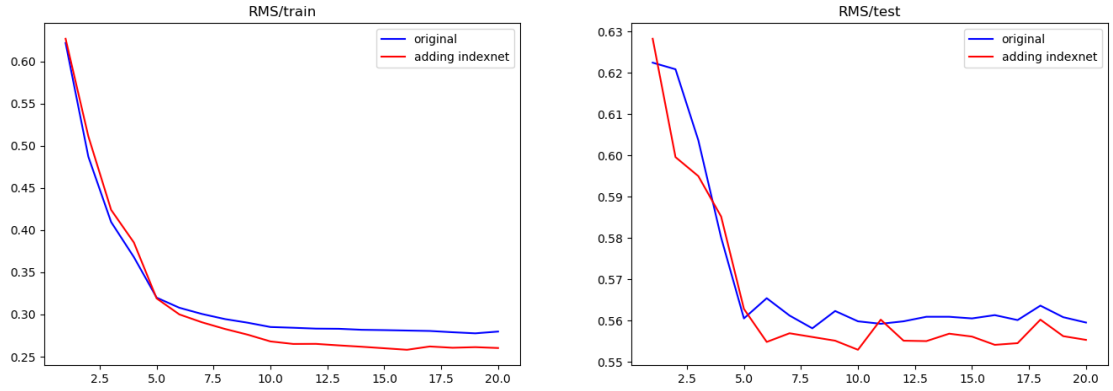

Figure 2: Learning curves of RMS on the NYUDv2 dataset.

- root mean squared error (rms): $\sqrt{\frac{1}{T}\sum_{i=1}^{T}(d_i - g_i)^2}$
- mean relative error (rel): $\frac{1}{T}\sum_{i=1}^{T}\frac{\|d_i - g_i\|_1}{g_i}$
- mean log 10 error (log 10): $\frac{1}{T}\sum_{i=1}^{T}\|\log_{10} d_i - \log_{10} g_i\|_1$
- thresholded accuracy: percentage(%) of d_i, s.t. $\max\left(\frac{d_i}{g_i}, \frac{g_i}{d_i}\right) = \delta <$ threshold.

Experimental results are reported in Table 2 and Fig. 3. From Fig. 3, we can see that IndexNet improves the performance of baseline, especially on capturing edges. The training process of the modified architecture is also more stable than the baseline, as shown by the learning curves of RMS in Fig. 2.

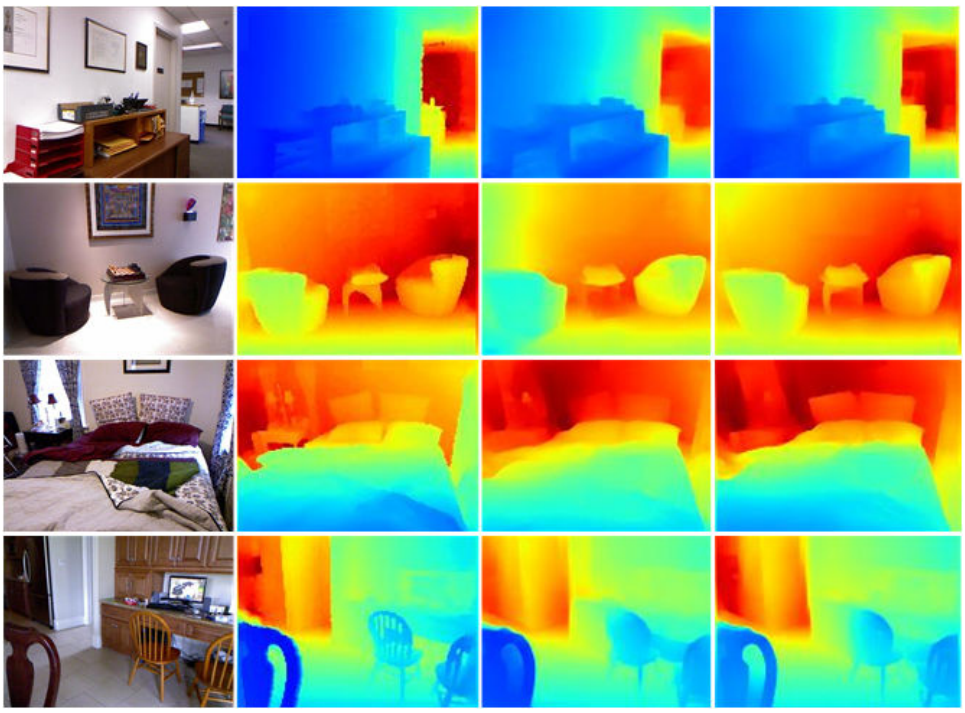

Figure 3: Depth prediction results on the NYUv2 dataset. From left to right, the original image, ground-truth, baseline, and our modified architecture with IndexNet (NonLinear HIN).

Network	IndexNet	Mean IoU
SegNet	–	31.50
SegNet + $\mathcal{IP}$ + $\mathcal{IU}$	Nonlinear HIN	32.70
SegNet + $\mathcal{IP}$ + $\mathcal{IU}$	O2O Linear DIN	32.47
SegNet + $\mathcal{IP}$ + $\mathcal{IU}$	M2O Nonlinear DIN	**32.71**

Table 3: Quantitative comparison on the SUN RGB-D dataset. The best performance is boldfaced

4. Scene Understanding

Here we extend IndexNet to the task of scene understanding. We choose SegNet [2] as our baseline and evaluate three different IndexNet structures—Nonlinear HIN, O2O Linear DIN, and M2O NonLinear DIN. All models are trained from scratch with a batch size 16 for 100 epochs. The learning rate is initially set to 0.01, and reduced by $\times 10$ at the 70-th epoch and the 90-th epoch, respectively. As shown in Table 3, all three types of IndexNet bring $3\% \sim 4\%$ relative improvements. According to the qualitative results shown in Fig. 4, we observe that our modified models significantly suppress some discrete predictions appeared in the baseline. It is worth noting that, IndexNet also speeds up the convergence of training process, which can be observed from the learning curves presented in Fig. 5.

5. Visualization of Index Maps

Here we present further visualization results of index maps learned by the IndexNet (M2O DIN with 'Nonlinear+Context') in Fig. 6. Note that only index maps generated for the decoder are visualized here because the magnitude of index maps for the encoder is suppressed due to the use of softmax. It can be seen from the results that the IndexNet automatically learns to capture contours and details.

6. Qualitative Results of Alpha Mattes

Here we show some further results on the testing set of `alphamatting.com` online benchmark in Fig. 7 and results on the Composition-1k testing set in Figures 8 and 9. These results further present the efficacy of our method, such as recovering textures and details, and extracting transparent foreground objects.

7. Ablation Study on Index Normalization

Details about the design of IndexNet have been illustrated in main part of the paper. Here we supplement results on the effect of different normalization choices to the index maps. Aside from sigmoid function for decoder and sigmoid+softmax function for encoder, we compare other three different combinations of normalization functions, as

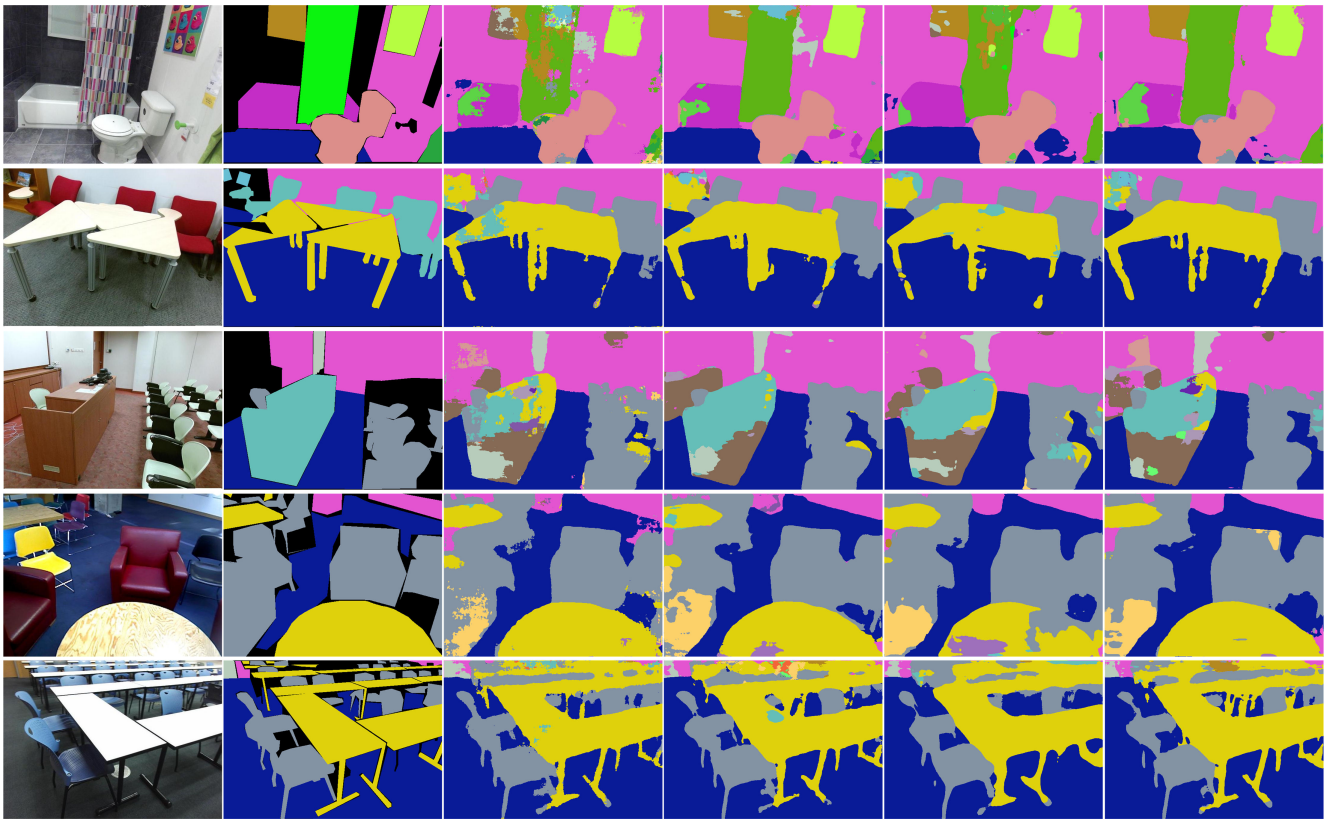

Figure 4: Scene understanding results on the SUNRGB-D dataset. From left to right, the original image, ground-truth, SegNet, HIN with 'Nonlinear+Context', O2O DIN with 'Linear+Context', M2O DIN with 'Nonlinear+Context'.

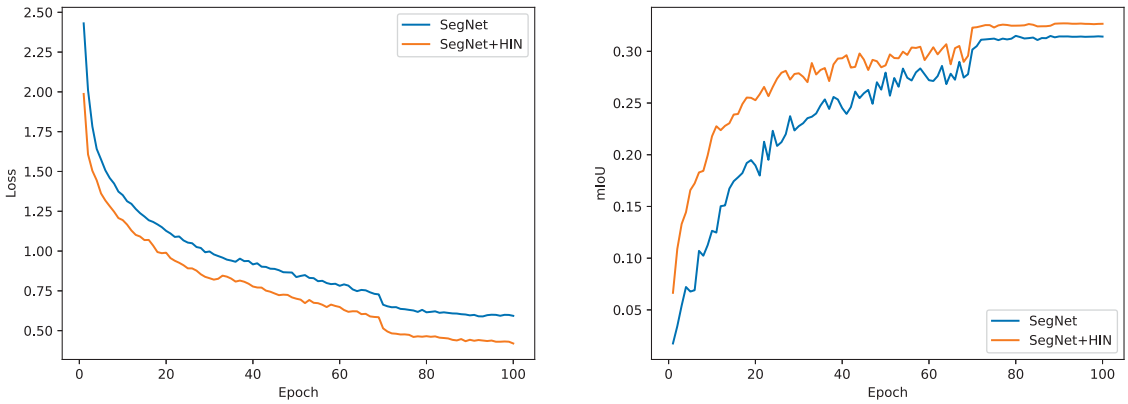

Figure 5: Learning curves of training loss and test mIoU on the SUN RGB-D dataset.

listed in Table 4. This experiment is conducted based on the M2O DIN with "Nonlinear+Context". We observe that, the other three types of design all perform worse than the normalization design we choose, which suggests it is important to keep the magnitude consistency during indexed pooling.

8. Failure Cases

Here we show some failure cases of our method, which are not included in the main text due to page limitation. As shown in Fig. 10, our method may achieve unsatisfactory results when the foreground and the background have similar colors. Indeed, we consider such a case is hard even

Figure 6: Further visualization results of index maps for upsampling. From left to right, learned index maps from shallow to deep layers.

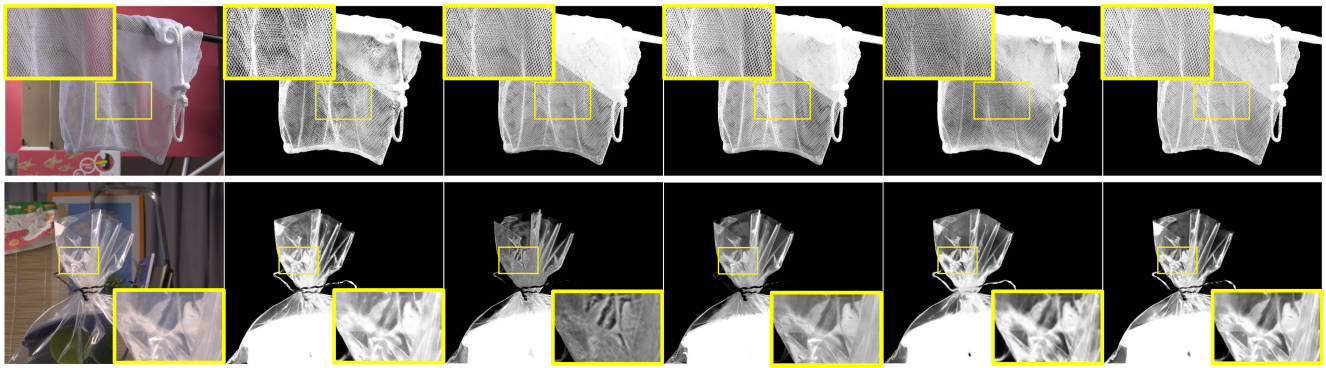

Figure 7: Further results on the `alphamatting.com` testing set. Here we focus on the matting of transparent objects. From left to right, the original image, AlphaGAN matting [12], DCNN matting [5], Information-flow matting [1], Deep matting [16], Ours.

Encoder	Decoder	SAD	MSE	Grad	Conn
sigmoid	sigmoid	52.7	0.016	29.3	52.4
softmax	softmax	51.6	0.015	29.2	51.6
softmax+sigmoid	softmax	57.3	0.016	43.5	57.3
sigmoid+softmax	sigmoid	**45.8**	**0.013**	**25.9**	**43.7**

Table 4: Ablation study of different normalization choices.

for manual matting. A potential solution may be to do local contrast enhancement.

References

[1] Yagiz Aksoy, Tunc Ozan Aydin, and Marc Pollefeys. Designing effective inter-pixel information flow for natural image matting. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 29–37, 2017. 5, 6, 7

[2] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. SegNet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12):2481–2495, 2017. 3

[3] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proc. European Conference on Computer Vision (ECCV)*, 2018. 6, 7

[4] Qifeng Chen, Dingzeyu Li, and Chi-Keung Tang. KNN matting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(9):2175–2188, 2013. 6, 7

[5] Donghyeon Cho, Yu-Wing Tai, and Inso Kweon. Natural image matting using deep convolutional neural networks. In *Proc. European Conference on Computer Vision (ECCV)*, pages 626–643. Springer, 2016. 5

[6] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry

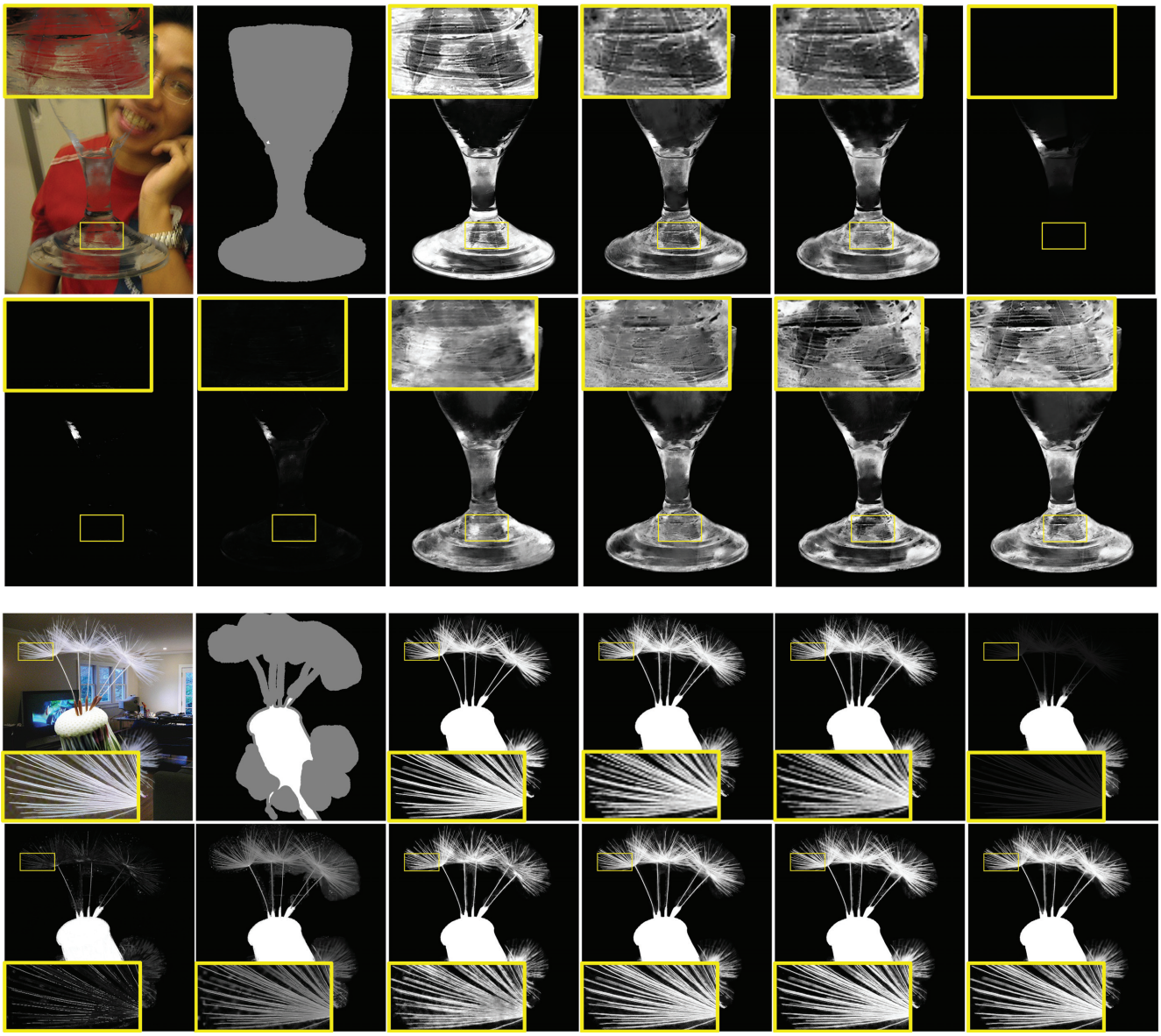

Figure 8: Further results on the Composition-1k testing set. Here we highlight textural and subtle details. From top left to bottom right, the original image, trimap, ground-truth alpha matte, Deeplabv3+ [3], RefineNet [11], Closed-form matting [10], KNN matting [4], Information-flow matting [1], Deep matting [16], Ours (HIN with "Nonlinear+Context"), Ours (O2O DIN with "Linear+Context"), Ours (M2O DIN with "Nonlinear+Context").

Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv*, 2017. 1

[7] Junjie Hu, Mete Ozay, Yan Zhang, and Takayuki Okatani. Revisiting single image depth estimation: toward higher resolution maps with accurate object boundaries. In *Proc. IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1043–1051. IEEE, 2019. 1, 2

[8] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009. 1

[9] Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 1

[10] Anat Levin, Dani Lischinski, and Yair Weiss. A closed-form solution to natural image matting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2):228–242, 2008. 6, 7

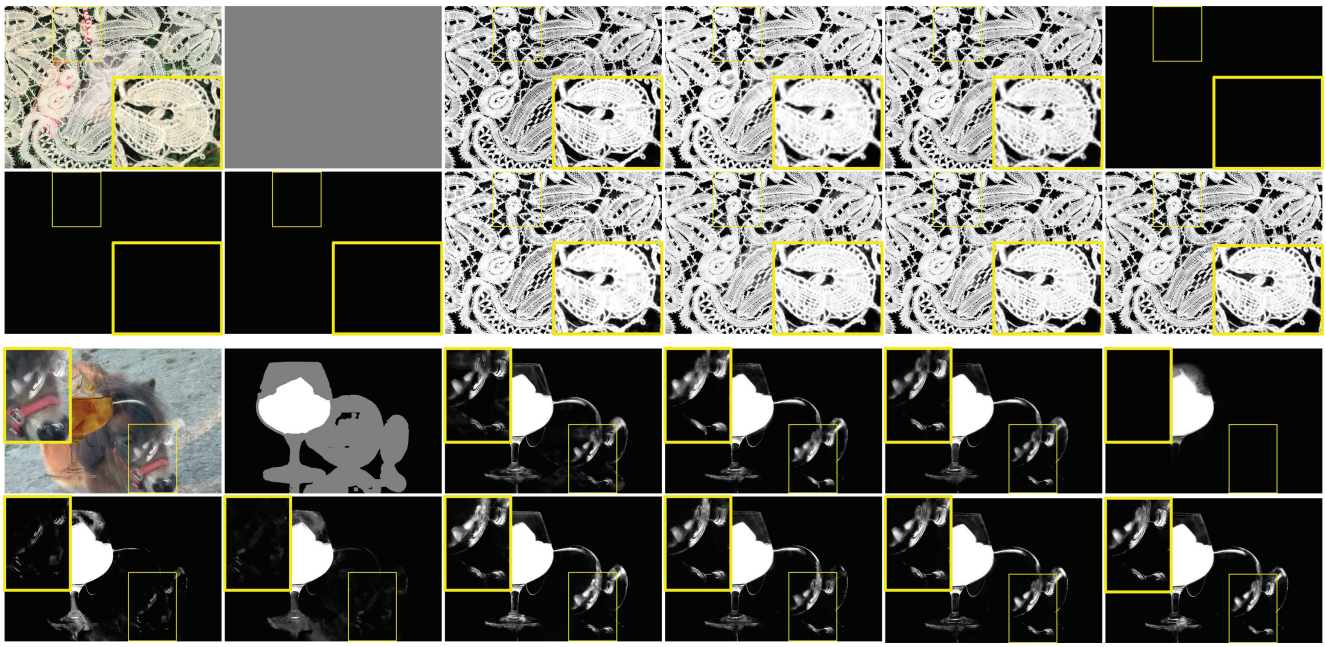

Figure 9: Further results on the Composition-1k testing set. Here we highlight details and transparency. From top left to bottom right, the original image, trimap, ground-truth alpha matte, Deeplabv3+ [3], RefineNet [11], Closed-form matting [10], KNN matting [4], Information-flow matting [1], Deep matting [16], Ours (HIN with "Nonlinear+Context"), Ours (O2O DIN with "Linear+Context"), and Ours (M2O DIN with "Nonlinear+Context").

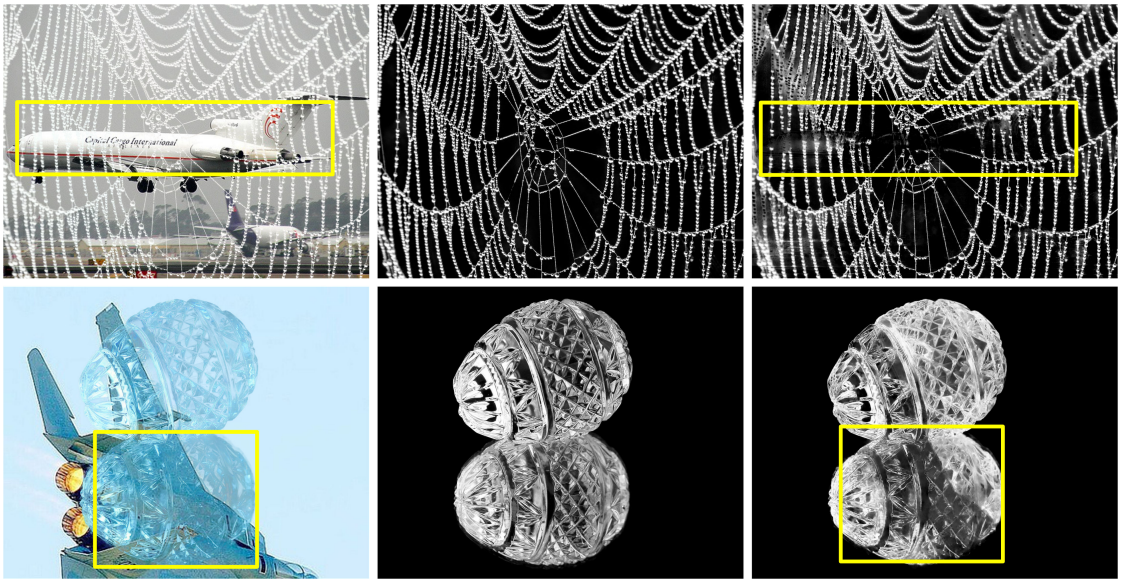

Figure 10: Failure cases where the foreground is similar to the background. From left to right, the original image, ground-truth alpha matte, Ours.

[11] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. RefineNet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1925–1934, 2017. 6, 7

[12] Sebastian Lutz, Konstantinos Amplianitis, and Aljosa Smolic. AlphaGAN: Generative adversarial networks for natural image matting. In *Proc. British Machince Vision*

Conference (BMVC), 2018. 5

[13] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *Proc. European Conference on Computer Vision (ECCV)*, pages 746–760. Springer, 2012. 1

[14] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proc. International Conference on Learning Representations (ICLR)*, 2014. 1

[15] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. SUN RGB-D: A RGB-D scene understanding benchmark suite. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 567–576, 2015. 1

[16] Ning Xu, Brian Price, Scott Cohen, and Thomas Huang. Deep image matting. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2970–2979, 2017. 5, 6, 7