Cascade Image Matting with Deformable Graph Refinement: Supplementary Materials

Zijian Yu[1,2]*, Xuhui Li[1]*, Huijuan Huang[2], Wen Zheng[2], Li Chen[1]†
[1]School of Software, BNRist, Tsinghua University [2]Y-tech, Kuaishou Technology
{zj-yu19,lixh20}@mails.tsinghua.edu.cn, {huanghuijuan,zhengwen}@kuaishou.com, chenlee@tsinghua.edu.cn

1. Overview

In this supplementary material, we provide more details of the main paper. The sections and contents are summarized as follows:

- Section 2: more details of the CasDGR and visual results of the CasDGR with different neighbors' number and iteration times.
- Section 3: more comparison results and analysis on Adobe testing dataset.
- Section 4: more difficult cases and some failure cases on the real-world human images.

2. More Details and Results of CasDGR

2.1. Details of the Backbone Network

We design the backbone network in each stage of CasDGR based on the RSU block [8]. The overall pipeline has been introduced in the main paper. We show the details of each network part in Table 1. In and Out layers in Table 1 are 3×3 convolutional layers. Backbone layers are RSU-structure network.

2.2. Initial Coordinates of Neighbors

In the main paper, we provide the ablation studies of the DGR module and test different models on the Adobe testing dataset. The DGR module assume that neighbors of each pixel in the feature map have initial locations $\mathbf{p}_0$ and use a 3×3 convolutional layer to predict a 2D offset $\Delta\mathbf{p}$ for each neighbor. The adjusted coordinates of neighbors are $\mathbf{p}_0 + \Delta\mathbf{p}$.

We choose the coordinates of K adjacent pixels in space as the initial coordinates $\mathbf{p}_0$. The details are shown in Figure 1, where the colored squares represent the neighbors of the center pixel.

*Joint first authors.
†The corresponding author is Li Chen.

Layers	Input Size	Output Size
In1	$32 \times 32 \times 3$	$32 \times 32 \times 64$
Backbone1	$32 \times 32 \times 64$	$32 \times 32 \times 256$
DGR1	$32 \times 32 \times 256$	$32 \times 32 \times 64$
Out1	$32 \times 32 \times 64$	$32 \times 32 \times 1$
In2	$64 \times 64 \times 3$	$64 \times 64 \times 64$
Backbone2	$64 \times 64 \times 128$	$64 \times 64 \times 256$
DGR2	$64 \times 64 \times 256$	$64 \times 64 \times 64$
Out2	$64 \times 64 \times 64$	$64 \times 64 \times 1$
In3	$128 \times 128 \times 3$	$128 \times 128 \times 64$
Backbone3	$128 \times 128 \times 128$	$128 \times 128 \times 128$
DGR3	$128 \times 128 \times 128$	$128 \times 128 \times 64$
Out3	$128 \times 128 \times 64$	$128 \times 128 \times 1$
In4	$256 \times 256 \times 3$	$256 \times 256 \times 64$
Backbone4	$256 \times 256 \times 128$	$256 \times 256 \times 128$
DGR4	$256 \times 256 \times 128$	$256 \times 256 \times 64$
Out4	$256 \times 256 \times 64$	$256 \times 256 \times 1$
In5	$512 \times 512 \times 3$	$512 \times 512 \times 64$
Backbone5	$512 \times 512 \times 128$	$512 \times 512 \times 64$
Out5	$512 \times 512 \times 64$	$512 \times 512 \times 1$

Table 1. Details of CasDGR. The order of size is $h \times w \times c$

2.3. More Results of CasDGR

We show the comparison results of CasDGR with defferent neighbors' number and iteration times in Figure 2, 3. According to the ablation study in the main paper, CasDGR with $K = 5$ can achieve state-of-the-art results on evaluation metrics on Adobe testing dataset. The first wedding dress case in Figure 2 can illustrate the results. Compared with visualized alpha masks of $K = 1$ and $K = 9$, alpha masks of $K = 5$ have the minimum artifacts of the sidewalks in the background. However, all CasDGR models cannot handle well on the case of lady with curly hair in Figure 3. Though the network can get more precise boundaries with more neighbors, more artifacts appear in the predicted alpha mask.

For iteration times, more iteration times can further improve the details of foreground human. Due to the limitation

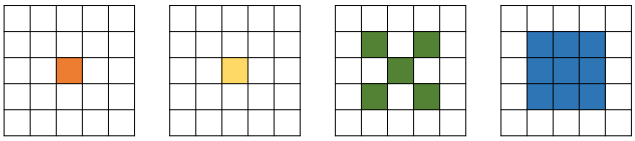

Figure 1. Initial coordinates of different number of neighbors.

of memory cost and training stability, we set the maximum number of iterations to 2.

3. More Results on the Adobe Testing Dataset

Similar to the experiments in the main paper, we compare our approach on the constructed Adobe human image dataset with three kinds of available approaches. **The traditional methods**: Closed-Form Matting (CFM) [5], Learning Based Matting (LBM) [13], KNN Matting (KNNM) [2], Random Walks Matting (RWM) [3], and Large Kernels Matting (LKM) [4]. **The trimap-based learning methods**: Deep Image Matting (DIM) [10], IndexNet Matting (IM) [7], and Guided Contextual Attention Matting (GCAM) [6]. **The automatic learning methods**: Late Fusion Matting (LFM) [12] and Background Matting (BGM) [9].

Similarly, during the evaluation, we resize input images to 512×512 resolution to predict the alpha mattes and compute four metrics between the predicted results and ground truths. Nevertheless, different from the experiments in the main paper, we perform erosion and dilation operations on GT alpha mattes to obtain the corresponding trimaps. In order to explore the influence of trimaps quality on this kind of methods, we use 10 and 20 iterations of dilation to generate different trimaps and feed them into the model. In addition, the manner of generating segmentation results and disturbed backgrounds are also the same as those in BGM [9].

The quantitative results are shown in Table 2. The implications of our experimental results are as follows:

After trying different kinds of trimaps, our CasDGR can still achieve state-of-the-art results on all metrics among all testing approaches on Adobe testing dataset, i.e., the traditional methods, trimap-based, and automatic methods mentioned above.

Most of the trimap-based methods, including traditional methods and deep learning methods, are sensitive to the quality of trimaps. When the quality of trimaps decreases, the alpha mattes will also decline in quality significantly. It is time-consuming and labor-consuming to construct a high-quality trimap and a low-quality trimap may lead to a much worse result. Similarly, BGM [9] is also influenced by the quality of the input backgrounds. When the disturbances in input backgrounds increases, all of the metrics of predicted alpha mattes will decrease. These comparison results show the advantages of automatic matting methods without any additional inputs.

Method	SAD	MSE	Grad	Conn
CFM [5] - *T-10*	4.46	0.0062	5.27	4.09
CFM [5] - *T-20*	6.67	0.0105	7.14	6.02
LBM [13] - *T-10*	4.58	0.0064	5.37	4.23
LBM [13] - *T-20*	6.75	0.0106	7.29	6.13
KNNM [2] - *T-10*	4.83	0.0062	4.95	4.46
NNM [2] - *T-20*	6.68	0.0093	6.15	5.96
RWM [3] - *T-10*	5.62	0.0105	10.13	5.48
RWM [3] - *T-20*	7.16	0.0142	11.58	6.87
LKM [4] - *T-10*	6.47	0.0073	6.37	5.37
LKM [4] - *T-20*	8.17	0.0104	7.57	6.81
IM [7] - *T-10*	2.44	0.0028	3.54	2.24
IM [7] - *T-20*	3.42	0.0048	5.48	3.14
DIM [10] - *T-10*	3.83	0.0040	4.68	3.31
DIM [10] - *T-20*	5.72	0.0063	5.88	4.77
GCAM [6] - *T-10*	1.95	0.0018	2.17	1.74
GCAM [6] - *T-20*	2.09	0.0020	2.43	1.85
BGM [9] - *Seg,* B'	2.30	0.0025	2.34	2.10
BGM [9] - *Seg,* B	2.28	0.0024	2.29	2.08
LFM [12]	4.35	0.0067	4.01	3.98
Ours-Baseline	3.78	0.0065	4.67	3.73
Ours-Cascade	2.92	0.0046	2.85	2.77
Ours-CasDGR	**1.76**	**0.0015**	**1.66**	**1.54**

Table 2. Results on the Adobe testing dataset. *T-10*, *T-20*: the input trimaps are generated through 10 or 20 iterations of dilation from GT alpha mattes. *Seg,* B', B: coarse segmentation results, disturbed backgrounds with Gaussian noises, and original backgrounds for Background-Matting [9]. Ours-Baseline: the single encoder–decoder network contains 1 RSU block. Ours-Cascade: the cascade network consists of 5 stages without DGR module. Ours-CasDGR: the cascade network with DGR modules.

4. More Results on Real-World Human Images

In this section, we provide more real-world matting results, including some difficult cases in Figure 4. The testing images are from 1) human matting dataset [1] and 2) Real World Portrait-636 dataset [11]. The visual results demonstrate that our CasDGR can also achieve good performance on real-world images even in some complex scenes.

Although our CasDGR has a good matting effect on many real pictures, there are still some failure cases when the situation is intractable. we report some typical failure cases on real-world images and explore the reasons below.

Several typical failure cases are shown in Figure 5. The failure case in the left image are due to the fusion of human hair and the background, which makes it difficult to extract fine boundaries. In this case, it is even difficult for humans to distinguish the hair boundary from the background pre-

cisely. The woman in middle image have hair with very complex shapes. Our CasDGR failed to generate the detail of all hairs. In the right image, the woman has her back to the carema, which is not common in training dataset. Besides, the area below the image is very dark, making it difficult to distinguish human from the beach.

In the future work, we will try to make our method overcome the above difficulties and achieve better results on real-world images.

References

[1] Quan Chen, Tiezheng Ge, Yanyu Xu, Zhiqiang Zhang, Xinxin Yang, and Kun Gai. Semantic human matting. *ACM International Conference on Multimedia*, pages 618–626, 2018.

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

[3] Leo Grady, Thomas Schiwietz, Shmuel Aharon, and Rüdiger Westermann. Random walks for interactive alpha-matting. *VIIP*, 2005:423–429, 2005.

[4] Kaiming He, Jian Sun, and Xiaoou Tang. Fast matting using large kernel matting laplacian matrices. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1:2165–2172, 2010.

[5] 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] Yaoyi Li and Hongtao Lu. Natural image matting via guided contextual attention. *AAAI*, pages 11450–11457, 2020.

[7] Hao Lu, Yutong Dai, Chunhua Shen, and Songcen Xu. Indices matter: Learning to index for deep image matting. *International Conference on Computer Vision (ICCV)*, pages 3265–3274, 2019.

[8] Xuebin Qin, Zichen Vincent Zhang, Chenyang Huang, Masood Dehghan, Osmar R. Zaïane, and Martin Jägersand. U^2-net: Going deeper with nested u-structure for salient object detection. *Pattern Recognition*, 106:107404, 2020.

[9] Soumyadip Sengupta, Vivek Jayaram, Brian Curless, Steven M. Seitz, and Ira Kemelmacher-Shlizerman. Background matting: The world is your green screen. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2288–2297, 2020.

[10] Ning Xu, Brian L. Price, Scott Cohen, and Thomas S. Huang. Deep image matting. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 311–320, 2017.

[11] Qihang Yu, Jianming Zhang, He Zhang, Yilin Wang, Zhe Lin, Ning Xu, Yutong Bai, and Alan L. Yuille. Mask guided matting via progressive refinement network. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[12] Yunke Zhang, Lixue Gong, Lubin Fan, Peiran Ren, Qixing Huang, Hujun Bao, and Weiwei Xu. A late fusion CNN for digital matting. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7469–7478, 2019.

[13] Yuanjie Zheng and Chandra Kambhamettu. Learning based digital matting. *International Conference on Computer Vision (ICCV)*, pages 889–896, 2009.

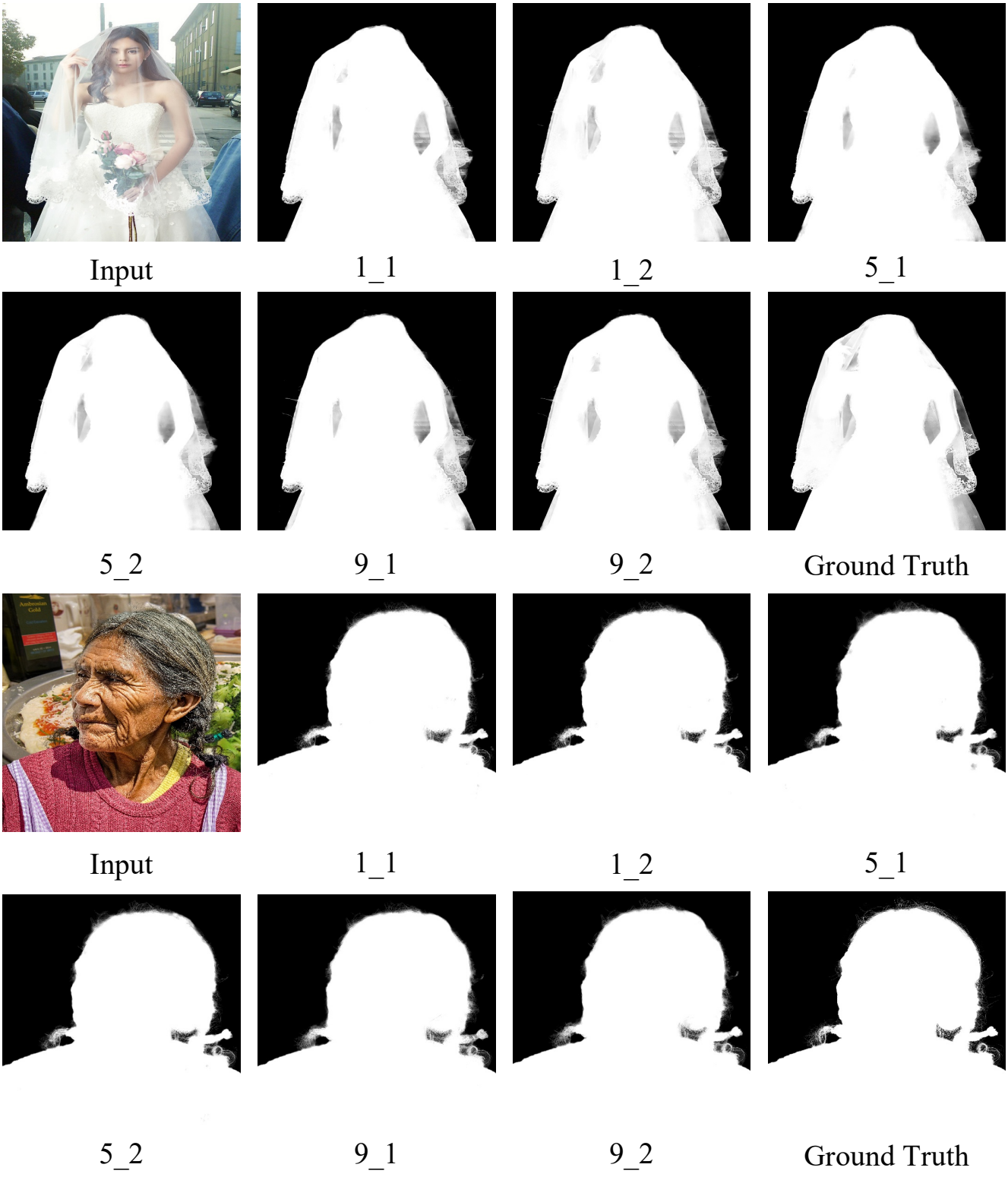

Figure 2. More results of CasDGR. 1_2 means $K = 1$ and $layers = 2$ (iterations).

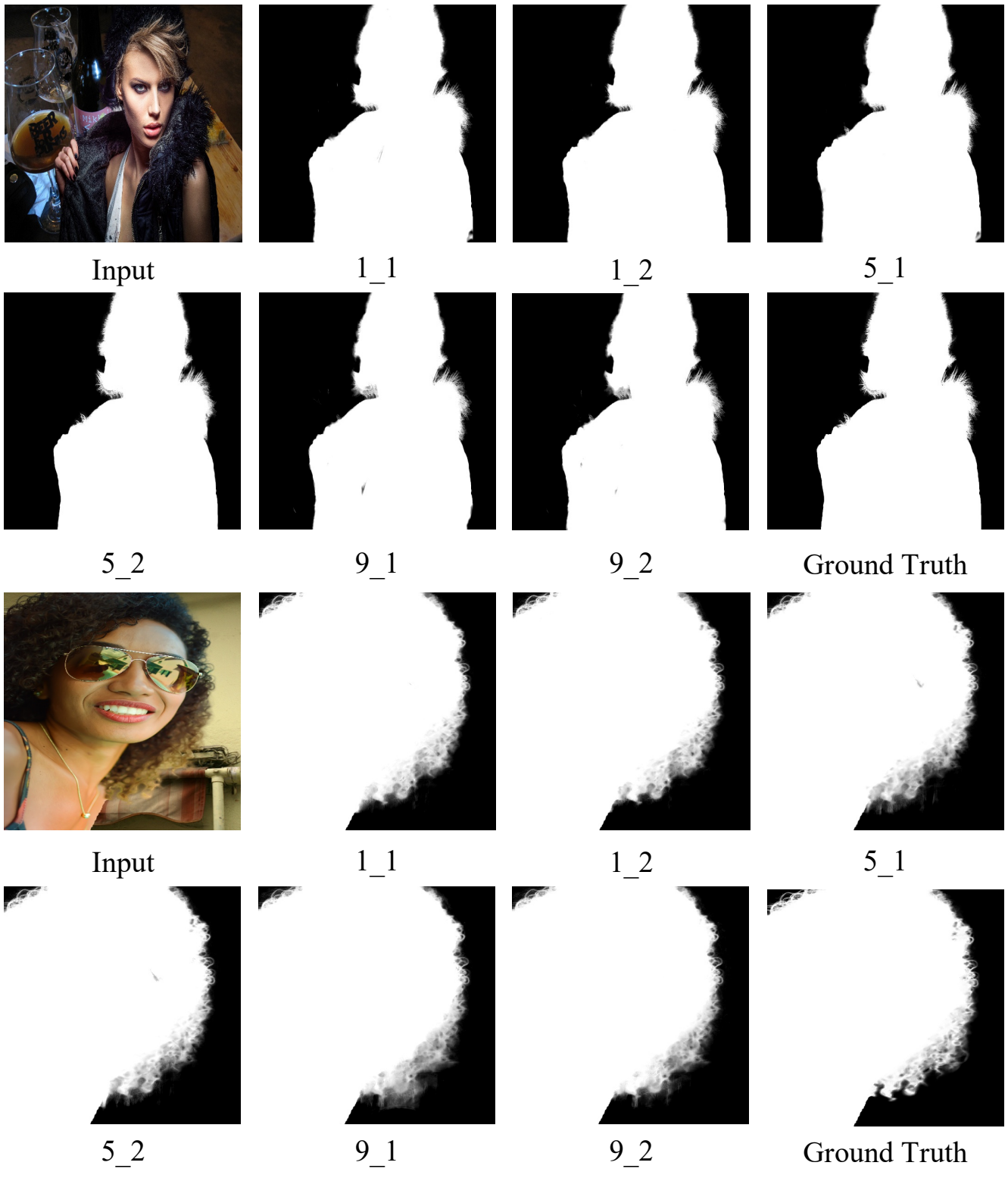

Figure 3. More results of CasDGR. 1_2 means $K = 1$ and $layers = 2$ (iterations).

Figure 4. More real-world cases. Row 1-3: Real World Portrait-636 dataset. Row 4: human matting dataset.

Figure 5. Some failure cases on real-world images.