Semantic Image Matting

Supplementary Material

1. Dataset

The amount of training samples on 20 classes in our Semantic Image Matting Dataset in listed in Table 1.

class	no.	class	no.	class	no.
defocus	32	motion	30	sharp	65
fire	42	net	52	smoke_cloud	46
fur	46	flower	25	lace	25
glass_ice	35	leaf	15	silk	35
hair_easy	66	tree	18	water_drop	41
hair_hard	36	spider_web	22	water_spray	38
insect	22	plastic_bag	30		

Table 1. Amount of 20 classes in Semantic Image Matting Dataset.

Methods	SAD	MSE	Grad	Conn
FBA [4]	**26.4**	**5.4**	**10.6**	**21.5**
FBA (rep) [4]	32.8	7.3	13.4	27.9
SIM (Ours)	28.0	5.8	10.8	24.8

Table 2. Comparison with FBA-Net [4] on the Composition-1K [11]. FBA and FBA (rep) are respectively the public released model and our reproduced model.

Methods	SAD	MSE	Grad	Conn
FBA [4]	35.08	7.5	14.54	29.12
FBA (rep) [4]	30.50	5.5	12.03	22.72
SIM (Ours)	**27.87**	**4.7**	**11.57**	**20.83**

Table 3. Comparisons with FBA-Net [4] on the Semantic Image Matting Dataset.

2. Implementation

2.1. Classifier

The architecture of our two classifiers is derived from ResNet-50 [5]. They are trained by Adam optimizer with an initial learning rate at 1e-2 and a linear decay by 0.1 in every 20 epochs. We train each classifier for totally 60 epochs with a batch size 16. We augment the training samples with random scaling within range $[160, 640]$, random rotation within range $[-45, 45]$, random horizontal flip and color jittering including contrast, brightness as well as hue.

2.2. Encoder-Decoder Network

In experiments, for the encoder, we change one 7×7 convolutional layer in ResNet-50 to three 3×3 convolutional layers and keep the rest unchanged. For the ASPP module, we use $(12, 24, 36)$ as the atrous rates to enlarge receptive fields. After the decoder, we use 4 prediction heads comprising of 3 convolutional layers for predicting F, B, α and learnable weights. Except for these prediction heads, we use batch normalization in the network. We use clipping and sigmoid activation respectively for training α-head and F, B-head. We initialize our encoder network with pretrained weights on ImageNet [3] dataset and the decoder network with Xavier random variables.

2.3. Training Schedule

We train our network for totally 50 epochs with batch size of 6 and initial learning rate as 6e-5 which decays at a rate of 0.1 in the last 5 epochs. In each epoch, we traverse all foreground images with random background images for 100 times. RAdam [7] optimizer is adopted to update parameters for the whole network with a weight decay of 0.005.

3. Experiments

3.1. Results on alphamatting.com

In Table 4, we provide the evaluation results on SAD, MSE and Gradient of our method and 5 representative state-of-the-art models across all test samples. Our method ranks first on most of the samples under different trimap sizes.

3.2. Comparisons with FBA-Net

Table 2 and Table 3 respectively list the quantitative comparison results with FBA-Net [4] on Composition-1k dataset and our Semantic Image Matting Dataset. We provide two evaluation results of FBA. One is the public released model, and the other is our own reproduced model.

3.3. More Qualitative Results

In the paper, we visualize the comparisons among our method and existing state-of-the-art methods including DIM [11], IndexNet [8] and GCA [6]. More qualitative results on Semantic Image Matting Dataset are shown in Figure 9 to Figure 14, where the first and second rows respectively show predicted alpha mattes and composition results. From left to right are Image (trimap), the results of DIM, IndexNet, GCA, SIM and groundtruth. In addition, Figure 15 shows more qualitative results on real-world images of various matting scenarios. Extensive results demonstrate the

Methods	SAD																							
	Troll			Doll			Donkey			Elephant			Plant			Pineapple			Plastic bag			Net		
	S	L	U	S	L	U	S	L	U	S	L	U	S	L	U	S	L	U	S	L	U	S	L	U
DIM [11]	10.7	11.2	11	4.8	5.8	5.6	2.8	2.9	2.9	1.1	1.1	2	6	7.1	8.9	2.7	3.2	3.9	19.2	19.6	18.7	21.8	23.9	24.1
SampleNet [10]	9.1	9.7	9.8	**4.3**	**4.8**	**5.1**	3.4	3.7	3.2	0.9	1.1	2	5.1	6.8	9.7	2.5	4	3.7	18.6	19.3	19.1	20	21.6	23.2
AdaMatting [2]	10.2	11.1	10.8	4.9	5.4	6.6	3.6	3.4	3.4	**0.9**	**0.9**	1.8	**4.7**	6.8	9.3	2.2	2.6	3.3	19.2	19.8	18.7	17.8	19.1	**18.6**
GCA [6]	8.8	9.5	11.1	4.9	4.8	5.8	3.4	3.7	3.2	1.1	1.2	1.3	5.7	6.9	7.6	2.8	3.1	4.5	18.3	19.2	18.5	20.8	21.7	24.7
Background [9]	9.3	10	10.1	4.5	5.1	6.7	2.9	3.3	2.9	1	1.2	2.2	5.2	6	7.8	2.8	3.4	4.3	16.4	17.3	16.4	19.5	20.9	27.9
SIM (Ours)	**8.3**	**8.7**	**9**	4.8	4.8	6	**2.2**	**2.2**	**2**	0.9	0.9	**1.1**	4.7	**5.1**	**6.3**	**2.2**	**2.3**	**2.5**	**15.9**	**16.3**	**16.3**	**17.8**	**18**	20.9

Methods	MSE																							
	Troll			Doll			Donkey			Elephant			Plant			Pineapple			Plastic bag			Net		
	S	L	U	S	L	U	S	L	U	S	L	U	S	L	U	S	L	U	S	L	U	S	L	U
DIM [11]	0.4	0.4	0.4	0.2	0.3	0.3	0.1	0.1	0.2	0	0	0.2	0.5	0.6	1	0.2	0.2	0.4	1.1	1.1	1.1	0.8	0.9	1
SampleNet [10]	0.3	0.3	**0.3**	**0.1**	**0.1**	**0.2**	0.2	0.2	0.2	0	0	0.1	0.4	0.6	1.2	0.1	0.3	0.3	1.1	1.1	1.2	0.7	0.8	0.8
AdaMatting [2]	0.3	0.4	0.4	0.2	0.2	0.3	0.2	0.2	0.2	**0**	**0**	0.1	0.4	0.6	1	**0.1**	0.2	0.3	1.1	1.2	1.1	0.6	0.6	**0.6**
GCA [6]	**0.3**	0.3	0.4	0.2	0.2	0.3	0.2	0.2	0.2	0	0	0.1	0.5	0.6	0.8	0.2	0.2	0.5	1	1.1	1.1	0.7	0.8	0.9
Background [9]	0.3	0.3	0.3	0.2	0.2	0.4	0.1	0.2	0.1	0	0	0.1	0.3	0.4	0.7	0.1	0.2	0.3	0.9	1	0.9	0.7	0.8	1.3
SIM (Ours)	0.3	**0.3**	0.3	0.2	0.2	0.3	**0.1**	**0.1**	**0.1**	0	0	**0**	**0.3**	**0.3**	**0.5**	0.1	**0.1**	**0.2**	**0.8**	**0.8**	**0.9**	**0.6**	**0.6**	0.9

Methods	Gradient																							
	Troll			Doll			Donkey			Elephant			Plant			Pineapple			Plastic bag			Net		
	S	L	U	S	L	U	S	L	U	S	L	U	S	L	U	S	L	U	S	L	U	S	L	U
DIM [11]	0.4	0.4	0.5	0.2	0.2	0.2	0.1	0.1	0.2	0.2	0.2	0.6	1.3	1.5	2.4	0.8	0.9	1.3	0.7	0.8	1.1	0.4	0.5	0.5
SampleNet [10]	**0.1**	**0.1**	0.2	0.1	0.1	**0.2**	0.2	0.3	0.3	0.1	0.2	0.5	1.1	1.5	2.7	0.6	0.9	1	0.8	0.9	0.9	0.4	0.4	0.4
AdaMatting [2]	0.2	0.2	0.2	**0.1**	**0.1**	0.4	0.2	0.2	0.2	**0.1**	**0.1**	0.3	1.1	1.4	2.3	**0.4**	0.6	0.9	0.9	1	0.9	0.3	0.4	0.4
GCA [6]	0.1	0.1	0.2	0.1	0.1	0.3	0.2	0.2	0.2	0.2	0.2	0.3	1.3	1.6	1.9	0.7	0.8	1.4	0.6	0.7	0.6	0.4	0.4	0.4
Background [9]	0.2	0.2	0.2	0.1	0.2	0.3	**0.1**	0.2	0.2	0.2	0.2	0.5	**0.9**	1.1	1.7	0.6	0.8	1.2	0.6	0.7	0.8	0.2	0.3	0.4
SIM (Ours)	0.2	0.2	**0.2**	0.1	0.1	0.3	0.2	**0.1**	**0.2**	0.2	0.1	**0.3**	1	**1**	**1.4**	0.5	**0.5**	**0.7**	**0.4**	**0.5**	**0.5**	**0.2**	**0.2**	**0.3**

Table 4. Quantitative results of our method and representative state-of-the-art methods on alphamatting.com [1] across 8 test samples. "S", "L", "U" denote different trimap sizes, small, large and user. Best results are in bold.

superiority of our model to other methods.

References

[1] alphamatting. http://www.alphamatting.com/. 2

[2] Shaofan Cai, Xiaoshuai Zhang, Haoqiang Fan, Haibin Huang, Jiangyu Liu, Jiaming Liu, Jiaying Liu, Jue Wang, and Jian Sun. Disentangled image matting. In *ICCV*, 2019. 2

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 1

[4] Marco Forte and François Pitié. F, b, alpha matting. *arXiv preprint arXiv:2003.07711*, 2020. 1

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1

[6] Yaoyi Li and Hongtao Lu. Natural image matting via guided contextual attention. In *AAAI*, 2020. 1, 2

[7] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. On the variance of the adaptive learning rate and beyond. In *ICLR*, 2020. 1

[8] Hao Lu, Yutong Dai, Chunhua Shen, and Songcen Xu. Indices matter: Learning to index for deep image matting. In *ICCV*, 2019. 1

[9] Soumyadip Sengupta, Vivek Jayaram, Brian Curless, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Background matting: The world is your green screen. In *CVPR*, 2020. 2

[10] Jingwei Tang, Yagiz Aksoy, Cengiz Oztireli, Markus Gross, and Tunc Ozan Aydin. Learning-based sampling for natural image matting. In *CVPR*, 2019. 2

[11] Ning Xu, Brian Price, Scott Cohen, and Thomas Huang. Deep image matting. In *CVPR*, 2017. 1, 2

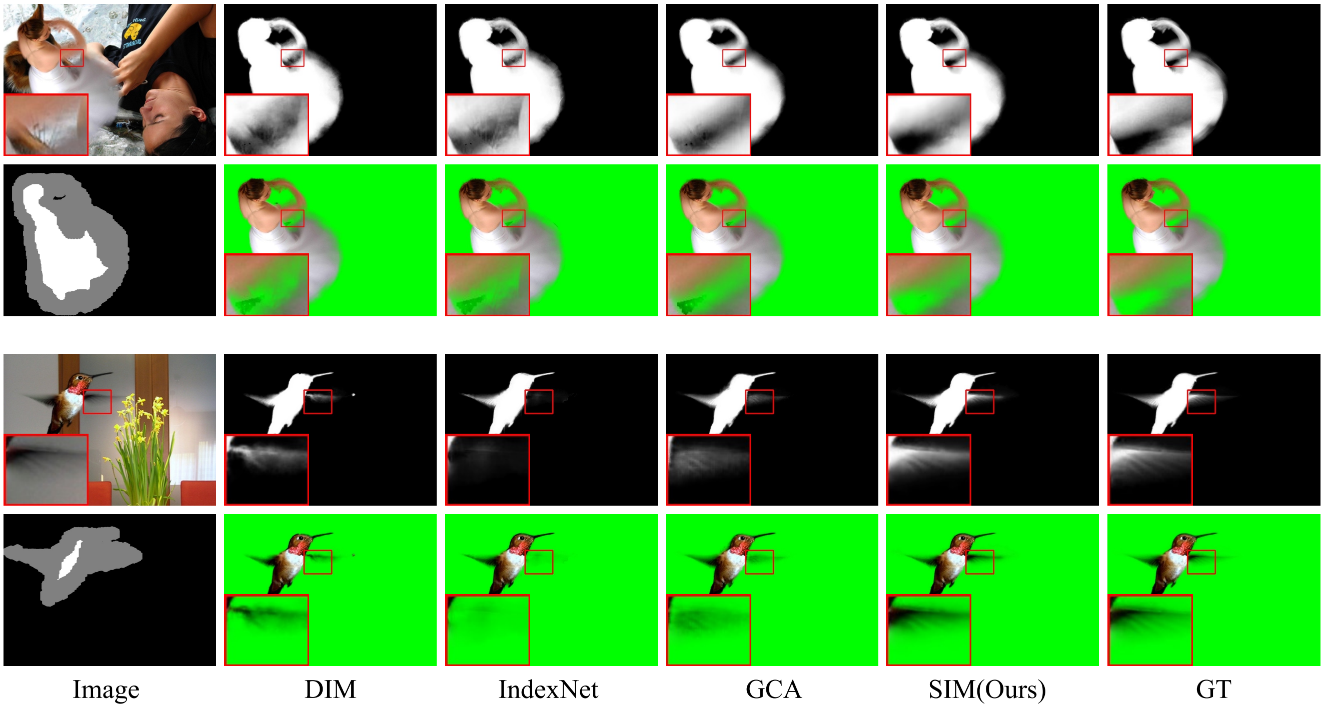

Figure 1. Qualitative comparisons on class $motion$.

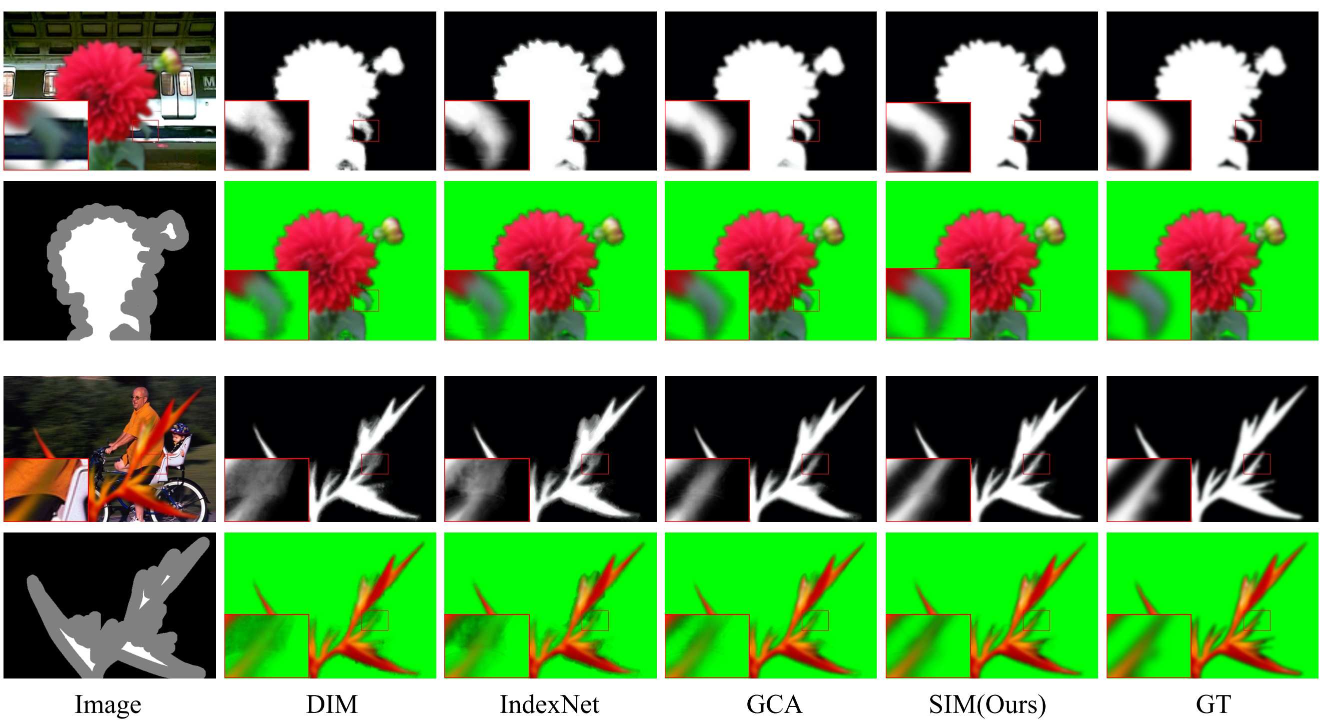

Figure 2. Qualitative comparisons on class $defocus$.

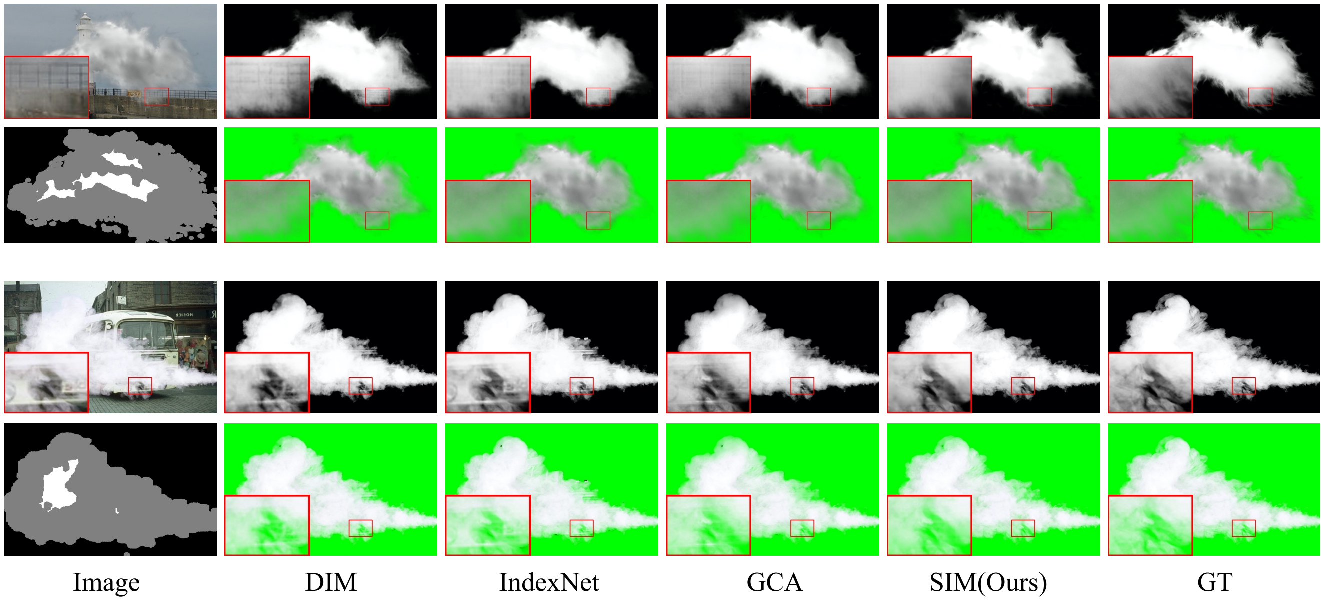

Figure 3. Qualitative comparisons on class $smoke\_cloud$.

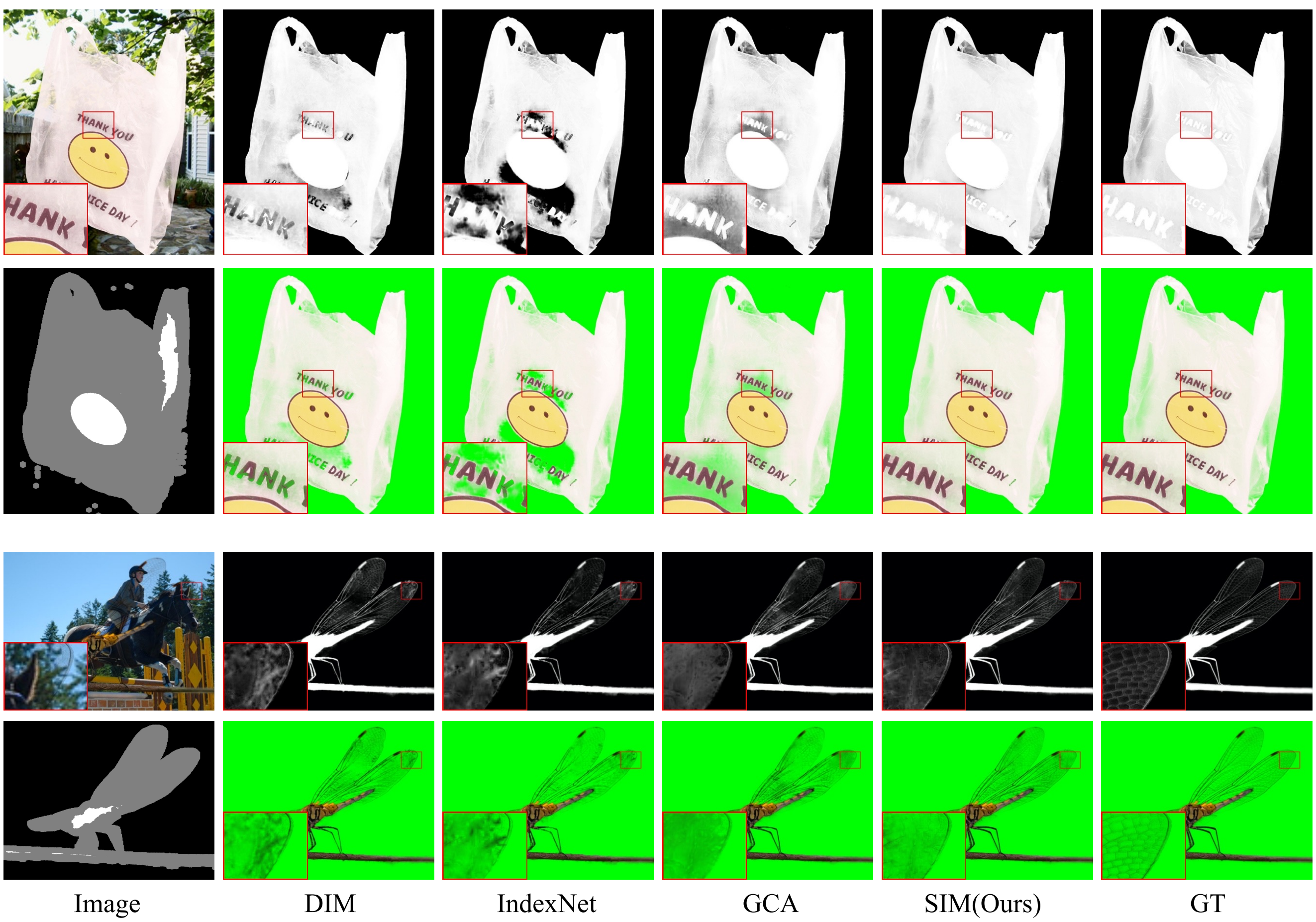

Figure 4. Qualitative comparisons on class $plastic\_bag$ (upper) and $insect$ (lower).

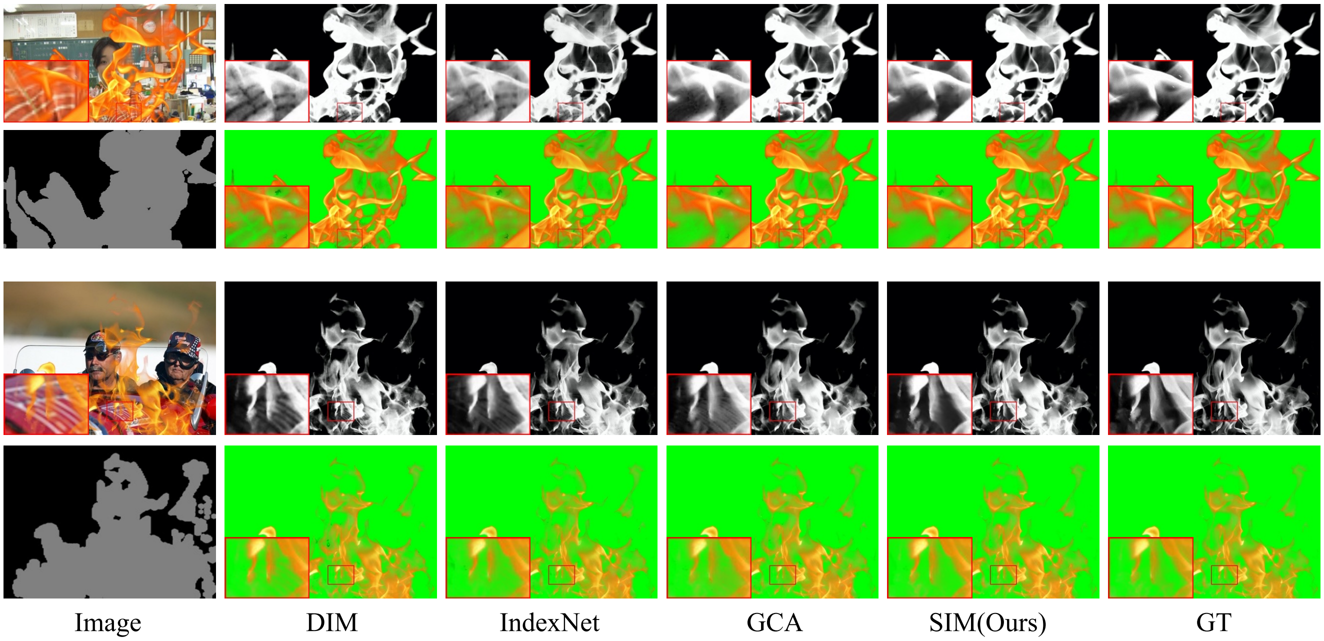

Figure 5. Qualitative comparisons on class $fire$.

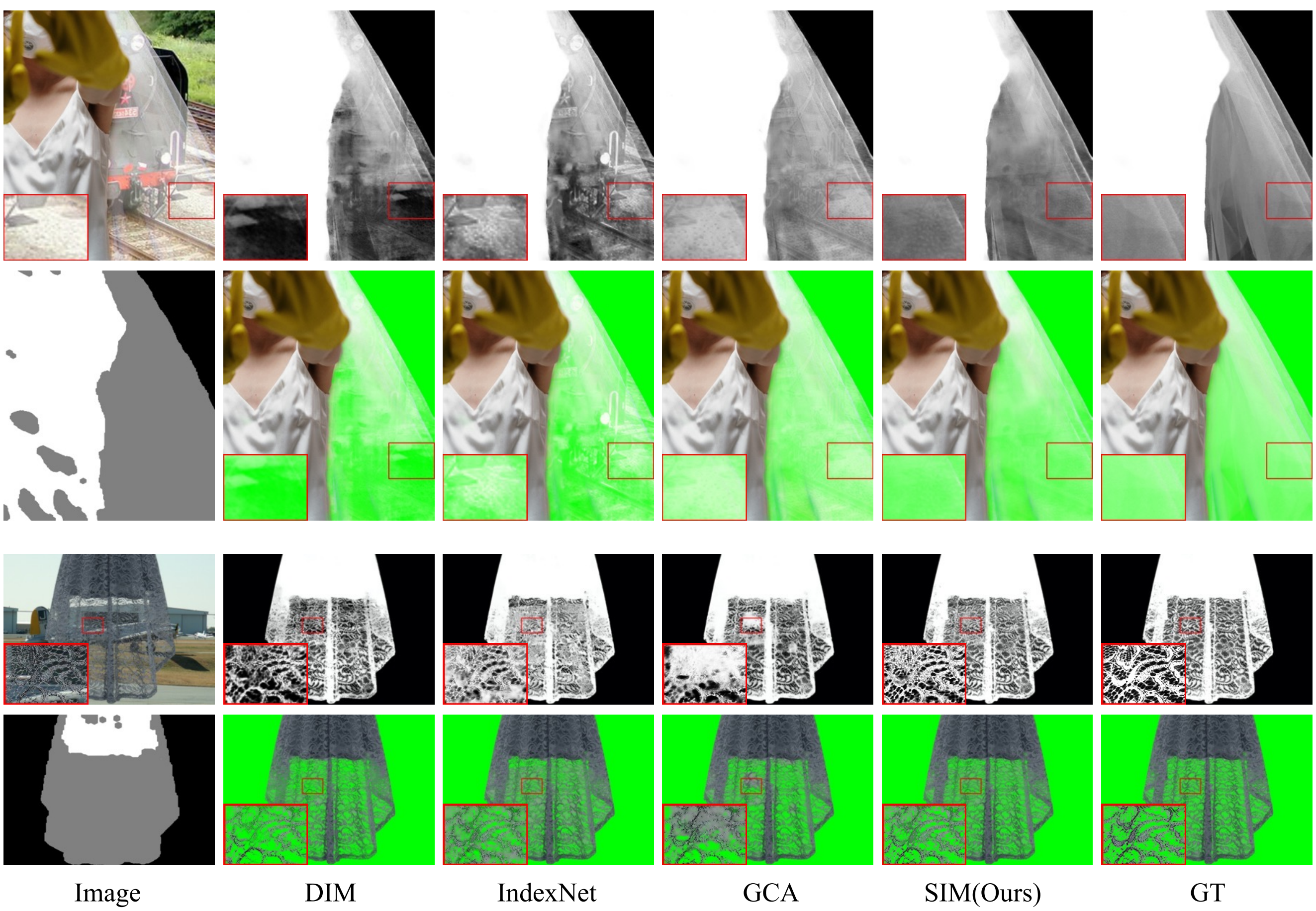

Figure 6. Qualitative comparisons on class $silk$ (upper) and $lace$ (lower).

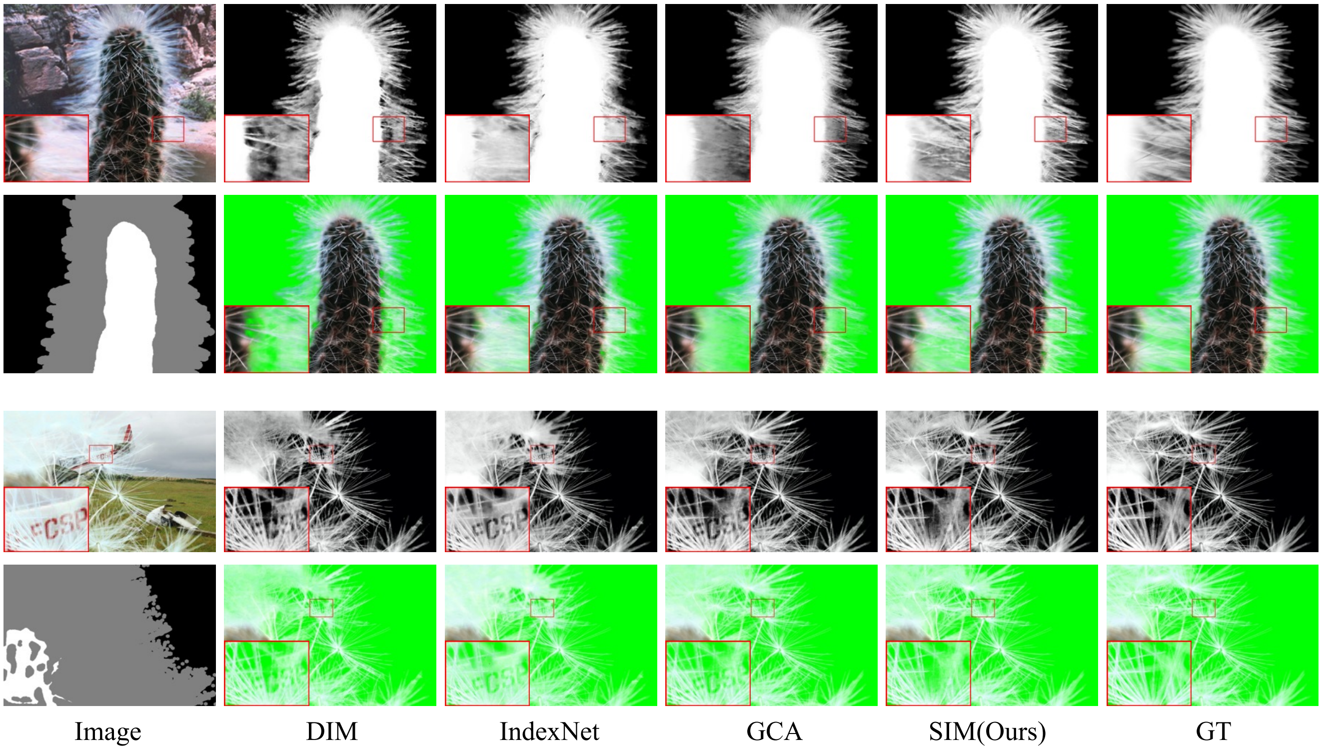

Figure 7. Qualitative comparisons on class $flower$.

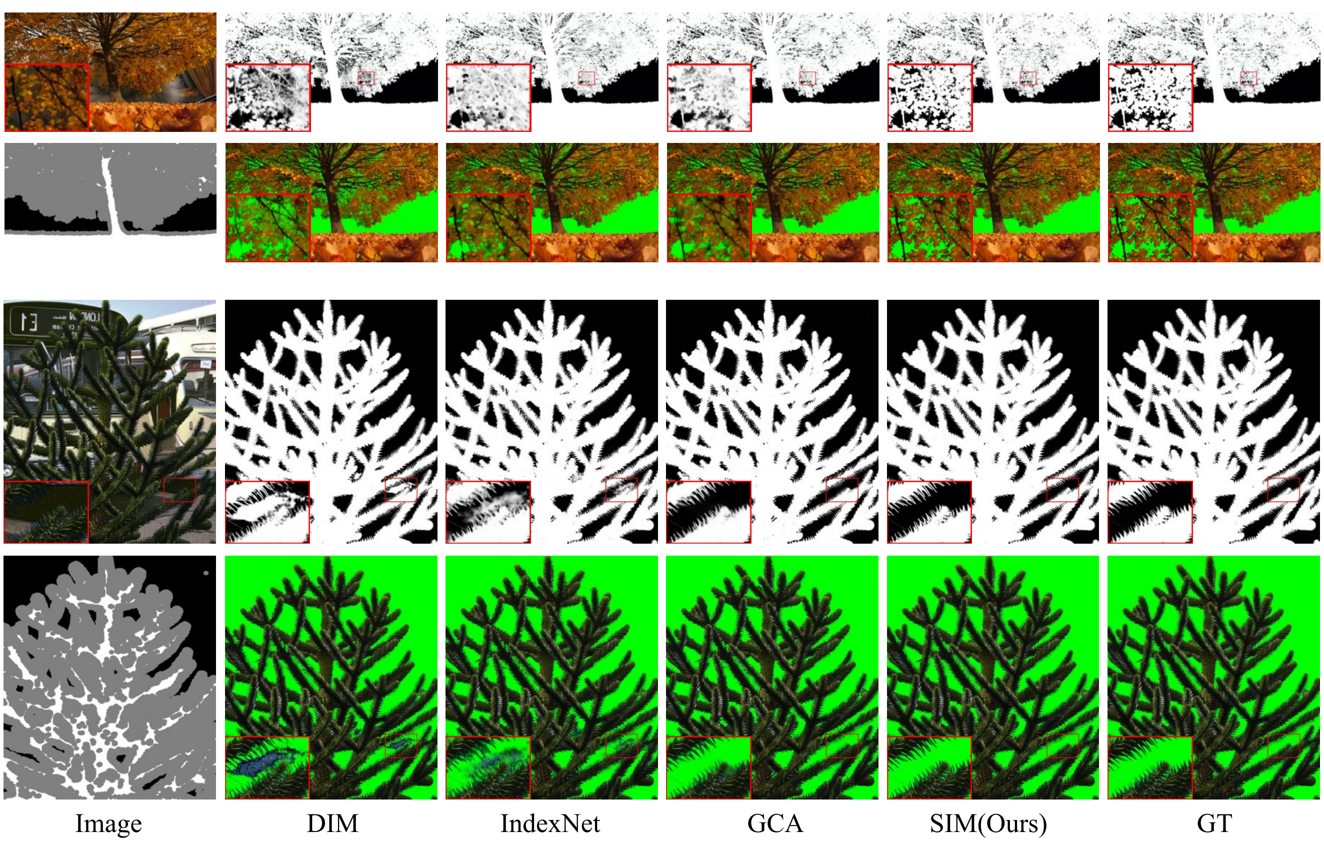

Figure 8. Qualitative comparisons on class $tree$ (upper) and $leaf$ (lower).

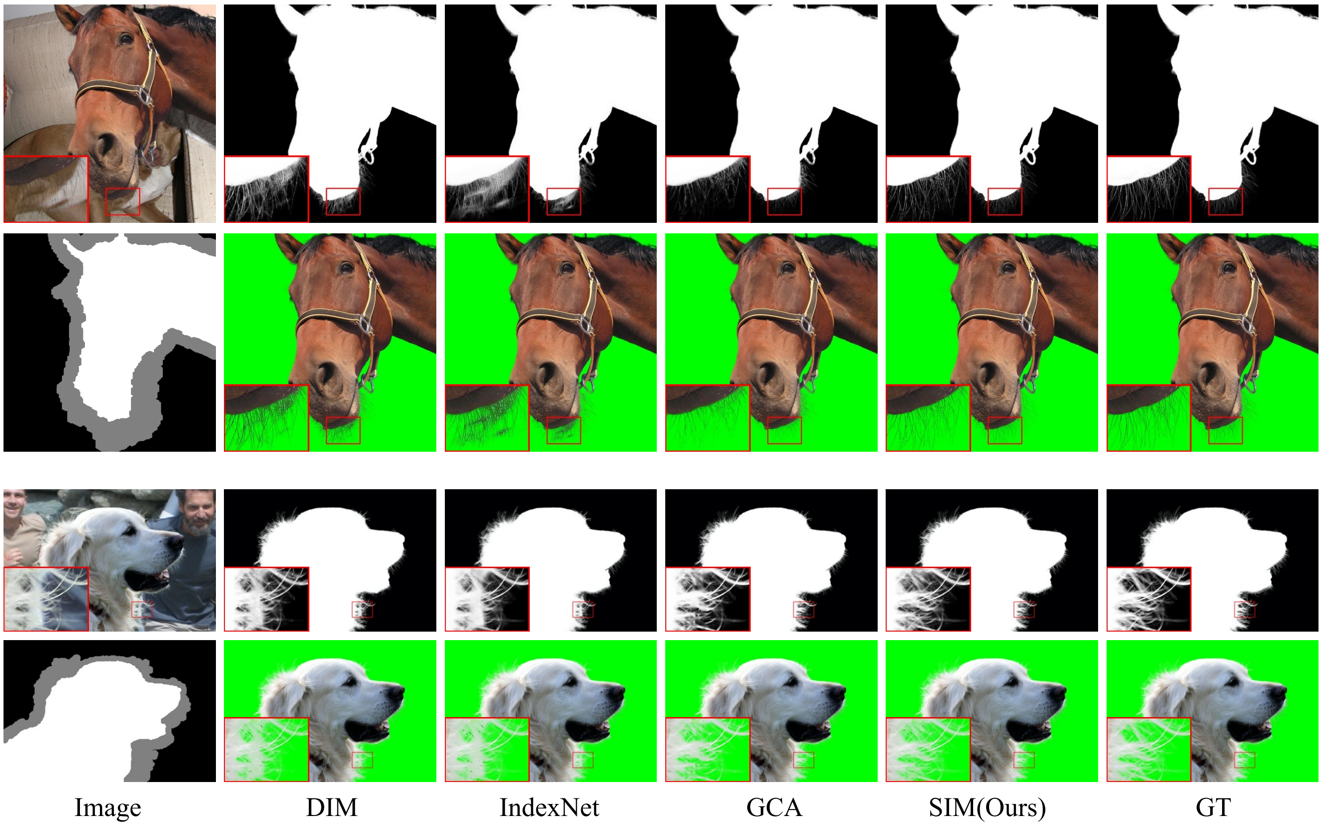

Figure 9. Qualitative comparisons on class fur.

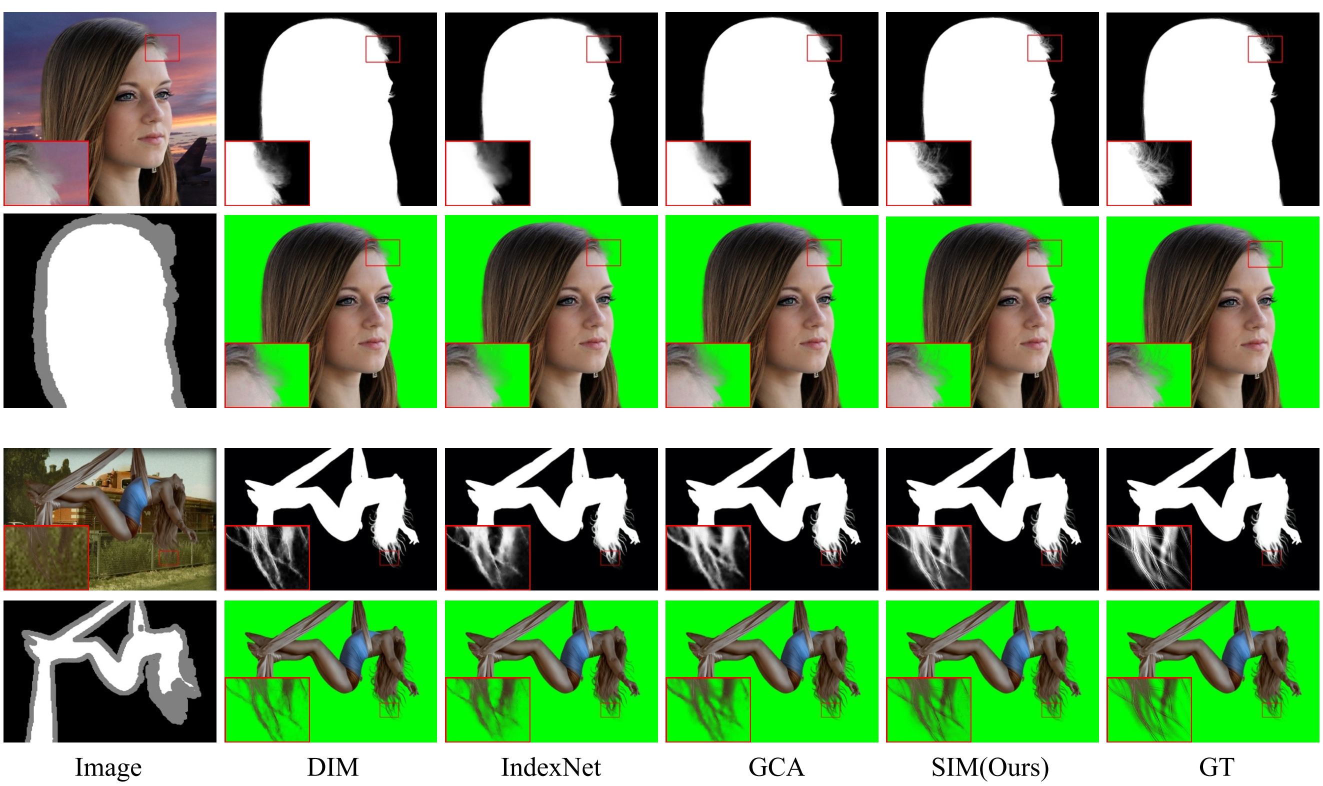

Figure 10. Qualitative comparisons on class $hair\_easy$ (upper) and $hair\_hard$ (lower).

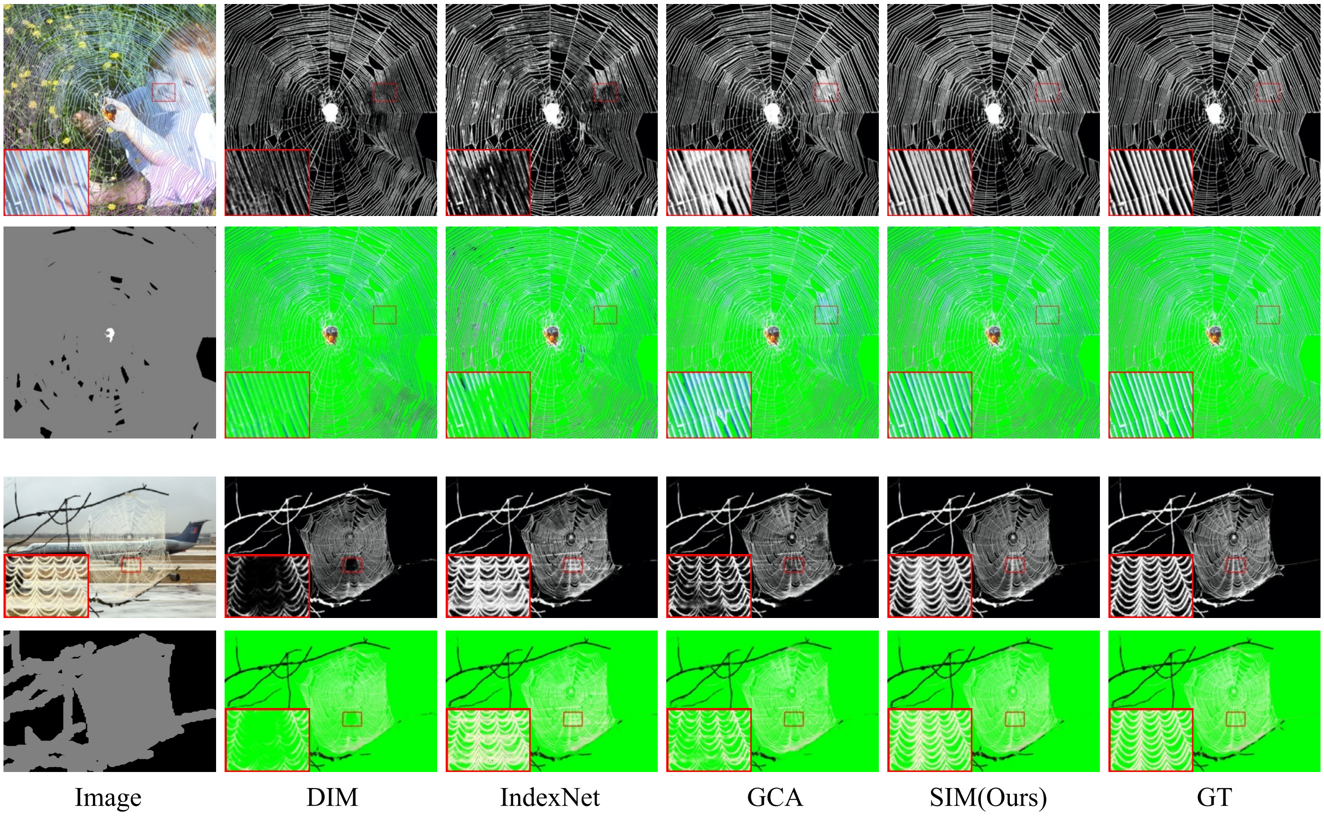

Figure 11. Qualitative comparisons on class $spider\_web$.

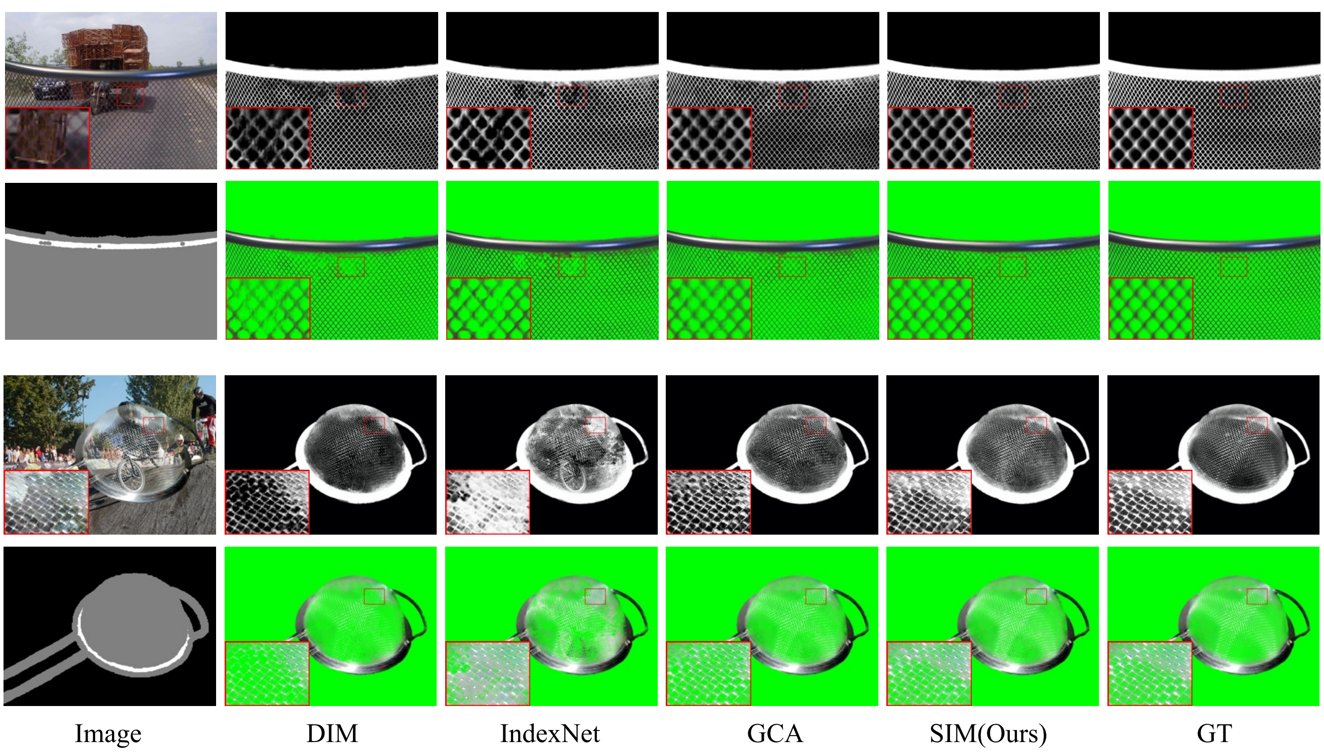

Figure 12. Qualitative comparisons on class net.

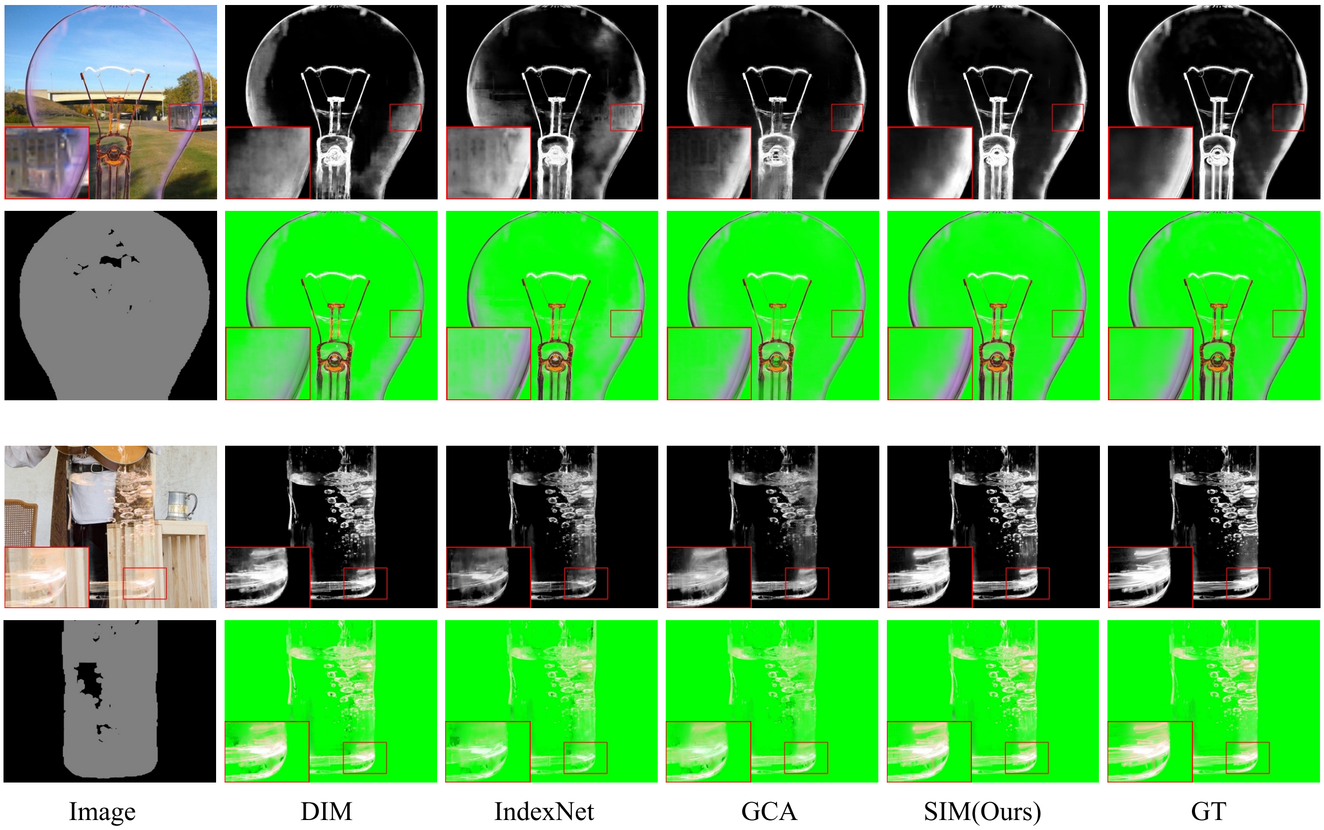

Figure 13. Qualitative comparisons on class $glass\_ice$.

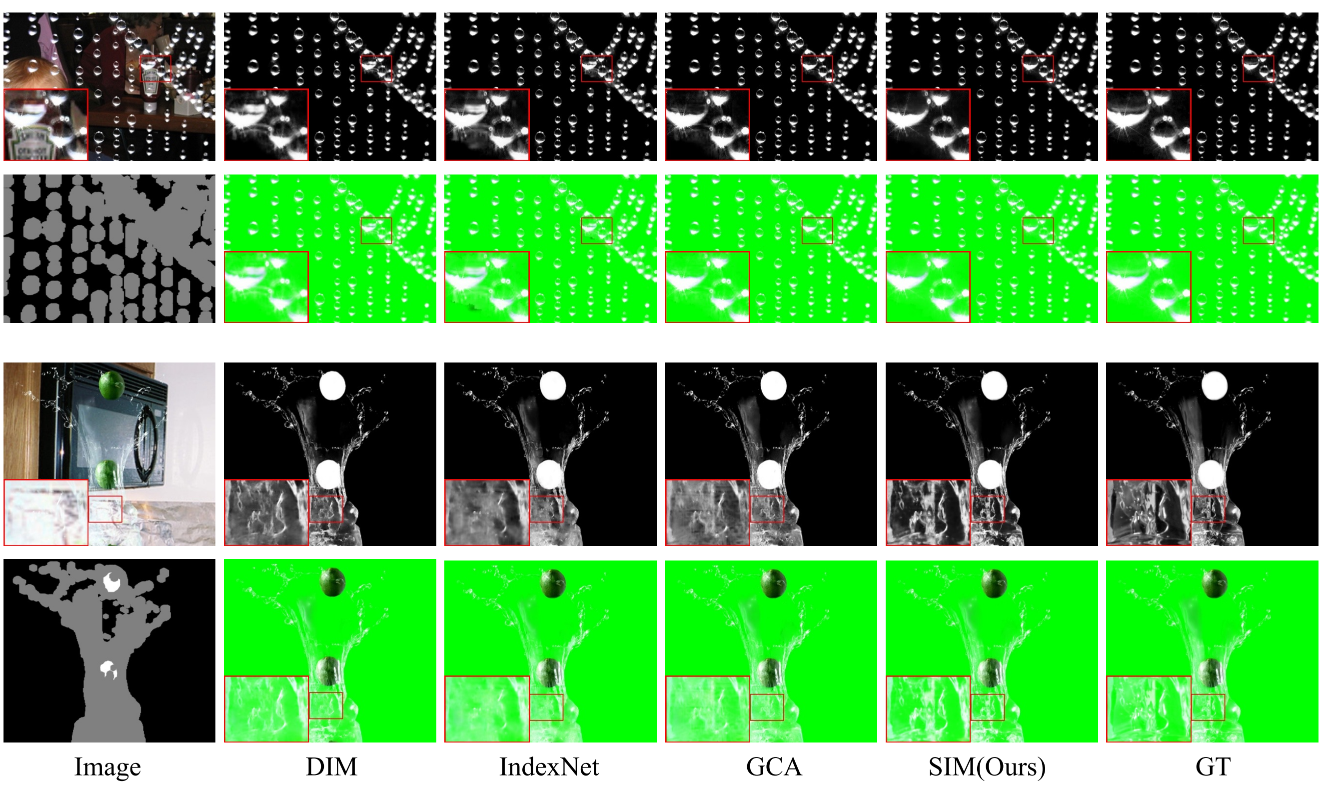

Figure 14. Qualitative comparisons on class $water\_drop$ (upper) and $water\_spray$ (lower).

Figure 15. Qualitative comparisons on real-world images.