Attention-Guided Hierarchical Structure Aggregation for Image Matting: Supplementary Materials

Yu Qiao[1,*], Yuhao Liu[1,*],Xin Yang[1,4,†],Dongsheng Zhou[2],Mingliang Xu[3],Qiang Zhang[2],Xiaopeng Wei[1,†]
[1] Dalian University of Technology, [2] Dalian University, [3] Zhengzhou University
[4] Beijing Technology and Business University
{coachqiao2018,yuhaoLiu7456}@gmail.com,{xinyang,zhangq,xpwei}@dlut.edu.cn,donyson@126.com
iexumingliang@zzu.edu.cn

1. Introduction

In this supplementary material, we first compare different strategies to extract pyramidal features and appearance cues, which can demonstrate the effectiveness of the proposed *HAttMatting*. Then we provide additional image matting results on the public Adobe Composition-1k dataset, our Distinctions-646 dataset and real-world images.

2. Alternative Features Extraction Strategies

In the pipeline of our *HAttMatting*, we extract high-level semantic features from ResNeXt [6] block4, then feed them to ASPP [1] module to capture pyramidal features. The original intention of this design is to consider that foreground objects always occupy the majority of the input images in image matting, and there is no need to design multi-scale framework to capture objects of various sizes.

Actually, we have tried other network architectures to obtain pyramidal features. As shown in Fig. 1, we attempt to concatenate the feature maps from block3 and block4 using these two potential network structures. Theoretically speaking, we can give the block3 branch a weight map 0 to degenerate the architectures in Fig. 1 into our pipeline. However, it is difficult to achieve the same accuracy as our pipeline by combining the semantic features of block3 in the training process. The quantitative comparisons are shown in Tab. 1, the implementation details of all models are the same with our *HAttMatting* and all the results are evaluated on the Composition-1k test set. The alpha mattes produced by our *HAttMatting* are better than the other features extraction strategies. In our analysis, we argue that the foreground object occupy most of the input image and the top layer of the network can represent the most important semantic information. The top semantics can suggest the foreground object in image matting, and the external branches like Fig. 1 can affect the convergence of the training process on the foreground to some extent.

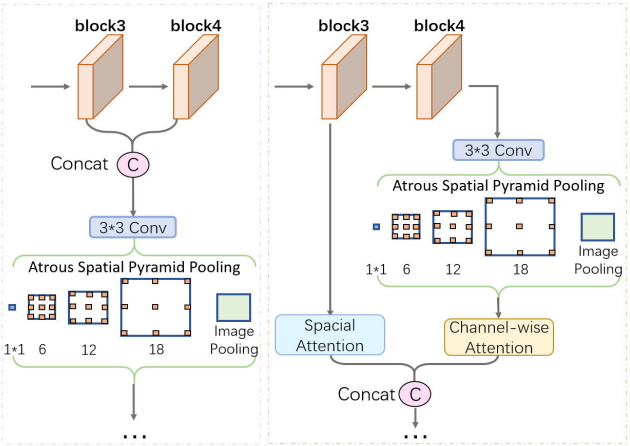


Figure 1: Some potential pyramidal features extraction architectures.

Methods	SAD↓	MSE↓	Grad↓	Conn↓
Input Image	115.23	0.0157	51.26	106.06
Strategy-1	77.29	0.0098	44.93	78.63
Strategy-2	64.37	0.0072	32.03	56.99
HAttMatting(Ours)	**44.01**	**0.0067**	**29.26**	**46.41**

Table 1: The comparisons with different features extraction strategies. The "Input Image" means we extract appearance cues from input image, and "Strategy-1" and "Strategy-2" refer to the alternative pyramidal features capture strategies on the left and right of Fig. 1 respectively.

3. Additional Alpha Mattes On Datasets

From Fig. 2 to Fig. 4, we illustrate more alpha mattes of the proposed *HAttMatting* on the Composition-1k dataset, and the relevant comparative methods are the same with our

[1]Joint first authors. [†]Joint corresponding authors, and they led this project. Project page: https://wukaoliu.github.io/HAttMatting/.

paper. Fig.5 and Fig. 6 shows the additional results on our Distinctions-646 dataset.

4. Additional Results On Real-world Images

Fig. 7 exhibit the real-world results of our *HAttMatting*. The model is trained on the Composition-1k dataset. The real-world images contain hair or fur, texture details or semi-transparent regions etc., and we can generate high-quality alpha mattes on them with the *HAttMatting*, which indicates the versatility of our method.

References

[1] L. C. Chen, G Papandreou, I Kokkinos, K Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE TPAMI*, 2018. 1

[2] Donghyeon Cho, Yu Wing Tai, and Inso Kweon. Natural image matting using deep convolutional neural networks. In *ECCV*, 2016. 3, 4, 5

[3] Qiqi Hou and Feng Liu. Context-aware image matting for simultaneous foreground and alpha estimation. In *ICCV*, 2019. 3, 4, 5

[4] Anat Levin, Dani Lischinski, and Yair Weiss. A closed-form solution to natural image matting. *IEEE TPAMI*, 2007. 3, 4, 5

[5] Hao Lu, Yutong Dai, Chunhua Shen, and Songcen Xu. Indices matter: Learning to index for deep image matting. In *ICCV*, 2019. 3, 4, 5

[6] Saining Xie, Ross Girshick, Piotr Dollar, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *CVPR*, 2017. 1

[7] Ning Xu, Brian Price, Scott Cohen, and Thomas Huang. Deep image matting. In *CVPR*, 2017. 3, 4, 5, 6, 7

[8] Yunke Zhang, Lixue Gong, Lubin Fan, Peiran Ren, Qixing Huang, Hujun Bao, and Weiwei Xu. A late fusion cnn for digital matting. In *CVPR*, 2019. 3, 4, 5

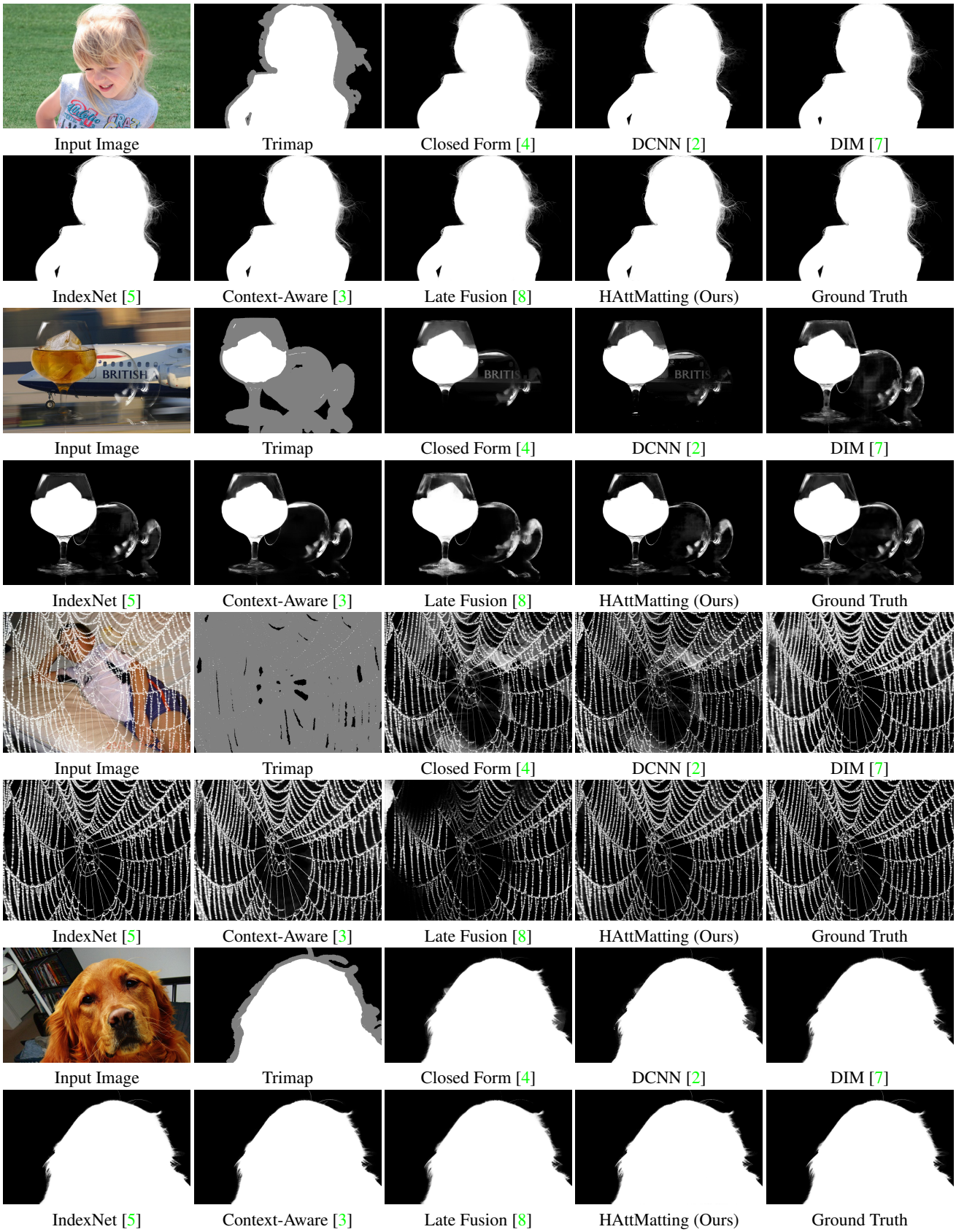


Figure 2: The visual comparisons on the Composition-1k test set.

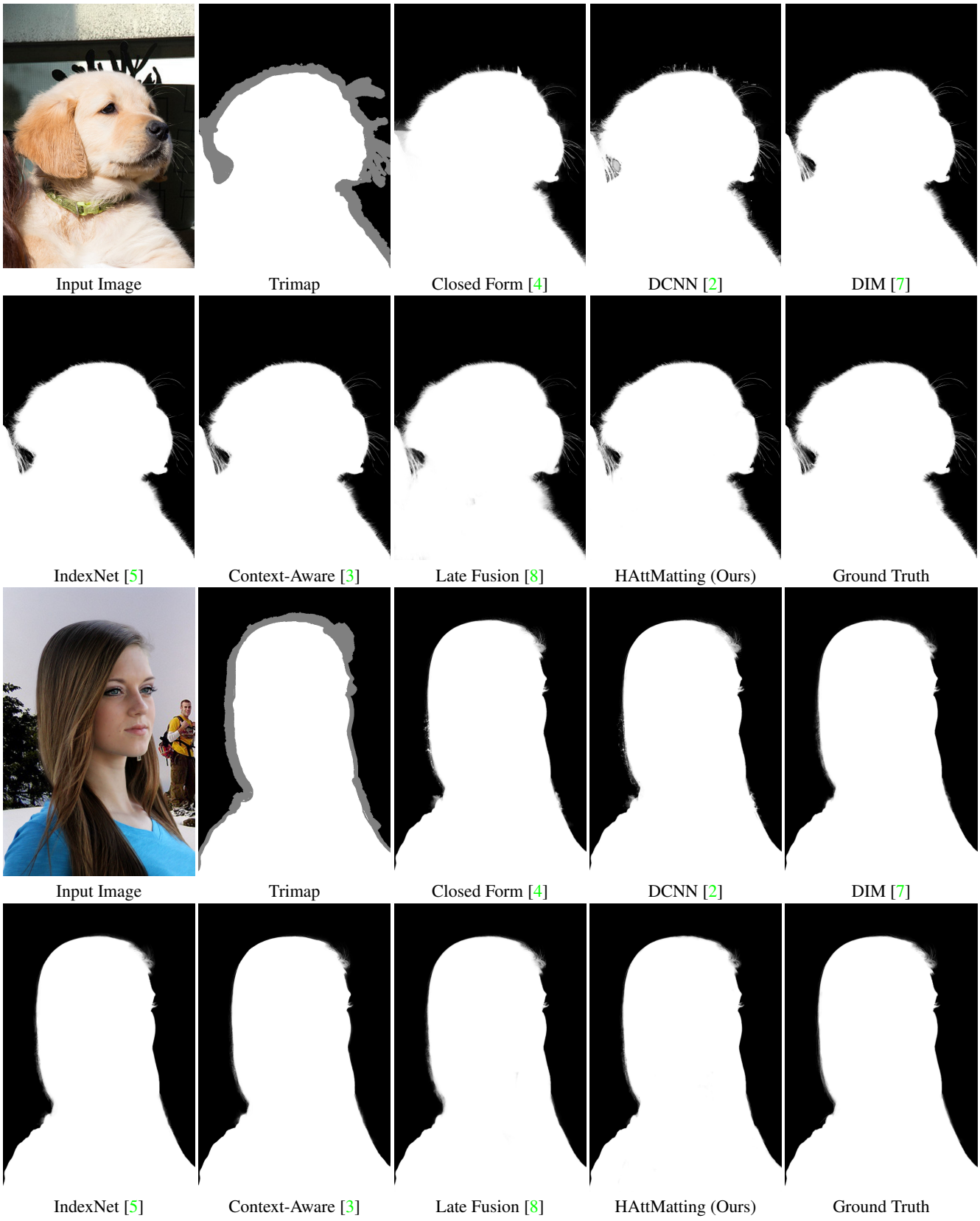


Figure 3: The visual comparisons on the Composition-1k test set.

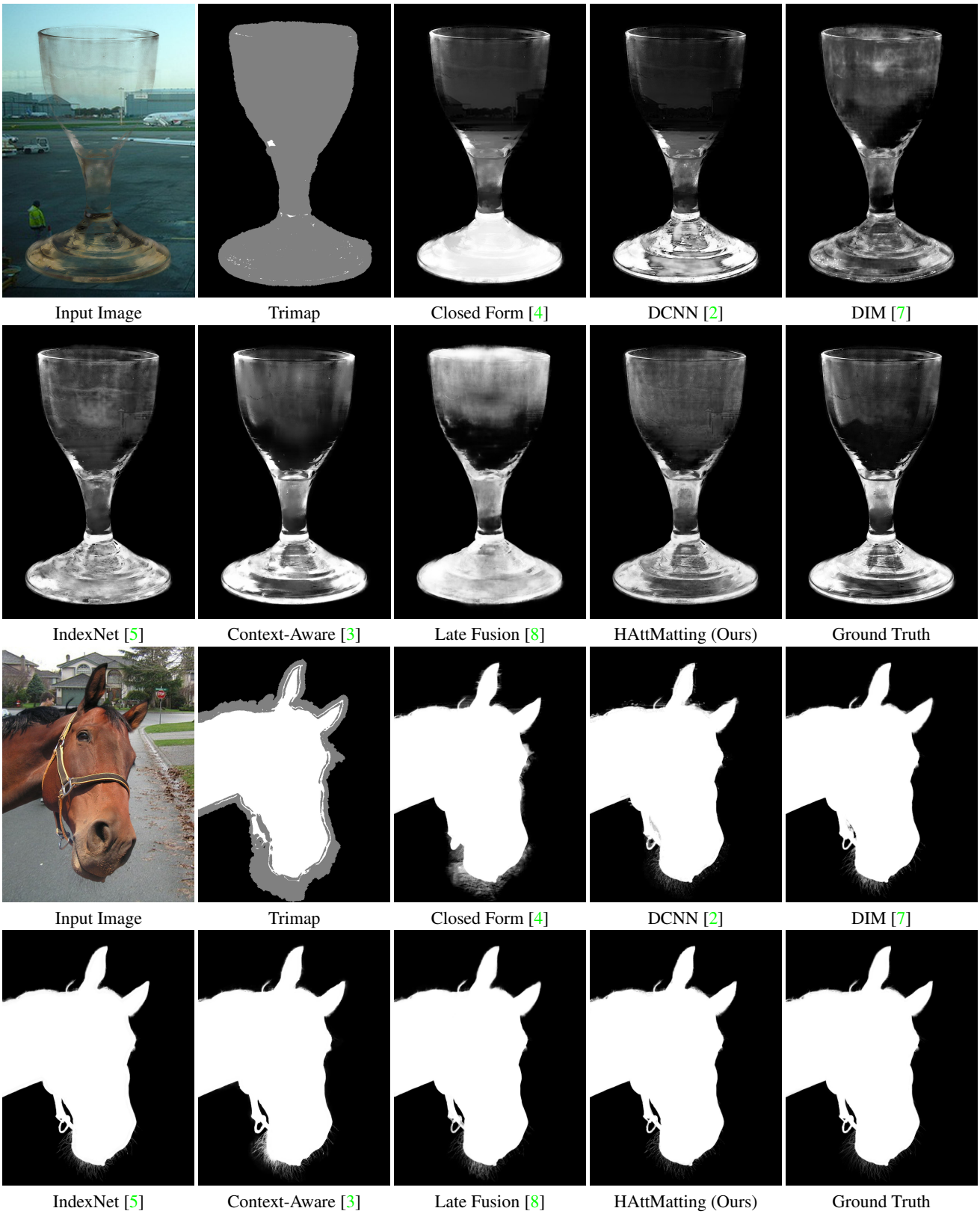
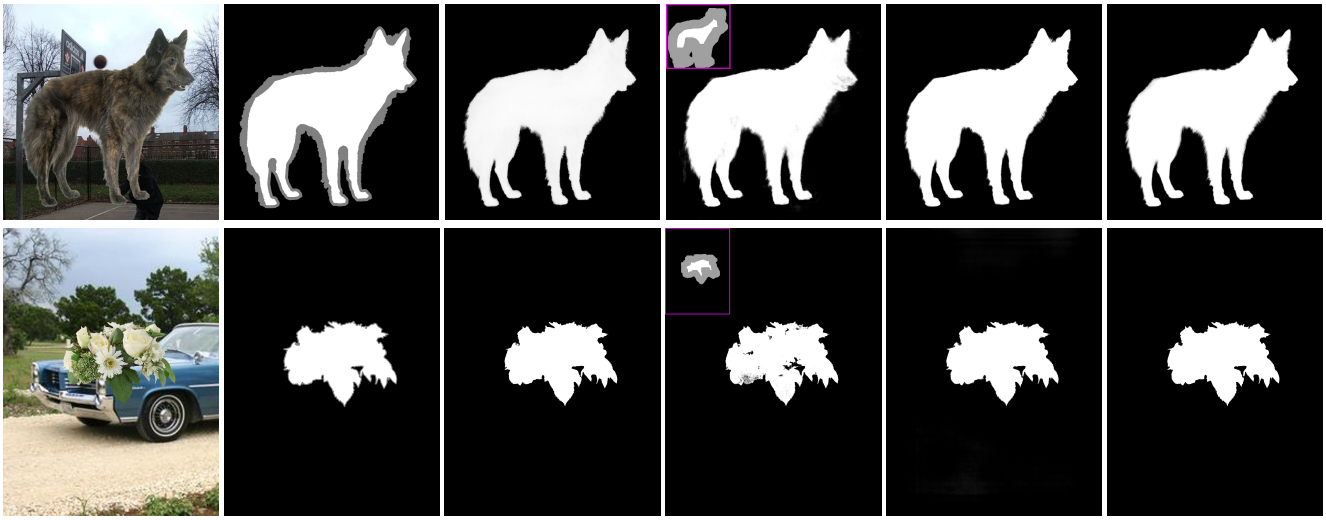


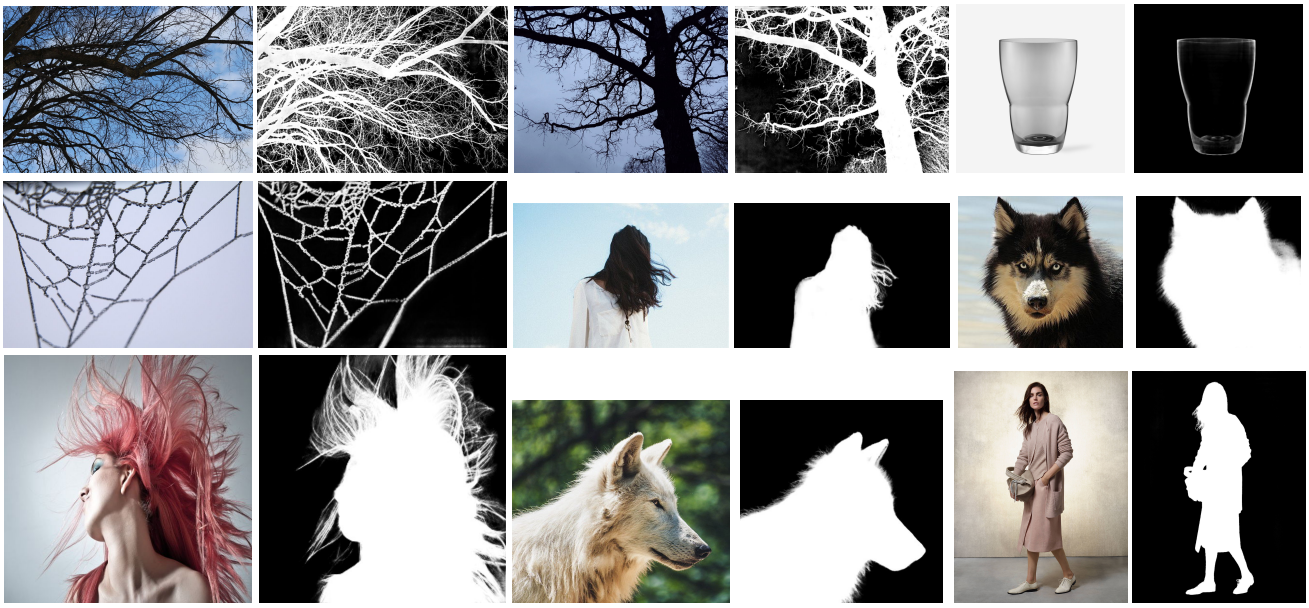
Figure 4: The visual comparisons on the Composition-1k test set.

Figure 5: The visual comparisons on our Distinctions-646 test set. The "DIM+Large" means that we feed DIM with trimaps that have larger transition region, while our method can generate high-quality alpha mattes without trimaps.



Input Image | Trimap | DIM [7] | DIM+Large | HAttMatting (Ours) | Ground Truth

Figure 6: The visual comparisons on our Distinctions-646 test set. The "DIM+Large" means that we feed DIM with trimaps that have larger transition region, while our method can generate high-quality alpha mattes without trimaps.



Input Image | Alpha matte | Input Image | Alpha Matte | Input Image | Alpha Matte

Figure 7: The alpha mattes produced by *HAttMatting* on real world images.