# Supplementary Material
# Human De-occlusion: Invisible Perception and Recovery for Humans

Qiang Zhou[1]*, Shiyin Wang[2], Yitong Wang[2], Zilong Huang[1], Xinggang Wang[1]†
[1]School of EIC, Huazhong University of Science and Technology [2]ByteDance Inc.
theodoruszq@gmail.com shiyinwang.ai@bytedance.com wangyitong@pku.edu.cn
zilong.huang2020@gmail.com xgwang@hust.edu.cn

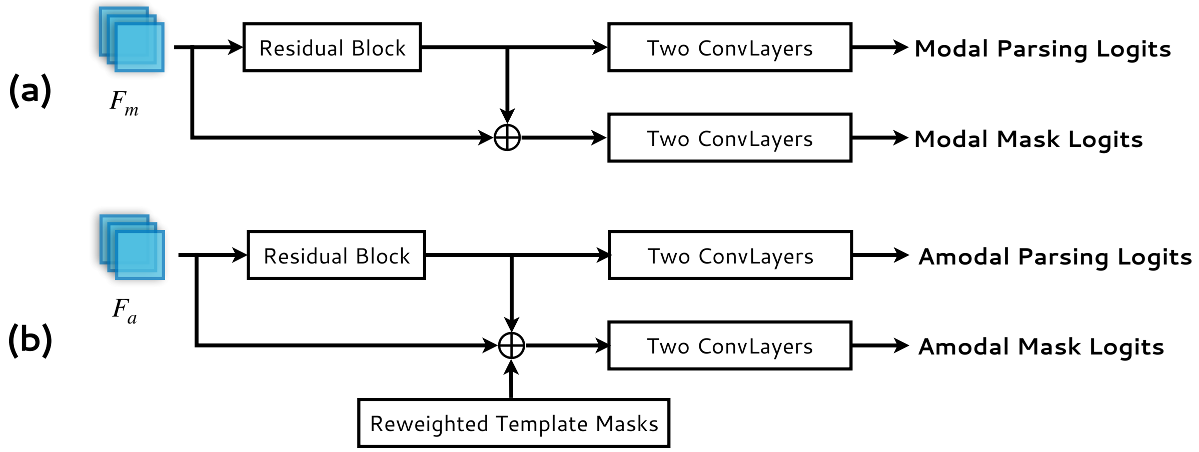

Figure 1. The detailed structures of the modal and amodal branches in the refined mask completion network. $F_m$ and $F_a$ denotes the output feature from the first hourglass module $N_m^{hg}$ and $N_a^{hg}$ respectively.

## 1. Network Details

We propose a two-stage framework to tackle human de-occlusion,which contains mask completion and content recovery. Here we provide more details of the two networks.

**Refined Mask Completion Network** We demonstrate the structures of the modal and amodal branches in Fig 1. The two branches share a similar pipeline. Take the modal branch for example, it uses a single residual block and two convolution layers with a ReLU in the middle to extract model parsing logits. To generate the modal mask logits, we concatenate the feature of $F_m$ and the auxiliary feature from the residual block [1] as the input of another two convolution layers to get the final modal mask logits. To obtain the amodal mask logits, we additionally concatenate the reweighted template masks and adjust the channel number of the subsequent layer. We also measure the inference time of our network and the two branches cost extra 2.5ms compared to our baseline. (baseline 13.37ms *vs.* ours 15.83ms)

**Parsing Guided Content Recovery Network** In the first attention stream of body parts, we distribute input feature to different body parts and extract useful part information. For the spatial attention, we add the one-hot parsing logits before obtaining key and query features, and a mutual spatial attention is applied to extract the relationship between the visible and the invisible regions. Because the differences of the key and the query mainly lie in the invisible points so that the network is able to find and recover them. The inference time increases by 5.82ms compared to the baseline. (baseline 8.02ms *vs.* ours 13.84ms) The extra time cost mainly comes from the multiplication operation to compute the matrix $R$. And the total inference time costs about 21.65ms ($\sim$ 46 FPS) so that speed is not a problem in real applications.

## 2. More Results

We show more comparison results of mask completion and content recovery in Fig 2 and Fig 3 respectively. And some results on real occluded humans are shown in Fig 4. Black borders are padded to fit the input size to $256 \times 256$.

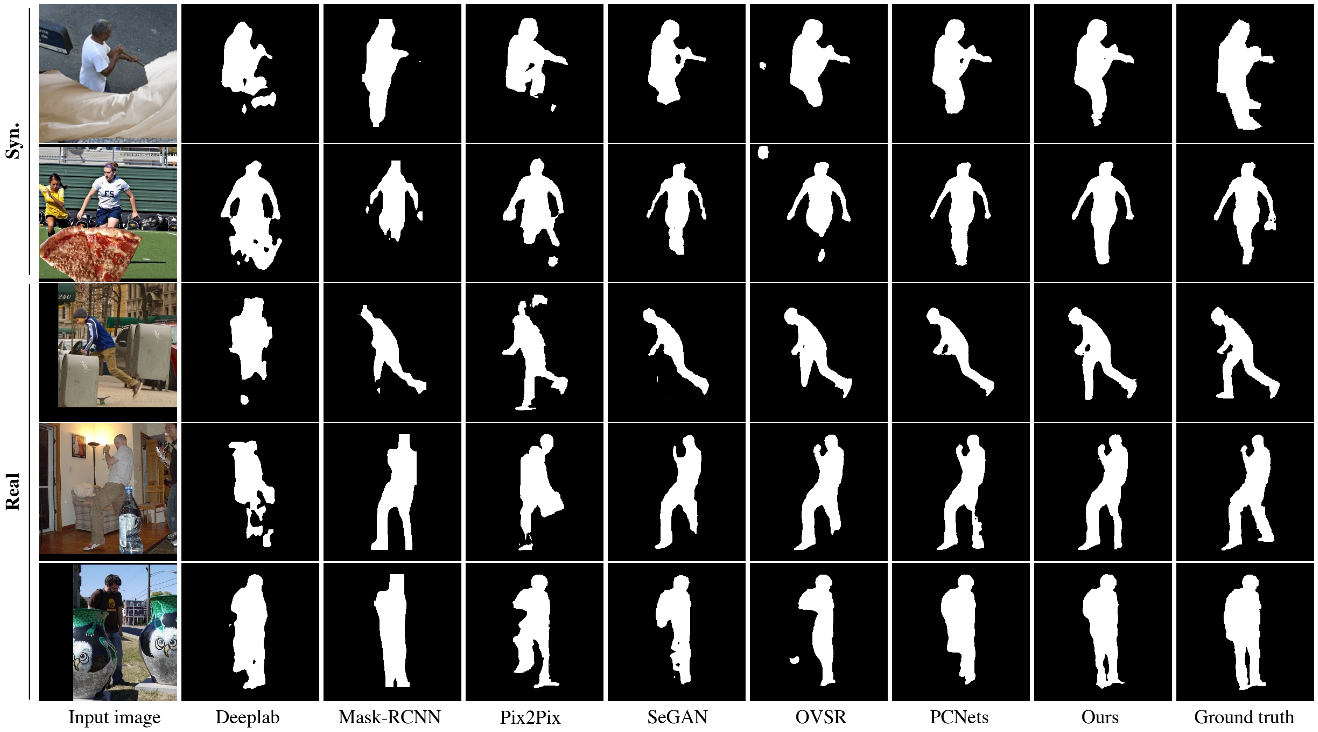

Figure 2. Some comparison results of mask completion on synthesized and real images.

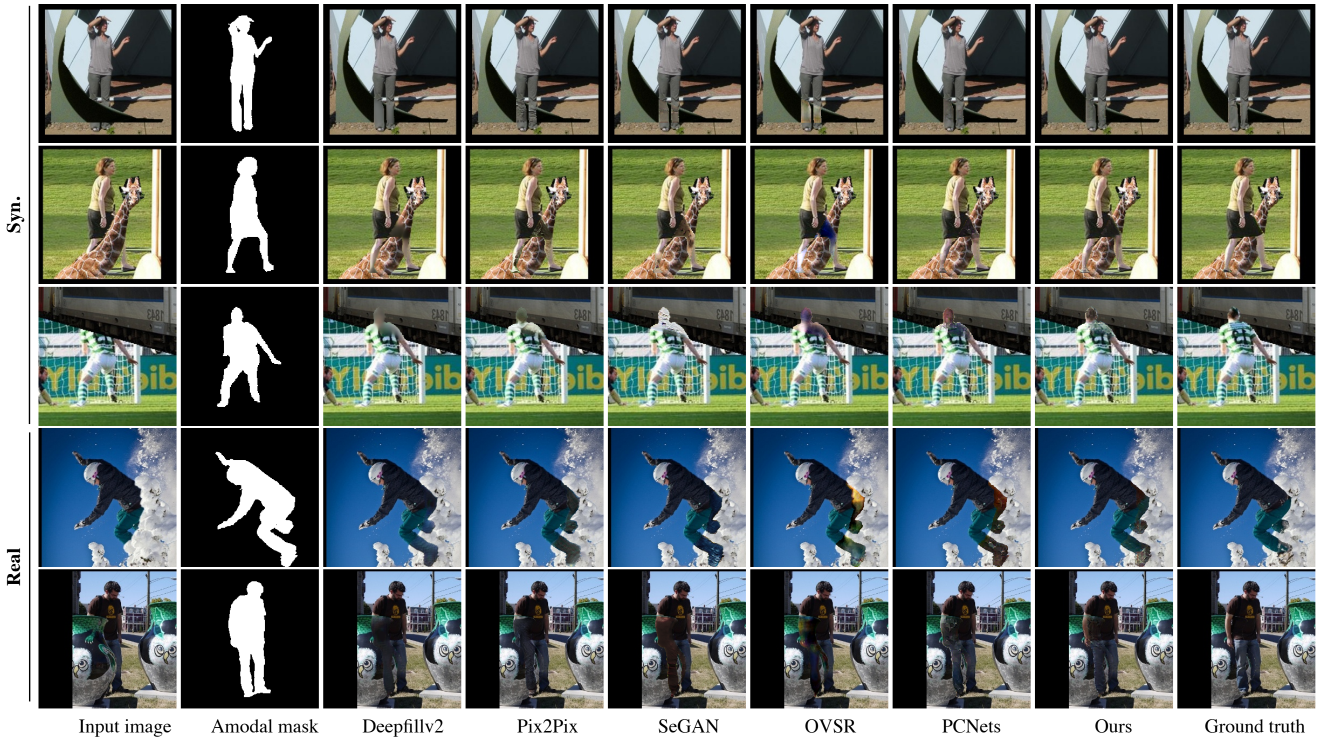

Figure 3. Some comparison results of content recovery on synthesized and real images.

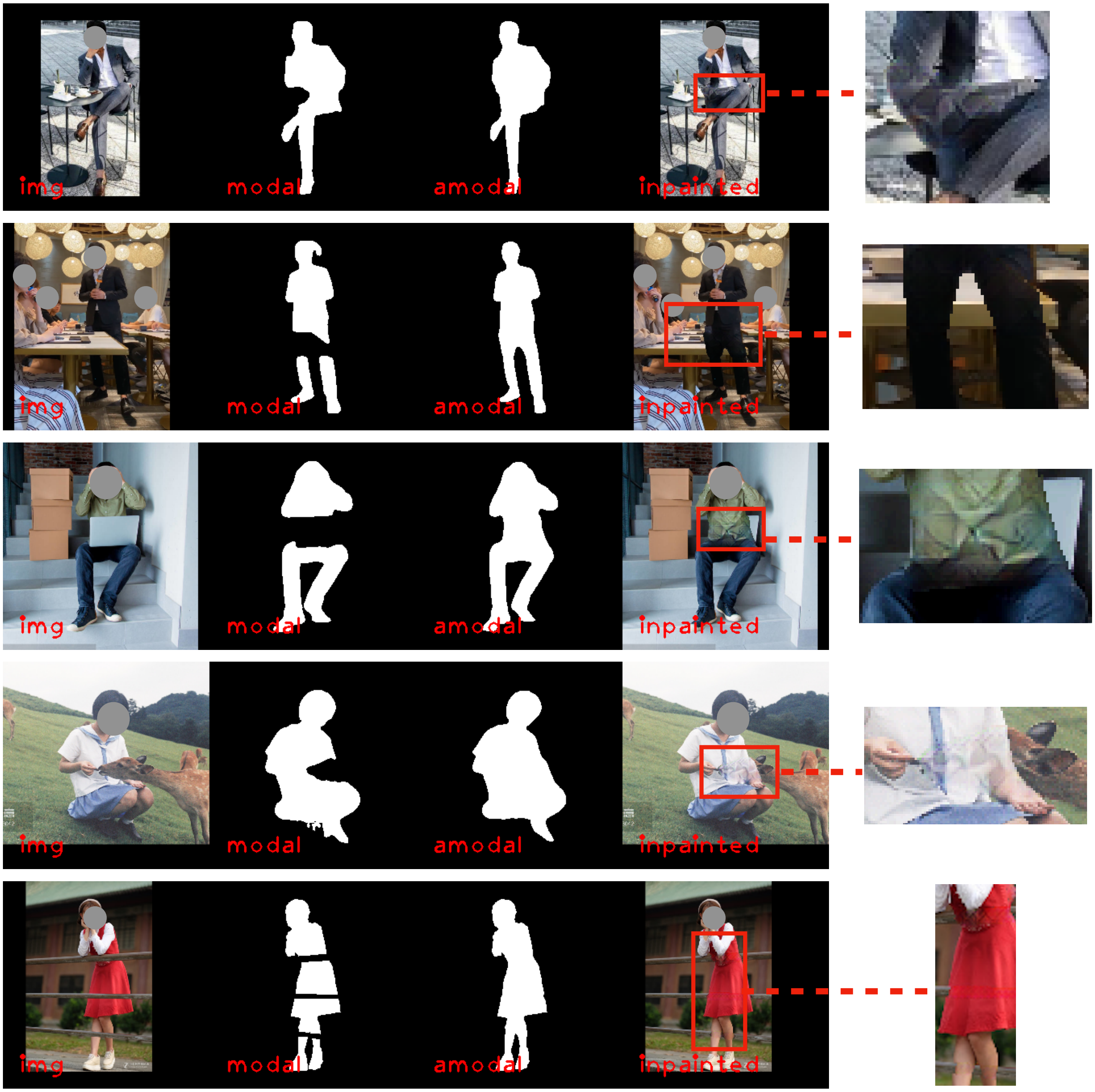

Figure 4. Some results in real human occlusion cases from the Internet. We mosaic the faces due to privacy issues.

## References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 1