6 Supplementary material

6.1 Architecture of the proposed generator

Encoder			Decoder		
layer name	output size	filter size	layer name	output	filter size
conv1	160×160	$7 \times 7, 64$, stride 2	bilinear	80×80	bilinear upsampling
conv2_x	80×80	3×3 max pool, stride 2 $\begin{bmatrix} 1 \times 1, & 64 \\ 3 \times 3, & 64 \\ 1 \times 1, & 256 \end{bmatrix} \times 3$	deconv1_x	80×80	skip from conv2_x, $1 \times 1, 48$ $\begin{bmatrix} 3 \times 3, & 256 \\ 3 \times 3, & 128 \\ 3 \times 3, & 64 \end{bmatrix}$
conv3_x	40×40	$\begin{bmatrix} 1 \times 1, & 128 \\ 3 \times 3, & 128 \\ 1 \times 1, & 512 \end{bmatrix} \times 4$	unpooling	160×160	2×2 unpool, stride 2
conv4_x	40×40	$\begin{bmatrix} 1 \times 1, & 256 \\ 3 \times 3, & 256 \quad r = 2 \\ 1 \times 1, & 1024 \end{bmatrix} \times 6$	deconv2_x	320×320	skip from conv1_x, $1 \times 1, 32$ $\begin{bmatrix} 3 \times 3, & 64 \\ 3 \times 3, & 64, \quad \text{stride} \frac{1}{2} \\ 3 \times 3, & 32 \end{bmatrix}$
conv5_x	40×40	$\begin{bmatrix} 1 \times 1, & 512 \\ 3 \times 3, & 512 \quad r = 4 \\ 1 \times 1, & 2048 \end{bmatrix} \times 3$	deconv3_x	320×320	skip from RGB image $\begin{bmatrix} 3 \times 3, & 32 \\ 3 \times 3, & 32 \end{bmatrix}$
aspp	40×40	$\begin{bmatrix} 1 \times 1, & 256 \\ 3 \times 3, r = 6, & 256 \\ 3 \times 3, r = 12, & 256 \\ 3 \times 3, r = 18, & 256 \\ \text{Image Pooling}, & 256 \end{bmatrix}$	deconv4_x	320×320	$3 \times 3, \quad 1$

Table 4: Architecture of the proposed generator. The encoder consists of the standard Resnet50 architecture with the last two layers removed and ASPP [3] module added to output 256 40×40 feature maps. The decoder is kept small and uses bilinear interpolation, unpooling and fractionally-strided convolution to upsample the feature maps back to 320×320. For the max-pooling operation in the encoder, the maximum indices are saved and used in the unpooling layer. All convolutional layers except the last one are followed by batch-normalization layers [17] and ReLU activation functions. The last convolutional layer is followed by a sigmoid activation function to scale the output between 0 and 1. r is the dilation rate of the convolution. The default stride or dilation rate is 1. Skip connections are added to retain localized information.

6.2 Examples from the Composition-1k dataset

Figure 4: Examples of non-realistic images introduced in the Composition-1k test dataset.

6.3 Additional comparison results on the Composition-1k test dataset

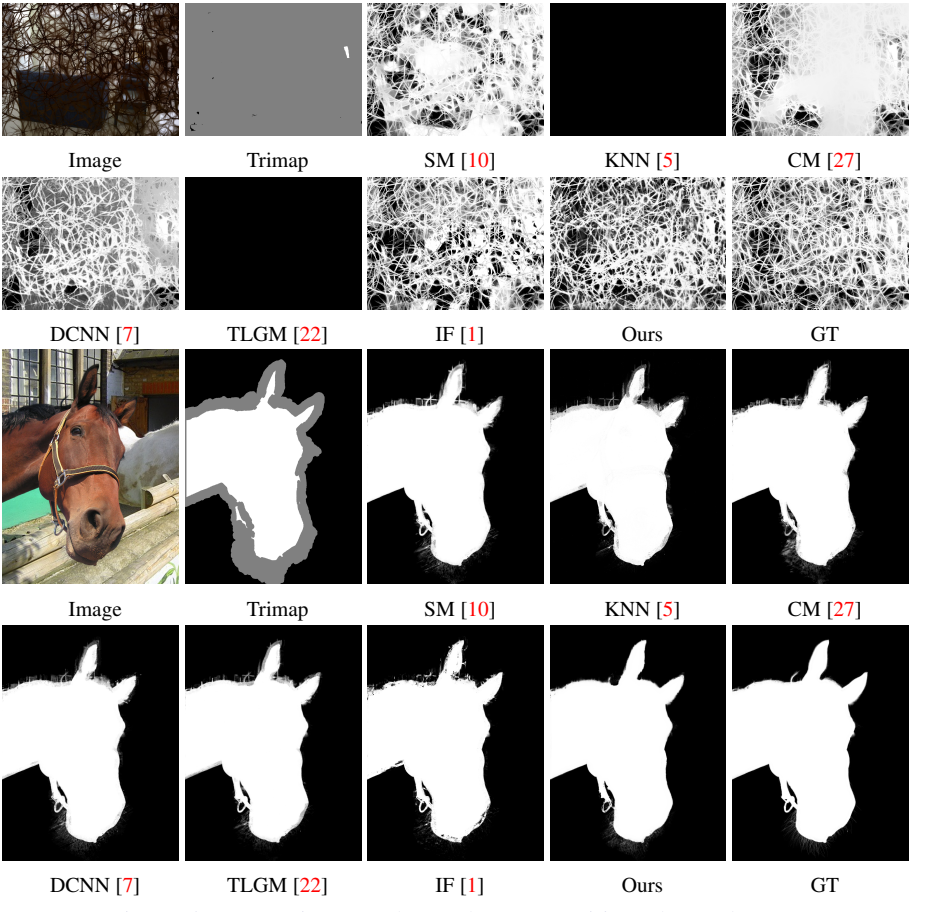

Figure 5: Comparison results on the Composition-1k test dataset.

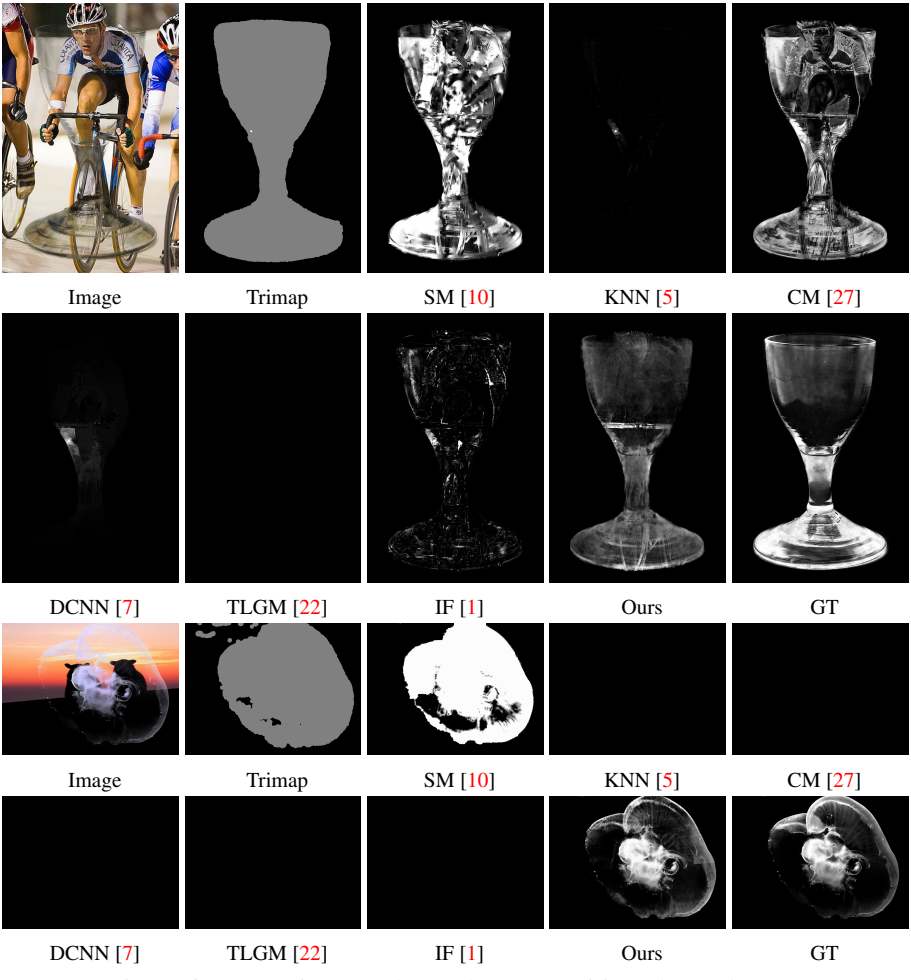

Figure 6: Comparison results on the Composition-1k test dataset.