Supplementary Materials: Generative Image Inpainting with Contextual Attention

Jiahui Yu[1] Zhe Lin[2] Jimei Yang[2] Xiaohui Shen[2] Xin Lu[2] Thomas S. Huang[1]

[1]University of Illinois at Urbana-Champaign
[2]Adobe Research

A. More Results on CelebA, CelebA-HQ, DTD and ImageNet

CelebA-HQ [?] We show results from our full model trained on CelebA-HQ dataset in Figure 1. Note that the original image resolution of CelebA-HQ dataset is 1024×1024. We resize image to 256×256 for both training and evaluation.

CelebA [?] We show more results from our full model trained on CelebA dataset in Figure 2. Note that the original image resolution of CelebA dataset is 218×178. We resize image to 315×256 and do a random crop of size 256×256 to make face landmarks roughly unaligned for both training and evaluation.

ImageNet [?] We show more results from our full model trained on ImageNet dataset in Figure 3.

DTD textures [?] We show more results from our full model trained on DTD dataset in Figure 4.

B. Comparisons with More Methods

We show more results for qualitative comparisons with more methods including Photoshop Content-Aware Fill [**?**], Image Melding [**?**] and StructCompletion [**?**] in Figure 5 and 6. For all these methods, we use default hyper-parameter settings.

C. More Visualization with Case Study

In addition to attention map visualization, we visualize which parts in the input image are being attended for pixels in holes. To do so, we highlight the regions that have the maximum attention score and overlay them to input image. As shown in Figure 7, the visualization results given holes in different locations demonstrate the effectiveness of our proposed contextual attention to borrow information at distant spatial locations.

D. Network Architectures

In addition to Section 3, we report more details of our network architectures. For simplicity, we denote them with K (kernel size), D (dilation), S (stride size) and C (channel number).

Inpainting network Inpainting network has two encoder-decoder architecture stacked together, with each encoder-decoder of network architecture:

K5S1C32 - K3S2C64 - K3S1C64 - K3S2C128 - K3S1C128 - K3S1C128 - K3D2S1C128 - K3D4S1C128 - K3D8S1C128 - K3D16S1C128 - K3S1C128 - K3S1C128 - resize ($2\times$) - K3S1C64 - K3S1C64 - resize ($2\times$) - K3S1C32 - K3S1C16 - K3S1C3 - clip.

Local WGAN-GP critic We use Leaky ReLU with $\alpha = 0.2$ as activation function for WGAN-GP critics.

K5S2C64 - K5S2C128 - K5S2C256 - K5S2C512 - fully-connected to 1.

Global WGAN-GP critic K5S2C64 - K5S2C128 - K5S2C256 - K5S2C256 - fully-connected to 1.

Contextual attention branch K5S1C32 - K3S2C64 - K3S1C64 - K3S2C128 - K3S1C128 - K3S1C128 - contextual attention layer - K3S1C128 - K3S1C128 - concat.

Figure 1: More inpainting results of our full model with contextual attention on CelebA-HQ faces. Each triad, from left to right, shows original image, input masked image and result image. All input images are masked from validation set (training and validation split is provided in released code). All results are direct outputs from same trained model without post-processing.

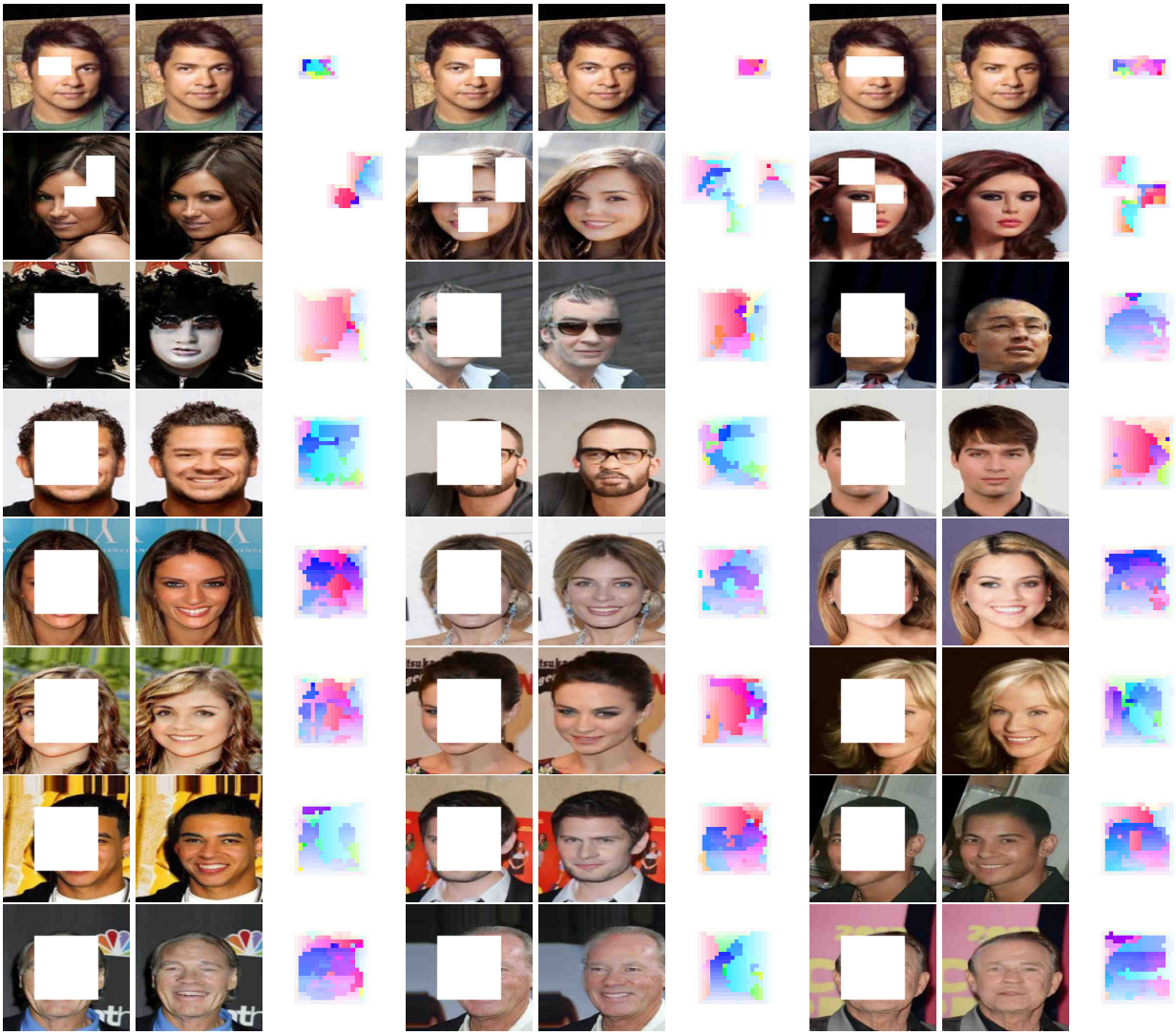

Figure 2: More inpainting results of our full model with contextual attention on CelebA faces. Each triad, from left to right, shows input image, result and attention map (upscaled $4\times$). All input images are masked from validation set (face identities are NOT overlapped between training set and validation set). All results are direct outputs from same trained model without post-processing.

Figure 3: More inpainting results of our full model with contextual attention on ImageNet. Each triad, from left to right, shows input image, result and attention map (upscaled $4\times$). All input images are masked from validation set. All results are direct outputs from same trained model without post-processing.

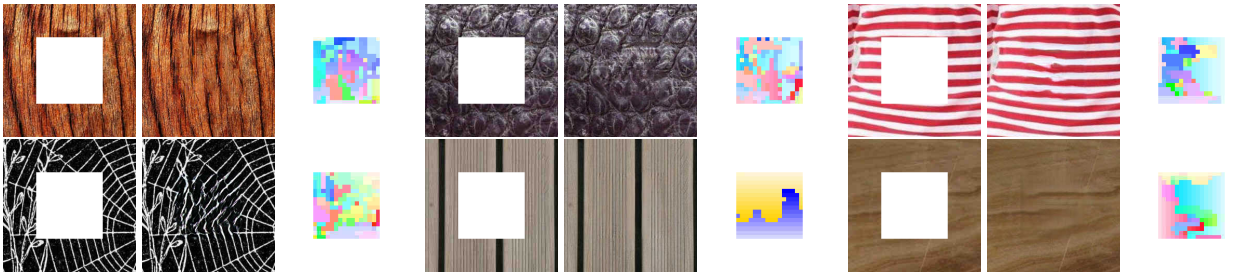

Figure 4: More inpainting results of our full model with contextual attention on DTD textures. Each triad, from left to right, shows input image, result and attention map (upscaled $4\times$). All input images are masked from validation set. All results are direct outputs from same trained model without post-processing.

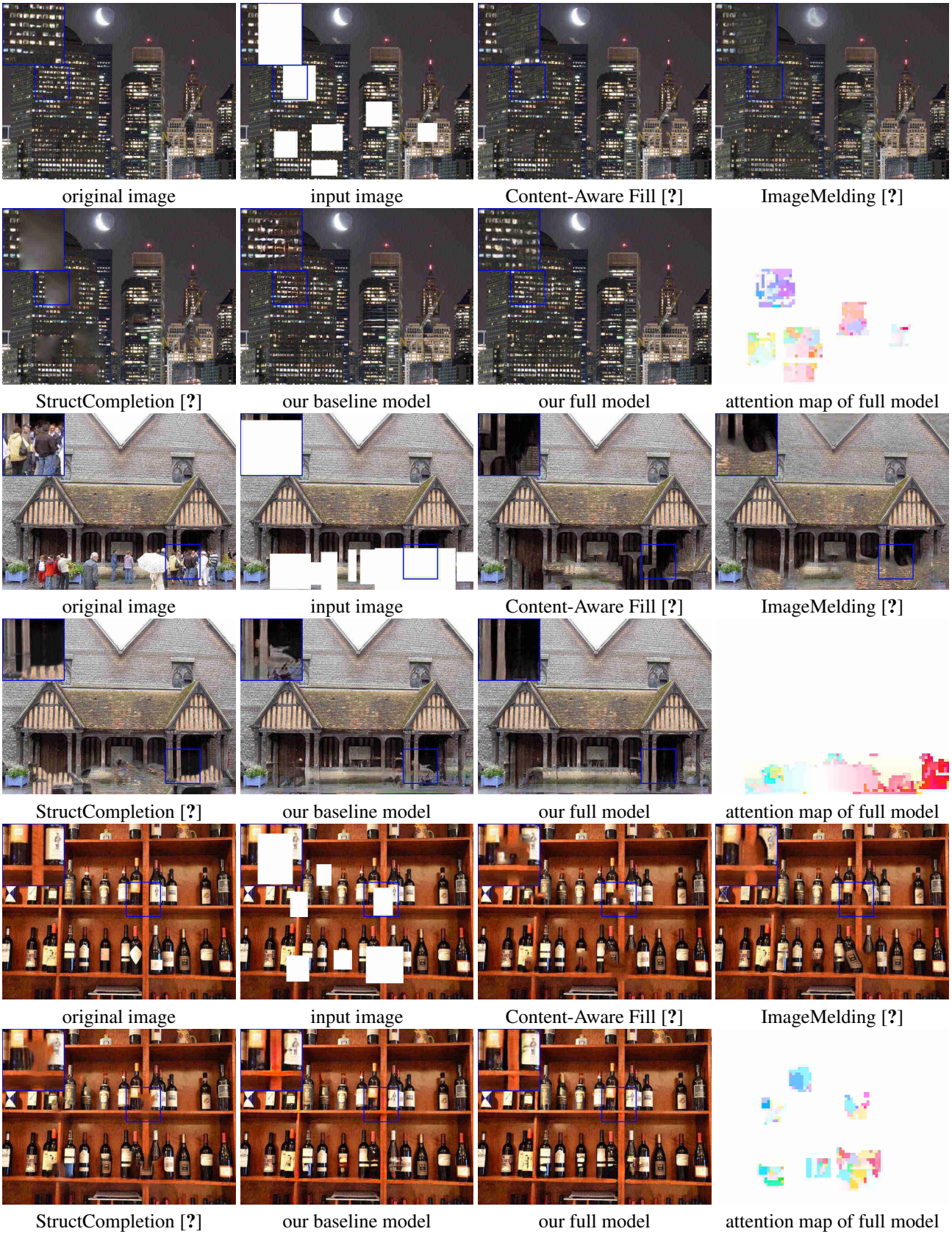

Figure 5: More qualitative results and comparisons. All input images are masked from validation set. All our results are direct outputs from same trained model without post-processing. Best viewed with zoom-in.

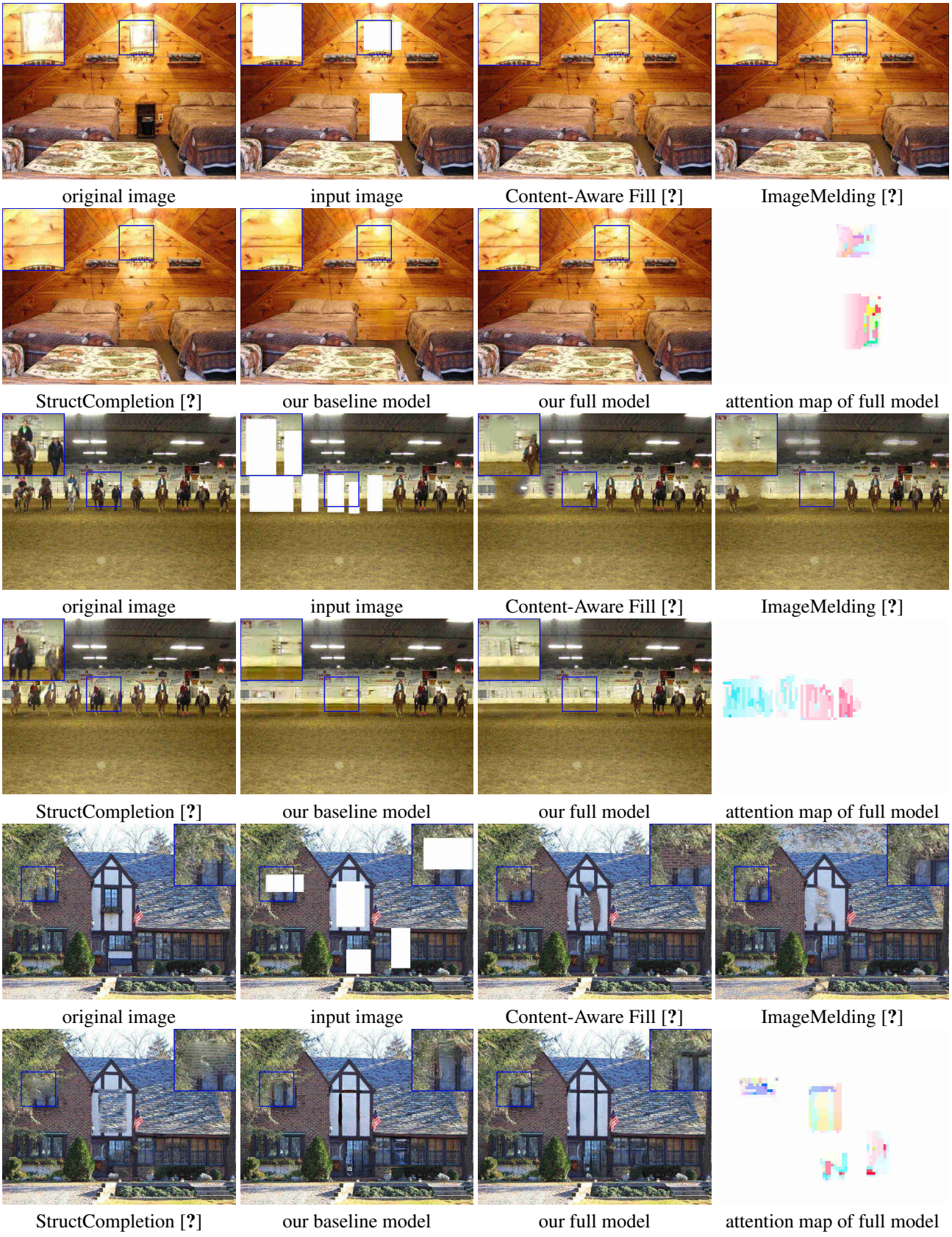

Figure 6: More qualitative results and comparisons. All input images are masked from validation set. All our results are direct outputs from same trained model without post-processing. Best viewed with zoom-in.

Figure 7: Visualization (highlighted regions) on which parts in input image are attended. Each triad, from left to right, shows input image, result and attention visualization.