Supplementary Material for Image Completion with Adaptive Multi-Temperature Mask-Guided Attention

Xiang Zhou[1]
zhoux2020@mail.sustech.edu.cn
Yuan Zeng[†2]
zengy3@sustech.edu.cn
Yi Gong[†3]
gongy@sustech.edu.cn

[1] Southern University of Science and Technology(SUSTech), China
[2] Academy for Advanced Interdisciplinary Studies, SUSTech, China
[3] University Key Laboratory of Guangdong Province, SUSTech, China
[†] Corresponding authors

This supplementary material is organized as follows: In Section 1, we present the architectural details of our completion network. In Section 2, we present additional qualitative comparison results on CelebA-HQ [3] and Paris StreetView [2], as well as quantitative comparison results on Paris StreetView.

1 Network Architecture

This section provides architectural details of our two-stage image completion network. It consists of three components: Coarse Network, Refinement Network and Discriminator. Their architectures are shown in Table 1, 2 and 3, respectively.

For column MODULE, "GConv" means gated convolution, which contains two convolutions for calculating intermediate output and soft mask, and they share the same setting. "GDeconv" consists of a $2\times$ nearest neighbor upsampling followed with a gated convolution. "Conv" indicates a convolutional layer and "FC" means a fully connected layer. All gated convolution are followed by batch normalization in our experiment, but this is not the case for convolutional layer and full connected layer. The proposed Adaptive multi-Temperature Mask-guided Attention module is shown in Table 2 with "ATMA".

For the other columns, KERNEL and DILATION lists specified kernel size and stride used in convolution, respectively. DILATION indicates dilation rate of convolution. NONLINEARITY represents type of non-linear activation function.

2 Additional Experimental Results

In Figures 1 and 2, we illustrate more examples on CelebA-HQ and Paris StreetView dataset. 256×256 images degraded by 128×128 squared central masks are used as input. Table 4 shows our quantitative comparison results in terms of mean l_1 loss, mean l_2 loss, peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM) on the validation set

Table 1: Architecture of the Coarse Network (Stage 1)

Module	Kernel	Stride	Dilation	Nonlinearity	Output Shape
Concat	—	—	—	—	$256 \times 256 \times 4$
GConv1	5	1	1	LeakyReLU(0.2)	$256 \times 256 \times 32$
GConv2	3	2	1	LeakyReLU(0.2)	$128 \times 128 \times 64$
GConv3	3	1	1	LeakyReLU(0.2)	$128 \times 128 \times 64$
GConv4	3	2	1	LeakyReLU(0.2)	$64 \times 64 \times 128$
GConv5	3	1	1	LeakyReLU(0.2)	$64 \times 64 \times 128$
GConv6	3	1	1	LeakyReLU(0.2)	$64 \times 64 \times 128$
GConv7	3	1	2	LeakyReLU(0.2)	$64 \times 64 \times 128$
GConv8	3	1	4	LeakyReLU(0.2)	$64 \times 64 \times 128$
GConv9	3	1	8	LeakyReLU(0.2)	$64 \times 64 \times 128$
GConv10	3	1	16	LeakyReLU(0.2)	$64 \times 64 \times 128$
GConv11	3	1	1	LeakyReLU(0.2)	$64 \times 64 \times 128$
GConv12	3	1	1	LeakyReLU(0.2)	$64 \times 64 \times 128$
GDeconv13	3	1	1	LeakyReLU(0.2)	$128 \times 128 \times 64$
GConv14	3	1	1	LeakyReLU(0.2)	$128 \times 128 \times 64$
GDeconv15	3	1	1	LeakyReLU(0.2)	$256 \times 256 \times 32$
GConv16	3	1	1	LeakyReLU(0.2)	$256 \times 256 \times 16$
Conv17	3	1	1	Tanh	$256 \times 256 \times 3$

Table 2: Architecture of the Refinement Network (Stage 2)

Network	Module	Kernel	Stride	Dilation	Nonlinearity	Output Shape
Conv Branch	Concat	—	—	—	—	$256 \times 256 \times 4$
	GConv1	5	1	1	LeakyReLU(0.2)	$256 \times 256 \times 32$
	GConv2	3	2	1	LeakyReLU(0.2)	$128 \times 128 \times 64$
	GConv3	3	1	1	LeakyReLU(0.2)	$128 \times 128 \times 64$
	GConv4	3	2	1	LeakyReLU(0.2)	$64 \times 64 \times 128$
	GConv5	3	1	1	LeakyReLU(0.2)	$64 \times 64 \times 128$
	GConv6	3	1	1	LeakyReLU(0.2)	$64 \times 64 \times 128$
	GConv7	3	1	2	LeakyReLU(0.2)	$64 \times 64 \times 128$
	GConv8	3	1	4	LeakyReLU(0.2)	$64 \times 64 \times 128$
	GConv9	3	1	8	LeakyReLU(0.2)	$64 \times 64 \times 128$
	GConv10	3	1	16	LeakyReLU(0.2)	$64 \times 64 \times 128$
Attention Branch	Concat	—	—	—	—	$256 \times 256 \times 4$
	GConv1	5	1	1	LeakyReLU(0.2)	$256 \times 256 \times 32$
	GConv2	3	2	1	LeakyReLU(0.2)	$128 \times 128 \times 64$
	GConv3	3	1	1	LeakyReLU(0.2)	$128 \times 128 \times 64$
	GConv4	3	2	1	LeakyReLU(0.2)	$64 \times 64 \times 128$
	GConv5	3	1	1	LeakyReLU(0.2)	$64 \times 64 \times 128$
	GConv6	3	1	1	LeakyReLU(0.2)	$64 \times 64 \times 128$
	ATMA	—	—	—	—	$64 \times 64 \times 128$
	GConv7	3	1	1	LeakyReLU(0.2)	$64 \times 64 \times 128$
	GConv8	3	1	1	LeakyReLU(0.2)	$64 \times 64 \times 128$
Decoder	Concat	—	—	—	—	$64 \times 64 \times 256$
	GConv11	3	1	1	LeakyReLU(0.2)	$64 \times 64 \times 128$
	GConv12	3	1	1	LeakyReLU(0.2)	$64 \times 64 \times 128$
	GDeconv13	3	1	1	LeakyReLU(0.2)	$128 \times 128 \times 64$
	GConv14	3	1	1	LeakyReLU(0.2)	$128 \times 128 \times 64$
	GDeconv15	3	1	1	LeakyReLU(0.2)	$256 \times 256 \times 32$
	GConv16	3	1	1	LeakyReLU(0.2)	$256 \times 256 \times 16$
	Conv17	3	1	1	Tanh	$256 \times 256 \times 3$

Table 3: Architecture of the Globally-and-Locally Consistent Discriminator

Network	Module	Kernel	Stride	Dilation	Nonlinearity	Output Shape
Global	Conv1	5	2	1	LeakyReLU(0.01)	$128 \times 128 \times 64$
	Conv2	5	2	1	LeakyReLU(0.01)	$64 \times 64 \times 128$
	Conv3	5	2	1	LeakyReLU(0.01)	$32 \times 32 \times 256$
	Conv4	5	2	1	LeakyReLU(0.01)	$16 \times 16 \times 512$
	Conv5	5	2	1	LeakyReLU(0.01)	$8 \times 8 \times 512$
	Conv6	5	2	1	LeakyReLU(0.01)	$4 \times 4 \times 512$
	FC1	—	—	—	LeakyReLU(0.01)	1024
Local	Conv1	5	2	1	LeakyReLU(0.01)	$64 \times 64 \times 128$
	Conv2	5	2	1	LeakyReLU(0.01)	$32 \times 32 \times 256$
	Conv3	5	2	1	LeakyReLU(0.01)	$16 \times 16 \times 512$
	Conv4	5	2	1	LeakyReLU(0.01)	$8 \times 8 \times 512$
	Conv5	5	2	1	LeakyReLU(0.01)	$4 \times 4 \times 512$
	FC2	—	—	—	LeakyReLU(0.01)	1024
Linear	Concat	—	—	—	—	2048
	FC3	—	—	—	—	1

of Paris StreetView. It shows that our completion network outperforms the other approaches.

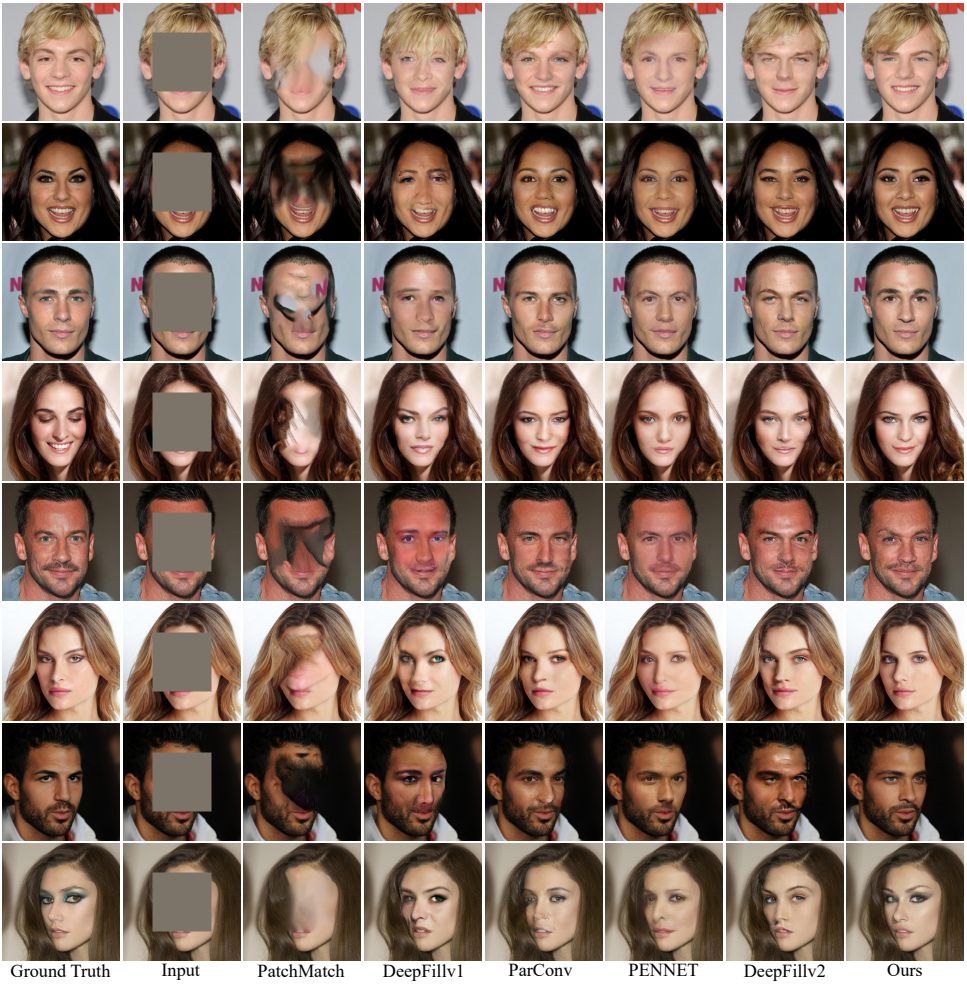

Figure 1: Qualitative comparisons on CelebA-HQ dataset. From left to right, ground truth, input image with hole, PatchMatch [1], DeepFillv1 [6], ParConv [4], PENNET [8], DeepFillv2 [7] and Ours.

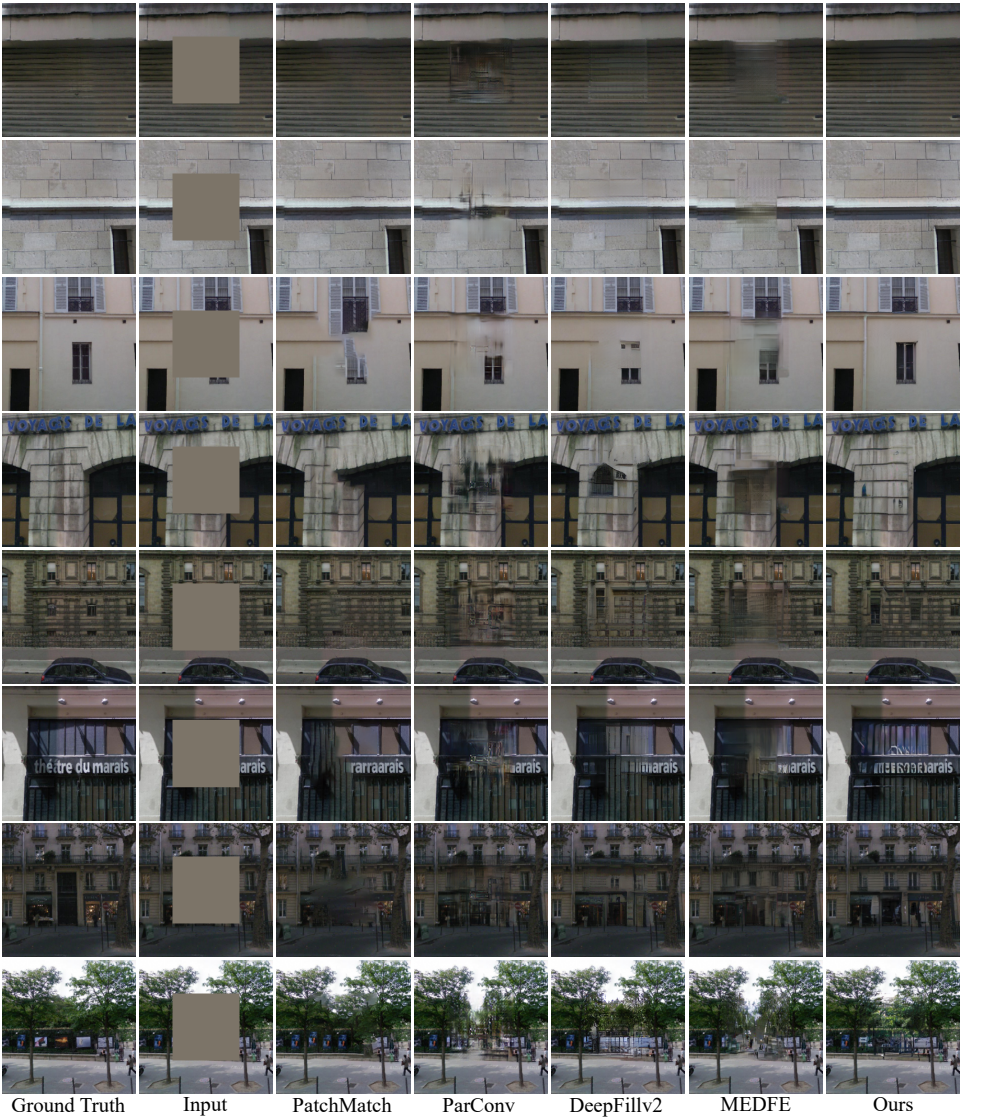

Figure 2: Qualitative comparisons on Paris StreetView dataset. From left to right, ground truth, input image with hole, PatchMatch [1], ParConv [4], DeepFillv2 [7], MEDFE [5] and Ours.

Table 4: Results of mean l_1, mean l_2, PSNR and SSIM on validation set on Paris StreetView.

	mean $l_1 \downarrow$	mean $l_2 \downarrow$	PSNR (dB) $\uparrow$	SSIM $\uparrow$
PatchMatch	2.53 %	0.62%	23.7	85.1 %
ParConv	2.63 %	0.59%	23.6	84.3 %
DeepFillv2	2.40 %	0.51%	24.3	85.3 %
MEDFE	2.58 %	0.56%	23.9	85.2 %
ours ($K = 2$)	**2.14** %	**0.49** %	**25.0**	**86.7** %

References

[1] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. Patchmatch: A randomized correspondence algorithm for structural image editing. *ACM Transactions on Graphics*, 28(3):1–11, July 2009.

[2] Carl Doersch, Saurabh Singh, Abhinav Gupta, Josef Sivic, and Alexei A. Efros. What makes paris look like paris? *ACM Transactions on Graphics (SIGGRAPH)*, 31(4):1–9, 2012.

[3] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[4] Guilin Liu, Fitsum A. Reda, Kevin J. Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In *Proceedings of the European Conference on Computer Vision*, September 2018.

[5] Hongyu Liu, Bin Jiang, Yibing Song, Wei Huang, and Chao Yang. Rethinking image inpainting via a mutual encoder-decoder with feature equalizations. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 725–741. Springer, 2020.

[6] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S. Huang. Generative image inpainting with contextual attention. In *Proceedings of the IEEE Computer Vision and Pattern Recognition*, pages 5505–5514. IEEE Computer Society, 2018.

[7] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S. Huang. Free-form image inpainting with gated convolution. In *Proceedings of the IEEE International Conference on Computer Vision*, October 2019.

[8] Yanhong Zeng, Jianlong Fu, Hongyang Chao, and Baining Guo. Learning pyramid-context encoder network for high-quality image inpainting. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1486–1494, 2019.