Supplementary Material for Parallel Multi-Resolution Fusion Network for Image Inpainting

Wentao Wang[1*], Jianfu Zhang[2*], Li Niu[1†], Haoyu Ling[1],Xue Yang[1],Liqing Zhang[1†]
[1] Department of Computer Science and Engineering, MoE Key Lab of Artificial Intelligence, Shanghai Jiao Tong University
[2] Tensor Learning Team, RIKEN AIP
{wwt117,ustcnewly,smallling,yangxue-2019-sjtu,lqzhang}@sjtu.edu.cn, jianfu.zhang@riken.jp

The supplementary material contains the following parts: (1) Details of Network: illustrate the architecture details of our network; (2) Experiments on High-Resolution Image Inpainting; (3) Additional Ablation Study: provide additional ablative studies for our network architecture modifications, inpainting priority and representation with high-resolution; (4) Attention Map Visualization: visualize the effect of attention fusion method; (5) User Study: provide user study results to compare our proposed method and other state-of-the-art methods; (6) Model Complexity and Inference Time; (7) Additional Qualitative Comparison: provide more visual comparison results on CelebA [4], Paris Street View [1], and Places2 [13] with regular and irregular holes; (8) Additional Quantitative Comparison: provide additional quantitative comparisons on CelebA.

1. Details of the Proposed Parallel Multi-Resolution Fusion Network

As shown in Figure 1 in the main text, our whole network contains a starting sub-network and a main body sub-network. The input of the starting sub-network includes masked image $\boldsymbol{I}_m$ of size $256 \times 256 \times 3$ and corresponding mask $\boldsymbol{M}$ of size $256 \times 256 \times 1$. After a 3×3 convolution to $\boldsymbol{I}_m$ and $\boldsymbol{M}$, the extracted features are processed with a group of 4 residual blocks and finally fed into the main body sub-network. The residual block consists of two duplications of PConv 3×3, BatchNorm, ReLU and the input features are summed with output features using a skip connection.

The main body sub-network (input size 256×256) consists of four parallel branches with four different resolutions, in which each branch consists of multiple sub-networks with one sub-network belonging to one stage. The information from different branches is exchanged at the end of each stage. The resolution of the feature map for each branch from high to low is 256, 128, 64, 32, respectively. For each stage in each branch, we use a group of 4 residual blocks to extract features. For each residual block, the input channels and output channels stay the same and the channel numbers are 32, 64, 128, 256 for resolution 256, 128, 64, 32, respectively. To enlarge the receptive field of the feature map, we use dilated convolutions in the second and the third residual block with dilation rates $(1, 2)$ and $(4, 8)$. The resolution of feature maps in each branch remains the same. Prior to the last stage, we use the self-attention learned from all resolution feature maps to guide the refinement of each resolution. All convolutions used in our network are partial convolutions. The discriminator used in our network is the same as [12], which contains a self-attention layer and several convolutional layers, for the detailed structure of the discriminator, please refer to [12].

[*] Equal Contributions.
[†] Corresponding author.

Settings	ℓ_1 (%)$^{\downarrow}$	SSIM$^{\uparrow}$	PSNR$^{\uparrow}$	FID$^{\downarrow}$
HF [8]	2.60	0.896	24.855	111.28
PF [10]	2.07	0.917	26.337	109.13
Ours	**1.83**	**0.925**	**26.821**	**107.05**

Table 1. Quantitative results on 1024×1024 images.

2. Performance on High-Resolution Image

We also explore the ability of our model on high-resolution image inpainting. We qualitatively and quantitatively compare our model with two state-of-the-art high-resolution inpainting networks: HF [8] and PF [10]. One hundred images with the size of 1024×1024 randomly selected from Places2 [13] are used as the test set. Our model is retrained on Places2 with 512×512 images. The quantitative results of the three methods are shown in Table 1 and our method achieves the best results. In Figure 1, we show the visual comparison results on 1024×1024 images. HF is good at processing images with simple scenes

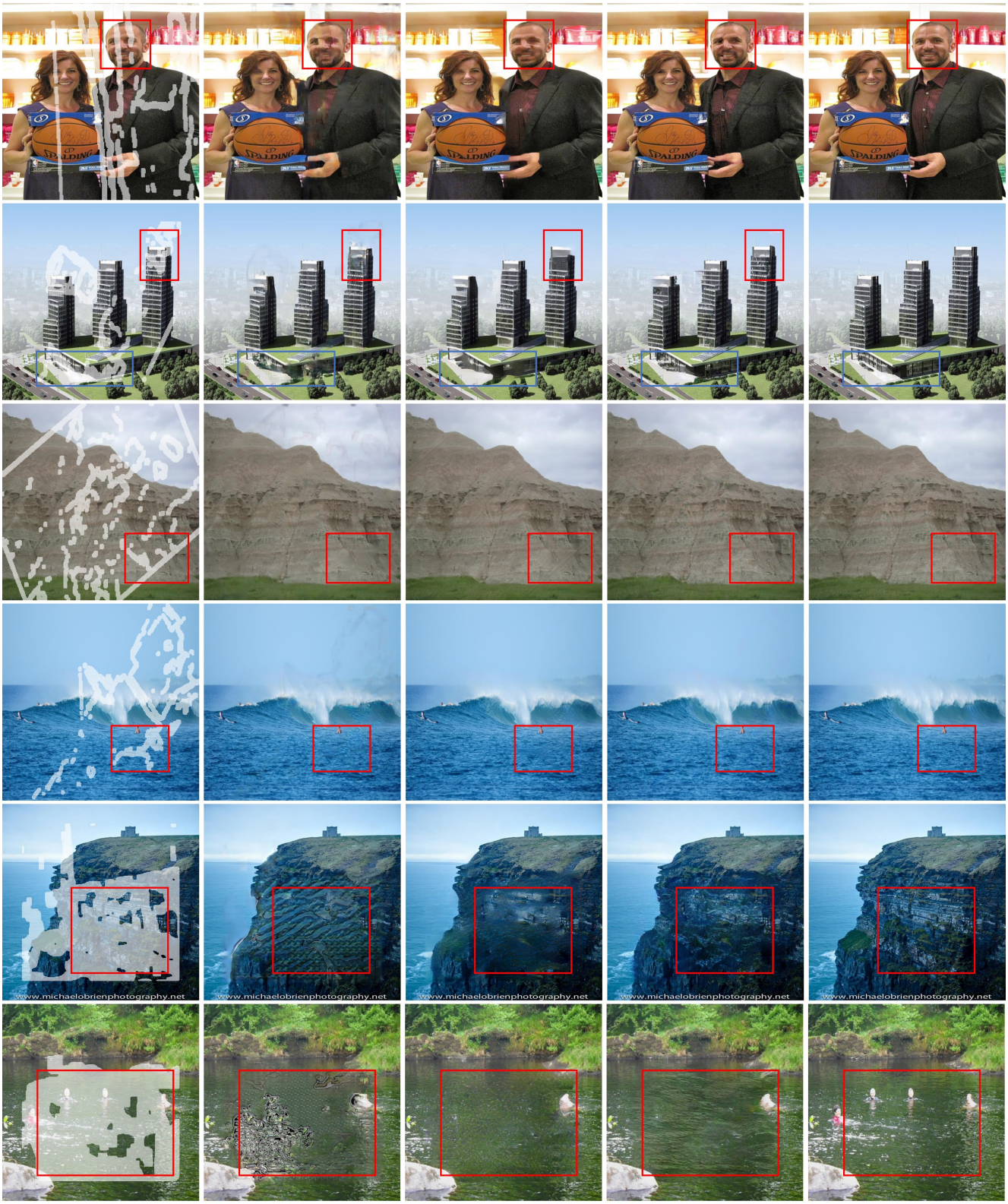


Figure 1. The visual comparison results on 1024 × 1024 images. Best viewed by zooming in.

Settings	ℓ_1 (%)$^\downarrow$	SSIM$^\uparrow$	PSNR$^\uparrow$	FID$^\downarrow$
start-first	3.69	0.8354	25.007	7.95
4-stage	3.73	0.8329	25.124	9.18
5-stage	3.67	0.8362	25.283	7.43
7-stage	3.56	0.8407	25.612	7.41
8-stage	3.58	0.8403	25.580	8.02
selected	3.54	0.8410	25.475	6.98

Table 2. Analyses for network architecture.

on ultra-high-resolution image, but has difficulty in dealing with complex scenes. PF can generate comparable results with ours, but our results have more delicate textures and consistent structures.

3. Additional Ablation Study

3.1. Slight Modifications

Recall that compared with [7], we make two slight modifications to the overall network architecture. Thus, we conduct two additional ablative studies to prove the effectiveness of our modifications. (1) The network architecture in [7] starts from a high-resolution sub-network as the first stage, and gradually adds high-to-low resolution sub-networks one by one to form more stages. Now, tailoring for image inpainting to focus on both local and global information earlier, our network starts from four resolutions at the beginning. To investigate the effectiveness of this modification, we compare the results of the network starting from one resolution ("start first") and our network ("selected"). As shown in Table 2, the selected network structure performs better than the network starting from one resolution. (2) We found that large missing regions are unable to be repaired completely in the shallow network. To guarantee adequate stages for inpainting all types of missing regions, we conduct experiments to verify the inpainting performance on 4-stage, 5-stage, 6-stage ("selected"), 7-stage, 8-stage networks, respectively. In Table 2, it can be seen that the results generated by 6-stage network are superior to 4-stage and 5-stage networks. We also observe that 6-stage network is competitive compared with 7-stage and 8-stage network. Considering the trade-off between model complexity and inpainting performance, we finally choose 6-stage network instead of the network with more stages.

3.2. Inpainting Priority

In this section, we analyze the effect of the inpainting priority. First, we vary the hyper-parameter δ in Eqn. 6 in the main text in $\{0, 0.1, 0.3, 0.5, 0.7, 0.9\}$. Note that our method uses $\delta = 0.5$ by default. $\delta = 0$ represents directly use PConv [2]. In Table 3, we report the inpainting metrics and the average number of stages the network takes to completely fill the missing region. We can see that when δ is large, the network inpaints the missing region slower (*i.e.*, with more stages). Then we vary the setting of inpainting priority by (a) only using common priority (CP); (b) only use resolution-specific priority (RP); (c) replace high-resolution priority with low-resolution priority (LP). All the results are summarized in Table 3. We can see that only using common priority (CP) or resolution-specific priority (RP) will accelerate inpainting the missing regions, but degrade the quality of generated images. When replacing high-resolution priority with low-resolution priority (LP), both inpainting speed and inpainting quality are compromised.

Settings	ℓ_1 (%)$^\downarrow$	SSIM$^\uparrow$	PSNR$^\uparrow$	FID$^\downarrow$	#Stages
32^2	3.86	0.832	25.024	12.26	-
64^2	3.81	0.831	24.871	10.49	-
128^2	3.61	0.838	25.299	8.61	-
$\delta = 0.9$	3.87	0.831	24.733	12.31	2.86
$\delta = 0.7$	3.62	0.838	25.373	7.43	2.55
$\delta = 0.3$	3.64	0.840	25.267	7.45	1.52
$\delta = 0.1$	3.68	0.836	25.010	8.53	1.00
$\delta = 0$	3.69	0.833	24.898	8.64	1.00
CP	3.65	0.831	24.871	8.51	1.18
RP	3.62	0.836	25.273	7.92	1.64
LP	3.58	0.841	25.370	7.32	2.11
Full-Fledged	3.54	0.841	25.475	6.98	1.95

Table 3. Analyses for inpainting priority and representation with high-resolution.

3.3. Representation with High-Resolution

We analyze the necessity of maintaining high-resolution representation in image inpainting by comparing the quality of images output by four branches. For the sake of fairness, we keep the whole network and use the final stage output of specific resolution to generate the inpainted images. Taking resolution 32^2 as an example, we upsample the feature map from resolution 32^2 to resolution 256^2, then add a convolution layer with kernel 1×1 to output the inpainted images. The results are shown in Table 3. We can see that results using low-resolution representations are worse than those using the highest-resolution representations (our full-fledged model), which shows the necessity of maintaining high-resolution representations.

We also show an example of masked image and the order of inpainting its missing region in Figure 2. In this figure, we show the masked image ("Masked"), the ground-truth image ("Truth"), and our inpainted image ("Output"). Results and masks in different branches (with different resolutions) at different stages are provided. We can see that at different resolution levels, the order of inpaint the missing

region has different preferences. The low-resolution branch firstly inpaints the missing regions around the border between sea and sky, while the high-resolution branches firstly inpaint the waves on the sea. The missing regions disappear after the third stage. More examples of masked image and inpainting order are shown in the rest part of Figure 2. The low-resolution branch firstly inpaints the missing regions with rich structure information (*e.g.*, borders), while the high-resolution branches firstly inpaint the regions with more texture information (*e.g.*, grass, hair).

4. Attention Map Visualization

In Figure 3, we visualize the attention map of our attention-guided representation fusion method. The attention maps are obtained by averaging the attention score maps a (in Eqn. 3) of pixels within the masked region. It can be seen that the masked region attends relevant information from unmasked region [11, 8], e.g., wheat land (resp., sky) for wheat land (resp., sky).

We also compared our method with a naive fusion method that simply concatenate the feature maps. The obtained relative L1, SSIM, PSNR, FID of latter are 3.59%, 0.836, 25.320, 8.21. In comparison, relative L1, SSIM, PSNR, FID of our attention-guide fusion method are 3.54%, 0.841, 25.475, 6.98, which proves the effectiveness of our attention-guided fusion module.

5. User study

Following [9], we conduct user study on 120 images randomly selected from both datasets, in which every 20 images are processed with one of six mask groups. 30 subjects with basic background in computer vision are invited to rank the subjective visual qualities of images. We perform three pairwise comparisons for each baseline with our method: (1) Our method *v.s.* GC, (2) Our method *v.s.* EC, (3) Our method *v.s.* SF, (4) Our method *v.s.* HF, (5) Our method *v.s.* MEDFE. A total of $120 \times 30 = 3600$ comparisons were conducted for each baseline. The study shows that 79.30% (2855 out of 3600), 75.47% (2717 out of 3600), 84.50% (3042 out of 3600), 88.14% (3173 out of 3600) and 82.38% (2966 out of 3600) of comparisons preferred our results over GC, EC, SF, HF, and MEDFE, respectively.

6. Model Complexity and Inference Time

We compare our model complexity and inference time with compared to other baseline methods as shown in Table 4.

Our method has a competitive model size and inference speed compared with EC. The reason is that our architecture is capable of parallel processing to reduce running time and the channels in higher resolution is smaller (L044-045 in the supplementary) which does not introduce too many parameters. SF has a large model with longer inference time. Although GC and HF has a small model, HF performs worse

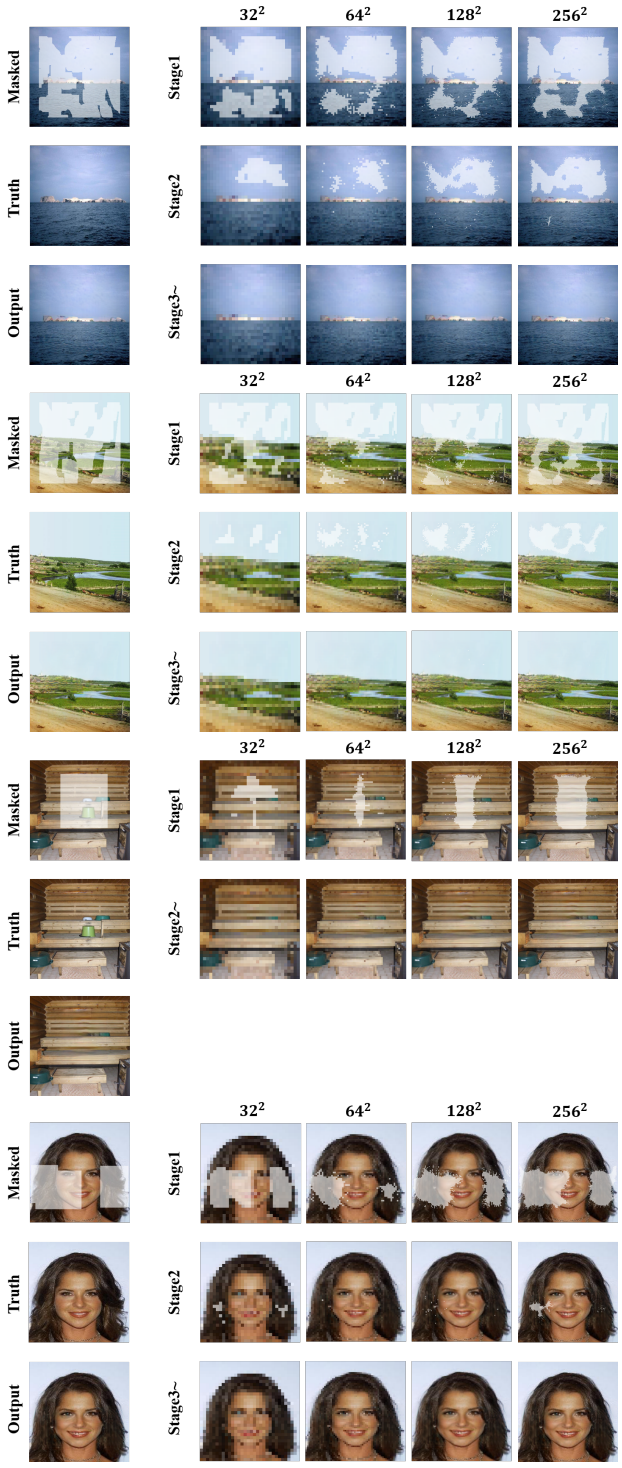


Figure 2. Samples for the process of inpainting priorities. "Stage2∼" and "Stage3∼" which are labeled with "∼" symbol represent that the missing area is completely filled in that stage.

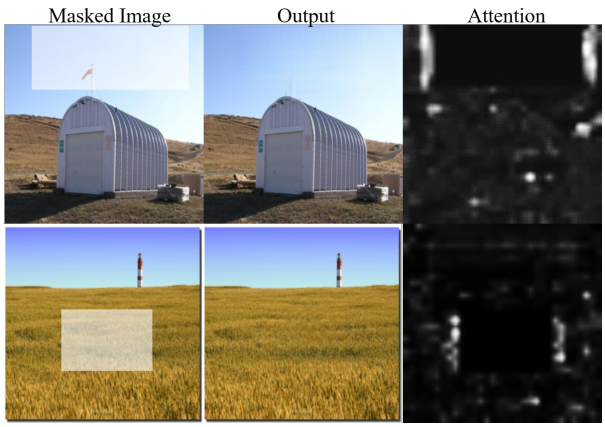


Figure 3. Attention map visualization. Best viewed by zooming in.

	GC	EC	SF	HF	MEDFE	Ours
Inference time (s/frame)	0.048	0.081	0.105	0.024	0.146	0.071
Model size (M)	10.0	27.1	93.7	2.7	130.3	24.3

Table 4. Comparison of model complexity and inference time.

than other methods (See Table 1 and Figure 3 in the main text) and GC struggles to deal with large mask circumstance compared to our method.

7. Additional Quantitative Comparison

We conduct an additional quantitative comparison with GC, EC, SF, and MEDFE on CelebA. The test setting on CelebA is the same as that on Places2. Note that the results of MEDFE on CelebA are terrible because the released pretrained model from MEDFE is trained on center regular mask. This is also the reason that we omit the qualitative results of MEDFE on CelebA in Figure 6. From Table 5, it can be seen that our method outperforms other methods for all evaluation metrics and all mask ratios.

8. Additional Qualitative Comparison

We conduct more qualitative comparison with GC [9], EC [5], SF [6], HF [8], MEDFE [3] on CelebA, Paris Street View, and Places2 with regular and irregular holes. The results of HF on CelebA and Paris Street View are omitted since only the pretrained test model on Places2 is officially released. Figure 4, Figure 5, Figure 6 show the comparison results on CelebA, Paris Street View, and Places2 respectively. From the results, GC, HF tend to generate results with distorted content or artifacts. EC is good at maintaining structural consistency by applying prior edge constraints but there exists color discrepancies in some results. The results of SF have severe color discrepancies. Although MEDFE takes texture and structure into account, there still exist some blurry textures and unreasonable semantics. Our

	Mask	GC [9]	EC [5]	SF [6]	MEDFE [3]	Ours
ℓ_1 (%)$\downarrow$	0-10%	0.79	0.67	1.52	1.43	**0.66**
	10-20%	1.19	1.08	2.03	2.95	**1.05**
	20-30%	1.83	1.66	2.67	5.16	**1.56**
	30-40%	2.63	2.45	3.36	7.64	**2.25**
	40-50%	3.67	3.47	4.18	10.37	**3.06**
	50-60%	5.78	5.48	5.71	13.91	**4.99**
	Ave%	3.29	3.01	4.09	7.67	**2.82**
SSIM$\uparrow$	0-10%	0.976	0.981	0.942	0.946	**0.983**
	10-20%	0.951	0.957	0.915	0.877	**0.961**
	20-30%	0.914	0.920	0.882	0.792	**0.931**
	30-40%	0.870	0.878	0.848	0.708	**0.893**
	40-50%	0.819	0.827	0.809	0.622	**0.852**
	50-60%	0.738	0.747	0.747	0.543	**0.767**
	Ave%	0.858	0.865	0.857	0.745	**0.875**
PSNR$\uparrow$	0-10%	37.175	37.478	33.715	27.922	**38.566**
	10-20%	32.437	32.629	30.623	23.155	**33.447**
	20-30%	28.838	29.092	28.200	19.918	**30.031**
	30-40%	26.131	26.406	26.281	17.725	**27.288**
	40-50%	23.849	24.120	24.521	16.071	**25.441**
	50-60%	20.714	21.103	21.883	14.482	**21.897**
	Ave%	28.945	29.291	27.537	21.202	**30.232**
FID$\downarrow$	0-10%	0.57	0.59	1.00	11.88	**0.48**
	10-20%	1.14	1.01	1.98	33.57	**0.88**
	20-30%	2.84	1.72	3.02	68.56	**1.32**
	30-40%	6.20	2.76	4.13	103.23	**2.19**
	40-50%	11.23	4.26	5.60	136.40	**3.98**
	50-60%	19.88	7.41	8.80	151.64	**5.34**
	Ave%	10.23	4.00	4.10	81.76	**2.91**

Table 5. Quantitative results of different methods on CelebA.

method generates the most appealing results, which have fine-grained textures and reasonable structures.

References

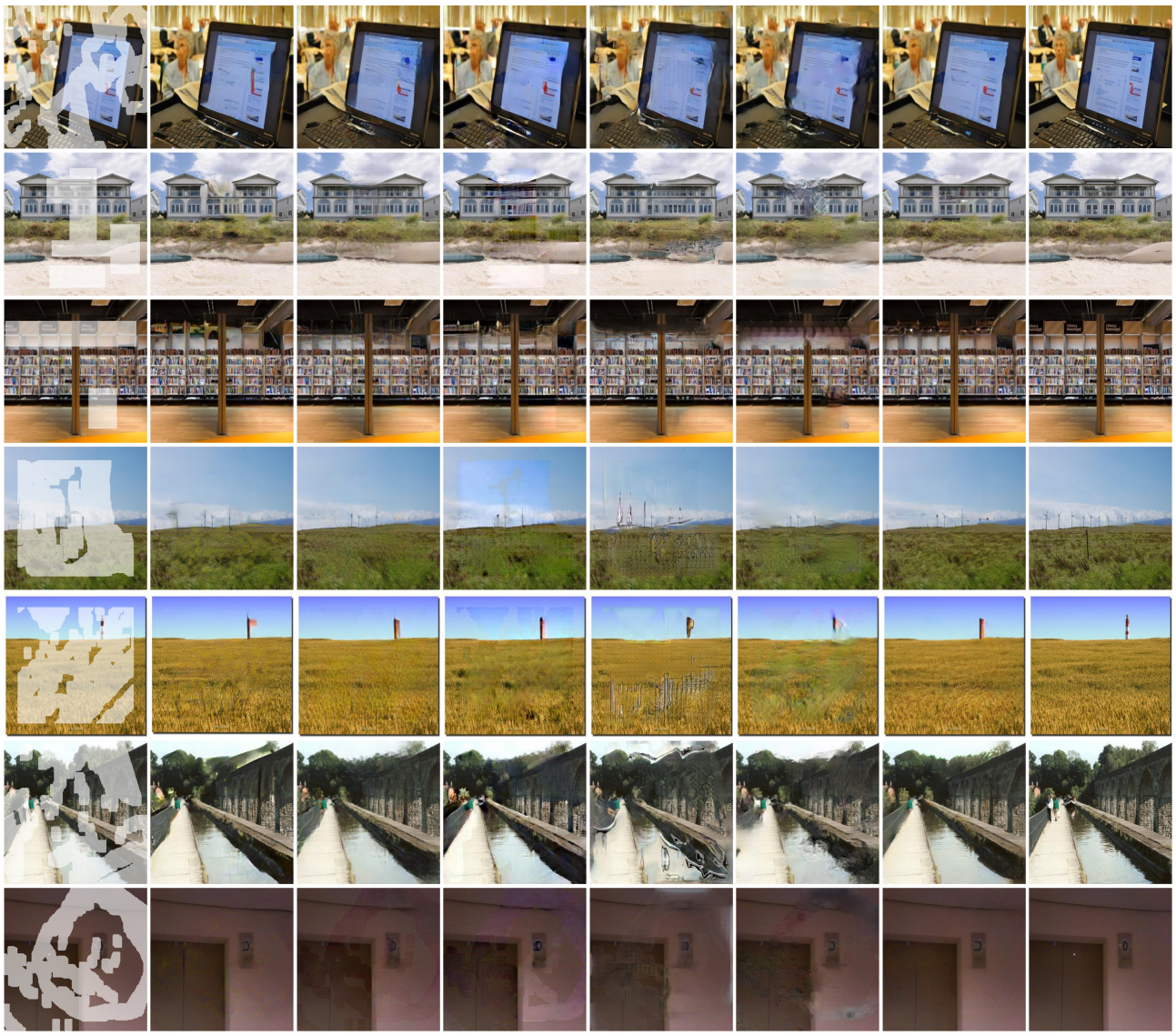
[1] Carl Doersch, Saurabh Singh, Abhinav Gupta, Josef Sivic, and Alexei Efros. What makes paris look like paris? *Acm Transactions on Graphics*, pages 1–9, 2012.

[2] 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*, pages 85–100, 2018.

[3] Hongyu Liu, Bin Jiang, Yibing Song, Wei Huang, and Chao Yang. Rethinking image inpainting via a mutual encoder-decoder with feature equalizations. In *Proceedings of the European Conference on Computer Vision*, pages 725–741, 2020.

[4] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3730–3738, 2015.

[5] Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Qureshi, and Mehran Ebrahimi. Edgeconnect: Structure guided image inpainting using edge prediction. In *Proceedings of the IEEE*



(a) Input image **(b) GC** **(c) EC** **(d) SF** **(e) HF** **(f) MEDFE** **(h) Ours** **(g) Ground Truth**

Figure 4. More visual comparison results on Places2 [13].

International Conference on Computer Vision Workshops, 2019.

[6] Yurui Ren, Xiaoming Yu, Ruonan Zhang, Thomas H Li, Shan Liu, and Ge Li. Structureflow: Image inpainting via structure-aware appearance flow. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 181–190, 2019.

[7] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5693–5703, 2019.

[8] Zili Yi, Qiang Tang, Shekoofeh Azizi, Daesik Jang, and Zhan Xu. Contextual residual aggregation for ultra high-resolution image inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7508–7517, 2020.

[9] 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*, pages 4471–4480, 2019.

[10] Yu Zeng, Zhe Lin, Jimei Yang, Jianming Zhang, Eli Shechtman, and Huchuan Lu. High-resolution image inpainting with iterative confidence feedback and guided upsampling. In *Proceedings of the European Conference on Computer Vision*, pages 1–17, 2020.

[11] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *International Conference on Machine Learning*, pages

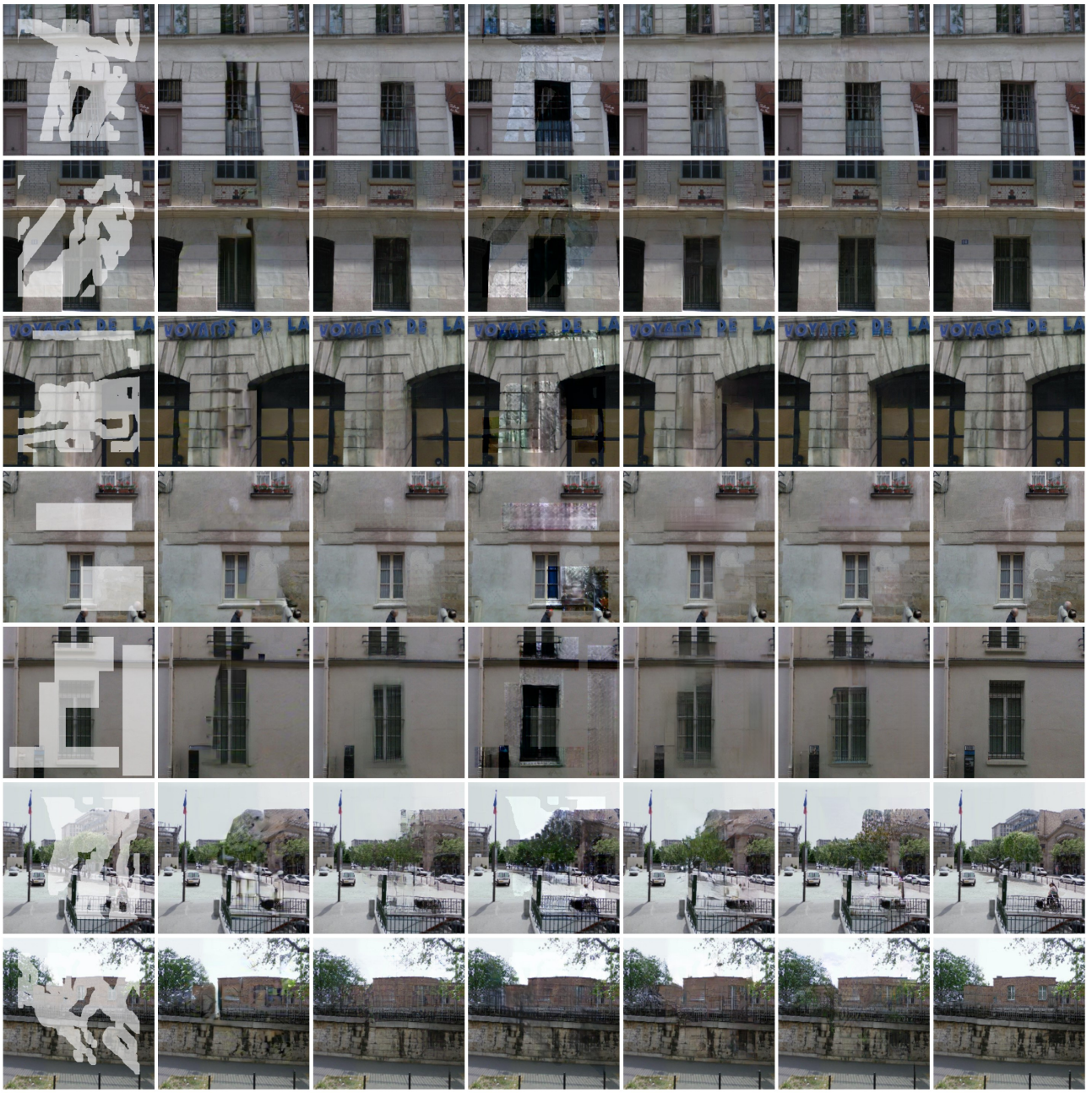


Figure 5. More visual comparison results on Paris Street View. [1]

7354–7363, 2019.

[12] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. Pluralistic image completion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1438–1447, 2019.

[13] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017.

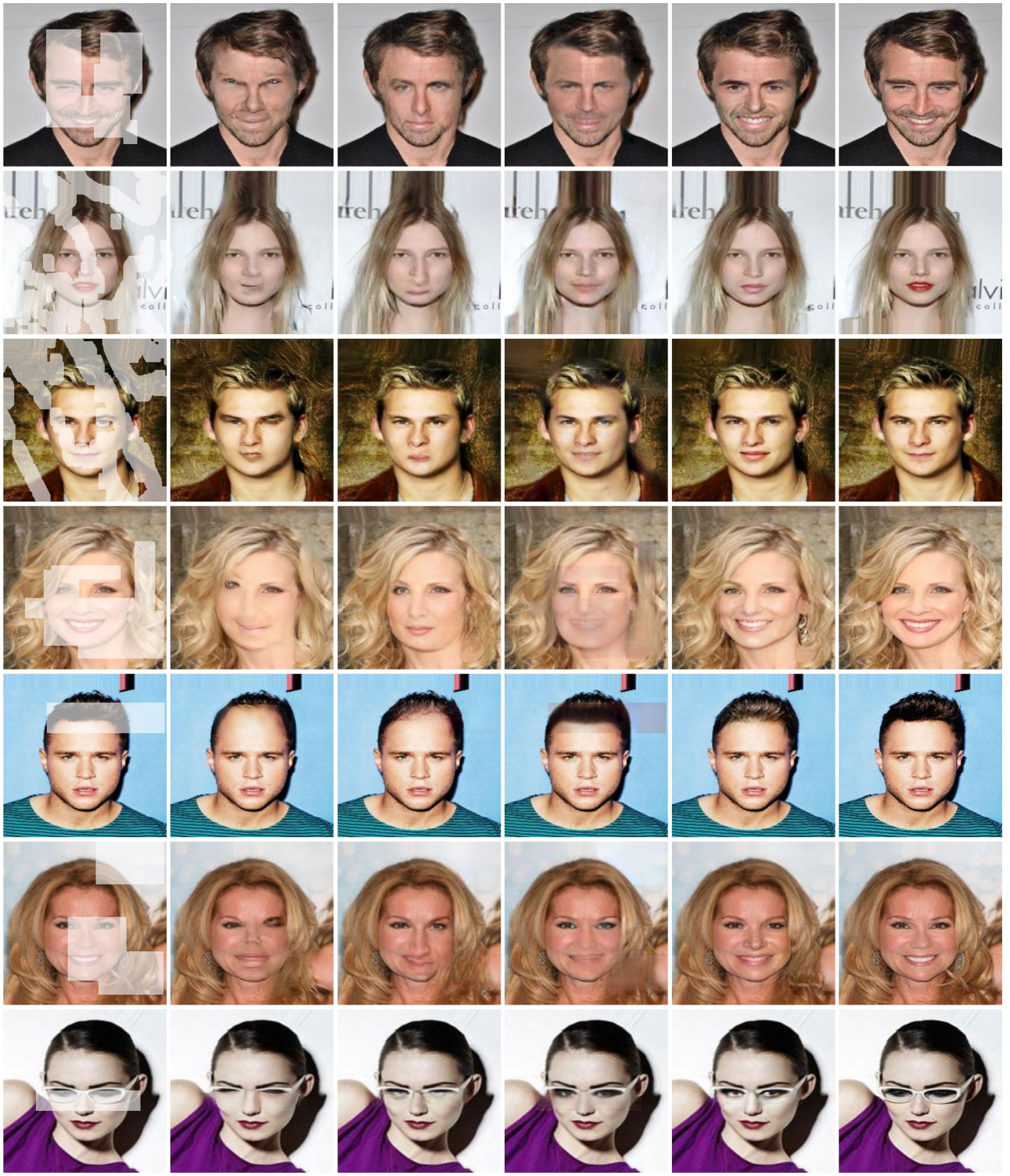


Figure 6. More visual comparison results on CelebA [4].