WaveFill: A Wavelet-based Generation Network for Image Inpainting

1. Appendix Outline

We provide more details on experimental settings, detailed WaveFill network architecture, more experimental results, and more ablation studies in the ensuing 4 sections.

2. Detailed Experimental Settings

Image Processing: We fix the image resolution at 256×256. Specifically, we use small images (256×256) of Places2 [13] and resize CelebA-HQ [2] images to 256×256. For Paris StreetView [7] images (936×537), we crop patches of 537×537 from the image center and then resize the cropped patches to 256×256. Given a raw image I_{gt} and its mask M (0 for valid pixels and 1 for invalid) of the target size, the input image is obtained by $I_{in} = I_{gt} \odot (1 - M)$, where $\odot$ is element-wise product. The WaveFill generator takes $[I_{in}, M]$ as inputs and produces the prediction I_{pred} in the spatial domain. The final output is $I_{out} = I_{in} + I_{pred} \odot M$.

Ablation Study Settings: In the ablation studies presented in the submitted manuscript, the baseline adopts a U-Net-like architecture [8] for image inpainting in spatial domain. It uses multiple residual blocks (as in the low-frequency branch of WaveFill generator) between the CNN encoder and decoder, and replaces all vanilla convolution by gated convolution [11]. For a fair comparison, it also performs feature fusion by concatenating low-level features from encoder with up-sampled high-level features from decoder followed by a gated convolution.

For DCT-based model, we apply Discrete Cosine Transform (DCT) to the entire image. As illustrated in Fig. 1, we split frequency bands into one low-frequency band and two high-frequency bands. We then concatenate the 3 square patches of high-frequency bands in the channel dimension, and obtain similar inputs as the wavelet-based model. With the frequency domain inputs, we adopt the same network structure of WaveFill to train the DCT-based model.

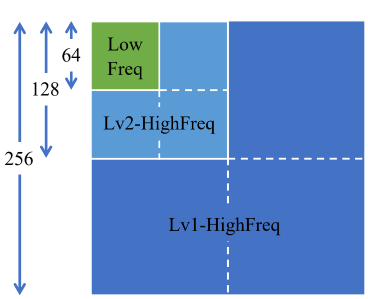

Figure 1. Illustration of frequency separation over the resultant frequency bands with DCT.

Frequency Bands	GMCNN [10]	EC* [6]	GC [11]	Ours
Low	9.03	9.97	**8.99**	9.76
High	43.86	39.60	43.42	**30.81**

Table 1. Histogram differences as measured by Earth Mover's Distance (EMD) between the prediction and ground truth (over CelebA-HQ [2] validation set with square masks).

3. Detailed Network Architectures

The discriminators D_1 and D_2 share the same structure for the predictions of *Lv1-HighFreq* and *Lv2-HighFreq*. Fig. 4 shows detailed discriminator structure. For the generator in WaveFill, Fig. 5 shows it detailed architecture where the IDWT part is not included for brevity.

4. More Experimental Results

Inter-Frequency Conflicts: As most state-of-the-art (SOTA) inpainting methods optimize different objectives in spatial domain concurrently, which often leads to inter-frequency conflicts and compromised inpainting. To visualize and quantify the inter-frequency conflicts, we compute the histogram of low-frequency and high-frequency components of SOTA predictions (as illustrated in Fig. 1 of the submitted manuscript) and measure their differences with that of ground truth via Earth Mover's Distance (EMD) [9]. In particular, the inpainting prediction is decomposed into LL, LH, HL, and HH with 1-level wavelet decomposition where LL is treated as low-frequency bands and the rest are assembled and treated as high-frequency bands. Table 1 shows the EMD over the validation set of CelebA-HQ [2]. It can be found that WaveFill achieves comparable EMD on low-frequency bands but outperforms SOTAs significantly on high-frequency bands. The EMD statistics show that separate generation of different frequency bands with corresponding objectives helps to produce better high-frequency

	Mask	EC [6]	MEDFE [5]	Ours
FID↓	10-20%	20.46	14.79	**12.22**
	20-30%	35.09	27.77	**21.98**
	30-40%	49.43	43.02	**34.02**
	40-50%	62.03	64.98	**46.50**
$\ell_1(\%)\downarrow$	10-20%	1.16	**1.02**	1.04
	20-30%	2.01	1.78	**1.63**
	30-40%	2.94	2.76	**2.37**
	40-50%	4.10	4.14	**3.34**
PSNR↑	10-20%	30.81	**32.80**	32.79
	20-30%	27.87	29.07	**29.86**
	30-40%	25.76	26.31	**27.72**
	40-50%	24.09	23.73	**25.78**
SSIM↑	10-20%	0.948	0.965	**0.966**
	20-30%	0.905	0.929	**0.938**
	30-40%	0.848	0.869	**0.894**
	40-50%	0.778	0.790	**0.839**

Table 2. Quantitative experimental results over Paris StreetView [7] validation images (100) with irregular masks [4].

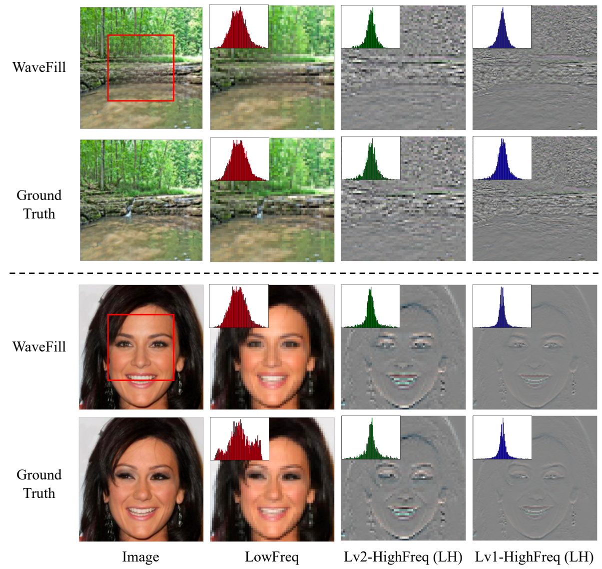

Figure 2. WaveFill inpainting vs GT in multi-frequency bands and histograms of inpainting regions (highlighted by the red box). The above sample is from Places2 [13] and the below one is from CelebA-HQ [2].

distributions and thus more realistic inpainting.

Quantitative Results: Table 2 shows the quantitative results over the Paris StreetView [7] with irregular masks [4] (no space to report it in the submitted manuscript). For small masks with mask ratios of 10-20%, WaveFill achieves superior FID scores and fair ℓ_1, PSNR, and SSIM as performance saturates over these three metrics. For larger and more challenging irregular masks, WaveFill outperforms the state-of-the-art [6, 5] consistently by large margins.

Qualitative Results:

Fig. 2 illustrates our synthesized multi-frequency components and their histograms. In addition, we compare WaveFill with a number of state-of-the-art methods [10, 6, 11, 5] qualitatively. Figs. 6, 7 and 8 show the inpainting of a few more sample images from CelebA-HQ [2], Places2 [13] and Paris StreetView [7], respectively.

Failure Case Analysis: Similar to existing work like EC [6] and GC [11], our method sometimes fails to generate reasonable semantics while handling large corrupted regions as shown in Fig. 3. This happens more for Places2 dataset [13] that captures thousands of objects and scenes of different types. The failure is largely attributed to the insufficient contextual information (around the corrupted regions) with which the model cannot recognize the semantics and produce reasonable contents.

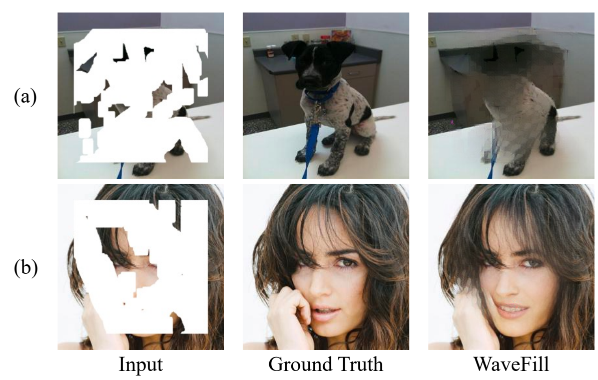

Figure 3. Failure cases from Places2 [13] (above) and CelebA-HQ [2] (below).

5. More Ablation Studies

Decomposition Levels: We studied different levels of wavelet decomposition over Paris StreetView [7] with irregular masks. As Table 3 shows, 2-level wavelet decomposition (ours) performs better than 1-level decomposition in most metrics except PSNR, which could be because 1-level decomposition does not disentangle low and high-frequency information sufficiently and so may still suffer from inter-frequency conflicts. 3-level decomposition performs worse than 2-level decomposition as well largely due to the increased complexity in generation. As we adopt the image size 256×256, the size of low-frequency bands is only 32×32 (in 3-level decomposition) which contains very limited information for guiding the generation of high-frequency contents.

Wavelet Filters: We adopted the Haar wavelet filter as the basis for the wavelet transform. To investigate the effects of wavelet filters over the inpainting performance, we

	FID↓	$\ell_1(\%)\downarrow$	PSNR↑	SSIM↑
Lv1 Wavelet	33.23	2.41	**29.04**	0.900
Lv3 Wavelet	34.36	2.48	28.49	0.898
Lv2 Wavelet	**31.02**	**2.34**	28.94	**0.904**

Table 3. Ablation study of different wavelet decomposition levels over Paris StreetView [7] validation images (100) with irregular masks [4].

	FID↓	$\ell_1(\%)\downarrow$	PSNR↑	SSIM↑
only *LowFreq*	62.32	3.77	25.84	0.820
w/o *Lv2-HighFreq*	54.48	4.18	23.26	0.784
w/o *Lv1-HighFreq*	42.25	2.80	27.87	0.882
Full	**31.02**	**2.34**	**28.94**	**0.904**

Table 4. Ablation study of frequency component over Paris StreetView [7] validation images (100) with irregular masks [4].

evaluated another two types of wavelet filters including *db2* from the Daubechies wavelet family and *bior2.2* from the biorthogonal wavelet family. Table 5 shows experimental results. It can be seen that the three types of wavelet filters achieve comparable inpainting performance under different evaluation metrics.

Frequency Component: The WaveFill generator consists of three branches for processing 3 frequency bands *LowFreq*, *Lv2-HighFreq* and *Lv1-HighFreq*. We trained three models to study the effects of the three branches: 1) Removing both high-frequency branches with only *LowFreq*; 2) Removing intermediate-frequency branch *Lv2-HighFreq*; 3) Removing high-frequency branch *Lv1-HighFreq*. As shown in Table 4, removing lower frequency bands *Lv2-HighFreq* leads to more FID drops as compared with removing *Lv1-HighFreq*, and the model with only *LowFreq* obtains the lowest FID. But for ℓ_1, PSNR and SSIM, removing *Lv2-HighFreq* affects more as compared with using *LowFreq* only. We conjecture that *Lv1-HighFreq* will be like noise when directly adding to *LowFreq* without the support of intermediate-frequency bands *Lv2-HighFreq*.

Loss Term Sensitivity: The complete WaveFill involves 4 losses in training. We performed ablation studies to examine the contribution of each loss by removing it from the full objectives. As shown in Table 6, each loss has its contribution to the overall performance, and $\mathcal{L}_{perc}$ contributes the most.

Filter	FID↓	$\ell_1(\%)\downarrow$	PSNR↑	SSIM↑
db2	32.12	2.35	28.96	**0.904**
bior2.2	31.18	2.36	**29.06**	0.903
haar	**31.02**	**2.34**	28.94	**0.904**

Table 5. Ablation study of wavelet filters over Paris StreetView [7] validation images (100) with irregular masks [4].

	FID↓	$\ell_1(\%)\downarrow$	PSNR↑	SSIM↑
w/o $\mathcal{L}_{LF}$	31.93	2.36	28.81	0.902
w/o $\mathcal{L}_{G}$	32.12	**2.34**	28.85	0.903
w/o $\mathcal{L}_{FM}$	31.66	2.36	28.66	0.901
w/o $\mathcal{L}_{perc}$	41.40	2.63	27.56	0.886
Full	**31.02**	**2.34**	**28.94**	**0.904**

Table 6. Ablation study of loss terms over Paris StreetView [7] validation images (100) with irregular masks [4]. $\mathcal{L}_{LF}$ denotes the low-frequency L1 loss, and $\mathcal{L}_G$ is the adversarial loss for high-frequency bands. $\mathcal{L}_{FM}$ and $\mathcal{L}_{perc}$ denote feature matching loss and perceptual loss respectively.

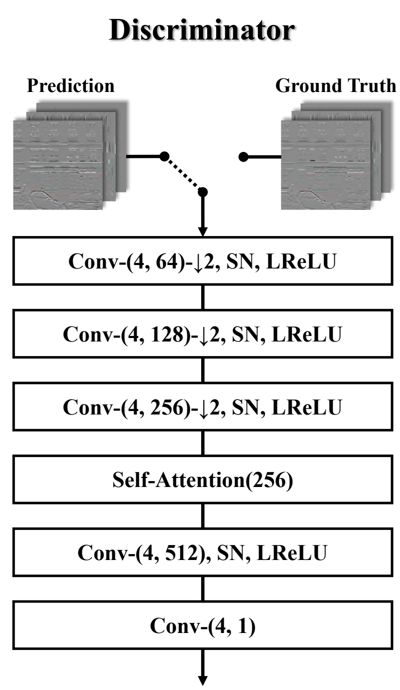

Figure 4. Detailed structures of WaveFill Discriminator: Conv-(k, c, d) denotes the vanilla convolution with kernel size of $k \times k$, number of output channel c and dilation rate d, d is neglected when $d = 1$; LReLU denotes the Leaky ReLU with a slope of 0.2; SN refers to Spectral & Instance Normalization; Self-Attention denotes the Self-Attention module used in [12].

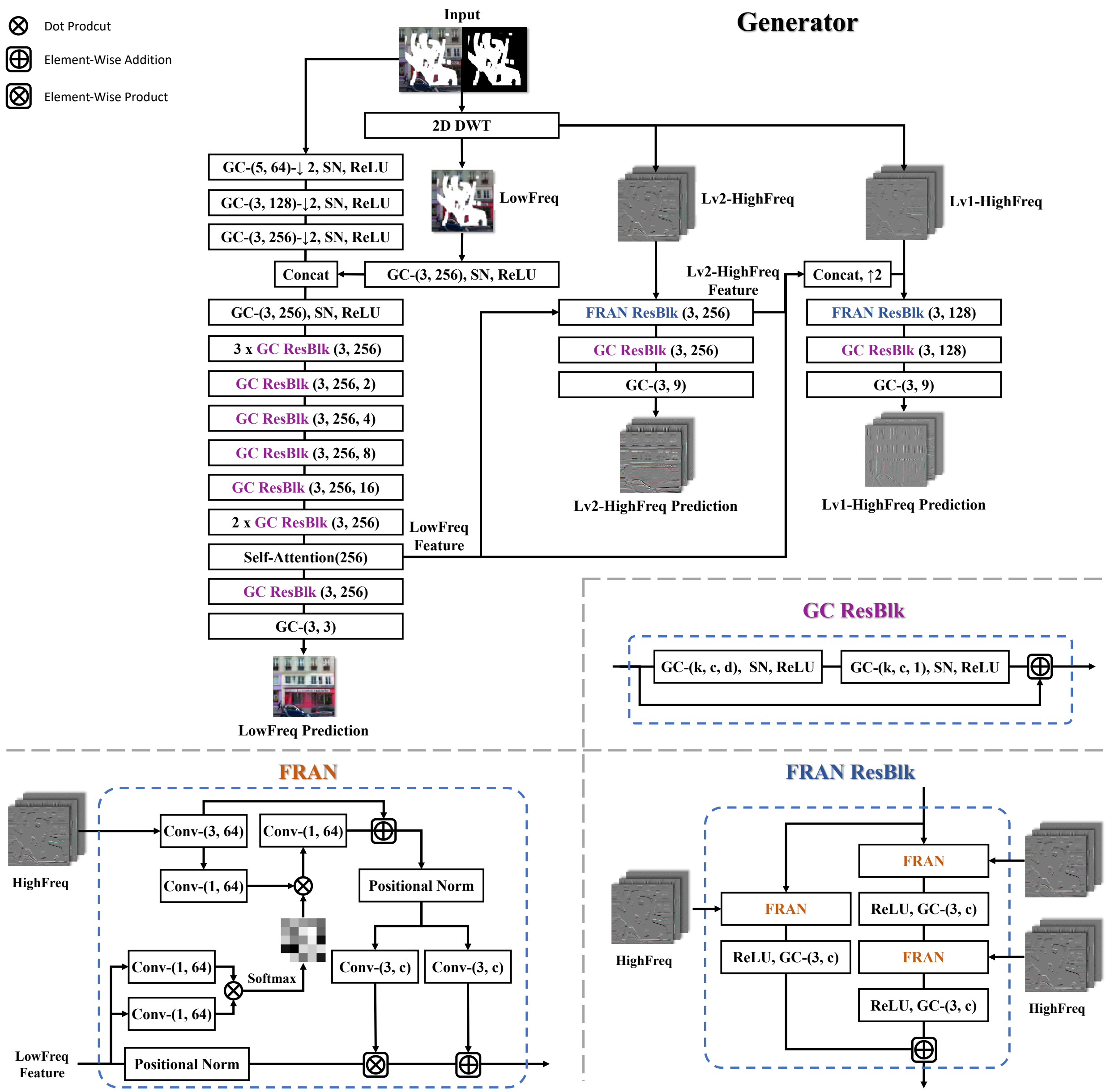

Figure 5. Detailed structures of WaveFill generator: DWT stands for Discrete Wavelet Transform; FRAN denotes the proposed Frequency Region Attention Normalization; GC/Conv-(k, c, d) denotes the gated convolution [11] or vanilla convolution with kernel size of $k \times k$, number of output channel c and dilation rate d, d is neglected when $d = 1$; Resblk is the abbreviation of Residual Block [1]; SN refers to Spectral & Instance Normalization; Positional Norm denotes Positional Normalization [3]; Self-Attention refers to the Self-Attention module used in [12].

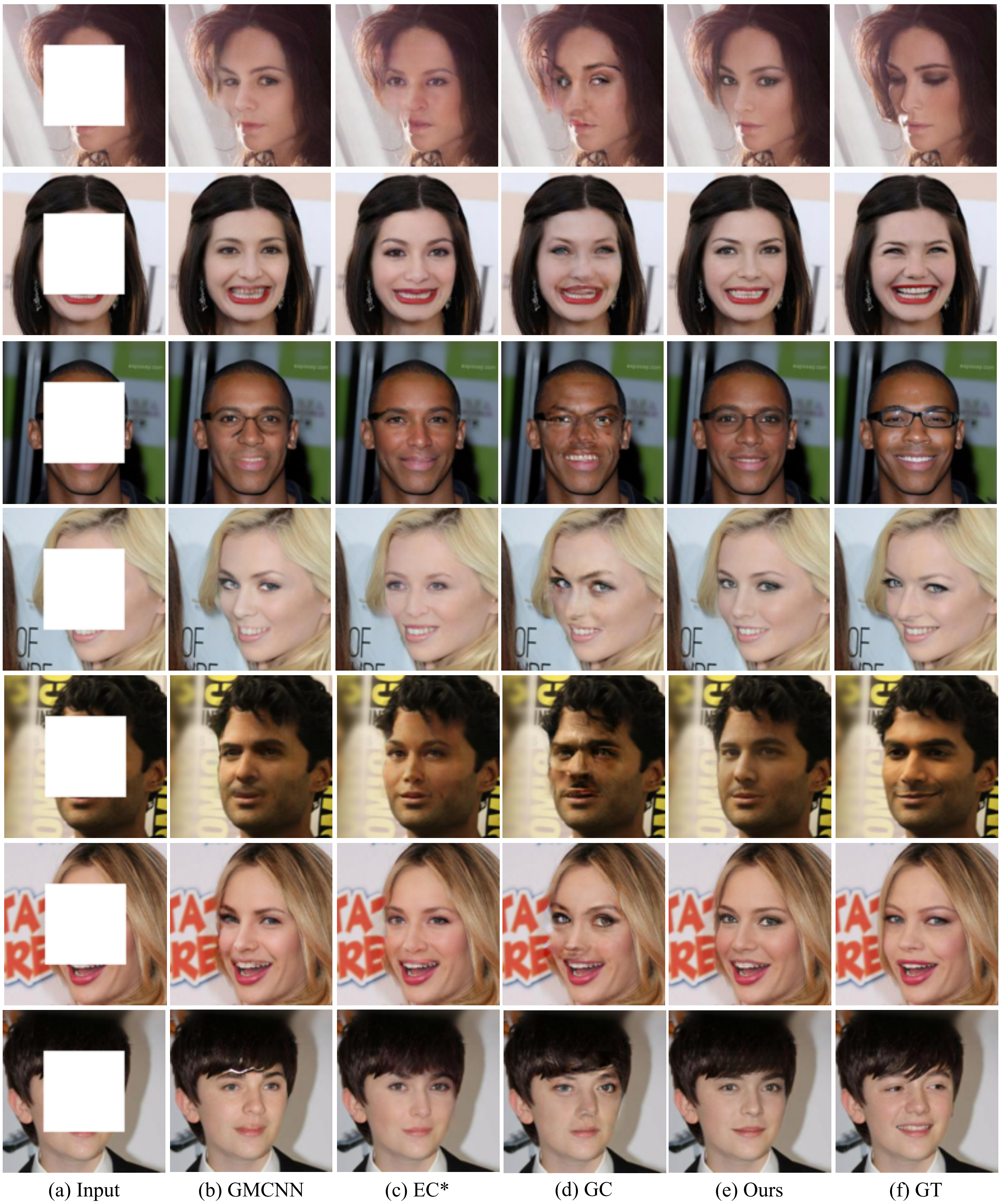

Figure 6. Qualitative comparisons of our WaveFill with several state-of-the-art inpainting methods [10, 6, 11] over the CelebA-HQ [2] validation set with central square masks.

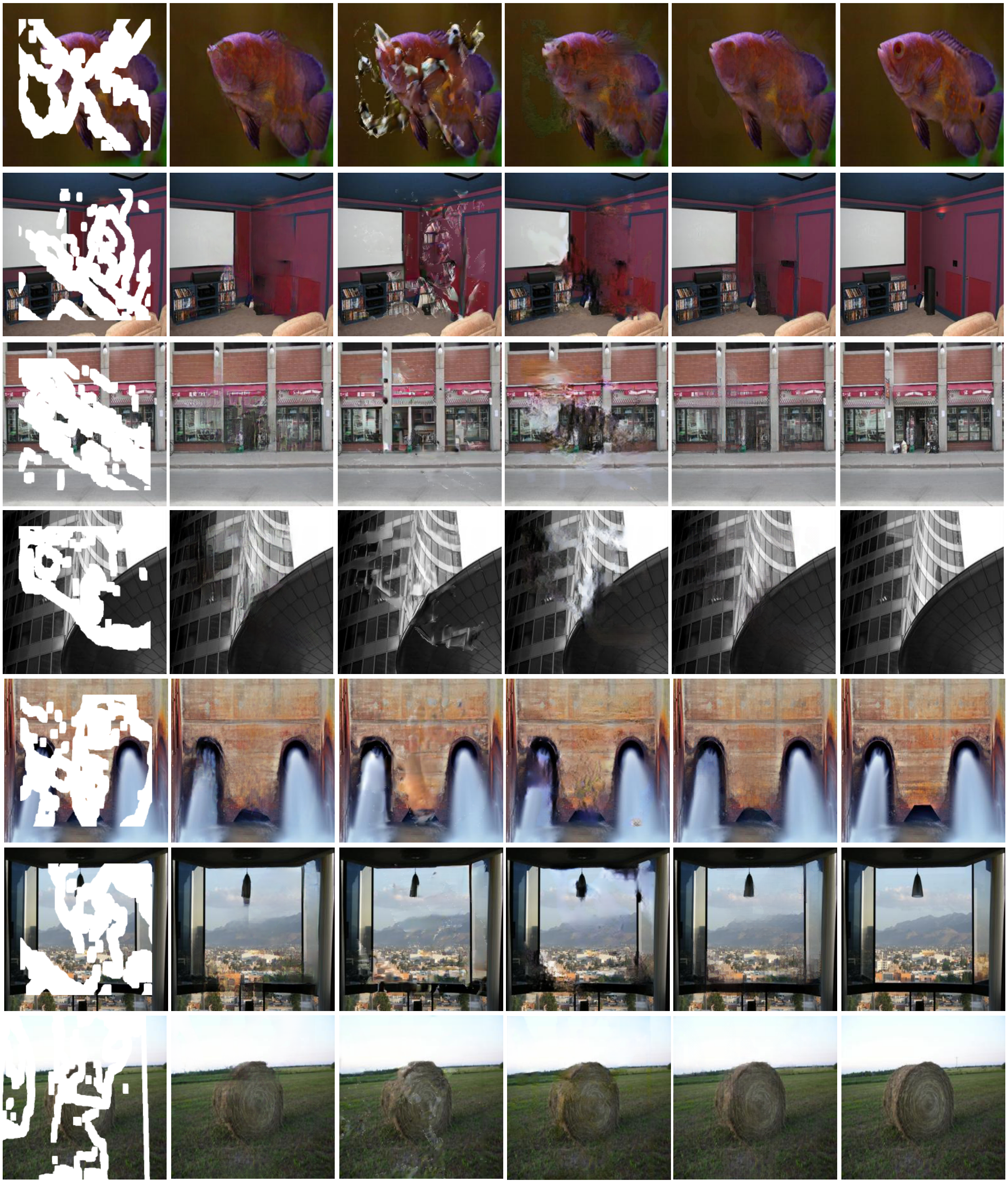

Figure 7. Qualitative comparisons of our WaveFill with several state-of-the-art inpainting methods [6, 11, 5] over the Places2 [13] validation set with irregular masks [4].

(a) Input (b) EC (c) MEDFE (d) Ours (e) GT

Figure 8. Qualitative comparisons of our WaveFill with several state-of-the-art inpainting methods [6, 5] over the Paris StreetView [7] validation set with irregular masks [4]. The red boxes are used to highlight the main differences across different approaches.

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. 4

[2] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *International Conference on Learning Representations*, 2018. 1, 2, 5

[3] Boyi Li, Felix Wu, Kilian Q Weinberger, and Serge Belongie. Positional normalization. In *Advances in Neural Information Processing Systems*, pages 1622–1634, 2019. 4

[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 *European Conference on Computer Vision*, pages 85–100, 2018. 2, 3, 6, 7

[5] Hongyu Liu, Bin Jiang, Yibing Song, Wei Huang, and Chao Yang. Rethinking image inpainting via a mutual encoder-decoder with feature equalizations. In *European Conference on Computer Vision*, 2020. 2, 6, 7

[6] Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Z Qureshi, and Mehran Ebrahimi. Edgeconnect: Generative image inpainting with adversarial edge learning. *arXiv preprint arXiv:1901.00212*, 2019. 1, 2, 5, 6, 7

[7] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2536–2544, 2016. 1, 2, 3, 7

[8] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *arXiv preprint arXiv:1505.04597*, 2015. 1

[9] Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. The earth mover's distance as a metric for image retrieval. *International journal of computer vision*, 40(2):99–121, 2000. 1

[10] Yi Wang, Xin Tao, Xiaojuan Qi, Xiaoyong Shen, and Jiaya Jia. Image inpainting via generative multi-column convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 331–340, 2018. 1, 2, 5

[11] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In *International Conference on Computer Vision*, pages 4471–4480, 2019. 1, 2, 4, 5, 6

[12] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *International Conference on Machine Learning*, pages 7354–7363. PMLR, 2019. 3, 4

[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. 1, 2, 6