

Supplementary Material: MAT: Mask-Aware Transformer for Large Hole Image Inpainting

Wenbo Li¹ Zhe Lin² Kun Zhou³ Lu Qi¹ Yi Wang⁴ Jiaya Jia¹

¹The Chinese University of Hong Kong ²Adobe Inc.

³The Chinese University of Hong Kong (Shenzhen) ⁴Shanghai AI Laboratory

{wenboli, luqi, leojia}@cse.cuhk.edu.hk

zlin@adobe.com kunzhou@link.cuhk.edu.cn wangyi@pjlab.org.cn

A. Network Architecture

As illustrated in Sec. 3.1, the proposed MAT is a two-stage framework, where the first stage consists of a convolutional head, a transformer body and a convolutional reconstruction tail while the second stage is a Conv-U-Net. And the discriminator follows the design of CoModGAN [14].

Given an $H \times W$ input, the head first applies a convolution to change the number of channels from 4 (image 3 + mask 1) to 180 and then adopts three strided convolutions (stride = 2) to downsample the feature size to $\frac{H}{8} \times \frac{W}{8}$. The feature is transformed to tokens as input to the transformer body. The body is composed of five stages of transformer blocks, where the block numbers are $\{2, 3, 4, 3, 2\}$ and the corresponding feature sizes are $\{\frac{H}{8} \times \frac{W}{8}, \frac{H}{16} \times \frac{W}{16}, \frac{H}{32} \times \frac{W}{32}, \frac{H}{16} \times \frac{W}{16}, \frac{H}{8} \times \frac{W}{8}\}$. The downsampling and upsampling are realized by convolutions. The detailed structure of a transformer block is shown in Sec. 3.3. Then the output tokens from the body are converted to a 2D feature, passed to the reconstruction tail. The convolutional tail upsamples the feature size from $\frac{H}{8} \times \frac{W}{8}$ to $H \times W$ and generates a completed image, during which style modulation is applied to all layers to enable pluralistic generation.

The second-stage Conv-U-Net takes in the coarse prediction and the input mask for subsequent high-fidelity detail rendering. It first downsamples the feature size to $\frac{H}{32} \times \frac{W}{32}$ and then upsamples the size back to $H \times W$. Shortcut connections are adopted at each resolution. The number of convolution channels in the encoder starts from 64 and is doubled after each downsampling, with a maximum of 512, while the decoder uses a symmetrical setting. Besides, all decoding layers are modulated by the image-conditional and noise-unconditional style representations.

B. Free-Form Mask Sampling and Statistics

Referring to DeepFill v2 [11], we sample rectangles and brush strokes with random sizes, shapes and locations to



Figure A.1. Examples of free-form masks (512×512). Visible and invisible pixels are in white and black colors.

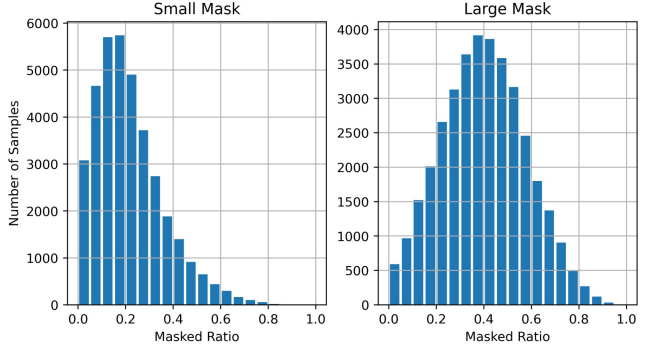


Figure A.2. Small and large mask (512×512) statistics on the Places Val set [15]. The are totally 36500 masks.

generate free-form masks. During training, we use a large mask sampling strategy. The number of up to full-size or half-size rectangles is uniformly sampled within $[0, 3]$ or $[0, 5]$. The number of strokes is randomly sampled within $[0, 9]$, with a random brush width within $[12, 48]$ and vertex number within $[4, 18]$. During testing, apart from the large mask setup, we also introduce a small mask sampling strategy, where the number of up to full-size or half-size rectangles is within $[0, 2]$ or $[0, 3]$ and the number of strokes is within $[0, 4]$, while other settings remain unchanged. Note that our model is trained on large masks and is evaluated on both small and large mask settings. As shown in Fig. A.2, we present the mask statistics on the Places Val set [15] that is used for evaluation. It is observed that large masks are very aggressive and diverse.

Model	FID↓	P-IDS (%)↑	U-IDS(%)↑
Stacked Conv. (Ours)	5.97	13.17	29.23
Linear Projection	10.54	5.77	20.86

Table C.1. Quantitative comparison between linear projection and stacked convolutions for token extraction. We use the same training setting as the ablation study (Sec. 4.3).

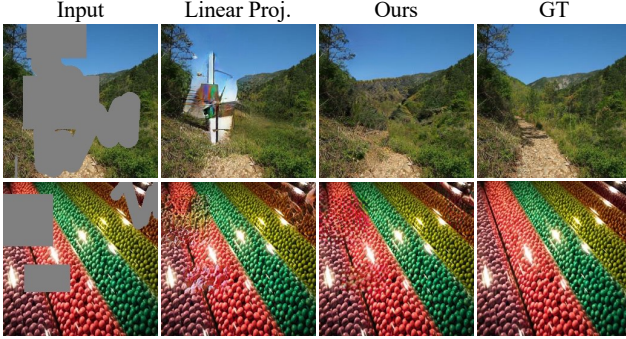


Figure C.3. Qualitative comparison between linear projection and stacked convolutions (ours) for tokenization.

C. Tokenization

As described in Sec. A, we adopt a stack of convolutions (the convolutional head) to extract tokens for the transformer body, which is specially tailored to the inpainting problem. Compared to the linear projection of ViT [2], our design owns two merits. First, stacked convolutions can gradually fill the holes, producing more effective tokens. Second, the multi-scale downsampled features can be passed to the decoder through shortcut connections, improving the optimization. As illustrated in Table C.1 and Fig. C.3, stacked convolutions obtain obviously superior results. The model using linear projection is more likely to generate unpleasing artifacts and fail to borrow surrounding textures to fill the holes, while our MAT successfully recovers high-fidelity contents. Both the quantitative and qualitative results demonstrate the effectiveness of our design.

D. Model Configuration

Following the same experimental setting as ablation study, we explore several model variants in terms of feature width, block number and window size of the transformer body, leaving Conv-U-Net unchanged. The results are shown in the Table D.2. The performance is positively correlated to the model capacity and attention range.

E. CelebA-HQ 256×256 Results

We provide the quantitative results on 256×256 CelebA-HQ [4]. As illustrated in Table F.3, our MAT yields significant improvements on FID [3], P-IDS [14] and U-IDS [13]

Model	Feature Dim.	Block Num.	Window Size	FID↓
Ours	180	{2, 3, 4, 3, 2}	{8, 16, 16, 16, 8}	5.97
V1	90	{2, 3, 4, 3, 2}	{8, 16, 16, 16, 8}	6.28
V2	180	{1, 1, 2, 1, 1}	{8, 16, 16, 16, 8}	6.18
V3	180	{2, 3, 4, 3, 2}	{8, 8, 8, 8, 8}	6.09

Table D.2. Ablation study on model configuration.

Method	Small Mask			Large Mask		
	FID↓	P-IDS↑	U-IDS↑	FID↓	P-IDS↑	U-IDS↑
MAT (Ours)	2.94	20.88	32.01	5.16	13.90	25.13
LaMa [8]	3.98	8.82	22.57	8.75	2.34	8.77
ICT [9]	5.24	4.51	17.39	10.92	0.90	5.23
MADF [16]	10.43	6.25	14.62	23.59	0.50	1.44
AOT GAN [12]	9.64	5.61	14.62	22.91	0.47	1.65
DeepFill v2 [11]	5.69	6.62	16.82	13.23	0.84	2.62
EdgeConnect [7]	5.24	5.61	15.65	12.16	0.84	2.31

Table F.3. Quantitative results on CelebA-HQ at 256×256 size. The results of P-IDS and U-IDS are shown in percentage (%).

Method	#Param. $\times 10^6$	CelebA-HQ		Places	
		Small	Large	Small	Large
MAT (Ours)	60	0.065	0.125	0.099	0.189
CoModGAN [14]†	109	0.073	0.140	0.101	0.192
LaMa [8]†	27/51	0.075	0.143	0.086	0.166
ICT [9]	150	0.105	0.195	-	-
MADF [16]	85	0.068	0.130	0.095	0.181
AOT GAN [12]	15	0.074	0.145	0.101	0.195
HFill [10]	3	-	-	0.148	0.284
DeepFill v2 [11]	4	0.117	0.221	0.113	0.213
EdgeConnect [7]	22	0.101	0.208	0.114	0.275

Table F.4. LPIPS [13] comparison on 512×512 CelebA-HQ [4] and Places [15] datasets. “†”: CoModGAN [14] and LaMa [8] use 8M and 4.5M Places images to train their models, while our model is only trained on Places365-Standard (1.8M images). The LaMa models on CelebA-HQ and Places are different in size.

metrics over other methods.

F. LPIPS Results

As discussed in Sec. 4.1, LPIPS [13] is not an appropriate measure for large mask inpainting, especially for pluralistic generation systems, since there could be numerous plausible solutions to fill the holes. Therefore, we provide the LPIPS results only for reference. As shown in Table F.4, our method achieves superior or comparable performance on the CelebA-HQ [4] and Places [15] datasets. *Note that we only use 22.5% of full data to train our Places model.*

Method	Training Data	FID↓	Precision↑	Recall↑
MAT (Ours)	1.8M	2.90	0.925	0.951
CoModGAN	8M	2.92	0.929	0.942

Table H.5. Precision and Recall results of our MAT and CoModGAN on Places.

G. Generalization to A Higher Resolution

Though trained on 512×512 images, our model generalizes well to larger resolutions. For example, we transfer our model and Big LaMa [8] trained at 512×512 resolution to 1024×1024 . Compared to Big LaMa (FID↓ 6.31, PIDS↑ 4.98%), our model (FID↓ 5.83, P-IDS↑ 9.51%) obtains superior results on Places under the large mask setting. We suggest that maintaining a resolution consistency during training and testing yields better visual quality.

H. Diversity-Fidelity Tradeoff

To evaluate the fidelity and diversity, apart from FID (depending on both diversity and fidelity), we also follow [1, 6] to use Improved Precision and Recall to separately measure sample fidelity (precision) and diversity (recall). As shown in Table H.5, our method obtains better FID, higher recall yet slightly lower precision compared to CoModGAN on Places. It is noted that we use much less training data.

I. Additional Qualitative Results

We present more visual comparisons on the Places [15] dataset between our MAT and other state-of-the-art methods. As shown in Fig J.4 and Fig J.5, our method generates more photo-realistic results with few artifacts, manifesting the effectiveness of MAT. Due to potential copyright issues with CelebA-HQ [4], we do not provide visual comparisons on this dataset. If necessary, you can process CelebA-HQ images with the provided code and model, or contact the authors.

J. Licenses of Face Images

All face images used in the paper and supplementary material are from the FFHQ [5] dataset. Here we provide the detailed information on source and license.

- Face image in Fig.1 of main paper, source: <https://www.flickr.com/photos/v63/5876049365/>, license: CC BY-NC 2.0 (<https://creativecommons.org/licenses/by-nc/2.0/>).
- Face image in Fig.2 of main paper, source: <https://www.flickr.com/photos/tbisaacs/4089001580/>, license: CC BY 2.0 (<https://creativecommons.org/licenses/by/2.0/>).

[//creativecommons.org/licenses/by/2.0/](https://creativecommons.org/licenses/by/2.0/)).

- The first face image in Fig.6 of main paper, source: <https://www.flickr.com/photos/southlanarkshirecouncil/8341157963/>, license: CC BY-NC 2.0 (<https://creativecommons.org/licenses/by-nc/2.0/>).
- The second face image in Fig.6 of main paper, source: <https://www.flickr.com/photos/afge/34804627253/>, license: CC BY 2.0 (<https://creativecommons.org/licenses/by/2.0/>).

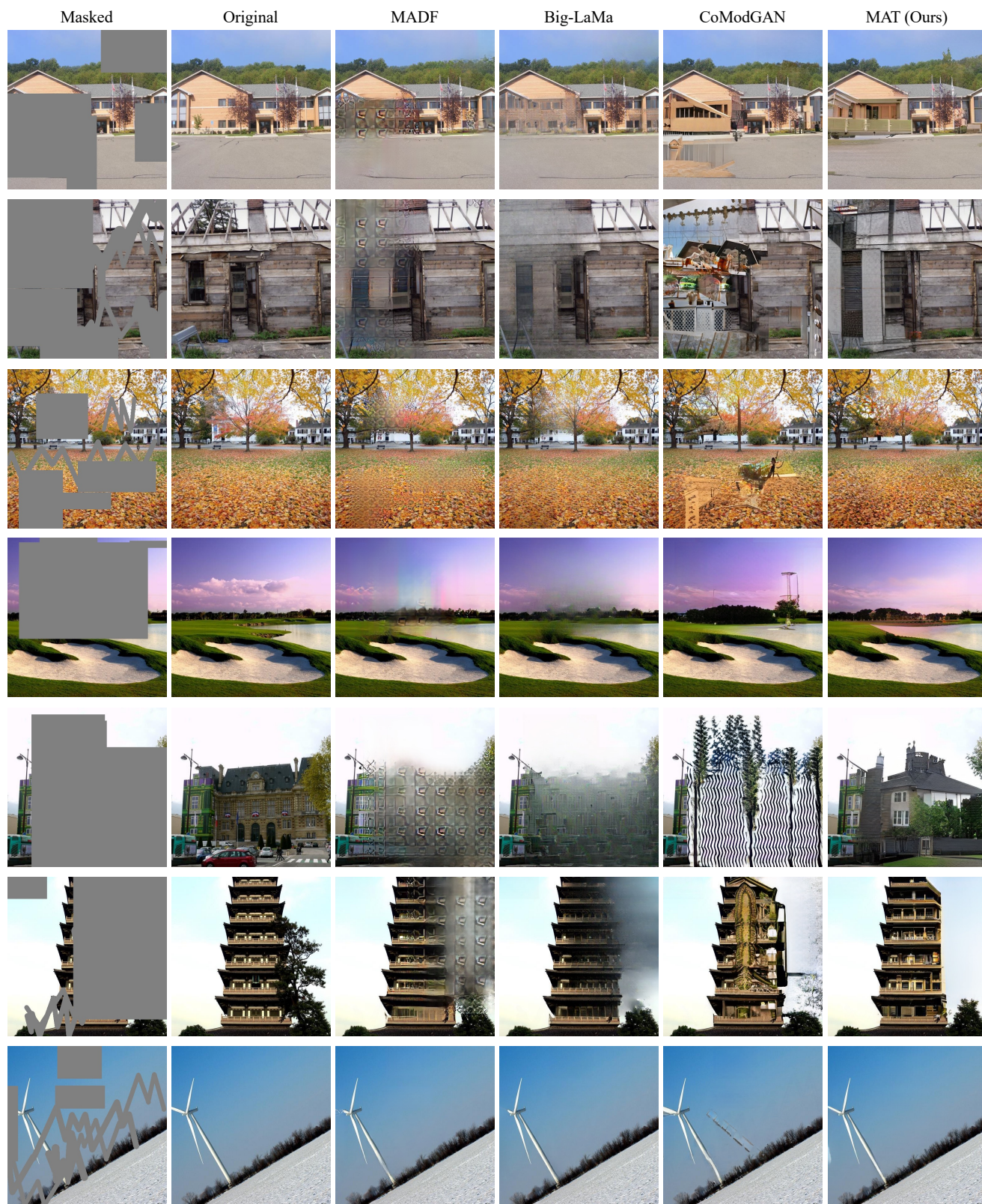


Figure J.4. Qualitative comparison (512×512) with state-of-the-art methods on the Places dataset. Zoom in for a better view.

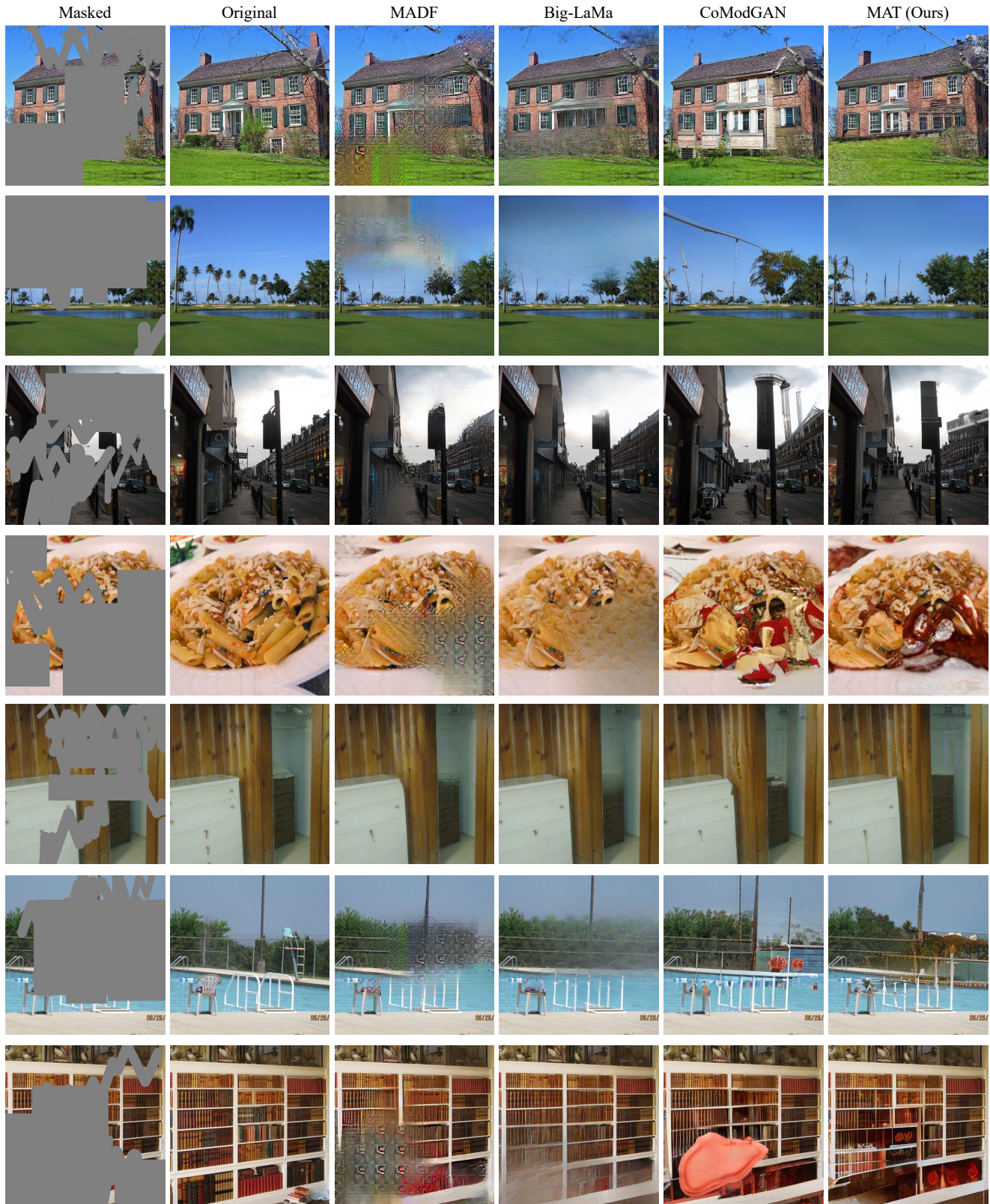


Figure J.5. Qualitative comparison (512×512) with state-of-the-art methods on the Places dataset. Zoom in for a better view.

References

- [1] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *NeurIPS*, 34, 2021. 3
- [2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2020. 2
- [3] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *NIPS*, 30, 2017. 2
- [4] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *ICLR*, 2018. 2, 3
- [5] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, pages 4401–4410, 2019. 3
- [6] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. *NeurIPS*, 32, 2019. 3
- [7] 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. 2
- [8] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. *arXiv preprint arXiv:2109.07161*, 2021. 2
- [9] Ziyu Wan, Jingbo Zhang, Dongdong Chen, and Jing Liao. High-fidelity pluralistic image completion with transformers. *arXiv preprint arXiv:2103.14031*, 2021. 2
- [10] Zili Yi, Qiang Tang, Shekoofeh Azizi, Daesik Jang, and Zhan Xu. Contextual residual aggregation for ultra high-resolution image inpainting. In *CVPR*, pages 7508–7517, 2020. 2
- [11] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In *ICCV*, pages 4471–4480, 2019. 1, 2
- [12] Yanhong Zeng, Jianlong Fu, Hongyang Chao, and Baining Guo. Aggregated contextual transformations for high-resolution image inpainting. *arXiv preprint arXiv:2104.01431*, 2021. 2
- [13] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018. 2
- [14] Shengyu Zhao, Jonathan Cui, Yilun Sheng, Yue Dong, Xiao Liang, I Eric, Chao Chang, and Yan Xu. Large scale image completion via co-modulated generative adversarial networks. In *ICLR*, 2020. 1, 2
- [15] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *PAMI*, 40(6):1452–1464, 2017. 1, 2, 3
- [16] Manyu Zhu, Dongliang He, Xin Li, Chao Li, Fu Li, Xiao Liu, Errui Ding, and Zhaoxiang Zhang. Image inpainting by end-to-end cascaded refinement with mask awareness. *TIP*, 30:4855–4866, 2021. 2