Appendix of Learning Prior Feature and Attention Enhanced Image Inpainting

Chenjie Cao*, Qiaole Dong*, and Yanwei Fu†

School of Data Science, Fudan University
{20110980001,qldong18,yanweifu}@fudan.edu.cn

1 Social Impact

All generated results of both the main paper and the appendix are based on learned statistics of the training dataset. Therefore, the results only reflect biases in those released data without our subjective opinion, especially for the face images from FFHQ. This work is only researched for the algorithmic discussion, and related societal impacts should not be ignored by users.

2 Detailed Network Settings

We provide some details for different model components in this section. **Gated Convolution Block (GC)**. For the GC block used for upsampling prior features from MAE, which contains GateConv2D [12]→BatchNorm→ReLU. And the GateConv2D works with stride=2.
Encoder and Decoder of ACR. The encoder and decoder of ACR are consisted of vanilla Conv2D→BatchNorm→ReLU.
Fast Fourier Convolution Block (FFC). As illustrated in [9], features for FFC are split into local ones encoded by vanilla convolutions and global ones encoded by the spectral transform. The spectral transform is consisted of Fast Fourier Transform (FFT), Conv2D→BatchNorm→ReLU, and the inverse FFT. And both the real and imaginary parts are confirmed in the Conv2D after FFT. After the inverse FFT, local and global features are combined as the final output.

3 Loss Functions of ACR

We provide some details about the loss functions of ACR, which are referred to LaMa [9] and include L1 loss, adversarial loss, feature match loss, and high receptive field (HRF) perceptual loss [9]. L1 loss is only calculated between the unmasked regions as

$$\mathcal{L}_{L1} = (1 - \mathbf{M}) \odot |\hat{\mathbf{I}} - \tilde{\mathbf{I}}|_1, \tag{1}$$

where $\mathbf{M}$ indicates 0-1 mask that 1 means masked regions; $\odot$ means the element-wise multiplication; $\hat{\mathbf{I}}, \tilde{\mathbf{I}}$ indicate the ground truth and predicted images respectively. The adversarial loss is consisted of a PatchGAN [6] based discriminator

loss $\mathcal{L}_D$ and a WGAN-GP [3] based generator loss $\mathcal{L}_G$ as

$$\begin{aligned}\mathcal{L}_D = &- \mathbb{E}_{\hat{\mathbf{I}}}\left[\log D(\hat{\mathbf{I}})\right] - \mathbb{E}_{\tilde{\mathbf{I}},\mathbf{M}}\left[\log D(\tilde{\mathbf{I}}) \odot (1-\mathbf{M})\right] \\ &- \mathbb{E}_{\tilde{\mathbf{I}},\mathbf{M}}\left[\log(1 - D(\tilde{\mathbf{I}})) \odot \mathbf{M}\right], \\ \mathcal{L}_G = &-\mathbb{E}_{\tilde{\mathbf{I}}}\left[\log D(\tilde{\mathbf{I}})\right], \\ \mathcal{L}_{adv} = &\ \mathcal{L}_D + \mathcal{L}_G + \lambda_{GP}\mathcal{L}_{GP},\end{aligned} \tag{2}$$

where $\mathcal{L}_{GP} = \mathbb{E}_{\hat{\mathbf{I}}}||\nabla_{\hat{\mathbf{I}}} D(\hat{\mathbf{I}})||^2$ is the gradient penalty [3] and $\lambda_{GP} = 1e-3$. Moreover, the feature match loss [10] $\mathcal{L}_{fm}$, which is based on L1 loss between discriminator features D_f of true and fake samples as

$$\mathcal{L}_{fm} = \mathbb{E}(|D_f(\hat{\mathbf{I}}) - D_f(\tilde{\mathbf{I}})|_1). \tag{3}$$

Furthermore, we use the HRF loss $\mathcal{L}_{hrf}$ in [9] as

$$\mathcal{L}_{hrf} = \mathbb{E}(\left[\phi_{hrf}(\hat{\mathbf{I}}) - \phi_{hrf}(\tilde{\mathbf{I}})\right]^2), \tag{4}$$

where ϕ_{hrf} indicates a pretrained segmentation ResNet50 with dilated convolutions, which shows superior performance in inpainting compared with vanilla VGG as discussed in [9]. The final loss of our model can be written as

$$\mathcal{L}_{final} = \lambda_{L1}\mathcal{L}_{L1} + \lambda_{adv}\mathcal{L}_{adv} + \lambda_{fm}\mathcal{L}_{fm} + \lambda_{hrf}\mathcal{L}_{hrf}, \tag{5}$$

where $\lambda_{L1} = 10, \lambda_{adv} = 10, \lambda_{fm} = 100, \lambda_{hrf} = 30$ set by the experience.

4 More Implement Details

High-Resolution (HR) Finetuning. To save the computation, we find that dynamic finetune the inpainting model from 256 to 512 resolutions can still achieve competitive results. We gradually reduce the resolution from 512 to 256, and then let them back to 512, which can be seen as a cycle. For each epoch in Places2, we finetune the model with 64 cycles.

The Subset of Places2 for Ablations. To flexibly evaluate our ablation studies, we choose to use a subset of Places2 with 5 scenes of 'bow_window', 'house', 'village', 'dining_room', and 'viaduct' with about 25,000 training images and 500 validation images. This subset contains indoor, outdoor, and natural scenes, which are comprehensive to evaluate the inpainting performance.

Detailed Settings of Places2. We reorganize detailed settings of main experiments and ablations in Tab. 1.

5 User Study

To test the effectiveness of our model with human perception, we conduct user studies on several models. We specifically ask 12 participants who are unfamiliar with image inpainting to judge the quality of inpainted images. FFHQ is

Table 1. Settings of main experiments (Main Exp.) and ablations. 'Res.', 'F.T.' mean resolution and finetuned. Models without finetuning are trained from scratch with pretrained MAE. 'HR subset' indicate validated images of 512×512. Related data scales are in brackets.

Exp.	Res.	F.T. from	Places2(train)		Places2(eval)		
			whole (1.8M)	subset (25,000)	whole (36,500)	subset (500)	HR subset (1,000)
MAE	256	–	✓		–	–	–
Main Exp.-1	256	–	✓		✓		
Main Exp.-2	512	Main Exp.-1	✓				✓
Ablations-1	256	–		✓		✓	
Ablations-2	512	Main Exp.-1	✓				✓

compared with three models: Co-Mod [14], LaMa [9] and ours, while Places2 is compared with four methods: Co-Mod, LaMa, CTSDG [4] and ours. We randomly shuffle and combine the outcomes of these algorithms except the masked inputs. After that, volunteers must choose the best one from each group. On both datasets, as shown in Fig. 1, our technique outperforms other competitors. Although Co-Mod also achieves competitive results in Places2, it is trained with extra 6.2 million images from Places365-Challenge, which is much larger than the training set of other competitors.

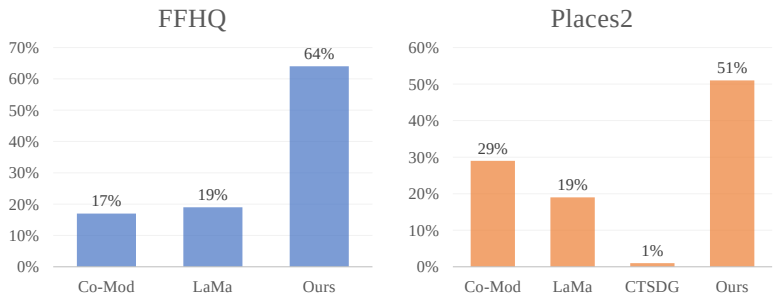


Fig. 1. Average scores of FFHQ and Places2 for user studies, which are collected from volunteers who select the best one from shuffled inpainted images.

6 Complete Quantitative Results

Tab. 2 presents more quantitative results for different masking rates ranging from 10% to 50% on datasets FFHQ and Places2. Except for the FID on the FFHQ, our model beats other state-of-the-art methods on other metrics, which demonstrates the superiority of our model.

Table 2. Quantitative results on FFHQ and Places2 with different mask ratios.

		FFHQ (256×256)			Places2 (256×256)				
	Mask	Co-Mod	LaMa	Ours	EC	Co-Mod	LaMa	CTSDG	Ours
PSNR↑	10~20%	28.45	29.84	**30.11**	26.61	26.40	28.23	26.73	**28.36**
	20~30%	26.04	27.52	**27.78**	24.26	23.61	25.31	24.37	**25.48**
	30~40%	24.29	25.82	**26.07**	22.60	21.73	23.44	22.71	**23.60**
	40~50%	22.93	24.48	**24.71**	21.28	20.28	22.03	21.41	**22.18**
	Mixed	25.25	26.60	**26.81**	23.31	22.57	24.37	23.43	**24.53**
SSIM↑	10~20%	0.938	0.950	**0.951**	0.913	0.926	**0.942**	0.913	**0.942**
	20~30%	0.909	0.924	**0.926**	0.872	0.880	0.901	0.872	**0.903**
	30~40%	0.876	0.897	**0.899**	0.828	0.831	0.859	0.828	**0.861**
	40~50%	0.843	0.869	**0.872**	0.783	0.781	0.814	0.782	**0.818**
	Mixed	0.889	0.903	**0.906**	0.839	0.843	0.869	0.835	**0.871**
FID↓	10~20%	**3.22**	3.60	3.42	1.95	0.52	0.45	2.44	**0.41**
	20~30%	**4.66**	5.20	4.94	3.79	1.00	0.95	5.62	**0.81**
	30~40%	**5.68**	6.57	6.14	6.98	1.65	1.73	11.43	**1.40**
	40~50%	**7.04**	8.69	8.12	11.50	2.38	2.82	19.88	**2.20**
	Mixed	**5.85**	6.38	6.15	6.21	1.49	1.63	11.18	**1.31**
LPIPS↓	10~20%	0.049	0.045	**0.043**	0.073	0.053	0.047	0.085	**0.042**
	20~30%	0.069	0.062	**0.059**	0.111	0.098	0.083	0.133	**0.073**
	30~40%	0.091	0.082	**0.077**	0.152	0.140	0.121	0.185	**0.106**
	40~50%	0.113	0.101	**0.095**	0.194	0.184	0.161	0.237	**0.141**
	Mixed	0.085	0.078	**0.074**	0.149	0.246	0.155	0.185	**0.101**

7 Quantitative Comparisons on DIV2K

We further give quantitative high-resolution results on 100 DIV2K [1] validation images with 2k resolutions in Tab. 3. Following the evaluation protocol of DIV2K in Tab. 3. Our model beats other HR inpainting methods.

Table 3. Quantitative results on DIV2K with mixed masks.

	PSNR↑	SSIM↑	FID↓	LPIPS↓
HiFill	20.67	0.787	135.53	0.241
LaMa	21.24	0.865	118.80	0.200
Ours	**21.64**	**0.868**	**113.98**	**0.171**

8 Experiments with Center Square Masks

Both quantitative and qualitative results of 512×512 Places2 test set with 40% center square masks are shown in Tab. 4 and Fig. 2 respectively. Note that our method is trained *without any rectangular mask*, while masks of Co-Mod include rectangular ones. Co-Mod suffers from hallucinated artifacts and LaMa tends to

generate blur results. Our model beat others with best PSNR and LPIPS even without training on rectangular masks, benefited by MAE priors.

Table 4. Quantitative results with 40% center square masks on 512×512 Places2.

	PSNR↑	SSIM↑	FID↓	LPIPS↓
Co-Mod	17.59	0.755	**52.38**	0.262
LaMa	19.69	0.801	61.67	0.268
Ours	**19.82**	**0.804**	53.61	**0.214**

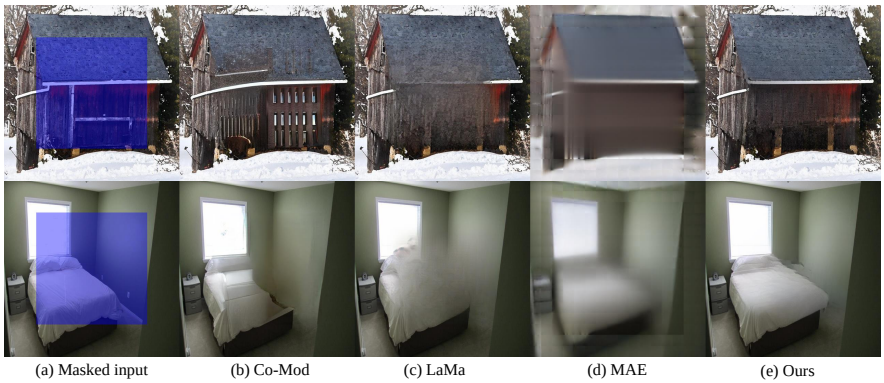


Fig. 2. Qualitative results with 40% center square masks on 512×512 Places2.

9 Ablations about Prior Attention from Different Layers

In Tab. 5, we test prior attentions of different layers from the start of the MAE decoder, and find that half-layer (4) just enjoys marginally better FID compared with all-layer (8) used in the main paper. These results show that using such attention priors from MAE is effective, while in general there is no significant difference in using attention from which layer.

10 Comparing with More SOTA Methods

We further compare our method with recently proposed ZITS [2] and MAT [7] on Places2 in Tab. 6. Our method can still outperform them with mixed masks.

11 More Qualitative Results

More 256×256 results of Places2 and FFHQ are shown in Fig. 3 and Fig. 4 respectively. For face images, we recommend to zoom-in for details near the eye

Table 5. Quantitative results of prior attention layers used from the start of MAE on the Places2 subset.

Attn layer	PSNR↑	SSIM↑	FID↓	LPIPS↓
-	24.34	0.860	26.84	0.117
2	24.50	0.863	25.61	**0.112**
4	24.51	0.862	**25.38**	0.113
6	**24.54**	0.863	25.67	0.114
8 (ours)	24.51	**0.864**	25.49	0.113

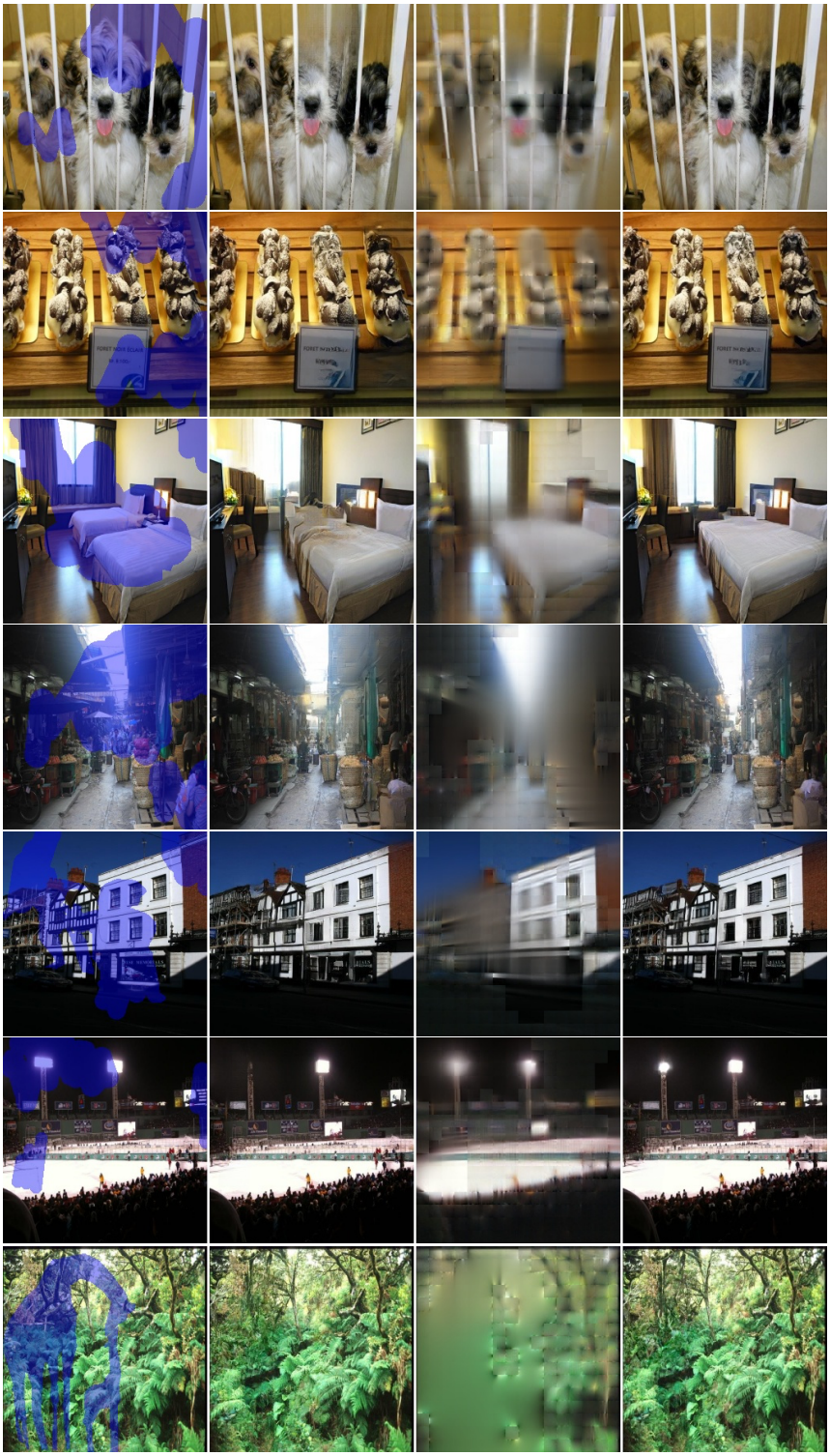
Table 6. Quantitative results compared with ZITS [2] and MAT [7] on Places2 with mixed masks.

	256×256				512×512			
	PSNR↑	SSIM↑	FID↓	LPIPS↓	PSNR↑	SSIM↑	FID↓	LPIPS↓
MAT	22.37	0.841	1.68	0.134	21.68	0.838	32.43	0.165
ZITS	24.42	0.870	1.47	0.108	24.23	**0.881**	26.08	0.133
Ours	**24.53**	**0.871**	**1.31**	**0.101**	**24.33**	0.880	**25.39**	**0.119**

regions. Our method tends to generate consistent eyes for face inpainting. We also provide more 512×512 results of Places2 in Fig. 5, and some 1024×1024 results from DIV2K [1] in Fig. 6. For the HR inpainting, we find an interesting phenomenon that the MAE enhanced results enjoy larger receptive fields for the structural recovery in HR cases as shown in the first row of Fig. 6. Besides, for a better reading experience, 1k results shown in the main paper are slightly compressed. We show the high quality ones in Fig. 7.

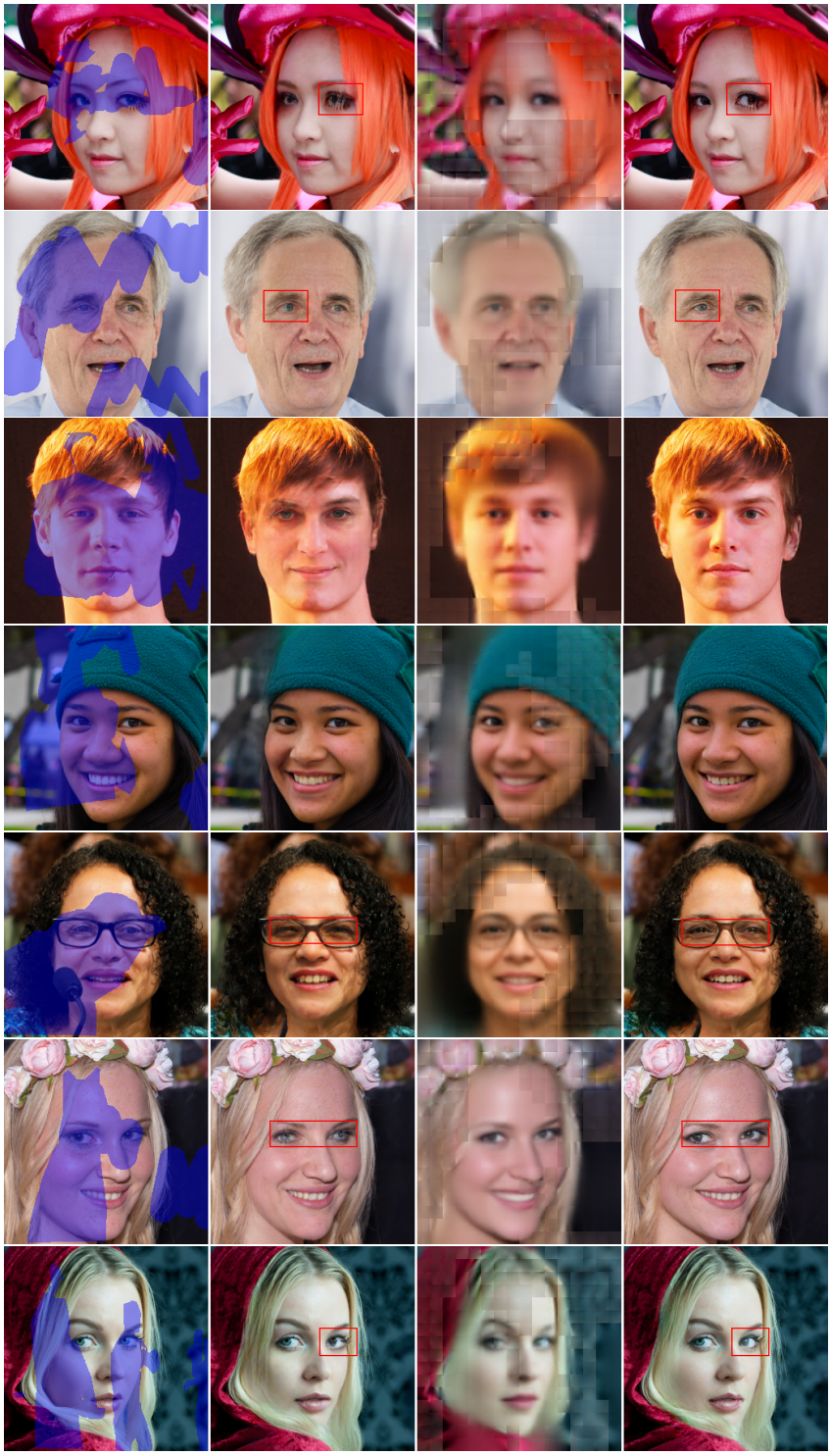
12 Limitations and Future Works

Although our proposed FAR is powerful enough to inpaint impressive results, it still suffers from fail cases as shown in Fig. 8. MAE has some difficulty in exactly recovering the object/building boundaries or some complex man-made structures, which leads to some ambiguity. To tackle this problem, we think that structure priors can provide more exact boundaries for high-fidelity results. Besides, as mentioned in the main paper, an interesting future work would be exploring features from different MAE layers for inpainting. In our opinion, such improvements are orthogonal to other proposed components in this paper. Our pre-trained Places2 MAE will be released, which is benefit for the community to further study the representation learning based image inpainting. Moreover, although our MAE pre-trained on Places2 is generalized enough for the inpainting, pre-training MAEs on larger datasets (such as ImageNet-22K [8] or even JFT-3B [13]) may achieve superior downstream performance.



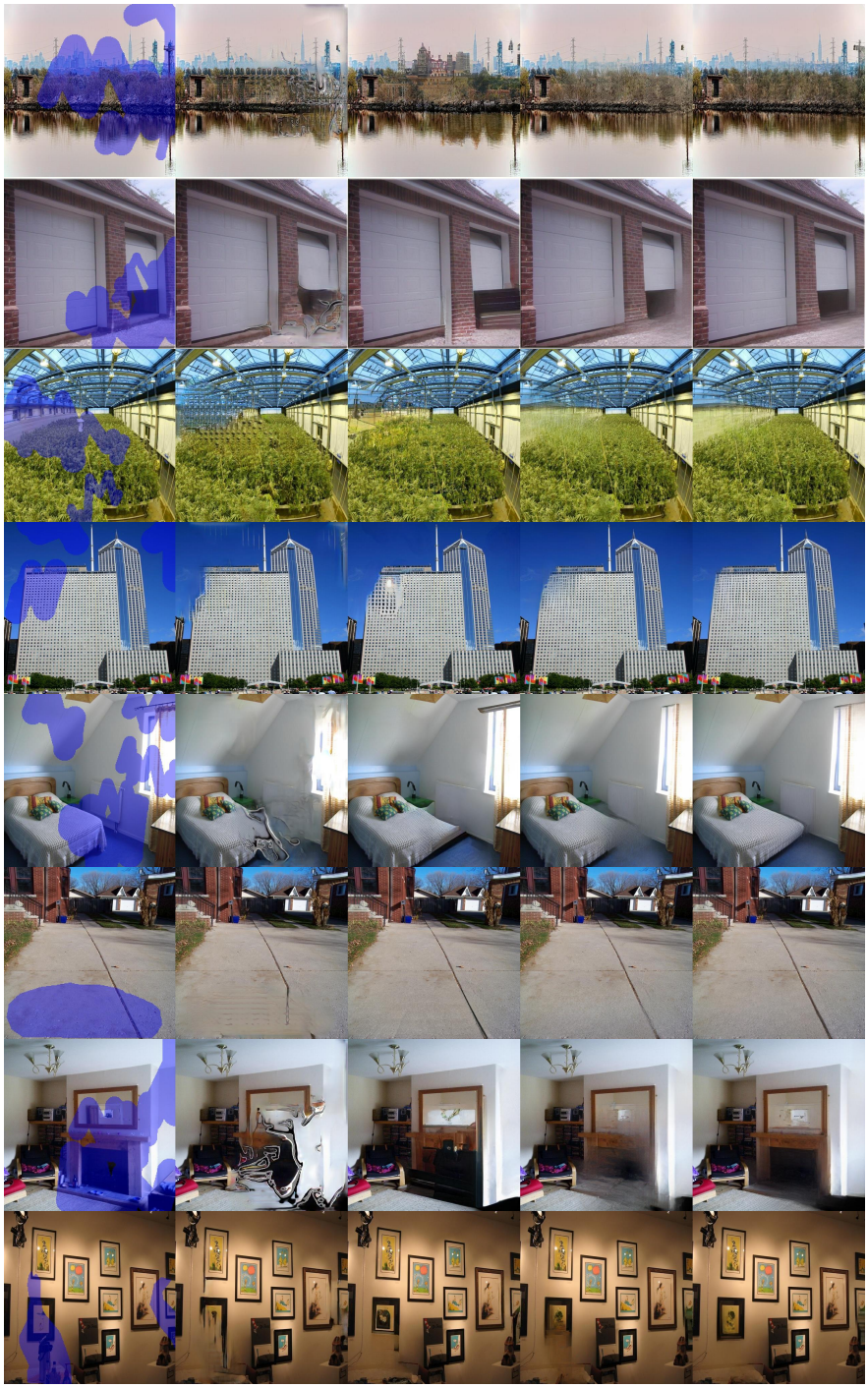
(a) Masked input (b) LaMa (c) MAE (d) Ours

Fig. 3. Qualitative results of places2 256×256 images. From left to right are masked inputs, LaMa [9], MAE [5], and our results. Please zoom-in for details.



(a) Masked input (b) LaMa (c) MAE (d) Ours

Fig. 4. Qualitative results of places2 256×256 images. From left to right are masked inputs, LaMa [9], MAE [5], and our results. Please zoom-in for details.



(a) Masked image (b) HiFill (c) Co-Mod (d) LaMa (e) Ours

Fig. 5. Qualitative results of 512×512 images from Places2. From left to right are masked image, HiFill [11], Co-Mod [14], LaMa [9], and our results. Please zoom-in for details.

(a) Masked input (b) MAE (c) LaMa (d) Ours

Fig. 6. Qualitative results of 1024×1024 images from DIV2K. From left to right are masked inputs, MAE [5], LaMa [9], and our results. Please zoom-in for details. For the first picture, both LaMa and our method fill all holes in the first row of the Colosseum, but our method still remains the holes in the second row. Because the MAE result is learned with a global receptive field, which guides our method to inpaint a more reasonable result rather than copying meaningless textures nearby.

(a) Masked input (b) MAE (c) LaMa (d) Ours

Fig. 7. Qualitative results of 1024×1024 images, which have also been shown in the main paper. From left to right are masked inputs, MAE [5], LaMa [9], and our results. Please zoom-in for details.

(a) GT (b) Masked input (c) MAE (d) Ours

Fig. 8. Failed cases of our method. GT indicates ground truth images

References

1. Agustsson, E., Timofte, R.: Ntire 2017 challenge on single image super-resolution: Dataset and study. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (July 2017)
2. Dong, Q., Cao, C., Fu, Y.: Incremental transformer structure enhanced image inpainting with masking positional encoding (2022)
3. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.: Improved training of wasserstein gans. arXiv preprint arXiv:1704.00028 (2017)
4. Guo, X., Yang, H., Huang, D.: Image inpainting via conditional texture and structure dual generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 14134–14143 (October 2021)
5. He, K., Chen, X., Xie, S., Li, Y., Doll'ar, P., Girshick, R.B.: Masked autoencoders are scalable vision learners. ArXiv **abs/2111.06377** (2021)
6. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1125–1134 (2017)
7. Li, W., Lin, Z., Zhou, K., Qi, L., Wang, Y., Jia, J.: Mat: Mask-aware transformer for large hole image inpainting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10758–10768 (2022)
8. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV) **115**(3), 211–252 (2015). https://doi.org/10.1007/s11263-015-0816-y
9. Suvorov, R., Logacheva, E., Mashikhin, A., Remizova, A., Ashukha, A., Silvestrov, A., Kong, N., Goka, H., Park, K., Lempitsky, V.: Resolution-robust large mask inpainting with fourier convolutions. arXiv preprint arXiv:2109.07161 (2021)
10. Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional gans. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8798–8807 (2018)
11. Yi, Z., Tang, Q., Azizi, S., Jang, D., Xu, Z.: Contextual residual aggregation for ultra high-resolution image inpainting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7508–7517 (2020)
12. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Free-form image inpainting with gated convolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4471–4480 (2019)
13. Zhai, X., Kolesnikov, A., Houlsby, N., Beyer, L.: Scaling vision transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12104–12113 (2022)
14. Zhao, S., Cui, J., Sheng, Y., Dong, Y., Liang, X., Chang, E.I., Xu, Y.: Large scale image completion via co-modulated generative adversarial networks. arXiv preprint arXiv:2103.10428 (2021)