Supplemental Materials on
"Continuously Masked Transformer for Image Inpainting"

S-1. Implementation details

The number of channels is set to 768 in CMT. Each MSAU block contains two MLP layers, as shown Figure S-1(a). It has the same structure as ViT [2], but the attention layer is replaced by the proposed masked attention. In each MLP layer, we also update the mask using the error propagator ϕ, as done in (2) and (8). On the other hand, Figure S-1(b) shows the detailed structure of the refinement network where we set $C = 32$. Here, Swin blocks are labeled as '(n, k)', where n is the number of blocks and k is the window size. The number of tokens decreases by half and increases by a factor of two through the patch merging layer [7] and the up-sampling layer, respectively. The up-sampling layer consists of one convolutional layer followed by bilinear interpolation. We use the Adam optimizer [4] with a learning rate of 1×10^{-4}. The proposed algorithm is trained with mask patterns generated by the free-form mask generator in [9] and resized images from the Places2 and CelebA-HQ datasets.

The running times on a 256×256 image are 0.018s and 0.027s for coarse and refinement networks, respectively, and the number of parameters are 73M and 70M.

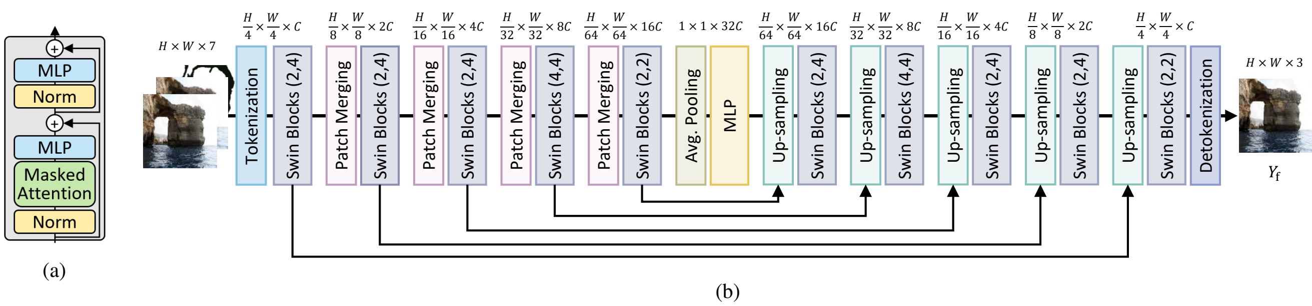


Figure S-1: The structures of (a) the MSAU block and (b) the refinement network.

S-2. Comparison on high-resolution images

We compare the proposed algorithm with HiFill [8] and MAT [5] on 512×512 images. We randomly select 12,000 test images from 36,500 validation images in Places2 [10] and use the six mask sets in the irregular mask dataset [6]. Again, the proposed CMT performs the best in all tests for all H2I ratio ranges with no exception.

Table S-1: Quantitative comparison on 512×512 images from the Places2 dataset [10] according to the hole-to-image (H2I) area ratios.

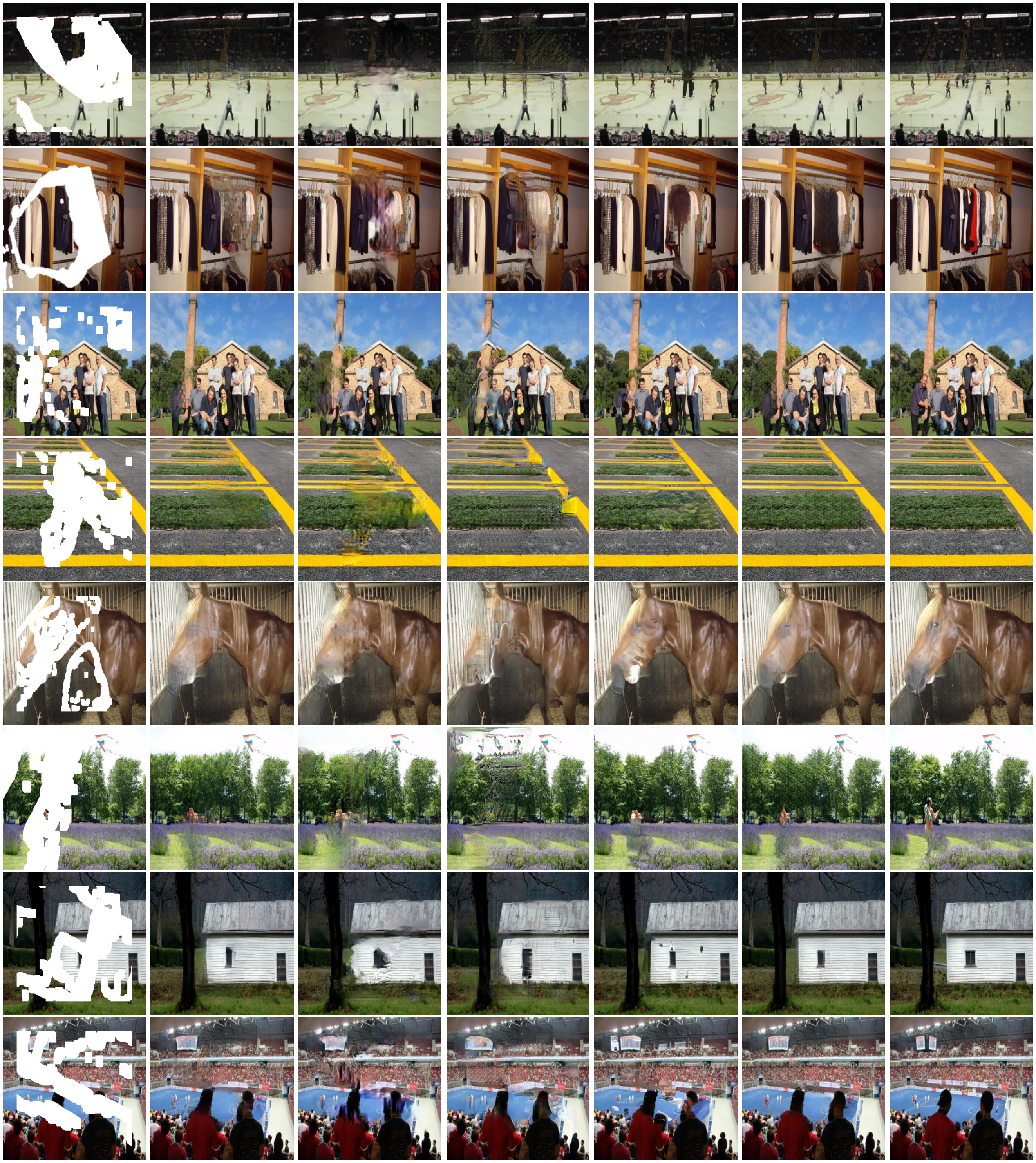
	H2I $\in (0.01, 0.1]$			H2I $\in (0.1, 0.2]$			H2I $\in (0.2, 0.3]$		
	PSNR(↑)	SSIM(↑)	FID(↓)	PSNR(↑)	SSIM(↑)	FID(↓)	PSNR(↑)	SSIM(↑)	FID(↓)
HiFill [8]	29.56	0.9656	7.20	24.31	0.9170	16.72	21.54	0.8624	28.33
MAT [5]	34.05	0.9838	2.59	27.53	0.9544	6.74	24.01	0.9161	11.77
CMT (Proposed)	**34.80**	**0.9844**	**2.55**	**28.63**	**0.9568**	**6.64**	**25.29**	**0.9211**	**11.69**
	H2I $\in (0.3, 0.4]$			H2I $\in (0.4, 0.5]$			H2I $\in (0.5, 0.6]$		
	PSNR(↑)	SSIM(↑)	FID(↓)	PSNR(↑)	SSIM(↑)	FID(↓)	PSNR(↑)	SSIM(↑)	FID(↓)
HiFill [8]	19.68	0.8076	42.23	17.95	0.7422	64.20	15.85	0.6565	98.68
MAT [5]	21.68	0.8746	16.31	19.80	0.8271	21.29	17.16	0.7547	29.46
CMT (Proposed)	**23.14**	**0.8832**	**15.79**	**21.39**	**0.8406**	**20.88**	**19.01**	**0.7775**	**29.19**

S-3. More qualitative comparisons

Figures S-2 to S-6 compare qualitative results of the proposed CMT algorithm with those of conventional algorithms.

(a) Input (b) RN (c) ICT (d) BAT (e) MAT (f) CMT (Proposed) (g) Ground-Truth

Figure S-2: Qualitative comparison of inpainted images on the Places2 dataset [10].



(a) Input (b) Edge-Connect (c) MEDFE (d) HiFill (e) MAT (f) CMT (Proposed) (g) Ground-Truth

Figure S-3: Qualitative comparison of inpainted images on the Places2 dataset [10].

Figure S-4: Qualitative comparison of inpainted images on the CelebA-HQ dataset [3].

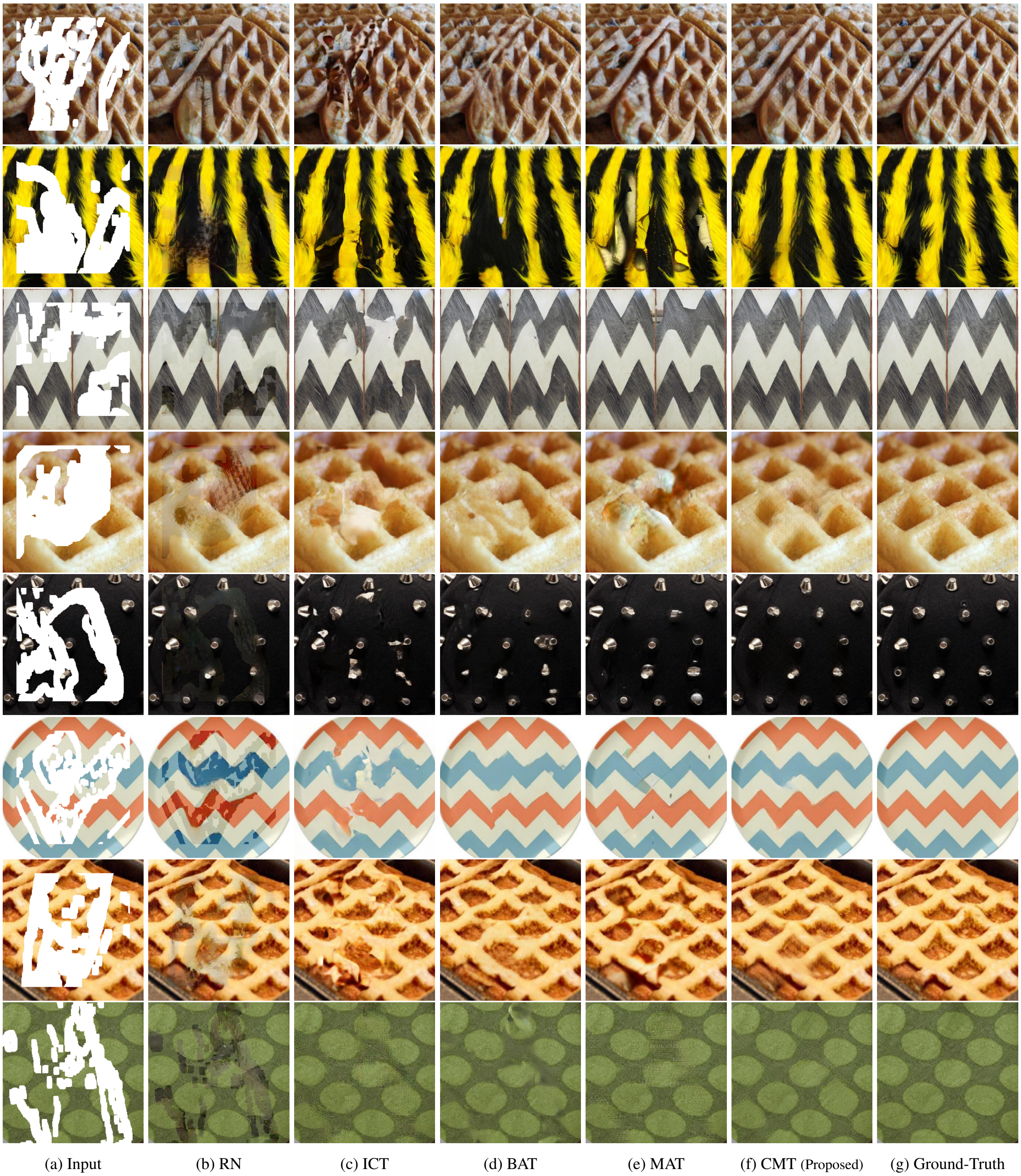


Figure S-5: Qualitative comparison of inpainted images on the DTD dataset [1].

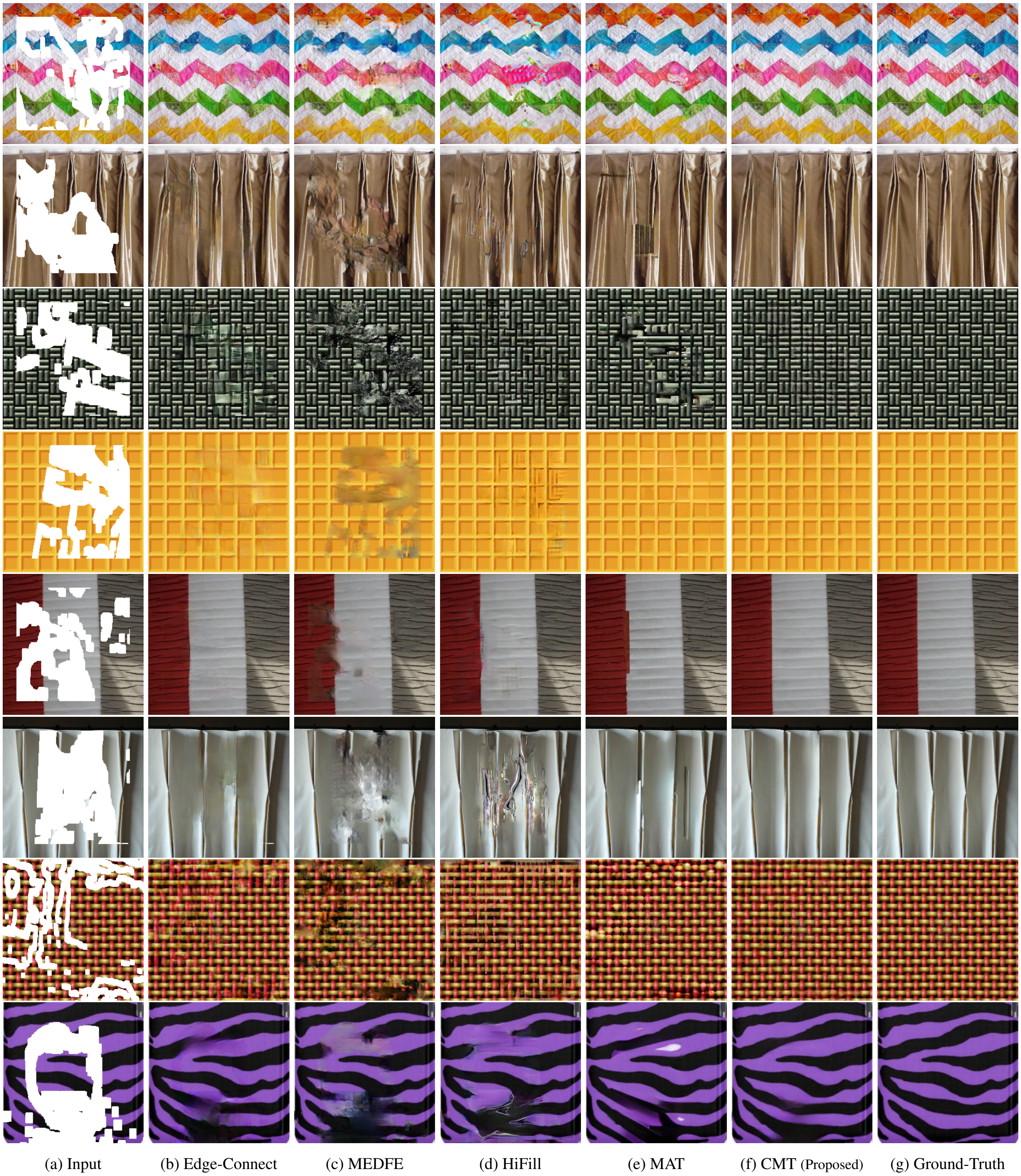


Figure S-6: Qualitative comparison of inpainted images on the DTD dataset [1].

References

[1] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, 2014. S-5, S-6

[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. S-1

[3] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *ICLR*, 2017. S-4

[4] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. S-1

[5] Wenbo Li, Zhe Lin, Kun Zhou, Lu Qi, Yi Wang, and Jiaya Jia. MAT: Mask-aware transformer for large hole image inpainting. In *CVPR*, 2022. S-1

[6] 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 *ECCV*, 2018. S-1

[7] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. S-1

[8] Zili Yi, Qiang Tang, Shekoofeh Azizi, Daesik Jang, and Zhan Xu. Contextual residual aggregation for ultra high-resolution image inpainting. In *CVPR*, 2020. S-1

[9] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In *ICCV*, 2019. S-1

[10] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(6):1452–1464, 2017. S-1, S-2, S-3