Method	#Param. $\times 10^6$	Places (512 × 512)						CelebA-HQ (512 × 512)					
		Small Mask			Large Mask			Small Mask			Large Mask		
		FID↓	P-IDS(%)↑	U-IDS(%)↑	FID↓	P-IDS(%)↑	U-IDS(%)↑	FID↓	P-IDS(%)↑	U-IDS(%)↑	FID↓	P-IDS(%)↑	U-IDS(%)↑
MAT (Ours)†	62	0.78	31.72	43.71	1.96	23.42	38.34	2.86	21.15	32.56	4.86	13.83	25.33
MAT (Ours)		1.07	27.42	41.93	2.90	19.03	35.36						
CoModGAN [75]†	109	1.10	26.95	41.88	2.92	19.64	35.78	3.26	19.65	31.41	5.65	11.23	22.54
LaMa [51]†	51/27	0.99	22.79	40.58	2.97	13.09	32.29	4.05	9.72	21.57	8.15	2.07	7.58
ICT [55]	150	-	-	-	-	-	-	6.28	2.24	9.99	12.84	0.13	0.58
MADF [79]	85	2.24	14.85	35.03	7.53	6.00	23.78	3.39	12.06	24.61	6.83	3.41	11.26
AOT GAN [70]	15	3.19	8.07	30.94	10.64	3.07	19.92	4.65	7.92	20.45	10.82	1.94	6.97
HFill [65]	3	7.94	3.98	23.60	28.92	1.24	11.24	-	-	-	-	-	-
DeepFill v2 [67]	4	3.02	9.17	32.56	9.27	4.01	21.32	10.11	3.11	9.52	24.42	0.17	0.42
EdgeConnect [40]	22	4.03	5.88	27.56	12.66	1.93	15.87	10.58	4.14	12.45	39.99	0.10	0.22

Table 2. Quantitative comparison on Places [78] and CelebA-HQ [25]. "†": Our Mat, CoModGAN [75] and LaMa [51] use 8M, 8M and 4.5M training images on Places, respectively, while our other model (without "†") is only trained on a subset (1.8M images). The LaMa models on Places and CelebA are different in size. The results of LPIPS and 256 × 256 CelebA are provided in the supplementary. The best and second best results are in red and blue.

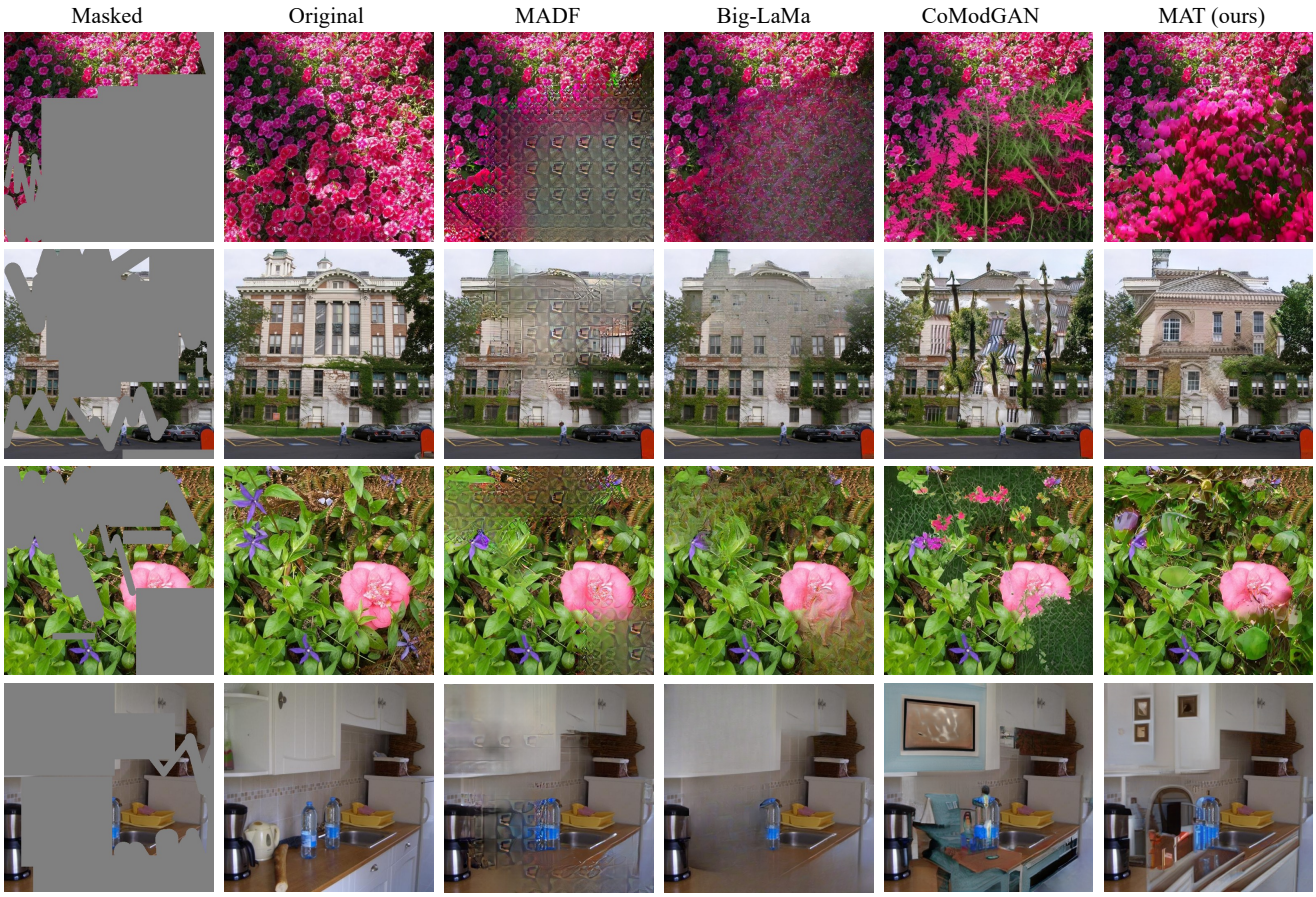

Figure 8. Qualitative comparison (512×512) with state-of-the-art methods. Our results are more visually realistic, containing more details.

defined window sizes in attention, we need to pad or resize an image to make its size a multiple of 512.

5. Conclusion

We have presented a mask-aware transformer (MAT) for pluralistic large hole image inpainting. Taking advantage of