MTADiffusion: Mask Text Alignment Diffusion Model for Object Inpainting

Supplementary Material

1. Ablation studies for the edge loss and style loss

To evaluate the influence of different training losses of our method, we first sampled 1 million subsets from our dataset for fast convergence. The model was then trained for 100,000 iterations using data from the subset with different loss functions. Table A1 shows the quantitative results on BrushBench. As we can see, adding edge prediction loss and style-consistency loss brings about improvements in image quality and masked region preservation, especially in the metrics of IR [39] and PSNR.

Figure A1 shows the qualitative influence of different losses. It is obvious that the edge loss enhanced structural stability, while the style loss improved style consistency. Combining noise loss, style loss, and edge loss yielded the best results.

Table A1. Comparison of different loss functions. NL denotes Noise Loss, SL denotes Style Loss, and EL denotes Edge Loss. For clarity, IR is scaled by 10, and LPIPS and MSE by 1000.

Loss	IR ↑	AS ↑	PSNR ↑	LPIPS ↓	MSE ↓	CLIP Sim ↑
NL	12.52	6.38	31.52	19.23	0.83	26.41
NL+EL	12.57	6.39	31.65	19.19	0.83	26.48
NL+EL+SL	12.63	6.39	31.72	19.09	0.82	26.44

Figure A1. Influence of different loss functions.

2. Qualitative results on BrushBench and EditBench

Figures A2 and A3 show the visual results of our method in BrushBench [9] and EditBench [34], respectively. We compared our method with SDI [25], CNI [42], PowerPaint [45], and BrushNet [9]. It can be seen that our results show superiority in semantic alignment, structure stability and style consistency, which resulted in reasonable and natural images.

Figure A2. Qualitative Results on BrushBench.

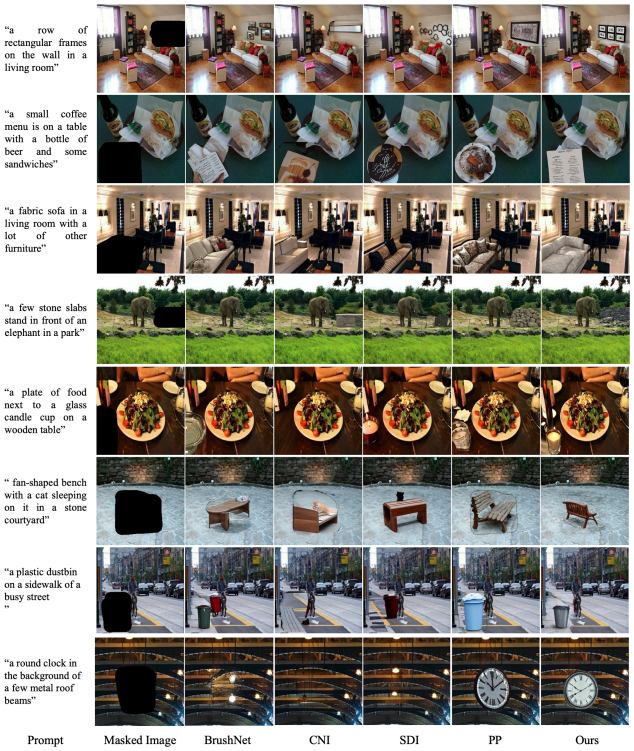

Figure A3. Qualitative Results on EditBench.