RePaint: Inpainting using Denoising Diffusion Probabilistic Models Supplementary Material

Andreas Lugmayr Martin Danelljan Andres Romero Fisher Yu Radu Timofte Luc Van Gool

Computer Vision Lab

ETH Zürich, Switzerland

Appendix

In this appendix, we provide additional details and analysis of our approach. We give more explanation on our user study in Section 1. Further, we present additional details on how we implemented the diffusion time schedule for jumps in Section 2. Visual results for our ablation for jump size and the number of resamplings are provided in Section 3. The evaluation on the second part of the LaMa Benchmark on Places2 is presented in Section 4. Furthermore, to compare the diversity of the inpaintings for RePaint compared with state-of-the-art, we provide a quantitative analysis in Section 5. Details on failure cases and data bias on the ImageNet dataset are provided in Section 6. For gaining a better intuitive understanding of the evolution of the latent space, we provide a video of the inference in Section 7. And finally, we show additional visual examples in Section 9.

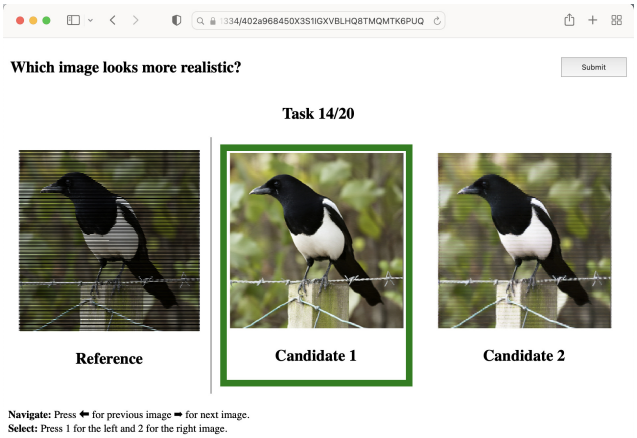


Figure 1. **User Study Interface.** Example of the user-study interface. Based on the reference image on the Left, the user selects the image that looks more realistic.

```
t_T = 250
jump_len = 10
jump_n_sample = 10

jumps = {}
for j in range(0, t_T - jump_len, jump_len):
    jumps[j] = jump_n_sample - 1

t = t_T
ts = []

while t >= 1:
    t = t-1
    ts.append(t)

    if jumps.get(t, 0) > 0:
        jumps[t] = jumps[t] - 1
        for _ in range(jump_len):
            t = t + 1
            ts.append(t)

ts.append(-1)
```

Figure 2. **Diffusion Time Schedule.** Pseudo code to generate diffusion time steps for jump length $j = 10$ and resample $r = 10$.

1. User Study

As described in Section 5.2 in the main paper, we conduct a user study to determine which method is best perceived to the human eye. In Figure 1, we depict the user interface, where the user selects the most realistic solution from an input reference. To reduce bias, we show the two candidate images in random order. Additionally, to improve the consistency of the user decision and prevent answers with low effort, we show every example twice. The users that agree in less than 75% of their own votes are discarded.

2. Algorithm for jump size larger than one

In addition to the resampling introduced in Algorithm 1 in the main paper, we use jumps in diffusion time as de-

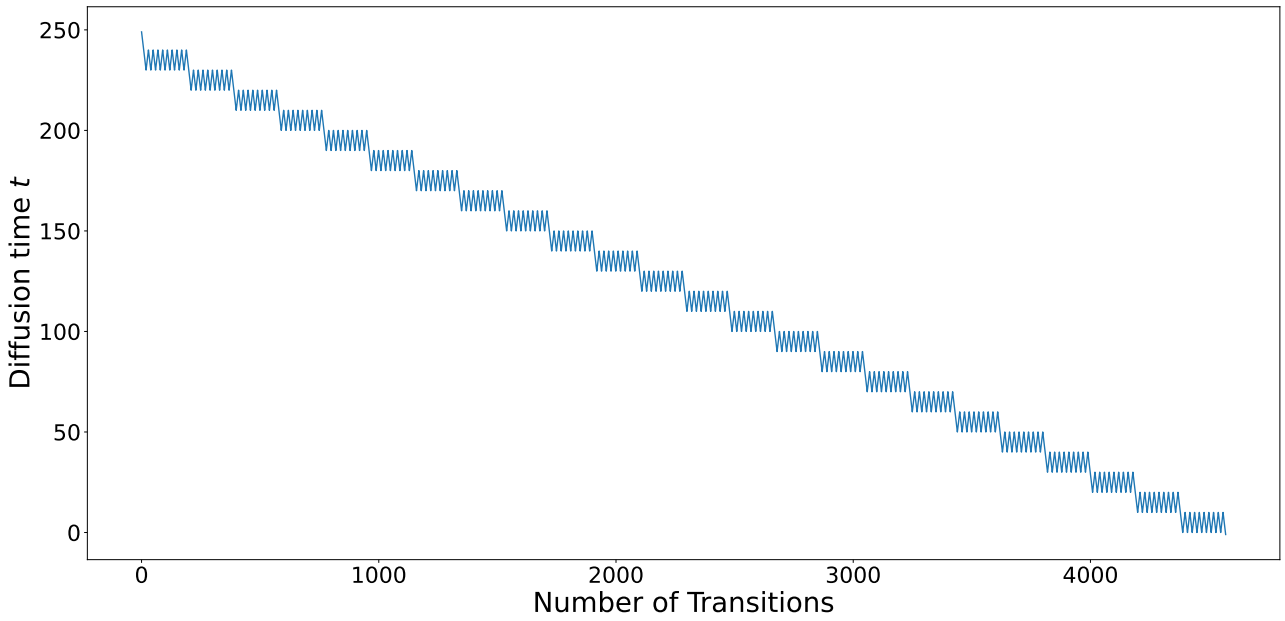


Figure 3. **Diffusion time during inference.** The diffusion time t that a sample x_t is transiting during the inference process with jump length $j = 10$ and resampling $r = 10$.

scribed in Section 4.2 in the main paper. Figure 2 shows a pseudo-code to further clarify the generation of state transitions. Note that each transition increases or decreases the diffusion time t by one. For example, for a chosen jump length of $j = 10$ shown in Figure 4, we apply ten forward transitions before applying ten reverse transitions. The diffusion time t for the latent vector x_t is plotted in Figure 3.

3. Ablation

In addition to the quantitative analysis in Table 3 in the main paper, this section shows visual examples for different jump lengths j and number or resamplings r. As discussed in Section 5.5 in the main paper, smaller jump lengths j tend to produce blurrier images as shown in Figure 5, and an increased number or resamplings r improves the overall image consistency.

```
times = get_schedule()
x = random_noise()

for t_last, t_cur in zip(times[:-1], times[1:]):
    if t_cur < t_last:
        # Apply Equation 8 (Main Paper)
        x = reverse_diffusion(x, t, x_known)
    else:
        # Apply Equation 1 (Main Paper)
        x = forward_diffusion(x, t)
```

Figure 4. **Inference Process.** Pseudo code of RePaint inference process using a precalculated time schedule.

4. Evaluation on Places2

For a more comprehensive experimental framework, in this section, we provide the second part of the benchmark proposed in LaMa [4], which is over the Places2 [8] dataset. The experiments on Places2 were conducted using an unconditional model that we trained for 300k iterations with batch size four on four V100, taking about six days in total. All other training settings were kept as originally [1] used for ImageNet. The model checkpoint will be published. We will clarify these aspects and add further details in the paper. We use the same mask generation procedure and settings described in the main paper. The results shown in Table 1 are in line with those on CelebA and ImageNet in Table 1 of the main paper. RePaint outperforms all other methods for all masks with significance 95% except for one inconclusive case. This case is when comparing RePaint to LaMa on Wide Masks, where the users vote in 52.4% for RePaint, but the significance interval overlaps with the 50% border. The visual comparison on the and Wide and Narrow mask is shown in Figure 15. Moreover, the visual results further confirm the robustness against sparse masks as shown in Figure 16. The mask pattern is clearly visible in all competing methods, while RePaint shows better harmonization. Regarding large masks, RePaint is able to inpaint semantically meaningful content such as the companion in the Bar

Datasets	Wide		Narrow		Super-Resolve 2×		Altern. Lines		Half		Expand	
Methods	LPIPS	Votes [%]	LPIPS	Votes [%]	LPIPS	Votes [%]	LPIPS	Votes [%]	LPIPS	Votes [%]	LPIPS	Votes [%]
AOT [7]	0.112	35.4 ± 3.0	0.062	36.0 ± 3.0	0.560	2.2 ± 0.9	0.399	0.8 ± 0.6	0.263	34.0 ± 2.9	0.686	0.7 ± 0.5
DSI [3]	0.101	27.4 ± 2.8	0.054	33.1 ± 2.9	0.157	8.4 ± 1.7	0.083	6.9 ± 1.6	0.265	33.7 ± 2.9	0.565	13.8 ± 2.1
ICT [5]	0.101	35.7 ± 3.0	0.057	33.7 ± 2.9	0.776	0.9 ± 0.6	0.672	1.3 ± 0.7	0.256	26.0 ± 2.7	0.554	26.6 ± 2.7
Deep Fill v2 [6]	0.097	29.7 ± 2.8	0.051	33.0 ± 2.9	0.120	15.8 ± 2.3	0.070	15.4 ± 2.2	0.254	32.8 ± 2.9	0.550	12.9 ± 2.1
LaMa [4]	0.078	47.7 ± 3.1	0.039	43.3 ± 3.1	0.369	7.5 ± 1.6	0.138	21.5 ± 2.6	0.233	34.0 ± 2.9	0.512	39.4 ± 3.0
RePaint	0.105	*Reference*	0.044	*Reference*	0.099	*Reference*	0.051	*Reference*	0.286	*Reference*	0.615	*Reference*

Table 1. **Places2 Quantitative Results.** We compute the LPIPS (lower is better) and *votes* for five different mask settings. *Votes* refers to the ratio of votes in favor our RePaint.

Mask	Wide		Narrow		SR 2x		Alter. Lines		Half		Expand	
Measure	LPIPS	DS	LPIPS	DS	LPIPS	DS	LPIPS	DS	LPIPS	DS	LPIPS	DS
DSI [3]	0.0639	16.68	0.0454	18.74	0.1404	12.38	0.0591	4.78	0.2348	15.30	0.5458	14.33
ICT [5]	0.0596	15.77	0.0402	18.65	0.5427	8.70	0.3916	8.16	0.1817	16.40	0.4779	17.25
RePaint	0.0552	16.40	0.0337	23.79	0.0327	19.84	0.0106	23.00	0.1839	17.31	0.4832	17.11

Table 2. **Diversity Score.** The Diversity Score (DS) and LPIPS calculated on CelebA-HQ on various masks for 32 images.

in the same age, and overall lightning conditions as shown in the second row of Figure 17.

5. Diversity

For our quantitative evaluation in the main paper, we sample a single image per input. However, since our method is stochastic, we can sample from it. To compare the diversity among the stochastic methods, we use the Diversity Score as described in [2] (higher is better). In contrast to the standard diversity metric [3, 5] that only computes the mean LPIPS across pair of outputs, this score is designed to describe meaningful diversity yet also weighting the overall performance in LPIPS. It aims at measuring the diversity of the generations inside the manifold of plausible predictions. In detail, too extreme predictions or failures are therefore penalized. As shown in Table 2, for "Wide" and "Half", there is no method with both best LPIPS and Diversity Score and for "Expand" ICT beats RePaint by 0.81% in Diversity Score and 1.1% in LPIPS. RePaint is superior by a large margin in both LPIPS and Diversity Score for the thin structured masks "Narrow", "Super-Resolution $2\times$", and "Alternating Lines" to both ICT [5] and DSI [3].

6. Failure Cases

As depicted in Figure 6, RePaint sometimes confuses the semantic context and mixes non-matching objects. Our model on ImageNet seems to be biased towards inpainting dogs more frequently than expected. Since ImageNet has many different breeds of dogs for classification tasks, dogs are over-represented in the training set, hence our model bias.

7. Attached Video

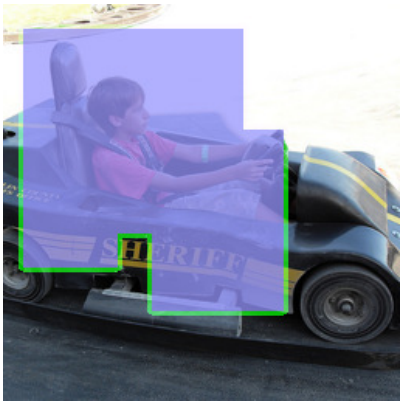
To inspect the latent space of the diffusion space, we provide a video in the attachment as shown in the screenshot in Figure 7. There we show the Ground Truth and the latent space x_t after every transition in the diffusion process. Note that the diffusion time t, shown on top, jumps up and down according to the following schedule: The jump length is $j = 5$, and the number of resamplings is $r = 9$. To focus more on the visually interesting part of the diffusion process we set the number of diffusion steps to $T = 100$ and start resampling below $t = 50$.

8. Experiment on larger resolution

As shown in Figure 8, our inpainting method also works on pretrained model from [1] for 512×512. However, we were not able to conduct our full analysis on that resolution due to limited computational resources.

9. Additional Visual Results

We also provide additional visual examples for CelebA-HQ and ImageNet, comparing our approach to the same state-of-the-art methods as in the main paper. We show the results for Wide and Narrow masks in Figures 9 and 12, respectively, for the sparse masks "Super-Resolve $2\times$" and "Alternating Lines" in Figures 10 and 13 and for "Half" and "Expand" in Figures 11 and 14.

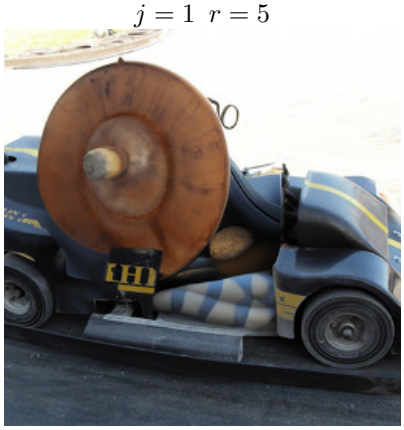


Input

$j = 1 \;\; r = 5$

$j = 1 \;\; r = 10$

$j = 1 \;\; r = 15$



$j = 5 \;\; r = 5$

$j = 5 \;\; r = 10$

$j = 5 \;\; r = 15$

$j = 10 \;\; r = 5$

$j = 10 \;\; r = 10$

$j = 10 \;\; r = 15$

Figure 5. **Ablation Study.** Analysis of length of the jumps j and number of resamplings r on ImageNet validation set with LaMa [4] Benchmark mask setting Wide.

Figure 6. **Failure Cases on ImageNet.** When applying RePaint trained on ImageNet for inpainting it is more likely to inpaint dogs, due to the data bias. Zoom-in for better details.

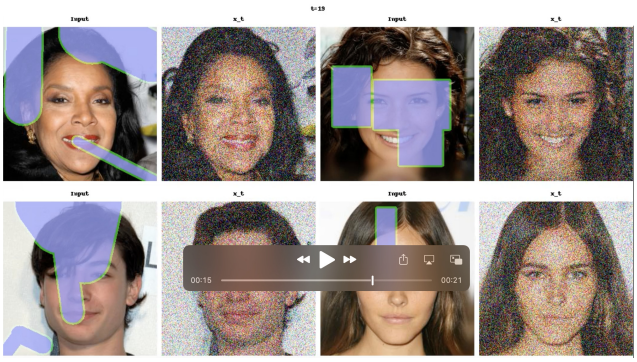


Figure 7. **Video of Diffusion Process.** In the attachment we show the video of the denoising diffusion process on the CelebA-HQ validation set.

Figure 8. Visual results on ImageNet 512×512 for thin mask.

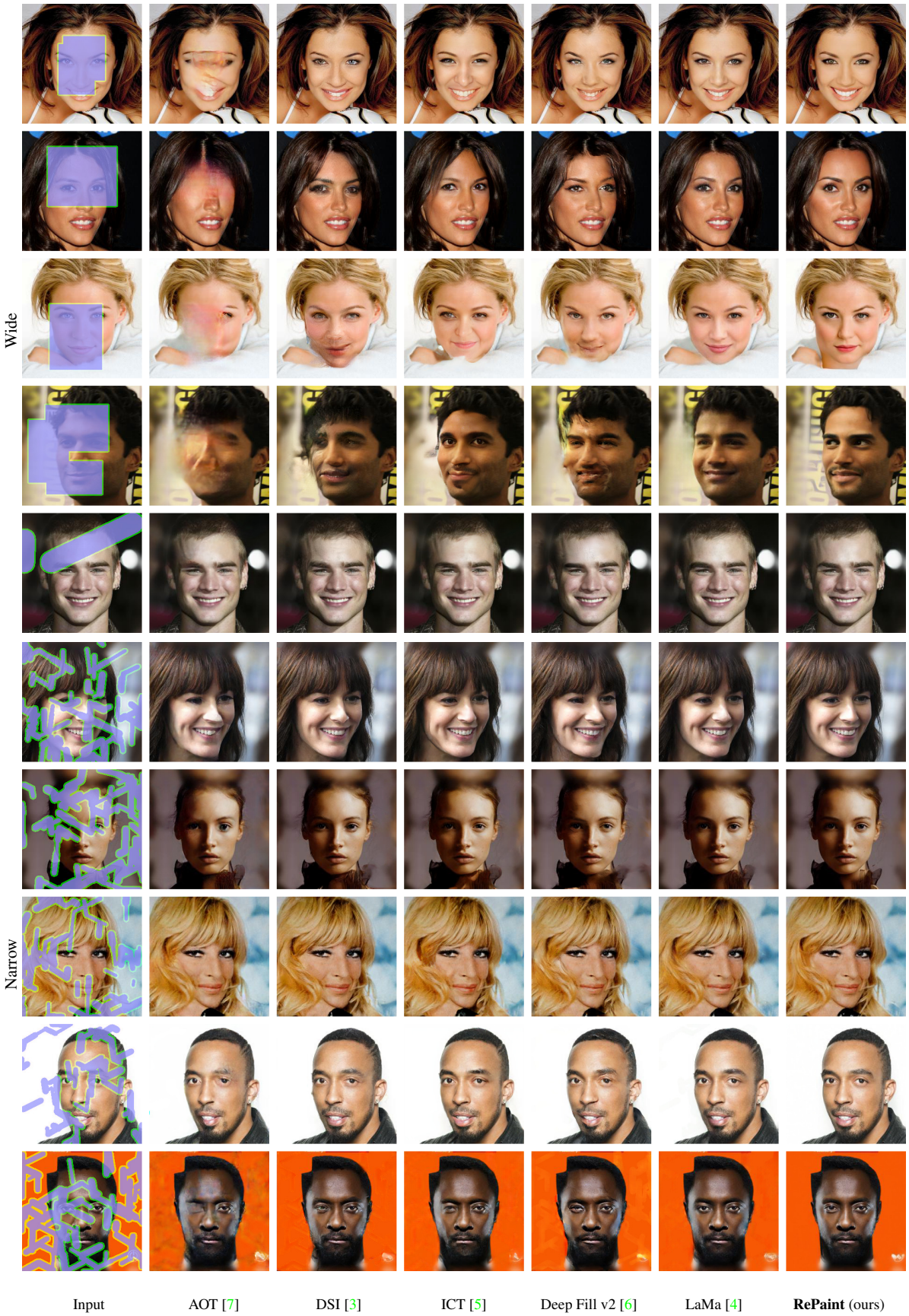


Figure 9. **CelebA-HQ Qualitative Results.** Comparison against the state-of-the-art methods for face inpainting. Zoom for better details.

Figure 10. **CelebA-HQ Qualitative Results.** Comparison against the state-of-the-art methods for face inpainting. Zoom for better details.

Figure 11. **CelebA-HQ Qualitative Results.** Comparison against the state-of-the-art methods for face inpainting. Zoom for better details.

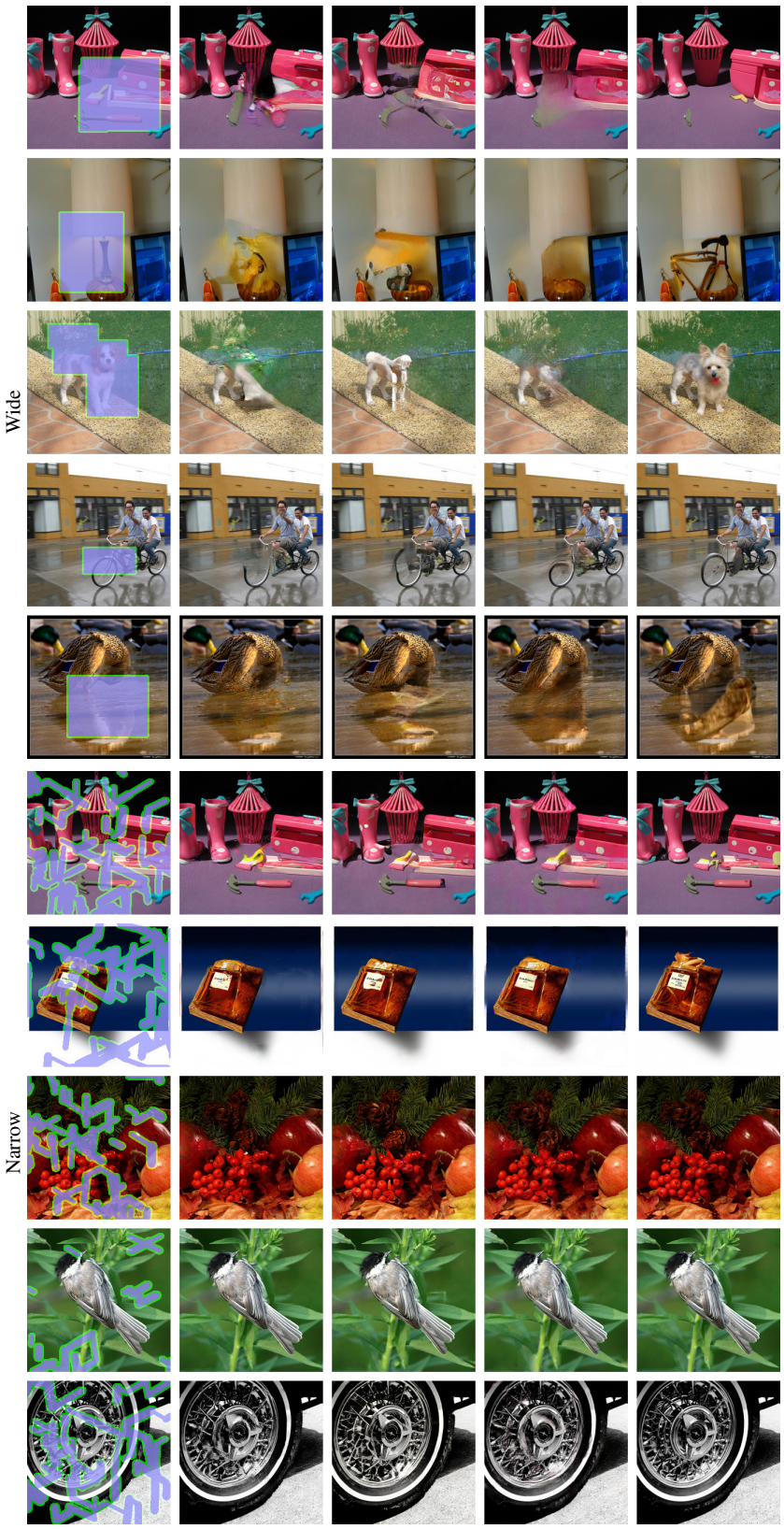


Figure 12. **ImageNet Qualitative Results.** Comparison against the state-of-the-art methods for diverse inpainting. Zoom for better details.

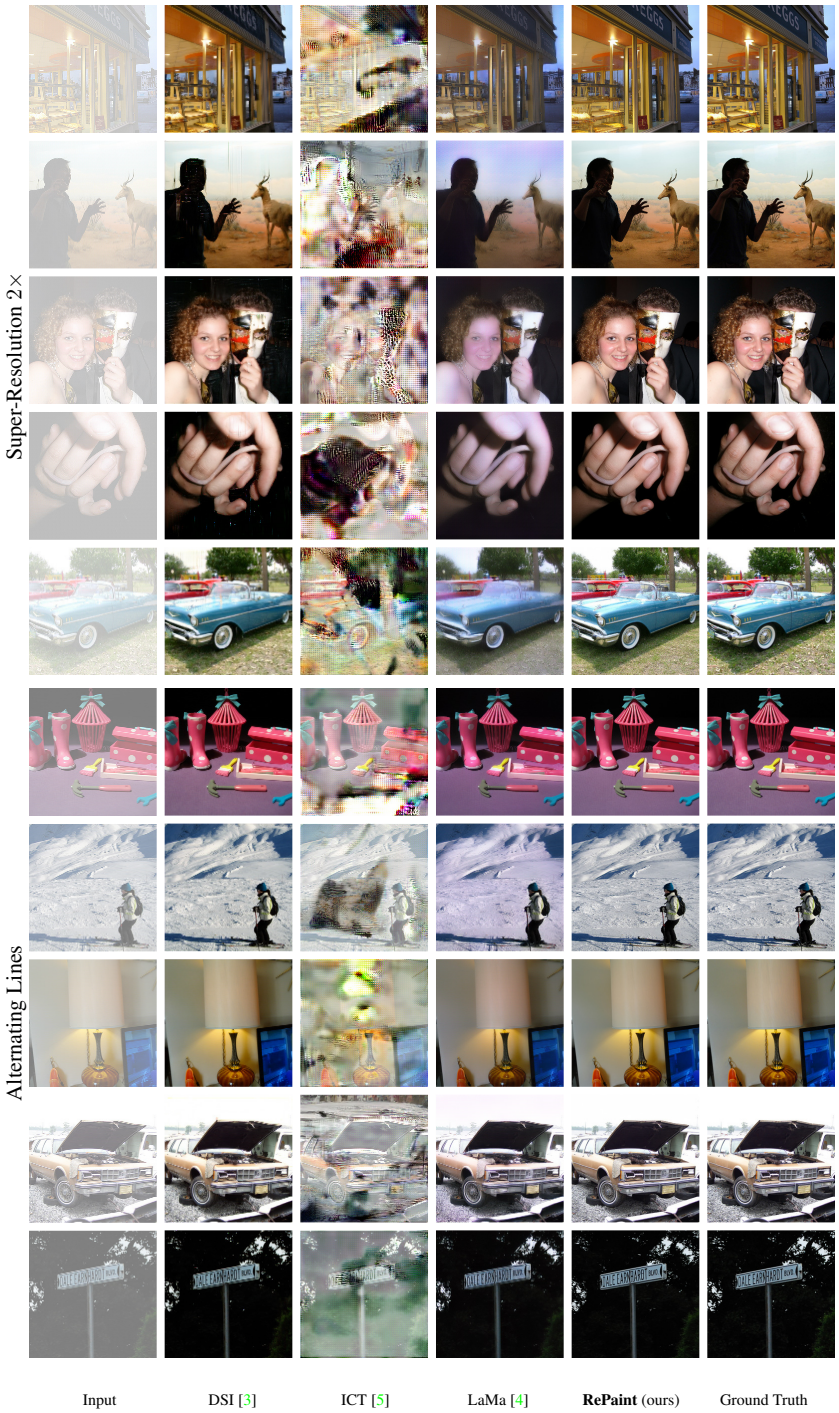


Figure 13. **ImageNet Qualitative Results.** Comparison against the state-of-the-art methods for diverse inpainting. Zoom for better details.

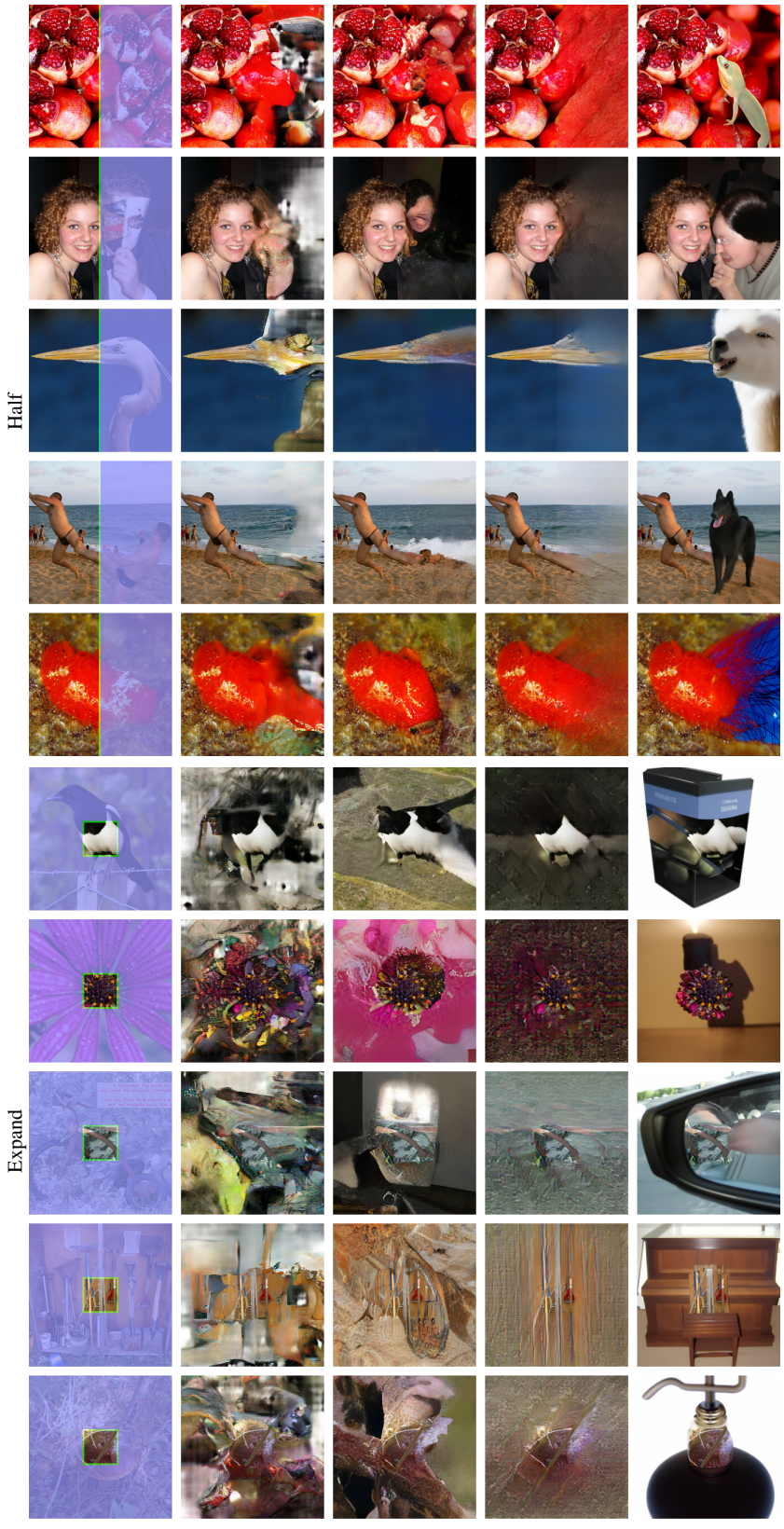


Figure 14. **ImageNet Qualitative Results.** Comparison against the state-of-the-art methods for diverse inpainting. Zoom for better details.

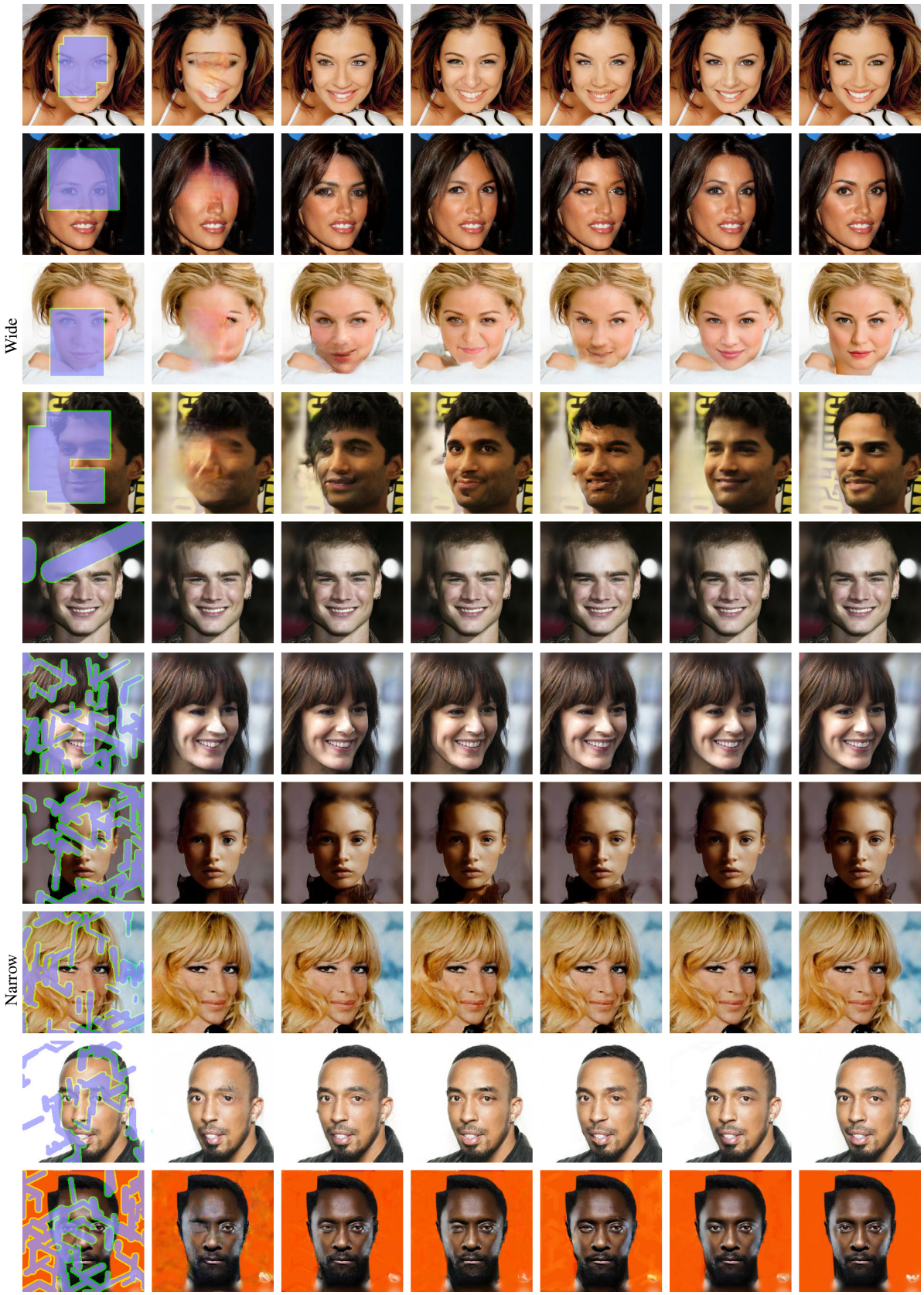


Figure 15. **Places2 Qualitative Results.** Comparison against the state-of-the-art methods for diverse inpainting. Zoom for better details.

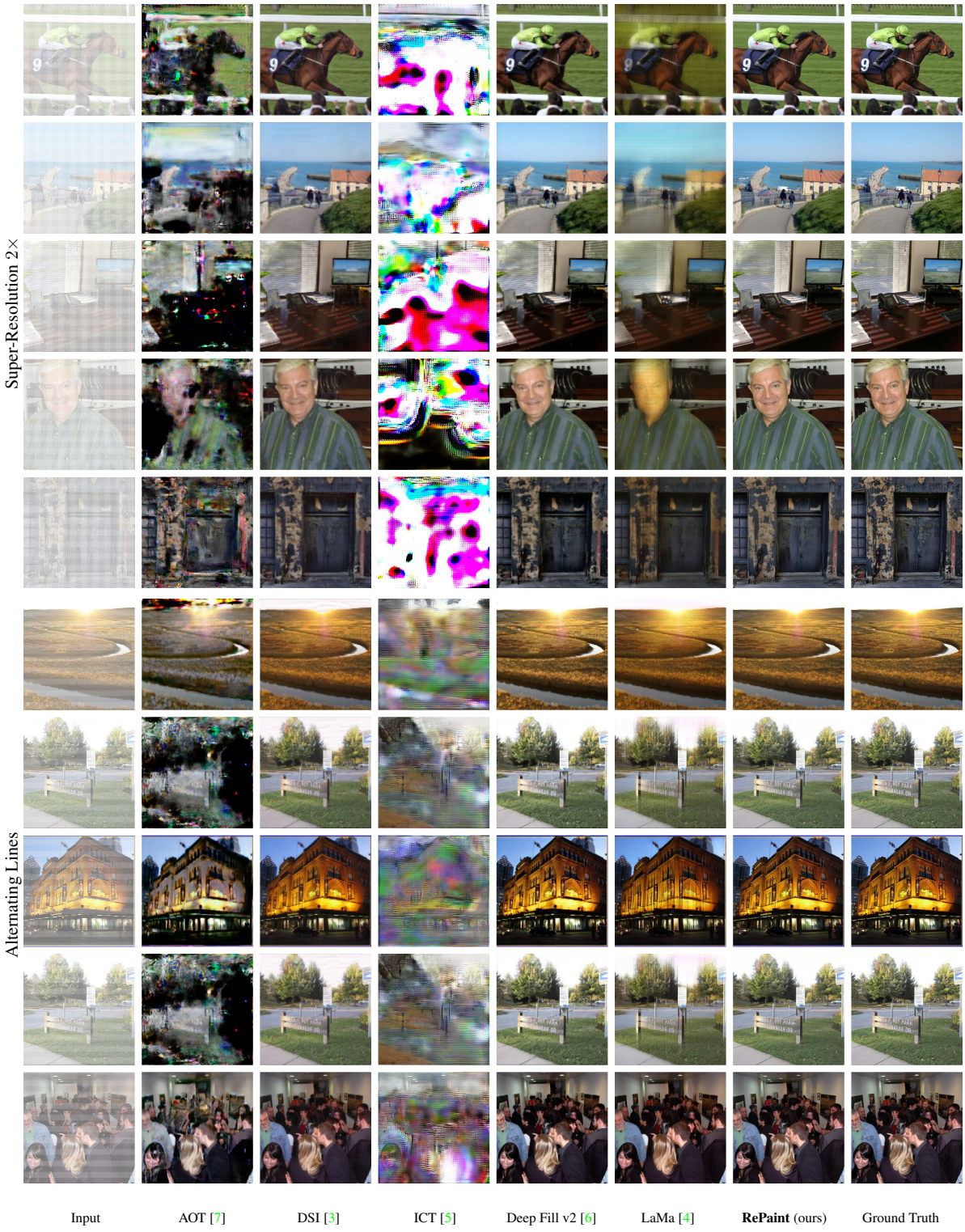


Figure 16. **Places2 Qualitative Results.** Comparison against the state-of-the-art methods for diverse inpainting. Zoom for better details.

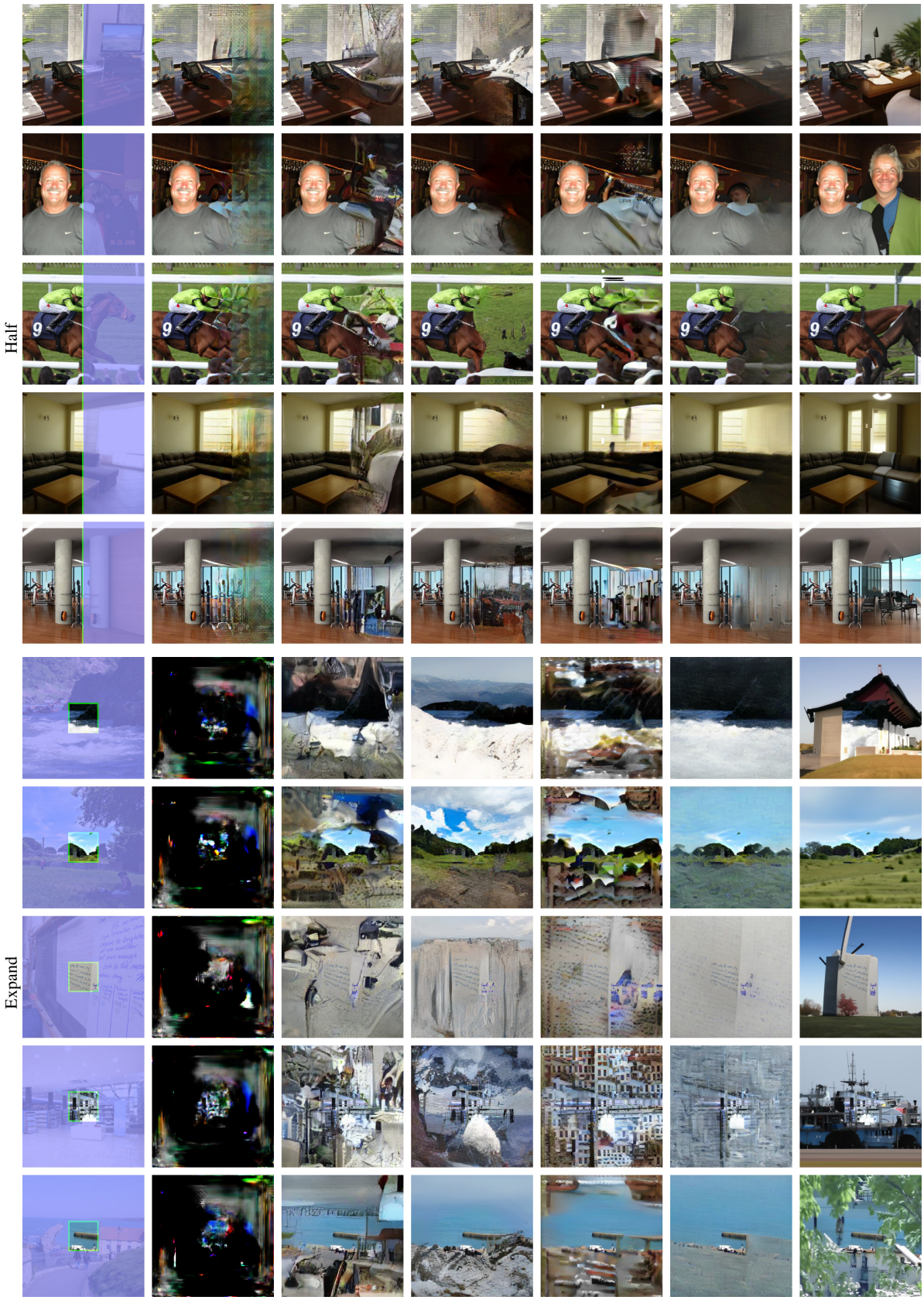


Figure 17. **Places2 Qualitative Results.** Comparison against the state-of-the-art methods for diverse inpainting. Zoom for better details.

References

[1] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis. *arXiv preprint arXiv:2105.05233*, 2021. 2, 3

[2] Andreas Lugmayr, Martin Danelljan, and Radu Timofte. Ntire 2021 learning the super-resolution space challenge. In *CVPRW*, 2021. 3

[3] Jialun Peng, Dong Liu, Songcen Xu, and Houqiang Li. Generating diverse structure for image inpainting with hierarchical vq-vae. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10775–10784, 2021. 3, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14

[4] 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, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14

[5] Ziyu Wan, Jingbo Zhang, Dongdong Chen, and Jing Liao. High-fidelity pluralistic image completion with transformers. *arXiv preprint arXiv:2103.14031*, 2021. 3, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14

[6] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. *arXiv preprint arXiv:1801.07892*, 2018. 3, 5, 6, 7, 8, 12, 13, 14

[7] Yanhong Zeng, Jianlong Fu, Hongyang Chao, and Baining Guo. Aggregated contextual transformations for high-resolution image inpainting. *arXiv preprint arXiv:2104.01431*, 2021. 3, 5, 6, 7, 8, 12, 13, 14

[8] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 2