Improving Detail in Pluralistic Image Inpainting with Feature Dequantization

A. Encoder-decoder Arcitecture

In our proposed method, encoder-decoder has the same architecture as PUT [4].

Encoder. The encoder is composed of 8 linear residual blocks and 2 linear layers. Each block consists of a linear layer followed by a ReLU activation function. After the linear layer, output is added to the input of the block and passed through a ReLU activation function. In the encoding process, initially, the masked image $\hat{\mathbf{x}} \in \mathbb{R}^{H \times W \times 3}$ is transformed from RGB channels (3) to feature channels (256) through linear layers with ReLU activation functions. Subsequently, the feature undergoes computation with linear residual blocks. Finally, to convert features channels into the codebook vector channels (256), it passes through a linear layer followed by a ReLU activation function.

Decoder. The decoder comprises two paths: a feature upsampling path and a masked image downsampling path.

The feature upsampling path consists of 8 convolutional residual blocks and 3 upsampling layers. Each block includes a 3×3 convolutional layer followed by a ReLU activation function and a 1×1 convolutional layer. Subsequently, the output of the 1×1 convolutional layer is added to the input of the block and passed through a ReLU activation function. Upsampling is conducted using 4×4 deconvolutional layers with a stride of 2.

The masked image downsampling path comprises 3 downsampling layers. Downsampling is executed using 3×3 convolutional layers followed by ReLU activation. In the feature upsampling path, features traverse through residual blocks before undergoing upsampling. Subsequently, the features are upsampled to match the image resolution. In the proposed method, the feature resolution is 32×32, and the image resolution is 256×256, thus 3 upsampling steps are performed.

At each upsampling step, the upsampled features are combined with the downsampled masked image using the following equation:

$$\mathbf{f}_n = \mathbf{f}'_n \otimes (1 - \mathbf{m}_n) + \hat{\mathbf{x}}_n \otimes \mathbf{m}_n \tag{1}$$

where $\mathbf{f}'_n$ represents the upsampled feature in the n-th step, $\hat{\mathbf{x}}_n$ denotes the downsampled masked image in the n-th step, and $\mathbf{m}_n$ signifies the downsampled mask in the n-th step.

Finally, to convert features channels (256) into the RGB channels (3), upsampled features passes through a 3×3 convolutional layer.

B. Comparison with Deterministic Method

We compare the proposed method with the following state-of-the-art deterministic inpainting approach LaMa [6]. Table 1 displays the quantitative results comparing our proposed approach, FDM, with LaMa. FDM shows competitive performance in terms of FID compared to the state-of-the-art model LaMa. Particularly, it demonstrates better performance, especially in the case of large masks.

Figure 1 illustrates the inference results of FDM and Lama. It can be observed that Lama's inpainting performance deteriorates when the mask ratio is high. On the contrary, the proposed method generates natural images even as the mask ratio increases.

LaMa sometimes fills the mask with a single color, as seen in the example when the mask ratio is wide. In such cases, if the model generates unnatural images, users have no way to improve them. In contrast, PII offers various generated results, thereby expanding the user's choice and increasing the likelihood of obtaining satisfactory results.

C. Additional Quantitative Results

Table 2 presents a comparison of methods across MAE, PSNR and SSIM. These metrics evaluate pixel-wise similarity between an output image and the ground-truth image, without considering diversity or alignment with human perception. Therefore, they are not suitable for evaluating pluralistic inpainting, as discussed in Section 4.1.

FDM converts quantized features into continuous features based on the predicted image structure by the feature sampler, without making them closer to the ground-truth features. Therefore, FDM does not significantly improve the performance of PUT [4] in terms of PSNR, SSIM [8], and MAE, unlike in FID [2] and LPIPS [9].

As discussed in Section 4.4, the Paris Street View dataset [1] contains a lot of noise, resulting in a decrease in both the performance improvement capability of FDM and the

Table 1. Comparison with deterministic inpainting model

Methods	Places				Paris Street View			
	FID		LPIPS		FID		LPIPS	
	small	large	small	large	small	large	small	large
LaMa	**17.30**	29.86	**0.120**	**0.213**	**10.89**	21.79	**0.124**	0.236
Ours (PUT +FDM)	18.46	**29.67**	0.127	0.230	11.63	**18.66**	0.131	**0.234**

Figure 1. Comparison with deterministic models across various mask ratios.

Table 2. Quantitative results of different methods. **Bold** indicates the best score in PII methods.

Type	Dataset	Places [10]						Paris Street View [1]					
	Metric	MAE		SSIM		PSNR		MAE		SSIM		PSNR	
	Mask Ratio	small	large	small	large	small	large	small	large	small	large	small	large
DII	LaMa	0.024	0.045	0.868	0.707	26.19	22.38	0.025	0.053	0.897	0.764	25.94	21.94
PII	ICT [7]	0.033	0.059	0.821	0.625	24.33	20.53	0.035	0.065	0.857	0.693	24.20	20.47
	MAT [3]	0.028	**0.049**	**0.873**	**0.700**	**26.48**	**22.09**	0.031	0.063	0.859	0.688	24.22	20.22
	LDM [5]	**0.024**	0.049	0.849	0.662	25.47	21.24	0.032	0.063	0.849	0.689	23.87	20.23
	PUT [4]	0.028	0.054	0.840	0.649	25.07	20.88	0.031	0.059	0.875	0.729	24.90	21.11
	Ours (PUT +FDM)	0.026	0.052	0.844	0.653	25.24	20.98	**0.030**	**0.058**	**0.877**	**0.733**	**25.03**	**21.22**

performance of MAT [3] and LDM [5]. However, as discussed in Section 4.2, the proposed method demonstrates much greater diversity than MAT and achieves better FID scores than LDM with large masks. Therefore, our proposed method has been proven to generate diverse and natural-looking images.

D. Additional Qualitative Results

Figure 2 provides a more detail comparison between PUT [4] and our proposed method. PUT produce color discrepancies and distorted structures or fails to properly represent texture. For example, in row 1 and 3, PUT generates the window grilles inconsistently or unclearly. However, our proposed method generates the window grilles in a straight line without interruption.

Figure 3 and Figure 4 provide more visual comparison of diverse inpainting results among PII methods. In ICT [7] and LDM [5], artifacts have been generated, such as blurring or structural ambiguity. Although MAT [3] shows few artifacts, it often generates structurally similar images, resulting in limited diversity in the results. In contrast, our proposed method has successfully generated diverse images while preserving naturalness.

References

[1] Carl Doersch, Saurabh Singh, Abhinav Gupta, Josef Sivic, and Alexei Efros. What makes paris look like paris? *ACM Transactions on Graphics*, 31(4), 2012. 1, 2

[2] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 1

[3] Wenbo Li, Zhe Lin, Kun Zhou, Lu Qi, Yi Wang, and Jiaya Jia. Mat: Mask-aware transformer for large hole image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 2

[4] Qiankun Liu, Zhentao Tan, Dongdong Chen, Qi Chu, Xiyang Dai, Yinpeng Chen, Mengchen Liu, Lu Yuan, and Nenghai Yu. Reduce information loss in transformers for pluralistic image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11347–11357, 2022. 1, 2

[5] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2

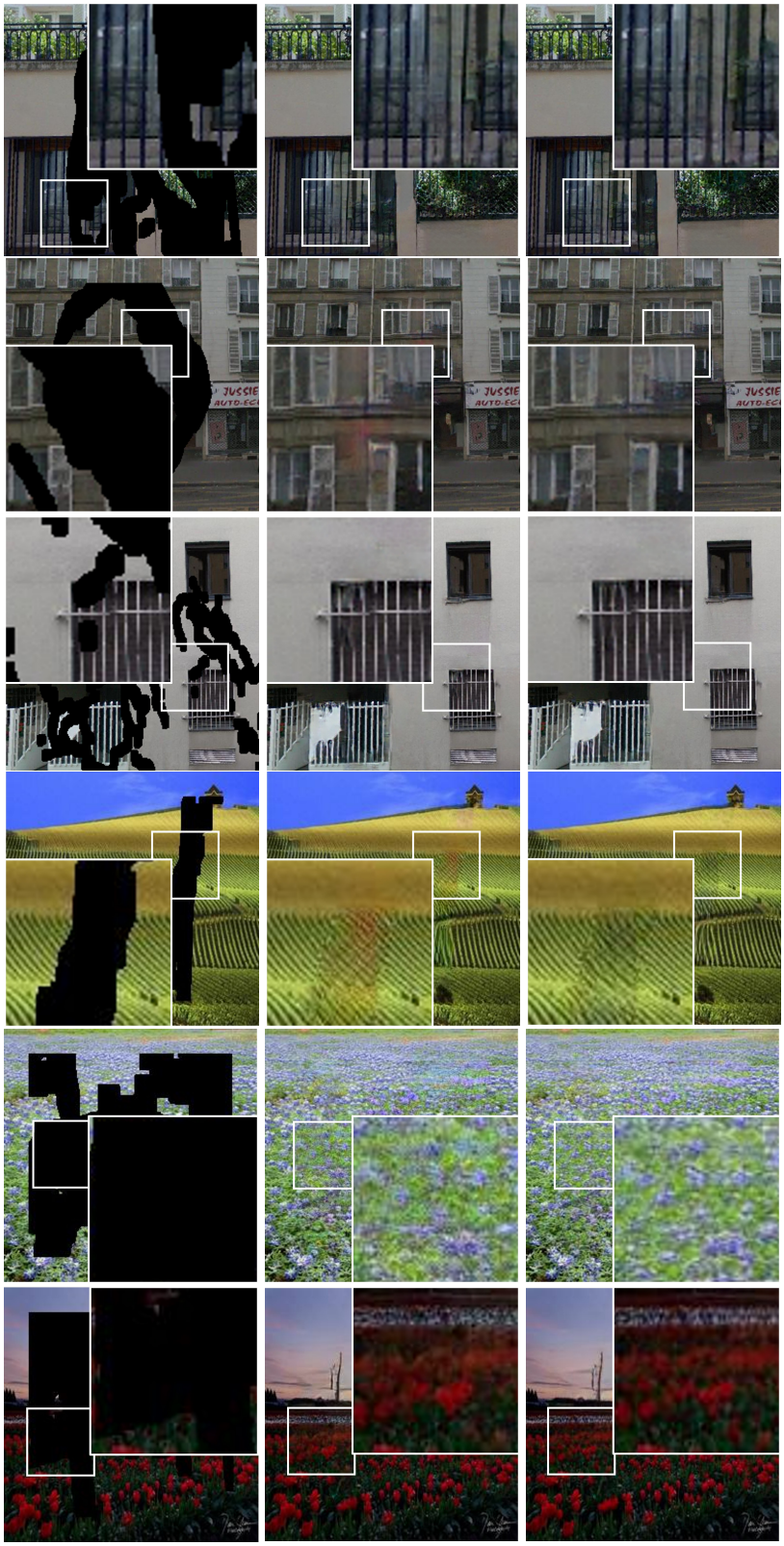
[6] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2149–2159, 2022. 1

[7] Ziyu Wan, Jingbo Zhang, Dongdong Chen, and Jing Liao. High-fidelity pluralistic image completion with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4692–4701, 2021. 2

[8] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 1

[9] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 1

[10] 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*, 40(6):1452–1464, 2017. 2



Input PUT Ours

Figure 2. Detail comparison between proposed method and PUT.

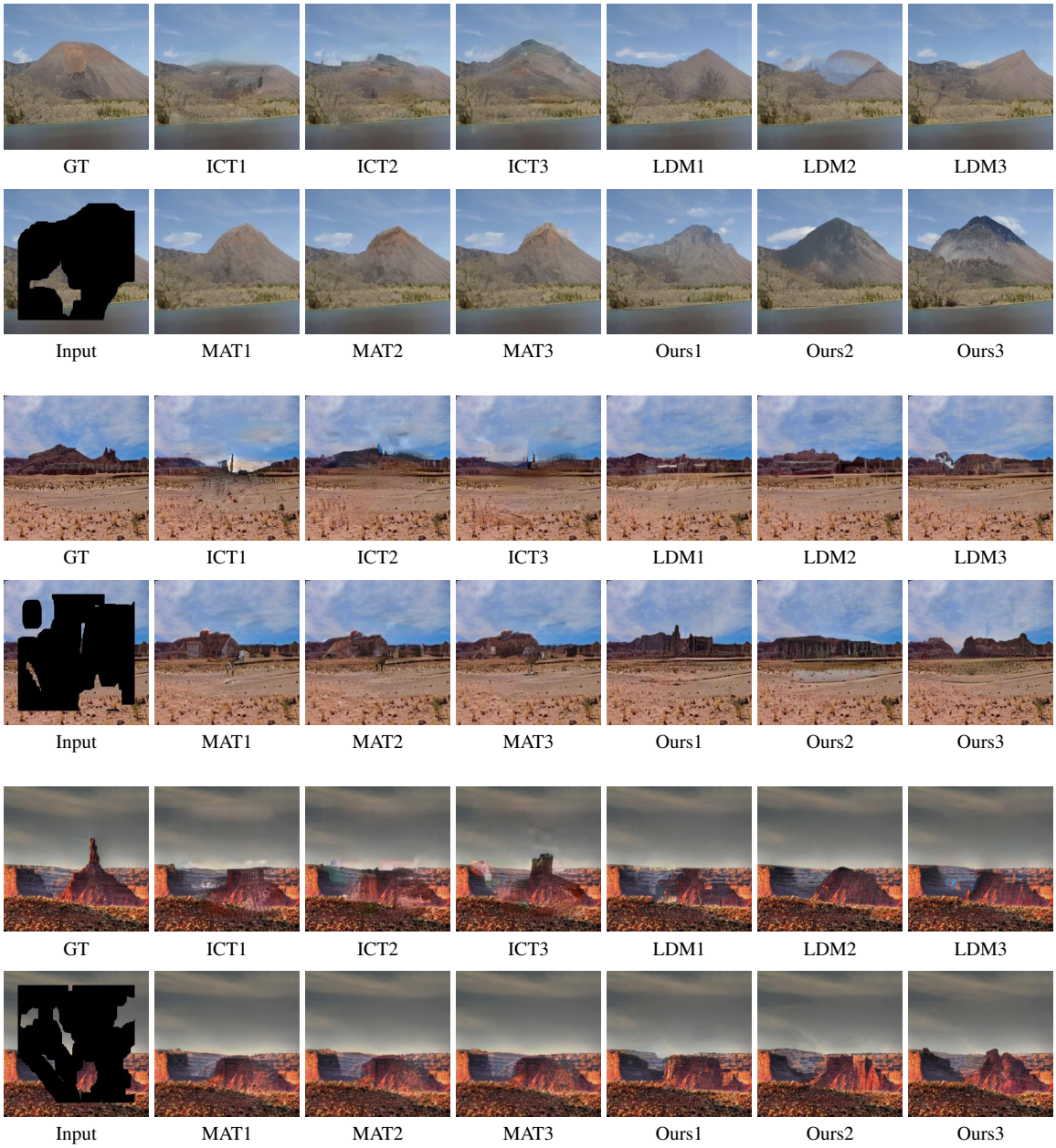


Figure 3. Visual comparison of diverse inpainting results in Places

Figure 4. Visual comparison of diverse inpainting results in Paris Street View