High-Fidelity Pluralistic Image Completion with Transformers Supplementary Material

Ziyu Wan[1] Jingbo Zhang[1] Dongdong Chen[2] Jing Liao[1*]

[1]City University of Hong Kong [2]Microsoft Cloud + AI

1. Overview

In this supplemental material, additional experimental details, analysis and results are provided, including:

- more details about network architectures(Section 2);
- more qualitative comparisons on different datasets (Section 3);
- more analysis on proposed method (Section 4).

2. Network Architecture

2.1. Transformer

We plot the details of transformer layer in Figure. 1.

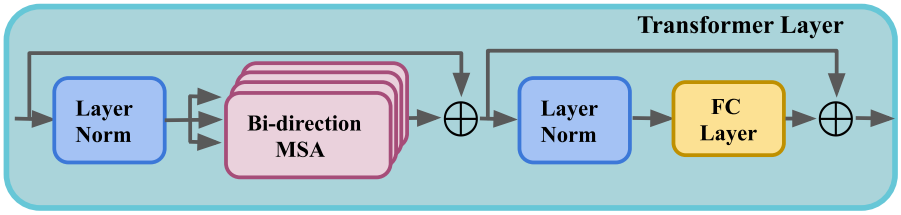

Figure 1: **Details of transformer layer. MSA**: Multi-head self-attention.

The detailed configurations of different transformer models are shown in Table. 1. During training, we try to use Automatic Mixed Precision (AMP) to speed up the training period, but find AMP would easily lead to NAN parameters. Thus we disable AMP in all experiments.

Experiment	h	d	N	$\mathbb{L}$	Parameter #
FFHQ [5]	8	512	30	48x48	97M
PLACES2 [12]	8	512	35	32x32	112M
IMAGENET[9]	8	1024	35	32x32	443M

Table 1: **Transformer parameter setting across different experiments.** h: Head number of bi-directional attention. d: The dimension of embedding space. N: Number of transformer layer. $\mathbb{L}$: The length of appearance prior.

*Corresponding author.

Module	Layer	Kernel size / stride	Output size
Encoder E	Conv	$7 \times 7/1$	$256 \times 256 \times 64$
	Conv	$4 \times 4/2$	$128 \times 128 \times 128$
	Conv	$4 \times 4/2$	$64 \times 64 \times 256$
Decoder D	Deconv	$4 \times 4/2$	$128 \times 128 \times 128$
	Deconv	$4 \times 4/2$	$256 \times 256 \times 64$
	Conv	$7 \times 7/1$	$256 \times 256 \times 3$
ResBlock $R\times$ 8	Dilated Conv	$3 \times 3/1$	$64 \times 64 \times 256$
	Dilated Conv	$3 \times 3/1$	$64 \times 64 \times 256$

Table 2: **Detailed guided upsampling network $\mathcal{F}$ structure.** In the residual block, we employ the dilated conv with dilation $= 1$, which is shown in gray.

Module	Layer	Kernel size / stride	Output size
Discriminator $\mathcal{D}$	Conv	$4 \times 4/2$	$128 \times 128 \times 64$
	Conv	$4 \times 4/2$	$64 \times 64 \times 128$
	Conv	$4 \times 4/2$	$32 \times 32 \times 256$
	Conv	$4 \times 4/1$	$31 \times 31 \times 512$
	Conv	$4 \times 4/1$	$30 \times 30 \times 1$

Table 3: **Detailed discriminator $\mathcal{D}$ structure.**

2.2. Guided Upsampling Network

Table. 2 and Table. 3 show the employed architecture of guided upsampling network $\mathcal{F}$ and discriminator $\mathcal{D}$, which are fixed in all experiments. For each convolution layer of $\mathcal{D}$, we use spectrum normalization (SN) [7] to stabilize the training procedure.

3. More Results

We show more qualitative comparisons in Figure. 2, Figure. 3 and Figure. 4.

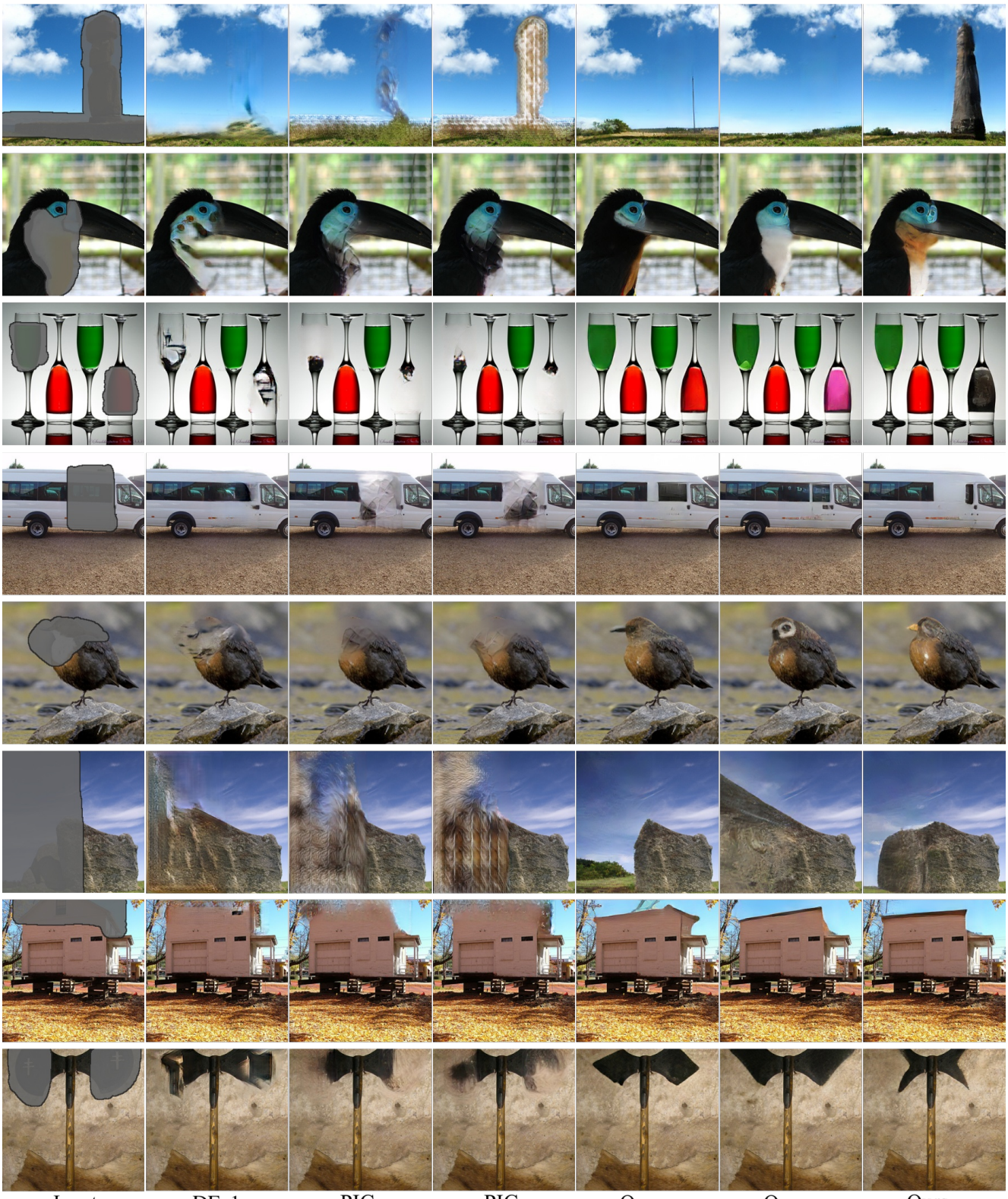

Figure 2: **Comparisons on ImageNet.**

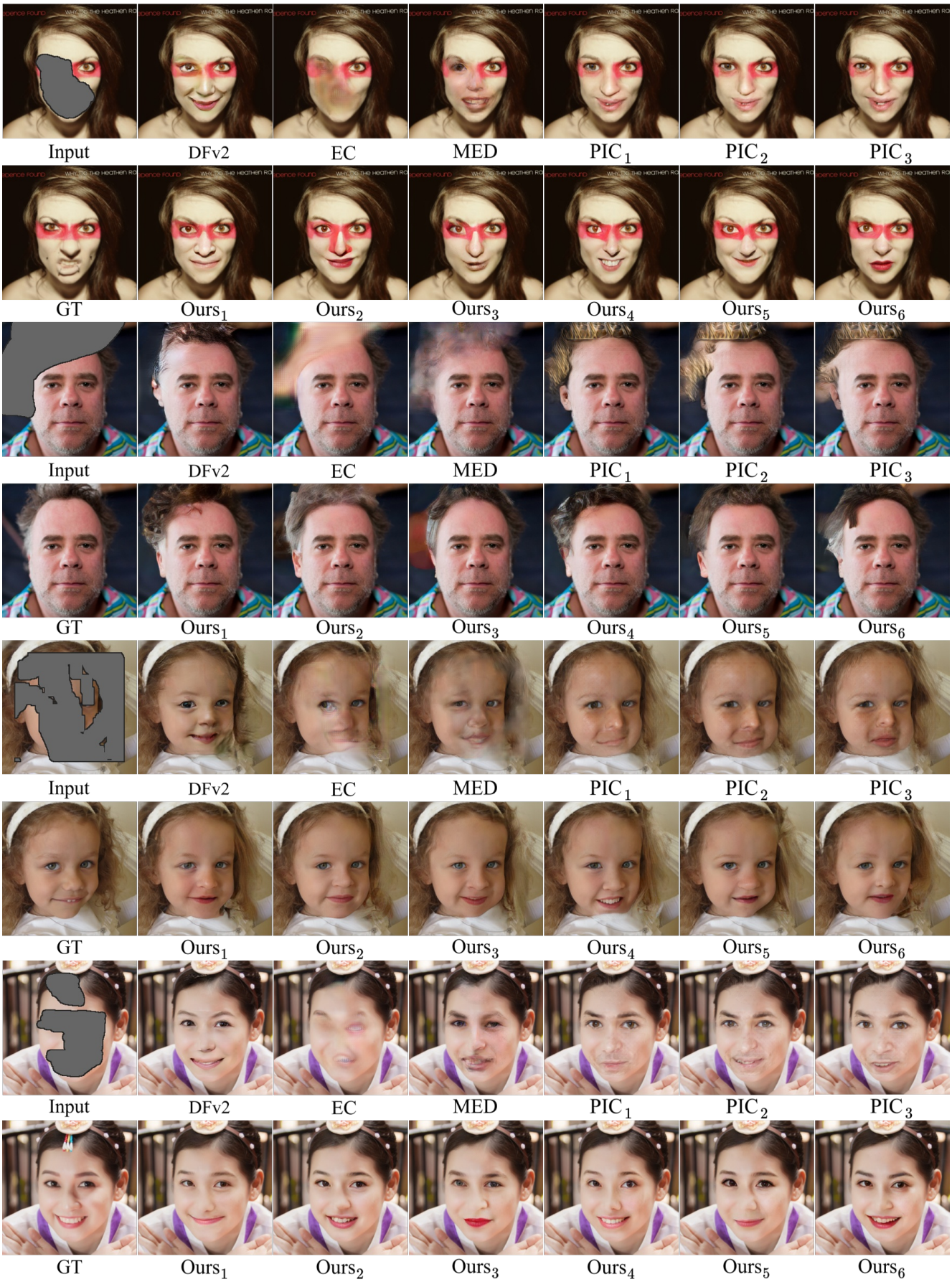

Figure 3: **Comparisons on FFHQ.**

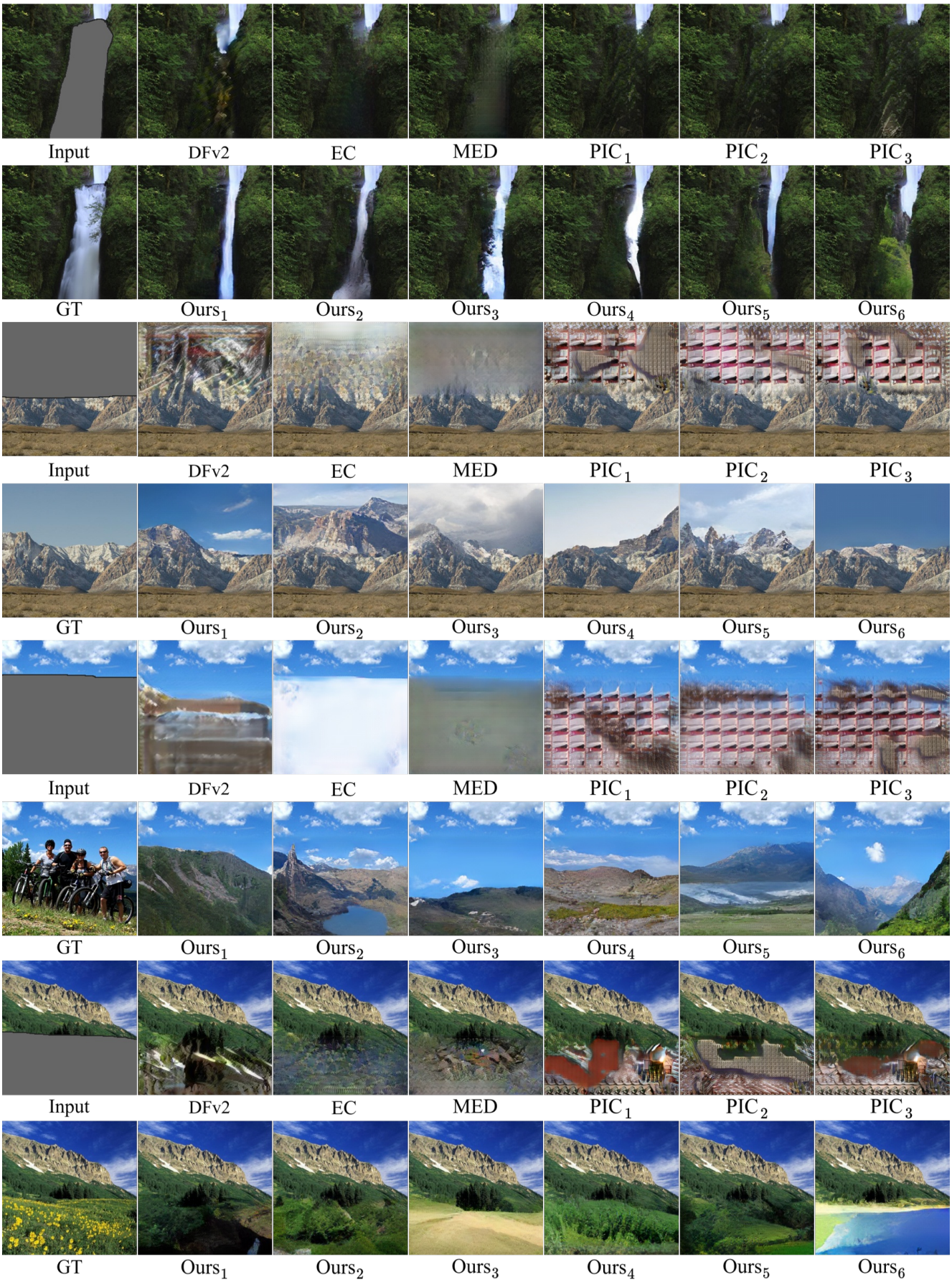

Figure 4: **Comparisons on Places2.**

4. More Analysis

4.1. Visualization of Appearance Prior

To further understand the effectiveness of transformer and the appearance priors, we give some random reconstructed results in this section. As shown in Figure. 5, although these priors produced by transformer are low-resolution and just composed with discrete RGB tokens, they could faithfully represent the information of structures and textures of one complete image, as well as containing abundant diversities.

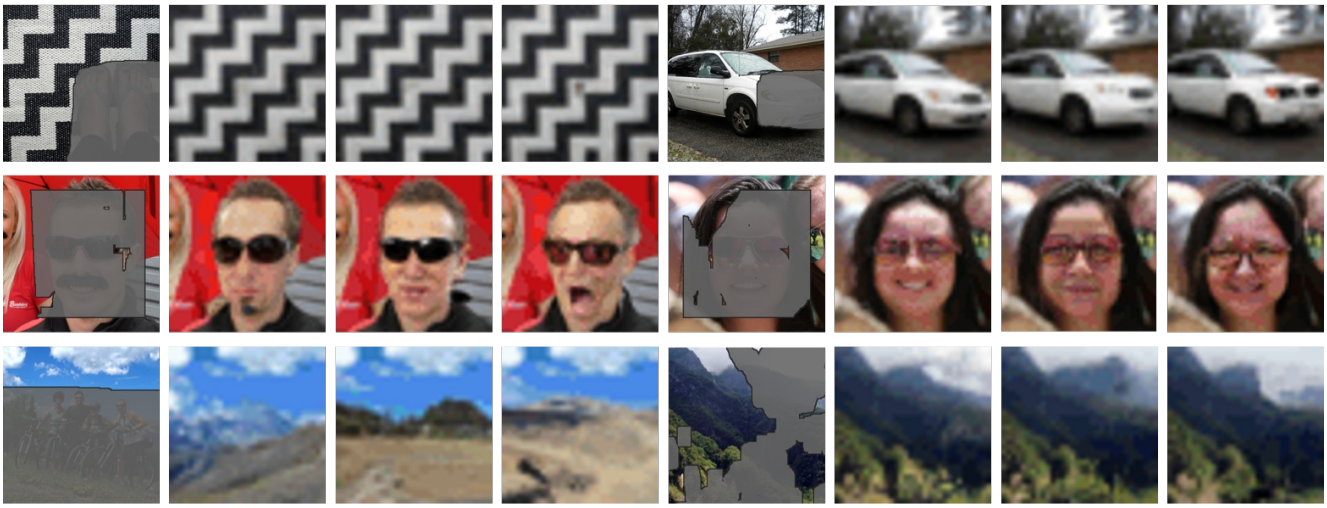

Figure 5: **Reconstructed appearance priors from the transformer.**

4.2. From Image Completion to Unconditional Image Generation

In this setting, we try to let the completion transformer perform unconditional image generation task, which means all pixels of input image are erased. Some random generation examples are shown in Figure. 6. We could observe that part of generated images like hair region lose some texture details. This phenomenon may be caused by the upsampling network, since here no useful pixels are regarded as guidance. One potential solution is to add extra full zero masks into the training procedure. Overall, the generated results are reasonable and diverse.

Figure 6: **Examples of unconditional image generation.**

4.3. Comparison with IGPT [1]

IGPT [1] has demonstrated great abilities for model pre-training and image generation. However, they could not handle the image completion with arbitrary masks well. In Table 4 and Figure 7, we compare our method with IGPT model. By leveraging bi-direction information and the texture enhancement capability of CNN, our method clearly outperforms IGPT by a large margin.

Method	Mask Ratio	PSNR↑	SSIM↑	MAE↓	FID↓
IGPT	Random	19.407	0.671	0.0624	103.088
Ours		**23.775**	**0.835**	**0.0358**	**35.842**

Table 4: Quantitative comparison with IGPT on ImageNet.

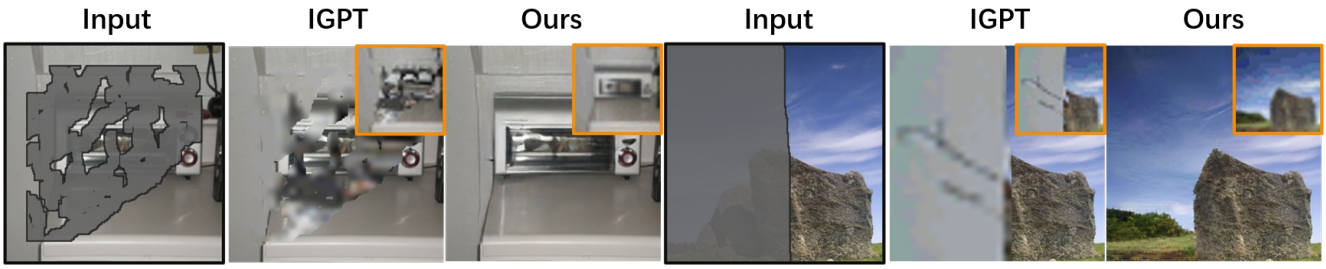

Figure 7: Qualitative comparison with IGPT on ImageNet.

4.4. Discussion of Limitation

Inference Speed Currently the main limitation of our proposed method is inference speed, which is also the common issue of auto-regressive method [10] and transformer-based model [8, 1]. We provide the speed statistics in Table. 5. All the inference experiments are conducted on single RTX 2080Ti GPU. Specifically, the transformer model trained on *FFHQ* [5] could generate 6.5 tokens per second. On *Places2* [12], the number of processed tokens in each second are increased to 19.9 since the input length becomes shorter. As the largest trained model, on ImageNet it could produce 8.3 tokens per second. By contrast, in the second guided upsampling stage, the inference only requires 1.3 seconds for each input images, which is less than the transformer model, meanwhile demonstrating the high efficiency of CNNs.

There are two interesting directions to alleviate this problem: 1) **More efficient attention mechanism.** Since the employed attention owns $O\left(\mathbb{L}^2\right)$ time complexity, we could reduce this computational cost using recent fast attention techniques [3] to $O(\mathbb{L}\sqrt{\mathbb{L}})$ [11, 2]; 2) **Faster sampling strategy.** Unlike the general auto-regressive model, our bi-directional transformer model is not limited to fixed sampling order. For example, in each iteration, we could update the several grids simultaneously with high confidence. We will explore these methods in the future.

Experiments	FFHQ [5]	PLACES2 [12]	IMAGENET [9]
Transformer (Token # /sec)	6.5	19.9	8.3
Upsampling (sec /frame)	1.3	1.3	1.3

Table 5: **Inference speed of different components.**

Upsampling Artifacts We notice that sometimes the upsampled results will have some slight blurriness artifacts, as shown in the green bounding box of Figure. 8. This may be caused by the GAN training. In the future, we will explore employing more advanced upsampled network [6], more powerful discriminator and adversarial objective [4] to alleviate this problem.

References

[1] Mark Chen, Alec Radford, Rewon Child, Jeff Wu, Heewoo Jun, Prafulla Dhariwal, David Luan, and Ilya Sutskever. Generative pretraining from pixels. 2020.

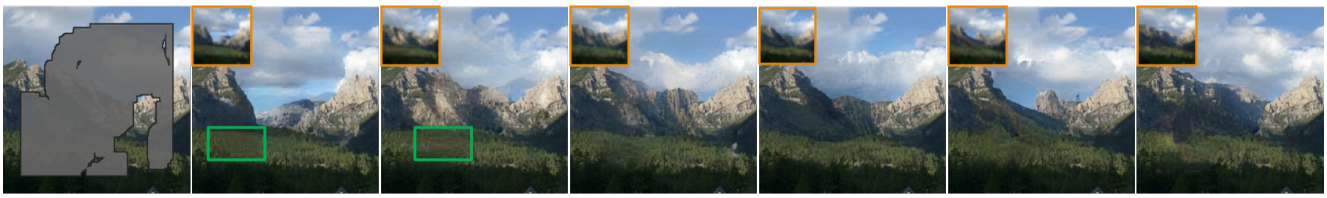

Figure 8: **Upsampling artifacts.** First column: input. Rest: Completion results of our method.

[2] Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, et al. Rethinking attention with performers. *arXiv preprint arXiv:2009.14794*, 2020.

[3] Jonathan Ho, Nal Kalchbrenner, Dirk Weissenborn, and Tim Salimans. Axial attention in multidimensional transformers. *arXiv preprint arXiv:1912.12180*, 2019.

[4] Alexia Jolicoeur-Martineau. The relativistic discriminator: a key element missing from standard gan. *arXiv preprint arXiv:1807.00734*, 2018.

[5] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019.

[6] Jie Liu, Wenjie Zhang, Yuting Tang, Jie Tang, and Gangshan Wu. Residual feature aggregation network for image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2359–2368, 2020.

[7] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.

[8] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. In *International Conference on Machine Learning*, pages 4055–4064, 2018.

[9] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[10] Tim Salimans, Andrej Karpathy, Xi Chen, and Diederik P Kingma. Pixelcnn++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications. *arXiv preprint arXiv:1701.05517*, 2017.

[11] Sinong Wang, Belinda Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*, 2020.

[12] 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.