Supplementary: Keys to Better Image Inpainting: Structure and Texture Go Hand in Hand

Jitesh Jain[1,2,3*†] Yuqian Zhou[4*†] Ning Yu[5] Humphrey Shi[1,3]

[1]SHI Lab @ University of Oregon [2]IIT Roorkee
[3]Picsart AI Research (PAIR) [4]Adobe Inc. [5] Salesforce Research

https://praeclarumjj3.github.io/fcf-inpainting/

Appendix

We provide ablation studies on applying the FaF-Syn module to different resolutions and loss functions in Sec. 1. We also provide a quantitative comparison at different masked ratios in Sec. 2. Lastly, we provide more qualitative comparisons in Sec. 3.

1. Additional Ablation Studies

Ablation on Resolution for FFC Residual Blocks. We experiment with application of our FaF-Syn block at lower resolutions with the setting $\{L_{32} : 1, , L_{64} : 1, L_{128} : 1, L_{256} : 1\}$. For each experiment we set $L_{\text{res}} = 1$ for res $\in \{8, 16\}$. We observe that adding FFC to lower resolutions harms the performance as shown in Tab. I. We reason that the lower resolution features contain insufficient spatial information required for modeling the global context. The coarse-level features input to the FFC are magnified with noise, thus leading to a drop in performance and even instability during training (4×4).

8×8	16×16	32×32	64×64	128×128	256×256	FID	LPIPS
			✓	✓	✓	12.14	0.266
		✓	✓	✓	✓	**11.33**	0.264
	✓	✓	✓	✓	✓	11.69	0.263
✓	✓	✓	✓	✓	✓	12.24	0.269
✓			✓	✓	✓	11.83	0.266
	✓		✓	✓	✓	11.44	0.262

Table I: **Ablation on resolution for FFC Residual Blocks.** Applying FFC to lower resolution coarse-features harms the performance.

Loss Functions. We ablate the effect of different loss terms on the inpainting performance of our framework. We remove the $\mathcal{L}_{\text{rec}}$ and $\mathcal{L}_{\text{HRFPL}}$ from the total loss to study the importance of pixel and feature level supervision, respectively. We use the L_{adv} and the R_1 regularization as usual. We trained our models for 10M images and evaluated them on 10k images with free-form masks [6] sampled from the Places2 [8] val dataset. We observe an increase in the FID and LPIPS scores when removing the loss terms. The major drop in performance (increase in FID and LPIPS score) happens when we remove both the loss terms. We also conclude that using only adversarial loss while training models based on FFC [1] can lead to major drop as FFC requires supervision from the frequency signal present in the images as shown in Tab. II.

$\mathcal{L}_{\text{rec}}$	$\mathcal{L}_{\text{HRFPL}}$	FID ↓	LPIPS ↓
		16.83	0.297
✓		14.14	0.279
	✓	12.52	0.270
✓	✓	**11.33**	**0.264**

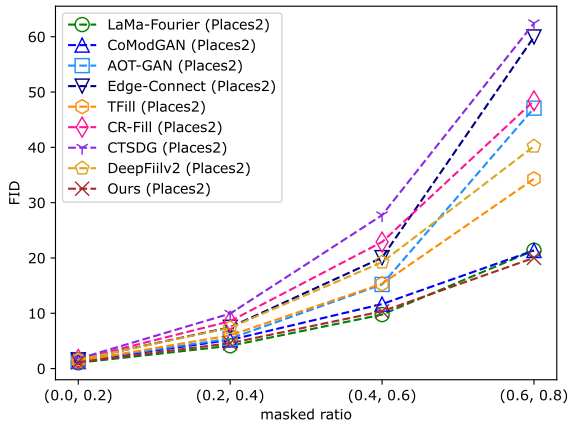
Table II: **Ablation on Loss Functions.** We study the impact of reconstruction and HRFPL losses during training. We observe that pixel and feature level supervision is critical to the success of FFC based networks.

2. Quantitative comparison at different Masked Ratios.

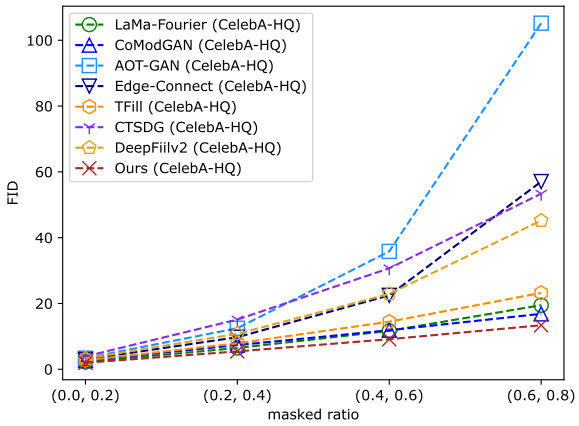
We study the quantitative performance with different hole ratios in Fig. I. A larger hole means it is more challenging to complete the structure. We use a free-form mask generation strategy to generate 10k samples for Places2 and 2k samples for CelebA-HQ during evaluation. The results showed that only Ours, LaMa [5], and CoModGAN† [6] performed consistently well as the hole size increased. Other state-of-the-arts still struggle to fill complex structures. Among them, TFill [7] with transformer-based network structures works better. Ours are robust enough for

* Equal Contribution.

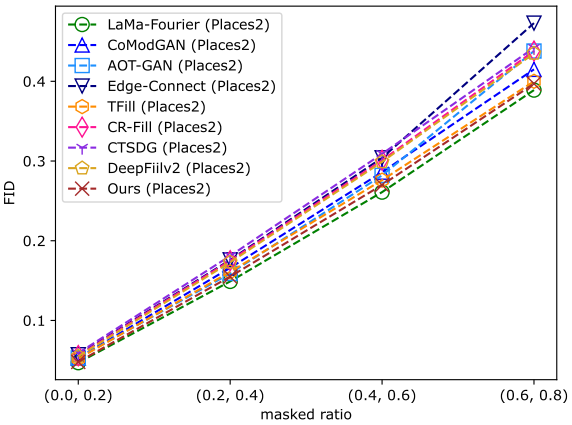
† This work started when Jitesh interned at SHI Lab @ University of Oregon, and Yuqian was a Ph.D. student at IFP @ UIUC.



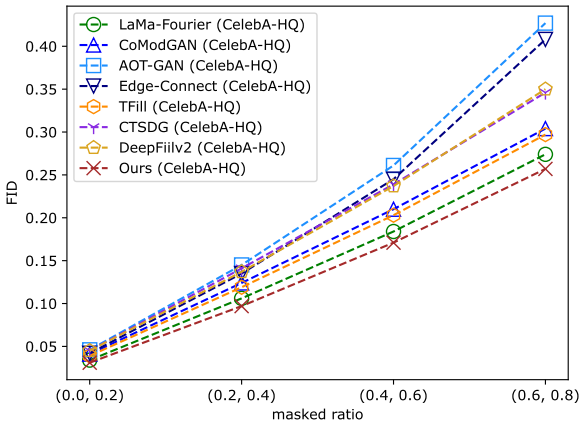
(a) FID comparison on Places2



(b) FID Comparison on CelebA-HQ



(c) LPIPS comparison on Places2



(d) LPIPS comparison on CelebA-HQ

Figure I: **Evaluation on ratio-wise masks.** We plot and compare the FID and LPIPS scores of our framework to all baselines with respect to masked ratios. Larger masks bring more challenging cases in completing structures. Ours, as well as LaMa and CoModGAN†, perform consistently well than other baselines.

both Places2 [8] and CelebA-HQ [3] datasets.

3. More Qualitative Results

We provide more qualitative results on Places2 [8] and CelebA-HQ [3] in Fig. II and Fig. III, respectively. We compare our FcF framework to TFill [7], CTSDG [2], LaMa-Fourier [5] and CoModGAN† (our PyTorch [4] implementation).

We also provide qualitative comparisons for our model trained on 512×512 resolution to the official publicly released models: LaMa-Fourier [5], Big-LaMa [5] and CoModGAN [6] in Fig. VI, Fig. IV, and Fig. V.

References

[1] Lu Chi, Borui Jiang, and Yadong Mu. Fast fourier convolution. In *NIPS*, 2020.

[2] Xiefan Guo, Hongyu Yang, and Di Huang. Image inpainting via conditional texture and structure dual generation. In *ICCV*, 2021.

[3] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *ICLR*, 2018.

[4] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An im-

Figure II: **Qualitative examples for image completion on** 256×256 **Places2.** We compare texture and structure completion among TFill [7], CTSDG [2], LaMa [5], CoModGAN† [6], and FcF (*Ours*)

Figure III: **Qualitative examples for image completion on** 256×256 **CelebA-HQ.** We compare the face structure completion among TFill [7], CTSDG [2], LaMa [5], CoModGAN† [6], and FcF (*Ours*)

perative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[5] 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 *WACV*, 2022.

[6] Shengyu Zhao, Jonathan Cui, Yilun Sheng, Yue Dong, Xiao Liang, Eric I Chang, and Yan Xu. Large scale image com-

pletion via co-modulated generative adversarial networks. In *ICLR*, 2021.

[7] Chuanxia Zheng, Tat-Jen Cham, Jianfei Cai, and Dinh Phung. Bridging global context interactions for high-fidelity image completion. In *CVPR*, 2022.

[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.

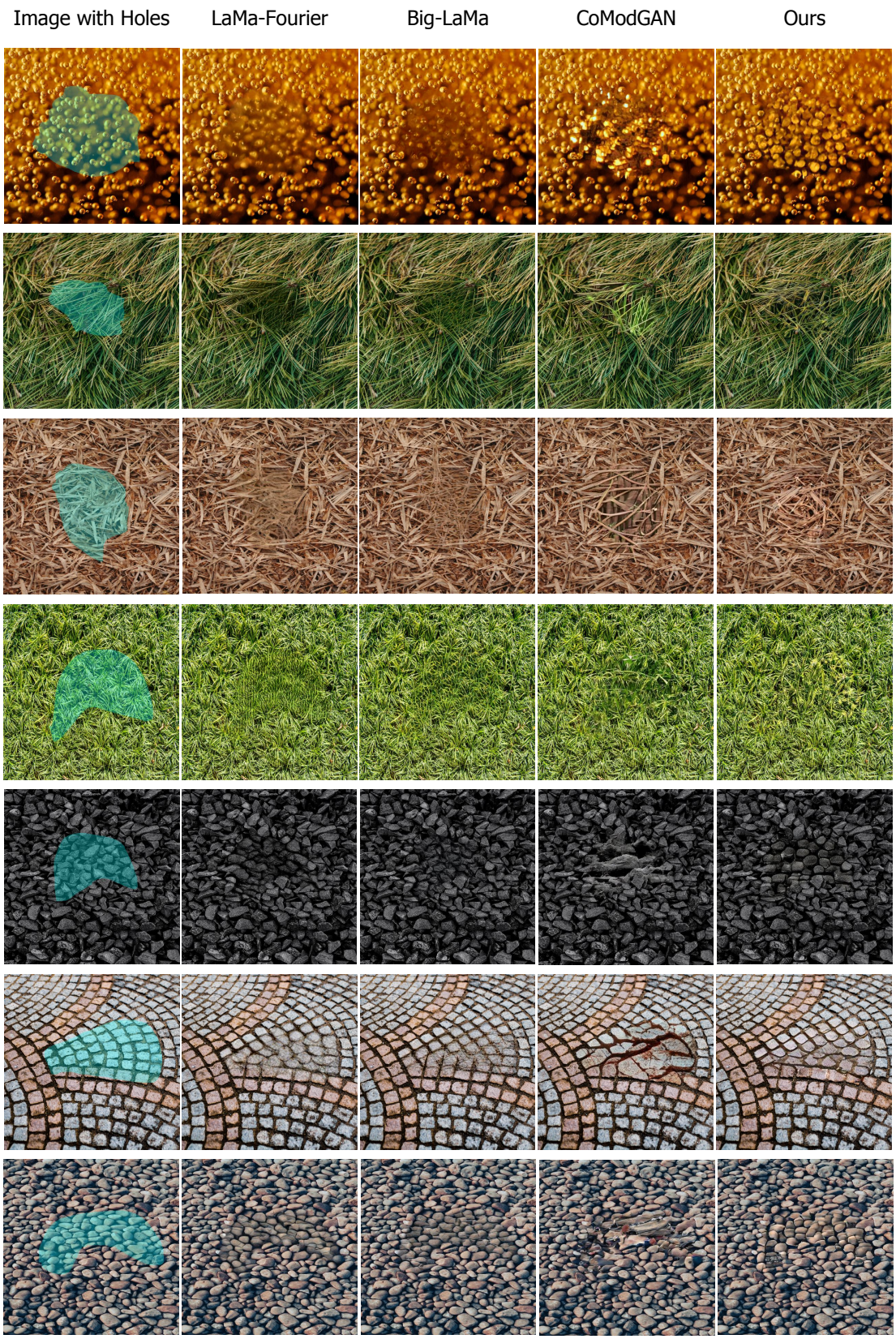


Figure IV: **Qualitative examples for image completion on** 512×512 **Texture Images.** We compare texture and structure completion among LaMa-Fourier [5], Big-LaMa [5], CoModGAN† [6], and FcF (*Ours*). Zoom-in for best view.

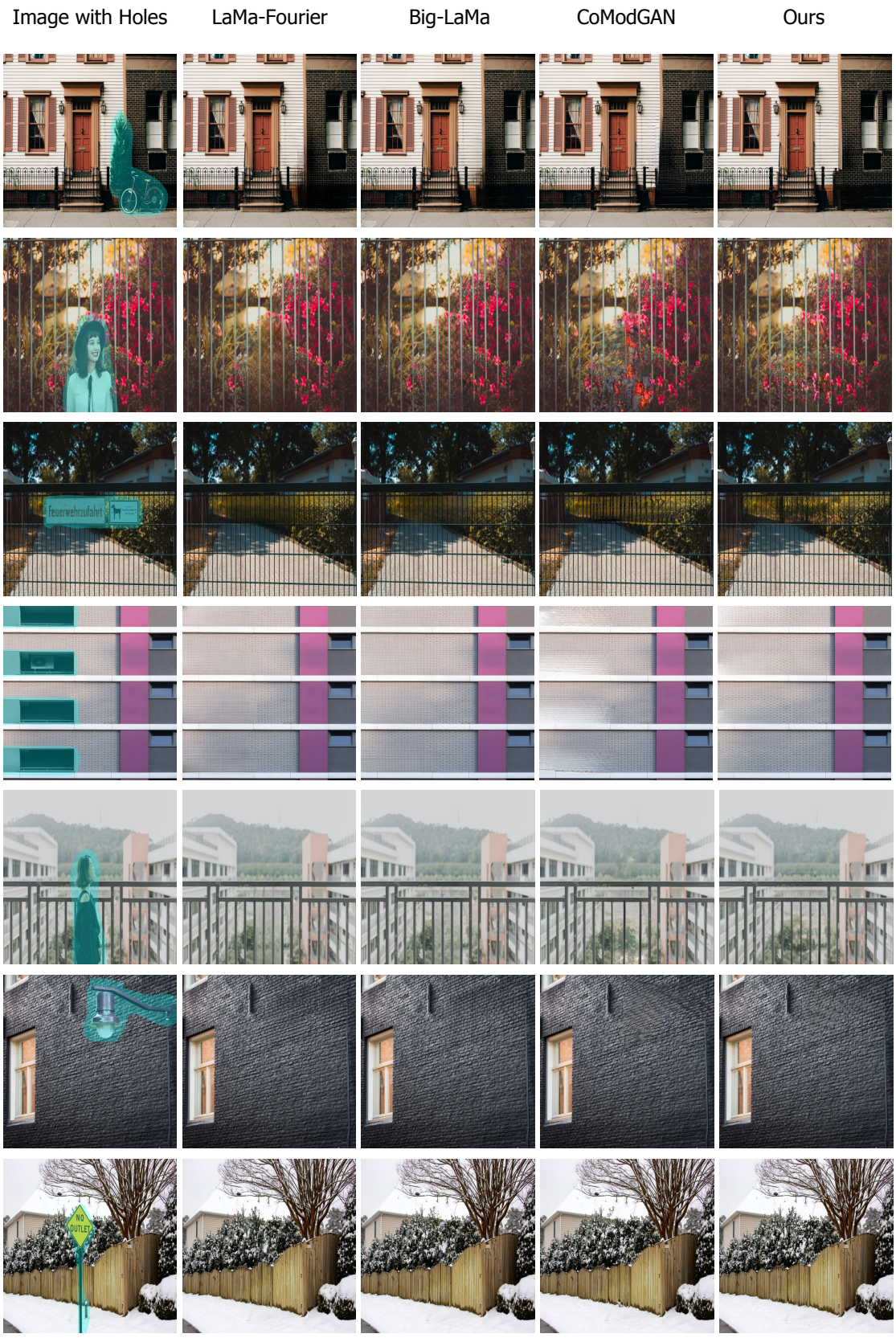


Figure V: **Qualitative examples for image completion on** 512×512 **images.** We compare texture and structure completion among LaMa-Fourier [5], Big-LaMa [5], CoModGAN† [6], and FcF (*Ours*). Zoom-in for best view.

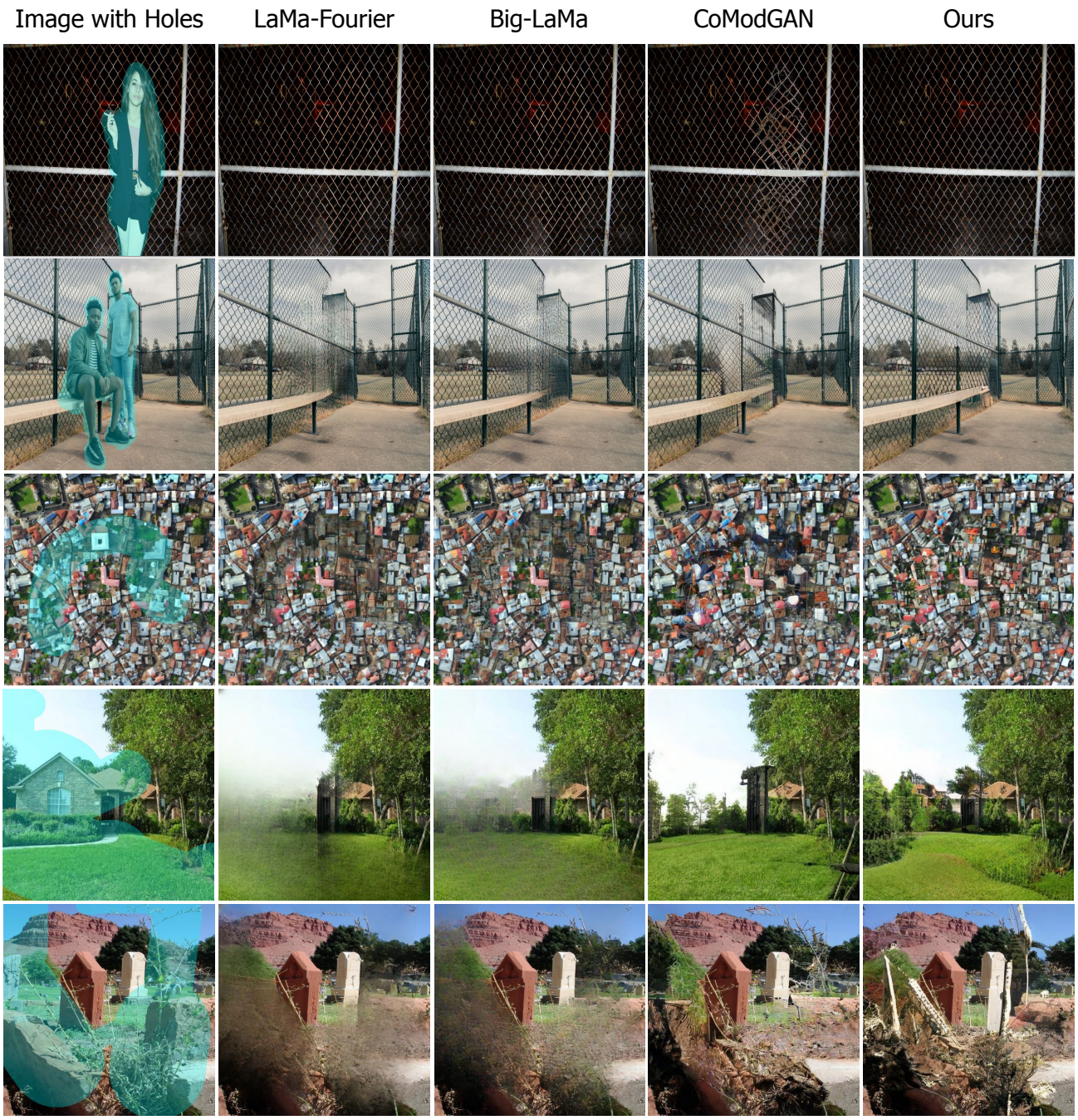


Figure VI: **Qualitative examples for image completion on** 512×512 **images.** We compare texture and structure completion among LaMa-Fourier [5], Big-LaMa [5], CoModGAN† [6], and FcF (*Ours*). Zoom-in for best view.