Boundless: Generative Adversarial Networks for Image Extension - Supplementary Material

1. Network Training and Architecture Details

1.1. Generator Network G

Layer ID	Type	Act.	K	S	D	Out	Skip
1	Gated Conv	ELU[2]	5	1	1	32	None
2	Gated Conv	ELU	3	2	1	64	None
3	Gated Conv	ELU	3	1	1	64	None
4	Gated Conv	ELU	3	2	1	128	None
5	Gated Conv	ELU	3	1	1	128	None
6	Gated Conv	ELU	3	1	1	128	None
7	Gated Conv	ELU	3	1	2	128	None
8	Gated Conv	ELU	3	1	4	128	None
9	Gated Conv	ELU	3	1	8	128	None
10	Gated Conv	ELU	3	1	16	128	None
11	Gated Conv	ELU	3	1	1	128	5
12	Gated Conv	ELU	3	1	1	128	4
13	Resize (2x)	n/a	n/a	n/a	n/a	n/a	n/a
14	Gated Conv	ELU	3	1	1	64	3
15	Gated Conv	ELU	3	1	1	64	2
16	Resize (2x)	n/a	n/a	1	n/a	n/a	n/a
17	Gated Conv	ELU	3	1	1	32	1
18	Gated Conv	ELU	3	1	1	16	None
19	Conv	None	3	1	1	3	None
20	Clip	n/a	n/a	n/a	n/a	n/a	n/a

Table 1. The generator architecture. Act. stands for activation type, K stands for kernel size, S for stride, D for dilation, Out for number of channels in convolutional layers and number of units in fully connected units, and Skip represents the layer-id which is concatenated into the output of the given layer. All resize operations use bilinear interpolation. In the Generator, all convolutional layers use 'Same' padding.

1.2. Discriminator Network D

The discriminator applies spectral normalization [5] at all layers, and consists of the the common tower ($\boldsymbol{D_N}$, Table 2), which feeds into the non-conditional branch ($\boldsymbol{f_N}$, Table 3) and projection discriminator branch ($\boldsymbol{f_C}$, Table 4). These two branches produce scalars, which are then summed to produce a single network output. We invite the reader to see Section 3 of the main paper for more in depth discussion of the model.

The scalar outputs of the main and projection discriminator are summed and passed to the adversarial loss.

Common Tower D_N

Layer ID	Type	Act.	K	S	Padding	Out Size
1	Conv	LeakyReLU[4]	5	2	Same	64
2	Conv	LeakyReLU	5	2	Same	128
3	Conv	LeakyReLU	5	2	Same	256
4	Conv	LeakyReLU	5	2	Same	256
5	Conv	LeakyReLU	5	2	Same	256
6	Conv	LeakyReLU	5	2	Same	256
7	Conv	LeakyReLU	5	1	Valid	256
8	Flatten	n/a	n/a	n/a	n/a	n/a

Table 2. The base of the discriminator. It takes generated and ground truth images as input. Act. stands for activation type, K stands for kernel size, S for stride, Out for number of channels in convolutional layers and number of units in fully connected units.

Non-Conditional Branch f_N

Layer ID	Type	Act.	Out Size
1	Fully Connected No Bias	None	1

Table 3. The non-conditional branch of the discriminator, taking the common tower from Table 2 as input and outputting a single scalar value. Act. stands for activation type.

Projection Discriminator Branch f_C

Layer ID	Type	Act.	Out Size
1	Normalize	None	1000
2	Fully Connected No Bias	None	256
3	Inner Product w/Common Tower	None	1

Table 4. The projection discriminator [6] branch of the network. The input is logits of a pretrained classification network, for which we used an InceptionV3 [7] network trained on ImageNet [3]. The output is a single scalar, which is summed with the output of the non-conditional branch and passed to the hinge loss.

1.3. Training details:

We take the training set of Places365-Challenge dataset [9], select the top 50 classes by number of samples, and

create a holdout validation set from this. This creates about 39,000 training and 930 test samples per class, for a total training set size of 1,953,624 and testing set size of 46376. The classes selected are:

- amusement park
- aquarium
- athletic field
- baseball field
- bathroom
- beach
- bridge
- building facade
- car interior
- church - indoor
- church - outdoor
- cliff
- coast
- corridor
- dining room
- embassy
- forest
- forest path
- golf course
- harbor
- highway
- industrial area
- lagoon
- lake
- lighthouse
- living room
- lobby
- mansion
- mountain
- ocean
- office building
- palace
- parking lot
- pier
- pond
- porch
- railroad track
- rainforest
- river
- skyscraper
- stadium
- staircase
- swamp
- swimming hole
- swimming pool
- train station
- underwater
- valley
- vegetable garden
- water park

Before passing the training image into the generator we resize the image to 257 x 257, and also concatenate the mask channel. The mask size is randomly sampled from a uniform distribution, which is the target size plus/minus 4 pixels, so the model doesn't overfit to a specific mask size.

Following the code of DeepFill [8], we concatenate a channel of 1's to the input of the generator. This enables the generator to see the edge of the image after 0 padding the inputs, although we do not verify this in this work.

We take generator and discriminator steps in a 1:1 ratio, with the steps executed jointly.

Please see Section 3 of the main paper for more discussion of the loss and optimizer.

2. Qualitative Results

We show additional samples from on the 25%, 50%, and 75% mask image extension experiments, and refer the reader to Figures 2, 3, and 4. We also show additional results from in-painting experiment in Figure 5 and more panorama results in Figure 6.We also demonstrate the suitability of our method on freeform masks in Figure 1.

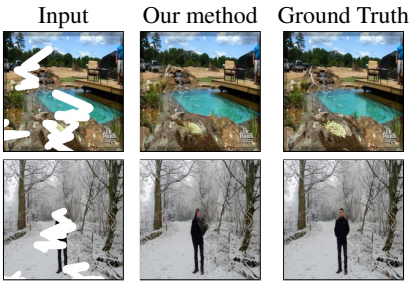

Figure 1. Results on freeform masks.

3. Exploring the Space of Plausible Extensions

We invite the reader to view the accompanying video derived from a sample from the YouTube8m dataset [1] at https://drive.google.com/file/d/1x6FCYPmoqSuCdeLJTD0UpQ_MQhBPv7_e/view?usp=sharing. Please refer to the main paper for details on how it was created. We encourage the reader to pause the video at arbitrary frames to see how the model produces different plausible completions as the result of tiny perturbations of the original frame.

4. Failure Cases

In Figure 7 we examine some of the failure modes of our image extension model. We note that our model is much better at textures than objects; for example vehicles, people, and furniture are challenging for the model. Addressing this is left to future work.

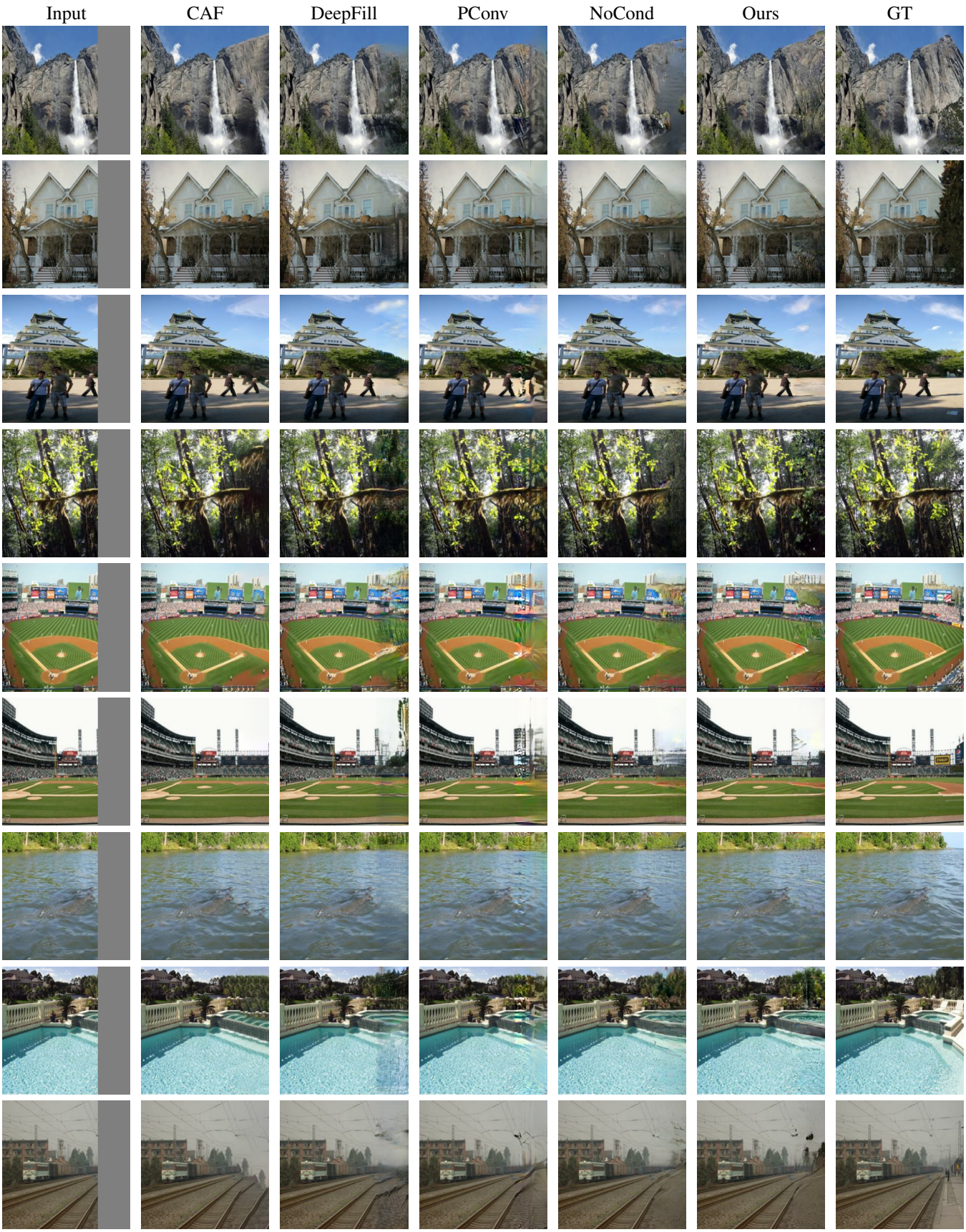

Figure 2. Extending images from masks which are 25% of the image width. We note that edges and structure are better defined in our method. For instance, edge of the roof in the second row.

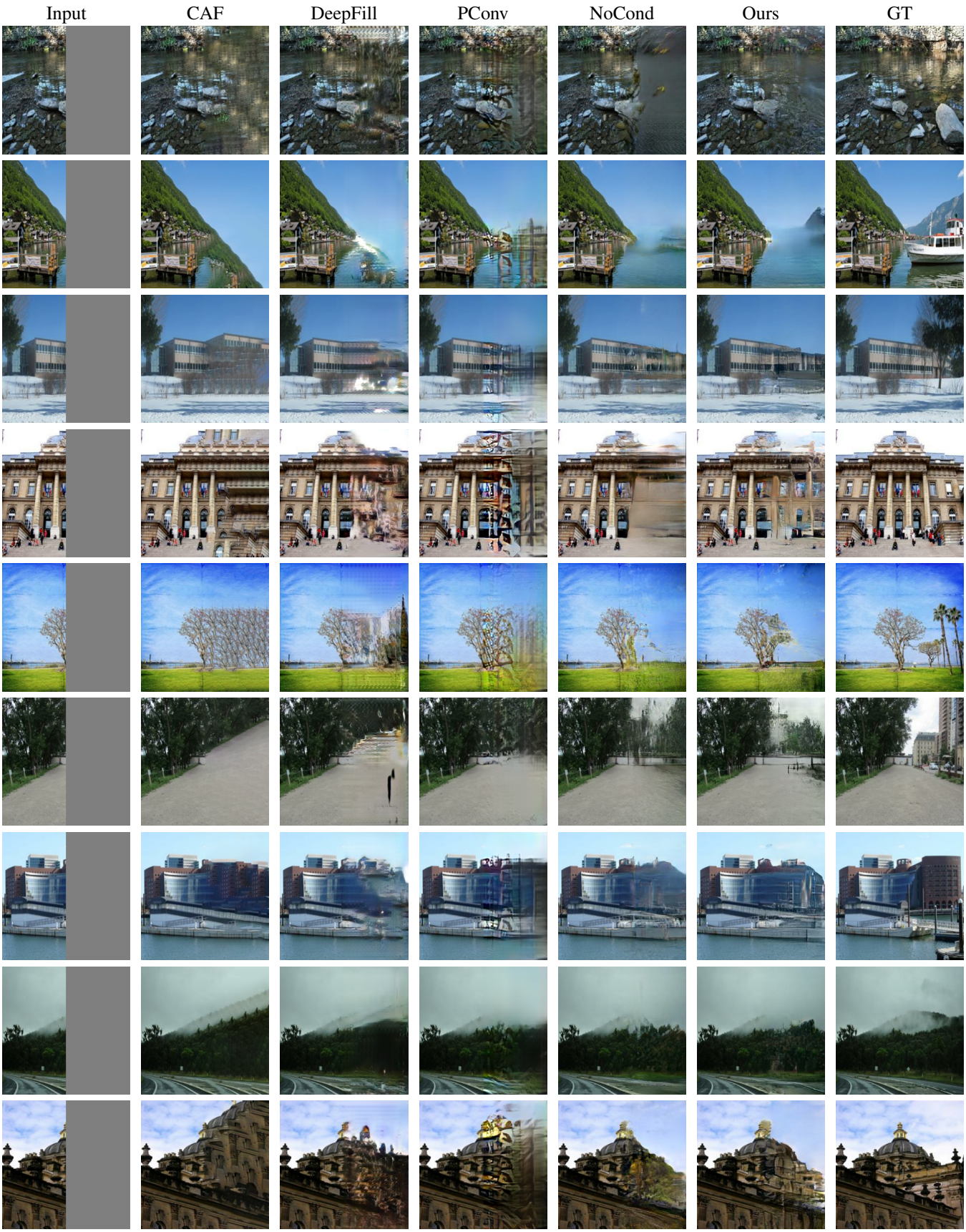

Figure 3. Extending images from masks which are 50% of the image width.

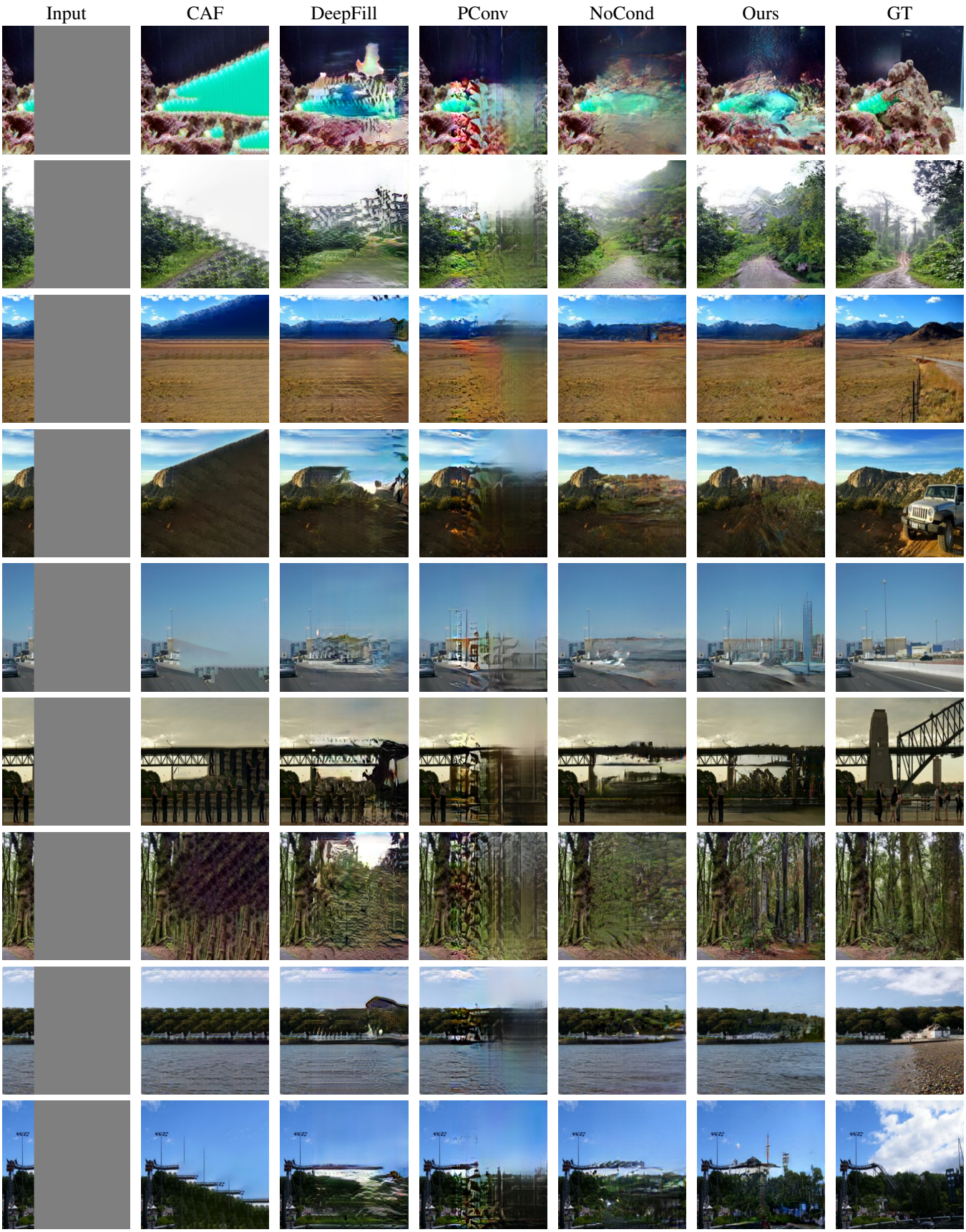

Figure 4. Extending images from masks which are 75% of the image width.

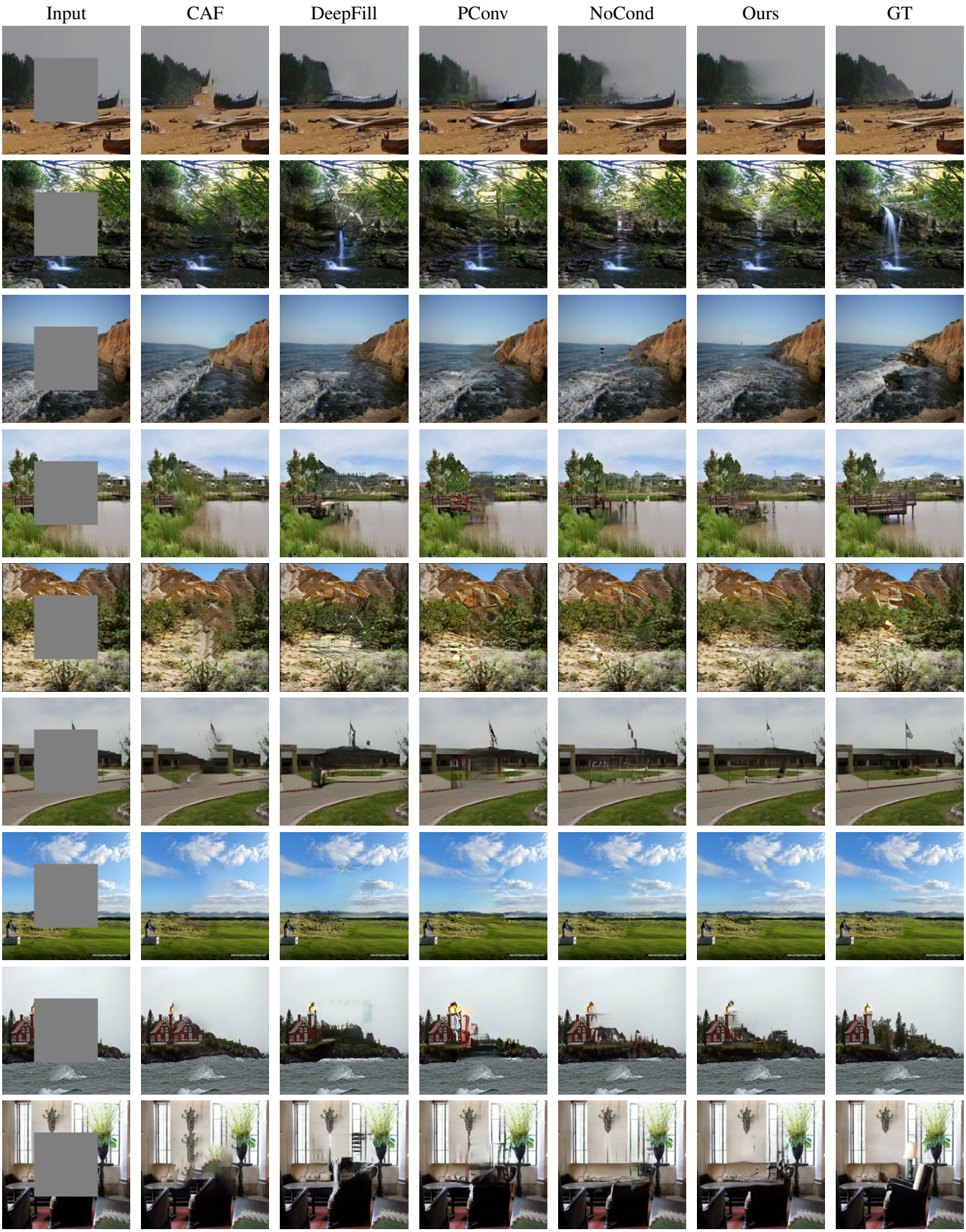

Figure 5. Center Inpainting.

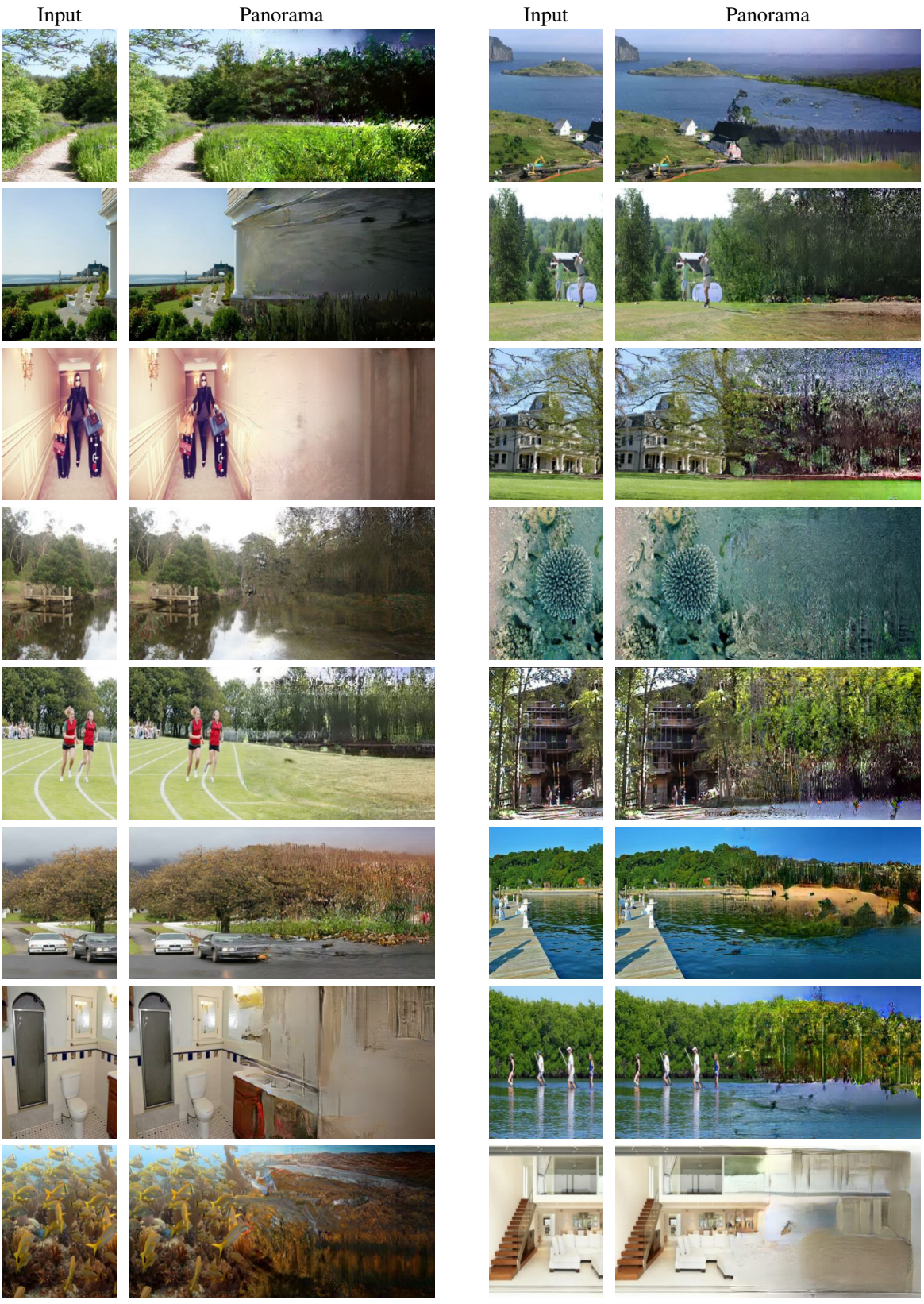

Figure 6. Additional panorama results

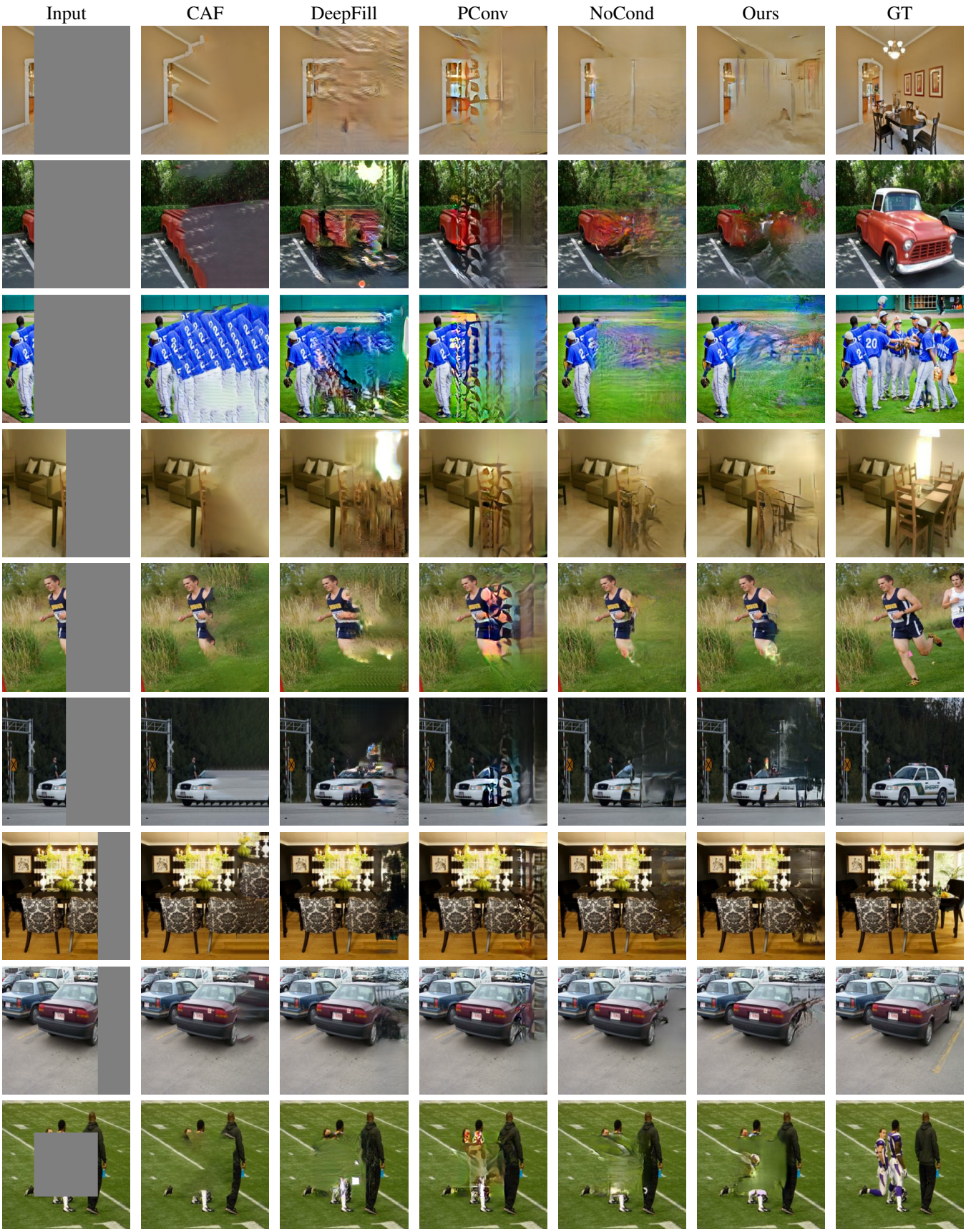

Figure 7. Failure cases. The network struggles with objects; especially cars, humans, and furniture.

References

[1] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016. 2

[2] D.-A. Clevert, T. Unterthiner, and S. Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). In *International Conference on Learning Representations*, 2016. 1

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1

[4] A. L. Maas, A. Y. Hannun, and A. Y. Ng. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, volume 30, page 3, 2013. 1

[5] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018. 1

[6] T. Miyato and M. Koyama. cGANs with projection discriminator. In *International Conference on Learning Representations*, 2018. 1

[7] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 1

[8] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang. Generative image inpainting with contextual attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5505–5514, 2018. 2

[9] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 1