Supplementary Material: Image Inpainting with Cascaded Modulation GAN and Object-Aware Training

Haitian Zheng[1,2], Zhe Lin[2], Jingwan Lu[2], Scott Cohen[2], Eli Shechtman[2], Connelly Barnes[2], Jianming Zhang[2], Ning Xu[2], Sohrab Amirghodsi[2], and Jiebo Luo[1]

[1] University of Rochester
[2] Adobe Research

Appendix

We provided more analysis, visual results and implementation details in the appendix, including analysis of object-aware training (Sec. 1), effect of the masked-R_1 regularization (Sec. 2), visual comparison to other methods (Sec. 3), visual comparison on other types of masks (Sec. 4) and more implementation details (Sec. 5). We also introduce comparisons to the recent transformer-based approach TFill [11]. All visual results are in high-resolution and **best viewed by zoom-in on screen**.

1 Visual Effects of Object-aware Training on Other Models

To analyze the generalization of object-aware training to other recent inpainting methods [4,10] while complementing the numerical results in Table 2 of the main paper, we provide the supplemental visual effect of object-aware training in Figure 1 and Figure 2.

Figure 1 presents the visual comparison of LaMa [4] and CoModGAN [10] trained without or with object-aware training (OT). Object-aware training in general improves other state-of-the-art models including LaMa and CoModGAN on retaining object boundaries and background under the distractor removal scenario.

Furthermore, Figure 2 presents the visual comparison of our method and the state-of-the-art models trained with object-aware training, including LaMa-OT and CoModGAN-OT. CM-GAN with object-aware training achieves better performance than other state-of-the-art models trained with object-aware training, validating the strong generation capacity of CM-GAN.

2 The Effect of The Masked-R_1 Regularization

Figure 3 visualizes the effect of masked-R_1 during the training. Specifically, we visualize the baseline model trained with masked-R_1 regularization (red) and R_1 regularization (orange). The masked-R_1 regularization helps the model achieve lower FID scores, higher discriminator classification loss and makes discriminator harder to distinguish fake samples.

3 Additional Qualitative Results

In the following, we provide the supplementary qualitative results of methods evaluated in Table 1 of the main paper. In addition, we include evaluation of a recent transformer-based approach, i.e. TFill [11] in Section 3.2.

Fig. 1: **The effect of object-aware training on other models.** Object-aware training (OT) improves other models including LaMa [4] and CoModGAN [10] on achieving sharper boundaries and clearer background. **Best viewed by zoom-in on screen.**

Fig. 2: **Results of CM-GAN in comparison to LaMa [4] and CoModGAN [10] trained with object-aware training (OT). Best viewed by zoom-in on screen.**

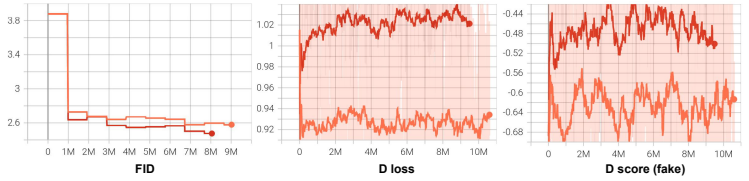

Fig. 3: **The convergence curves of the baseline models trained with masked-R_1 regularization (red) and R_1 regularization (orange).** The masked-R_1 regularization help the model achieve lower FID scores, higher discriminator classification loss and makes the discriminator harder to distinguish fake samples.

3.1 More Visual Comparisons with ProFill, LaMa, and CoModGAN

We present additional visual comparisons to ProFill [9], LaMa [4] and CoModGAN [10] in Figures 4 to 7 to supplement Figure 5 of the main paper.

3.2 Visual Comparisons to Transformer-based Methods

CM-GAN is based on the GAN framework. With transformers becoming popular in computer vision, several recent works [3,5,11] leverage transformer-based architectures for inpainting. In this section, we present the visual comparison to DS [3], ICT [5] and the recently proposed TFill [11] to analyze the visual quality of those approaches. As observed in Figure 8, CM-GAN achieves consistently better visual results in terms of holistic structures and local textures, which is coherent to the FID scores reported in Table 1 of the main paper [3].

3.3 Visual Comparisons to Other Remaining Methods

Figure 9 presents the visual comparisons of CM-GAN to other methods including Edge-Connect [2], MEDEF [1], DeepFillv2 [7], HiFill [6], CRFill [8]. Our method achieves substantially better visual quality than all these compared methods.

4 Visual Comparisons on Other Types of Masks

To supplement Table 3 of the main paper, we provide the visual comparisons on the LaMa mask [4] in Figure 10 and CoModGAN mask [10] in Figure 11, respectively. Consistent to the numerical result in the main paper, our method generates better global structure and more coherent textures than others, demonstrating robustness of CM-GAN on different mask types.

5 Implementation Details

5.1 The Details of Spatial Modulation

We provide the detail implementation of the Affine Parameters Networks (APN) and our spatial modulation operation in pseudo code.

[3] The FID score of TFill is 7.435.

Fig. 4: **Results of CM-GAN in comparison to ProFill [9], LaMa [4] and CoModGAN [10].** **Best viewed by zoom-in on screen.**

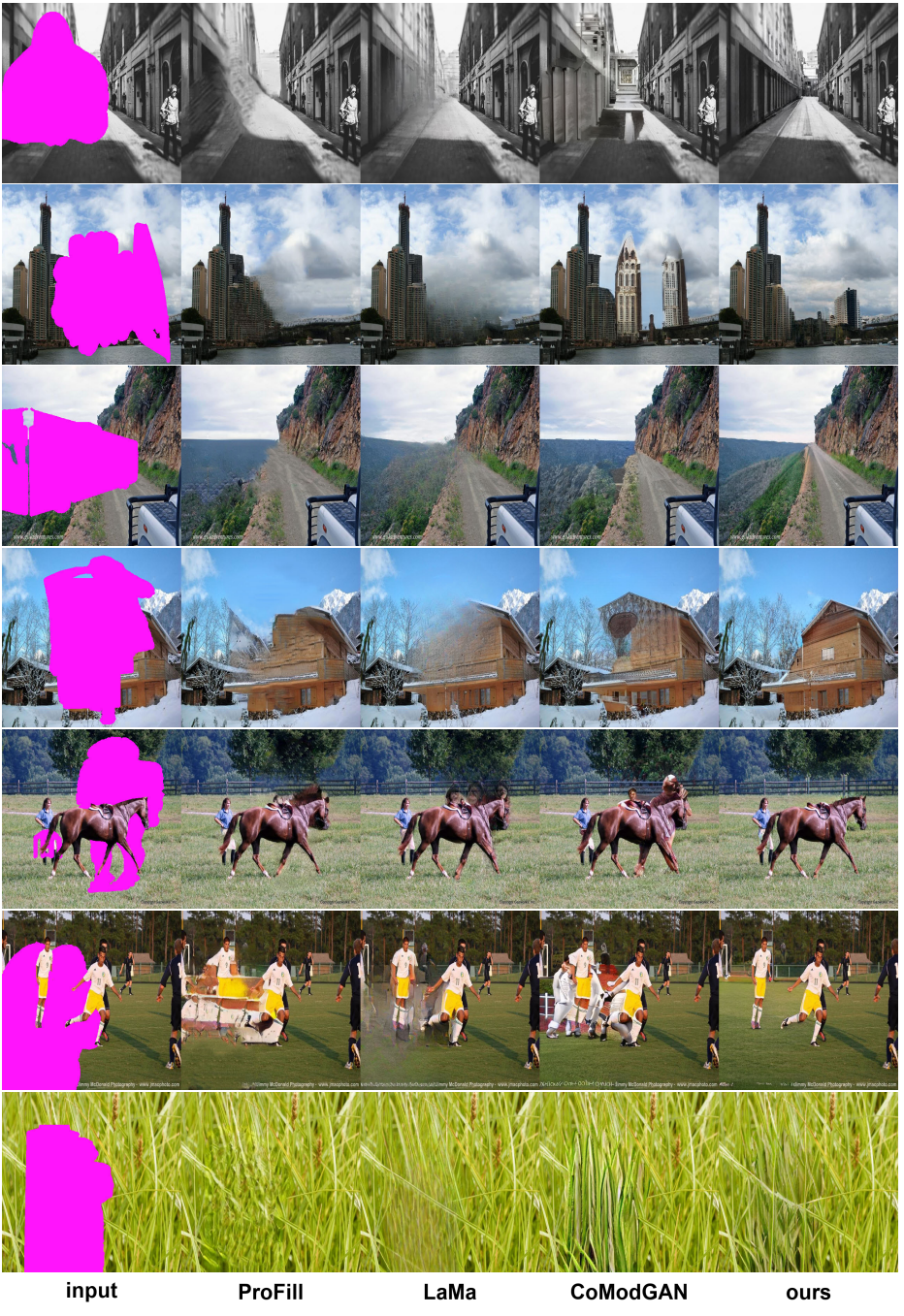

Fig. 5: **Results of CM-GAN in comparison to ProFill [9], LaMa [4] and CoModGAN [10].** **Best viewed by zoom-in on screen.**

Fig. 6: **Results of CM-GAN in comparison to ProFill [9], LaMa [4] and CoModGAN [10].** **Best viewed by zoom-in on screen.**

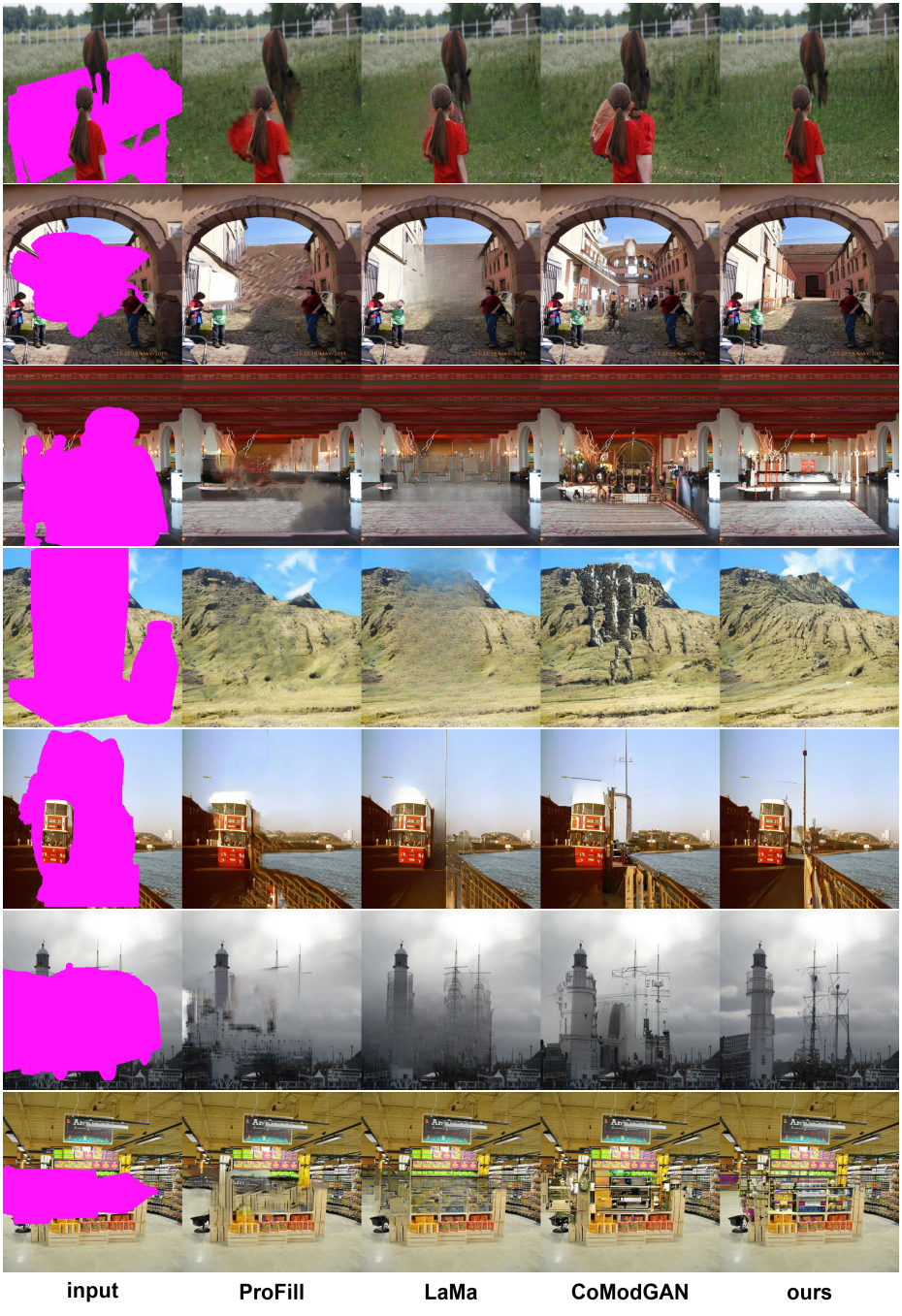

Fig. 7: **Results of CM-GAN in comparison to ProFill [9], LaMa [4] and CoModGAN [10].** **Best viewed by zoom-in on screen.**

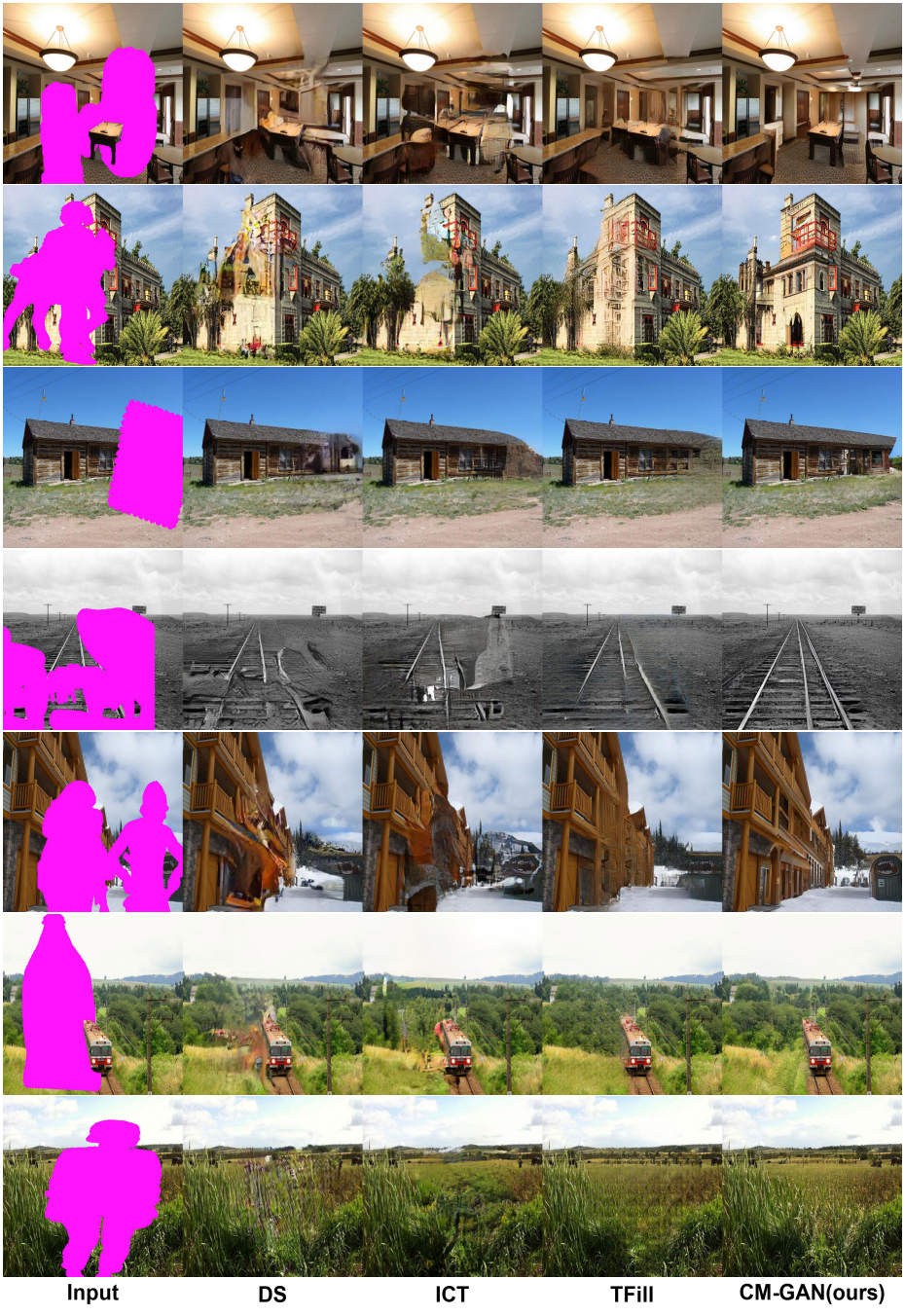

Fig. 8: **Results of CM-GAN in comparison to transformer-based approaches including DS [3], ICT [5] and TFill [11]. Best viewed by zoom-in on screen.**

Fig. 9: **Results of CM-GAN in comparison to other methods including EdgeConnect [2], MEDEF [1], DeepFillv2 [7], HiFill [6], CRFill [8]. Best viewed by zoom-in on screen.**

Fig. 10: **Visual comparison on the mask of LaMa [4]. Best viewed by zoom-in on screen.**

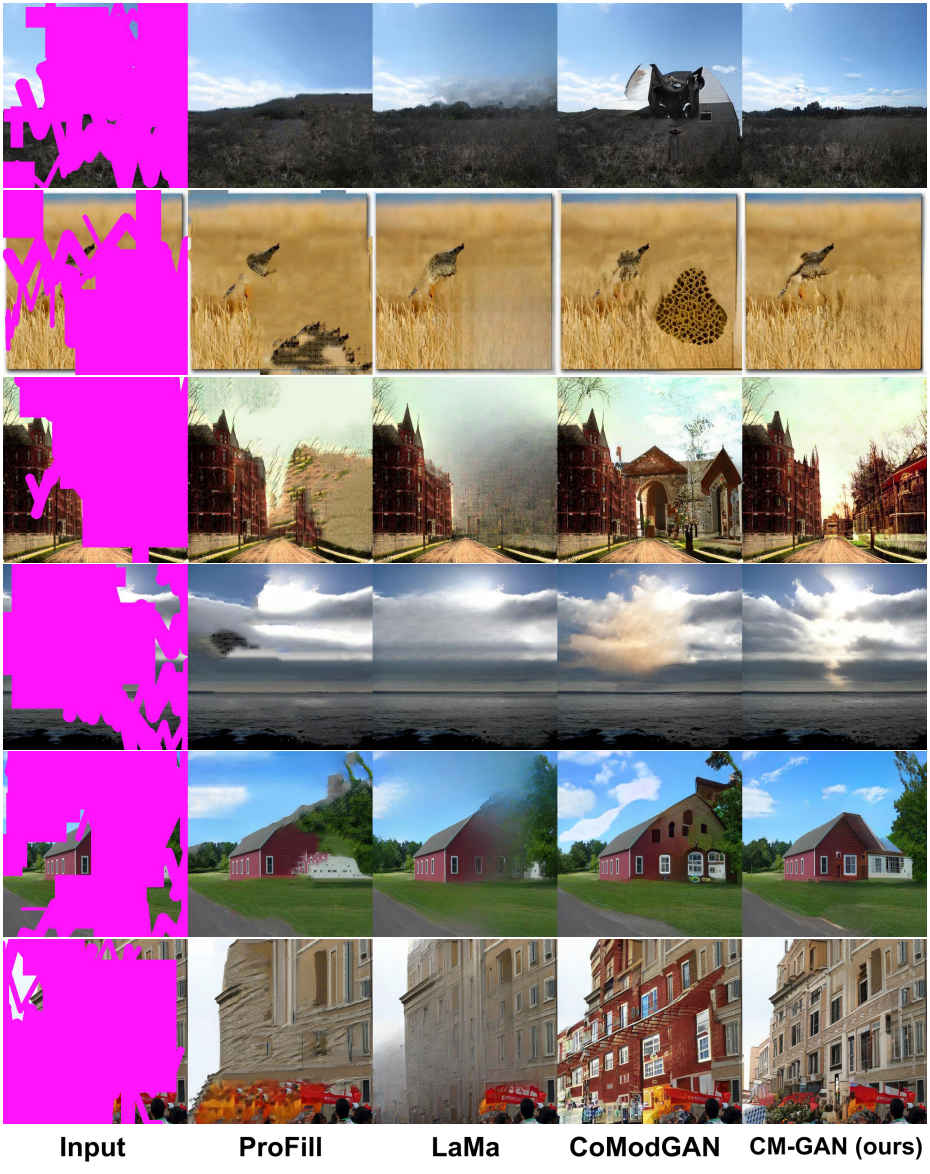

Fig. 11: **Visual comparison on the mask of CoModGAN [10]. Best viewed by zoom-in on screen.**

The Affine Parameters Network (APN). The affine parameters network (APN) is implemented as a stack of convolutional layer that takes X as input to generate scaling parameters A and shifting parameters B.

```
def APN(X):
    # the 1x1 input layer
    t1 = self.conv1_1x1(X)
    # the 3x3+1x1 middle layer
    t2 = self.conv2_3x3(t1)
    t2 = t2 + self.conv2_1x1(t1)
    # the 1x1 output layer
    A = self.conv_A_1x1(t)
    B = self.conv_B_1x1(t)
    return A, B
```

Spatial Modulation. Next, the spatial modulation takes feature maps X, Y, global code $\boldsymbol{g}$, the convolutional kernel weight $\boldsymbol{w}$ and the noise $\boldsymbol{n}$ as inputs to modulate Y:

```
import torch.nn.functional as F
def spatial_mod_ops(X, Y, g, w, noise):
    bs = X.size(0) # batch size
    # predicting the spatial code
    A0, B = self.APN(X)
    # merging with the global code
    A = A0 + self.fc(g).reshape(bs,-1,1,1)
    # spatial modulation
    Y = Y.mul(A)
    # convolution
    Y = F.conv2d(Y, w)
    # spatially-aware demodulation
    w = w.unsqueeze(0)
    A_avg_var = A.square().mean([2,3]).reshape(bs,1,-1,1,1)
    D = (w.square().mul(A_avg_var).sum(dim=[2,3,4]) + 1e-8).rsqrt()
    Y = Y.mul(D.reshape(bs, -1, 1, 1))
    # adding bias and noise
    Y = Y + B + noise
    return Y
```

5.2 Details of the Object-Aware Mask Generation Procedure

Our object-aware mask generation scheme is based on the following pipeline to sample a mask for image x:

1. *Generating the initial mask.* We sample an initial mask $\boldsymbol{m}$ with either irregular masks proposed by [10], object masks proposed by [9] or random overlapping rectangle with probablity 0.45, 0.45 and 0.1, respectively. We further augment the object mask by random circular translation and dilate the mask using random width.
2. *Occluding foreground instance.* For each object instance $\boldsymbol{s}_i$ from image x, we compute the overlapping ratio $r_i = \text{Area}(\boldsymbol{m}, \boldsymbol{s}_i)/\text{Area}(\boldsymbol{s}_i)$ between the instance and the initial mask. If the overlapping ratio r_i is larger than 0.5, we exclude instance

$\boldsymbol{s}_i$ from the initial mask $\boldsymbol{m}$, namely $\boldsymbol{m} \leftarrow \boldsymbol{m} - \boldsymbol{s}_i$, to mimic the distractor removal use case.

3. *Rejecting small mask.* If the area of the mask is less than 0.05 of the area of the entire image, we repeating the sampling procedure, until the maximal sampling iteration 5 is reached. Otherwise, the sampled mask is returned.

References

1. Hongyu Liu, Bin Jiang, Y.S.W.H., Chao, Y.: Rethinking image inpainting via a mutual encoder-decoder with feature equalizations. In: Proceedings of the European Conference on Computer Vision (2020) 4, 10
2. Nazeri, K., Ng, E., Joseph, T., Qureshi, F.Z., Ebrahimi, M.: Edgeconnect: Generative image inpainting with adversarial edge learning. arXiv preprint arXiv:1901.00212 (2019) 4, 10
3. Peng, J., Liu, D., Xu, S., Li, H.: Generating diverse structure for image inpainting with hierarchical vq-vae. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10775–10784 (2021) 4, 9
4. Suvorov, R., Logacheva, E., Mashikhin, A., Remizova, A., Ashukha, A., Silvestrov, A., Kong, N., Goka, H., Park, K., Lempitsky, V.: Resolution-robust large mask inpainting with fourier convolutions. arXiv preprint arXiv:2109.07161 (2021) 1, 2, 3, 4, 5, 6, 7, 8, 11
5. Wan, Z., Zhang, J., Chen, D., Liao, J.: High-fidelity pluralistic image completion with transformers. arXiv preprint arXiv:2103.14031 (2021) 4, 9
6. Yi, Z., Tang, Q., Azizi, S., Jang, D., Xu, Z.: Contextual residual aggregation for ultra high-resolution image inpainting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7508–7517 (2020) 4, 10
7. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Free-form image inpainting with gated convolution. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4471–4480 (2019) 4, 10
8. Zeng, Y., Lin, Z., Lu, H., Patel, V.M.: Cr-fill: Generative image inpainting with auxiliary contextual reconstruction. In: Proceedings of the IEEE International Conference on Computer Vision (2021) 4, 10
9. Zeng, Y., Lin, Z., Yang, J., Zhang, J., Shechtman, E., Lu, H.: High-resolution image inpainting with iterative confidence feedback and guided upsampling. arXiv preprint arXiv:2005.11742 (2020) 4, 5, 6, 7, 8, 13
10. Zhao, S., Cui, J., Sheng, Y., Dong, Y., Liang, X., Chang, E.I., Xu, Y.: Large scale image completion via co-modulated generative adversarial networks. arXiv preprint arXiv:2103.10428 (2021) 1, 2, 3, 4, 5, 6, 7, 8, 12, 13
11. Zheng, C., Cham, T.J., Cai, J.: Tfill: Image completion via a transformer-based architecture. arXiv preprint arXiv:2104.00845 (2021) 1, 4, 9