Supplementary material: Image Inpainting via Generative Multi-column Convolutional Neural Networks

1 Network Architectures

We give more details of our network architecture. In the following, K denotes kernel size, S denotes stride size, C denotes channel number, and D denotes dilation ratio. Our Generative Multi-column Convolutional Neural Network (**GMCNN**) has three encoder branches denoted as B1, B2 and B3, and one decoding module. The outputs of the three encoder branches are denoted as EB1, EB2, and EB3. Their specifications are as follows.

B1: K7S1C32 - K7S2C64 - K7S1C64 - K7S2C128 - K7S1C128 - K7S1C128 - K7D2S1C128 - K7D4S1C128 - K7D8S1C128 - K7D16S1C128 - K7S1C128 - K7S1C128 - upsampling($4\times$) - EB1.
B2: K5S1C32 - K5S2C64 - K5S1C64 - K5S2C128 - K5S1C128 - K5S1C128 - K5D2S1C128 - K5D4S1C128 - K5D8S1C128 - K5D16S1C128 - K5S1C128 - K5S1C128 - upsampling($2\times$) - K5S1C64 - K5S1C64 - upsampling($2\times$) - EB2.
B3: K3S1C32 - K3S2C64 - K3S1C64 - K3S2C128 - K3S1C128 - K3S1C128 - K3D2S1C128 - K3D4S1C128 - K3D8S1C128 - K3D16S1C128 - K3S1C128 - K3S1C128 - upsampling($2\times$) - K5S1C64 - K5S1C64 - upsampling($2\times$) - K3S1C64 - K3S1C64 - EB3.

Decoding Module: Concate([EB1, EB2, EB3]) - K3S1C16 - K3S1C3 - clipped to [-1, 1].

Local Discriminator: K5S2C64 - K5S2C128 - K5S2C256 - K5S2C512 - K5S2C256 - K5S2C128 - FC (output channel is 1).

Global Discriminator: K5S2C64 - K5S2C128 - K5S2C256 - K5S2C512 - K5S2C256 - K5S2C128 - FC (output channel is 1).

2 A Simple Example about How ID-MRF Works

Given two 2-dimension feature sets $\mathbf{X}$ and $\mathbf{Y}$, which are distributed as in Figure 1. Suppose $h = 0.5$ and $\epsilon = 1e - 5$, the ID-MRF losses between $\mathbf{X}$ and $\mathbf{Y}$ are 0.6012, 0.3754, and 0.2815 for those in Figures 1(a), 1(b), and 1(c), respectively. As explained in the paper, Figure 1 with our computed ID-MRF loss manifests that the smaller ID-MRF loss between $\mathbf{X}$ and $\mathbf{Y}$ is, the more diverse the nearest neighbors found in $\mathbf{Y}$ of each feature in $\mathbf{X}$ are. Thus ID-MRF regularization can diversify the searching results.

3 About Comparison and Corresponding Results

We conduct pairwise comparisons with CE [1], MSNPS [2], and CA [3]. The corresponding codes are provided by the authors. Specifically, the results of CE and MSNPS on Paris street view and ImageNet are tested with provided codes and models. For CA, the inpainting results on Places2, ImageNet, CelebA, and CelebA-HQ are also produced by their codes and models. However, their model on Pairs street view dataset is trained by ourselves with released codes. We first train its coarse network. After it converges, we train the full model using the default hyper-parameters.

32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.

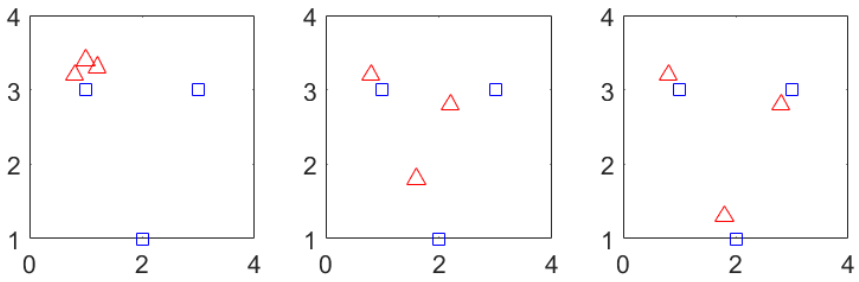


Figure 1: The feature distribution visualization of two 2-dimension feature sets $\mathbf{X}$ (red triangles) and $\mathbf{Y}$ (blue squares). The features here are just coordinates in the figure.

4 More Ablation Study

More visual comparisons on using different structures and receptive fields (Figures 2 and 3), different reconstruction losses (Figures 4 and 5), and the effect of ID-MRF regularization (Figures 6 and 7) are given below.

5 More Inpainting Results

5.1 Visual comparisons with Other Methods

We give more comparisons with methods of [1–3] on Pairs street view (Figures 8, 9, 10, and 11), ImageNet (Figures 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, and 24), Places2 (with 256×256 images) (Figures 25, 26, and 27), Places2 (with 512×680 images) (Figures 28, 29, 30, 31, and 32) CelebA (Figure 33), and CelebA-HQ (Figures 12 and 13).

5.2 Inpainting for Multiple Random Masks

More examples for filling multiple random regions on Paris street view (Figure 34), CelebA-HQ (Figures 35, 36, 37, 38, and 39), and Places2 (Figures 28, 29, 30, 31, and 32) are given.

5.3 Filling 256×256 Central Region for 512×512 Images on CelebA-HQ

Examples about completing 256×256 regions for faces are given in Figures 40, 41, and 42. The model uses the same structure described in our paper and yet the numbers of all convolutional filters are doubled.

(a) (b) (c) (d) (e)

Figure 2: Visual comparison of CNNs with different structures. (a) Input image. (b) Single encoder-decoder. (c) Coarse-to-fine structure [3]. (d) GMCNN with the fixed receptive field in all branches. (e) GMCNN with varied receptive fields.

(a) (b) (c) (d) (e)

Figure 3: Visual comparison of CNNs with different structures. (a) Input image. (b) Single encoder-decoder. (c) Coarse-to-fine structure [3]. (d) GMCNN with the fixed receptive field in all branches. (e) GMCNN with varied receptive fields.

Figure 4: Visual comparisons of different reconstruction losses. (a) Input image. (b) Spatial discounted loss [3]. (c) Confidence-driven reconstruction loss.

Figure 5: Visual comparisons of different reconstruction losses. (a) Input image. (b) Spatial discounted loss [3]. (c) Confidence-driven reconstruction loss.

Figure 6: Visual comparison of results using ID-MRF and not with it. (a) Input image. (b) with ID-MRF. (c) without ID-MRF.

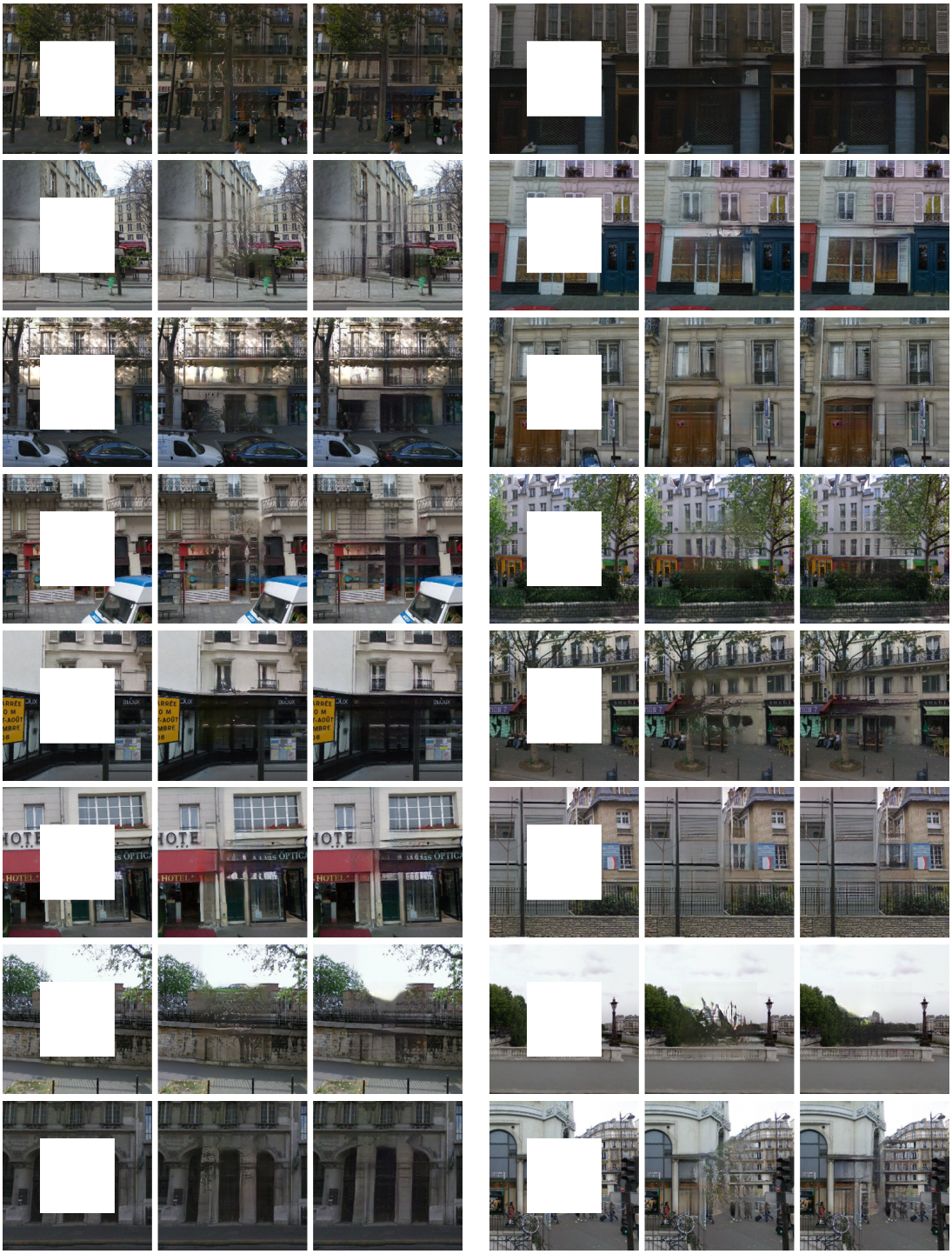


Figure 7: Visual comparison of results using ID-MRF and not with it. (a) Input image. (b) with ID-MRF. (c) without ID-MRF.

Figure 8: Visual comparisons on Paris street view. (a) Input image. (b) CE [1], (c) MSNPS [2]. (d) CA [3]. (e) Our results. Best viewed with zoom-in.

(a) (b) (c) (d) (e)

Figure 9: Visual comparisons on Paris street view. (a) Input image. (b) CE [1], (c) MSNPS [2]. (d) CA [3]. (e) Our results. Best viewed with zoom-in.

(a) (b) (c) (d) (e)

Figure 10: Visual comparisons on Paris street view. (a) Input image. (b) CE [1], (c) MSNPS [2]. (d) CA [3]. (e) Our results. Best viewed with zoom-in.

Figure 11: Visual comparisons on Paris street view. (a) Input image. (b) CE [1], (c) MSNPS [2]. (d) CA [3]. (e) Our results. Best viewed with zoom-in.

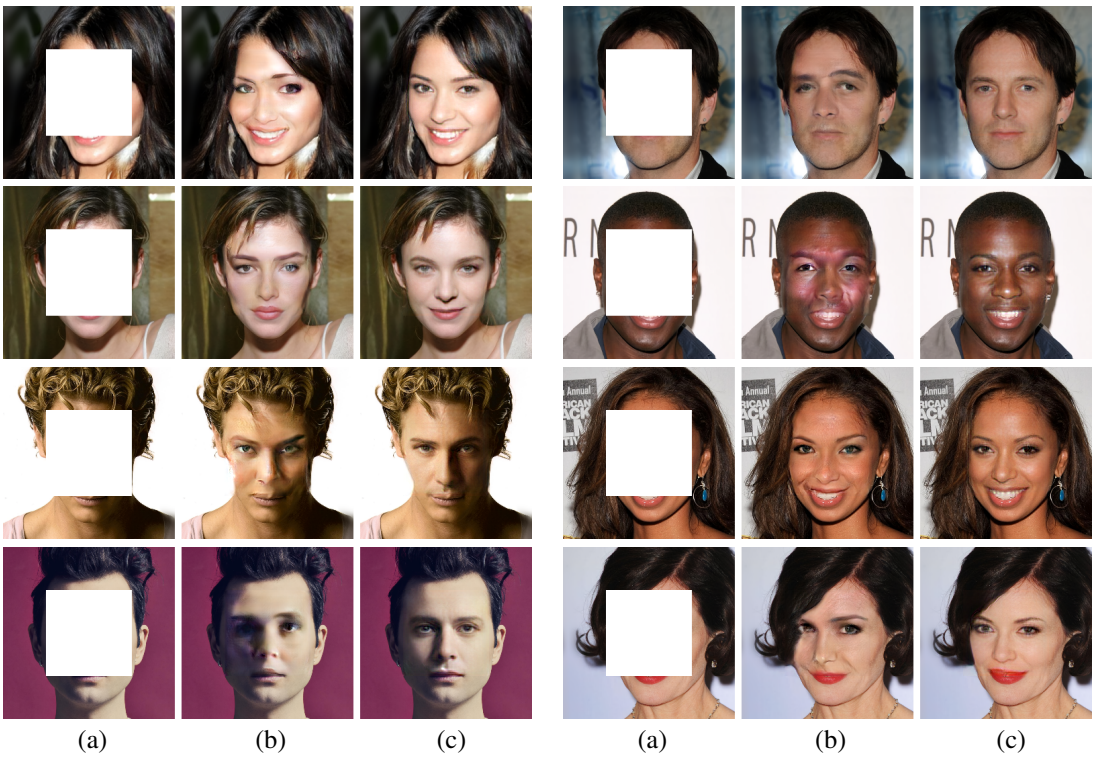


Figure 12: Visual comparisons on CelebA-HQ. (a) Input image. (b) CA [3]. (c) Our results. Best viewed with zoom-in.

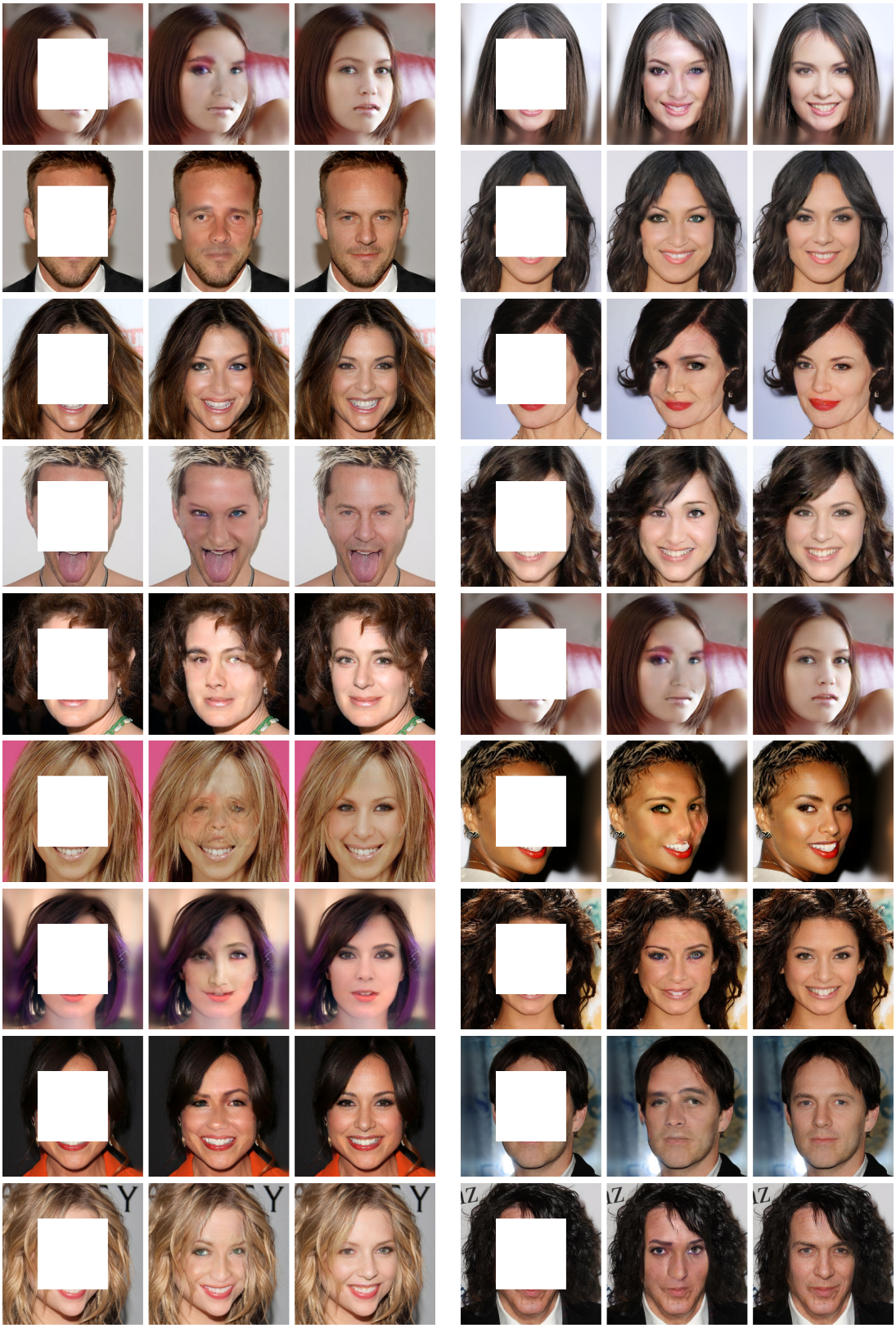


Figure 13: Visual comparisons on CelebA-HQ. (a) Input image. (b) CA [3]. (c) Our results. Best viewed with zoom-in.

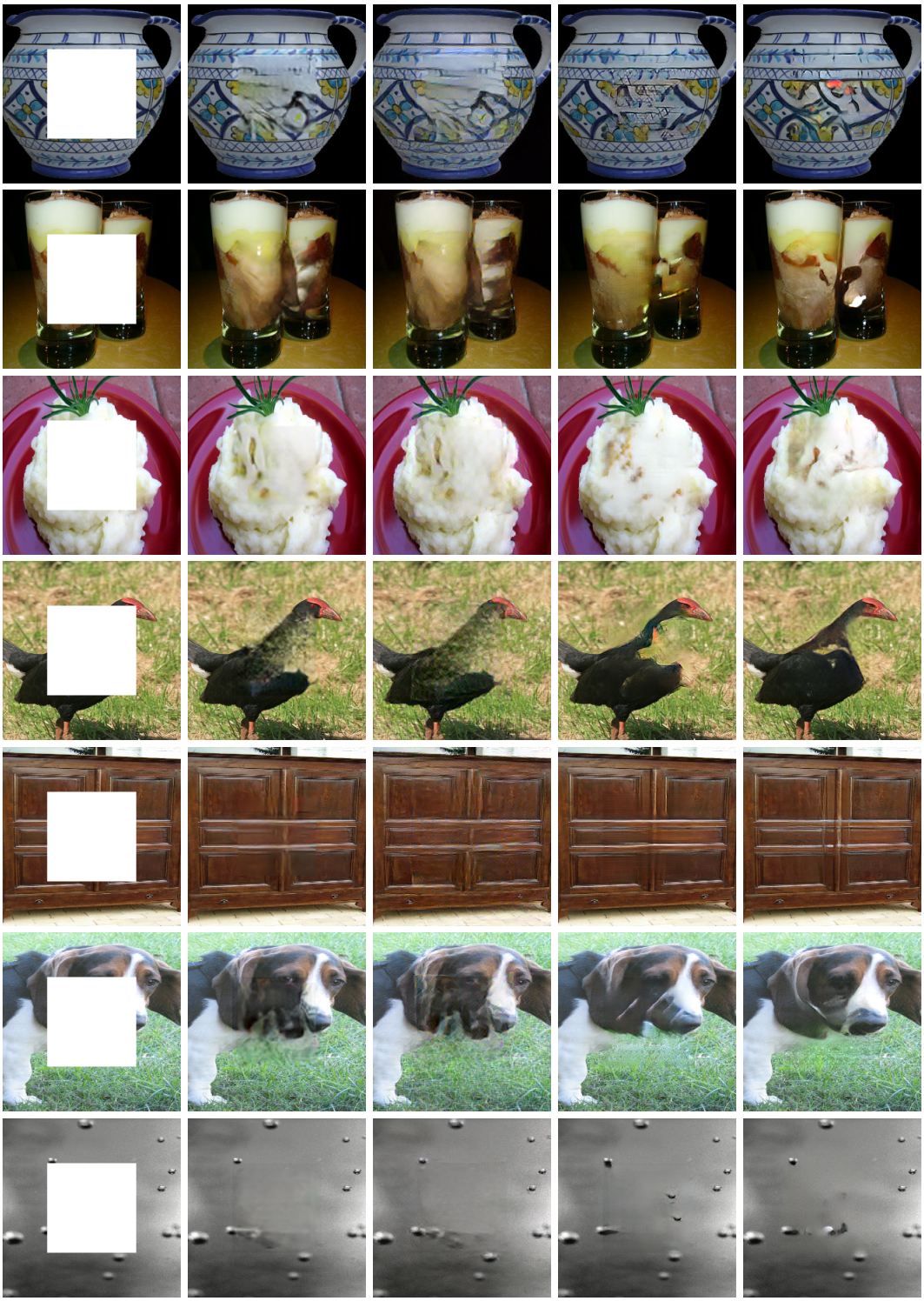


Figure 14: Visual comparisons on ImageNet. (a) Input image. (b) CE [1]. (c) MSNPS [2]. (d) CA [3]. (e) Our results. Best viewed with zoom-in.

(a) (b) (c) (d) (e)

Figure 15: Visual comparisons on ImageNet. (a) Input image. (b) CE [1]. (c) MSNPS [2]. (d) CA [3]. (e) Our results. Best viewed with zoom-in.

(a) (b) (c) (d) (e)

Figure 16: Visual comparisons on ImageNet. (a) Input image. (b) CE [1]. (c) MSNPS [2]. (d) CA [3]. (e) Our results. Best viewed with zoom-in.

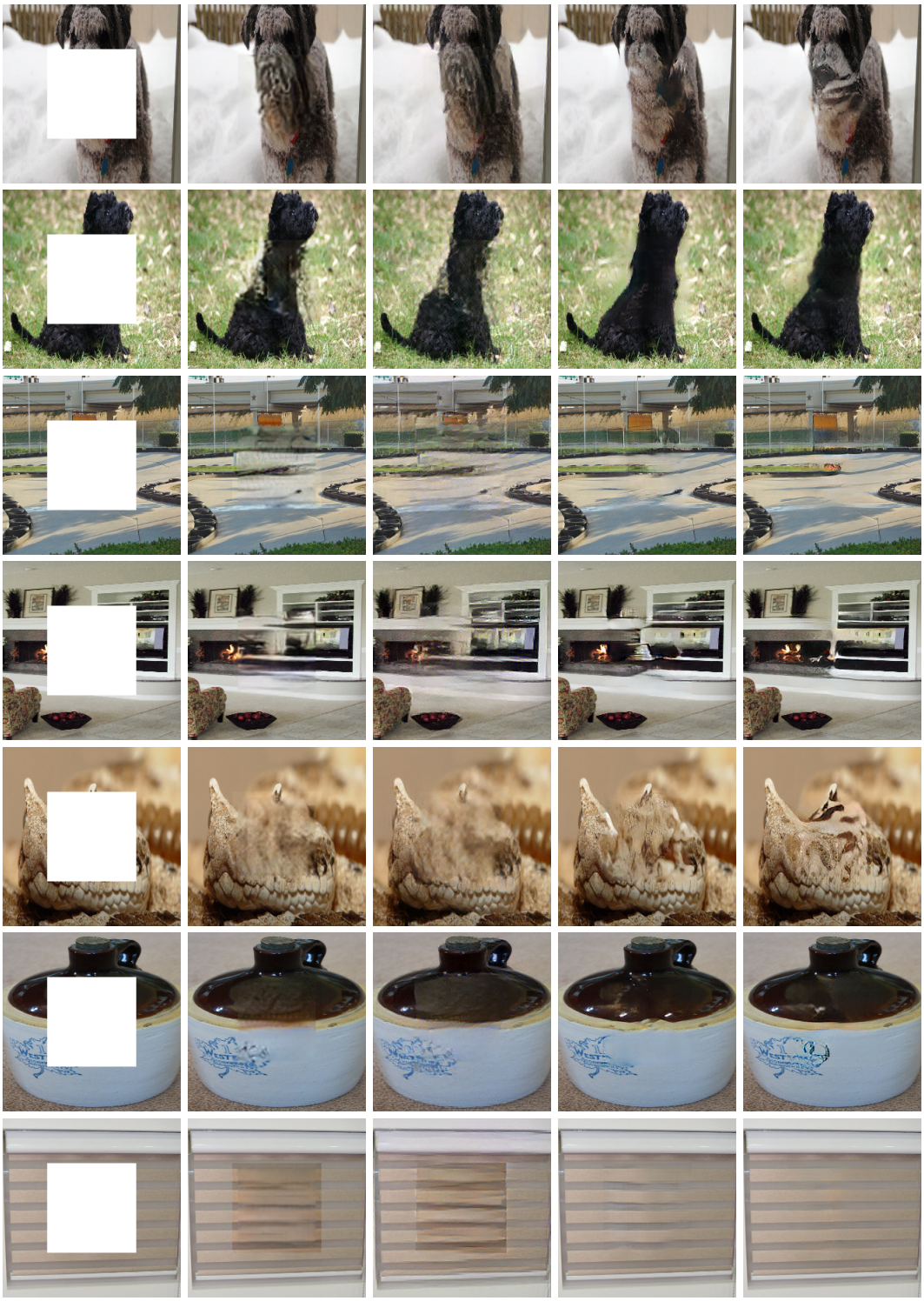
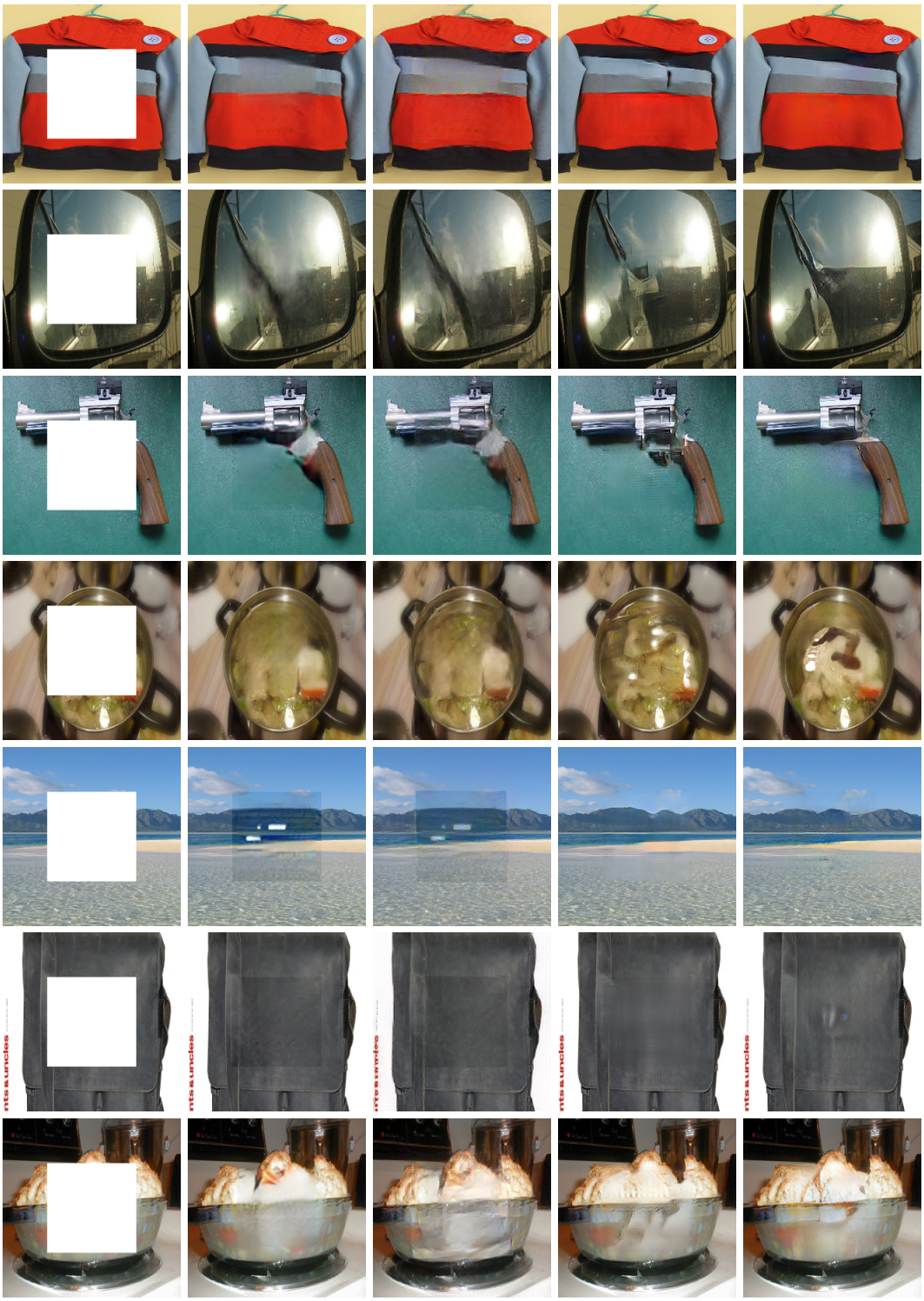


Figure 17: Visual comparisons on ImageNet. (a) Input image. (b) CE [1]. (c) MSNPS [2]. (d) CA [3]. (e) Our results. Best viewed with zoom-in.



(a) (b) (c) (d) (e)

Figure 18: Visual comparisons on ImageNet. (a) Input image. (b) CE [1]. (c) MSNPS [2]. (d) CA [3]. (e) Our results. Best viewed with zoom-in.

(a) (b) (c) (d) (e)

Figure 19: Visual comparisons on ImageNet. (a) Input image. (b) CE [1]. (c) MSNPS [2]. (d) CA [3]. (e) Our results. Best viewed with zoom-in.

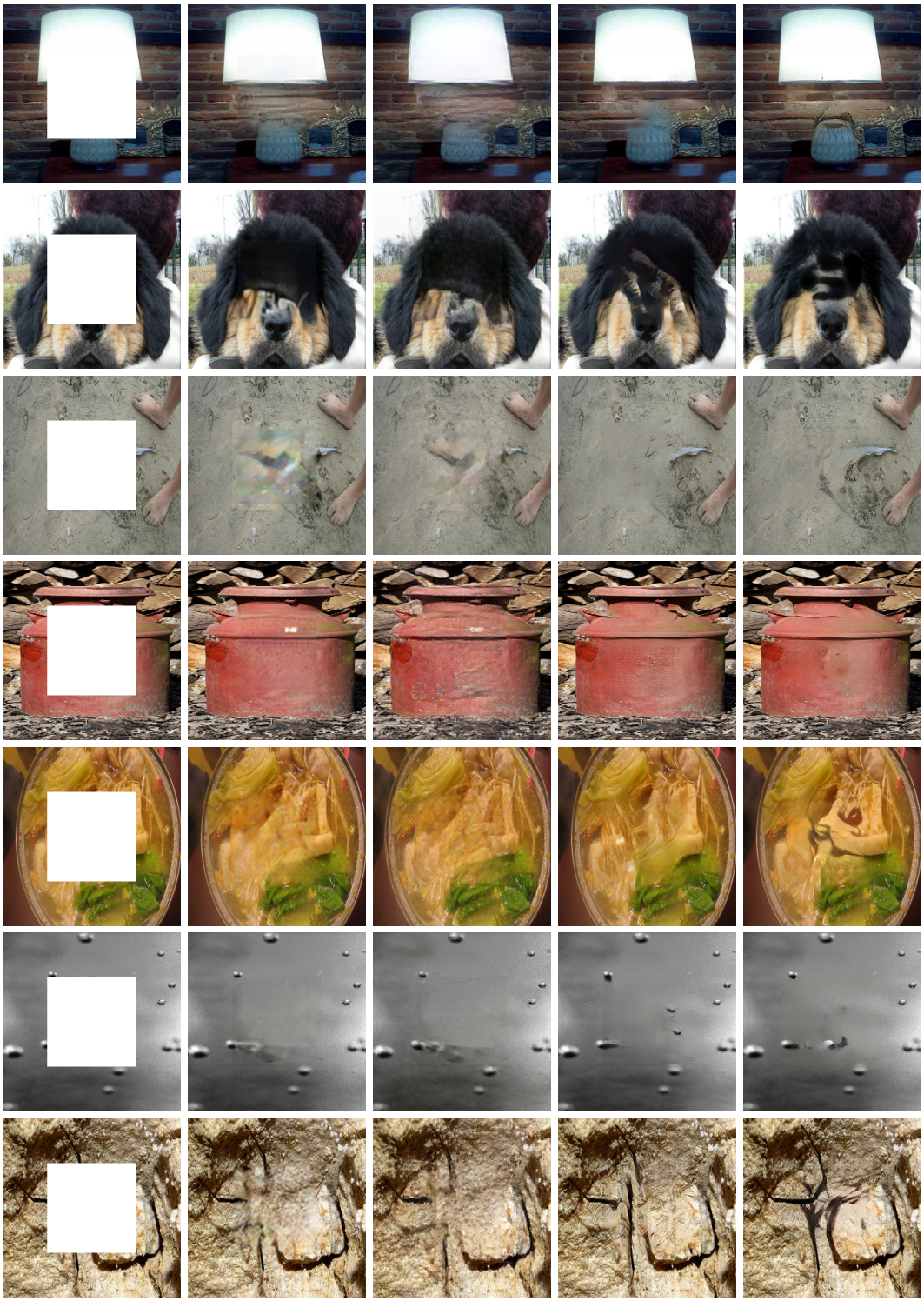


Figure 20: Visual comparisons on ImageNet. (a) Input image. (b) CE [1]. (c) MSNPS [2]. (d) CA [3]. (e) Our results. Best viewed with zoom-in.

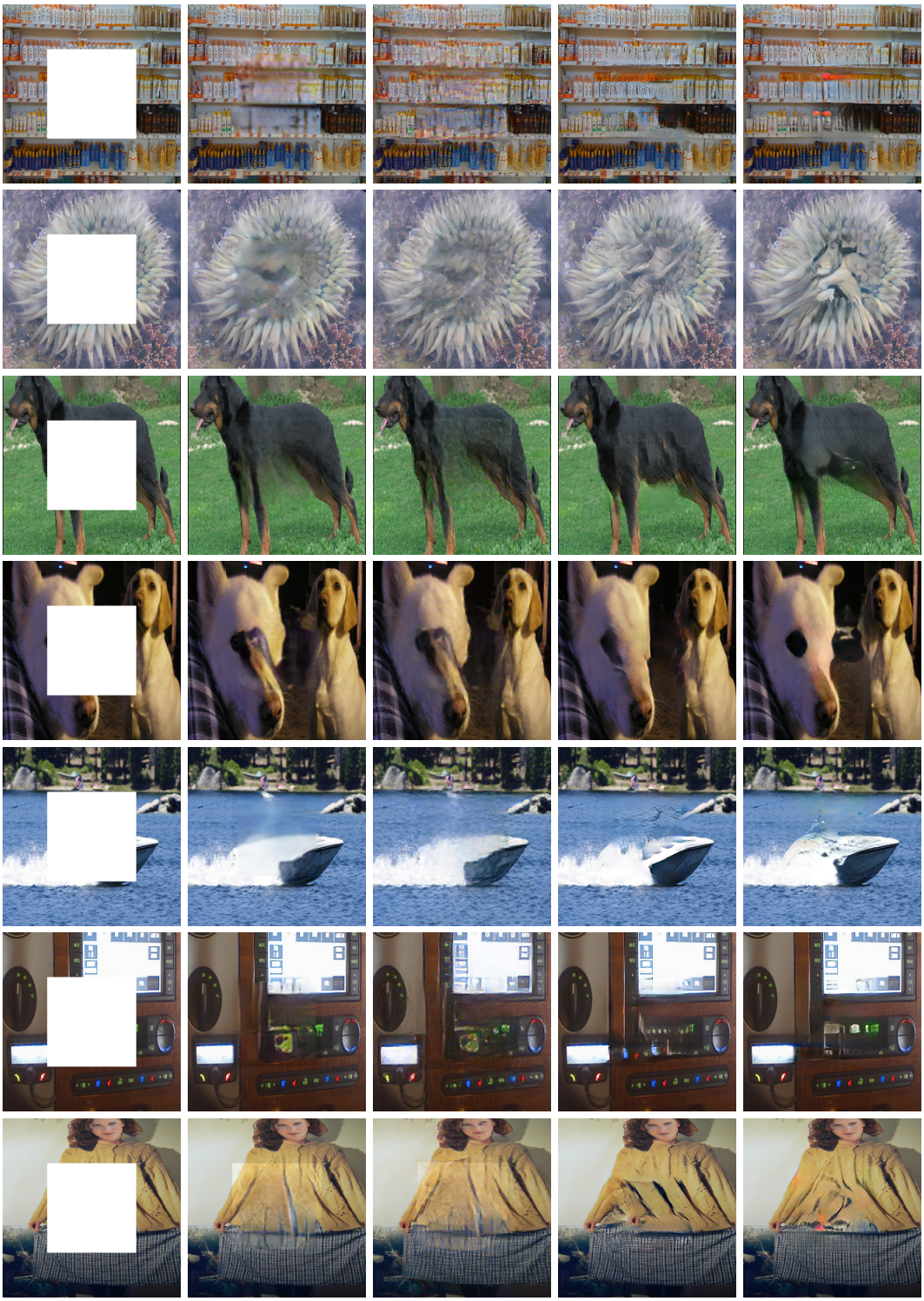


Figure 21: Visual comparisons on ImageNet. (a) Input image. (b) CE [1]. (c) MSNPS [2]. (d) CA [3]. (e) Our results. Best viewed with zoom-in.

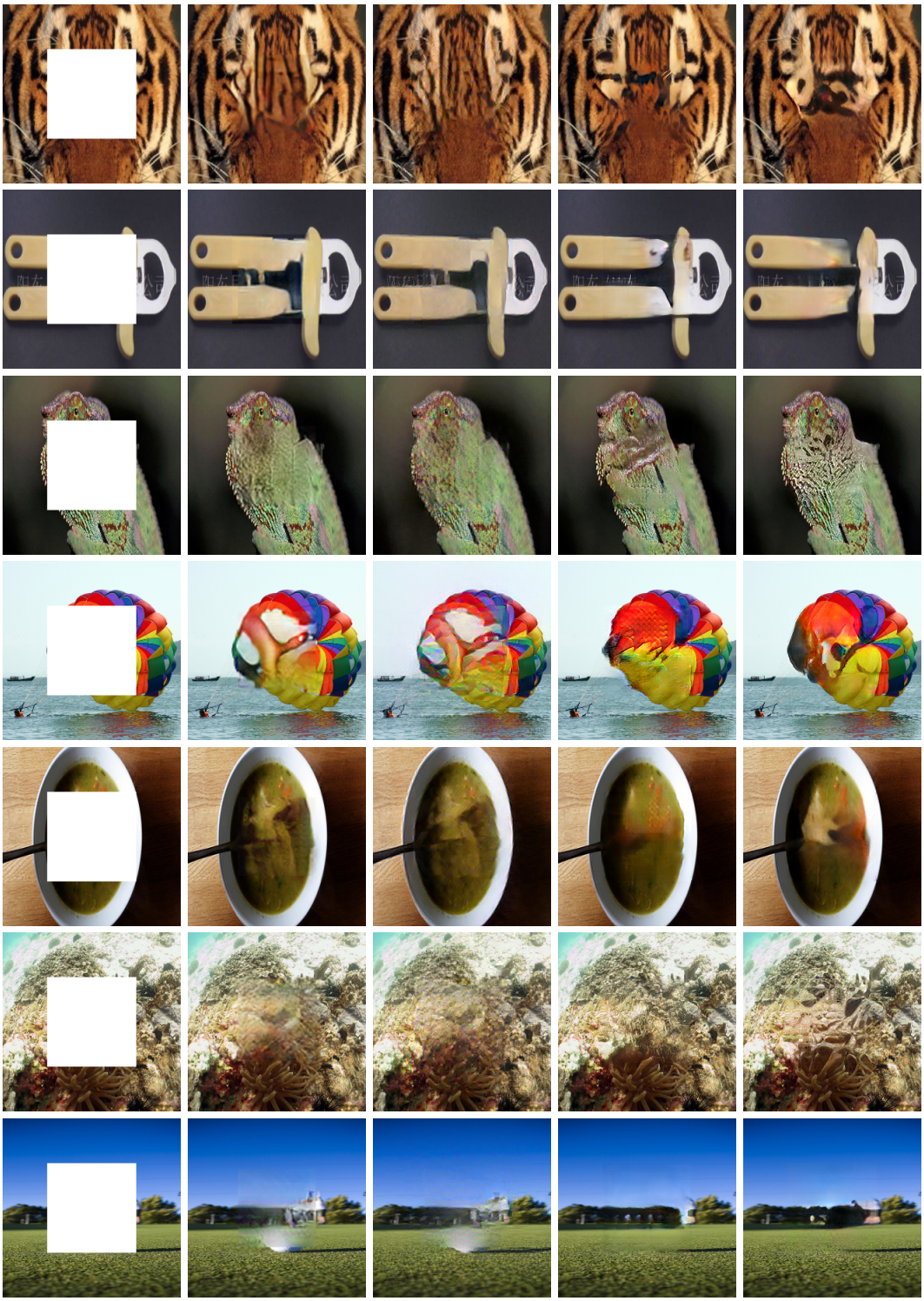
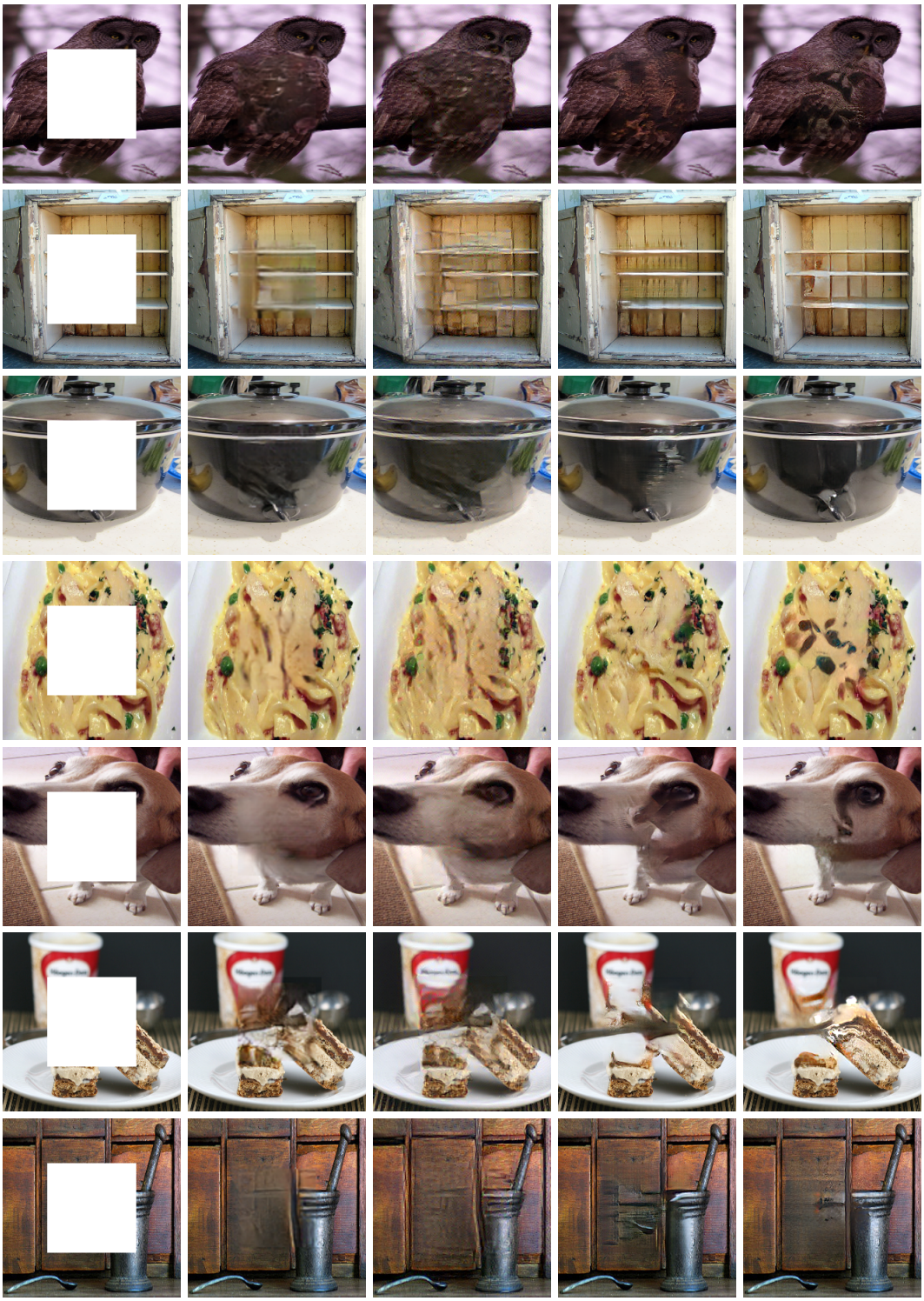
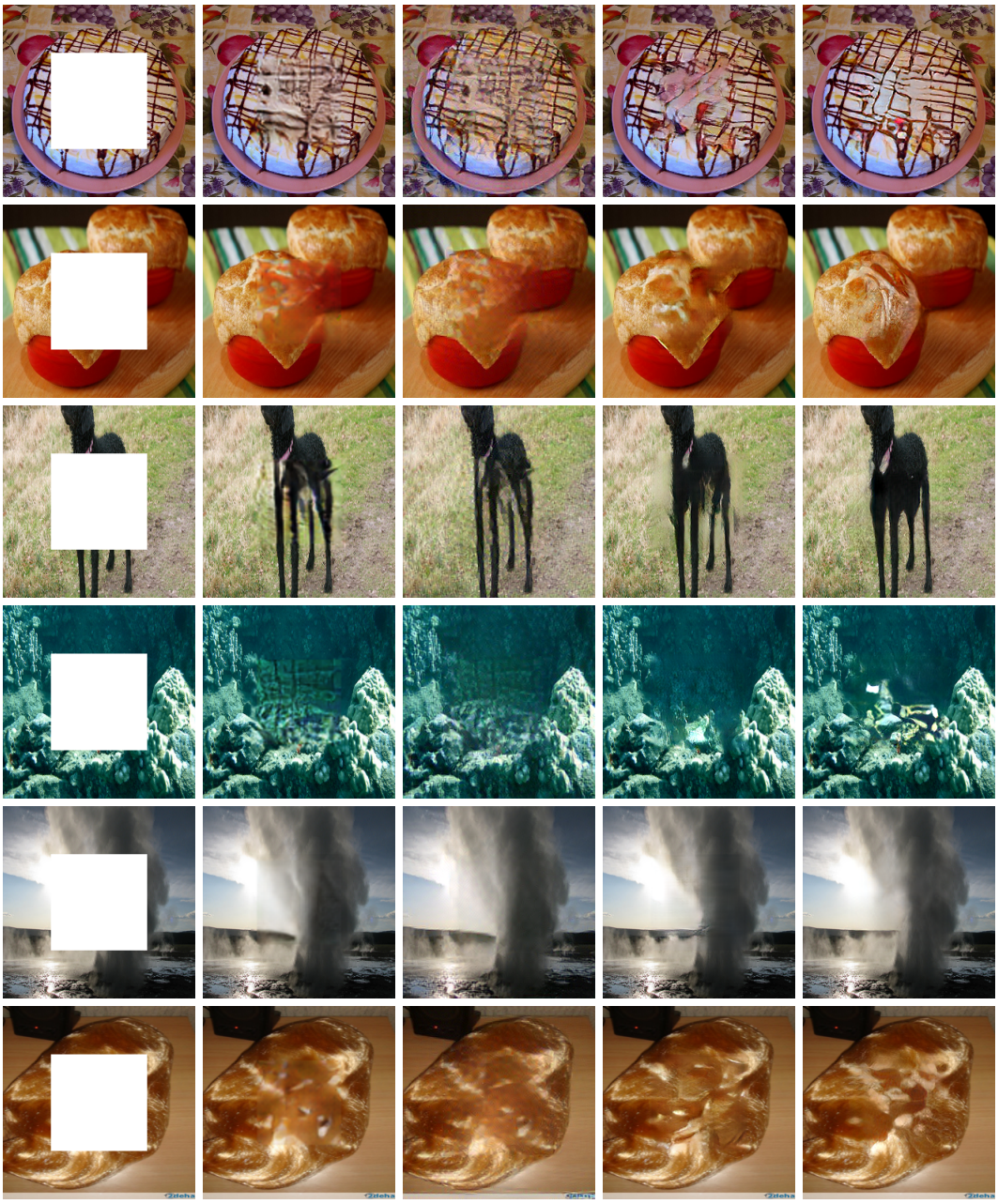


Figure 22: Visual comparisons on ImageNet. (a) Input image. (b) CE [1]. (c) MSNPS [2]. (d) CA [3]. (e) Our results. Best viewed with zoom-in.



(a) (b) (c) (d) (e)

Figure 23: Visual comparisons on ImageNet. (a) Input image. (b) CE [1]. (c) MSNPS [2]. (d) CA [3]. (e) Our results. Best viewed with zoom-in.



(a) (b) (c) (d) (e)

Figure 24: Visual comparisons on ImageNet. (a) Input image. (b) CE [1]. (c) MSNPS [2]. (d) CA [3]. (e) Our results. Best viewed with zoom-in.

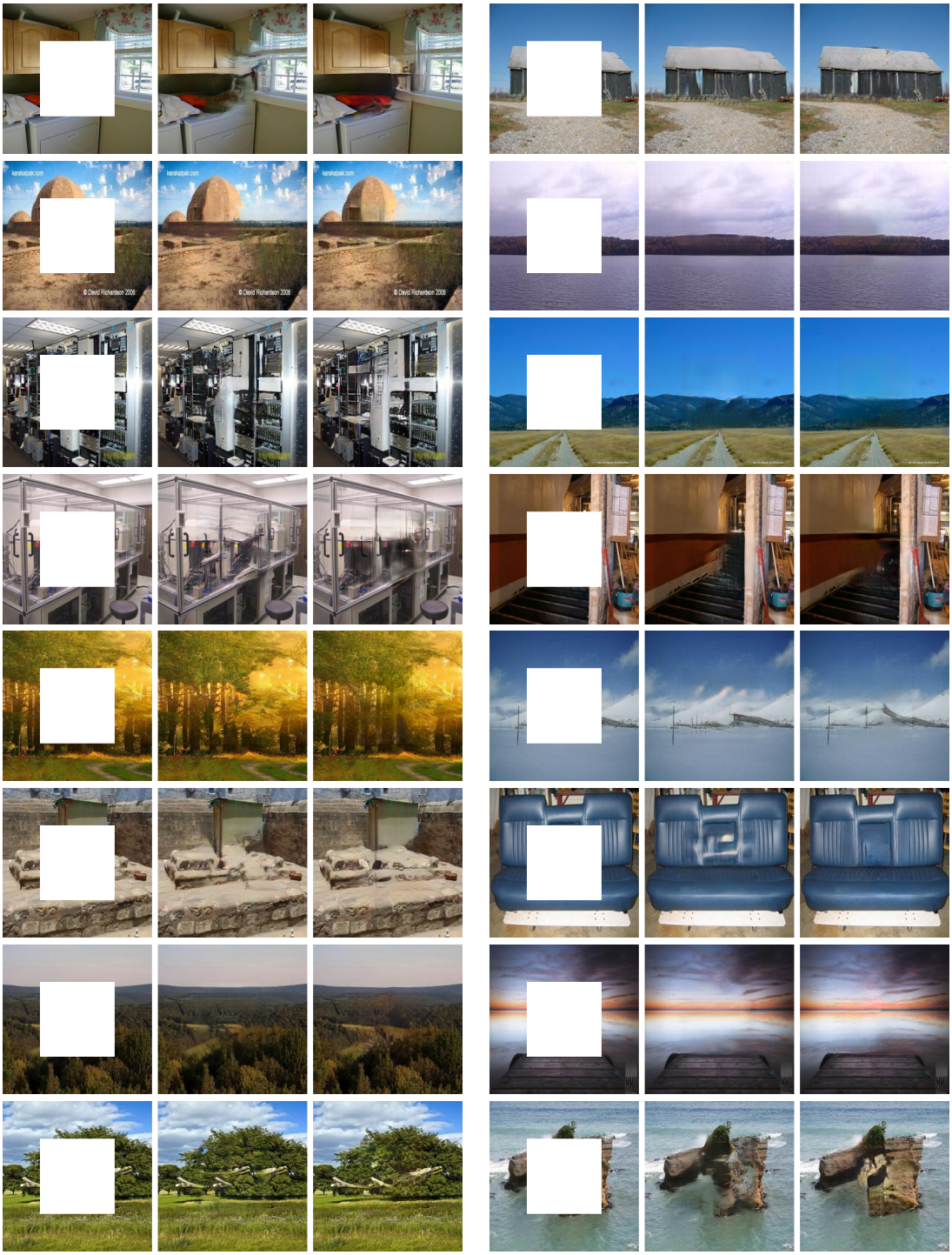


Figure 25: Visual comparisons on Places2. (a) Input image. (b) CA [3]. (c) Our results. Best viewed with zoom-in.

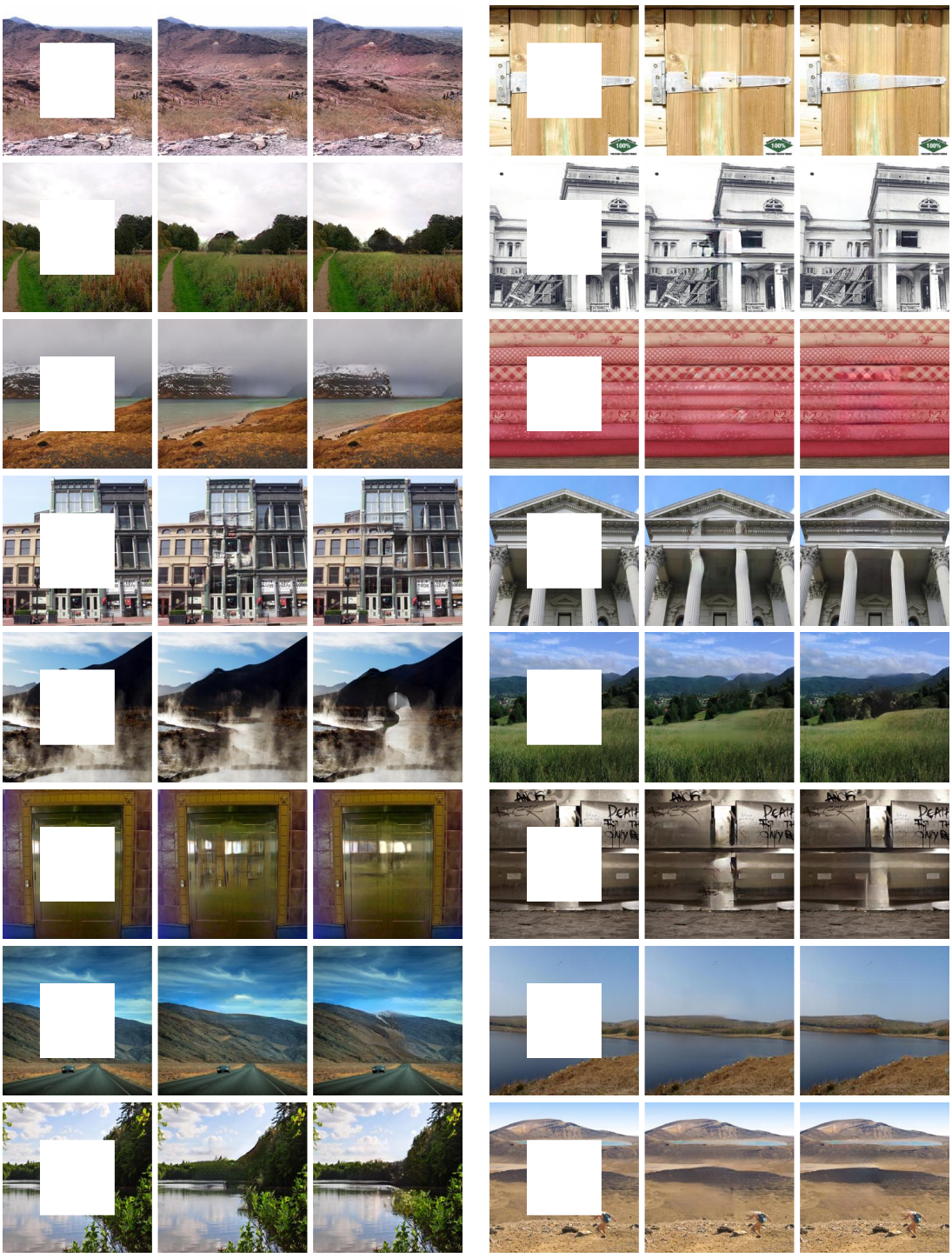


Figure 26: Visual comparisons on Places2. (a) Input image. (b) CA [3]. (c) Our results. Best viewed with zoom-in.

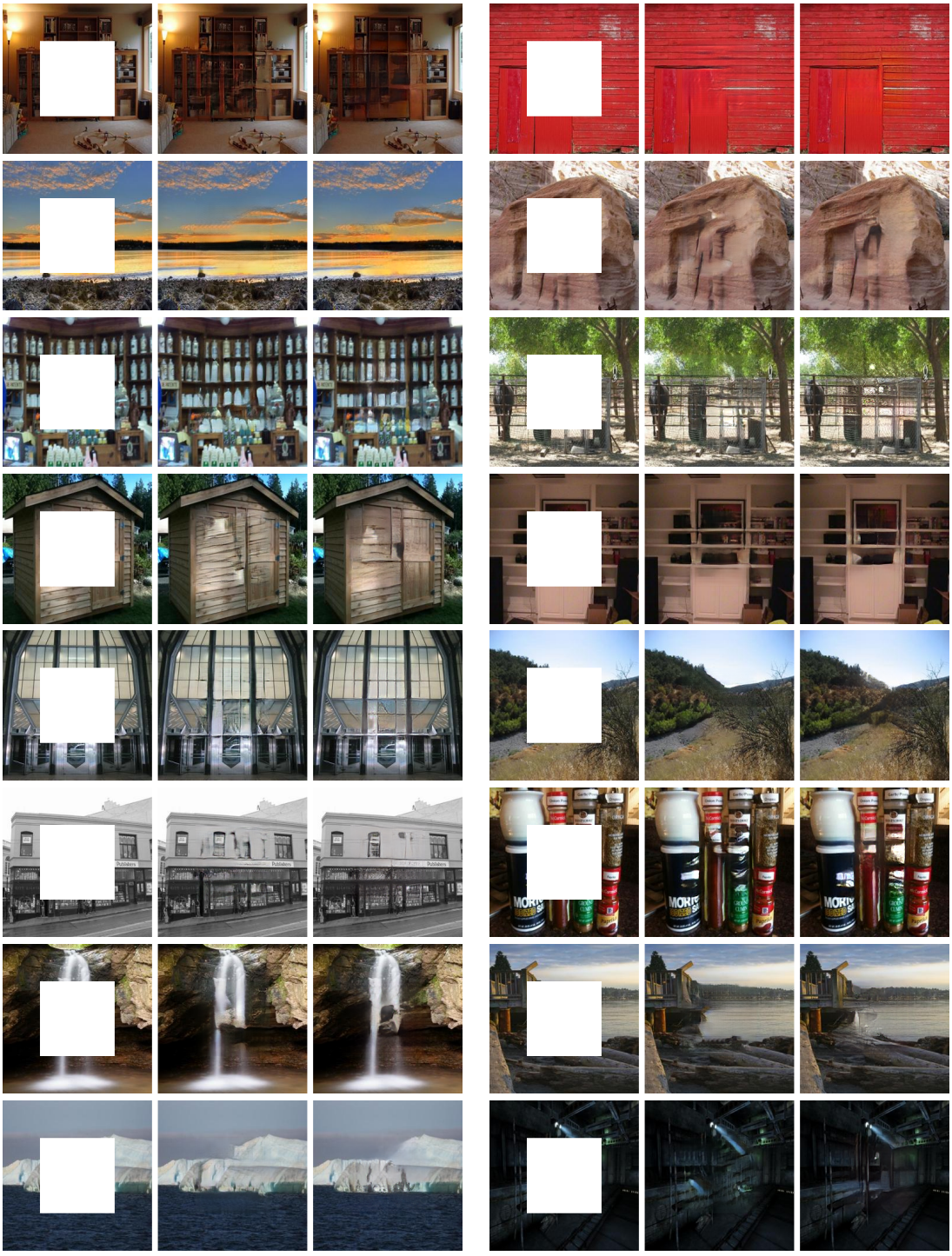
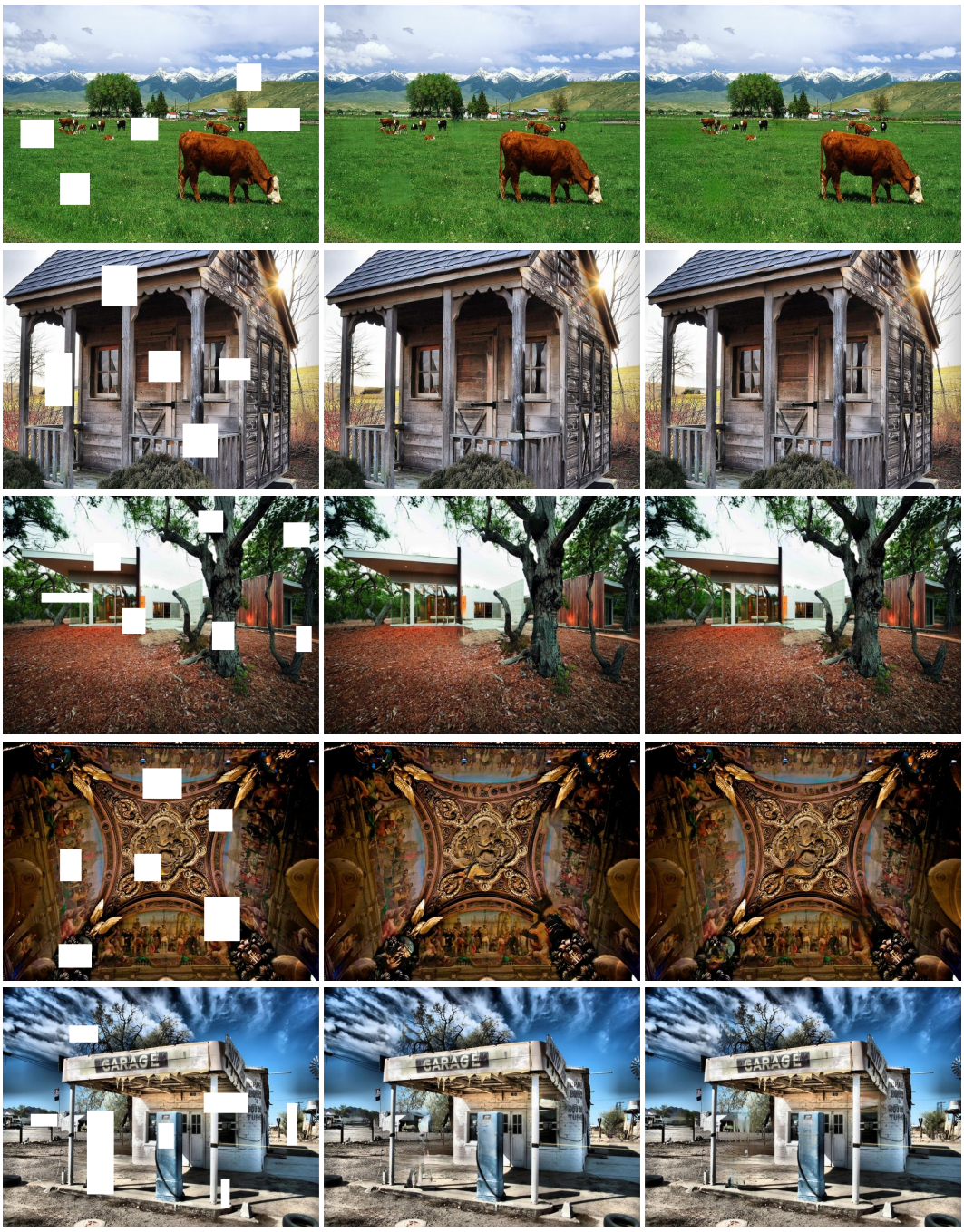
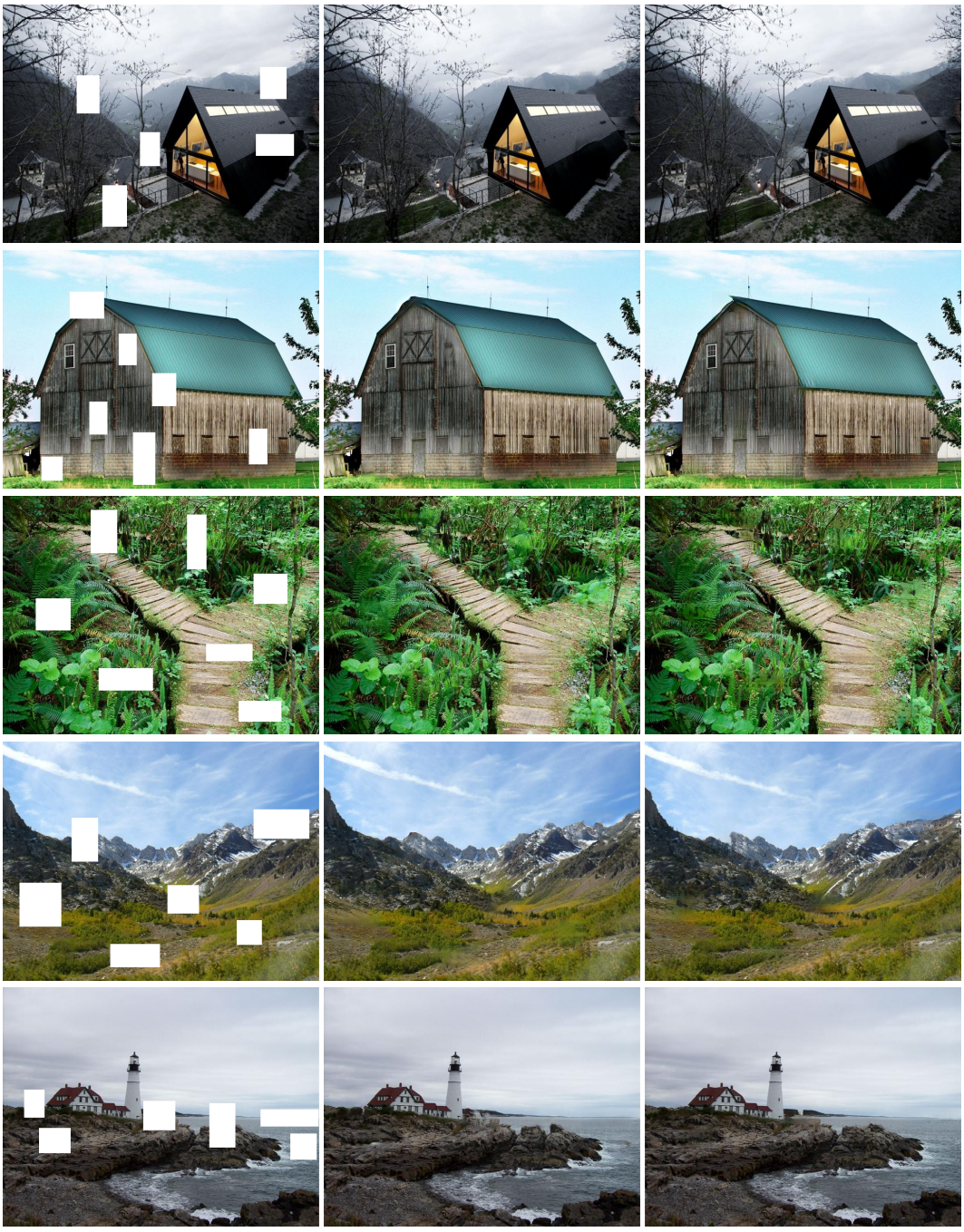


Figure 27: Visual comparisons on Places2. (a) Input image. (b) CA [3]. (c) Our results. Best viewed with zoom-in.



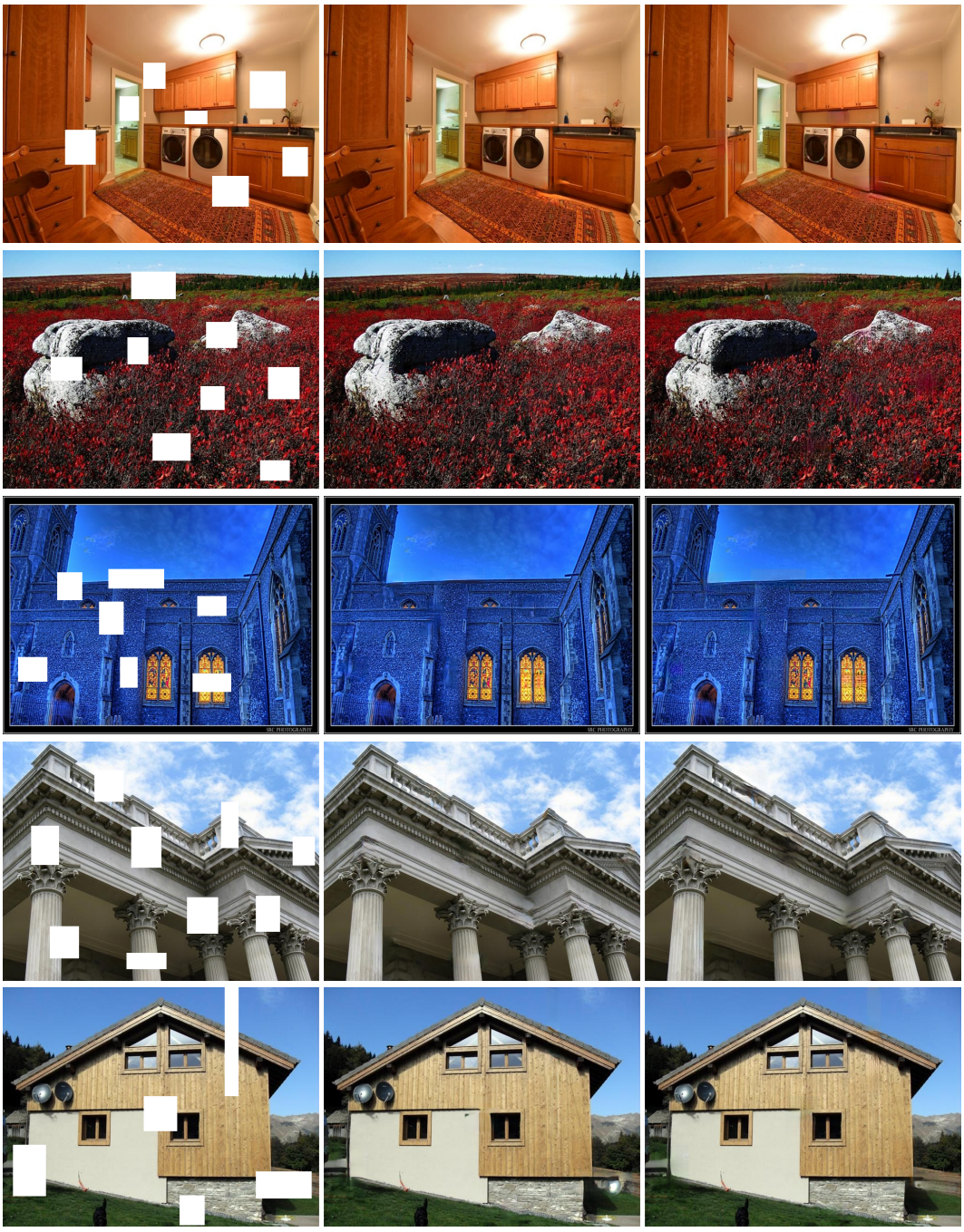
(a) (b) (c)

Figure 28: Inpainting images of 512×680 with random masks from Places2. (a) Input image. (b) Results by CA [3]. (c) Our results. Best viewed with zoom-in.



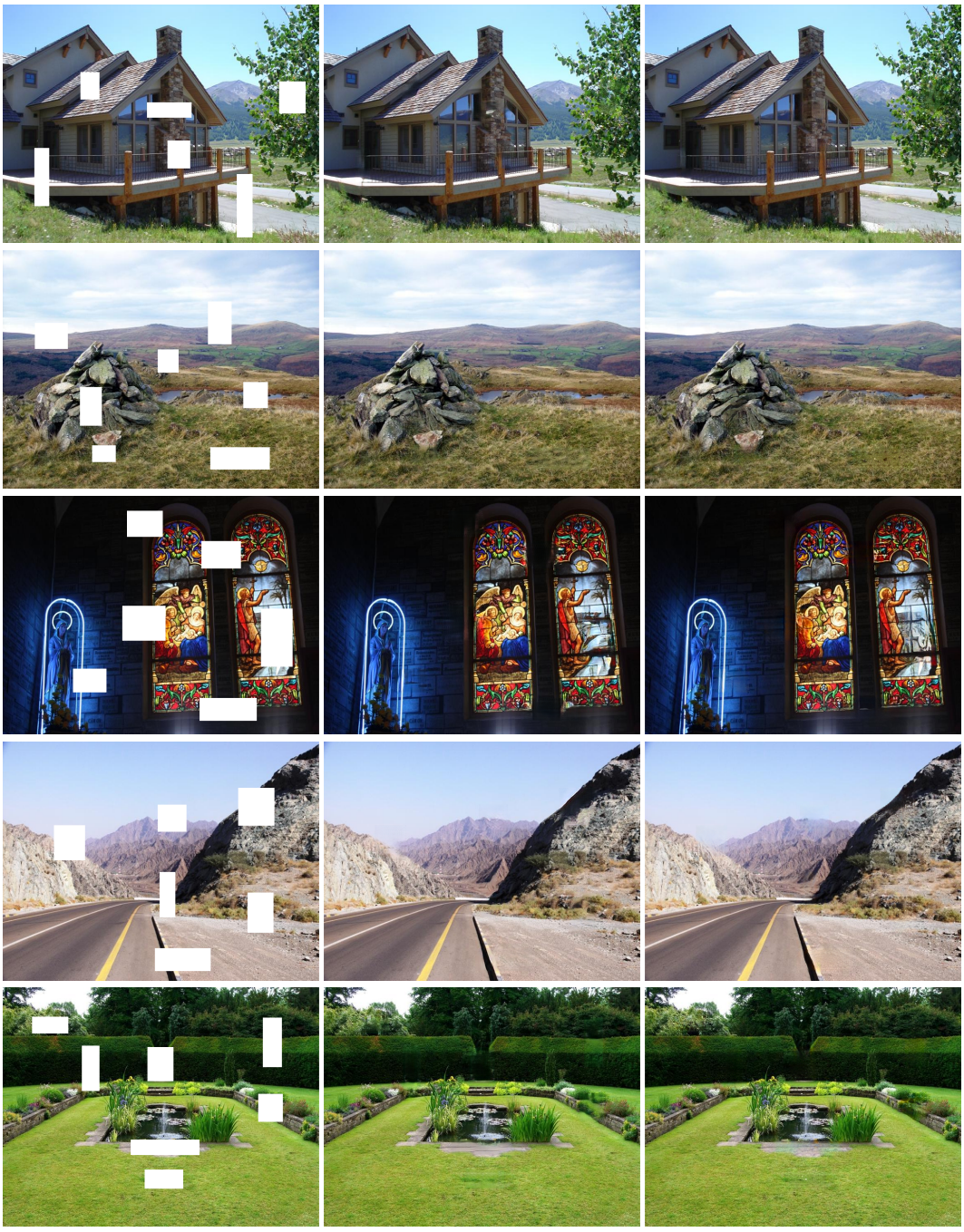
(a) (b) (c)

Figure 29: Inpainting images of 512×680 with random masks from Places2. (a) Input image. (b) Results by CA [3]. (c) Our results. Best viewed with zoom-in.



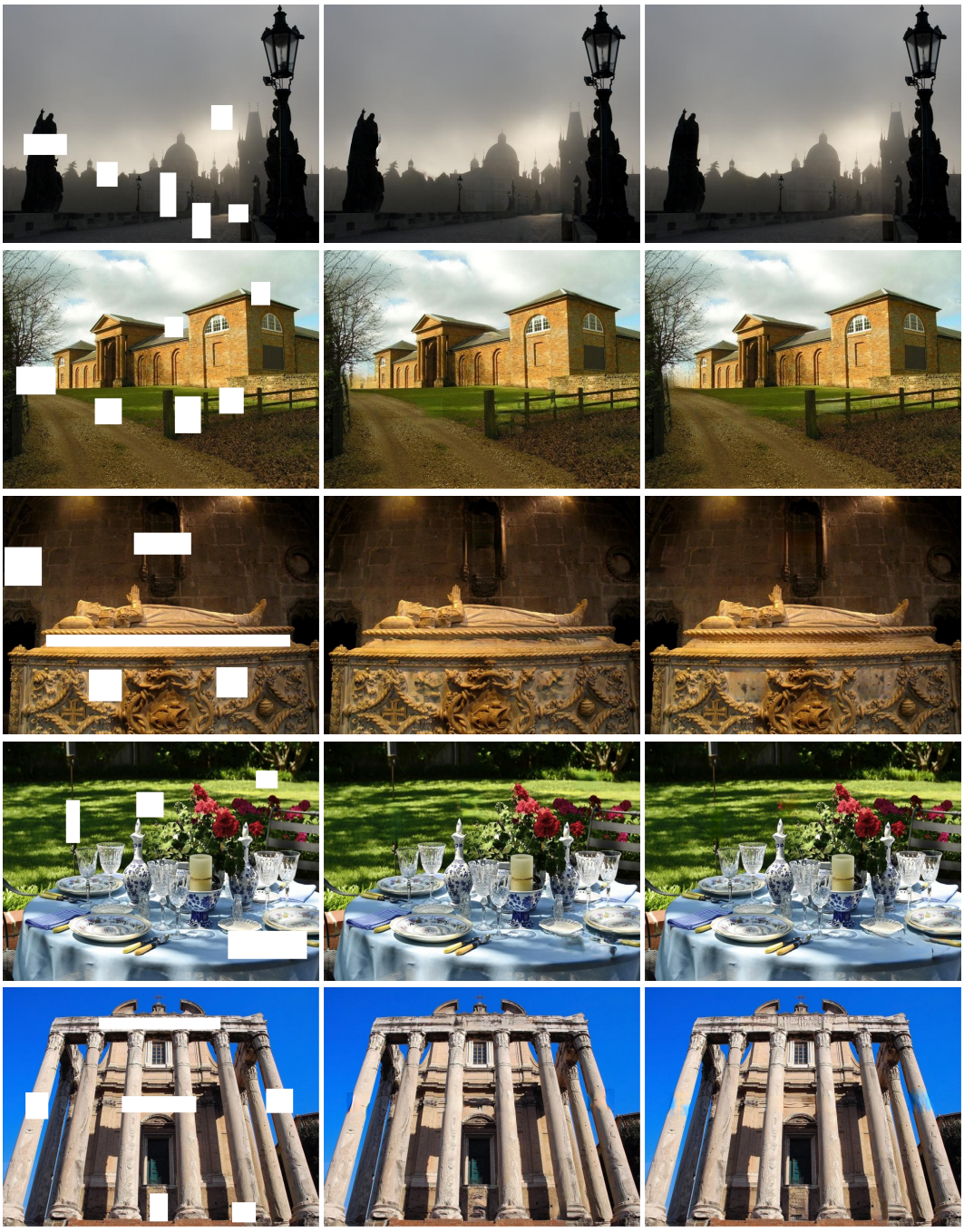
(a) (b) (c)

Figure 30: Inpainting images of 512×680 with random masks from Places2. (a) Input image. (b) Results by CA [3]. (c) Our results. Best viewed with zoom-in.



(a) (b) (c)

Figure 31: Inpainting images of 512×680 with random masks from Places2. (a) Input image. (b) Results by CA [3]. (c) Our results. Best viewed with zoom-in.



(a) (b) (c)

Figure 32: Inpainting images of 512×680 with random masks from Places2. (a) Input image. (b) Results by CA [3]. (c) Our results. Best viewed with zoom-in.

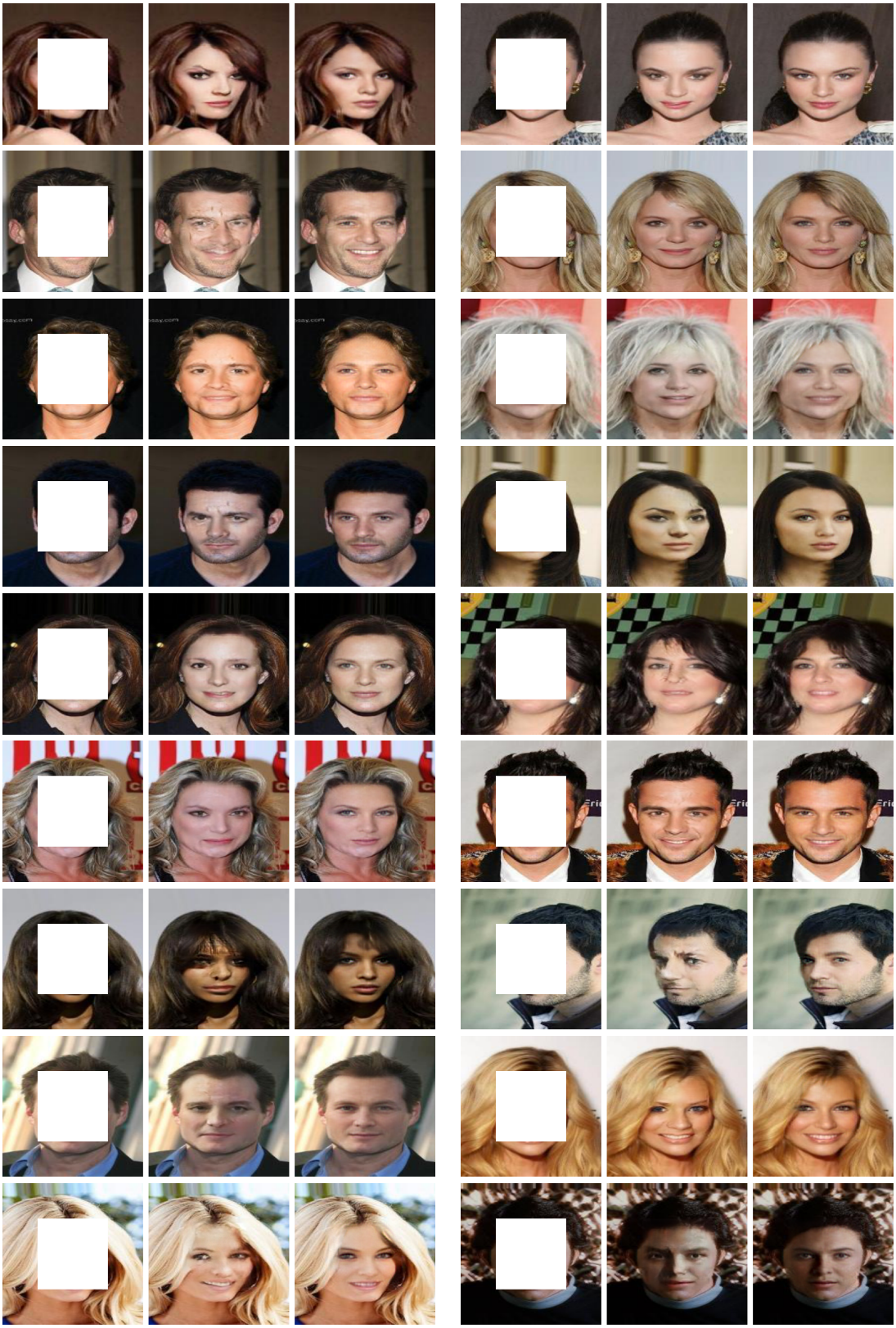


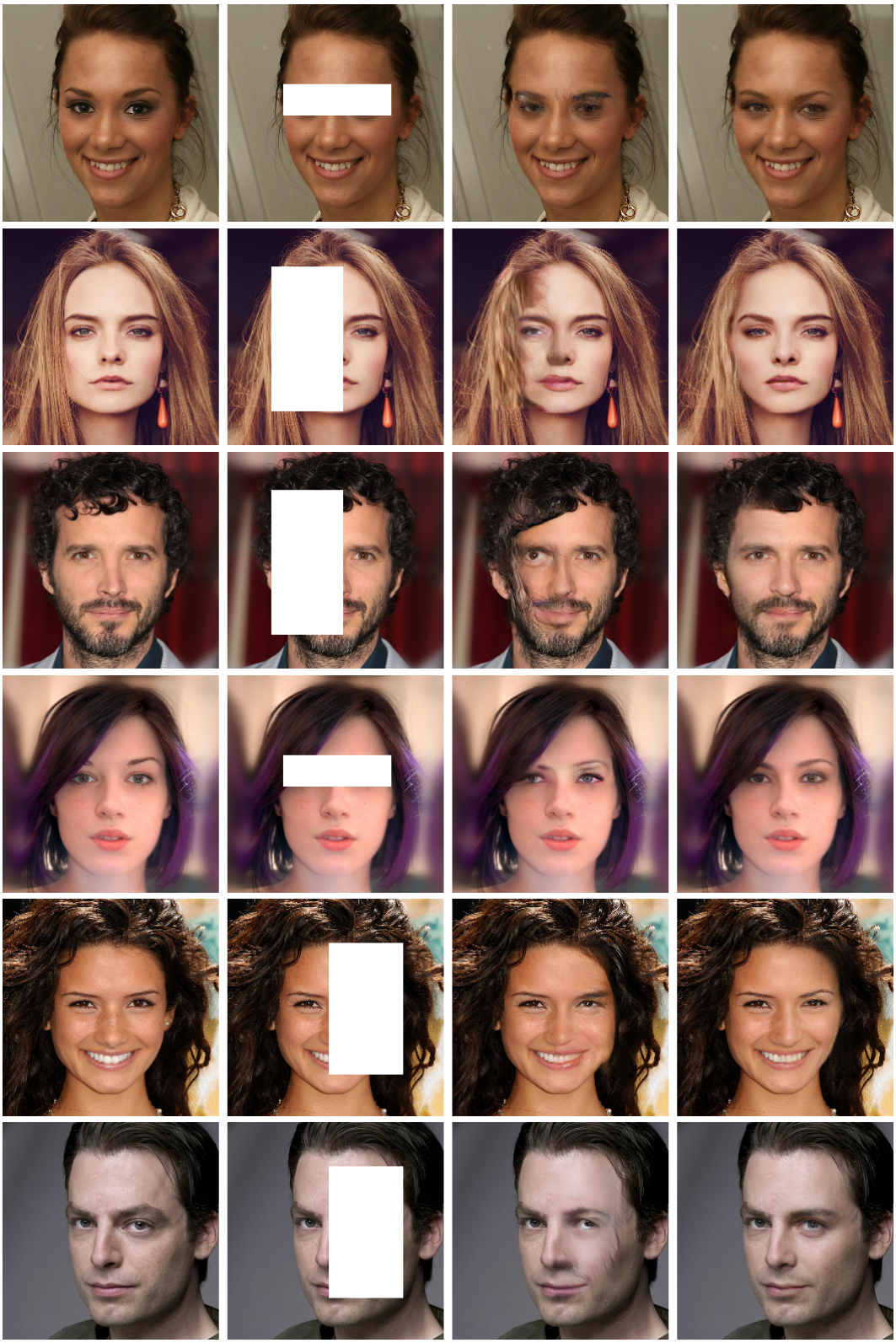
Figure 33: Visual comparisons on CelebA. (a) Input image. (b) CA [3]. (c) Our results. Best viewed with zoom-in.

(a) (b) (c) (d)

Figure 34: Inpainting images with random masks on Paris street view. (a) Ground truth. (b) Input images. (c) CA [3]. (d) Our results. Best viewed with zoom-in.

(a) (b) (c) (d)

Figure 35: Inpainting images with random masks on CelebA-HQ. (a) Ground truth. (b) Input images. (c) CA [3]. (d) Our results. Best viewed with zoom-in.



(a) (b) (c) (d)

Figure 36: Inpainting images with random masks on CelebA-HQ. (a) Ground truth. (b) Input images. (c) CA [3]. (d) Our results. Best viewed with zoom-in.

(a) (b) (c) (d)

Figure 37: Inpainting images with random masks on CelebA-HQ. (a) Ground truth. (b) Input images. (c) CA [3]. (d) Our results. Best viewed with zoom-in.

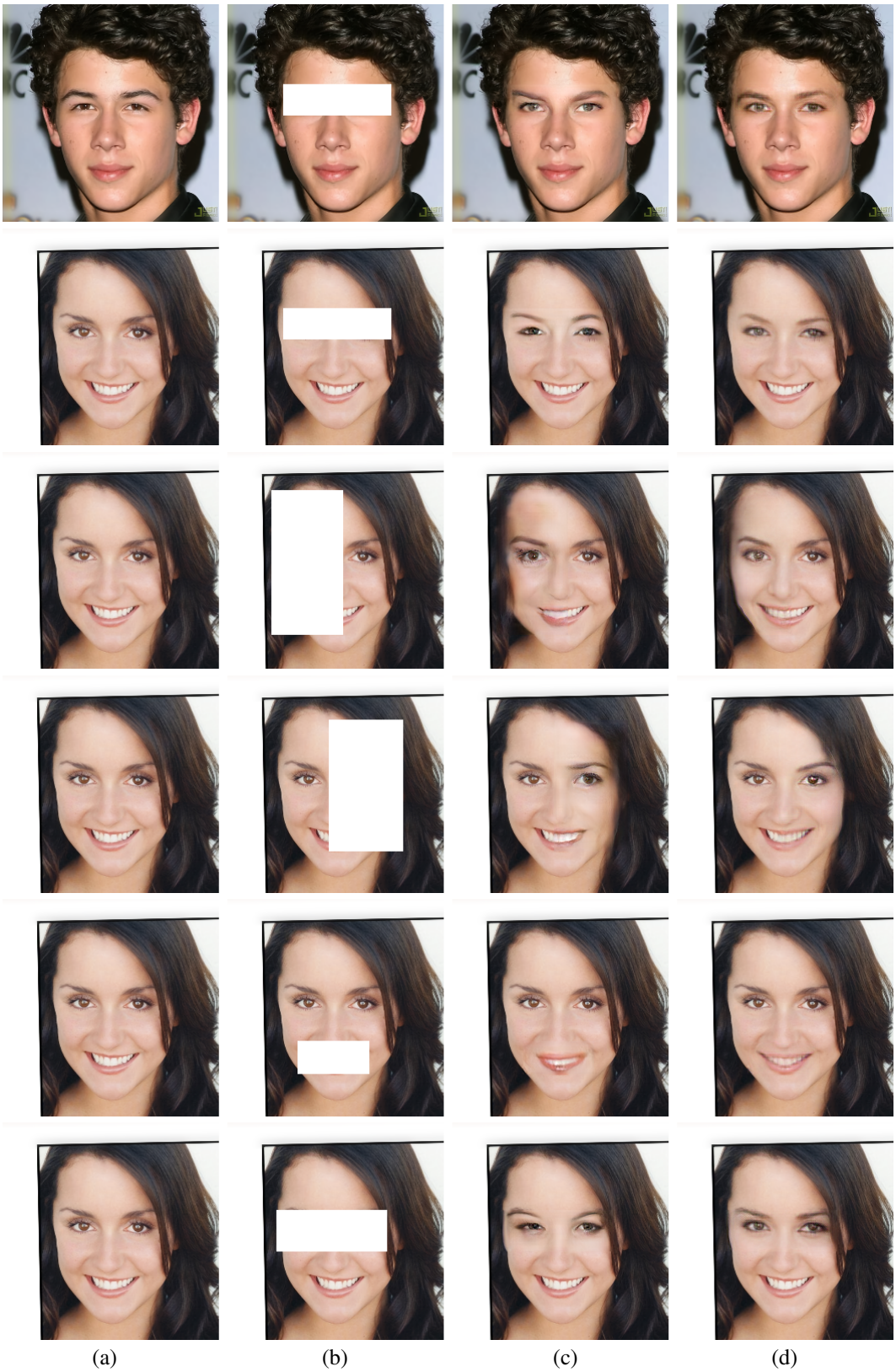
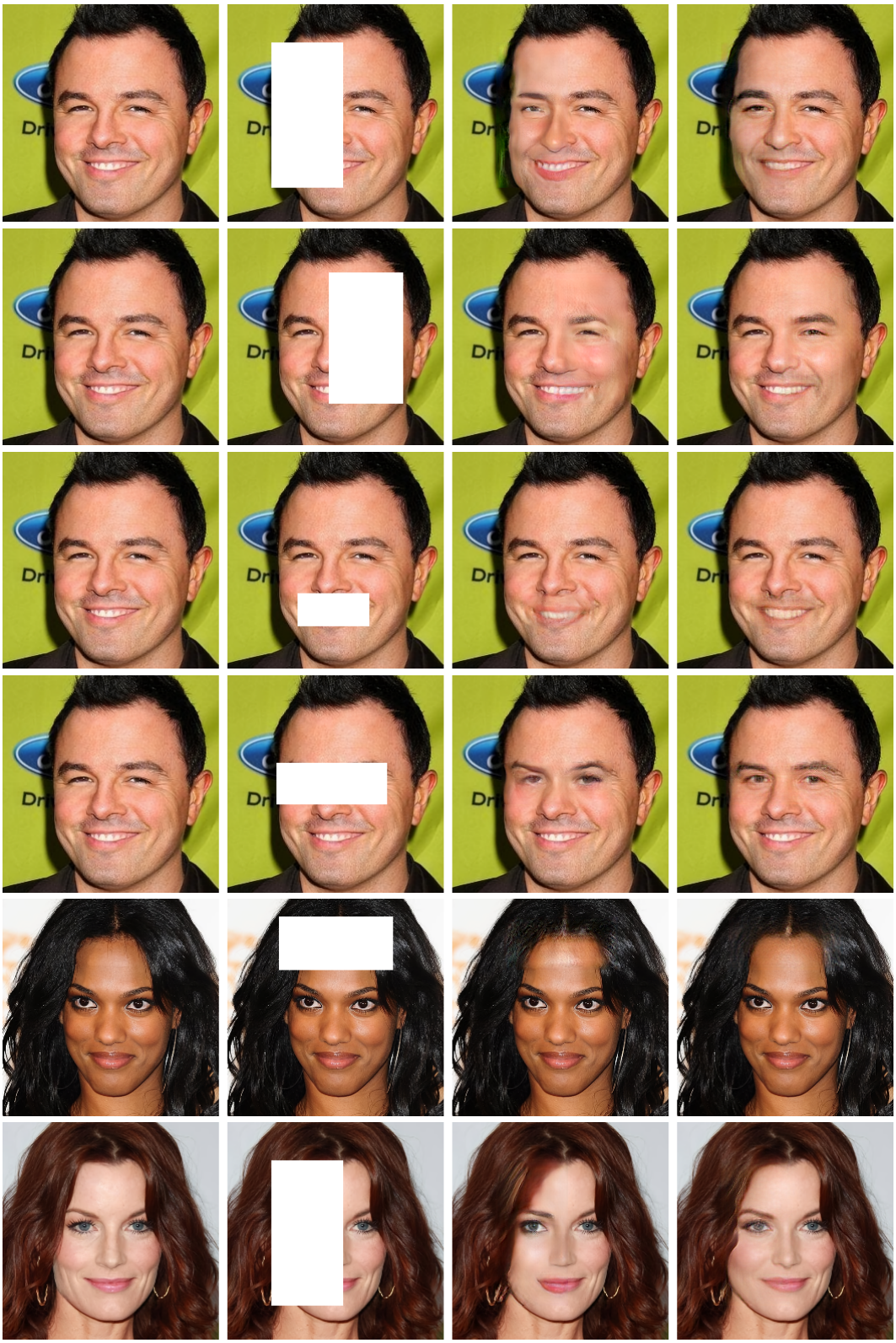


Figure 38: Inpainting images with random masks on CelebA-HQ. (a) Ground truth. (b) Input images. (c) CA [3]. (d) Our results. Best viewed with zoom-in.



(a) (b) (c) (d)

Figure 39: Inpainting images with random masks on CelebA-HQ. (a) Ground truth. (b) Input images. (c) CA [3]. (d) Our results. Best viewed with zoom-in.

(a) (b) (c)

Figure 40: Inpainting images of 512×512 with central mask from CelebA-HQ. (a) Input images. (b) Ground truth. (c) Our results. Best viewed with zoom-in.

(a) (b) (c)

Figure 41: Inpainting images of 512×512 with central mask from CelebA-HQ. (a) Input images. (b) Ground truth. (c) Our results. Best viewed with zoom-in.

(a) (b) (c)

Figure 42: Inpainting images of 512×512 with central mask from CelebA-HQ. (a) Input images. (b) Ground truth. (c) Our results. Best viewed with zoom-in.

References

[1] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros. Context encoders: Feature learning by inpainting. In *CVPR*, pages 2536–2544, 2016.

[2] C. Yang, X. Lu, Z. Lin, E. Shechtman, O. Wang, and H. Li. High-resolution image inpainting using multi-scale neural patch synthesis. In *CVPR*, volume 1, page 3, 2017.

[3] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang. Generative image inpainting with contextual attention. *arXiv preprint arXiv:1801.07892*, 2018.