Supplementary Material for: Hierarchical Contrastive Learning for Pattern-Generalizable Image Corruption Detection

1. Overview

In this supplementary material, we provide more implementation details, model architecture, experimental results, and analysis, including:

- Detailed architecture of Transformer-based Restoration model (Section 2).
- Training details of proposed hierarchical contrastive learning framework(Section 3).
- Additional results of ablation study and user study (Section 4).
- More visual comparisons in various experimental settings, including typical blind image inpainting, bidirectional blind image inpainting, model generalization, image watermark removal, and shadow removal (Section 5).

2. Network Architecture

As illustrated in Figure 2 in the main paper, our Transformer-based restoration model follows the encoder-decoder framework. Given an input image, it first employs a convolution with 5×5 kernel to extract image tokens. Table 1 lists the detailed architecture of our Transformer-based restoration model, which mainly consists of three parts, encoder, bottleneck, and decoder. Since encoder needs to predict corruption masks by hierarchical contrastive learning, as well as extracting features from the input image, it contains more transformer blocks. Finally, decoder employs a 1×1 convolution to generate output images, which are further sent to an additional Conv-U-Net to refine high-frequency details of output results, leaning upon the local texture refinement capability and efficiency of CNNS. Besides, our model employs stride-2 convolution to downsample feature maps in encoder and nearest neighboring interpolation to upsample feature maps in decoder.

In hierarchical interaction mechanism, our model leverages a projection head to produce input features in the current stage from features in the previous stage, which comprises two fully connected layers and a *GELU* [4] nonlinearity in between. Thus, it is able to perform contrastive learning and clustering analysis in the projected feature space. Besides, our model also employs a linear mapping to decrease the feature dimension.

Table 1: Architecture of Transformer-based Inpainting Model

Module	Stage	Dim	Resolution	Blocks	Heads
Encoder	T_1	64	256×256	6	1
	T_2	128	128×128	4	2
	T_3	256	64×64	2	4
Bottleneck	T_4	256	64×64	2	4
Decoder	T_5	256	64×64	2	4
	T_6	128	128×128	2	2
	T_7	64	256×256	2	1

Our framework employs a Conv-U-Net to improve the quality of image details, following previous methods [7, 3] for image inpainting. It takes the reconstructed image and the predicted mask as input, gradually downsampling the feature maps by five convolution-based residual blocks, and then upsampling back to the original size. The number of feature channels starts from 64 and is doubled after each downsampling with a maximum of 512. All the convolutions are gated convolutions [10], which allows the network to automatically refine the content of corruption regions.

3. Implementation Details

The training procedure of hierarchical contrastive learning contains two phases. In the first phase, we train the Transformer-based inpainting model with loss coefficients $\lambda_1=0.1$,$\lambda_2=1$,$\lambda_3=0.1$ and $\lambda_4=0$. The learning rate is 0.001. Before the convergence of training contrastive learning, we feed the groundtruth masks to decoder for stabilizing the training of image inpainting. Note that we simultaneously train encoder and decoder together to avoid losing low-level details of the input image by contrastive learning. After the training of contrastive learning in all stages is converged, we jointly train the Transformer-based inpainting model and the refinement network with loss coefficients $\lambda_1=0.1$,$\lambda_2=1$,$\lambda_3=0.1$,$\lambda_4=1$. The learning rate of the refinement network and the discriminator is 0.0001. The $\mathcal{L}_1$

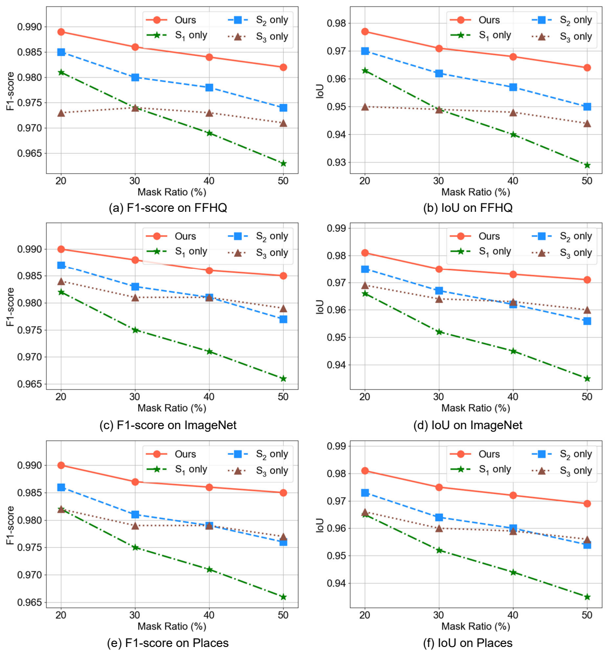


Figure 1: Comparison with single stage contrastive learning.

Table 2: Experiments on different depths.

Depth	1	2	3	4
F1 ↑	0.967	0.984	**0.986**	0.982
IoU ↑	0.938	0.970	**0.973**	0.966

loss and perceptual loss are employed on both outputs of two networks, while the adversarial loss is only employed on the refined results.

4. Additional Experimental Results

4.1. Ablation study.

Effect of hierarchical contrastive learning. In Figure 1, we further compare the performance of our hierarchical contrastive learning-based model with single-stage contrastive learning-based models on three benchmark datasets [11, 6, 2]. The results demonstrate that our proposed method improves the performance of mask detection for blind image inpainting.

Investigation on the depth of hierarchical contrative learning. To explore which depth is optimal to detect the corrupted mask, we train the model with different depths of hierarchical contrastive learning on FFHQ [6] dataset. The experimental results are illustrated in Table 2, which shows deploying three depths of hierarchical contrastive learning achieves the best performance.

Table 3: Effect of capacity on generalization.

Corruption pattern		Random constant		CelebA-HQ [5]	
Mask ratio (%)		0-30	30-60	0-30	30-60
Acc ↑	VCNet (1.37M)	0.981	0.977	0.975	0.974
	Ours-small (808K)	**0.987**	**0.981**	**0.981**	**0.975**
IoU ↑	VCNet (1.37M)	0.977	0.960	0.969	0.954
	Ours-small (808K)	**0.985**	**0.965**	**0.976**	**0.956**

Table 4: ConvNet vs. Transformer.

Methods	Segmenter [8]	VCNet [9]	ConvNet+HCL (Ours)	Transformer+HCL (Ours)
ACC↑	0.974	0.975	0.978	0.980
IoU↑	0.962	0.963	0.969	0.970

Table 5: Computational overhead.

Metrics	Params (M)	Memory usage (GB)	GMACs	Inference time (ms)
VCNet	3.88	1.65	11.4	15.3
Ours	3.57	1.97	10.6	29.1 (28.7 for Transformer)

Effect of capacity on generalization. To investigate the effect of model capacity on the generalization, we reduce the size of our model to only 0.6× of VCNet, and find that our model still outperforms VCNet distinctly, as shown in Table 3.

ConvNet vs. Transformer. We further investigate the effect of the backbone network by using ConvNet, with similar structure and model size as VCNet, as the backbone for our Hierarchical Contrastive Learning, denoted as 'ConvNet+HCL'. The performance of corruption detection on Places dataset in Table 4 validates the effectiveness of our model.

Computational complexity. The results listed in Table 5 lists show that 1) our model has comparable parameters, memory usage and MACs with VCNet; 2) our model has longer inference time than VCNet, most of which are consumed by Transformer.

4.2. User study.

Since quantitative metrics have their bias for the quality evaluation of restored images, to standardize evaluation process, we further perform user study in our experiment. We randomly select 50 test images of FFHQ [6] and Places [11] respectively, and present reconstructed results of two methods to 20 human subjects for manual ranking of image quality. Figure 3 lists the voting results of this user study. For the FFHQ dataset, our model reaches 80.2% votes among total 50×20=1000 rankings, which is much higher than VCNet. In addition, we count winning samples of each method, and our model wins on 41 test samples and VCNet altogether 9 samples. As for the Places dataset, our model reaches 72.1% votes among 1000 rankings, which is significantly higher than VCNet as well. And our model wins on 37 test samples and VCNet 13 samples. As a result, our model is able to precisely detect the corruption mask and fill in more realistic content for corrupted regions.

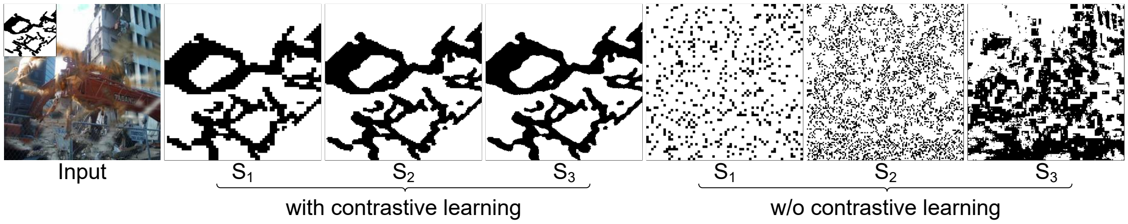


Figure 2: Visualization of masks in different stages.

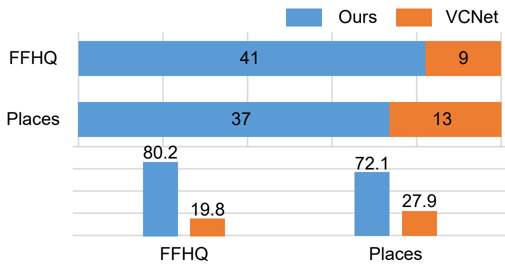


Figure 3: The results of user studies. First row: winning samples. Second row: share of the vote.

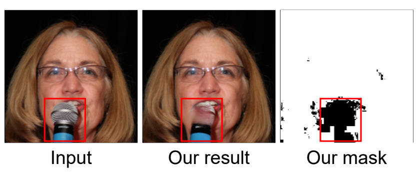


Figure 4: The microphone is mis-detected as corruption.

5. More Qualitative Results

Visualization of detected masks across stages. We visualize the detected masks across stages with and without our hierarchical contrastive learning in Figure 2, which reveals the effectiveness of our hierarchical contrastive learning.

Failure cases. A foreground object that differs semantically from the other context, especially rarely appeared in the training data, tends to be mis-detected as corruption. Figure 4 illustrates such an example, in which the microphone is wrongly recognized as corruption.

Typical blind image inpainting We present more visual results on three benchmark datasets [6, 2, 11] in Figure 5-11.Note that our method not only achieves higher-quality results while coping with different corruption ratios, but also gets impressive results in the bidirectional blind image inpainting setting.

Generalization w.r.t. corruption patterns We also supplement the visual results of generalization experiments on novel corruption patterns, including random noise (shown in Figure 13 and Figure 14), single constant (shown in Figure 15 and Figure 16), and CelebA-HQ [5] (shown in Figure 17 and Figure 18).

Other tasks of image restoration To show the great potential of proposed hierarchical contrastive learning method in real-life applications, we also present the experimental results on other tasks of image restoration, such as image watermark removal (shown in Figure 19) and image shadow removal (shown in Figure 20).

References

[1] Xiaodong Cun and Chi-Man Pun. Split then refine: stacked attention-guided resunets for blind single image visible watermark removal. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1184–1192, 2021. 19

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2, 3, 7, 8

[3] Qiaole Dong, Chenjie Cao, and Yanwei Fu. Incremental transformer structure enhanced image inpainting with masking positional encoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11358–11368, 2022. 1

[4] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 1

[5] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018. 2, 3, 17, 18

[6] 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. 2, 3, 5, 6, 11

[7] Wenbo Li, Zhe Lin, Kun Zhou, Lu Qi, Yi Wang, and Jiaya Jia. Mat: Mask-aware transformer for large hole image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10758–10768, 2022. 1

[8] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7262–7272, 2021. 2

[9] Yi Wang, Ying-Cong Chen, Xin Tao, and Jiaya Jia. Vcnet: A robust approach to blind image inpainting. In *European Conference on Computer Vision*, pages 752–768. Springer, 2020. 2, 19, 20

[10] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4471–4480, 2019. 1

[11] 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*, 40(6):1452–1464, 2017. 2, 3, 9, 10

[12] Yurui Zhu, Jie Huang, Xueyang Fu, Feng Zhao, Qibin Sun, and Zheng-Jun Zha. Bijective mapping network for shadow removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5627–5636, 2022. 20

Figure 5: Visual results of blind image inpainting on FFHQ dataset [6].

Figure 6: Visual results of blind image inpainting on FFHQ dataset [6].

Figure 7: Visual results of blind image inpainting on ImageNet dataset [2].

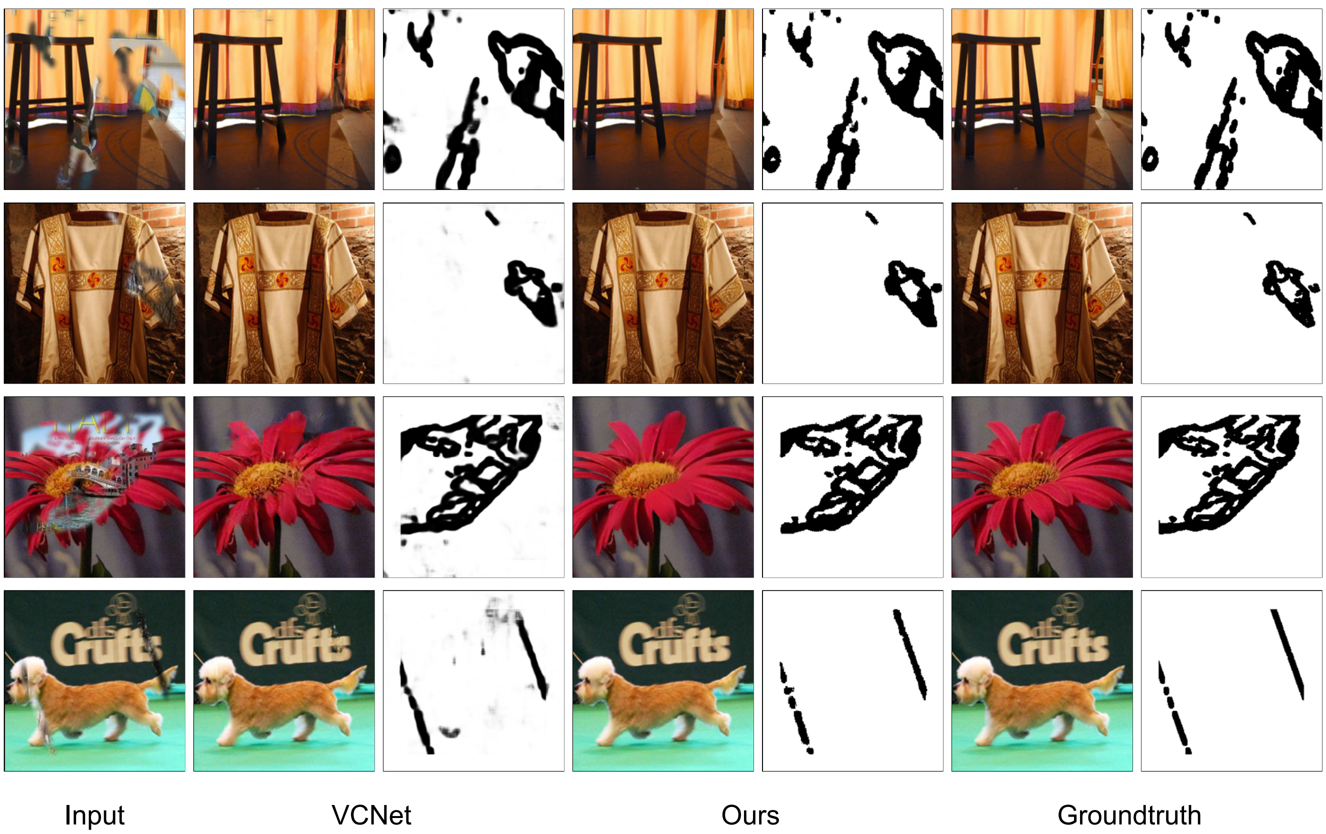


Figure 8: Visual results of blind image inpainting on ImageNet dataset [2].

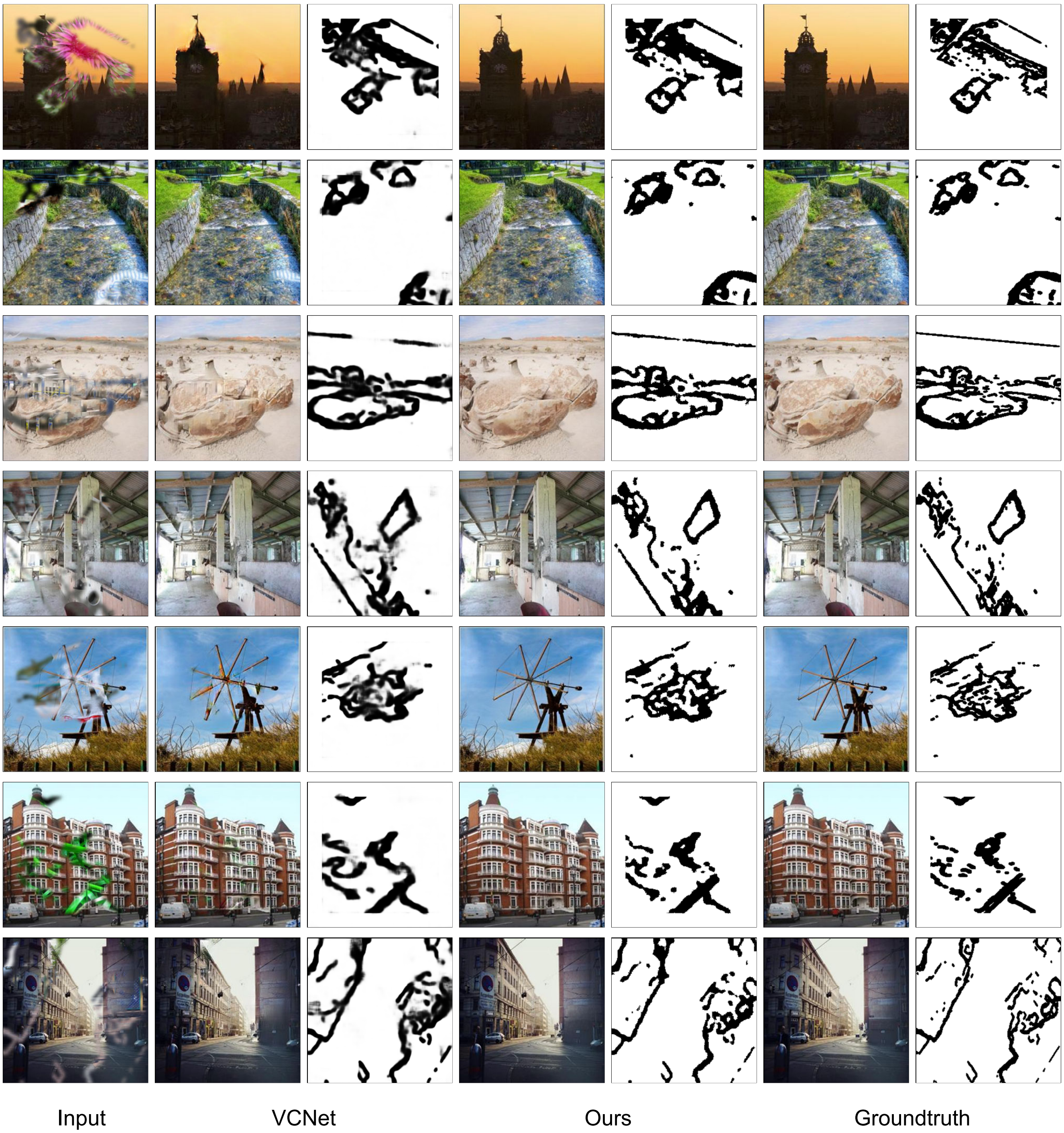


Figure 9: Visual results of blind image inpainting on Places dataset [11].

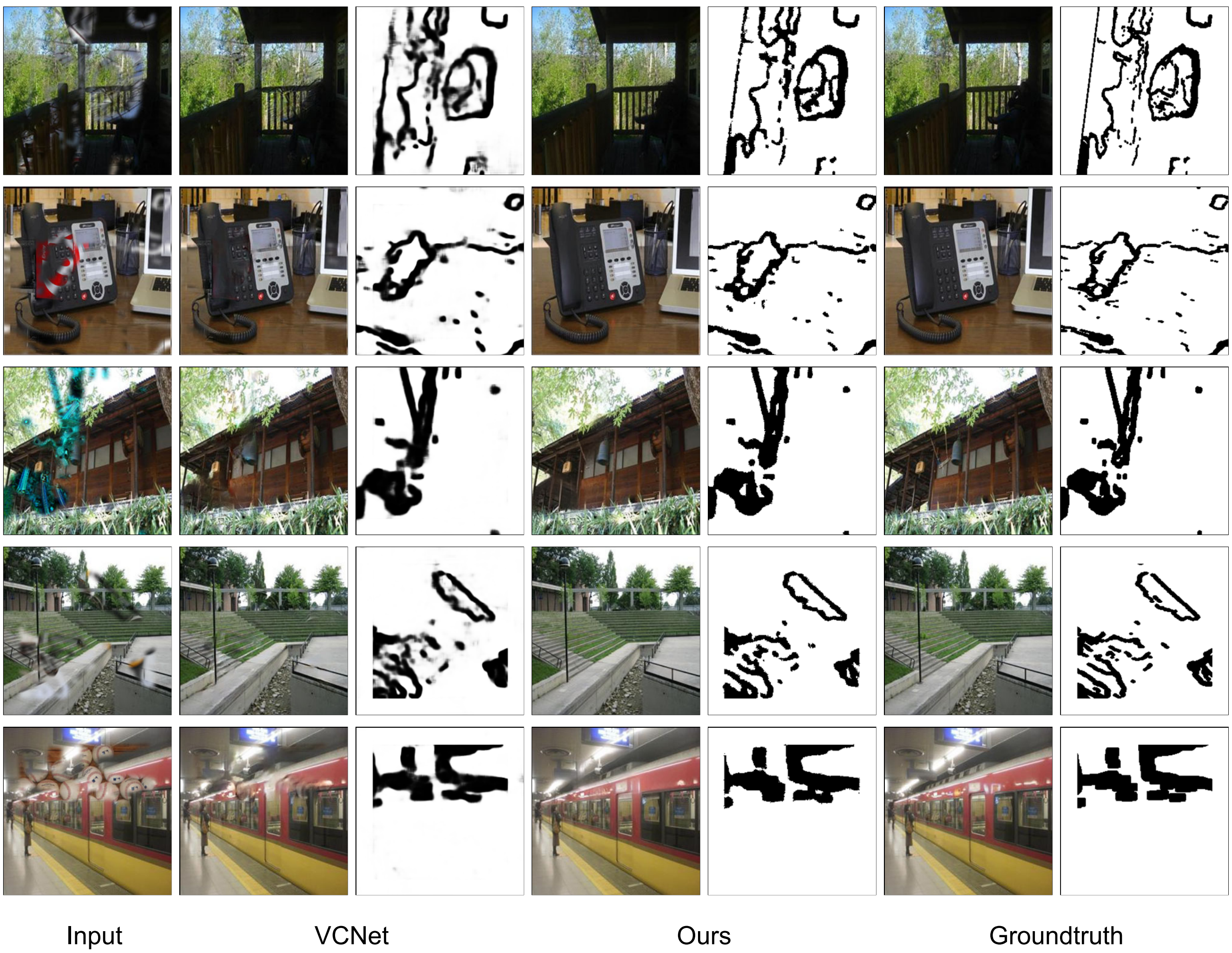


Figure 10: Visual results of blind image inpainting on Places dataset [11].

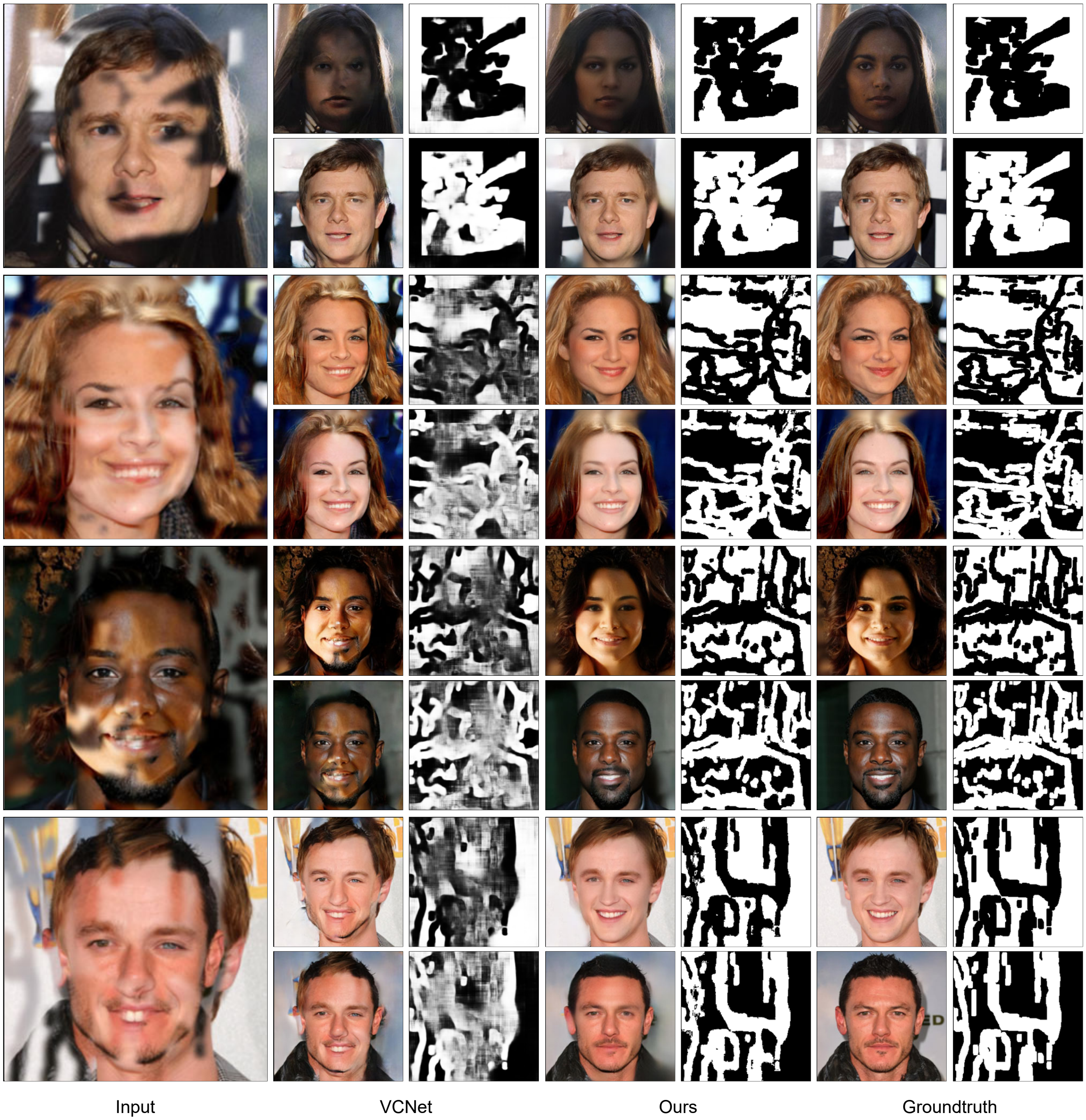


Figure 11: Visual results of bidirectional blind image inpainting on FFHQ dataset [6].

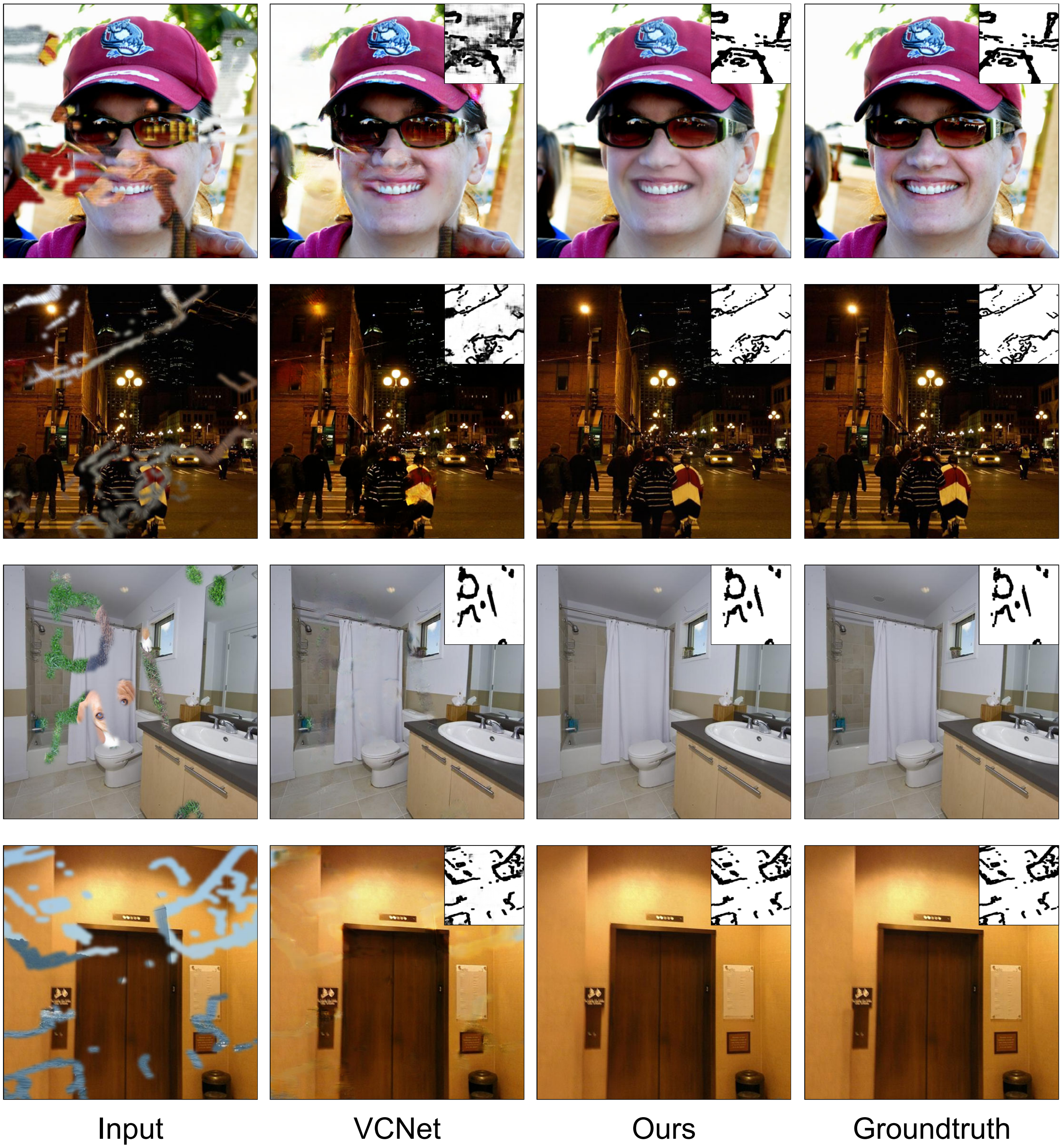


Figure 12: Visual results of blind image inpainting on 512×512 images.

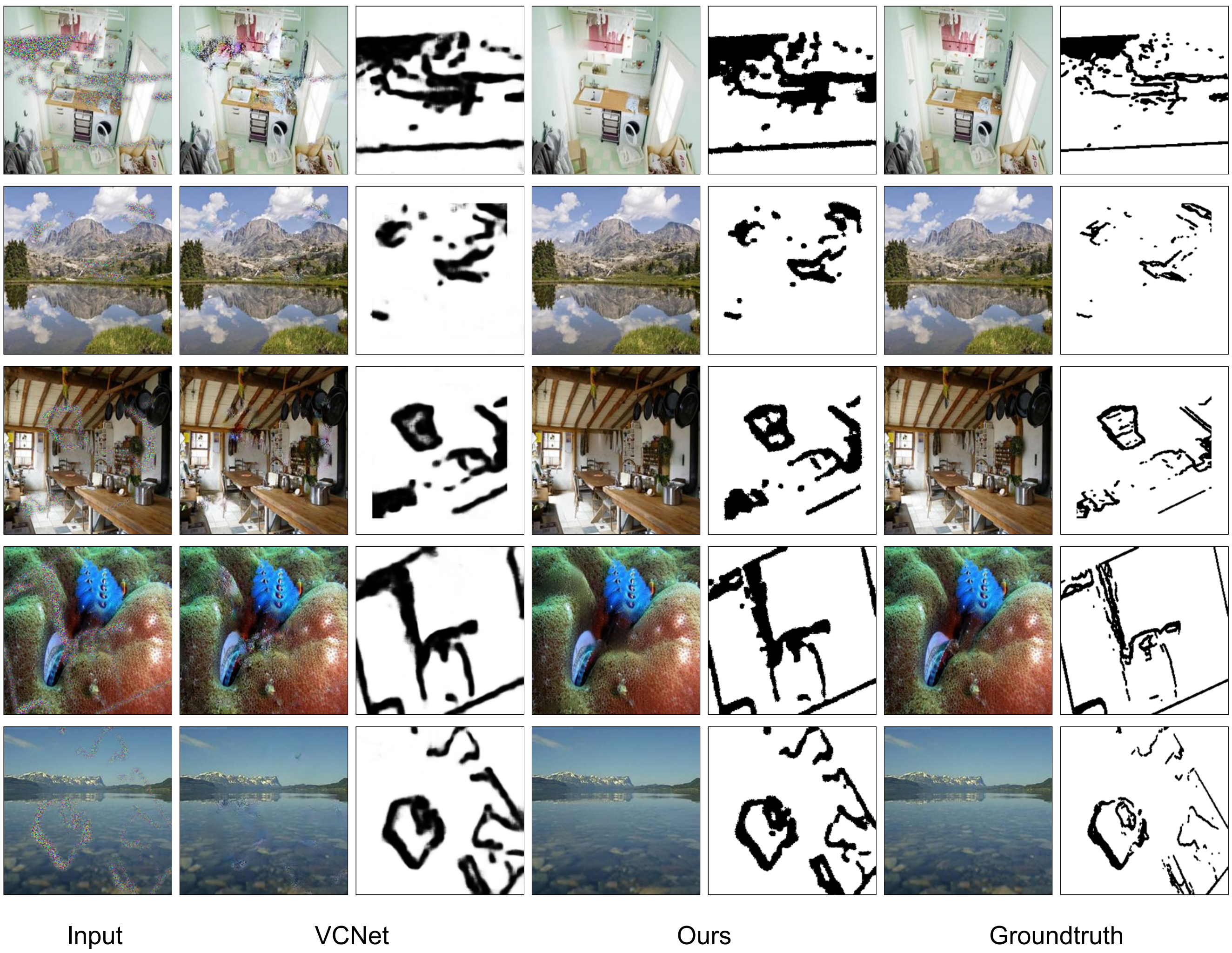


Figure 13: Visual results of model generalization to the unseen corruption: noise corruption.

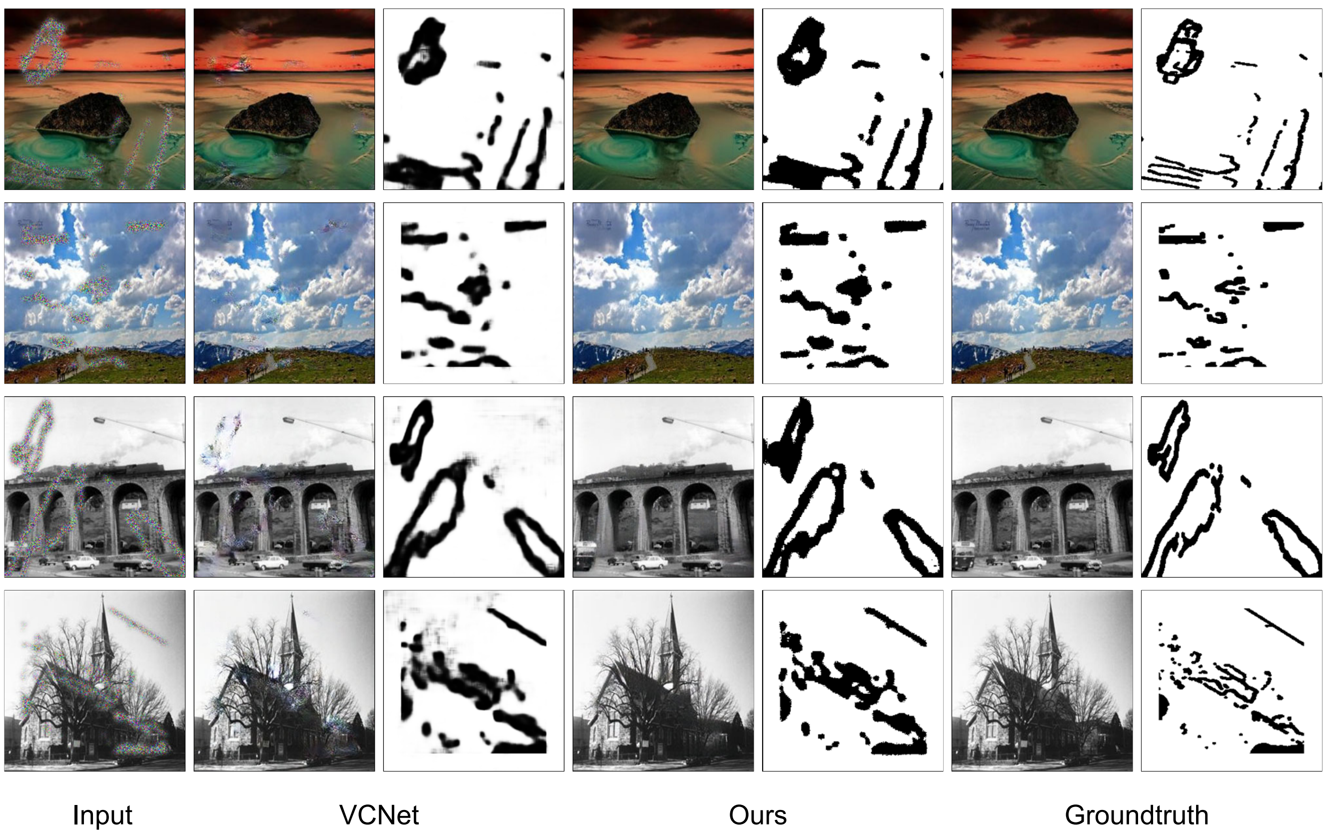
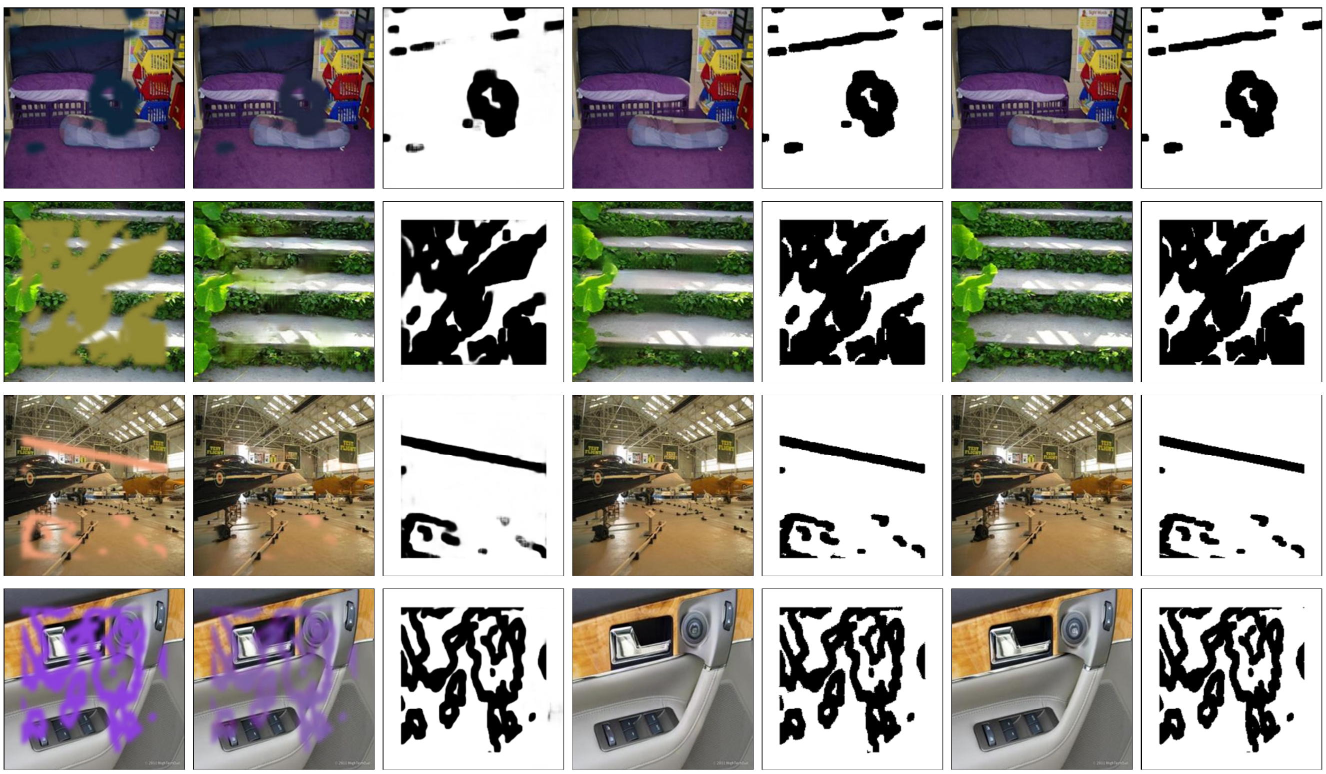


Figure 14: Visual results of model generalization to the unseen corruption: noise corruption.



Input VCNet Ours Groundtruth

Figure 15: Visual results of model generalization to the unseen corruption: single-value constant corruption.

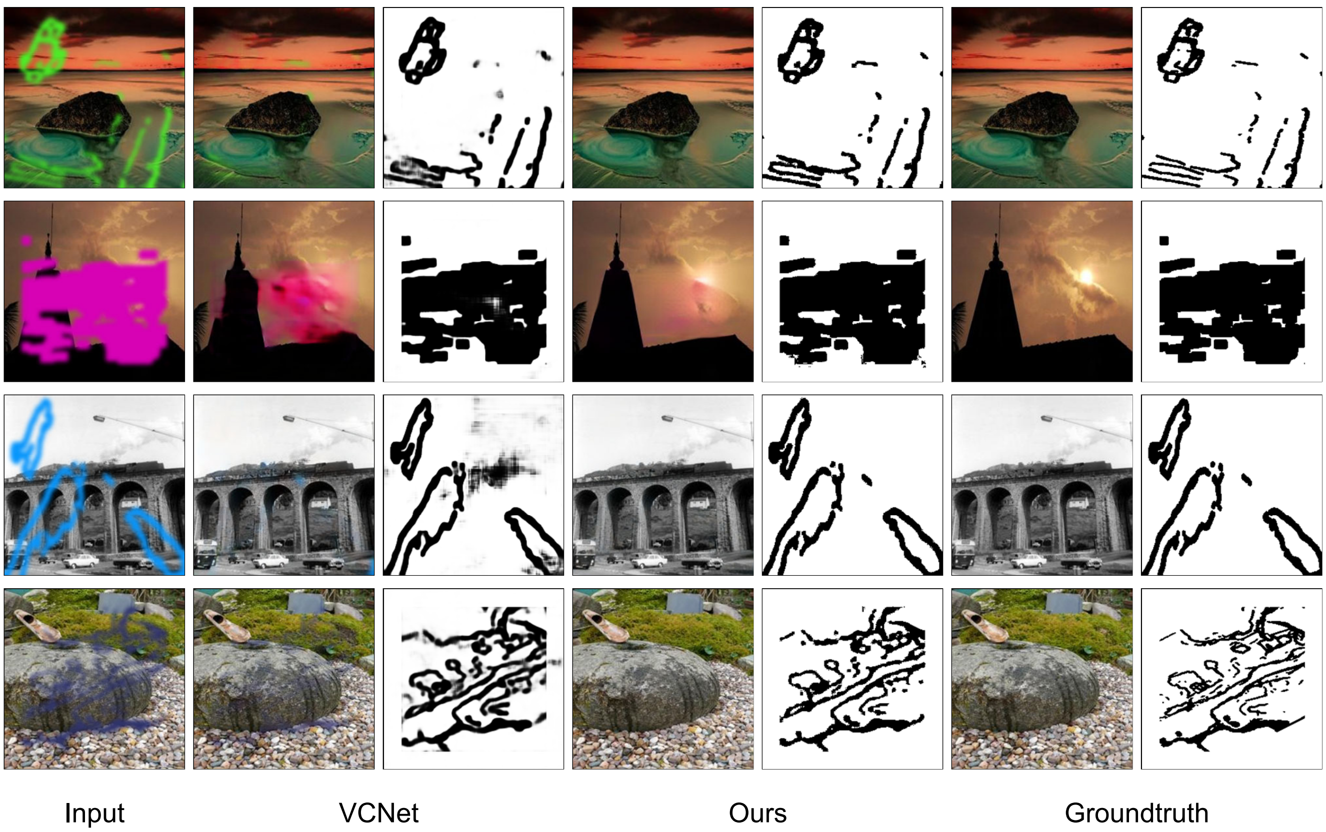


Figure 16: Visual results of model generalization to the unseen corruption: single-value constant corruption.

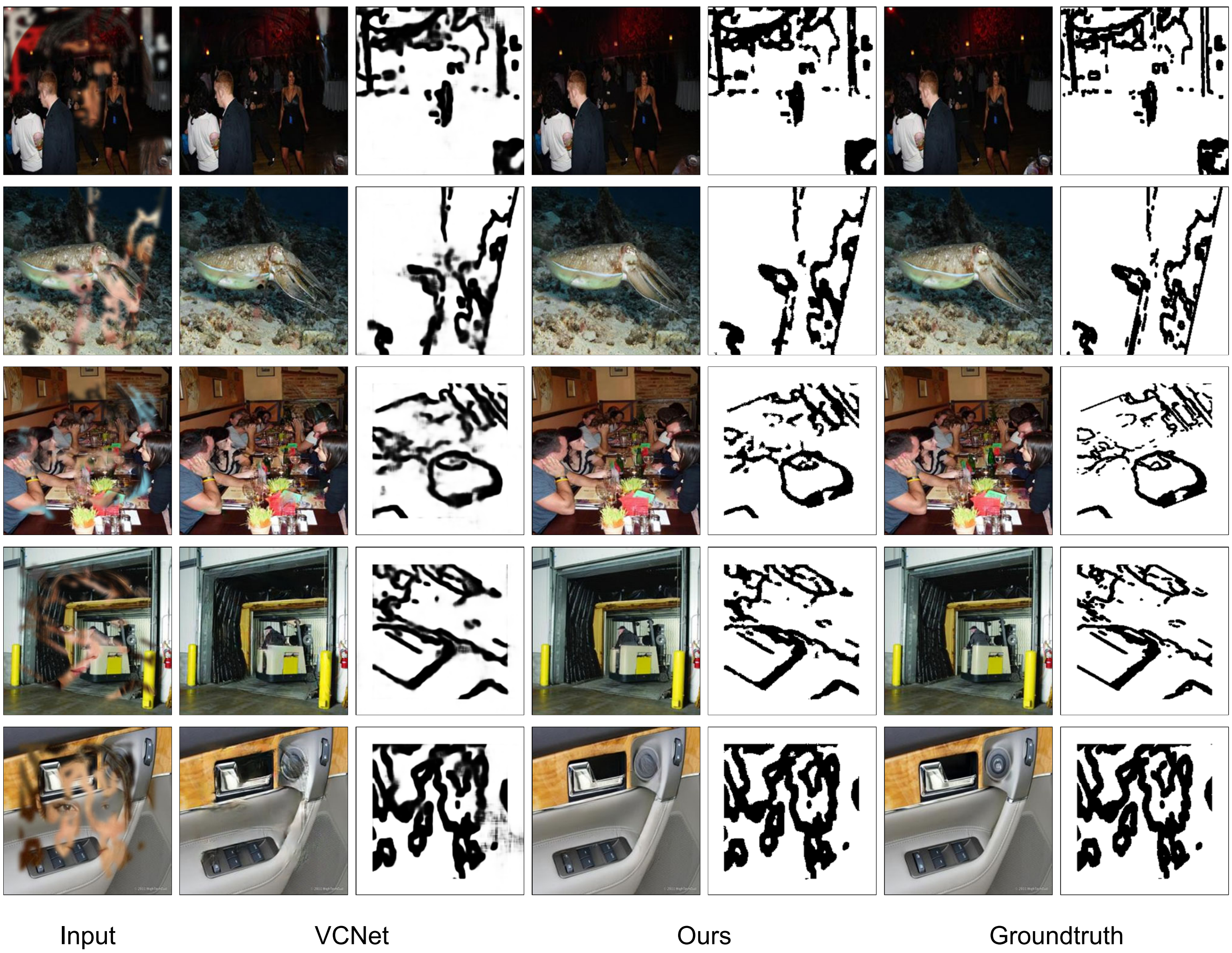


Figure 17: Visual results of model generalization to the unseen corruption: graffiti with CelebA-HQ [5] images.

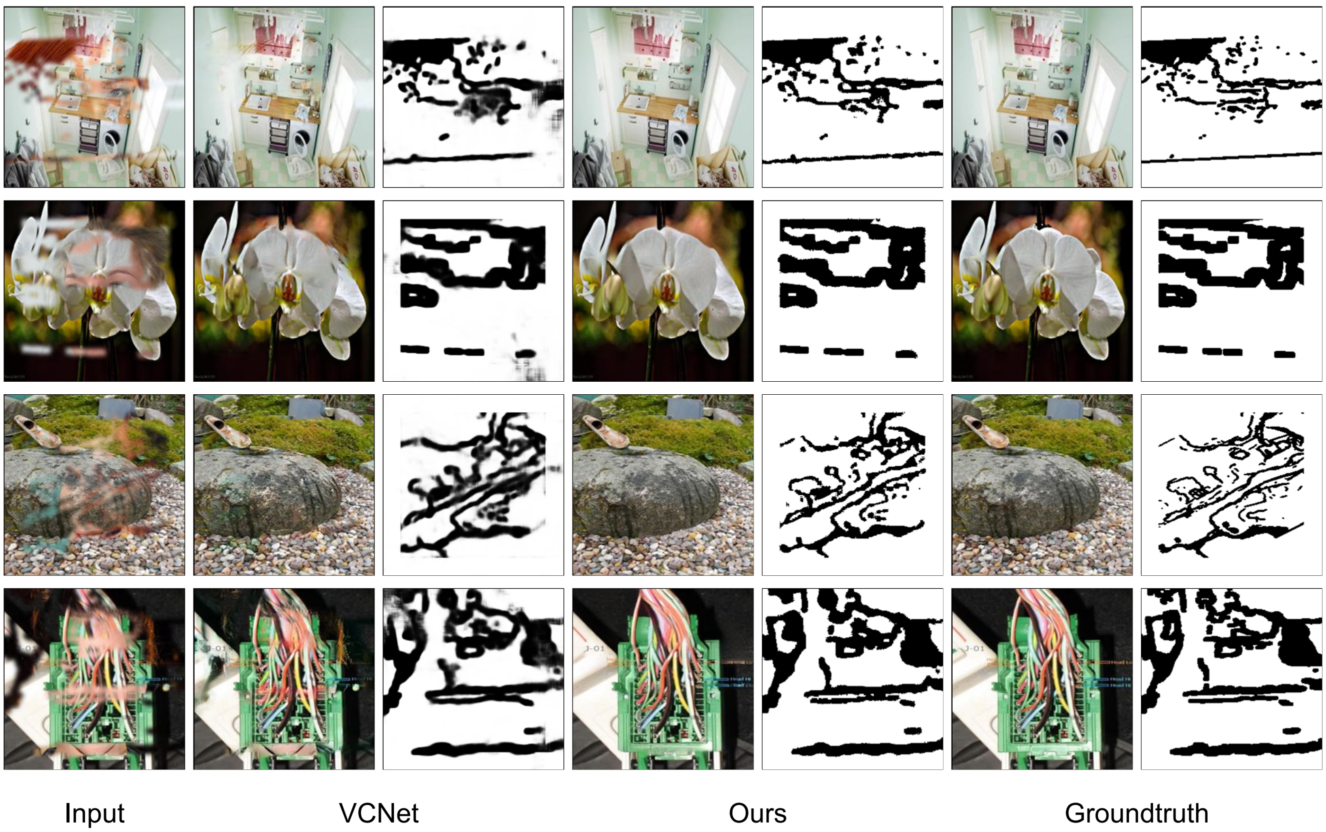


Figure 18: Visual results of model generalization to the unseen corruption: graffiti with CelebA-HQ [5] images.

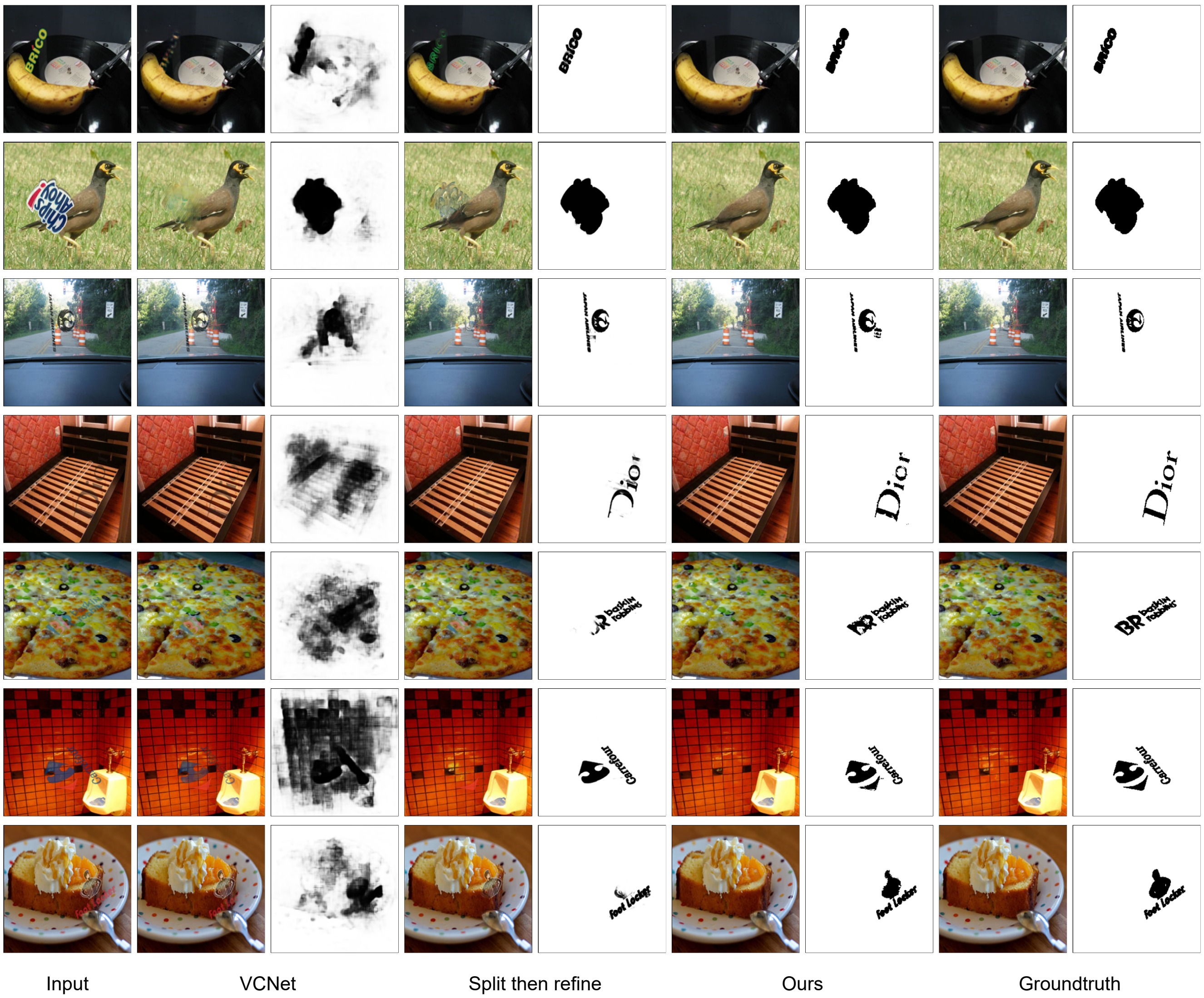


Figure 19: Visual comparison with state-of-the-art methods [9, 1] for image watermark removal. ‘Split then refine’ [1] is a specialized model for image watermark removal.

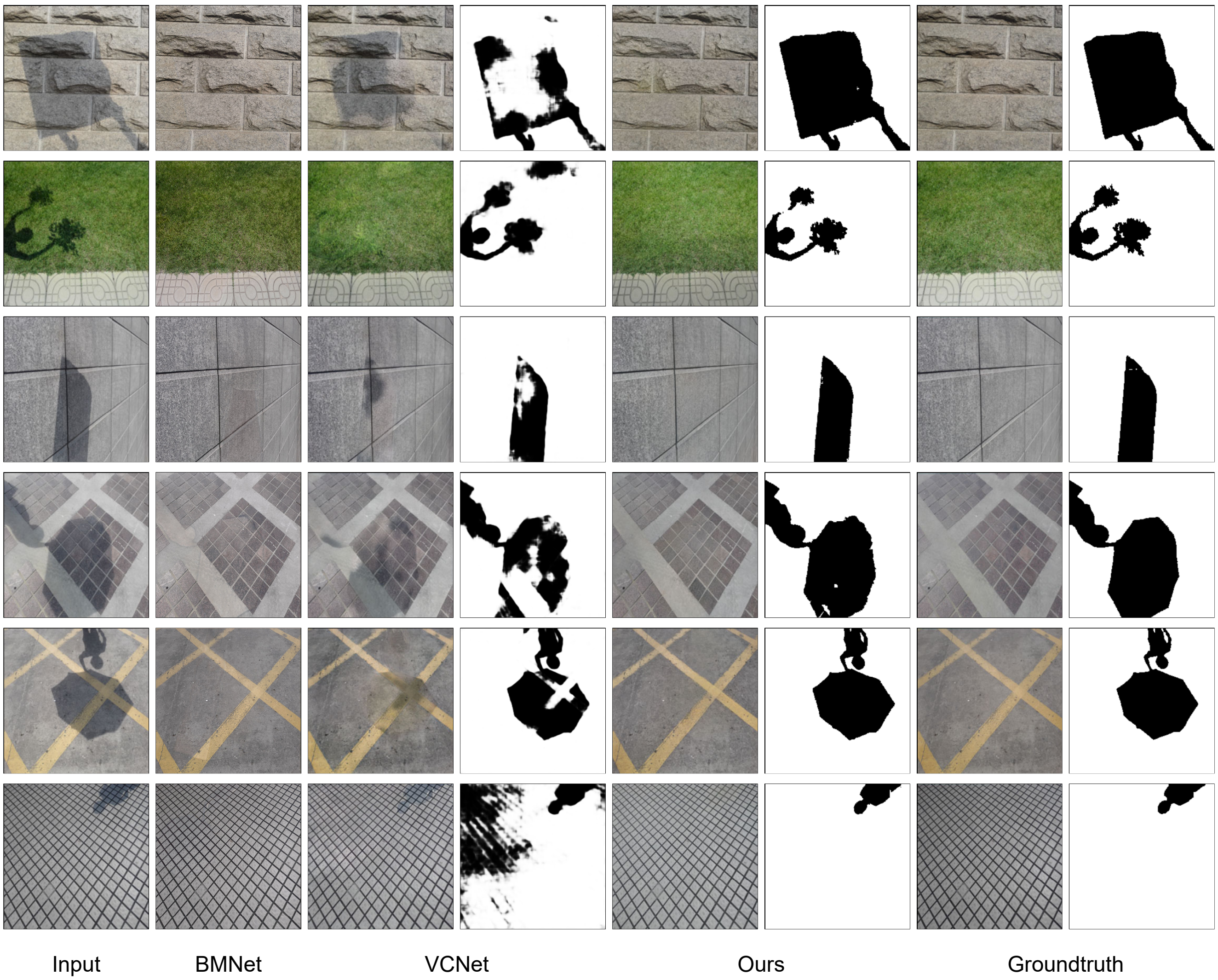


Figure 20: Visual comparison with state-of-the-art methods [9, 12] for image shadow removal. BMNet [12] is a specialized model for image watermark removal.