I See-Through You: A Framework for Removing Foreground Occlusion in Both Sparse and Dense Light Field Images Supplementary materials

Jiwan Hur*, Jae Young Lee*, Jaehyun Choi, and Junmo Kim
School of Electrical Engineering, KAIST, South Korea
{jiwan.hur, mcneato, chlwogus, junmo.kim}@kaist.ac.kr

1. Comparison between Sparsely and Densely Sampled Light Fields

Fig. 1 shows the different characteristics between sparsely sampled (sparse) and densely sampled (dense) LF images. We draw the overlapped images between center-view (CV) images and the rightmost images of LF. Both LF images have the angular resolution $(U, V) = (5, 5)$, where U and V represent the horizontal and vertical angular resolutions. The corresponding angular coordinate (u, v) is $(2, 2)$ and $(4, 2)$, respectively. It is clear that the occlusion in the sparse LF image shows larger disparities than occlusion in the dense LF image. In the visible map, with respect to the CV images $(u, v) = (2, 2)$, while the green label denotes the occluded regions visible in other views, the red label signifies the occluded regions invisible in other views. We labeled the visible maps by ourselves. In the yellow dotted circle, the dense LF image has more invisible regions even both LF images have a similar occlusion size. On the right of gray dotted lines, there are de-occlusion outputs of DeOccNet [9] and ours, both trained on the dense LF dataset. The proposed model could reconstruct clear occlusion-free CV image with reduced occlusion artifacts in both sparse and dense LF images.

The table 1 shows the comparison of the number of scenes in the sparse and dense LF datasets. Clearly, the dense LF datasets have the larger number of scenes than the sparse LF datasets because it is easier to collect the dense LF scenes using the portable LF camera than the sparse LF scenes. Thus, it is reasonable to train a model using the dense LF datasets to make the model learn various features in scenes.

*Equal contribution.

Category	Dataset	# of Scenes
Sparse LF	Stanford	30
	DeOccNet Train [9]	60
Dense LF	DUTLF [8]	1462
	DUTLF-V2 [4]	4204
	LFSD [3]	100

Table 1. The number of training and test scenes of publicly available real-world LF datasets. Dense LF datasets usually have larger number of scenes compared to sparse LF datasets.

2. Experiments on Various Fusion Methods

The proposed framework combines the LF features ($\mathbf{F}_{LF}$) to the decoder features of occlusion inpainter ($\mathbf{F}_{dc}$) to reconstruct the occlusion-free CV image through the background information from LFs as well as context information. Since the $\mathbf{F}_{LF}$ includes not only useful background information but also occlusion information, which may cause artifacts, a careful fusion method is required. We mainly focus on the attention based feature fusion methods to filter out occlusion artifact from $\mathbf{F}_{LF}$ and only background information is combined to fused features f_{Fuse}. The self-attention, which is used in LF feature extractor (LFE) of our proposed framework, is experimented as self-attention fusion (*SA fusion*). The output features of k^{th} layer of encoder in LFE (f^k_{LF}) is concatenated to the k^{th} layer of decoder in occlusion inpainter (f^k_{dc}). The self-attention output is calculated by the concatenated features and residually added to the f^k_{dc} with learnable parameter γ, which is initially set to 0.25. Our occlusion inpainter (OI) has reverse mask attention in the decoder (A_{RM}) which has feature-level information about the occlusion. Inspired by an encoder-decoder attention used in the Transformer [7], in which the decoder refers the encoder feature to generate a sentence, we design a mask-feature attention which refers the occlusion mask information to fuse the features. (*M-F fusion*). Based on the self-attention module, we replace the input of key convolution as A^k_{RM}, the k^{th} layer of reverse mask attention. Figure 2 shows the detailed architecture of each fusion methods. Even though attention based fusion methods requires more parameters and computational power, simple 1x1 convolution shows similar performance.

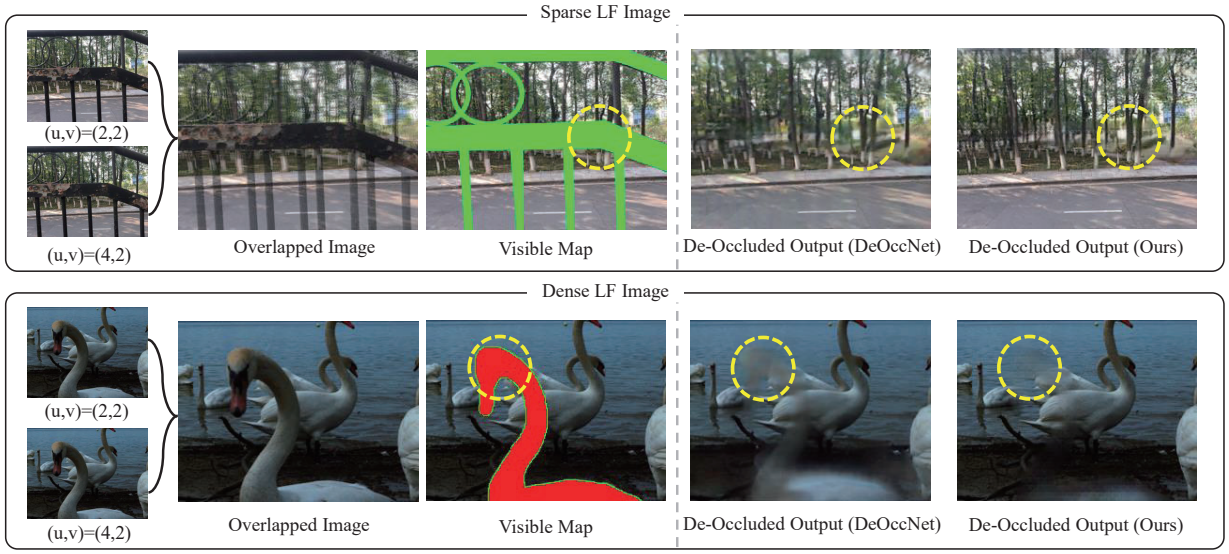


Figure 1. Illustration of the different characteristics of sparse LF image (top) and the dense LF image (bottom).

Since 1x1 convolution is more efficient and could be easily applied to other architectures, we adopt 1x1 convolution as a fusion method for proposed framework.

3. Mask Embedding Method

3.1. Light Field Reparameterization

Light field (LF) reparameterization [2, 1] can be expressed as

$$L_d(x, y, u, v) = L_0(x + ud, y + vd, u, v), \qquad (1)$$

where L_d and L_0 signify the reparameterized LF and input LF images, respectively. By controlling the disparity plane d, the zero-disparity plane (focused plane) can be moved. In the mask embedding of our training step, a single occlusion mask is copied to each view image, then the copied multiple occlusion masks are reparameterized on an arbitrary disparity plane d. By doing so, a mask can have the disparity information in the 4-D LF manifold.

3.2. Settings for Reparameterization

In the qualitative results on the dense LFs in the main manuscript, we control the zero-disparity plane of input LF images to make the foreground occlusions have positive disparity, by reparameterizing the LF scenes from EPFL-10 [6] and Stanford Lytro dataset [5]. Figs. 4 and 3 show the output de-occluded images generated by the DeOccNet* and the proposed framework from various disparity planes d in Eq. 1, where DeOccNet* denotes the DeOccNet [9] trained on the same training dataset with ours. If d is too small, foreground objects are not considered as occlusion and if d is too large, background objects are also considered as occlusion. With the proper LF reparametrization, the foreground occlusion is properly removed with the proposed framework.

3.3. Detailed Implementation of Mask Embedding

In this subsection, we describe the detailed mask embedding approach used in this paper. Although the existing methods [9] generated the mask embedded scenes before training as pre-pocessing, we generate various scenes using a set of occlusion mask templates in training time for the data augmentation.

Original mask embedding randomly embed the 1-3 masks to deal with multi disparity occlusion scenario, and they randomly select the disparity plane d from $[0, D]$, $[D, 2D]$, and $[2D, 3D]$ for the first, second and third occlusion. Different from the existing methods, we embed more occlusion in low disparity plane in order that the model could deal with dense LFs. At the same time, the model should be also trained on a large disparity occlusion scenarios to work on the sparse LFs. Therefore, we randomly select 1-3 occlusions with disparity plane from $[0, 1]$, $[1, 4]$, and $[4, 9]$, respectively, not uniformly selecting the disparity plane.

4. Evaluation Details

Qualitative results. In this part, we provide the enlarged input, output images used in qualitative results of the paper for a detailed comparison (Figs. 5, 6, 7, and 8).

Quantitative results. In this part, we provide the input, output and ground truth images used in quantitative re-

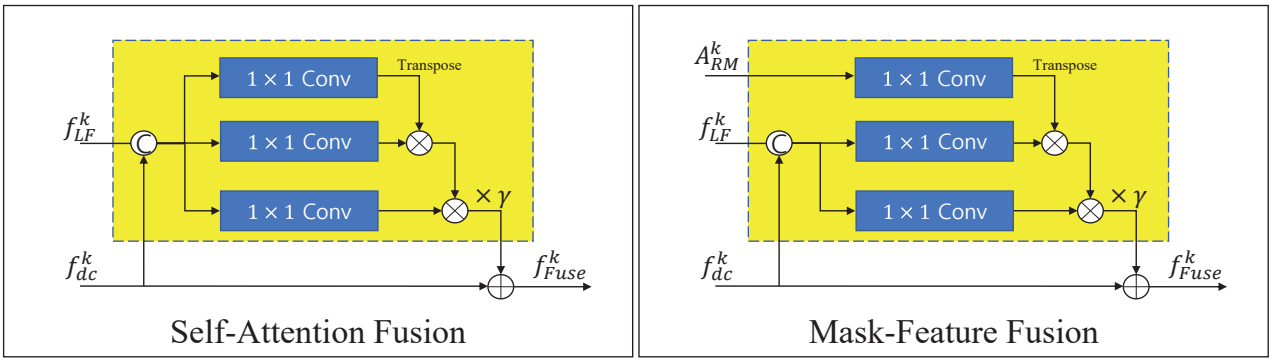


Figure 2. Illustration of the fusion methods used in ablation study, self-attention based fusion and mask-feature based fusion.

sults of the paper. We calculate peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM) for quantitative evaluation. Figs. 9 and 10 show the input CV images, ground truth occlusion-free CV images, and LF-DeOcc outputs. Note that all scenes are synthetic sparse LF images, where most of the background objects are visible in other LF views. As shown in qualitative results, our model generates the clear and accurate de-occlusion outputs with less occlusion artifacts.

5. Experimental Results on Various Scenes

In this section, we provide various real-world LF-DeOcc outputs generated by existing and proposed LF-DeOcc methods. Figs. 11 and 12 shows the LF-DeOcc outputs of the dense LF images in EPFL-10 dataset [6] and Stanford Lytro dataset [5]. Different from the existing LF-DeOcc methods, the proposed framework can remove the large occlusions in dense LF images, effectively preventing the artifacts from occlusions. However, still some limitations can be seen as discussed in the main manuscript. Since the inpainting knowledge is sub-optimal due to the relatively limited number of dataset compared to single RGB dataset (1418 scenes), the removed regions are sometimes unnatural if the occlusion is extremely large. OI pre-trained on single RGB dataset may be effective, but the inpainting knowledge is catastrophically forgot during the LF-DeOcc training time. We expect some continual learning approach may solve this problem in exchange of memory, parameters, or training times.

References

[1] J. Y. Lee and R.-H. Park. Separation of foreground and background from light field using gradient information. OSA Applied Optics, 56(4):1069–1078, 2017.

[2] M. Levoy. Light fields and computational imaging. Computer, 39(8):46–55, 2006.

[3] Nianyi Li, Jinwei Ye, Yu Ji, Haibin Ling, and Jingyi Yu. Saliency detection on light field. In 2014 IEEE Conference on Computer Vision and Pattern Recognition, pages 2806–2813, 2014.

[4] Yongri Piao, Zhengkun Rong, Shuang Xu, Miao Zhang, and Huchuan Lu. Dut-lfsaliency: Versatile dataset and light field-to-rgb saliency detection. arXiv preprint arXiv:2012.15124, 2020.

[5] A. S. Raj, M. Lowney, and R. Shah. Light-field database creation and depth estimation, 2016.

[6] Martin Rerabek and Touradj Ebrahimi. New light field image dataset. In 8th International Conference on Quality of Multimedia Experience (QoMEX), number CONF, 2016.

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. Advances in neural information processing systems, 30, 2017.

[8] Tiantian Wang, Yongri Piao, Xiao Li, Lihe Zhang, and Huchuan Lu. Deep learning for light field saliency detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 8838–8848, 2019.

[9] Yingqian Wang, Tianhao Wu, Jungang Yang, Longguang Wang, Wei An, and Yulan Guo. Deoccnet: Learning to see through foreground occlusions in light fields. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 118–127, 2020.

[10] Shuo Zhang, Zeqi Shen, and Youfang Lin. Removing foreground occlusions in light field using micro-lens dynamic filter. In Zhi-Hua Zhou, editor, Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21, pages 1302–1308. International Joint Conferences on Artificial Intelligence Organization, 8 2021. Main Track.

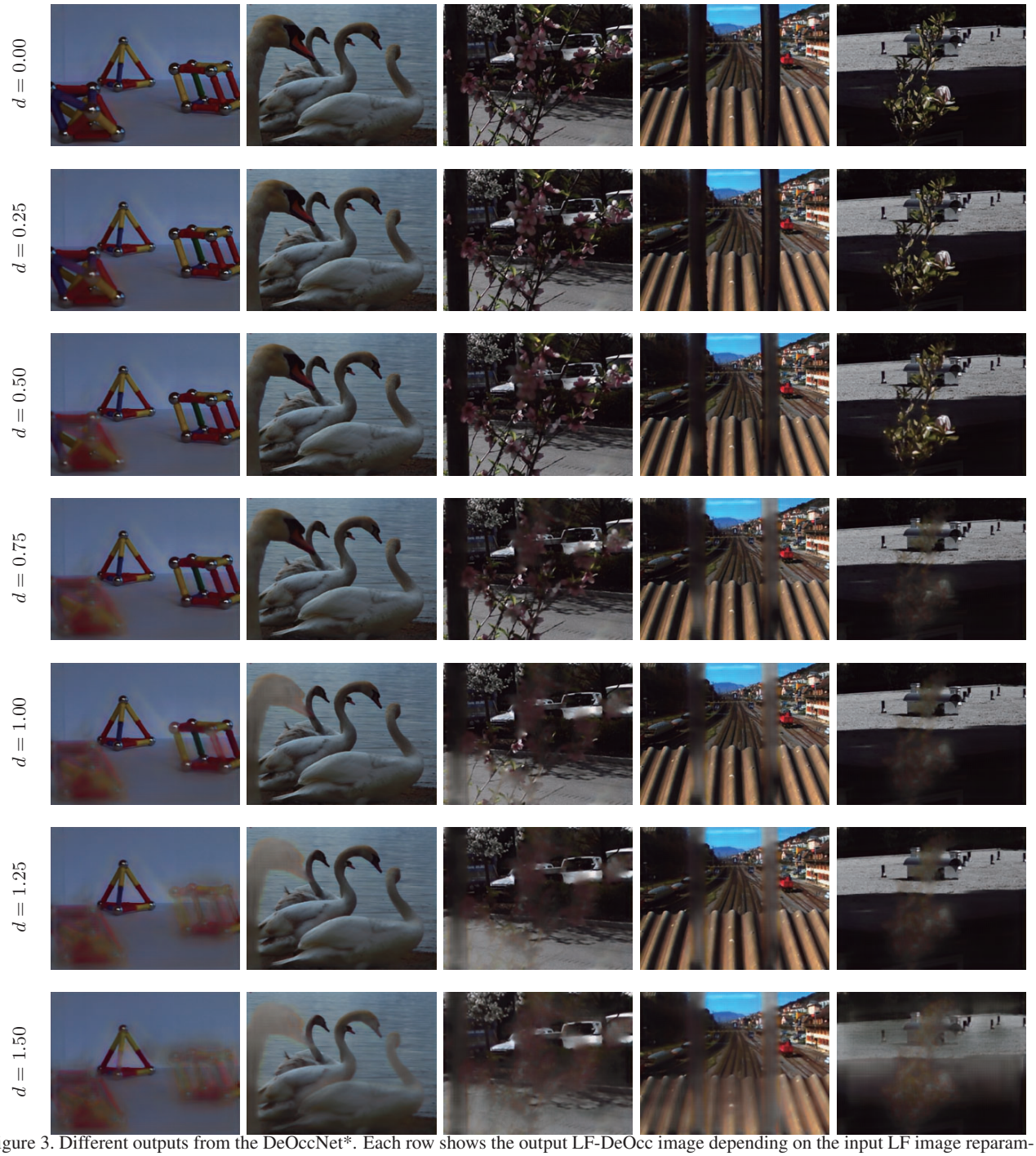


Figure 3. Different outputs from the DeOccNet*. Each row shows the output LF-DeOcc image depending on the input LF image reparameterized by the parameter d.

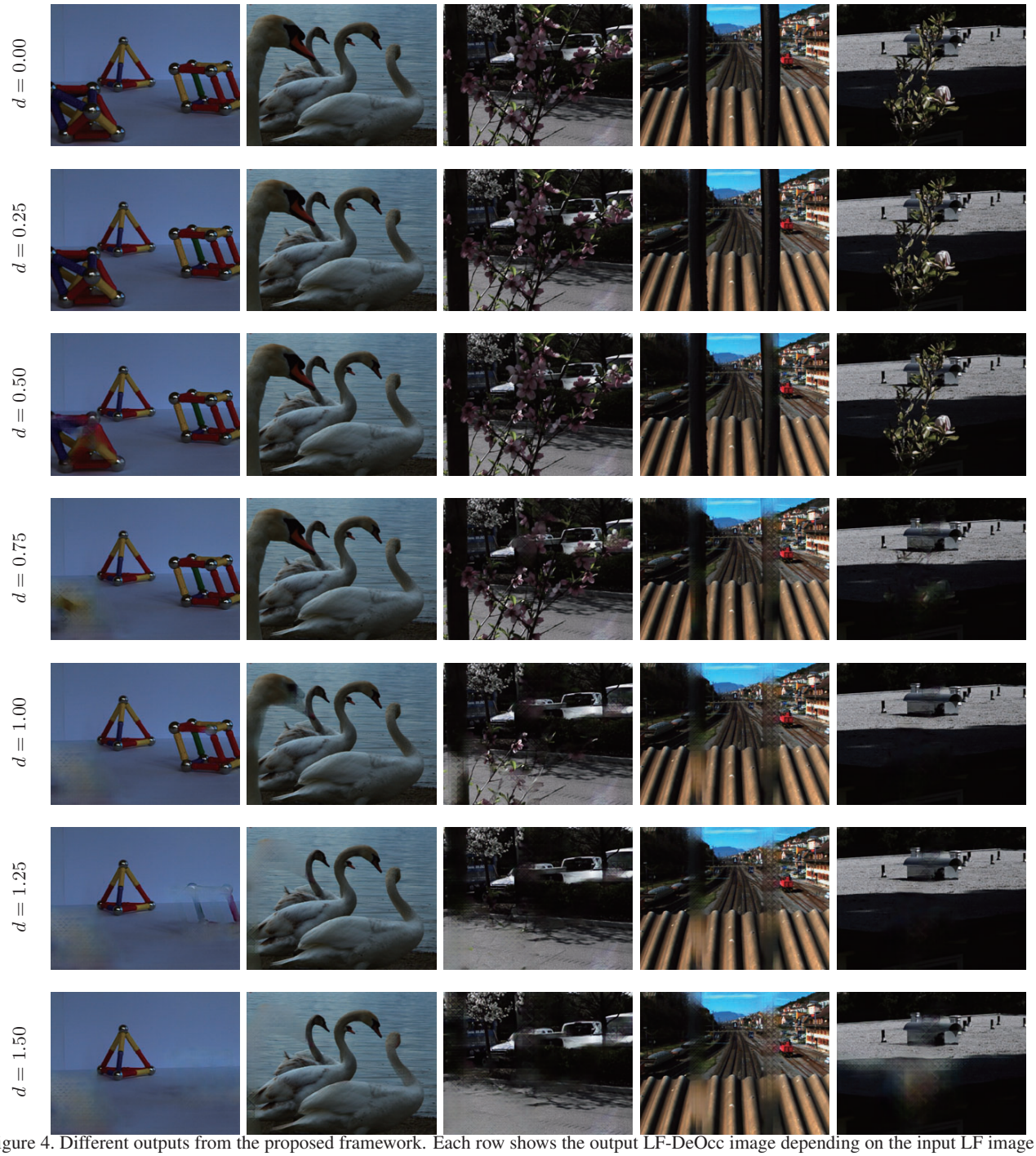


Figure 4. Different outputs from the proposed framework. Each row shows the output LF-DeOcc image depending on the input LF image reparameterized by the parameter d.

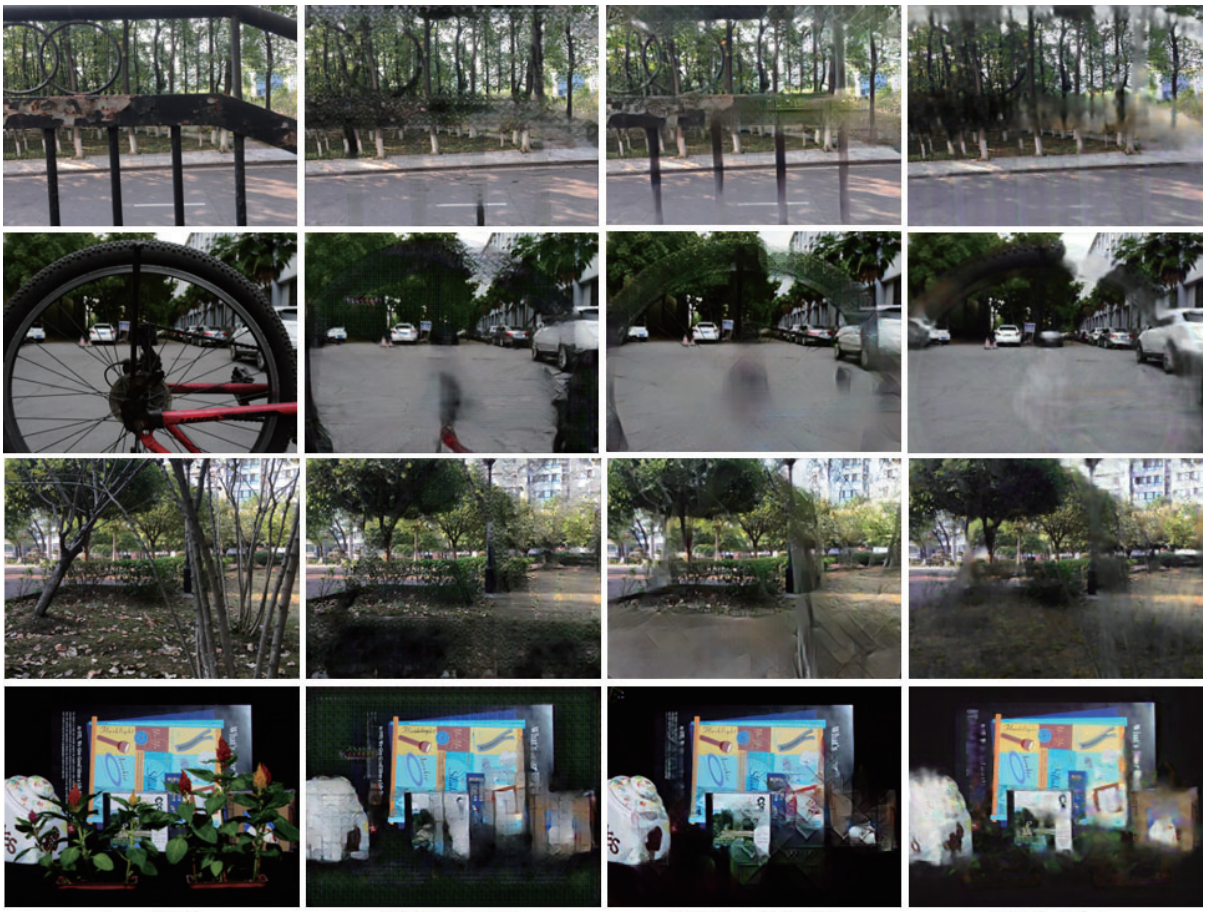


Figure 5. The enlarged outputs used in the qualitative results of main manuscript.

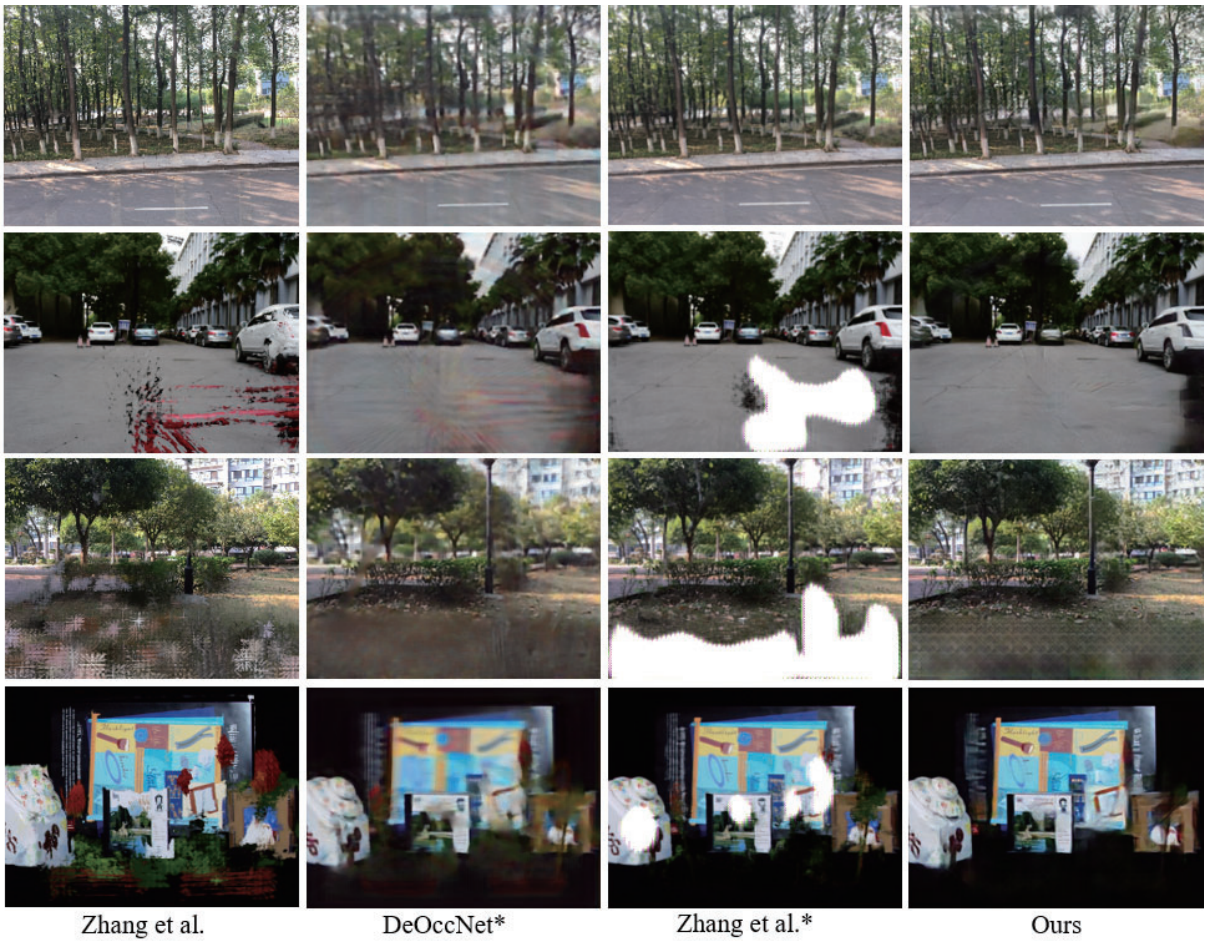


Figure 6. The enlarged outputs used in the qualitative results of main manuscript.

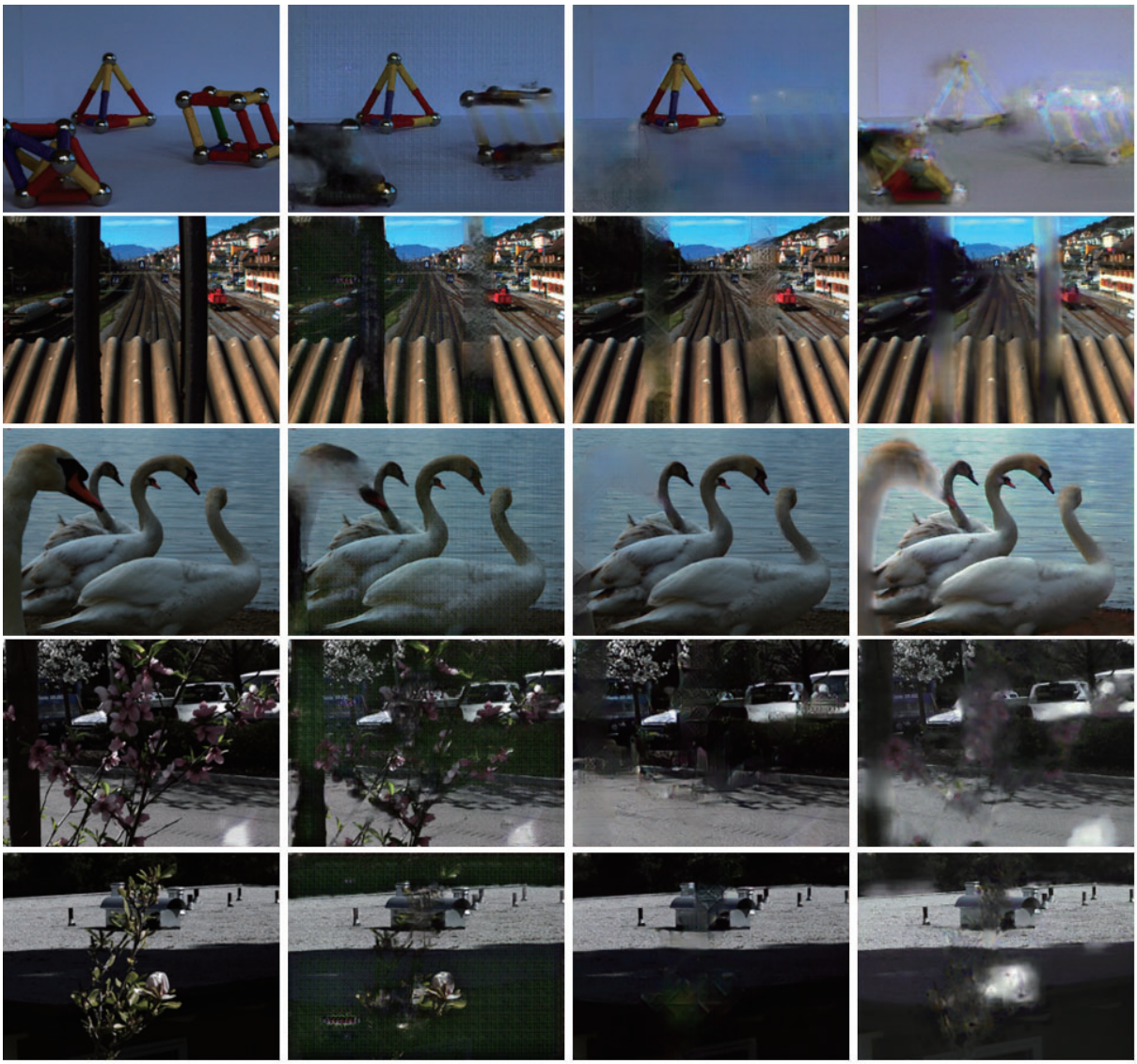


Figure 7. The enlarged outputs used in the qualitative results of main manuscript.

Figure 8. The enlarged outputs used in the qualitative results of main manuscript.

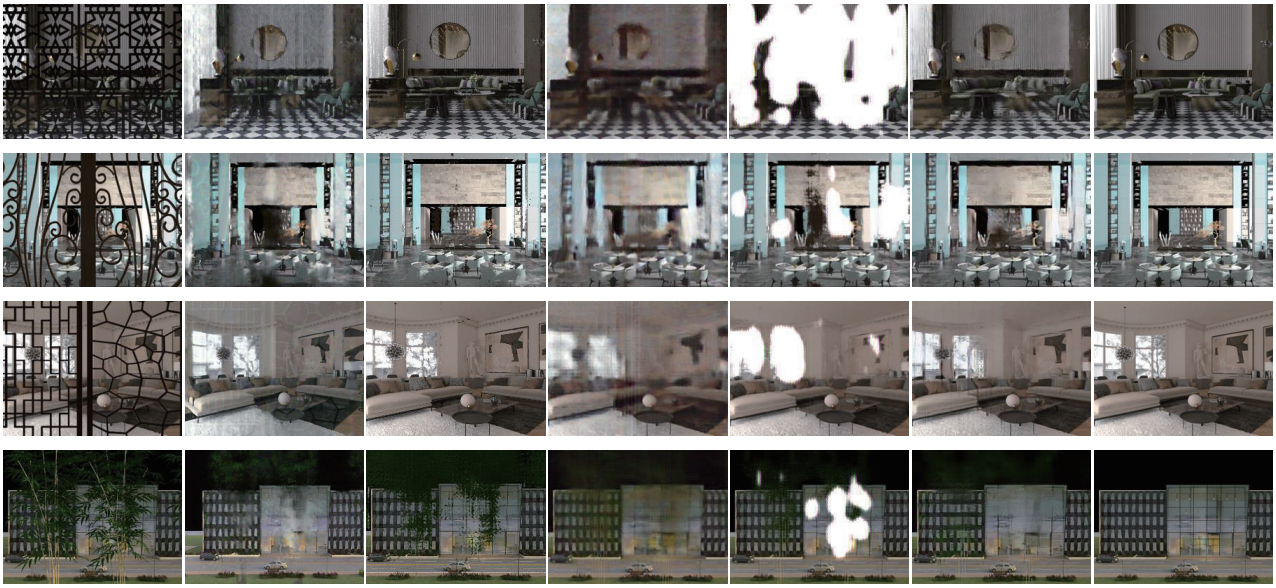
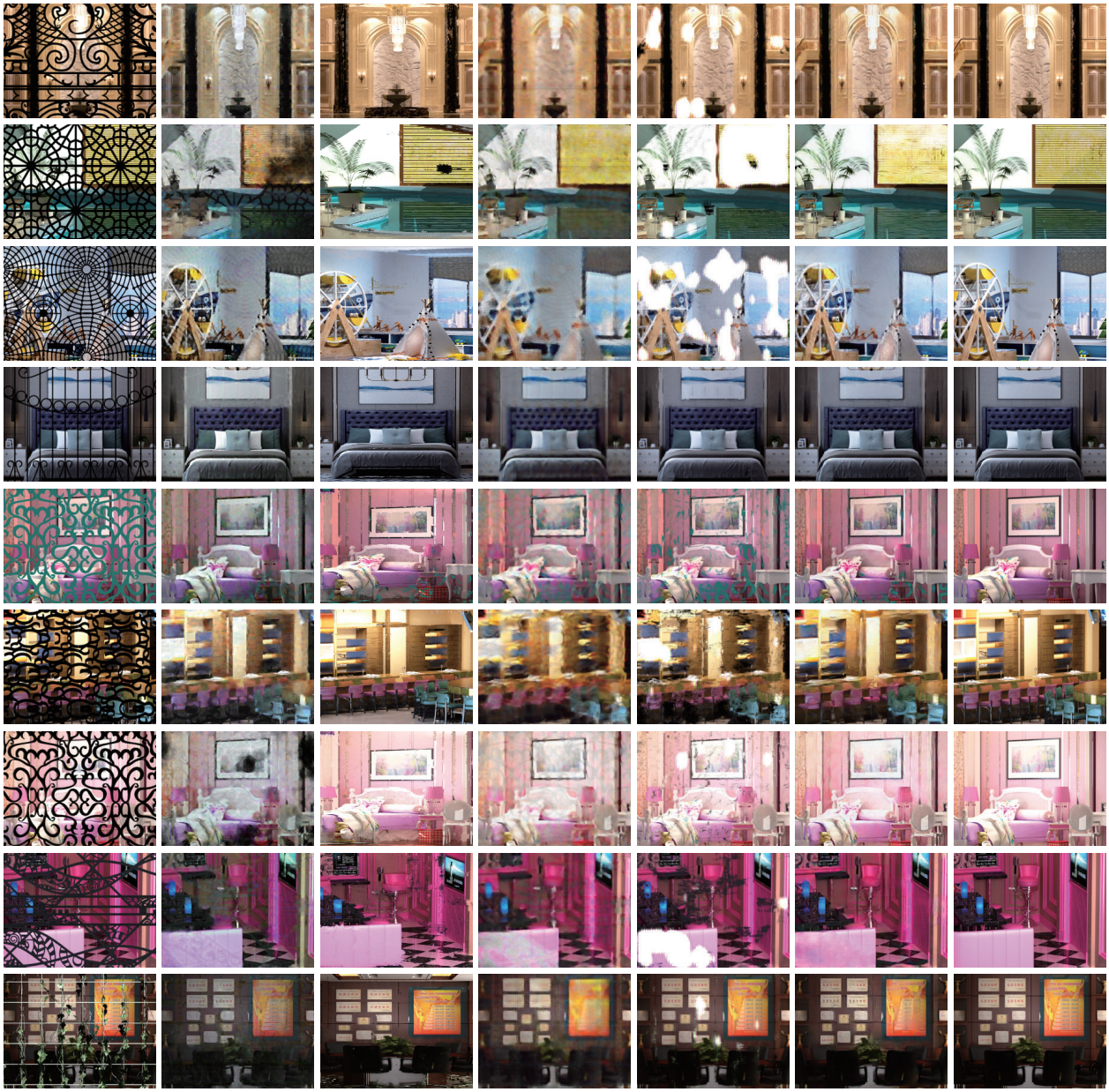
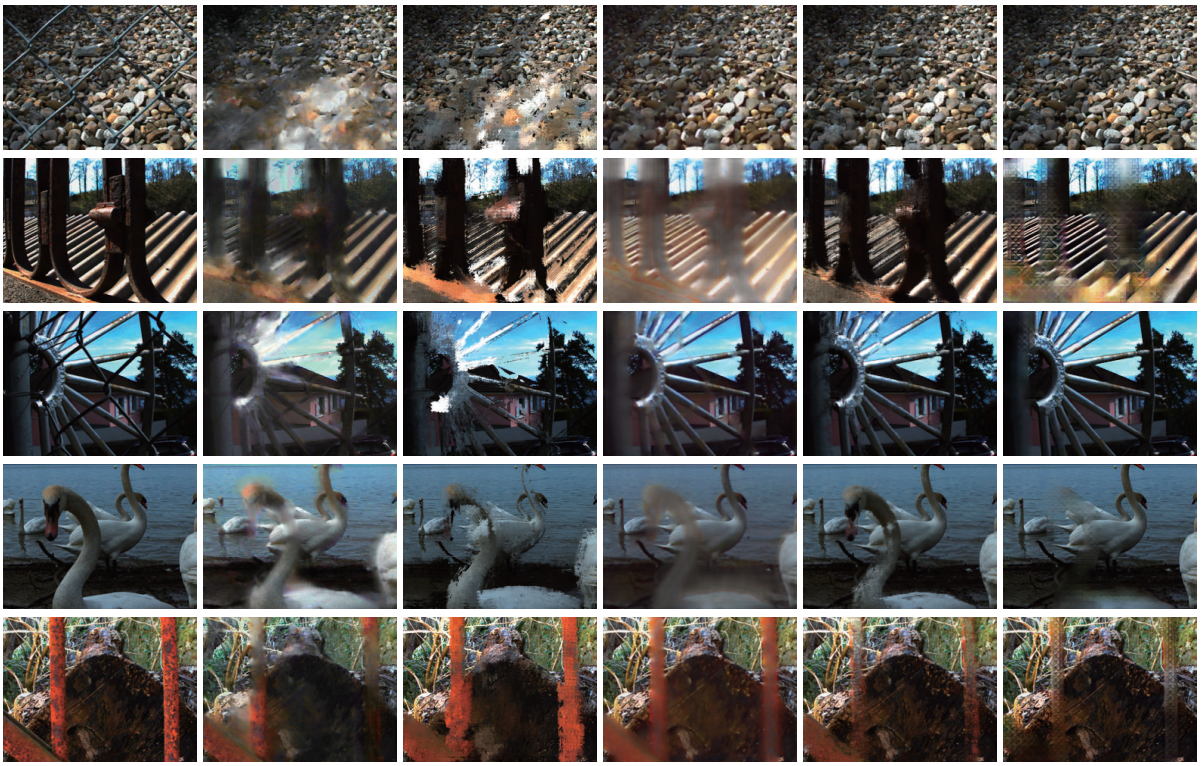


Figure 9. De-occlusion outputs on 4-syn dataset[9] using various LF-DeOcc methods, which is used for quantitative results.



Input CV DeOccNet Zhang et al. DeOccNet* Zhang et al.* Ours GT

Figure 10. De-occlusion outputs on 9-syn dataset[10] using various LF-DeOcc methods, which is used for quantitative results.



Input CV DeOccNet Zhang et al. DeOccNet* Zhang et al.* Ours

Figure 11. De-occlusion outputs on various real-world occlusion scenes in EPFL-10 dataset[6] which is a real-world dense LF dataset.

Input CV DeOccNet Zhang et al. DeOccNet* Zhang et al.* Ours

Figure 12. De-occlusion outputs on various real-world occlusion scene in Stanford Lytro dataset[5] which is a real-world dense LF dataset.