Self-improving Multiplane-to-layer Images for Novel View Synthesis

Supplementary Materials

1. Additional details

Architectures. Detailed diagrams of the steps of our pipeline are shown in Fig. 1. The architectures of the aggregating modules are described in Fig. 2. The implementation of our system is released on the website https://samsunglabs.github.io/MLI.

MLI vs LDI. Since the seminal work on layered depth images (LDI) [7], multiple papers considered equipping each pixel of the reference image with a stack of depth values. However, the original definition of this representation [7] assumed that the size of this stack may differ between pixels. In addition, any connections between pixels were not imposed. Although later manuscripts introduced explicit local connectivity of neighboring pixels [8], we stick to another terminology and refer to our representation as a layered mesh [1] or a multilayer image (MLI) [4]. The latter name is preferred, as it reveals the relation to multiplane images [11]: just like them, our proxy geometry contains semitransparent RGBA textures. On the contrary, many methods that have been reported to employ LDI representation do not use the opacity channel [2, 9, 8, 5]. Besides, MLI contains a predefined number of layers; therefore, each pixel gets the same number of depth values. While the layers are non-overlapping by design, a ray from a novel camera can intersect each layer in several points, justifying the usage of z-buffer during the rasterization step. In contrast, the original LDI representation did not need the z-buffer, and McMillan's warp ordering algorithm was used instead.

2. Additional results

Fig. 3 demonstrates the results of our SIMPLI method in the case of 4 layers. Also we provide an additional comparison with the baseline methods on publicly available datasets [10, 6]. We show visual results for 2 input views in Fig. 4, for 5 input views in Fig. 5 and for 8 – in Fig. 6. These results correlate with the metrics reported in the main text. For two or five input views, our method clearly outperforms all baselines and produces the most visually pleasant results. Also, most of the crops for the eight input images show that our method is at least on par with existing approaches. Note that all the demonstrated crops and scenes are uncurated.

Fig. 7 demonstrates a comparison of our model with the DeepView system [3] on the Spaces dataset [3]. We show the results for the small and large camera baselines separately. As may be seen from the figures, our model produces slightly blurrier results than DeepView does. However, it allows us to get a much more compact scene representation, as was discussed in the main text. We demonstrate the visual comparison for SIMPLI with a different number of layers in the MLI representation in Figs. 4 to 6. We observe the degradation of the quality with decreasing of the number of layers in the final representation.

References

[1] Michael Broxton, John Flynn, Ryan Overbeck, Daniel Erickson, Peter Hedman, Matthew Duvall, Jason Dourgarian, Jay Busch, Matt Whalen, and Paul Debevec. Immersive light field video with a layered mesh representation. In *ACM TOG*, 2020. 1

[2] Helisa Dhamo, Keisuke Tateno, Iro Laina, Nassir Navab, and Federico Tombari. Peeking behind objects: Layered depth prediction from a single image. *Pattern Recognition Letters*, 125:333–340, 2019. 1

[3] John Flynn, Michael Broxton, Paul Debevec, Matthew DuVall, Graham Fyffe, Ryan Styles Overbeck, Noah Snavely, and Richard Tucker. Deepview: High-quality view synthesis by learned gradient descent. In *CVPR*, 2019. 1

[4] Taras Khakhulin, Denis Korzhenkov, Pavel Solovev, Gleb Sterkin, Timotei Ardelean, and Victor Lempitsky. Stereo magnification with multi-layer images. In *CVPR*, 2022. 1

[5] Johannes Kopf, Kevin Matzen, Suhib Alsisan, Ocean Quigley, Francis Ge, Yangming Chong, Josh Patterson, Jan-Michael Frahm, Shu Wu, Matthew Yu, Peizhao Zhang, Zijian He, Peter Vajda, Ayush Saraf, and Michael Cohen. One shot 3d photography. In *ACM TOG*, 2020. 1

[6] Ben Mildenhall, Pratul P. Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. In *ACM TOG*, 2019. 1

[7] Jonathan Shade, Steven Gortler, Li-wei He, and Richard Szeliski. Layered depth images. In *ACM TOG*, 1998. 1

[8] M. L. Shih, S. Y. Su, J. Kopf, and J. B. Huang. 3d photography using context-aware layered depth inpainting. In *CVPR*, 2020. 1

[9] Shubham Tulsiani, Richard Tucker, and Noah Snavely. Layer-structured 3d scene inference via view synthesis. In *ECCV*, 2018. 1

[10] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P. Srinivasan, Howard Zhou, Jonathan T. Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. IBRNet: Learning Multi-View Image-Based Rendering. In *CVPR*, 2021. 1

[11] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. In *ACM TOG*, 2018. 1

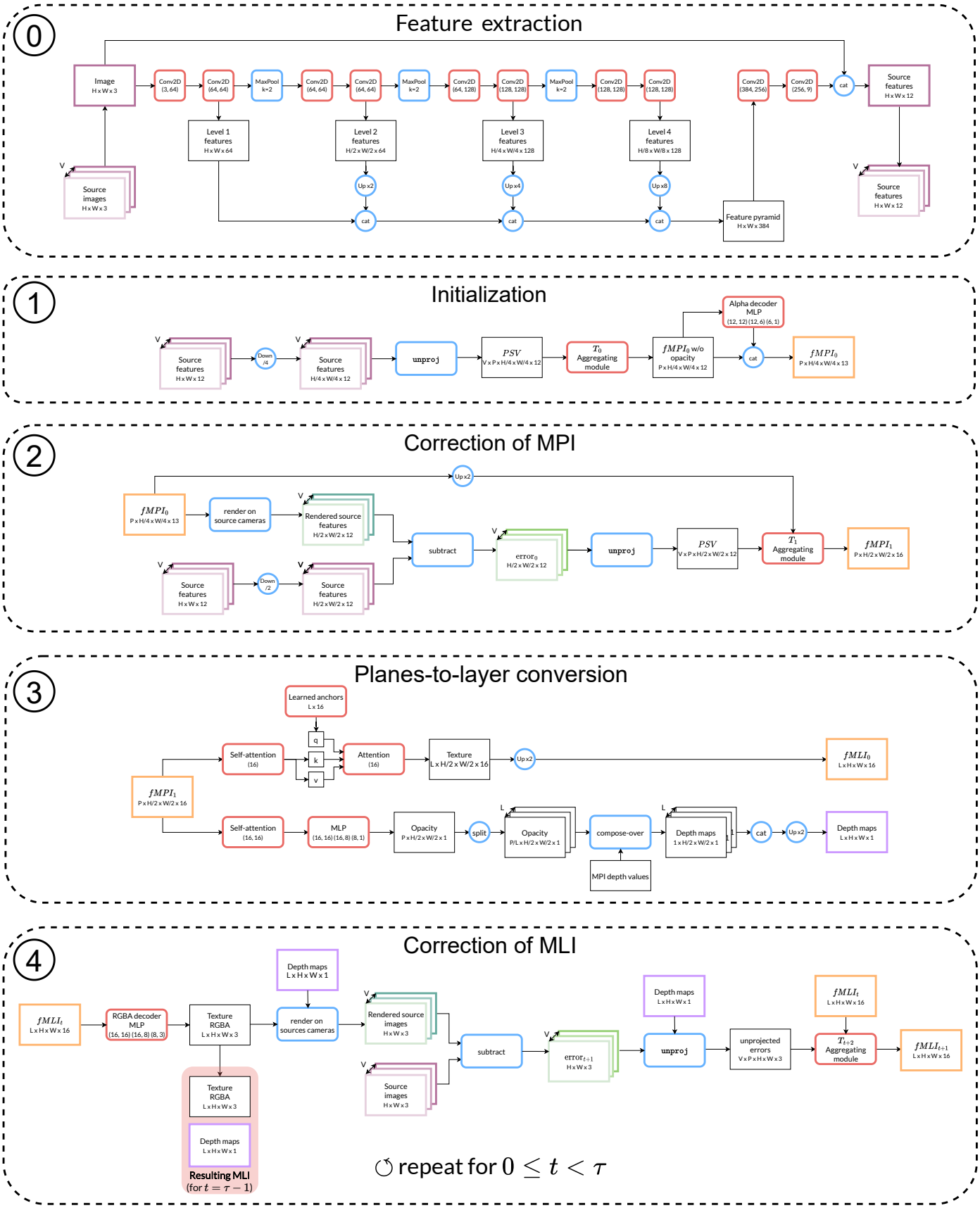

Figure 1. The detailed diagram of our pipeline. Please zoom in for details.

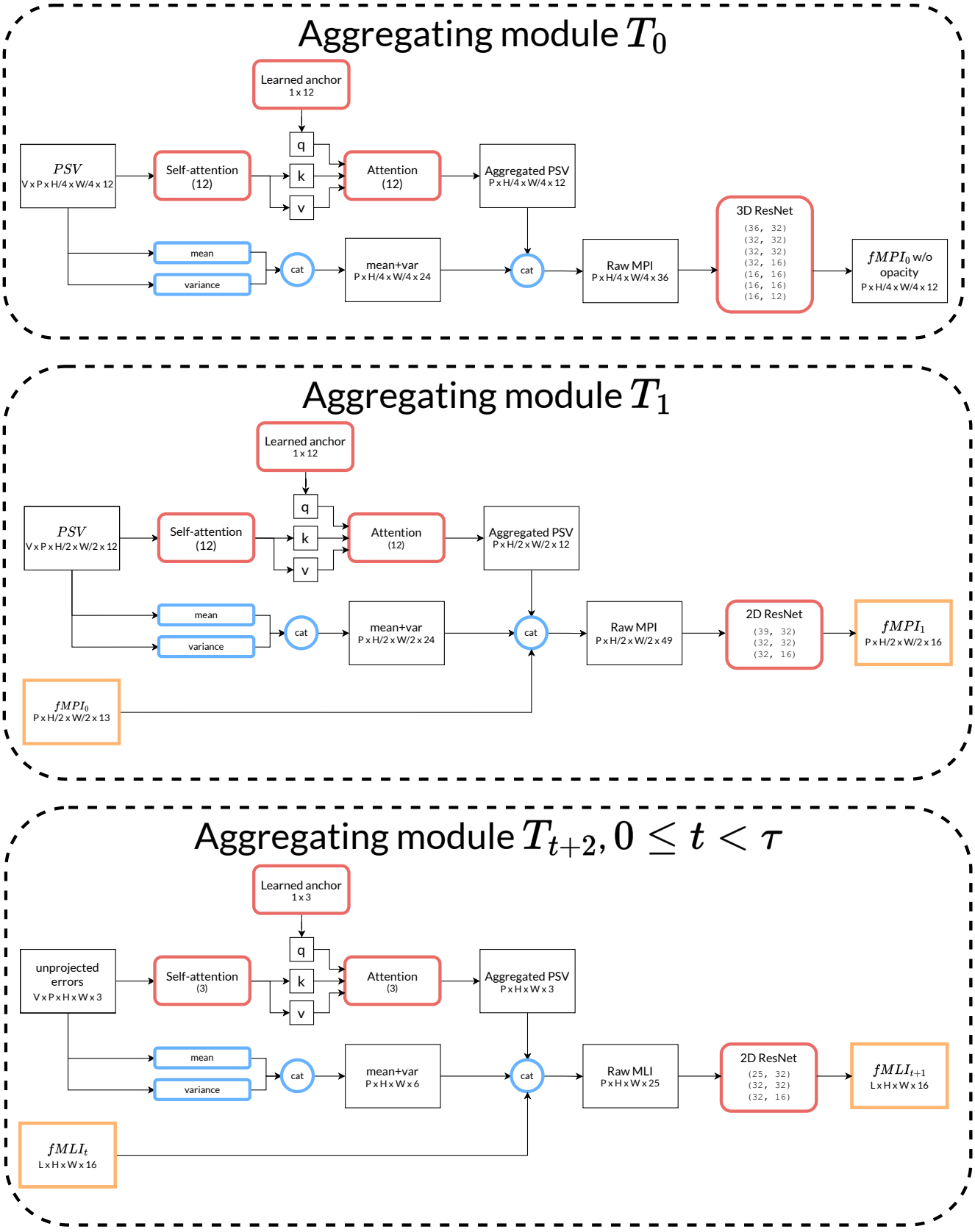

Figure 2. Architecture of aggregating modules. Please zoom in for details.

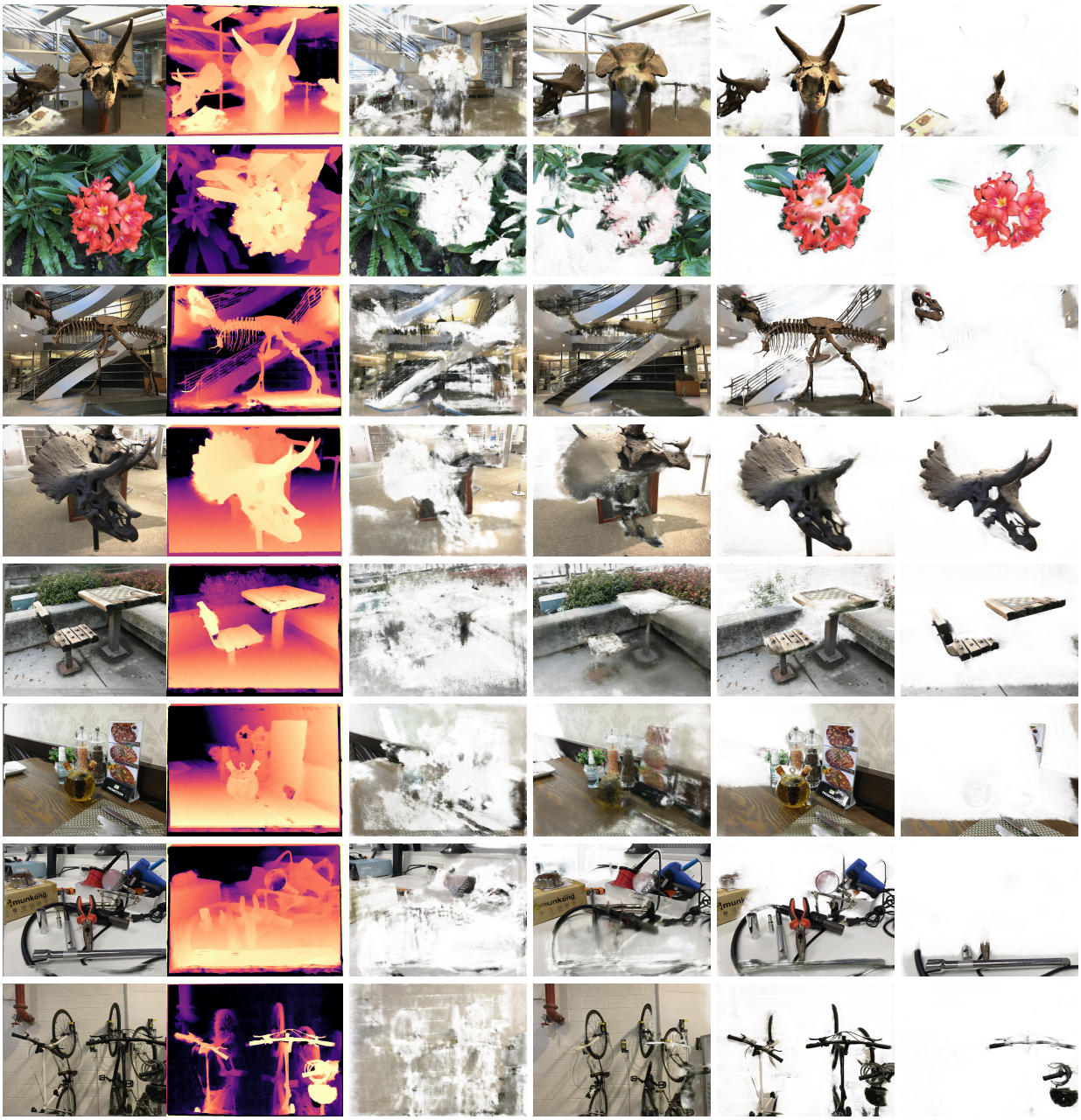

Figure 3. Extension of Fig. 1 from the main text. The textures of MLI representation with 4 deformable layers estimated by SIMPLI. Left to right: generated novel view, corresponding depth map, four semitransparent textures in the back-to-front order. The inferred depth map is computed by overcomposing the per-layer depth maps w.r.t. the opacity extracted from the corresponding RGBA textures.

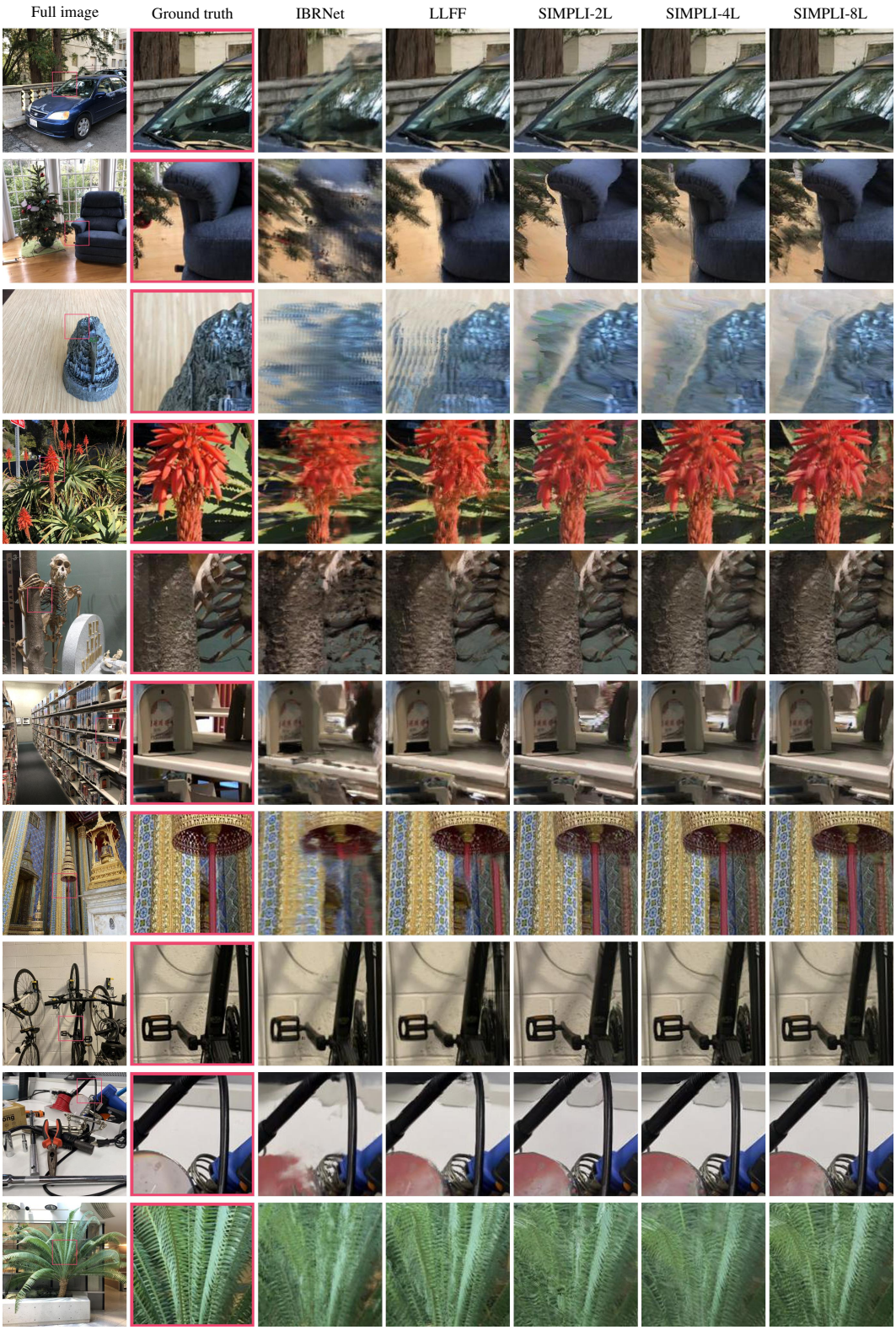

Figure 4. Results for real novel cameras with **two** input views given. Note that IBRNet cannot produce any information for areas unobserved from the source views.

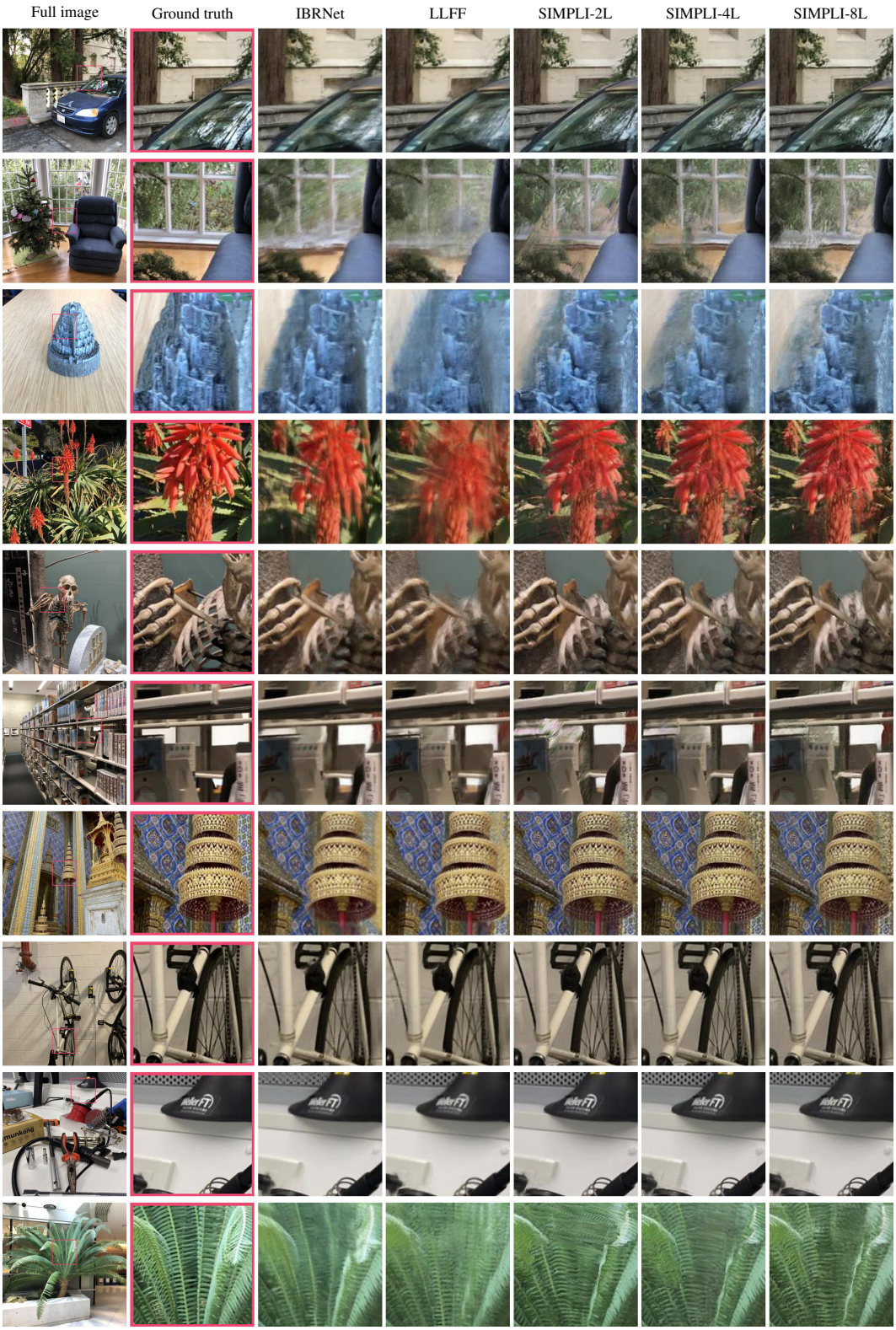

Figure 5. Results for real novel cameras with **five** input views given. The outputs of SIMPLI-8L are the most similar to the ground truth frames and have less artifacts than other models.

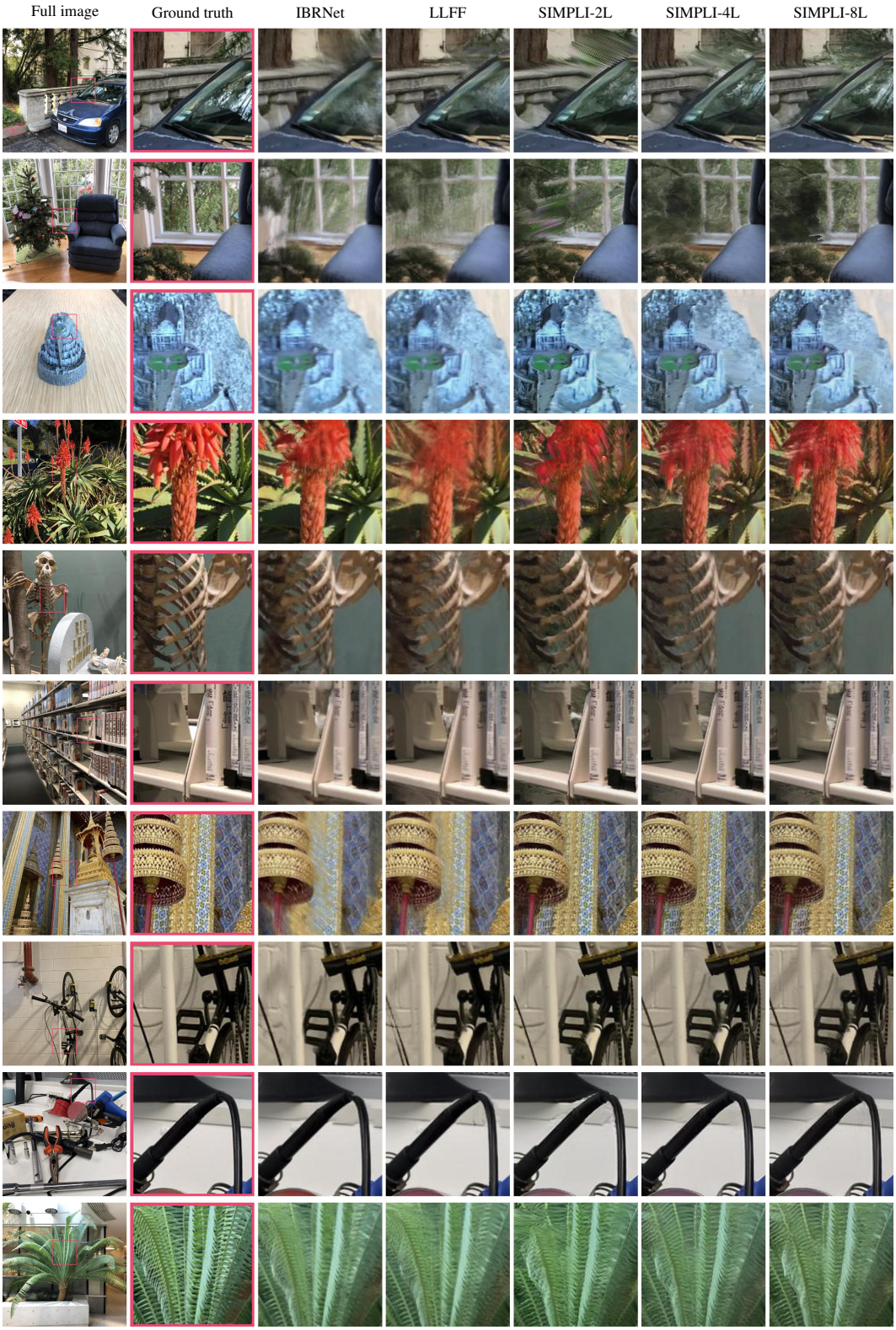

Figure 6. Results for real novel cameras with **eight** input views given. This is the most competitive scenario. Note that both IBRNet and LLFF tend to produce many artifacts, *e.g*. blurriness, while frames rendered by SIMPLI are more sharp.

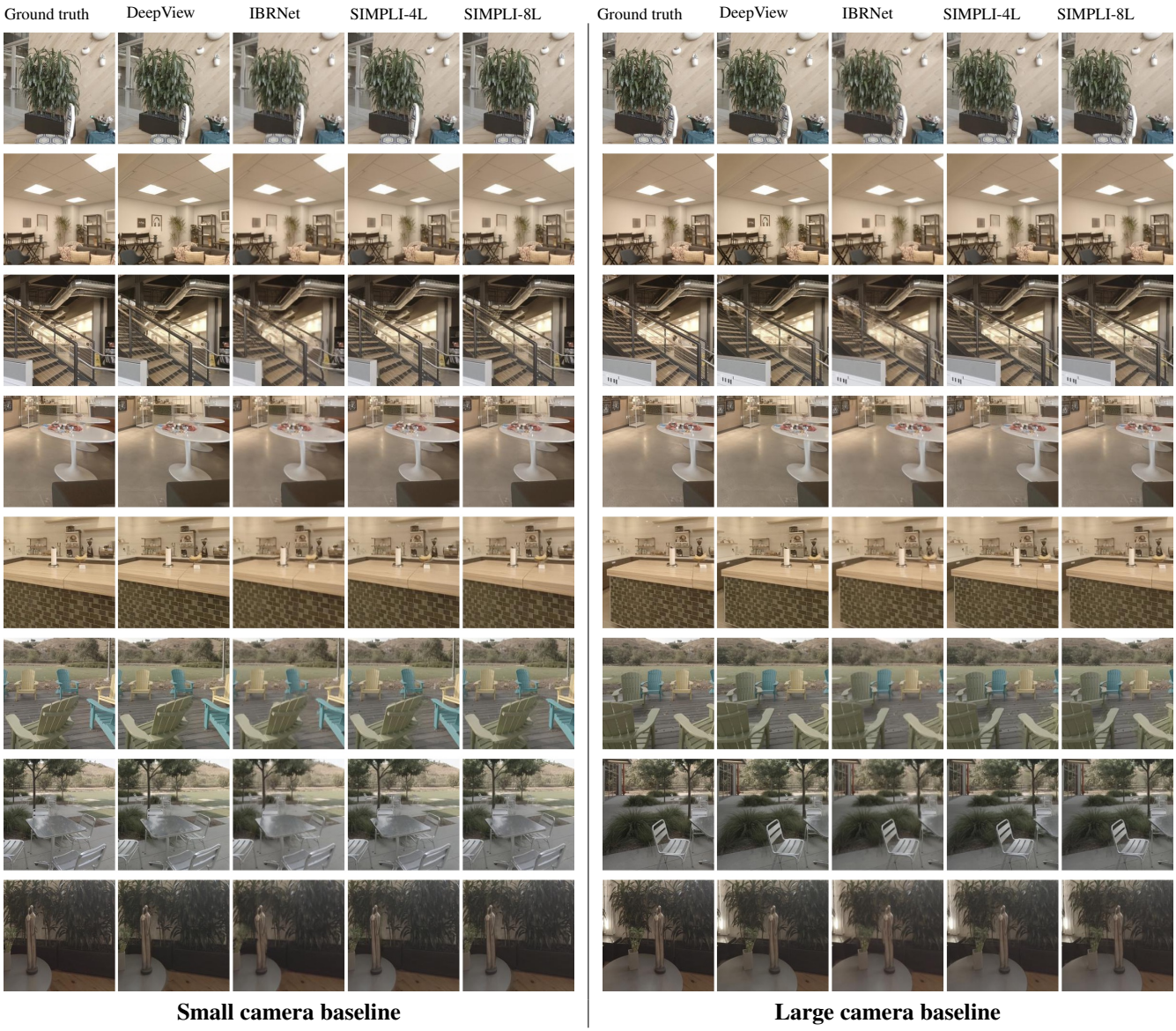

Figure 7. Results for the Spaces dataset for DeepView (40 planes) and SIMPLI with 4 and 8 layers. Our model produces more blurry and less bright results, trading off for more compact representation of the scene.