Aperture Supervision for Monocular Depth Estimation: Supplementary Material

Pratul P. Srinivasan[1] * Rahul Garg[2] Neal Wadhwa[2] Ren Ng[1] Jonathan T. Barron[2]
[1]UC Berkeley, [2]Google Research

1. Network Architectures

Here, we provide detailed descriptions of the convolutional neural network (CNN) architectures used in our method.

$C_i S_j D_k$ denotes a convolution layer with i 3x3 filters, a spatial stride of j in each dimension, and a dilation rate of k. Additionally, E denotes an exponential linear unit activation function [1] and I denotes instance normalization [2]. Finally, R denotes a residual connection where the current tensor is added to the tensor output from the previous instance normalization layer.

For our light field dataset experiments, the monocular depth estimation CNN $f_{\boldsymbol{\theta}_d}(\cdot)$ contains 12 convolutional layers structured as:

$$C_8S_1D_1\text{-}E\text{-}I\text{-}C_{32}S_1D_1\text{-}E\text{-}I\text{-}C_{64}S_1D_1\text{-}E\text{-}I\text{-}C_{128}S_1D_1\text{-}E\text{-}I\text{-}C_{128}S_1D_2\text{-}E\text{-}I\text{-}R\text{-}C_{128}S_1D_4\text{-}E\text{-}I\text{-}R\text{-}C_{128}S_1D_8\text{-}E\text{-}I\text{-}R\text{-}C_{128}S_1D_{16}\text{-}E\text{-}I\text{-}R\text{-}C_{128}S_1D_{32}\text{-}E\text{-}I\text{-}R\text{-}C_{64}S_1D_1\text{-}E\text{-}I\text{-}C_{32}S_1D_1\text{-}E\text{-}I\text{-}C_1S_1D_1. \quad (1)$$

For our SLR dataset experiments, the monocular depth estimation CNN $f_{\boldsymbol{\theta}_d}(\cdot)$ contains 12 convolutional layers structured as:

$$C_4S_2D_1\text{-}E\text{-}I\text{-}C_8S_2D_1\text{-}E\text{-}I\text{-}C_{16}S_1D_1\text{-}E\text{-}I\text{-}C_{64}S_1D_1\text{-}E\text{-}I\text{-}C_{64}S_1D_2\text{-}E\text{-}I\text{-}R\text{-}C_{64}S_1D_4\text{-}E\text{-}I\text{-}R\text{-}C_{64}S_1D_8\text{-}E\text{-}I\text{-}R\text{-}C_{64}S_1D_{16}\text{-}E\text{-}I\text{-}R\text{-}C_{64}S_1D_{32}\text{-}E\text{-}I\text{-}R\text{-}C_{64}S_1D_1\text{-}E\text{-}I\text{-}R\text{-}C_{32}S_1D_1\text{-}E\text{-}I\text{-}C_1S_1D_1. \quad (2)$$

When using our light field aperture rendering function, the output of the monocular depth estimation network is passed through a scaled $tanh(\cdot)$ activation function to restrict the disparities to $[-10, 10]$ pixels between adjacent views. When using our compositional aperture rendering function, the number of filters in the last convolutional layer is modified to be the number of discrete depth planes n.

For all experiments using the light field aperture rendering function, the depth expansion CNN $g_{\boldsymbol{\theta}_e}(\cdot)$ contains 3 convolutional layers structured as:

$$C_mS_1D_1\text{-}E\text{-}I\text{-}C_mS_1D_1\text{-}E\text{-}I\text{-}C_nS_1D_1 \quad (3)$$

where m is the total number of views in the light field.

2. Additional Results

Below, we display additional qualitative results for both our light field and DSLR experiments.

References

[1] D. A. Clevert, T. Unterthiner, and S. Hochreiter. Fast and accurate deep network learning by exponential linear units (ELUs). *ICLR*, 2016.

[2] D. Ulyanov, A. Vedaldi, and V. Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016.

*Work done while interning at Google Research.

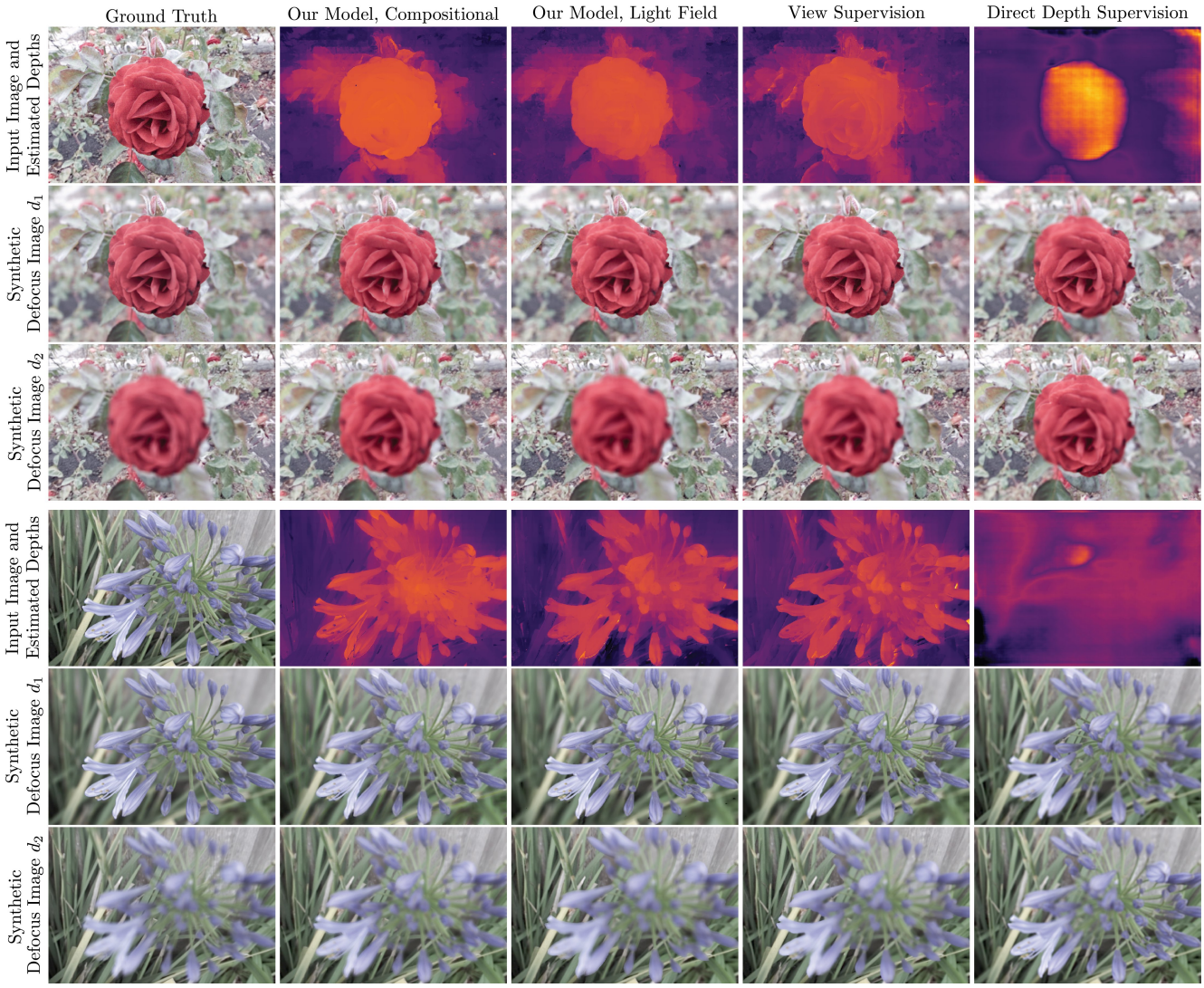

Figure 1. Additional qualitative comparison of monocular depth estimation and synthetic defocus results on images from the test set of our light field experiments. Our aperture supervision models are able to estimate high-quality detailed depths and render convincing shallow-depth-of-field images. The depths estimated by a network trained by view synthesis supervision are reasonable, but typically have artifacts around occlusion edges, causing false edges and artifacts in their rendered shallow depth-of-field images. We recommend that readers view these figures digitally and zoom in to see fine details and differences between the various methods.

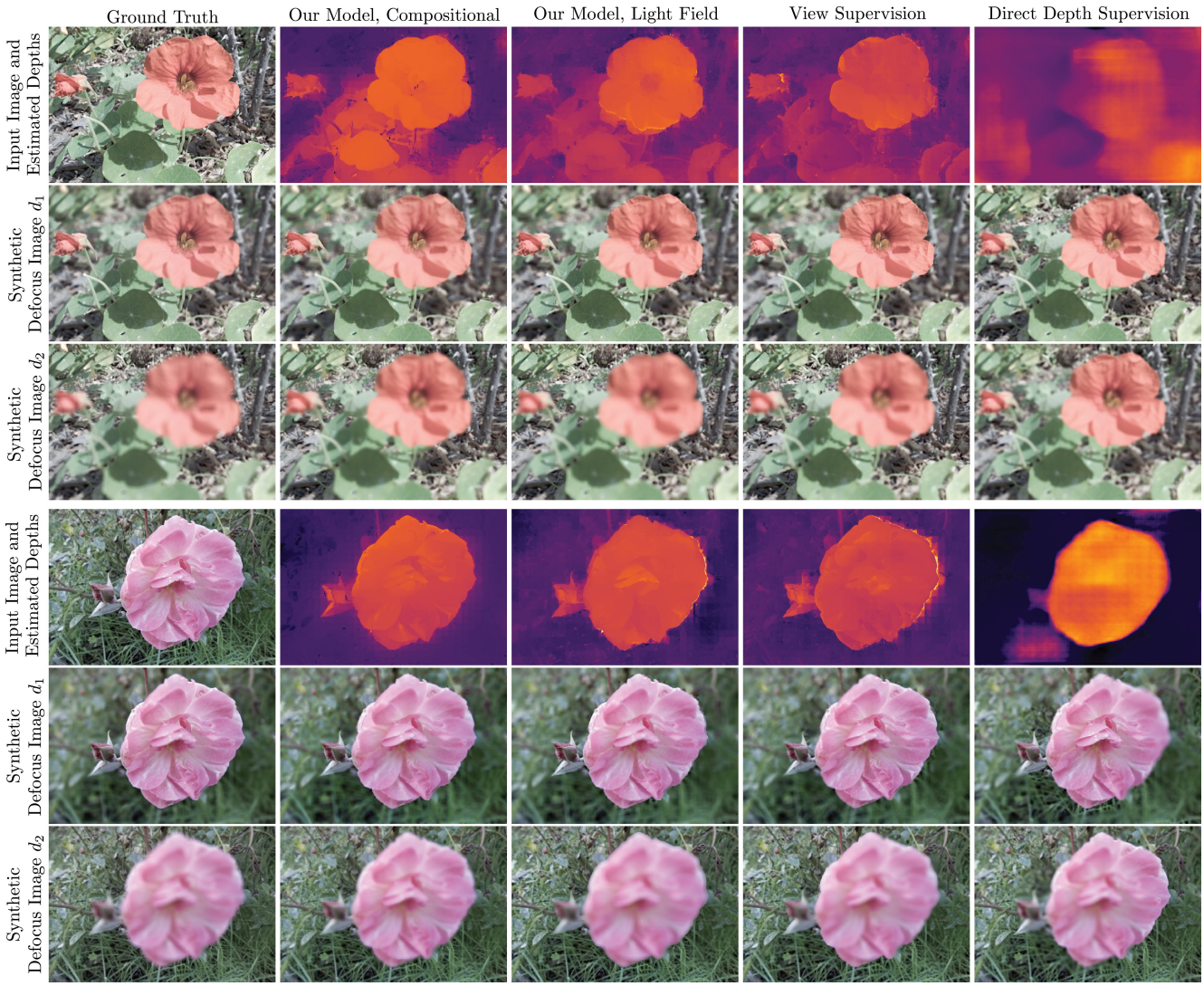

Figure 2. Additional qualitative comparison of monocular depth estimation and synthetic defocus results on images from the test set of our light field experiments. Our aperture supervision models are able to estimate high-quality detailed depths and render convincing shallow-depth-of-field images. The depths estimated by a network trained by view synthesis supervision are reasonable, but typically have artifacts around occlusion edges, causing false edges and artifacts in their rendered shallow depth-of-field images. We recommend that readers view these figures digitally and zoom in to see fine details and differences between the various methods.

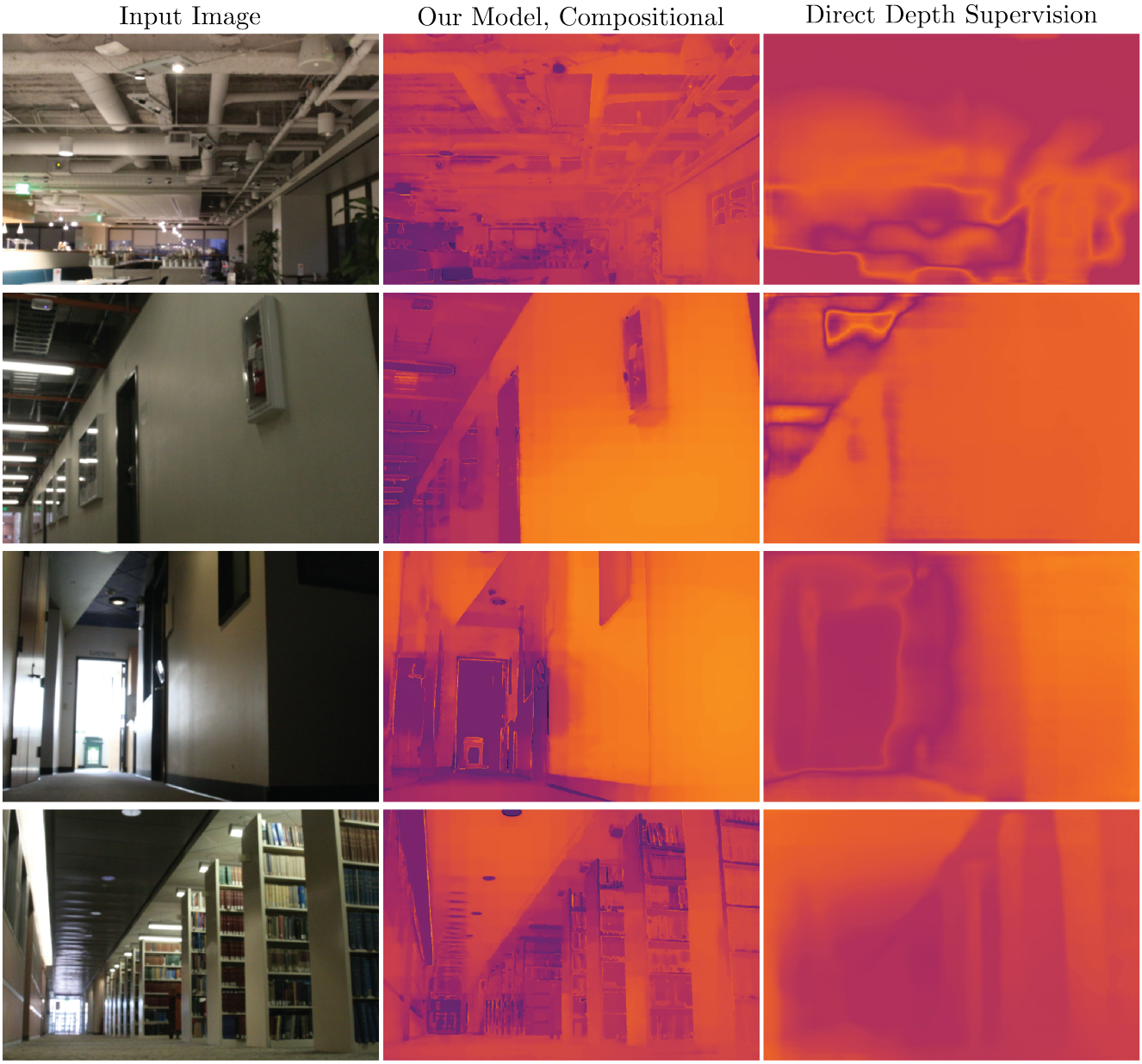

Figure 3. Additional qualitative comparison of monocular depth estimation results on images from the test set of our DSLR dataset experiments. Our aperture supervision model is able to estimate more detailed depth maps than the direct depth supervision baseline.