A. Appendix

A.1. Synthetic Focal Stack Dataset for Robust Training

As mentioned in Sec. 5.2, we generate a denser focal stack and sample five images in each iteration to train a more robust DAIF-Net model. To be specific, we render the dense focal stack datasets from NYUv2 dataset [17]. The dense focal stacks contain 100 images whose focus distance is uniformly distributed from $1m$ to $9m$. In the training stage, the dense focal stack is separated into 5 bins and one image is randomly sampled from each bin to form the sparse input focal stack. Therefore, our model benefits from the robustness of various focus distances. The same rendering strategy is applied to the evaluation set for best-model selection after each training epoch.

A.2. Focus Distance Generalization

Currently, all of the DFD works tend to train their models on datasets with fixed focus distances. In order to generalize our model to varying focus distances, we also propose a robust training strategy. When preparing the training data, we generate a denser focal stack, and at each iteration, we sample five images from the dense focal stack as the input. Since the focal stack is no longer a fixed input, the model can learn how different focus distances can affect the defocus better. To illustrate this point, we perform the robust training on the NYUv2 Dataset. The results, as shown in Figure 6, indicate that the robustly trained model, although having lower performance compared with the original models, can generalize better to focal stacks with different focus distances and with different numbers of images.

A.3. Ablation Study for Scales of the Losses

Our ablation study focuses on the relative scale reconstruction loss and the blurriness loss for coarse AIFs. The results are shown in Table 4. We can see from the table that the reconstruction loss contributes the most to the framework, while appropriate coarse AIF blurriness loss can further boost the model performance.

A.4. Qualitative Results on DefocusNet Dataset

We provide some qualitative results of our method compared with the state-of-the-art supervised works [15, 27] on our rendered DefocusNet dataset [15] in Figure 7. Note our method is the only method that can predict AIF images together with the depth map. Also, we observe that our predicted depth map is on par with the supervised method at a closer distance.

cAIF $\mathcal{L}_{blur}$	$\mathcal{L}_{recon}$			
	1	10	100	1000
0.1	-	0.904	0.934	-
1	-	-	-	0.917
10	-	-	**0.950**	-

Table 4. The ablation study of loss scales. The reported numbers are the δ accuracies. - means the set of parameters fails.

A.5. Qualitative Results on NYUv2 Dataset

A.5.1 Predicted Depth Map and AIF images

Our DAIF-Net predict depth maps and the AIF images from the sparse focal stack. Figure 8 shows some results of our framework. The results shows that our model performs well in the scenes with rich textures. Note we also present the coarse AIF images produced by picking the sharpest point in the focal stacks. The sharp coarse AIF images indicate a good quality of predicted depth maps.

A.5.2 Defocus Map and Focal Stack

We also provide the visualization of the defocus maps produced by the thin-lens module, and reconstruct the focal stack generarted by the PSF convolution layer. Figure 9 shows some examples of the reconstructed focal stack along with their calculated defocus map. From the figure we can see that our optical model can reconstruct the focal stacks in a physical-realistic way, which is critical to the accurate prediction of the depth maps and AIF images.

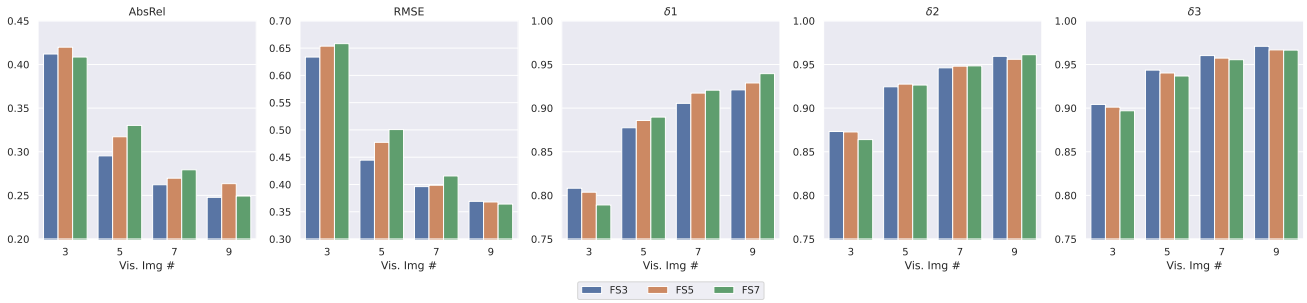

Figure 6. Robustly trained model generalizing to focal stack with different size and focus distances. Different colors indicate the different sizes of the training focal stack. Vis. Img # indicates the size of the testing focal stack.

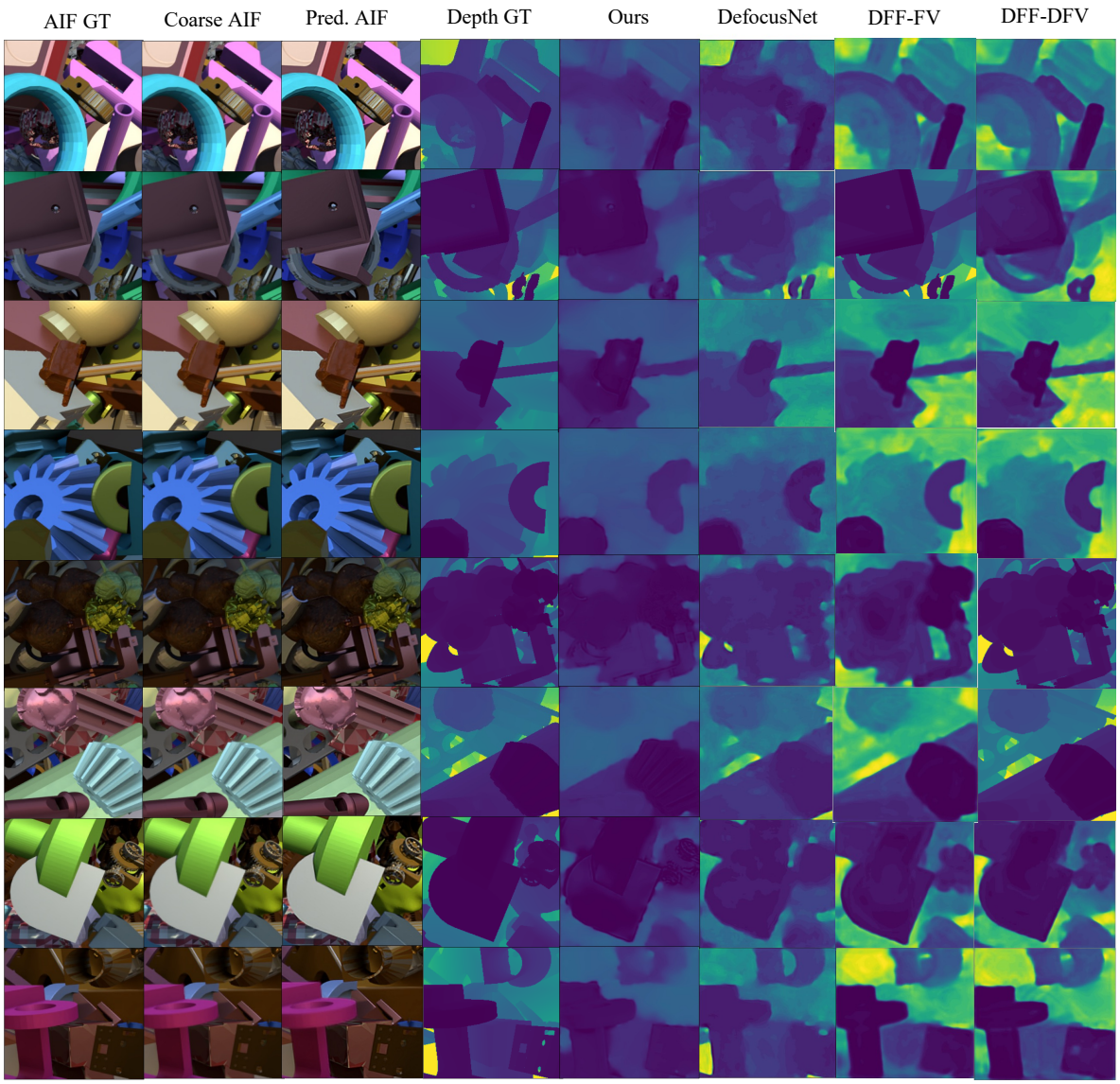

Figure 7. Some examples of the framework outputs comparing with the state-of-the-art supervised works. The outputs are produced from the input focal stacks with 5 images. For the depth map, lighter colors indicate farther distances.

Figure 8. Some examples of the framework outputs. The outputs are produced from the input focal stacks with 5 images. For the depth map, lighter colors indicate farther distances.

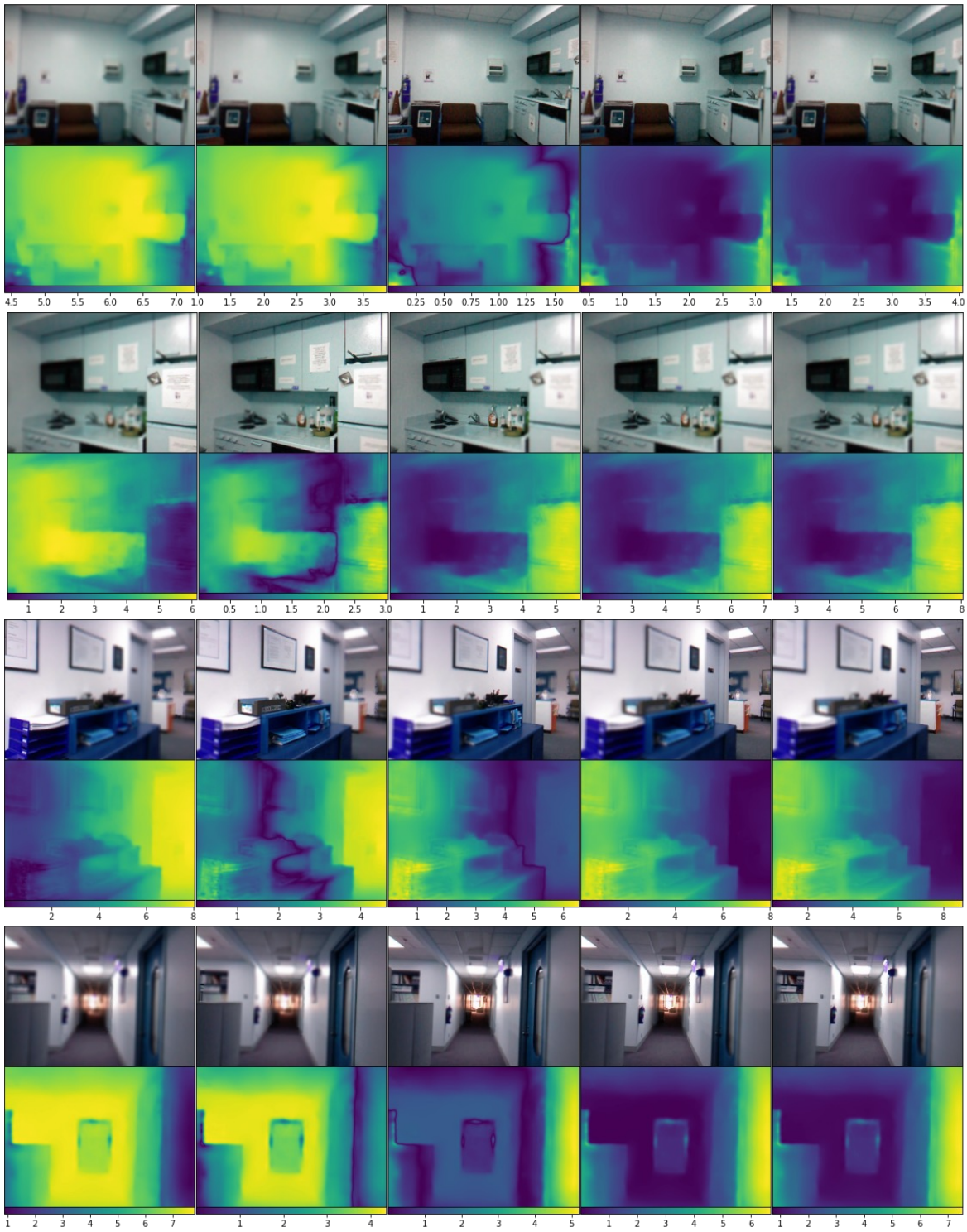

Figure 9. Some examples of the focal stack and their defocus maps. The focus distances are $1m$, $1.5m$, $2.5m$, $4m$ and $6m$ from left to right. Darker colors indicate smaller defocus values.