Towards Photorealistic Reconstruction of Highly Multiplexed Lensless Images —Supplementary—

Salman S. Khan[1], Adarsh V. R.[1], Vivek Boominathan[2], Jasper Tan[2], Ashok Veeraraghavan[2], and Kaushik Mitra[1]

[1] IIT Madras, India
[2] Rice University, USA

Abstract

In this supplementary material, we provide qualitative as well as quantitative comparisons as mentioned in section 5 of the main paper along with additional details of the proposed architecture and display capture setup.

1. Display Capture Setup

To capture a display-captured image, the image is resized to 1280x1280 using bicubic interpolation and displayed on the center 1280x1280 pixels of a 25-inch monitor with 2560x1440 pixels placed approximately one foot from the FlatCam device. With this setup, all 1280x1280 pixels are within the sensors field of view. This setup is fixed for all image captures such that the alignment of the monitor pixels to the camera pixels is uniform throughout both training and test. The FlatCams white balance setting is fixed to be the white balance setting obtained in the FlatCams (i.e. PointGrey Flea3) automatic white balance mode when an all-white image is displayed on the monitor. The exposure time is set to PointGreys automatic mode, and the cameras gain is set to 0dB. Figure 1 shows the setup.

2. More detail on trainable inversion stage

The dimension of W_1 in our experiments was 256×500 while that of W_2 was 620×256 as our network was trained to reconstruct scenes of spatial resolution 256×256 from measurements of spatial resolution 500×620. The perceptual enhancement stage has the same input and output spatial resolution.

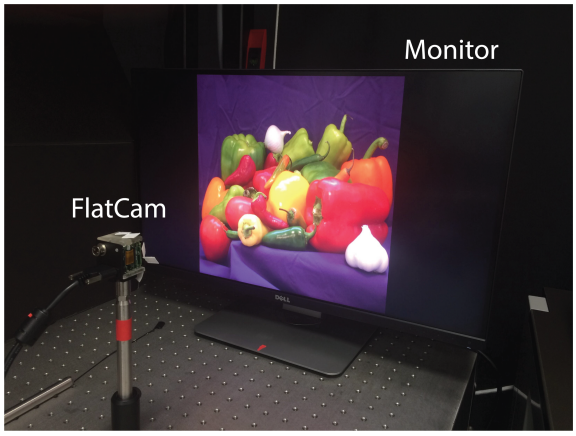
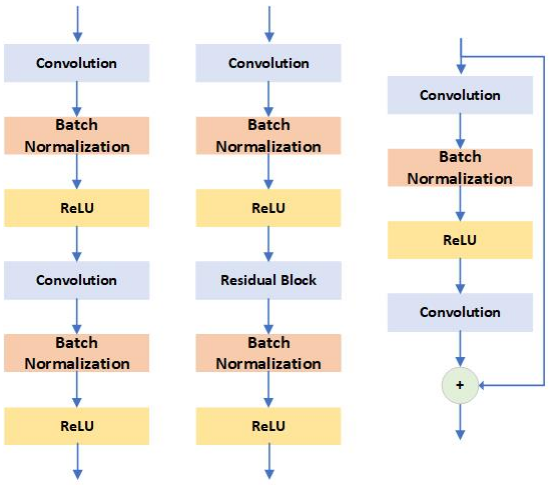


Figure 1. The display capture setup.

3. Comparison of different architectures for perceptual enhancement stage

In our proposed model, we use U-Net[3] for the perceptual enhancement stage. In this section, we compare the performance of U-Net[3] with 3 different architectures and justify its efficacy over them. For fairness in comparison, we keep the same combination loss as defined for the proposed model. We use the transpose initialization scheme described in section 3.1.1 of the original paper. We report PSNR, SSIM and Ma Score[2] for the 100 test images from the display captured dataset. Perceptual score in table 1 refers to the Ma score. Qualitative comparison is provided in figure 3 and 4. The architectures we compared for the perceptual enhancement stage are:

U-Net-Residual: This is a dense version of U-Net that we experimented. We replace one of the convolutional blocks in each stage of the encoder and decoder of U-Net with residual blocks. Each of the residual blocks has 2 convolutional layers. We call this U-Net-Residual. Figure



(a) U-Net (b) U-Net-Residual (c) Residual Block

Figure 2. Basic blocks for U-Net used in the proposed architecture and U-Net-Residual.

2 shows the basic block of the U-Net we used for our perceptual enhancement stage along with the basic block of the U-Net-Residual. It can be seen that U-Net-Residual is much denser compared to U-Net used in the proposed model.

DnCNN[5]: We used a 19-layered model of the popular denoiser network DnCNN for comparison.

RCAN[6]: This is the state of the art for super-resolution. We use the model with 5 residual groups and 10 residual blocks.

Architecture	PSNR (in dB)	SSIM	Perceptual score	Time taken (in sec)
RCAN[6]	17.55	0.56	5.12	0.02
DnCNN[5]	17.07	0.55	5.27	0.005
U-NET-Residual	19.55	0.64	6.15	0.01
U-NET	**19.62**	**0.64**	**6.48**	**0.006**
Ground Truth	-	1	8.04	-

Table 1. Quantitative comparison for different perceptual enhancement stage architectures. RCAN[6] and DnCNN[5] show much lower PSNR, SSIM and perceptual scores than U-Net and its variant U-Net-Residual. Although, U-Net-Residual is close to U-Net used in the proposed architecture in terms of PSNR, SSIM and perceptual score, it takes almost twice the time for a single forward pass.

4. Comparison with compressive image recovery methods

We also compare our proposed model with two state of the art deep learning based compressive image recovery algorithms [1, 4]. Table 2 presents the comparative study of these methods.

Method	PSNR (in dB)	SSIM	Perceptual score
ISTA-Net[4]	14.57	0.54	2.57
Deep Pixel Prior[1]	13.46	0.38	2.49
Proposed	**19.62**	**0.64**	**6.48**

Table 2. Comparison with compressive image recovery algorithms. Average PSNR, SSIM and perceptual scores comparison for display captured measurements.

5. Video reconstruction

In this section we present our result on the temporal stability of the proposed reconstruction pipeline for FlatCam videos. Though we do not explicitly enforce temporal stability, we observe stability over most of the FOV. However, some temporal instability is observed in the very dark regions of the FOV. Figure 5 shows some of the measurement frames along with their reconstructions. Although both traditional and proposed algorithm provide reasonable temporal stability, proposed algorithm clearly extracts finer details. Full video is provided along with the supplementary file.

References

[1] Akshat Dave, Anil Kumar Vadathya, Ramana Subramanyam, Rahul Baburajan, and Kaushik Mitra. Solving inverse computational imaging problems using deep pixel-level prior. *IEEE Transactions on Computational Imaging*, 5(1):37–51, 2018. 2

[2] Chao Ma, Chih-Yuan Yang, Xiaokang Yang, and Ming-Hsuan Yang. Learning a no-reference quality metric for single-image super-resolution. *Computer Vision and Image Understanding*, 158:1–16, 2017. 1

[3] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 1

[4] Jian Zhang and Bernard Ghanem. Ista-net: Interpretable optimization-inspired deep network for image compressive sensing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1828–1837, 2018. 2

[5] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of

(a) Ground Truth (b) RCAN[6] (c) DnCNN[5] (e) U-Net-Residual (f) U-Net

Figure 3. Comparison for display captured measurements. RCAN[6] and DnCNN[5] are unable to restore finer details such as the text in top row. U-Net-Residual is outperformed by U-Net used in the proposed architecture in restoring these fine details.

deep cnn for image denoising. *IEEE Transactions on Image Processing*, 26(7):3142–3155, 2017. 2, 3, 4

[6] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 286–301, 2018. 2, 3, 4

(a) RCAN[6] (b) DnCNN[5] (c) U-Net-Residual (d) U-Net

Figure 4. Comparison for direct captured measurements. Both RCAN[6] and DnCNN[5] result in reconstructions with hazy appearance and color artifacts. This, however, is not seen in both U-Net and U-Net-Residual. It should be noted that U-Net-Residual does not lead to any improvement over the U-Net used in the proposed architecture despite having a very dense structure.

Figure 5. Reconstruction of video frames. Top row shows the captured measurement, while the second and third row show the Tikhonov regularized reconstruction and proposed reconstruction respectively.