Towards Photorealistic Reconstruction of Highly Multiplexed Lensless Images

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

Recent advancements in fields like Internet of Things (IoT), augmented reality, etc. have led to an unprecedented demand for miniature cameras with low cost that can be integrated anywhere and can be used for distributed monitoring. Mask-based lensless imaging systems make such inexpensive and compact models realizable. However, reduction in the size and cost of these imagers comes at the expense of their image quality due to the high degree of multiplexing inherent in their design. In this paper, we present a method to obtain image reconstructions from mask-based lensless measurements that are more photorealistic than those currently available in the literature. We particularly focus on FlatCam [2], a lensless imager consisting of a coded mask placed over a bare CMOS sensor. Existing techniques for reconstructing FlatCam measurements suffer from several drawbacks including lower resolution and dynamic range than lens-based cameras. Our approach overcomes these drawbacks using a fully trainable non-iterative deep learning based model. Our approach is based on two stages: an inversion stage that maps the measurement into the space of intermediate reconstruction and a perceptual enhancement stage that improves this intermediate reconstruction based on perceptual and signal distortion metrics. Our proposed method is fast and produces photo-realistic reconstruction as demonstrated on many real and challenging scenes.

1. Introduction

Cameras have become ubiquitous in the present world. Devices ranging from consumer products to high-end scientific tools use cameras in one form or another. With the proliferation of applications like augmented reality (AR), surveillance, Internet of Things, etc., the purpose of cameras has changed from merely taking photographs to also being sensors for inferential inputs. Some of these novel imaging applications impose stringent constraints in the size of the cameras. Consider, for example, the integration of cameras into wearables like AR glasses or smartwatches. This has resulted in a growing need for the miniaturization of cameras. However, reduction in size and cost is possible only to a certain extent for lens-based cameras. Lensless cameras come to the rescue in such scenarios.

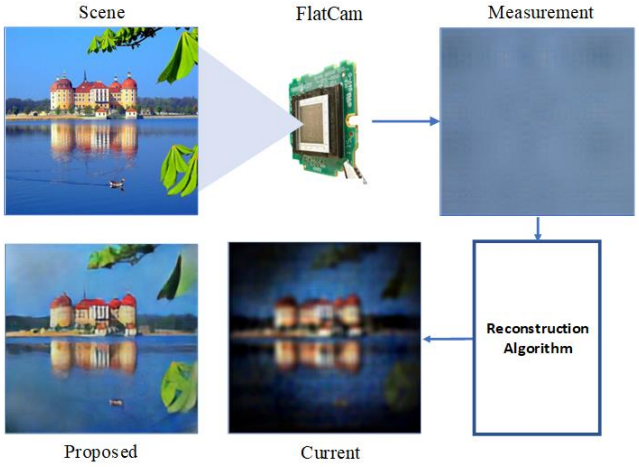


Figure 1. The FlatCam framework: The FlatCam is a thin lensless camera that captures a globally multiplexed measurement of the scene and requires a reconstruction step to recover the true scene. In this work, we propose a novel reconstruction algorithm that gives higher resolution images with improved dynamic range and robustness to noise than current methods.

Recent advancements in sensor technologies and computational imaging techniques have resulted in the emergence of lensless imaging systems. These imaging systems differ from the conventional imaging system in the sense that they encode the incoming light to the sensor (instead of directly focusing it) [2, 1]. A reconstruction algorithm is then required to decode the scene from the measurements.

Lensless imaging systems provide numerous benefits over lens-based cameras. First, lensless imaging systems eliminate the need for a lens, which is the major contribu-

tor towards the size and weight of the camera. In addition, a lensless design permits a broader class of sensor geometries, allowing sensors to have more unconventional shapes (e.g. spherical or cylindrical) or to be physically flexible [37]. Moreover, lensless cameras can be produced with traditional semiconductor fabrication technology and therefore exploit all its scaling advantages, yielding low-cost, high-performance cameras [4]. Earlier instances of using lensless coded aperture imaging systems for X-ray and gamma ray [11, 13, 5, 12, 6] are proofs that lensless imagers have better wavelength scaling as well.

However, the absence of a focusing element and the requirement of a reconstruction algorithm in lensless cameras result in three major challenges. First, lensless design results in an ill-conditioned system, yielding imperfect reconstructions. Second, poor design of the reconstruction algorithm may greatly amplify noise in the images. Third, the reconstruction algorithm's runtime adds a delay to the imaging pipeline which needs to be minimal for applications like virtual or augmented reality. Therefore, lensless cameras need efficient algorithms to overcome these challenges.

In this paper, we focus on developing a reconstruction algorithm for the FlatCam lensless imaging system, which consists of a coded mask placed above the bare imaging sensor [2]. In this design, the sensor records the scene multiplexed by the mask pattern, requiring a demultiplexing algorithm for the recovery of the underlying scene. Existing methods use traditional approaches such as Tikhonov or total variation regularized least squares to reconstruct the scene from FlatCam measurements [2]. However, these methods suffer from several drawbacks including high noise sensitivity, low resolution, low dynamic range, and poor perceptual quality (see figure 1). Apart from these, they also require careful calibration [2], and any error in the process can result in severe degradation in the reconstruction performance.

One way to improve the reconstruction performance would be to exploit the natural image statistics within the data using data-driven techniques like convolutional neural networks [23]. Keeping this in mind, we build our approach on the recent success of deep learning [22] and Generative Adversarial Networks [14]. A simple approach to incorporate a deep network for reconstructing the scene from a FlatCam measurement would be to pre-process the measurement using a hand-crafted prior based reconstruction algorithm (such as Tikhonov regularized least squares) and then refine this reconstruction using the local filtering operations of convolutional neural networks. However, such an approach is dependent on the chosen hand-crafted prior, which may not be suitable for the particular imaging system. Therefore, we go a step further and present a fully trainable deep architecture that provides fast and high quality reconstructions for FlatCam measurements. To the best of our knowledge, our work is the first to use deep learning for mask based lensless image reconstruction and also to incorporate the lensless camera model into the deep architecture. Our major contributions include:

- We propose a fast and novel non-iterative end-to-end deep learning based framework for FlatCam image reconstruction.
- We improve the quality of FlatCam reconstructions by utilizing a weighted loss function based on perceptual and distortion metrics.
- We propose an initialization scheme for our network that removes the need for camera calibration.
- We show the effectiveness of our algorithm by evaluating it on challenging real scenarios. Especially, we show that, compared to existing techniques, proposed technique can recover higher dynamic range scenes.

1.1. Related work

1.1.1 Data driven image reconstruction

Recently, deep learning has shown remarkable performance in many computer vision and image processing tasks. This has largely been driven by the ability of these learning based methods to exploit the structure in the data. However, the application of deep learning in computational imaging is still in a nascent stage. Recently, Boominathan *et al*. [3] proposed a deep network based on conditional GANs to solve the Fourier Ptychography problem while Kulkarni *et al*. [20] and Mousavi *et al*. [28] developed non-iterative deep architectures for reconstructing true signals from compressively sensed measurements. Another class of data-driven methods used for imaging problems involves designing data-dependent priors [8] or data-driven proximal mapping stages [31]. These methods, although more general in comparison to the previous ones, are iterative in nature and are therefore very slow during inference. In this paper, we develop a fast and non-iterative deep architecture to reconstruct scenes from FlatCam measurements with high perceptual quality.

1.1.2 Lensless imaging

Lensless imaging involves capturing an image of a scene without physically focusing the incoming light with a lens. It has been widely used in the past for X-ray and gamma ray imaging for astronomy [11, 6], but its use for visible spectrum has only recently been studied. In a lensless imaging system, the scene is captured either directly on the sensor [18] or after being modulated by a mask element. Types of masks that have been used include phase gratings [36], diffusers [1], amplitude masks [34, 2], compressive samplers [16] and spatial light modulators [7, 10]. Replacing lens

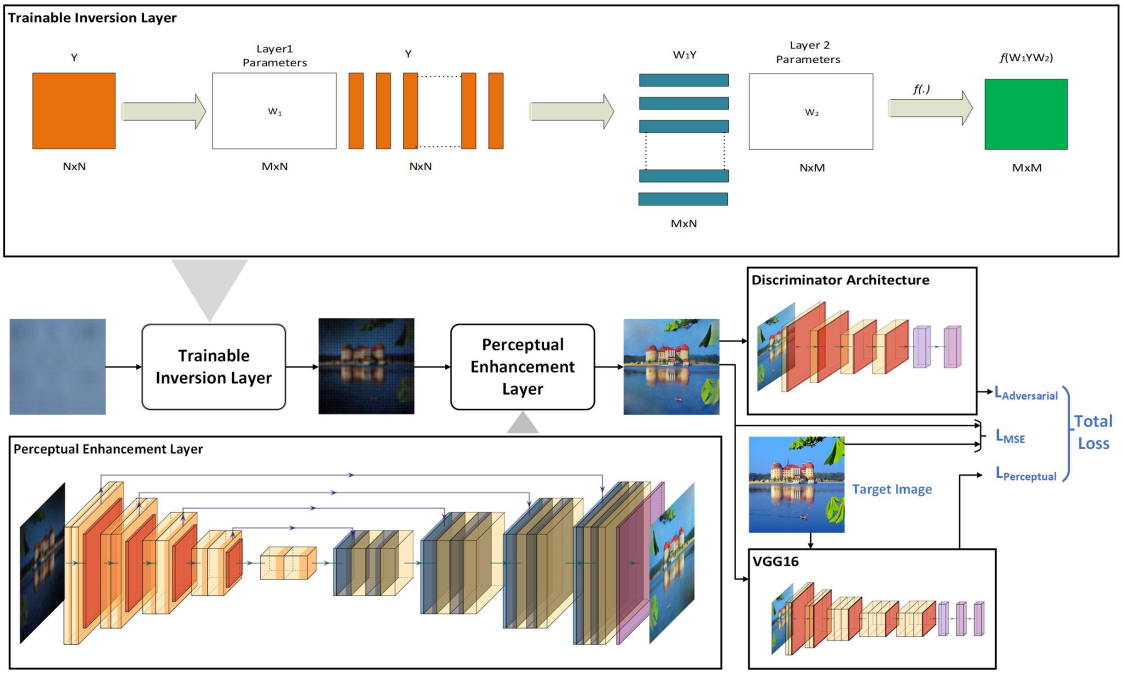


Figure 2. Overall architecture of the proposed system. Our network consists of a generator and a discriminator. In the generator, the measurement is first mapped into an intermediate image space using a trainable inversion layer. A U-Net then enhances the perceptual quality of the intermediate reconstruction. We use a weighted combination of three losses in training our system: a perceptual loss[17] using a VGG16 network[35], mean-square error (MSE), and adversarial loss using a discriminator neural network [14].

with the above masks result in muddled sensor capture that lacks any resemblance to the scene imaged. A recognizable image is then recovered using a computational reconstruction algorithm. In this paper, we develop a deep learning based reconstruction for a particular lensless camera called FlatCam [2], which has a low complexity image formation model due to the separable property of the amplitude mask design.

2. FlatCam

FlatCam is a lensless imaging system developed by Asif *et al*. [2] that consists of an amplitude mask placed above the CMOS sensor. As the mask is made up of multiple apertures/pinholes, the resultant measurement recorded at the sensor is a superposition of the images formed due to each pinhole. The mask pattern is an outer product of two different maximum length sequences and as a result is a rank one matrix. This causes the FlatCam to have a separable point spread function (PSF) and the following forward model,

$$Y = \Phi_L X \Phi_R^T + N. \quad (1)$$

Here, Φ_L and Φ_R are the corresponding matrix representation of the separable PSF kernel, X is the scene irradiance, Y is the recorded measurement, and N models additive noise. The separability of the PSF provides a means to efficiently calibrate as well as invert the sensor measurements [2, 10]. The current reconstruction method in [2] is to find the Tikhonov regularized scene X that minimizes the error in (1).However, in reality, the true physical process of lensless image capture contains non-idealities (such as diffraction) that (1) does not completely account for. Furthermore, the operation tends to have poor conditioning, leading to high sensitivity to noise. This makes recovering the true scene radiance from these measurements very challenging.

3. End-to-end network for FlatCam reconstruction

To address the difficulties in accurate FlatCam reconstruction, we take a data-driven approach at recovering the true scene from the highly multiplexed measurements. We exploit the physics of the forward model along with the efficiency of a deep neural network in learning a photorealistic mapping from the measurement space to the natural image space. Following the success of Generative Adversarial Nets [14], our proposed network has two main components: a generator network that learns to output a visually meaningful reconstruction from the measurement and a discriminator network that tries to distinguish this reconstruction from real images. Both the networks are finally trained in an adversarial setup. Figure 2 shows the generalized block diagram for our method. In the following subsections, we describe each of these steps in more details.

3.1. Generator architecture

Our generator network has two basic stages: the *trainable inversion stage* maps the FlatCam measurements into a space of intermediate reconstructions, and the *perceptual enhancement stage* refines this mapping into a semantically meaningful image. It should be noted here that both stages of our approach are trained in an end-to-end way.

3.1.1 Trainable inversion stage

In the first stage, we use two layers of trainable left and right matrix multiplications on the two-dimensional sensor measurments followed by a non-linearity:

$$X_{\text{interm}} = f(W_1 Y W_2), \tag{2}$$

where X_{interm} is the output of this stage, f is a pointwise nonlinearity, Y is the input measurement, and W_1 and W_2 are the corresponding weight matrices for this stage. Figure 2 gives a diagrammatic overview of this stage. For the nonlinearity, we use the leaky ReLU[27]. The dimension of the weight matrices depends on the dimension of the measurement and the scene dimension we want to recover. It is important to initialize the weight matrices of this stage properly, so that the network does not get stuck in local minima. This can be done in two ways.

Transpose initialization: For this approach, we initialize our weight matrices (W_1 and W_2) with the adjoint of the calibration matrices, akin to back-projection. These calibration matrices are approximations of Φ_L and Φ_R in (1) physically obtained by the method described in [2]. This mode of initialization leads to faster convergence while training.

Random initialization: Calibration of FlatCam require careful alignment with display monitor [2], which can be a time consuming and inconvenient process especially for large volumes of FlatCams. Even a small error in calibration can lead to severe degradation in the performance of the reconstruction algorithm. To overcome the problems involved in calibration, we also propose a calibration-free approach by initializing the weight matrices with carefully designed pseudo-random matrices.

Initializing with any pseudo-random matrices of appropriate size does not yield successful reconstruction. To carefully design the random initialization, we make the following two observations regarding the FlatCam forward model: the calibration matrices have a 'toeplitz-like' structure and the slope of constant entries in the 'toeplitz-like' structure can be approximately determined using the FlatCam geometry, in particular the distance between the mask and the sensor and the pixel pitch. As the FlatCam's geometry is known apriori, we can construct the pseudo-random 'toeplitz-like' matrices with appropriate slope, and size, thereby making our approach calibration free. The weight matrices (W_1 and W_2) are initialized with the adjoint of such constructed random matrices. We observed that the training time increased slightly for this initialization in comparison to transpose initialization.

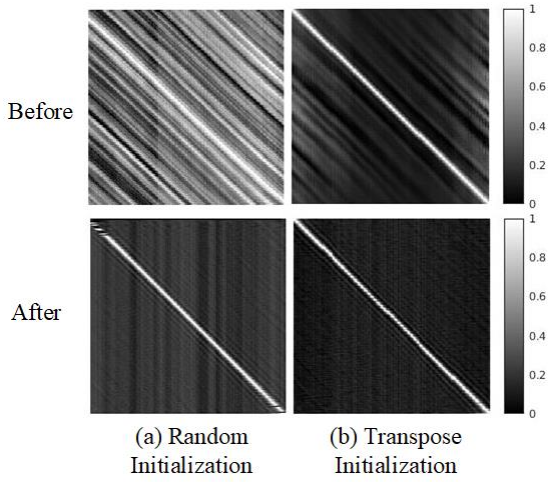


Figure 3. Product of the left weight matrix from the trainable inversion stage for with the calibration matrix ($W_1 \times \Phi_L$) before and after training. The top row shows the initial product at the beginning of training while the bottom row shows it after training the network. a) Random initialization. b) Transpose initialization.

In our experiments, we found that the products of the learned matrices W_1 and W_2 with the forward model calibration matrices Φ_L and Φ_R closely resemble an identity matrices, implying that this stage has tried to invert the FlatCam forward model. This is shown in figure 3. The left and right matrix multiplication ensures that the global multiplexing of the forward model is captured by this stage which may not have been possible using local operations like convolution. Moreover, this stage can also be interpreted as the forward model dependent stage and can be replaced suitably when the model changes.

3.1.2 Perceptual enhancement stage

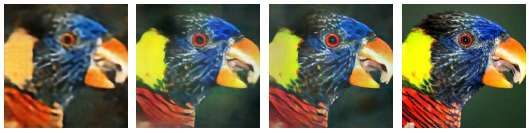
Once we obtain the output of the trainable inversion stage, which is of same dimension as that of the natural image, we use a fully convolutional network to map it into the natural image space. Owing to its large scale success in image-to-image translation problems and its multi-resolution structure, we choose a U-Net [32] to map the intermediate reconstruction to the final perceptually enhanced image. We keep the kernel size fixed at 3×3 while the number of filters is gradually increased from 128 to 1024 in the encoder and then reduced back to 128 in the decoder. In the end, we map the signal back to 3 RGB channels.

3.2. Discriminator architecture

The trainable inversion and the perceptual enhancement stage form the generator of our architecture. We then use a discriminator framework to classify our generator's output

as real or fake. We find that using a a discriminator network improves the perceptual quality of our reconstruction. We use 4 layers of 2-strided convolution followed by batch normalization and the swish activation function [30] in our discriminator.

3.3. Loss function



(a) M (b) M+P (c) M+P+A (d) GT

Figure 4. Effect of the different weighted loss functions. a) Using only MSE leads to blurry reconstruction. b) Using the weighted combination of MSE and perceptual leads to a sharper output with color artifacts. c) Addition of adversarial loss further improves the perceptual quality. d) Corresponding ground truth for reference.

An appropriate loss function is required to optimize our system to provide the desired output. Pixelwise losses like mean absolute error (MAE) or mean squared error (MSE) have been successfully used to capture signal distortion. However, they fail to capture the perceptual quality of images. As our objective is to obtain high quality photorealistic reconstructions from lensless measurements, perceptual quality matters. Thus, we use a weighted combination of signal distortion and perceptual losses. The losses used for our model are given below:

Mean squared error: We use MSE to measure the distortion between the ground truth and the estimated output. Given the ground truth image I_{true} and the estimated image I_{est}, this is given as:

$$\mathcal{L}_{\text{MSE}} = \|I_{\text{true}} - I_{\text{est}}\|_2^2 . \tag{3}$$

Perceptual loss: To measure the semantic difference between the estimated output and the ground truth, we use the perceptual loss introduced in [17]. We use a pre-trained VGG-16 [35] model for our perceptual loss. We extract feature maps between the second convolution (after activation) and second max pool layers, and between the third convolution (after activation) and the fourth max pool layers. We call these activations ϕ_{22} and ϕ_{43}, respectively. This loss is given as,

$$\mathcal{L}_{\text{percept}} = \|\phi_{22}(I_{\text{true}}) - \phi_{22}(I_{\text{est}})\|_2^2 + \|\phi_{43}(I_{\text{true}}) - \phi_{43}(I_{\text{est}})\|_2^2 . \tag{4}$$

Adversarial loss: Adversarial loss [14, 24] was added to further bring the distribution of the reconstructed output close to those of the real images. Given a discriminator D, this loss is given as,

$$\mathcal{L}_{\text{adv}} = -log(D(I_{\text{est}})). \tag{5}$$

Total generator loss: Our total loss for the generator is a weighted combination of the three losses and is given as,

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{MSE}} + \lambda_2 \mathcal{L}_{\text{percept}} + \lambda_3 \mathcal{L}_{\text{adv}}. \tag{6}$$

where, λ_1, λ_2 and λ_3 are weights assigned to each loss.

Discriminator loss: Given I_{est}, I_{true} and discriminator D, the discriminator was trained using the following loss,

$$\mathcal{L}_{\text{disc}} = -log(D(I_{true})) - log(1 - D(I_{\text{est}})). \tag{7}$$

4. Experiments and results

In this section we describe all our experimental results. It should be noted that all our experiments are performed on real data, which demonstrates the practicality of our approach. For collecting real world data, we use two setups: the display capture setup and the direct capture setup.

Collecting a large scale dataset of lensless measurements along with their aligned ground truth is a challenging task. To overcome this challenge, the first setup we use to capture real images is the display capture setup. In this setup, we place a monitor in front of the FlatCam and capture the images displayed on it. For the ground truth, we randomly selected 10 images from each of the ImageNet [33] classes and created a dataset of 10000 images. Out of this, we kept 9900 images from 990 classes for training and the rest 10 classes or 100 images for testing. We call these measurements display captures. More detail on display captured setup is provided in the supplementary material.

It is important to visually evaluate the performance of our reconstruction network on a direct real world setup. For collecting data for this setup, we place objects in front of FlatCam and directly capture the measurement. For this setup we do not have a corresponding ground truth. We call these measurements direct captures.

4.1. Implementation details

The FlatCam prototype used is Point Grey Flea3 camera with 1.3MP e2v EV76C560 CMOS sensor with a pixel size of 5.3 μm. All the ground truth images were resized to 256×256 as the FlatCam is calibrated to produce 256×256 output images. This ensures that there is no misalignment among the input and ground truth pairs. We directly used the Bayer measurements of 4 channels (R,Gr,Gb,B) as our input to the network and convert them into 3 channel RGB within the network. FlatCam measurements of dimension $500 \times 620 \times 4$ in batches of 4 were used as inputs for training. A smaller batch size was used due to memory constraints. We set λ_1 as 1, λ_2 to be 1.2 and λ_3 to be 0.6. For transpose initialization, we trained our model for 45K iterations while for random initialization, we trained it for 60K iterations. The Adam [19] optimizer was used for all models. We started with a learning rate of 0.0001 and gradually reduced it by half every 5000 iterations. We train the

(a) Ground Truth (b) Tikhonov[2] (c) TVAL3[25] (d) Naive (e) Proposed-R (f) Proposed-T

Figure 5. Reconstruction of display captured measurements using various approach. Inset shows the finer region in each image. a) Ground truth image for reference. Finer details like the text in the first image and leg of the insect in the second image are lost in b) Tikhonov regularized and c) TVAL3 reconstruction. d) Naive network improves the Tikhonov reconstruction but still lacks details. Finer details are better preserved in our end-to-end model for both e) random and f) transpose initializations.

Method	PSNR (in dB)	SSIM	Perceptual score	Time taken (in sec)
Tikhonov[2]	10.95	0.33	2.25	0.03
TVAL3[25]	11.81	0.36	3.38	45.28
Naive	18.90	0.62	5.72	0.016
Proposed-R	19.06	0.62	5.86	0.006
Proposed-T	**19.62**	**0.64**	**6.48**	**0.006**
Ground Truth	-	1	8.04	-

Table 1. PSNR, SSIM and perceptual score comparison for display captured measurements. Proposed method with transpose initialization (Proposed-T) gives the best result. The comparable performance of Proposed-R indicates that our approach can be used for situations where careful calibration isn't possible.

discriminator and the generator alternatively as is done for conventional GANs [14]. We use PyTorch [29] to implement our model.

4.2. Comparison with other approaches

We present a comparison of the performance of our method with that of other techniques. The following are the description of these techniques.

4.2.1 Traditional techniques

In this category, we use the closed form solution of Asif *et al*. [2] for Tikhonov regularized reconstruction. This is given as,

$$\hat{X} = V_L[(\Sigma_L^T U_L^T Y U_R \Sigma_R)./(\sigma_L \sigma_R^T + \lambda \mathbf{1}\mathbf{1}^T)]V_R^T, \quad (8)$$

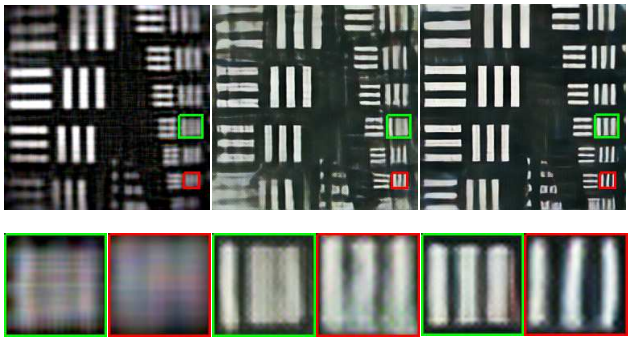
where $\Phi_L = U_L \Sigma_L V_L^T$ and $\Phi_R = U_R \Sigma_R V_R^T$ and λ is the regularization parameter. We set λ to 0.0003 for our reconstruction. We also used TVAL3[25] for reconstruction which imposes sparsity on the gradients of the reconstruction.

4.2.2 Naive deep learning based approach

The Tikhonov solution of (8) is extremely fast to compute as shown in [2]. A naive way to obtain higher quality reconstruction from FlatCam measurements would be to ob-

(a) Tikhonov[2] (b) TVAL3[25] (c) Naive (d) Proposed-R (e) Proposed-T

Figure 6. Reconstruction of direct captured measurements using various approaches. a) Tikhonov regularized reconstruction kills off the detail in the border and darker region. b) TVAL3 reconstructs the border but is unable to restore the sharpness. c) Naive network improves Tikhonov reconstruction but is highly sensitive to noise and the regularization parameter. The proposed end-to-end models for both d) random and e) transpose initialization produce the best reconstructions. These methods are robust to noise and does not contain any regularization parameter.



(a) Tikhonov (b) Naive (c) Proposed

Figure 7. Reconstruction of resolution chart. Inset shows the finer details. a) Tikhonov regularized reconstruction fails to capture high frequency details. b) Naive approach improves upon Tikhonov regularized reconstruction but is restricted by the information lost in Tikhonov reconstruction. c) Proposed method is able to reconstruct the fine details.

tain the Tikhonov regularized reconstruction and then use an image restoration framework to refine the reconstruction. To implement this, we pass the Tikhonov regularized reconstruction through our perceptual enhancement block (described in section 3.1.2). We use the same loss that is defined in (6).

Qualitative comparison: Figure 5 shows the comparison of our approach with the traditional and naive approach on some of the display captured images. In all figures and tables, Proposed-R refers to the model using random initialization and Proposed-T refers to the model using transpose initialization as explained in section 3.1.1. Unless explicitly mentioned, Proposed refers to Proposed-T. The Tikhonov regularized reconstructions are prone to severe vignetting effects which is somewhat reduced in the TVAL3 results. As the naive approach is trained on the Tikhonov reconstructions, it fails to reconstruct the regions lost in Tikhonov regularization. Inset images in figure 5 show the preservation of finer details in our approach. Figure 6 shows the comparison of our approach for direct captured measurements. It should be noted that Tikhonov regularization has a tendency to suppress low signal values and as a result has difficulty restoring the poorly illuminated background for most of the scenes in figure 6. The performance of TVAL3[25] is also similar. As the naive model is based on Tikhonov reconstruction, it heavily relies on the regularization parameter. It also requires calibration and is sensitive

to noise. This is verified by the distinguishable artifacts that appear in the naive reconstructions in figure 6. Apart from these disadvantages, it also lacks in ability to adapt to challenging low dynamic range scenes. This is further verified in section 4.3. As we also show comparable performance using random initialization, our method does not depend on calibration unlike the rest of the approaches. Figure 7 gives a better idea about the superiority of the proposed approach in reconstructing finer details.

Quantitative comparison: Table 1 shows the quantitative evaluation of our approach. We provide the test results on the 100 display captured test images we had previously separated from the 10,000 ImageNet captures. We use PSNR, SSIM and the no-reference image quality metric of Ma *et al*. [26] for signal distortion and perception evaluation. Higher PSNR, SSIM and Ma score indicate superior performance. From the metrics in table 1, it is clear that our approach using transpose initialization (Proposed-T) outperforms all the other reconstruction techniques for FlatCam. The next best approach is the proposed method using random initialization (Proposed-R) and it is worth noting that unlike all other methods, Proposed-R is calibration-free. The naive network also performs remarkably better than the existing FlatCam reconstruction techniques but is clearly outperformed by the proposed model for both cases of initialization. We also compare the time complexities for the various approaches in table 1. The Tikhonov and TVAL3[25] regularized reconstructions are computed on Intel Core i7 CPU with 16 GB RAM while the rest of the approaches are evaluated on Nvidia GTX 1080 GPU.

4.3. Handling bright light sources

For a highly multiplexed lensless imager like FlatCam, every pixel receives light from every point in the scene. Hence, if there is any really bright object (like a highly reflective object or a lamp) in the scene, the light from the object can dominate the pixel intensities and result in severe reconstruction artifacts on the dimmer objects. We show that, using our proposed reconstruction algorithm, the artifacts are minimized resulting in a higher quality reconstruction of the scene.

We show the bright object problem by introducing an LED into the scene. Figure 8 shows two scenes, with an LED introduced. With the LED, we see severe artifacts in the Tikhonov reconstruction. The artifacts appear along the row and column of the LED location due to the separable model of FlatCam. The Naive network fails to compensate for the information lost in Tikhonov reconstruction. However, the scene information for all levels of illumination is still present in the multiplexed measurement. Using our proposed reconstruction algorithm, we observe that the artifacts are reduced and a realistic representation of the scene is reconstructed. As our network is trained end-to-end directly on the measurements, it adapts to extremely challenging dynamic range scenes.

Figure 8. Handling bright objects. Arrows indicate the position of LED. The Tikhonov reconstructions for LED introduced scenes have been scaled in intensity for visualization. Strong separable mask artifact is seen in both Tikhonov regularized and naive approach, while proposed approach gives much cleaner and sharper results.

5. Discussion

For the perceptual enhancement stage, we experimented with several different architectures before settling down on U-Net. In earlier works on image restoration [39, 21], residual blocks[15] had shown to improve the performance by allowing finer details to pass. Thus, we replaced the original U-Net structure with residual blocks in both encoder and decoder while keeping rest of the architecture and losses unchanged. However, we did not observe any significant improvement over the original structure. We also experimented with state-of-the art image restoration techniques like DnCNN[39] and RCAN[40] and compressive image recovery techniques like ISTANet[38] and Deep pixel level prior[9]. Evaluation of these approaches is provided in the supplementary materials.

6. Conclusion

We present an approach to recover photorealistic images from highly multiplexed lensless images using a fully trainable deep network. Although lensless imaging promises a wide array of application due to its size, weight and cost benefits, standard optimization-based methods currently used for reconstruction yield outputs that suffer from low resolution and higher noise sensitivity. Our approach attempts to solve these problems as exhibited by our extensive experiments on real data. Moreover, as opposed to current methods, our reconstruction algorithm doesn't require the error-prone and meticulous calibration of the FlatCam system. In future, it will be interesting to look into the co-design of mask along with the reconstruction algorithm.

Acknowledgements: This work was supported by NSF CAREER: IIS-1652633, DARPA NESD: HR0011-17-C-0026 and NIH Grant: R21EY029459.

References

[1] Nick Antipa, Grace Kuo, Reinhard Heckel, Ben Mildenhall, Emrah Bostan, Ren Ng, and Laura Waller. Diffusercam: lensless single-exposure 3d imaging. *Optica*, 5(1):1–9, 2018.

[2] M Salman Asif, Ali Ayremlou, Aswin Sankaranarayanan, Ashok Veeraraghavan, and Richard G Baraniuk. Flatcam: Thin, lensless cameras using coded aperture and computation. *IEEE Transactions on Computational Imaging*, 3(3):384–397, 2017.

[3] Lokesh Boominathan, Mayug Maniparambil, Honey Gupta, Rahul Baburajan, and Kaushik Mitra. Phase retrieval for fourier ptychography under varying amount of measurements. *arXiv preprint arXiv:1805.03593*, 2018.

[4] Vivek Boominathan, Jesse K Adams, M Salman Asif, Benjamin W Avants, Jacob T Robinson, Richard G Baraniuk, Aswin C Sankaranarayanan, and Ashok Veeraraghavan. Lensless imaging: A computational renaissance. *IEEE Signal Processing Magazine*, 33(5):23–35, 2016.

[5] TM Cannon and EE Fenimore. Coded aperture imaging: many holes make light work. *Optical Engineering*, 19(3):193283, 1980.

[6] E Caroli, JB Stephen, G Di Cocco, L Natalucci, and A Spizzichino. Coded aperture imaging in x-and gamma-ray astronomy. *Space Science Reviews*, 45(3-4):349–403, 1987.

[7] Wanli Chi and Nicholas George. Optical imaging with phase-coded aperture. *Optics express*, 19(5):4294–4300, 2011.

[8] Akshat Dave, Anil Kumar, Kaushik Mitra, et al. Compressive image recovery using recurrent generative model. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 1702–1706. IEEE, 2017.

[9] 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.

[10] Michael J DeWeert and Brian P Farm. Lensless coded-aperture imaging with separable doubly-toeplitz masks. *Optical Engineering*, 54(2):023102, 2015.

[11] RH Dicke. Scatter-hole cameras for x-rays and gamma rays. *The astrophysical journal*, 153:L101, 1968.

[12] PT Durrant, M Dallimore, ID Jupp, and D Ramsden. The application of pinhole and coded aperture imaging in the nuclear environment. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 422(1-3):667–671, 1999.

[13] Edward E Fenimore and Thomas M Cannon. Coded aperture imaging with uniformly redundant arrays. *Applied optics*, 17(3):337–347, 1978.

[14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[16] Gang Huang, Hong Jiang, Kim Matthews, and Paul Wilford. Lensless imaging by compressive sensing. In *2013 IEEE International Conference on Image Processing*, pages 2101–2105. IEEE, 2013.

[17] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016.

[18] Ganghun Kim, Kyle Isaacson, Rachael Palmer, and Rajesh Menon. Lensless photography with only an image sensor. *Applied optics*, 56(23):6450–6456, 2017.

[19] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[20] Kuldeep Kulkarni, Suhas Lohit, Pavan Turaga, Ronan Kerviche, and Amit Ashok. Reconnet: Non-iterative reconstruction of images from compressively sensed measurements. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 449–458, 2016.

[21] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 624–632, 2017.

[22] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436, 2015.

[23] Yann LeCun, Patrick Haffner, Léon Bottou, and Yoshua Bengio. Object recognition with gradient-based learning. In *Shape, contour and grouping in computer vision*, pages 319–345. Springer, 1999.

[24] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017.

[25] Chengbo Li, Wotao Yin, Hong Jiang, and Yin Zhang. An efficient augmented lagrangian method with applications to total variation minimization. *Computational Optimization and Applications*, 56(3):507–530, 2013.

[26] 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.

[27] Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, volume 30, page 3, 2013.

[28] Ali Mousavi, Ankit B Patel, and Richard G Baraniuk. A deep learning approach to structured signal recovery. In *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1336–1343. IEEE, 2015.

[29] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Al-

ban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[30] Prajit Ramachandran, Barret Zoph, and Quoc V Le. Searching for activation functions. *arXiv preprint arXiv:1710.05941*, 2017.

[31] JH Rick Chang, Chun-Liang Li, Barnabas Poczos, BVK Vijaya Kumar, and Aswin C Sankaranarayanan. One network to solve them all–solving linear inverse problems using deep projection models. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5888–5897, 2017.

[32] 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.

[33] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[34] Takeshi Shimano, Yusuke Nakamura, Kazuyuki Tajima, Mayu Sao, and Taku Hoshizawa. Lensless light-field imaging with fresnel zone aperture: quasi-coherent coding. *Applied optics*, 57(11):2841–2850, 2018.

[35] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[36] David G Stork and Patrick R Gill. Lensless ultra-miniature cmos computational imagers and sensors. *Proc. SENSORCOMM*, pages 186–190, 2013.

[37] Eric J Tremblay, Ronald A Stack, Rick L Morrison, and Joseph E Ford. Ultrathin cameras using annular folded optics. *Applied optics*, 46(4):463–471, 2007.

[38] 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.

[39] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE Transactions on Image Processing*, 26(7):3142–3155, 2017.

[40] 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.