Robust Lensless Image Reconstruction via PSF Estimation

Joshua D. Rego
Arizona State University

Karthik Kulkarni
Arizona State University

Suren Jayasuriya
Arizona State University

Abstract

Lensless imaging is a new, emerging modality where image sensors utilize optical elements in front of the sensor to perform multiplexed imaging. There have been several recent papers to reconstruct images from lensless imagers, including methods that utilize deep learning for state-of-the-art performance. However, many of these methods require explicit knowledge of the optical element, such as the point spread function, or learn the reconstruction mapping for a single fixed PSF. In this paper, we explore a neural network architecture that performs joint image reconstruction and PSF estimation to robustly recover images captured with multiple PSFs from different cameras. Using adversarial learning, this approach achieves improved reconstruction results that do not require explicit knowledge of the PSF at test-time and shows an added improvement in the reconstruction model's ability to generalize to variations in the camera's PSF. This allows lensless cameras to be utilized in a wider range of applications that require multiple cameras without the need to explicitly train a separate model for each new camera.

1. Introduction

As imaging has become ubiquitous, there has been increased emphasis on reduced size, weight, power, and cost for cameras. While camera size and weight has reduced drastically over the past decade, it is still limited by the size/weight of optical components, namely the lens. In computational imaging research, lensless camera technology has been recently shown to be a promising solution for low size, weight, and cost [4, 39, 2, 14]. Without a lens, cameras can be reduced to consume a nearly flat form-factor to enable computer vision in small-scale robotics, embedded in wearable and Internet-of-Things applications, and even enable potential uses in biomedical microscopy.

Lensless cameras indirectly capture information from a scene by multiplexing light rays onto the sensor, and then reconstructing the image by solving an inverse problem. This multiplexing is done using various optical elements such as coded apertures, diffusers, or diffraction gratings. In previous research, lensless reconstructions have been shown to be near visual quality of lensed images using both optimization and machine learning methods [7]. Lensless cameras have also shown potential in 3D microscopy[26, 1] and depth estimation [44], thus augmenting imaging beyond traditional photography applications. Finally, lensless images without reconstruction are not visually identifying for human observers, and thus may have applications for privacy and security [42, 9].

However, lensless cameras have not yet been seen as a viable alternative to traditional cameras in many applications due to some specific limitations. Reconstruction of the scene image requires calibration to account for the optical element which performs light multiplexing onto the sensor, typically the point-spread-function (PSF) of the optical system. This calibration is sensor-specific, and most reconstruction algorithms to-date require this information in order to recover the scene. This need for calibration information limits the robustness of these reconstruction algorithms in practice for lensless cameras.

In this paper, we propose lensless camera reconstruction which does not require calibration of the point-spread-function (PSF) for a given optical element. Our neural network architecture performs joint image reconstruction and PSF estimation, and achieves comparable to state-of-the-art reconstruction performance while generalizing to multiple PSFs corresponding to multiple lensless cameras. We utilize a GAN-based deep neural network architecture, and estimate the PSF from a lensless input image to perform reconstruction without the need for calibration using a prior learned from training data. Our experimental results validates our network performance without calibration PSF data at test time along with low latency. This includes testing on a difficult, varying PSF dataset of lensless video from multiple lensless cameras collected from a prototype hardware setup. We hope this work shows the first steps toward calibration-free and robust lensless imaging in the future.

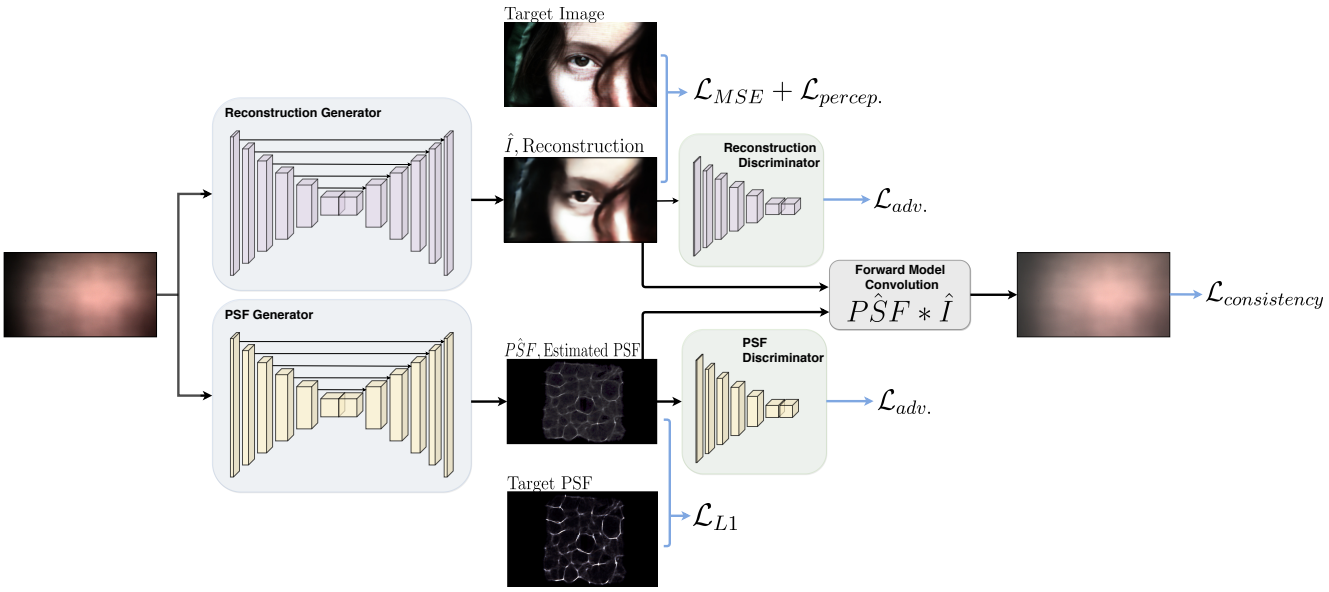

Figure 1: **Network Architecture:** Our lensless image reconstruction pipeline consisting of generative adversarial networks for image reconstruction and point spread function (PSF) estimation. In addition to regular and adversarial loss functions, a perceptual loss and a self-supervised consistency loss with respect to the forward imaging model is incorporated.

2. Related work

Lensless Cameras: The first lensless cameras are actually the oldest camera technology: pinhole cameras from classical times. While classical pinholes suffer from low-light throughput, the principle of coded aperture via multiple pinholes has been used extensively in astronomy and scientific imaging [12, 10]. Coded aperture cameras have been developed using mask patterns [27, 40] or diffraction gratings [18] to perform light field-based applications such as post-capture refocusing and novel view synthesis. Another parallel track of lensless imaging utilizes spatial light modulators and compressive sensing to recover images, commonly known as single-pixel cameras [13, 20].

In this paper, we are primarily concerned with lensless cameras which feature an optical element above the sensor to replace the lens, and typically these optical elements are thin and scalable to a small size [39, 4, 2, 14]. FlatCam [4] implements a coded mask with transparent and opaque features that multiplex light from the scene to cause unique mask shadows on the sensor. This concept has been extended for microscopy applications [1] as well as capturing both images and depth information [19, 44]. Since amplitude masks lose some light efficiency, newer designs have featured phase masks [6, 43] for improved performance. Diffraction gratings [14, 15] and Fresnel plates [37] have also been used to achieve small form factors. Finally, DiffuserCam utilizes a diffusion layer as the mask [2]. Images for the system are captured in a diffused pattern on the image sensor from which the scene can be recovered using an algorithm with the appropriately calibrated PSF image. In our paper, we build off the diffusion mask-based lensless camera and aim to improve reconstruction quality for this style of lensless cameras.

Lensless Image Reconstruction: Most of the cameras presented above utilize optimization algorithms to solve the inverse problem and recover back the original image of the scene, typically with a variation on either alternating direction method of multipliers (ADMM) [8] or some regularized ℓ_1 [5] or total variation regularization [33], where are adapted for lensless imaging [4, 2]. These optimization algorithms have the advantage of being data-agnostic, and thus have wide generality compared to networks, but as iterative methods suffer from computational inefficiency.

Recently deep learning has shown superior performance at lensless image reconstruction and other vision tasks. A ResNet architecture was used to construct images from bare sensor measurements in [34]. This model performs fairly well for reconstructions, but is limited to images captured at close distances which allows for high intensity features of the image on the sensor. Other deep learning for lensless optics works have been presented for glass diffusers [28], scattering media [29], and spatial light modulators [35]. For vision applications, FlatCam was able to expand to utilize deep learning to incorporate facial recognition on lensless captured images with several object detection and classification Convolution Neural Networks [38].

The two works most closest to our approach are Khan et al. [24] and Monakhova et al. [30]. In [24], recon-

struction on FlatCam images was achieved using generative adversarial networks and perceptual loss, and did not require a PSF at test time similar to our network. This paper achieves state-of-the-art reconstruction results, however it is designed specifically for the mask patterns of Flatcam, and is trained for only a single PSF. In contrast, our method can handle multiple PSFs at test-time through our PSF estimation sub-network.

In [30], a series of neural networks to perform unrolled ADMM optimization with learnable parameters coupled with (optional) U-Net denoisers was presented and evaluated on DiffuserCam images. This paper is the state-of-the-art at the moment and we utilize this for our comparison for experimental results. However, the method requires providing a PSF obtained via calibration to output high quality reconstructions. Our paper implements a GAN-based architecture instead to show that high reconstruction quality can be achieved without the need for a calibration PSF image, while also incorporating the ability to estimate the PSF in our architecture.

3. Approach

Forward Imaging Model: Our problem definition for lensless imaging formulates measurements from a lensless camera as a linear model. Namely, our forward model is that the captured lensless image $I_{lensless}$ is governed by the following convolution:

$$I_{lensless} = I * PSF, \tag{1}$$

where I is the true spatial image, and PSF is the point spread function of the optical element for the lensless camera. Note that this model is valid based on Fourier optics assuming Fraunhofer diffraction [17]. Thus we wish to recover I from $I_{lensless}$, which is a deconvolution problem. However, the difficulty of our particular problem is that the PSF is also unknown, and needs to be estimated at the same time as I.

Network Architecture: As noted in the section before, deep learning has become the state-of-the-art in lensless image reconstruction. In this paper, we describe our approach to building a network architecture for both image reconstruction and blind point spread function (PSF) estimation. Our full architecture is displayed in Figure 1, and will be explicated in this section below. Besides the usual mean-squared error loss on both image and PSF ground truths, we also utilize adversarial, perceptual, and a forward model consistency losses to help augment the performance of the network. In Section 4, we discuss our network implementation and training details, and demonstrate the results of our experiments in Section 5.

The basic building block for the network architecture is the U-Net [32] that forms the backbone of most image reconstruction architectures. We utilize a ten layer network (five encoding and five decoding convolution layers) with skip connections to help propagate high frequency information from the input image through the network. As we do not have a prior PSF information from calibration, we utilize two U-Net architectures: one for generating an estimate of the image reconstruction and one for the PSF estimate. These are labeled as generator networks in Figure 1 as they will form the basis for generative adversarial networks described later in this section to improve performance.

We train these networks using mean-squared error (MSE) by finding the loss between the reconstructed image and the ground truth image, and a ℓ_1 loss for the PSF estimation:

$$\mathcal{L}_{MSE}(I, \hat{I}) = ||I - \hat{I}||_2^2, \tag{2}$$
$$\mathcal{L}_1(PSF, P\hat{S}F) = ||PSF - P\hat{S}F||_1,$$

where I is the ground truth image, $\hat{I}$ is the generator estimated output image, PSF is the ground truth point spread function, and $P\hat{S}F$ is the generator estimated PSF. Note that the calibrated PSF is required at training, but will not be required at test time for the network. We found that the ℓ_1 norm was necessary for the PSF estimation rather than ℓ_2 especially for multiple PSFs, because otherwise the network would converge to an average PSF rather than learning to estimate the individual PSFs. For the image reconstruction, while MSE alone can get us most of way to the ground truth image, using it as the sole loss function still results in slightly blurred features in the output. We note that skip connections, while useful in general for preserving high frequency connections, are not as effective in lensless image reconstructions as the original images have significant blur, but still manage to help improve the performance empirically of the reconstruction.

Perceptual Loss: Since the image reconstruction handled by the network is ill-posed, particularly without known PSF information, we require additional losses to help improve performance. Similar to [24], we implement a perceptual loss [22] to help improve the visual fidelity of our reconstructions. Using a pre-trained VGG-16 network, we extract features $\phi(I)$ of an image and compute the following perceptual loss:

$$\mathcal{L}_{percep}(I, \hat{I}) = ||\phi(I) - \phi(\hat{I})||_2^2. \tag{3}$$

In training, we first train with only the MSE loss first for some epochs before adding in the perceptual loss (as well as the adversarial and consistency loss described below). We delay training for the perceptual loss so that the MSE can converge to an initial reconstruction with low frequency color information, and so that the perceptual loss does not overpower the MSE loss in favor of higher perceptual feature accuracy.

Adversarial Learning: Even using MSE + perceptual loss, our network did not achieve compelling lensless image reconstruction. This is probably due to the difficulty of estimating both the PSF and the reconstructed image simultaneously. Recently, adversarial losses have been shown to achieve photorealistic results via generative adversarial networks (GANs) [16]. GANs have enabled high visual fidelity in image synthesis and reconstruction tasks across computer vision. In particular, GANs were recently shown in [24] to work well for lensless image reconstruction. We also utilize GANs, however as opposed to [24], we do not first perform a preliminary reconstruction and then use GANs to clean the estimate in the second stage. Rather, we utilize adversarial learning for both the reconstruction and the PSF estimation simultaneously for a one-shot reconstruction at test time.

In Figure 1, we utilize a five convolutional layer discriminator for both the image and PSF generators. We utilize the Wasserstein GAN loss [3] for our loss function. We also perform delayed training with the adversarial loss to allow the network to first converge with MSE before adding in adversarial, perceptual, and consistency losses.

Model Consistency: Finally, as our problem differs from other lensless image reconstruction tasks in requiring blind PSF estimation, we also develop a custom loss function to check the forward model's consistency. Similar to the Reblur2Deblur [11] for motion deblurring, we utilize the following consistency loss that is semi-supervised:

$$\mathcal{L}_{consistency} = ||I_{lensless} - P\hat{S}F * \hat{I}||_2, \qquad (4)$$

where we take the outputs from the image and PSF generators, convolve them, and compare the result to the original blurred lensless input.

4. Implementation

Dataset: We utilize the DiffuserCam dataset by Monakhova et al. [30] for our comparative experiments. The dataset consists of 25000 images (24000 training, 1000 test images) from the MirFlickr dataset [21] that were displayed on a monitor and captured using both a regular camera as well as diffuser-based lensless camera, co-axially aligned via a beam-splitter. Both cameras utilized a Basler Dart (daA1920-30uc) sensor captured natively at 1920×1080 resolution, but then downsampled, flipped and cropped to 380×210 resolution to mitigate moire fringes from the screen on the lensed camera and ensure the monitor was covering the full field of view of the camera [30]. To create the simulated multiple PSF images from this dataset, we create 20 PSFs by randomly permuting sections of the original PSF area. We then use each PSF and convolve it with the ground truth data for 20 different variations of the original dataset with different PSFs used.

Network Architecture: For the generator network architecture, we use five encoding layers, five decoding layers and a convolution layer in the center and output. Skip connections are also used to transfer high-level features to the decoding layers. Each encoding layer uses two convolution layers and then performs batch normalization and a ReLU activation function. Each decoding layer uses an upsampled output from the previous layer and concatenates it with the corresponding encoding layer output before using three convolution layers. A kernel size, **k**, of 3×3 is used for all layers except the output layer and a stride, **s**, of 1 is used for all layers.

For the discriminator architecture, we use four convolution layers with batch normalization and the LeakyReLU activation function and an output layer that uses a Sigmoid activation function. Each layer uses a kernel size of 4×4 and a stride of 2 except the output layer that uses a stride of 1. Please refer to the supplemental material for details about the layer parameters for each sub-network.

Training Details: While training, we use a batch size of 20 images per iteration. This takes roughly $20 - 30$ minutes per epoch on 24000 training images. A lower batch size can also be used in case of memory constraints, but will result in slower training. We utilize a cyclic learning rate for the generator networks while training as shown in [36]. We set the step size of 2 epochs and train with cycle lengths in multiples of 4 epochs, and the minimum and maximum learning rates used were $\lambda = 1e-7$ and $\lambda = 5e-3$ respectively. While this style of learning rate scheduling results in oscillations in the training loss, the process trains much quicker than most other methods and assists with avoiding over-fitting into a particular local minima in the loss gradient. The discriminator's learning rate was fixed at $\lambda = 2e-4$. For the generator's loss functions, after pre-training with only $\mathcal{L}_{MSE}$, we assign weights to $\mathcal{L}_{MSE}$, $\mathcal{L}_{adv}$, $\mathcal{L}_{percep}$, and $\mathcal{L}_{consistency}$ of $\lambda_{MSE} = 1.0$, $\lambda_{adv} = 0.001$, $\lambda_{percep} = 0.7$, and $\lambda_{consistency} = 0 - 0.4$. This allowed the MSE loss to be dominant while still gaining desired contributions from the other losses. For the adversarial loss in particular, several values were attempted ranging from 0.001 to 0.5, to determine 0.001 as the best in our case. Finally, the Adam optimizer [25] was used with the cyclic learning rate scheduler to train the entire network.

The hardware used to train and evaluate the model uses an Intel Xeon W octa-core CPU with 64GB RAM and an Nvidia RTX 2080Ti GPU. All code to train and evaluate this model was done using Python 3 and utilized the PyTorch framework for machine learning functionality. All code, datasets, and trained models will be made available after review via Github.

Baselines and Metrics: To compare our network performance, we compare against several baseline methods for lensless reconstruction. The first is the alternating direction method of multipliers (ADMM) which solves the optimization problem [8]: $\text{argmin}_x ||y - Ax||_2 + \lambda ||x||_1$ where x is the

estimated image, A is the PSF (written in matrix form), and y is the lensless image. Our second baseline is a U-Net architecture [32] trained using the exact same architecture as our generator network. The final set of baselines are the Le-ADMM networks from [30] which is an unrolled ADMM algorithm with learnable parameters followed by an (optional) U-Net denoiser. Note that all the baselines require knowledge of the PSF from calibration to work, while our method performs on-par or slightly better than these techniques without the need for calibration at test-time.

Evaluating the quality of lensless image reconstruction is difficult, as both peak signal-to-noise ratio (PSNR) and SSIM [41] do not capture well the visual fidelity of an image with respect to human perception. For the sake of our paper, we decided to report both PSNR and SSIM, but we encourage the reader to look carefully at qualitative results to determine which reconstructions they prefer. Of course, these metrics and qualitative results can vary from image to image and dataset to dataset, but we believe general trends or observations can be gleaned from our experiments.

Real Hardware Setup: We also capture a real video dataset to capture lensless images using multiple PSFs in the lab. This dataset consists of videos captured using a Panasonic Lumix G85 mirrorless camera with a micro $4/3$ sensor. Shown in Figure 2, the diffuser is attached to the camera using a lens mount adapter with the diffusion layer taped to the back of the adapter, closest to the sensor with black electrical tape. A rectangular aperture, similar to DiffuserCam [2], is also created using the black electrical tape. The collected lensless video dataset uses the DAVIS dataset sequences as the ground truth and is recorded with the lensless camera through a monitor. The videos are recorded in 1920×1080 resolution, but downsampled to 480×270 to address moire fringes of the monitor. The final video dataset consists of 120 sequences with a total of 8,460 frames for the train set and 30 sequences with a total of 2,000 frames for the test set. Three different real PSFs were used for this dataset.

Figure 2: **Hardware Setup:** On the left, we show our custom lensless camera consisting of a Panasonic Lumix G85 mirrorless camera and diffusion layer attached with a lens mount adapter. On the right is our imaging setup with display monitor for capturing video.

5. Experimental Results

In this section, we present the results of our experiments, including both comparative studies with other methods on the DiffuserCam dataset [30] with simulated multiple PSF data as well as validation of our method for real multiple PSF data from a hardware prototype in the lab. We also test our method's robustness to additive noise, and analyze computational speed at inference.

DiffuserCam Results: In Table 1, we show the results of our model on the DiffuserCam dataset [30] as compared to our baselines. Note that our method has state-of-the-art performance on-par with Le-ADMM-U for image reconstruction (slightly lower for SSIM but slightly higher for PSNR). Further, our network has the fastest computation time at inference with 0.32 seconds, a $5\times$ improvement over Le-ADMM-U. Also our method does not require knowledge of the PSF at test-time but estimates it implicitly from the data, and thus achieving superior performance without calibration information.

Model Comparison			
Model	SSIM	PSNR (dB)	CPU Time (sec)
ADMM	0.47	11.97	36.38
Le-ADMM	0.51	11.89	1.26
Le-ADMM-S	0.59	14.61	4.22
Le-ADMM-U	0.77	20.46	1.62
U-NET	0.67	17.42	0.34
Ours	0.73	20.56	0.32

Table 1: **Single PSF Image Reconstruction Results:** Comparison of reconstruction quality as well as the computational time for network inference at test time on the DiffuserCam dataset [30]. Note that our model achieves state-of-the-art performance on-par with Le-ADMM-U, while also having the advantage of being the fastest network (0.32 seconds).

As we noted that PSNR and SSIM do not always track with perceptually better visual reconstructions. Therefore we also show qualitative reconstructions for this dataset in Figure 3. As we can see, our method visually reconstructs the lensless images fairly close to state-of-the-art methods, improving the quality over ADMM and UNet methods while being able to generalize for PSF variations. While the reconstructions are plausible, we note that there is a drop in visual quality for some images. We consider this to be the trade-off in our method between visual fidelity and generalizing to multiple PSFs. This seems to suggest that

Figure 3: **Comparison on DiffuserCam dataset [30]:** Reconstruction output of sample images from different models for the DiffuserCam dataset (with the original single PSF). Note that our model performs on-par with the state-of-the-art Le-ADMM-U model.

calibration-free lensless images can be as good as having knowledge of the PSF for deep learning reconstruction.

Noise Ablation Study: We observed that none of the models are designed to robustly handle sensor noise that is not seen during training. Simulating even a tiny amount of maximum Gaussian noise $\sigma^2 = 0.005$ led to significant degradation in SSIM and PSNR scores as seen in Figure 4. This speaks to the brittleness of these models with respect to their training data. To overcome this, we show that further training our model for about 100 epochs with images containing varying additional noise helps overcome this sensitivity to unseen noise at test-time and improve network performance. This is a useful strategy that can be used for all lensless reconstruction networks.

Generalizing to Multiple PSFs: One of the main characteristics of a diffusion-based lensless camera is that the mask's PSF are random caustic patterns. While this works well for solving the inverse problem for lensless cameras, it is limited by the need for PSF calibration which is unique to each individual diffusion mask. With vast improvement in reconstruction quality using DNNs, this further begs the question of whether it would be a feasible implementation to train a unique model for each camera or diffusion mask to

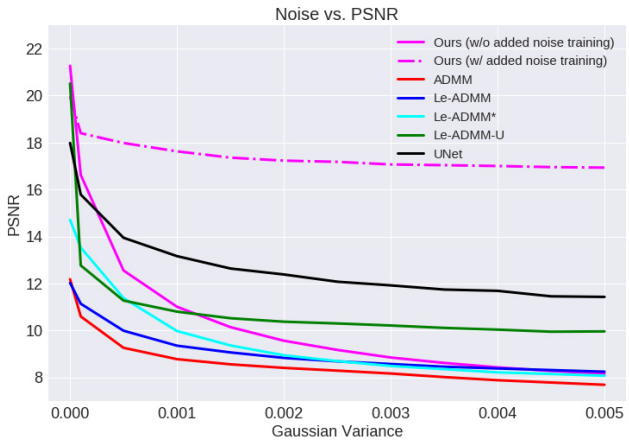

Figure 4: **Noise:** Ablative study of model performance with respect to additive Gaussian noise. Notice how even an extremely small amount of noise ($\sigma^2 = 0.005$) results in degradation of performance for all methods. However, specifically training our model with noise results in more robust performance.

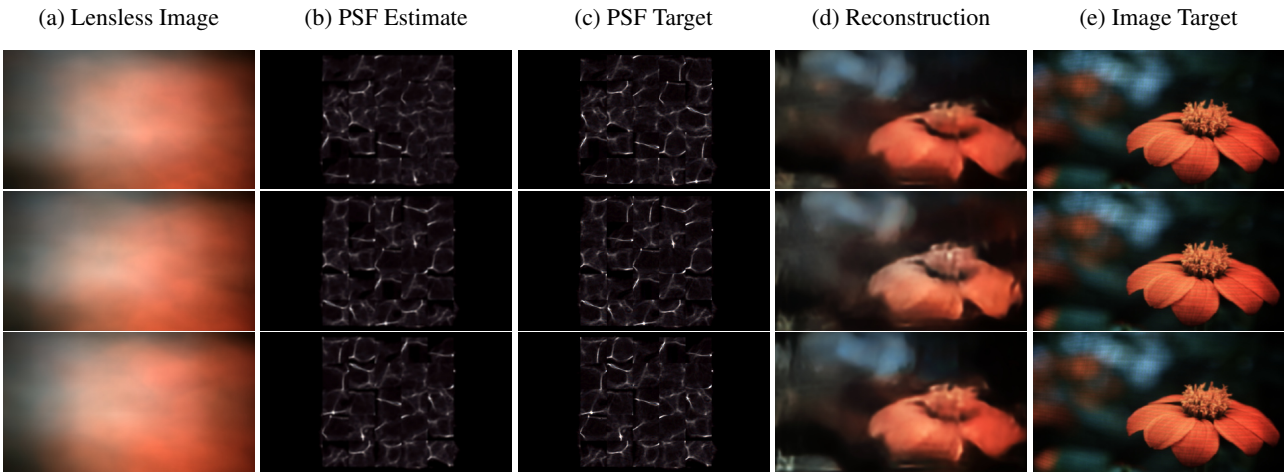

Figure 5: **Multiple PSF Reconstruction:** We display the PSF estimated as well as the reconstructed images for three different simulated PSFs on the DiffuserCam dataset [30]. Our method is able to generalize and reconstruct multiple PSFs, although the quality of the reconstruction is missing some high frequency information. However, the resulting reconstructions are uniform in quality across the three different PSFs. This shows the potential for generalization of our network for multiple lensless camera without training a separate model for each one.

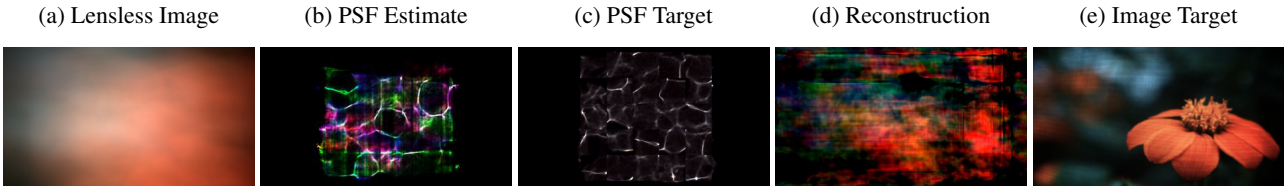

Figure 6: **Blind Deconvolution Reconstruction:** We display results from a blind deconvolution method in [31], which struggles to estimate the PSF or the image correctly. Blind deconvolution methods traditionally work well for estimating motion blur or more structured optical blur PSFs as opposed to the randomness of a diffusion PSF.

be used. Being able to generalize a single trained model to reconstruct for variations in the PSF will further help make lensless cameras more usable as reconstruction quality continues to improve towards photo-realistic quality.

Our results for training for multiple PSFs are shown in Figure 5 and give us promising results for both estimating the PSF and reconstructing the corresponding image. We see that the resulting PSF estimations are overall very close to the target PSF, with very slight deviations. These estimated PSFs are still useful for the network to perform image reconstruction including learning variations in the PSF through the self-supervised consistency loss. Quantitatively, we report an average PSNR of 35.24 dB and an average SSIM value of 0.812 for PSF estimation on our synthetic dataset.

The resulting image reconstructions show the ability for the generator to perform reconstruction for the different PSF, although with a drop in output quality when compared to the single PSF trained network. This may be an expected trade-off for output quality versus generalizability. Keeping in mind that the lensless data used for this section was simulated by performing the forward model convolution and true PSF variations from different masks can prove to be a much harder problem as shown with real captured data.

Comparison with Blind Deconvolution: Our image reconstruction problem is also very similar to the problem of blind deconvolution typically found in motion or image deblurring tasks, where an unknown blur kernel has been convolved across the image. However, we note that the blur kernels for motion or image blur are typically more structured and less random as lensless camera PSFs. Nevertheless, we utilize a blind deconvolution method [31] for our task in Figure 6. As you can see, the blind deconvolution method does not adequately reconstruct either the PSF or image as compared to our network architecture results.

Real Hardware Prototype Results: We also attempt to perform reconstruction on our own captured video dataset with three PSF variations, shown in Figure 7. While the results are not quite the same reconstruction quality as the DiffuserCam dataset, we can still see recovered regions that

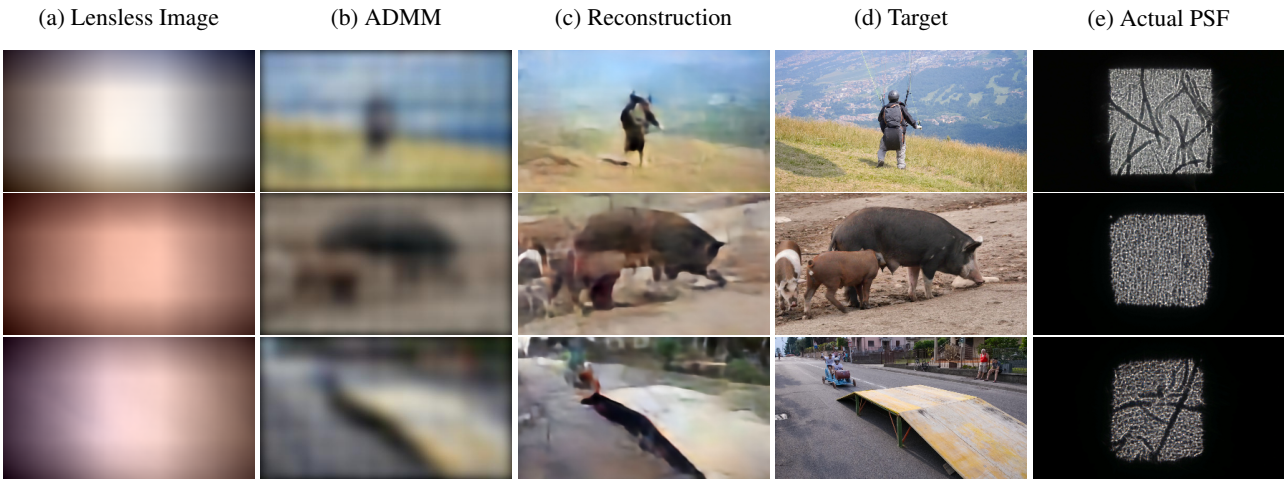

Figure 7: **Real Data with Multiple PSFs:** Example reconstructions of real image data captured with three different PSFs are shown in (c). For comparison, the ADMM algorithm, with the known PSFs, reconstructed the images in column (b). Reconstruction performance across both methods was not as visually pleasing, primarily due to the real PSFs (shown in (e)) quality in captured caustics. Despite this limitation, the network is still able to invert the measurements to achieve plausible reconstructions at higher fidelity than ADMM.

resemble the target image. For comparison, we also show the results of the ADMM algorithm on our real data, which performed worse than our network even when given the actual PSF at reconstruction time.

Trying to reconstruct for our captured dataset proved to be a much more challenging task. We hypothesize that this is primarily due to the diffusion layer used and the resulting PSF of each dataset. The DiffuserCam dataset utilizes a very small angle optical diffuser which results in more sparce and spread out caustics [2, 30], while our diffusion layer uses a plastic polyethylene terephthalate (PET) layer, commonly used for anti-static bags. The diffusion caused by this material, while still random, has a much larger spread of light with more dense caustics as seen in the figure. Even with this limitation, analysis for the trained model on the captured test set resulted in an average PSNR of 19.88 dB and and an average SSIM of 0.6769. While this is comparable to some of the better models compared in Table 1, we believe these metrics may not be the best comparisons for the perceptual quality of the reconstructed images as seen from the image results. It remains an avenue of future research to quantify which PSFs and their caustic patterns are ideal for lensless image reconstruction.

6. Conclusion

In this paper, we proposed a method for lensless image reconstruction that can generalize across varying point spread functions (PSFs) corresponding to different cameras. This allows our model to train and perform reconstruction without the need for any PSF calibration and does not simply learn parameters to a single PSF. This is achieved through a combination of adversarial training, PSF estimation, and a cycle consistency loss. We verify this approach on the DiffuserCam dataset, a simulated multi-PSF dataset, as well as our own captured video dataset. We show that our network achieves performance on-par with state-of-the-art methods without requiring PSF information at test-time, and with low computational latency. However, we point out challenges underlying the robustness of lensless image reconstruction including sensitivity to added noise and dependence on high quality caustics in the PSF as demonstrated in our real hardware prototype experiments. Despite these challenges, our network is able to achieve plausible visual reconstruction.

There are many avenues for future work to make lensless imaging more robust. This includes improving PSF estimation or using unsupervised or self-supervised techinques only so that calibration is not needed even during training. Our images, while at modest resolutions, are not yet reconstructed at full 1080p resolution, and techniques such as progressive GANs may help here [23]. Finally in addition to single frame reconstruction, lensless video reconstruction is another opportunity for deep learning techniques in general.

Acknowledgements: We wish to thank Celine Cheung and Yasser Dbeis for help with lensless camera simulation and video dataset creation respectively. This work was supported by NSF grant IIS-1909192, a seed grant from ASU's Herberger Research Initiative, as well as GPU resources from ASU Research Computing.

References

[1] Jesse K Adams, Vivek Boominathan, Benjamin W Avants, Daniel G Vercosa, Fan Ye, Richard G Baraniuk, Jacob T Robinson, and Ashok Veeraraghavan. Single-frame 3d fluorescence microscopy with ultraminiature lensless flatscope. *Science Advances*, 3(12):e1701548, 2017.

[2] 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, 01 2018.

[3] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 214–223, 2017.

[4] M. S. Asif, A. Ayremlou, A. Sankaranarayanan, A. Veeraraghavan, and R. G. Baraniuk. Flatcam: Thin, lensless cameras using coded aperture and computation. *IEEE Transactions on Computational Imaging*, 3(3):384–397, 09 2017.

[5] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

[6] Vivek Boominathan, Jesse Adams, Jacob Robinson, and Ashok Veeraraghavan. Phlatcam: Designed phase-mask based thin lensless camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[7] 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.

[8] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011.

[9] Thuong Nguyen Canh and Hajime Nagahara. Deep compressive sensing for visual privacy protection in flatcam imaging. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 3978–3986. IEEE, 2019.

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

[11] Huaijin Chen, Jinwei Gu, Orazio Gallo, Ming-Yu Liu, Ashok Veeraraghavan, and Jan Kautz. Reblur2deblur: Deblurring videos via self-supervised learning. In *2018 IEEE International Conference on Computational Photography (ICCP)*, pages 1–9. IEEE, 2018.

[12] RH Dicke. Scatter-hole cameras for x-rays and gamma rays. *The Astrophysical Journal*, 153:L101, 1968.

[13] Marco F Duarte, Mark A Davenport, Dharmpal Takhar, Jason N Laska, Ting Sun, Kevin F Kelly, and Richard G Baraniuk. Single-pixel imaging via compressive sampling. *IEEE Signal Processing Magazine*, 25(2):83–91, 2008.

[14] Patrick Robert Gill, Changhyuk Lee, Dhon-Gue Lee, Albert Wang, and Alyosha Molnar. A microscale camera using direct fourier-domain scene capture. *Optics Letters*, 36(15):2949–2951, 2011.

[15] Patrick R Gill and David G Stork. Lensless ultra-miniature imagers using odd-symmetry spiral phase gratings. In *Computational Optical Sensing and Imaging*, pages CW4C–3. Optical Society of America, 2013.

[16] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

[17] Joseph W Goodman. *Introduction to Fourier optics*. Roberts and Company Publishers, 2005.

[18] Matthew Hirsch, Sriram Sivaramakrishnan, Suren Jayasuriya, Albert Wang, Alyosha Molnar, Ramesh Raskar, and Gordon Wetzstein. A switchable light field camera architecture with angle sensitive pixels and dictionary-based sparse coding. In *2014 IEEE International Conference on Computational Photography (ICCP)*, pages 1–10. IEEE, 2014.

[19] Yi Hua, Shigeki Nakamura, Salman Asif, and Aswin Sankaranarayanan. Sweepcam-depth-aware lensless imaing using programmable masks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[20] 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.

[21] Mark J. Huiskes and Michael S. Lew. The mir flickr retrieval evaluation. In *MIR '08*, 2008.

[22] 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.

[23] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[24] Salman S Khan, VR Adarsh, Vivek Boominathan, Jasper Tan, Ashok Veeraraghavan, and Kaushik Mitra. Towards photorealistic reconstruction of highly multiplexed lensless images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7860–7869, 2019.

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

[26] Grace Kuo, Nick Antipa, Ren Ng, and Laura Waller. 3d fluorescence microscopy with diffusercam. In *Computational Optical Sensing and Imaging*, pages CM3E–3. Optical Society of America, 2018.

[27] Anat Levin, Rob Fergus, Frédo Durand, and William T Freeman. Image and depth from a conventional camera with a coded aperture. *ACM Transactions on Graphics (TOG)*, 26(3), 2007.

[28] Shuai Li, Mo Deng, Justin Lee, Ayan Sinha, and George Barbastathis. Imaging through glass diffusers using densely connected convolutional networks. *Optica*, 5(7):803–813, 2018.

[29] Yunzhe Li, Yujia Xue, and Lei Tian. Deep speckle correlation: a deep learning approach toward scalable imaging through scattering media. *Optica*, 5(10):1181–1190, 2018.

[30] Kristina Monakhova, Joshua Yurtsever, Grace Kuo, Nick Antipa, Kyrollos Yanny, and Laura Waller. Learned reconstructions for practical mask-based lensless imaging. *Opt. Express*, 27(20):28075–28090, 09 2019.

[31] Daniele Perrone and Paolo Favaro. Total variation blind deconvolution: The devil is in the details. In *EEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[32] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, 2015.

[33] Leonid I Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1-4):259–268, 1992.

[34] Ayan Sinha, Justin Lee, Shuai Li, and George Barbastathis. Lensless computational imaging through deep learning. *Optica*, 4(9):1117–1125, 09 2017.

[35] Ayan Sinha, Justin Lee, Shuai Li, and George Barbastathis. Lensless computational imaging through deep learning. *Optica*, 4(9):1117–1125, 2017.

[36] Leslie N. Smith. Cyclical learning rates for training neural networks. *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 3 2017.

[37] Kazuyuki Tajima, Takeshi Shimano, Yusuke Nakamura, Mayu Sao, and Taku Hoshizawa. Lensless light-field imaging with multi-phased fresnel zone aperture. In *2017 IEEE International Conference on Computational Photography (ICCP)*, pages 1–7. IEEE, 2017.

[38] J. Tan, L. Niu, J. K. Adams, V. Boominathan, J. T. Robinson, R. G. Baraniuk, and A. Veeraraghavan. Face detection and verification using lensless cameras. *IEEE Transactions on Computational Imaging*, 5(2):180–194, 06 2019.

[39] Jun Tanida, Tomoya Kumagai, Kenji Yamada, Shigehiro Miyatake, Kouichi Ishida, Takashi Morimoto, Noriyuki Kondou, Daisuke Miyazaki, and Yoshiki Ichioka. Thin observation module by bound optics (tombo): concept and experimental verification. *Applied Optics*, 40(11):1806–1813, 2001.

[40] Ashok Veeraraghavan, Ramesh Raskar, Amit Agrawal, Ankit Mohan, and Jack Tumblin. Dappled photography: Mask enhanced cameras for heterodyned light fields and coded aperture refocusing. In *ACM Transactions on Graphics (TOG)*, volume 26, page 69. ACM, 2007.

[41] Zhou Wang, Alan C Bovik, Hamid R Sheikh, Eero P Simoncelli, et al. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[42] Zihao W Wang, Vibhav Vineet, Francesco Pittaluga, Sudipta N Sinha, Oliver Cossairt, and Sing Bing Kang. Privacy-preserving action recognition using coded aperture videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[43] Yicheng Wu, Vivek Boominathan, Huaijin Chen, Aswin Sankaranarayanan, and Ashok Veeraraghavan. Phasecam3d—learning phase masks for passive single view depth estimation. In *2019 IEEE International Conference on Computational Photography (ICCP)*, pages 1–12. IEEE, 2019.

[44] Yucheng Zheng and M Salman Asif. Joint image and depth estimation with mask-based lensless cameras. *arXiv preprint arXiv:1910.02526*, 2019.