Unrolled, model-based networks for lensless imaging

Kristina Monakhova *
UC Berkeley
monakhova@

Joshua Yurtsever
UC Berkeley
joshua.yurtsever@

Grace Kuo
UC Berkeley
gkuo@

Nick Antipa
UC Berkeley
nick.antipa@

Kyrollos Yanny†
UC Berkeley and UCSF
kyrollosyanny@

Laura Waller
UC Berkeley
waller@

Abstract

We develop end-to-end learned reconstructions for lensless mask-based cameras, including an experimental system for capturing aligned lensless and lensed images for training. Various reconstruction methods are explored, on a scale from classic iterative approaches (based on the physical imaging model) to deep learned methods with many learned parameters. In the middle ground, we present several variations of unrolled *alternating direction method of multipliers* (ADMM) with varying numbers of learned parameters. The network structure combines knowledge of the physical imaging model with learned parameters updated from the data, which compensate for artifacts caused by physical approximations. Our unrolled approach is 20× faster than classic methods and produces better reconstruction quality than both the classic and deep methods on our experimental system.

1 Introduction

Mask-based lensless cameras can be small, light-weight, and capture higher-dimensional information, such as 3D and video, from a single shot [1–3, 8, 13]. Instead of using a lens, lensless cameras use a phase or amplitude mask which maps points in the world to a unique multiplexed pattern on the sensor (Fig. 1(a)). Typically, a reconstruction method based on convex optimization is used to iteratively solve for the scene from the multiplexed sensor data. In practice, iterative methods can be slow and the reconstruction quality is sensitive to errors from model mismatch, imperfect calibration, hand-tuned parameters, and hand-picked priors which are not necessarily representative of the data. Solving the inverse problem with deep methods offers a favorable alternative due to the decreased computation and the ability to directly optimize image quality. However, this comes at the price of thousands of training pairs, a loss of interpretability, and the inability to explicitly add prior knowledge, such as the imaging system physics, into the network.

Unrolled optimization has emerged as a promising middle-ground approach between classic and deep methods for a variety of inverse problems [5–7, 11, 12]. In unrolled optimization, a fixed number of iterations from a classic algorithm is interpreted as a deep network, with each iteration serving as a layer in the network. In each layer, if the parameters of the algorithm are differentiable with respect to the output, they can be optimized for a given loss function through backpropagation. Following this framework, we unroll the iterative *alternating direction method of multipliers* (ADMM) algorithm with a variable splitting specific for lensless imaging [1, 4]. This allows us to incorporate knowledge of the image formation process into the neural network as well as directly optimize

*Department of Electrical Engineering & Computer Sciences, University of California, Berkeley. All emails @berkeley.edu

†Department of Bioengineering, University of California, Berkeley and University of California, San Francisco

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

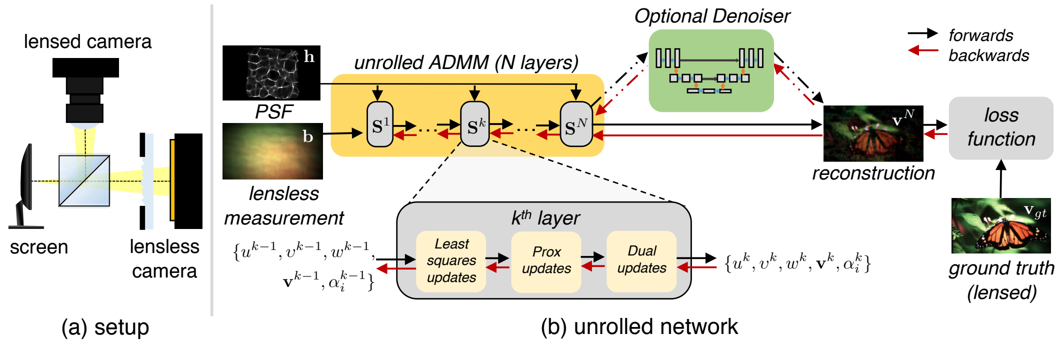

Figure 1: (a) Experimental Setup. We display images on a computer screen and use a beamsplitter to simultaneously record measurements on both a lensed and lensless camera, for training. (b) Unrolled network architecture. The input measurement and the calibration PSF are first fed into N layers of unrolled ADMM. At each layer, the updates corresponding to $\mathbf{S}^{k+1}$ in Eq. (3) are applied. The output of this can be fed into an optional denoiser network such as a U-Net [10]. The network parameters are updated based on a loss function comparing the output image to the lensed image. Red arrows represent backpropagation through the network parameters.

image reconstruction quality based on training examples. We present and experimentally test several variations of networks along the spectrum between classic and deep methods, by varying the number of trainable parameters.

To train our networks, we experimentally capture a large dataset of 25,000 aligned lensed and lensless images (Fig. 1(a)). Our full dataset and models are publicly available. We demonstrate a 20× speedup and 3× improvement in perceptual similarity for lensless imaging reconstructions, showing that our unrolled method outperforms both the classic and deep approaches in terms of visual quality. Finally, we demonstrate the generalization of our network to measurements taken in the wild.

2 Methods

To formulate our unrolled network, we first describe our imaging forward model and classic reconstruction algorithm. Our lensless imaging model can be approximated as a cropped convolution between the scene and the point spread function (PSF) of the system. The PSF is measured experimentally using an LED placed at the desired focal distance of the system. Assuming all points in the scene are incoherent with each other, our sensor measurement, $\mathbf{b}$, can be described as:

$$\begin{aligned}\mathbf{b}(x,y) &= \text{crop}[\mathbf{h}(x,y) * \mathbf{x}(x,y)] \\ &= \mathbf{CHx},\end{aligned} \tag{1}$$

where $\mathbf{h}$ is the system PSF, $\mathbf{x}$ represents the scene, and (x, y) are the sensor coordinates. Here, $*$ denotes 2D discrete linear convolution, which returns an array that is larger than both the scene and the PSF. Therefore, a crop operation restricts the output to the physical sensor size. This relation is represented compactly in matrix-vector notation with crop denoted as $\mathbf{C}$ and convolution with the PSF denoted as $\mathbf{H}$.

To efficiently solve the inverse problem, we use ADMM with a variable splitting that leverages the structure of the problem. The inverse problem is formulated as:

$$\begin{aligned}\hat{\mathbf{x}} = \underset{w\geq 0,u,v}{\arg\min}\ &\frac{1}{2}\|\mathbf{b} - \mathbf{C}v\|_2^2 + \tau\|u\|_1, \\ &s.t.\ v = \mathbf{Hx}, u = \Psi\mathbf{x}, w = \mathbf{x},\end{aligned} \tag{2}$$

where Ψ is a sparsifying transform, such as finite differences for total variation (TV) denoising, and τ is a tuning parameter that adjusts the sparsity level. The update equations for each iteration become:

$$
\mathbf{S}^{k+1} \leftarrow \begin{cases} u^{k+1} \leftarrow \mathcal{T}_{\tau^k}(\mathbf{\Psi}(\mathbf{x}^k) + \alpha_2^k/\mu_2^k) & \text{sparsifying soft-thresholding} \\ v^{k+1} \leftarrow (\mathbf{C^TC} + \mu_1 I)^{-1}(\alpha_1^k + \mu_1^k \mathbf{H}\mathbf{x}^k + \mathbf{C^T b}) & \text{least-squares update} \\ w^{k+1} \leftarrow \max(\alpha_3^k/\mu_3^k + \boldsymbol{x}^k, 0) & \text{enforce non-negativity} \\ \mathbf{x}^{k+1} \leftarrow (\mu_1^k \mathbf{H^TH} + \mu_2^k \mathbf{\Psi^T\Psi} + \mu_3^k I)^{-1} r^k & \text{least-squares update} \\ \alpha_1^{k+1} \leftarrow \alpha_1^k + \mu_1^k(\mathbf{H}\mathbf{x}^{k+1} - v^{k+1}) & \text{dual for v} \\ \alpha_2^{k+1} \leftarrow \alpha_2^k + \mu_2^k(\mathbf{\Psi}(\mathbf{x}^{k+1}) - u^{k+1}) & \text{dual for u} \\ \alpha_3^{k+1} \leftarrow \alpha_3^k + \mu_3^k(\mathbf{x}^{k+1} - w^{k+1}) & \text{dual for w} \end{cases} \tag{3}
$$

$$
\text{where } r^k = ((\mu_3^k w^{k+1} - \alpha_3^k) + \mathbf{\Psi^T}(\mu_2^k u^{k+1} - \alpha_2^k) + \mathbf{H^T}(\mu_1^k v^{k+1} - \alpha_1^k)).
$$

Here, α_1, α_2, and α_3 are the Lagrange multipliers, or dual variables, respectively associated with u, v, and w, and μ_1, μ_2, and μ_3 are scalar penalty parameters. $\mathcal{T}_{\tau/\mu_2}$ denotes vectorial soft-thresholding with parameter τ/μ_2. To unroll the network, we model each k^{th} iteration of ADMM as a layer in a neural network. We denote the collection of update equations at the k^{th} step of ADMM as $\mathbf{S}^k$.

We analyze three variations of unrolled ADMM, each having a different number of learned parameters, denoted by Θ (Fig. 1(b)). The three variations are summarized as:

- **Le-ADMM** (20 parameters, $\Theta = \{\mu_1^k, \mu_2^k, \mu_3^k, \tau^k\}$) - Learned ADMM has trainable tuning and hyper-parameters.
- **Le-ADMM*** (32,135 parameters, $\Theta = \{\mu_1^k, \mu_2^k, \mu_3^k, \boldsymbol{\mathcal{N}}\}$) - extends Le-ADMM by adding a trainable convolutional neural network (CNN) instead of a hand-tuned sparsifying transform. $\boldsymbol{\mathcal{N}}$ represents a learnable network and replaces the u^{k+1} update of Eq. (3).
- **Le-ADMM-U** (10,605,927 parameters, $\Theta = \{\mu_1^k, \mu_2^k, \mu_3^k, \tau^k, \boldsymbol{\mathcal{U}}\}$) - adds a trainable deep denoiser based on a CNN as the last layer of the Le-ADMM network, learning both the hyper-parameters of Le-ADMM as well as the denoiser.

For training, we simultaneously collect a set of lensless and ground truth (lensed camera) image pairs using an experimental setup with a beamsplitter to send the light to both, and computer monitor to display training images (see Fig. 1(a)). We use two Basler Dart (daA1920-30uc) sensors; our lensed camera has a 6mm focal length S-mount lens (*lensed*), and our lensless camera has an off-the-shelf phase mask (Luminit 0.5°) and laser-cut aperture (*lensless*). To achieve pixel-wise alignment between the image pairs, we first optically align the two cameras, then perform a digital calibration process to co-align both cameras' coordinate systems. We capture 25,000 images from the MirFlickr dataset [9]. After down-sampling and cropping, the final images are 380×210 pixels, separated into 24,000 training images and 1,000 test images. We use a combination of mean-squared error (MSE) and LPIPS from [14] for training.

3 Results

Figure 2 summarizes the performance of our networks on the test set. We compare against ADMM run to convergence (100 iterations), ADMM bounded to 5 iterations (similar run time to our unrolled network), as well as against an end-to-end trained U-Net [10]. All of our networks perform better or comparably to converged ADMM, with the best networks operating 20× faster. Le-ADMM-U has the most learned parameters out of our networks and also has the best performance, achieving a better MSE and LPIPS score than traditional ADMM as well as the U-Net. Figure 3 shows several sample reconstructions from our test set as well as some reconstructions of images taken in the wild without the beamsplitter and computer monitor.

Our work presents a preliminary analysis of using unrolled, model-based neural networks on a real experimental lensless imaging system. We show that it is favorable to use a network that incorporates both knowledge of the imaging system physics and trainable parameters to optimize the network performance. We can perform comparably to classic algorithms at a fraction of the speed using only a few learned parameters, but can greatly improve image quality when increasing the number of learned parameters. Our learned network is fast enough for interactive previewing of the scene and also produces visually appealing images, addressing two of the big limitations of lensless imagers.

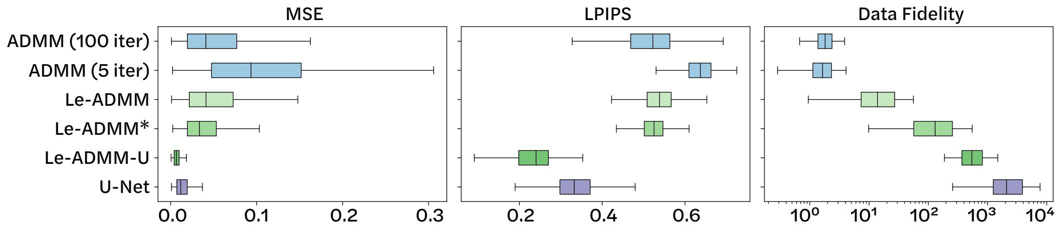

Figure 2: Network Performance on test set. On average, reconstructions from our learned networks (green) are more similar to the ground truth lensed images (lower MSE and LPIPS) than those from 5 iterations of ADMM. Furthermore, our networks have comparable or better performance than bounded ADMM (100 iterations), which takes 20× longer than Le-ADMM and Le-ADMM-U. The data fidelity term is higher for the learned methods, indicating that these reconstructions are less consistent with the image formation model. This suggests that the models are compensating for our simple forward model and are able to produce higher quality images.

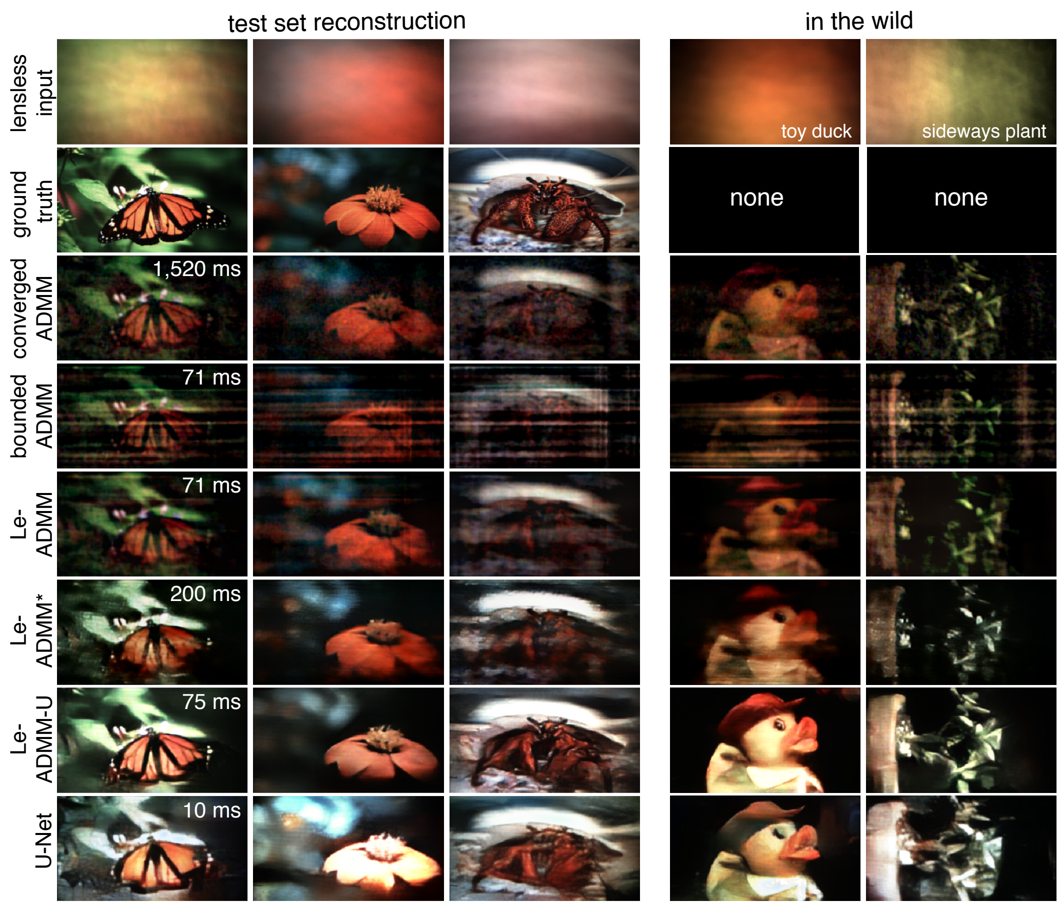

Figure 3: Reconstruction results for methods along the spectrum from classic to deep, with the raw lensless measurement (contrast stretched) and the ground truth images from the lensed camera for reference. Le-ADMM has similar image quality to converged ADMM and better image quality than bounded ADMM (5 iter). Le-ADMM* and Le-ADMM-U have noticeably better visual image quality. The fully deep U-Net by itself is unable to reconstruct the appropriate colors and lacks detail.

References

[1] N. Antipa, G. Kuo, R. Heckel, B. Mildenhall, E. Bostan, R. Ng, and L. Waller. DiffuserCam: lensless single-exposure 3D imaging. *Optica*, 5(1):1–9, 2018.

[2] N. Antipa, P. Oare, E. Bostan, R. Ng, and L. Waller. Video from Stills: Lensless Imaging with Rolling Shutter. *arXiv preprint arXiv:1905.13221*, 2019.

[3] M. S. Asif, A. Ayremlou, A. Veeraraghavan, R. Baraniuk, and A. Sankaranarayanan. FlatCam: Replacing lenses with masks and computation. In *Computer Vision Workshop (ICCVW), 2015 IEEE International Conference on*, pages 663–666. IEEE, 2015.

[4] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine learning*, 3(1):1–122, 2011.

[5] S. Diamond, V. Sitzmann, S. Boyd, G. Wetzstein, and F. Heide. Dirty pixels: Optimizing image classification architectures for raw sensor data. *arXiv preprint arXiv:1701.06487*, 2017.

[6] S. Diamond, V. Sitzmann, F. Heide, and G. Wetzstein. Unrolled optimization with deep priors. *arXiv preprint arXiv:1705.08041*, 2017.

[7] K. Gregor and Y. LeCun. Learning fast approximations of sparse coding. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, pages 399–406. Omnipress, 2010.

[8] R. Horisaki, S. Irie, Y. Ogura, and J. Tanida. Three-Dimensional Information Acquisition Using a Compound Imaging System. *Optical Review*, 14(5):347–350, 2007. ISSN 1349-9432. doi: 10.1007/s10043-007-0347-z. URL `http://dx.doi.org/10.1007/s10043-007-0347-z`.

[9] M. J. Huiskes and M. S. Lew. The MIR Flickr Retrieval Evaluation. In *MIR '08: Proceedings of the 2008 ACM International Conference on Multimedia Information Retrieval*, New York, NY, USA, 2008. ACM.

[10] O. Ronneberger, P. Fischer, and T. 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.

[11] U. Schmidt and S. Roth. Shrinkage fields for effective image restoration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2774–2781, 2014.

[12] J. Sun, H. Li, and Z. Xu. Deep ADMM-Net for compressive sensing MRI. In *Advances in neural information processing systems*, pages 10–18, 2016.

[13] J. Tanida, T. Kumagai, K. Yamada, S. Miyatake, K. Ishida, T. Morimoto, N. Kondou, D. Miyazaki, and Y. Ichioka. Thin observation module by bound optics: concept and experimental verification. *Applied Optics*, 40(11):1806–1813, 2001.

[14] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018.