End-to-End Computational Lensless Imaging with Perceptual Loss

Ya-Ti Chang Lee[1], Chung-Hao Tien[1]

yati.ee06@m365.nycu.edu.tw
[1]Department of Photonics, College of Electrical and Computer Engineering,
National Yang Ming Chiao Tung University, Hsinchu 30010, Taiwan
Keywords: Lensless imaging, perceptual loss, artificial neural network, coded aperture.

ABSTRACT

Recently, computational lensless imaging had been making progress with the evolution of artificial neural networks. Nonetheless, generative models for image reconstruction inherit challenge due to its ill-posed nature. We proposed a deep neural network based lensless imaging system by optimizing perceptual loss exclusively to end-to-end reconstruct images conforming human preference.

1 Introduction

Computational lensless imaging system, conventional lens set is replaced by other optical media, has becoming more promising due to the rapid progress in computational power and algorithm. Typically, a sensor receives an intermediate measurement either phase or intensity modulated by an engineered optical component from a scene. Then a purpose-built algorithm is applied to reconstruct the scene through the measurement. Coded aperture imaging, inspired by pinhole array, was originally developed for X-ray or gamma ray applications based on ray optical model [1]. In recent years, coded aperture imaging is extended toward visible light with specific purposes, such as compact form factor. Certainly, cutting edge computational techniques play a crucial role for such configuration [2]. Mathematically, the imaging formation can be represented as a forward model

$$y = \Phi x, \tag{1}$$

where x denotes the object, Φ denotes the lensless imager and y denotes the intermediate image. To retrieve the original information, it involves an ill-posed inverse problem. Conventionally, it is done by solving as a Tikhonov functional [3]

$$\hat{x} = \underset{x}{\operatorname{argmin}}\{\|\Phi x - y\|_2^2 + \lambda\gamma(x)\}, \tag{2}$$

where $\|\cdot\|_2$ denotes the L^2 norm and $\gamma(x)$ is the regularization term to make the estimation $\hat{x}$ matching the prior knowledge. However, evaluating inversion directly has drawbacks that it is very sensitive to noise perturbation and a calibration procedure is usually required, result in unpleasant reconstruction quality for real world applications. Such challenges have been much improved since the introduction of deep neural networks (DNNs) for scene retrieval. Then the reconstruction $\hat{x}$ is generated by the neural network $D(y)$, which can be written as

$$\hat{x} = D(y). \tag{3}$$

Many works have demonstrated that inversions modeled by DNNs can attain relative better visual fidelity [4]–[9]. Nevertheless, some good results were obtained under some circumstances, for example, trained models were only useable for specific scenes due to the characteristic of data driven methods. Some of these studies tried to improve model generality by training many natural scenes with prior knowledges or regularizations, which was effective but introduced more hyperparameters. This made DNNs harder for training and fine tuning.

In this work, we proposed a coded mask based lensless imaging system, combining with a DNN to restore scenes from intermediate measurements. We used natural scenes to train the model for system generality as well. In contrast, we aimed to minimize perceptual loss exclusively, where perceptual loss measures the human perceptual similarity between images. Although we optimized single objective, we could achieve good perceptual quality, and had an advantage of reducing hyperparameters significantly. We also designed to train the model with end-to-end manner, which made our system more feasible for real time applications.

2 Method

2.1 Coded Mask Lensless Imager

We utilized a coded mask as lensless imager for this research, where the intensity modulation was performed through the transfer matrix Φ. For a separable coded mask, the imaging formation can be described as

$$Y = \Phi_L X \Phi_R^T, \tag{4}$$

where X is 2-dimensional scene, Y is 2-dimensional measurement, Φ_L and Φ_R are Toeplitz matrices separated from Φ. Each element of matrices is determined by encoding scheme $\varphi_i \in \{0,1\}$, which can be written as

$$\Phi_L = \begin{bmatrix} \varphi_1 & \cdots & \varphi_i & 0 & 0 & \cdots & 0 \\ 0 & \varphi_1 & \cdots & \varphi_i & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \end{bmatrix}. \tag{5}$$

In this study, we employed the encoding vector from DeWeert's work [10], which can be represented as

$$\varphi_i = [1110001011100100001111111111111]. \tag{6}$$

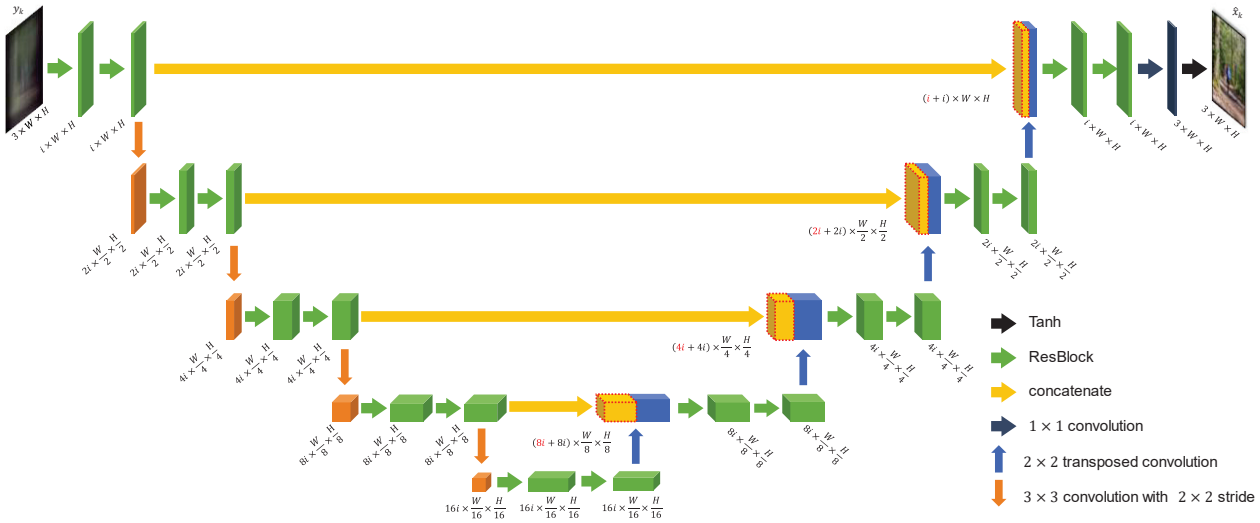

Fig. 1 Network architecture.

The network is composed of ResBlocks (green block), concatenations (yellow arrow), down sampling layers (orange block), up sampling layers (blue block) and a convolution layer (gray block) for output, where $i = 32$ in this work.

2.2 Network Architecture

Our proposed network architecture is shown in Fig. 1, which is inspired by U-Net architecture. U-Net was originally developed for image segmentation [11], but many researches had demonstrated that it is effective for image reconstruction tasks as well [4], [5], [8], [9], [12]. The key characteristic associated with U-Net is to introduced contracting path which concatenates features between layers. Such concatenations make rear layers can reuse features learned by fore layers, which improves overall performance. For improving reconstruction quality, we made some modifications to the original U-Net. We replaced convolution layers with residual blocks (ResBlock shown in Fig. 2) inspired by residual neural network and replaced max pooling layers with strided convolution layers for down sampling, which had been applied to many restoration tasks [13]–[15].

Classically, pixel wise loss functions like L1 or L2 loss, are chosen as optimization objective. But, it was shown that model tends to generate more burry reconstructions [13]. Instead, we chose perceptual loss, which measures perceptual similarity between images, as our optimization target. Perceptual loss had been introduced for image reconstruction works [4], [13], [15], where it is defined as the difference on feature maps of the pretrained VGG16 network [16]. The original loss function is written as L2 distance

$$L_{feat}^{\phi,j}(\hat{x}, x) = \frac{1}{C_j H_j W_j} \left\| \phi_j(\hat{x}) - \phi_j(x) \right\|_2^2, \qquad (7)$$

where ϕ_j is the j-th convolutional layer of the VGG16 and $C_j H_j W_j$ is the shape of ϕ_j. Instead of L2 distance, we used Charbonnier loss to evaluate the perceptual loss, where it is a stable modification of L1 loss [17]. Charbonnier loss is defined as

$$\text{Charbonnier loss}(x) = \sqrt{x^2 + \epsilon^2}, \qquad (8)$$

where $\epsilon = 1 \times 10^{-6}$ is a constant. Then the objective function to be minimized is

$$L_{feat}^{\phi,j}(\hat{x}, x) = \frac{1}{C_j H_j W_j} \sqrt{\left(\phi_j(\hat{x}) - \phi_j(x)\right)^2 + \epsilon^2}. \qquad (9)$$

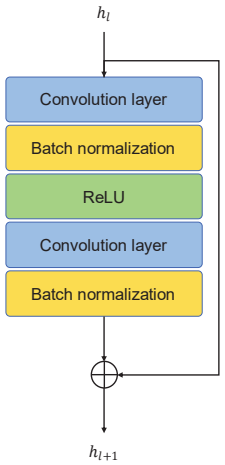

Fig. 2 ResBlock

A ResBlock consists of convolution layers, a ReLU activation function, batch normalization layers and a skip connection, where h_l is the output for l-th layer.

3 Experiment and result

The experimental setup for the proposed system is shown in Fig. 3, where the coded mask with the width of 1.6 mm was placed in front of the sensor at image distance $d_i = 1.5$ cm, and the distance between the mask and the monitor $d_o = 60$ cm. We employed Flickr2K dataset for training and DIV2K dataset for testing [18], [19], where each scene was rendered by a commercial display. The data augmentation with horizontal and vertical flipping was applied on training set,

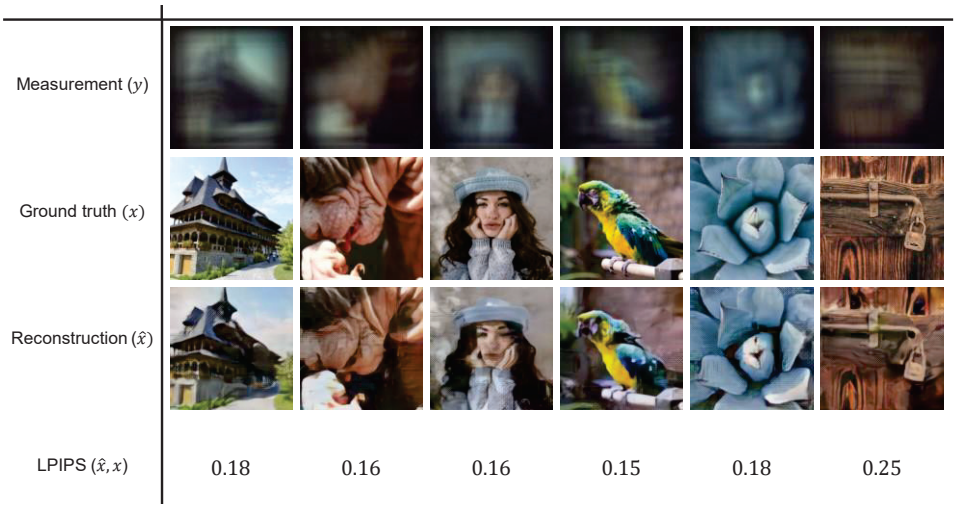

Fig. 4 Reconstruction samples

The first row indicates sensor measurements, the second row indicates ground truth images, the third row indicates reconstructions by our model and the fourth row indicates LPIPS scores.

thus leading to 7950 image pairs. A CMOS sensor (FLIR BFS-U3-70S7C-C) was employed to capture the intermediate images, afterward each measurement was cropped and resize to the resolution 128×128 for the model input. The network was trained to minimize feature loss at the layer "relu2_2" using Adam optimizer for 39800 iterations (100 epochs) [20], with the batch size 20 and 2×10^{-4} learning rate during training.

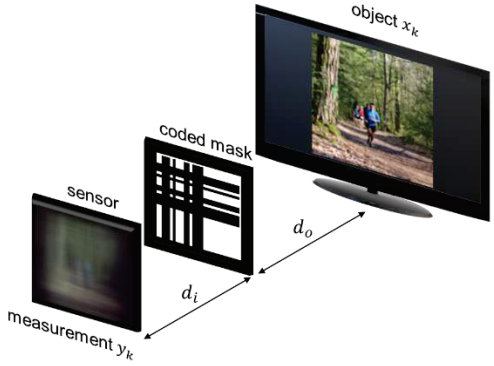

Fig. 3 Experiment setup

The coded mask was placed in front of the sensor with the image distance d_i, and the monitor for rendering scene was put at the object distance d_o from the coded mask.

Because we optimized the reconstructions via perceptual loss, conventional model-based metrics such as PSNR or SSIM were no longer suitable. Here we applied another metric, Learned Perceptual Image Patch Similarity (LPIPS), where LPIPS calculates the perceptual similarity between images [21]. Some reconstruction samples for testing data are shown in Fig. 4, where smaller LPIPS means high similarity through the human perception.

4 Conclusions

As the result, unlike the conventional coded mask scheme with necessity of estimated kernel function as the priors, we developed an end-to-end trained DNN through the perceptual loss function. Our model only minimized single perceptual loss without other regularization terms, which reduced the efforts for fine tuning many hyperparameters. The lensless configuration has an extreme compact form factor, which ease the constraints cause by the conventional lens-based imaging system accordingly.

Acknowledgement

This work was supported by the Ministry of Science and Technology (MOST) of Taiwan Government under contract MOST 111-2221-E-A49-059 -.

References

[1] E. Caroli, J. B. Stephen, G. D. Cocco, L. Natalucci, and A. Spizzichino, "Coded aperture imaging in X- and gamma-ray astronomy," *Space Science Reviews*, vol. 45, no. 3, pp. 349–403, 1987.

[2] 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*, vol. 3, no. 3, pp. 384–397, 2017.

[3] P. C. Hansen, J. G. Nagy, and D. P. O'Leary, *Deblurring Images: Matrices, Spectra, and Filtering*. 2006.

[4] S. S. Khan, V. Sundar, V. Boominathan, A. Veeraraghavan, and K. Mitra, "FlatNet: Towards Photorealistic Scene Reconstruction from Lensless Measurements," *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 1–1, 2020, doi: 10.1109/TPAMI.2020.3033882.

[5] K. Monakhova, J. Yurtsever, G. Kuo, N. Antipa, K.

Yanny, and L. Waller, “Learned reconstructions for practical mask-based lensless imaging,” *Opt. Express, OE*, vol. 27, no. 20, pp. 28075–28090, Sep. 2019, doi: 10.1364/OE.27.028075.

[6] X. Pan, X. Chen, S. Takeyama, and M. Yamaguchi, “Image reconstruction with transformer for mask-based lensless imaging,” *Opt. Lett., OL*, vol. 47, no. 7, pp. 1843–1846, Apr. 2022, doi: 10.1364/OL.455378.

[7] A. Sinha, J. Lee, S. Li, and G. Barbastathis, “Lensless computational imaging through deep learning,” *Optica, OPTICA*, vol. 4, no. 9, pp. 1117–1125, Sep. 2017, doi: 10.1364/OPTICA.4.001117.

[8] Y.-T. C. Lee, Y.-C. Fang, and C.-H. Tien, “Deep neural network for coded mask cryptographical imaging,” *Appl. Opt., AO*, vol. 60, no. 6, pp. 1686–1693, Feb. 2021, doi: 10.1364/AO.415120.

[9] Y. Yang *et al.*, “Transfer Learning in General Lensless Imaging through Scattering Media,” arXiv, arXiv:1912.12419, Dec. 2019. doi: 10.48550/arXiv.1912.12419.

[10] M. J. DeWeert and B. P. Farm, “Lensless coded-aperture imaging with separable Doubly-Toeplitz masks,” *Optical Engineering*, vol. 54, no. 2, pp. 23102–23102, 2015.

[11] 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*, 2015, pp. 234–241.

[12] D. Bae, J. Jung, N. Baek, and S. A. Lee, “Lensless Imaging with an End-to-End Deep Neural Network,” in *2020 IEEE International Conference on Consumer Electronics - Asia (ICCE-Asia)*, Nov. 2020, pp. 1–5. doi: 10.1109/ICCE-Asia49877.2020.9276865.

[13] C. Ledig *et al.*, “Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network,” arXiv, arXiv:1609.04802, May 2017. doi: 10.48550/arXiv.1609.04802.

[14] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, “Enhanced Deep Residual Networks for Single Image Super-Resolution,” *arXiv:1707.02921 [cs]*, Jul. 2017, [Online]. Available: http://arxiv.org/abs/1707.02921

[15] J. Johnson, A. Alahi, and L. Fei-Fei, “Perceptual Losses for Real-Time Style Transfer and Super-Resolution,” in *Computer Vision – ECCV 2016*, Cham, 2016, pp. 694–711. doi: 10.1007/978-3-319-46475-6_43.

[16] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” 2015.

[17] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, “Fast and Accurate Image Super-Resolution with Deep Laplacian Pyramid Networks,” *arXiv:1710.01992 [cs]*, Aug. 2018, [Online]. Available: http://arxiv.org/abs/1710.01992

[18] E. Agustsson and R. Timofte, “NTIRE 2017 Challenge on Single Image Super-Resolution: Dataset and Study,” Jul. 2017.

[19] R. Timofte *et al.*, “NTIRE 2017 Challenge on Single Image Super-Resolution: Methods and Results,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Jul. 2017, pp. 1110–1121. doi: 10.1109/CVPRW.2017.149.

[20] D. P. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” arXiv, arXiv:1412.6980, Jan. 2017. doi: 10.48550/arXiv.1412.6980.

[21] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, “The Unreasonable Effectiveness of Deep Features as a Perceptual Metric,” *arXiv:1801.03924 [cs]*, Apr. 2018, [Online]. Available: http://arxiv.org/abs/1801.03924