

To be published in Optics Letters:

Title: Image reconstruction with Transformer for mask-based lensless imaging

Authors: Xiuxi Pan,xiao chen,Saori Takeyama,Masahiro Yamaguchi

Accepted: 09 March 22

Posted 10 March 22

DOI: <https://doi.org/10.1364/OL.455378>

© 2022 Optica

OPTICA
PUBLISHING GROUP
Formerly OSA

Image reconstruction with Transformer for mask-based lensless imaging

XIUXI PAN*, XIAO CHEN, SAORI TAKEYAMA, AND MASAHIRO YAMAGUCHI

School of Engineering, Tokyo Institute of Technology, 4259-G2-28 Nagatsuta, Midori-ku, Yokohama, Kanagawa 226-8502, Japan

*Corresponding author: pan.x.aa@m.titech.ac.jp

Compiled March 9, 2022

A mask-based lensless camera optically encodes the scene with a thin mask and reconstructs the image afterward. The improvement of image reconstruction is one of the most important subjects in lensless imaging. Conventional model-based reconstruction approaches, which leverage knowledge of the physical system, are susceptible to imperfect system modeling. Reconstruction with pure data-driven deep neural network (DNN) avoids this limitation, thereby having potential to have better reconstruction quality. However, existing pure DNN reconstruction approaches for lensless imaging do not turn out a better result than model-based approaches. We reveal that the multiplexing property in lensless optics makes global features essential in understanding the optically encoded pattern. Whereas, all existing DNN reconstruction approaches apply fully convolutional networks (FCNs) which are not efficient in global feature reasoning. With this analysis, we in the first time propose a fully connected neural network with a Transformer for image reconstruction. The proposed architecture is better in global feature reasoning, hence enhancing the reconstruction. The superiority of the proposed architecture is verified by comparing with the model-based and FCN-based approaches in an optical experiment. © 2022 Optical Society of America

<http://dx.doi.org/10.1364/ao.XX.XXXXXX>

A mask-based lensless camera adopts a thin mask to optically encode the scene and records an encoded pattern on the image sensor. The image is afterward recovered from the encoded pattern with a reconstruction algorithm. Avoiding using lens components, the lensless camera can extremely reduce the thickness, weight and cost compared with the traditional lensed camera. The lensless camera is promising to be widely applied in internet of things where miniaturization is demanded. However, currently the lensless camera cannot produce images as high-quality as the lensed camera does. The improvement of image reconstruction remains a main subject in lensless imaging.

Model-based approaches reconstruct the image by leveraging knowledge of the physical system. Considering illuminating the mask-based lensless system with an ideal point light source, a

specific pattern is cast on the sensor. This pattern is determined by the physical system and is termed point spread function (PSF). Model-based approaches model mask-based lensless optics as the following linear system

$$\mathbf{i}_0 = \Phi \mathbf{o} + \mathbf{e}, \quad (1)$$

where \mathbf{i}_0 is the sensor measurement, \mathbf{o} is the object, Φ is the measurement matrix constructed by the PSF, and \mathbf{e} is noise. Φ is block Toeplitz if assuming that the PSF is shift-invariant. In the mask-based lensless optics, a sensor pixel measures multiplexed light from widely spread points in the scene. This property is known as multiplexing. The multiplexing property results in an ill-conditioned system which makes the reconstruction challenging.

Conventional iterative optimization approaches [1–5] reconstruct the image by iteratively minimizing a loss function [6–8]

$$\hat{\mathbf{o}} = \underset{\mathbf{o}}{\operatorname{argmin}} \|\mathbf{i}_0 - \Phi \mathbf{o}\|_{l_2} + \tau \Psi(\mathbf{o}), \quad (2)$$

where Ψ denotes a regularizer based on image sparsity prior and τ is a tuning parameter. Eq. 2 can be unrolled to a stack of neural network layers to accelerate the reconstruction process [9, 10]. Moiré decoding approaches [11–13] decode the object \mathbf{o} from Eq. 1 by using specially designed Fresnel zone apertures with different phases as the mask and decoding with the fringe scanning [14].

However, the physical model used in model-based approaches is too idealization and simplification. Firstly, the basic assumption of shift-invariance for PSF is broken when there is incident light with high angle. Secondly, an ideal PSF can hardly be acquired in practice. Additionally, the sparsity prior used in iterative optimization approaches does not work well for all scenes. The imperfect modeling results in errors and artifacts.

Though perceptual quality enhancement methods including the ones using DNN [9, 10] can be applied after reconstruction, they can only partially correct the errors and artifacts brought by imperfect manual modeling. Designing pure data-driven deep neural network (DNN) for image reconstruction in mask-based lensless imaging is attractive. This is because DNN can achieve more complex modeling than the manual one and it can learn stronger scene priors than the single sparsity prior. Existing pure DNNs for image reconstruction are all based on fully convolutional networks (FCNs) [15–17]. However, FCNs

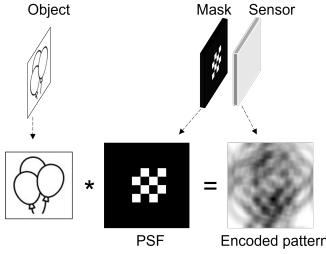


Fig. 1. A simulation example of the forward model in mask-based lensless imaging. Local information in the object is transformed into overlapping global information in the optically encoded pattern.

have limitation as the multiplexing property is not considered. The multiplexing property transforms local information in the scene to overlapping global information in the image sensor [18, 19], as a simulation example illustrated in Fig.1 where the object is a two-dimensional (2D) object and the mask is a 4×4 binary amplitude mask. FCNs are built on stacking small-kernel convolutions (such as 3×3 or 7×7) in a deep network. Such design is efficient in reasoning local features such as edges, ridges and corners which focus on a small group of pixels, but its access to global features is limited because it is only done in deeper layers. Meanwhile, long range dependencies are essential in the optically encoded pattern considering the multiplexing property of the mask-based lensless optics. Based on this observation, we argue that FCNs may not be the best choice here even they are currently dominate in most computer vision tasks.

Based on above discussion, we propose a fully connected neural network with a Transformer for pure data-driven image reconstruction, dubbed as "lensless imaging Transformer". The Transformer is free of the inductive bias of locality and considers global context straightly by encoding with self-attention mechanism, thus enhancing global feature reasoning [20–22]. The diagrammatic overview of proposed neural network is illustrated in Fig.2. Standardization, which rescales the data to sure the mean and the standard deviation to be 0 and 1 respectively, is performed for the input encoded pattern. Considering that the reconstruction should be identical for each color channel, the network receives and outputs one channel. For a colorful image, R, G, B channels are reconstructed separately and then concatenated. We follow SegFormer [23] to lay 4 encoder blocks to learn hierarchical feature representations from the input. Hierarchical feature representations are then upsampled to the target size, concatenated and decoded.

Each encoder block consists of the overlapped patchify and the Transformer. The overlapped patchify reduces size and increases depth of the input while the Transformer maintains the input shape. The overlapped patchify applies K -kernel size, S -stride, P -padding and C -number convolutions to extract overlapping embedded patches. K , S , P and C for each block are listed in Table 1. The overlapped patchify can preserve local

Table 1. Variants in the encoder.

Block No.	Patchify				Transformer	
	K	S	P	C	L	N
1	7	4	3	64	2	1
2	3	2	1	128	2	2
3	3	2	1	256	2	4
4	3	2	1	512	2	8

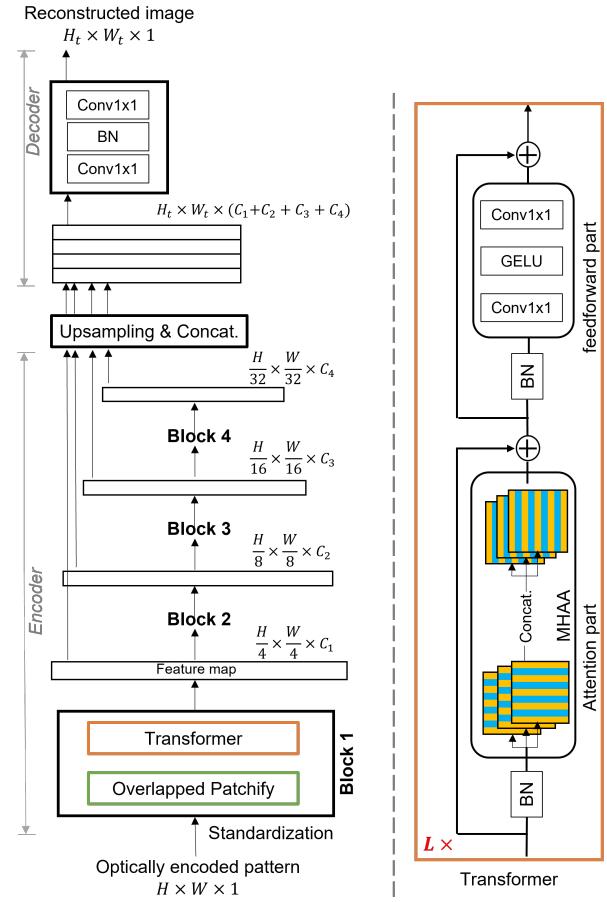


Fig. 2. The proposed lensless imaging Transformer.

continuity around patches which is beneficial for learning pixel-level transformation. The computation bottleneck lays in the Transformer. In order to be applicable for large input size, computation reduction in the Transformer is needed. We simplify the Transformer by replacing traditional self-attention with the axial-attentions [24]. The proposed Transformer has L layers. L for each block is listed in Table 1. Each layer consists of the attention part and the feedforward part. Residual connections are applied to both parts. The attention part includes batch normalization (BN) and multiheaded axial-attentions (MHA), as shown in the right part of Fig.2. The feedforward part includes BN, stride-(1×1), kernel-(1×1), number- d_{en} convolutions, Gaussian error linear unit (GELU) and stride-(1×1), kernel-(1×1), number- D convolutions in order. D is set as same as the input depth and d_{en} is set as $4D$.

In MHA, N axial-attentions are computed in parallel and finally concatenated. N for each block is listed in Table 1. Axial-attentions consequently performs self-attention on the feature map width axis and height axis, formally:

$$y_o = \sum_{p \in N_{1 \times m(o)}} \text{softmax}_p(q_o^T k_p + q_o^T r_{p-o}^q + k_p^T r_{p-o}^k)(v_p + r_{p-o}^v). \quad (3)$$

y_o is the output at position o . Queries $\mathbf{q} = \mathbf{W}_Q \mathbf{i}$, keys $\mathbf{k} = \mathbf{W}_K \mathbf{i}$, and values $\mathbf{v} = \mathbf{W}_V \mathbf{i}$ are linear projections of the input $\mathbf{i} \in \mathbb{R}^{H_i \times W_i \times C_i}$, where $\mathbf{W}_Q \in \mathbb{R}^{\frac{C_i}{2N} \times C_i}$, $\mathbf{W}_K \in \mathbb{R}^{\frac{C_i}{2N} \times C_i}$ and $\mathbf{W}_V \in \mathbb{R}^{\frac{C_i}{N} \times C_i}$ are learnable matrices. $r_{p-o}^q, r_{p-o}^k, r_{p-o}^v \in \mathbb{R}^{m \times m}$ are learn-

able parameters, measuring the compatibility from position p to o in queries, keys and values. The softmax_p signifies a softmax function applied to all possible p positions. $N_{1 \times m}(o)$ defines the row or the column where position o exists. $m = W_i$ for the width-axial attention, and $m = H_i$ for the height-axial attention. Different from traditional self-attention which processes the data as a 1D sequence, axial-attentions regards the data as 2D graphics. Taking advantage of geometry construction information, axial-attentions reduces the computational complexity by processing pixels in width and height axis separately. Replacing self-attention with axial-attentions reduces the computational complexity from $\mathcal{O}(H_i^2 W_i^2)$ to $\mathcal{O}(H_i W_i m)$.

The decoder is a feedforward network including stride-(1 \times 1), kernel-(1 \times 1), number- d_{de} convolutions, BN and stride-(1 \times 1), kernel-(1 \times 1), number-1 convolution in order. d_{de} is 256.

We compare reconstruction performance among the proposed lensless imaging Transformer, the traditional pure data-driven deep neural network FCN, and the model-based approach. For FCN, U-net [25] is implemented. U-net architecture is outlined in Table 2 where "up()" and "," represent bilinear up-sampling and concatenation, respectively. Each encode is the stride-(3 \times 3), kernel-(1 \times 1) convolution followed by BN, ReLu and the stride-(2 \times 2), kernel-(2 \times 2) pooling. Each decode is the stride-(3 \times 3), kernel-(1 \times 1) convolution followed by BN and ReLu. For the model-based approach, we choose the iterative optimization (Eq.2) which is considered as the most representative one. The iterative optimization employs alternating direction method of multipliers (ADMM) [8], the used regularization is total-variation [26] and the tuning parameter τ is set as 10^{-4} .

For optical experiment, a mask-based lensless camera is assembled, shown as Fig.3(a). The mask is placed 2.5 mm in front of the image sensor. Since we aim at evaluating reconstruction algorithms, a pseudorandom binary amplitude mask without any particular optimization is chosen as the most basic one. The proposed approach will be also effective for a phase only mask in which the multiplexing property is also significant. The mask is fabricated by chromium deposition in a synthetic-silica plate, with aperture size of $40 \times 40 \mu\text{m}$. The sensor is a 6.41 megapixels CMOS (Sony IMX178) with $2.4 \times 2.4 \mu\text{m}$ pixel pitch, and 12 bit color depth. The PSF is captured by illuminating the mask with a point LED with the diameter of 1 mm placed 15 cm away. The PSF is only used in the model-based approach.

The optically encoded pattern dataset is collected by shooting while displaying images on an LCD screen placed around 15 cm ahead the lensless camera. The original image dataset consists of MirFlickr dataset [27] (25,000 RGB images) and a part of ILSVRC-2012 ImageNet dataset [28] (35,000 RGB images are selected). 1,000 images are selected for validation and others for training. All displayed images are resized to 500×500 pixels, and captured encoded patterns occupy within 1600×1600 pixels on the sensor.

In the lensless imaging Transformer, the input and output size are set as 1600×1600 and 500×500 pixels respectively. Since the U-net requires input size and output size being the same and

the encoded pattern is sensitive to color depth precision, we set both input and output size of the U-net as 1600×1600 pixels. In training, the input is the captured encoded pattern. Loss is the mean squared error between the output and the corresponding original image which is resized to 500×500 pixels for the lensless imaging Transformer and 1600×1600 pixels for the U-net. Training of both networks are implemented on a machine equipped with an Intel Xeon E5-2698 v4 CPU (2.2 GHz), two NVIDIA TESLA V100 GPUs (2×32 GB), Python 3.6.5, Pytorch 1.7.1. The optimizer employs Adam [29] with $\beta_1=0.9$, $\beta_2=0.999$. Weight decay is 0.1. With a batch size of 4, learning rate of 6×10^{-5} , learning rate warmup of 15,000 steps and cosine learning rate decay, the lensless imaging Transformer and U-net converge within 250,000 steps and 150,000 steps respectively.

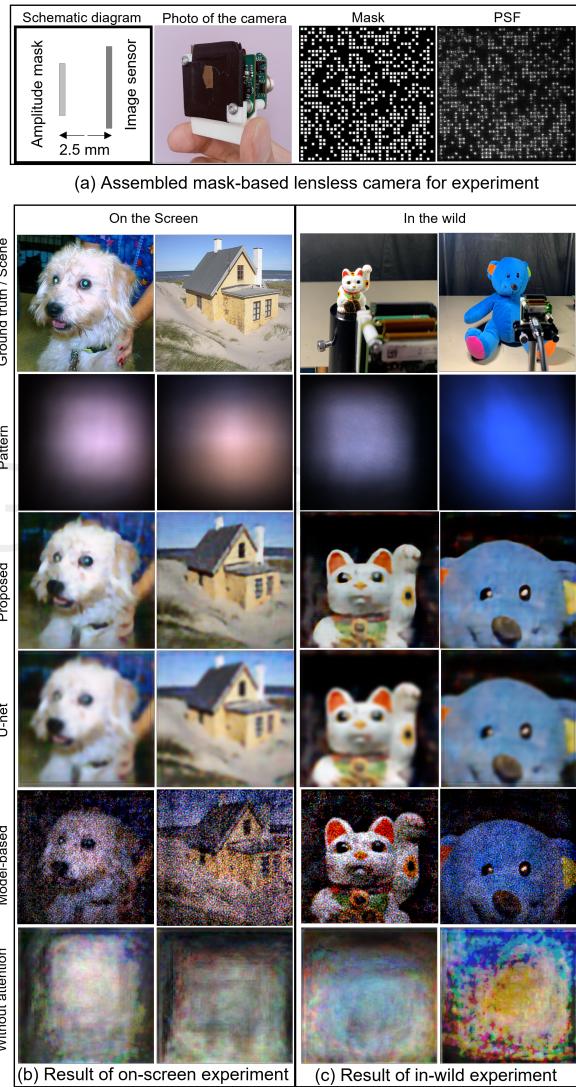


Fig. 3. Experimental setting and result. (a) shows the assembled mask-based lensless camera. (b) is the result of image-on-screen experiment. (c) is the result of object-in-wild experiment. Both the screen and objects are placed around 15 cm ahead of the camera. The proposed approach produces the most visually appealing images. The U-net [25] reconstructs images lacking details. The model-based approach is iterative optimization, employing ADMM [8] and total-variation [26]. It reconstructs images with evident streaky artifacts and noises, and color distortion in some areas. The result of the proposed network without attention part is listed in the last row.

Table 2. U-net architecture.

encode	cha. in/out	input	decode	cha. in/out	input
enc1	1/64	input	dec5	1024/512	up(conv1).enc5
enc2	64/128	enc1	dec4	512/256	up(dec5).enc4
enc3	128/256	enc2	dec3	256/128	up(dec4).enc3
enc4	256/512	enc3	dec2	128/64	up(dec3).enc2
enc5	512/1024	enc4	dec1	64/64	up(dec2).enc1
conv1	1024/1024	enc5	conv2	64/1	dec1

Fig.3(b), (c) show several samples for comparison. Fig.3(b) is the result of experiment with on-screen images. Peak signal to noise ratio (PSNR) and structural similarity index measure (SSIM) are calculated. PSNR/SSIM of the downy cat image is 29.18 dB/0.54 for the proposed approach and 28.76 dB/0.34 for the U-net. PSNR/SSIM of the house image is 28.40 dB/0.56 for the proposed approach and 27.61 dB/0.21 for the U-net. Fig.3(c) is the result of experiment with objects in wild. Both the screen and the objects are placed around 15 cm ahead of the camera. The iterative optimization approach reconstructs a 1600×1600 pixels encoded pattern into a reconstructed image with the same size. The ADMM runs 50 iterations ensuring convergence.

The result shows that the U-net reconstructs images with lacking details compared with the reconstructed images by the proposed approach. The model-based approach can reconstruct images with similar resolution as the proposed approach does. But model-based approach is unable to reconstruct correct color in some areas and it introduces evident streaky artifacts and noises. These factors make the reconstructed images have low visually quality. Overall, the proposed approach produces the most visually appealing images.

Ablation study on the Transformer component in the proposed network is also performed. By deleting the attention part in the Transformer component, the model with the same training conditions fails to reconstruct meaningful images, as shown in the last row of Fig.3. It verifies the main contribution of the Transformer in the proposed network.

This letter studies the reconstruction algorithm which is the most fundamental part in the lensless imaging. We have proposed the Transformer-based pure data-driven algorithm, which proven superiority in reconstruction quality compared with the FCN and the model-based algorithms. There are two points we recommend to consider when applying our proposed reconstruction algorithm. Firstly, pure data-driven approaches, including the proposed one, is restricted by the training data distribution. Learned image priors greatly rely on the fed training data. The domain difference between training data and application data should not be too large. Secondly, optimization of mask design and adding perceptual quality enhancement or denoise algorithm afterwards could further improve the image quality.

Though the proposed algorithm is designed for mask-based lensless imaging, it could potentially extended to more applications such as image reconstruction in mask-free lensless imaging [30], through-scattering imaging [31] and under-screen imaging.

Funding. JST SPRING JPMJSP2106; Super Smart Society Leadership Scholarship for Doctoral Students.

Acknowledgment. We are very grateful to Prof. Tomoya Nakamura for offering some experimental equipment.

Disclosures. The authors declare no conflicts of interest.

Data availability. Data and code underlying the results presented in this paper are available in Ref.[32].

REFERENCES

1. D. G. Stork and P. R. Gill, Int. J. on Adv. Syst. Meas. **7**, 201 (2014).
2. M. J. DeWeert and B. P. Farm, Opt. Eng. **9109**, 91090Q (2014).
3. M. S. Asif, A. Ayremiou, A. Sankaranarayanan, A. Veeraraghavan, and R. G. Baraniuk, IEEE Transactions on Comput. Imaging **3**, 384 (2016).
4. S. K. Sahoo, D. Tang, and C. Dang, Optica **4**, 1209 (2017).
5. N. Antipa, G. Kuo, R. Heckel, B. Mildenhall, E. Bostan, R. Ng, and L. Waller, Optica **5**, 1 (2018).
6. J. M. Bioucas-Dias and M. A. Figueiredo, IEEE Transactions on Image Process. **16**, 2992 (2007).
7. A. Beck and M. Teboulle, IEEE Transactions on Image Process. **18**, 2419 (2009).
8. S. Boyd, N. Parikh, and E. Chu, *Distributed optimization and statistical learning via the alternating direction method of multipliers* (Now Publishers Inc, 2011).
9. K. Monakhova, J. Yurtsever, G. Kuo, N. Antipa, K. Yanny, and L. Waller, Opt. express **27**, 28075 (2019).
10. S. S. Khan, V. Adarsh, V. Boominathan, J. Tan, A. Veeraraghavan, and K. Mitra, "Towards photorealistic reconstruction of highly multiplexed lensless images," in *Proceedings of the IEEE International Conference on Computer Vision*, (2019), pp. 7860–7869.
11. T. Shimano, Y. Nakamura, K. Tajima, M. Sao, and T. Hoshizawa, Appl. Opt. **57**, 2841 (2018).
12. T. Nakamura, T. Watanabe, S. Igarashi, X. Chen, K. Tajima, K. Yamaguchi, T. Shimano, and M. Yamaguchi, Opt. Express **28**, 39137 (2020).
13. X. Chen, T. Nakamura, X. Pan, K. Tajima, K. Yamaguchi, T. Shimano, and M. Yamaguchi, "Resolution improvement in fza lens-less camera by synthesizing images captured with different mask-sensor distances," in *2021 IEEE International Conference on Image Processing (ICIP)*, (IEEE, 2021), pp. 2808–2812.
14. D. Malacara, *Optical shop testing*, vol. 59 (John Wiley & Sons, 2007).
15. Y. Li, Y. Xue, and L. Tian, Optica **5**, 1181 (2018).
16. S. Li, M. Deng, J. Lee, A. Sinha, and G. Barbastathis, Optica **5**, 803 (2018).
17. R. Horisaki, Y. Okamoto, and J. Tanida, Opt. Lett. **45**, 3131 (2020).
18. X. Pan, T. Nakamura, X. Chen, and M. Yamaguchi, Opt. Express **29**, 9758 (2021).
19. X. Pan, X. Chen, T. Nakamura, and M. Yamaguchi, Opt. Express **29**, 37962 (2021).
20. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, (2017), pp. 5998–6008.
21. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, arXiv preprint arXiv:1810.04805 (2018).
22. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, arXiv preprint arXiv:2010.11929 (2020).
23. E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, arXiv preprint arXiv:2105.15203 (2021).
24. H. Wang, Y. Zhu, B. Green, H. Adam, A. Yuille, and L.-C. Chen, "Axial-deeplab: Stand-alone axial-attention for panoptic segmentation," in *European Conference on Computer Vision*, (Springer, 2020), pp. 108–126.
25. 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*, (Springer, 2015), pp. 234–241.
26. L. I. Rudin, S. Osher, and E. Fatemi, Phys. D: nonlinear phenomena **60**, 259 (1992).
27. M. J. Huiskes and M. S. Lew, "The mir flickr retrieval evaluation," in *Proceedings of the 1st ACM international conference on Multimedia information retrieval*, (2008), pp. 39–43.
28. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, (Ieee, 2009), pp. 248–255.
29. D. P. Kingma and J. Ba, arXiv preprint arXiv:1412.6980 (2014).
30. G. Kim, K. Isaacson, R. Palmer, and R. Menon, Appl. optics **56**, 6450 (2017).
31. A. P. Mosk, A. Lagendijk, G. Lerosey, and M. Fink, Nat. photonics **6**, 283 (2012).
32. X. Pan, "Lensless imaging transformer repository," https://github.com/BobPXX/Lensless_Imaging_Transformer. Accessed: 2021-12-17.

FULL REFERENCES

1. D. G. Stork and P. R. Gill, "Optical, Mathematical, and Computational Foundations of Lensless Ultra-Miniature Diffractive Imagers and Sensors," *Int. J. on Adv. Syst. Meas.* **7**, 201–208 (2014).
2. M. J. DeWeert and B. P. Farm, "Lensless coded aperture imaging with separable doubly Toeplitz masks," *Opt. Eng.* **9109**, 91090Q (2014).
3. 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 Comput. Imaging* **3**, 384–397 (2016).
4. S. K. Sahoo, D. Tang, and C. Dang, "Single-shot multispectral imaging with a monochromatic camera," *Optica* **4**, 1209–1213 (2017).
5. N. Antipa, G. Kuo, R. Heckel, B. Mildenhall, E. Bostan, R. Ng, and L. Waller, "Diffusercam: lensless single-exposure 3d imaging," *Optica* **5**, 1–9 (2018).
6. J. M. Bioucas-Dias and M. A. Figueiredo, "A new twist: Two-step iterative shrinkage/thresholding algorithms for image restoration," *IEEE Transactions on Image Process.* **16**, 2992–3004 (2007).
7. A. Beck and M. Teboulle, "Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems," *IEEE Transactions on Image Process.* **18**, 2419–2434 (2009).
8. S. Boyd, N. Parikh, and E. Chu, *Distributed optimization and statistical learning via the alternating direction method of multipliers* (Now Publishers Inc, 2011).
9. K. Monakhova, J. Yurtsever, G. Kuo, N. Antipa, K. Yanny, and L. Waller, "Learned reconstructions for practical mask-based lensless imaging," *Opt. express* **27**, 28075–28090 (2019).
10. S. S. Khan, V. Adarsh, V. Boominathan, J. Tan, A. Veeraraghavan, and K. Mitra, "Towards photorealistic reconstruction of highly multiplexed lensless images," in *Proceedings of the IEEE International Conference on Computer Vision*, (2019), pp. 7860–7869.
11. T. Shimano, Y. Nakamura, K. Tajima, M. Sao, and T. Hoshizawa, "Lensless light-field imaging with Fresnel zone aperture: quasi-coherent coding," *Appl. Opt.* **57**, 2841–2850 (2018).
12. T. Nakamura, T. Watanabe, S. Igarashi, X. Chen, K. Tajima, K. Yamaguchi, T. Shimano, and M. Yamaguchi, "Superresolved image reconstruction in fza lensless camera by color-channel synthesis," *Opt. Express* **28**, 39137–39155 (2020).
13. X. Chen, T. Nakamura, X. Pan, K. Tajima, K. Yamaguchi, T. Shimano, and M. Yamaguchi, "Resolution improvement in fza lens-less camera by synthesizing images captured with different mask-sensor distances," in *2021 IEEE International Conference on Image Processing (ICIP)*, (IEEE, 2021), pp. 2808–2812.
14. D. Malacara, *Optical shop testing*, vol. 59 (John Wiley & Sons, 2007).
15. Y. Li, Y. Xue, and L. Tian, "Deep speckle correlation: a deep learning approach toward scalable imaging through scattering media," *Optica* **5**, 1181–1190 (2018).
16. S. Li, M. Deng, J. Lee, A. Sinha, and G. Barbastathis, "Imaging through glass diffusers using densely connected convolutional networks," *Optica* **5**, 803–813 (2018).
17. R. Horisaki, Y. Okamoto, and J. Tanida, "Deeply coded aperture for lensless imaging," *Opt. Lett.* **45**, 3131–3134 (2020).
18. X. Pan, T. Nakamura, X. Chen, and M. Yamaguchi, "Lensless inference camera: incoherent object recognition through a thin mask with lbp map generation," *Opt. Express* **29**, 9758–9771 (2021).
19. X. Pan, X. Chen, T. Nakamura, and M. Yamaguchi, "Incoherent reconstruction-free object recognition with mask-based lensless optics and the transformer," *Opt. Express* **29**, 37962–37978 (2021).
20. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, (2017), pp. 5998–6008.
21. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805 (2018).
22. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929 (2020).
23. E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," arXiv preprint arXiv:2105.15203 (2021).
24. H. Wang, Y. Zhu, B. Green, H. Adam, A. Yuille, and L.-C. Chen, "Axial-deeplab: Stand-alone axial-attention for panoptic segmentation," in *European Conference on Computer Vision*, (Springer, 2020), pp. 108–126.
25. 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*, (Springer, 2015), pp. 234–241.
26. L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Phys. D: nonlinear phenomena* **60**, 259–268 (1992).
27. M. J. Huiskes and M. S. Lew, "The mir flickr retrieval evaluation," in *Proceedings of the 1st ACM international conference on Multimedia information retrieval*, (2008), pp. 39–43.
28. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, (Ieee, 2009), pp. 248–255.
29. D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980 (2014).
30. G. Kim, K. Isaacson, R. Palmer, and R. Menon, "Lensless photography with only an image sensor," *Appl. optics* **56**, 6450–6456 (2017).
31. A. P. Mosk, A. Lagendijk, G. Lerosey, and M. Fink, "Controlling waves in space and time for imaging and focusing in complex media," *Nat. photonics* **6**, 283–292 (2012).
32. X. Pan, "Lensless imaging transformer repository," https://github.com/BobPXX/Lensless_Imaging_Transformer. Accessed: 2021-12-17.