Cross-Net: Joint In-Line Holographic Image Reconstruction and Refocusing

Haitao Zhou[1], Mazen Mel[2*], Paul Springer[3], and Alexander Gatto[3]

[1] École polytechnique fédérale de Lausanne, Lausanne, Switzerland.
haitao.zhou@epfl.ch,
[2] University of Padova, Department of Information Engineering, Padova, Italy.
mazen.mel@dei.unipd.it,
[3] Sony Semiconductor Solutions Europe, Stuttgart Laboratory 1, Germany.
{paul.springer;alexander.gatto}@sony.com

Abstract. Holography enables quantitative phase sensing through light wave interference, allowing interesting imaging modalities applied in computational microscopy where phase shift is used as a way to generate image contrast. Recovering both amplitude and phase distributions of a light field in in-line holography is an ill-posed problem that requires post-processing and further computations. This paper introduces a novel learning-based model dubbed Cross-Net, which can extract latent representations used for patch-based image reconstruction and regression tasks, thus enabling joint holographic image reconstruction and computational refocusing. Compared to the standard Vision Transformer (ViT), Cross-Net uses a dedicated convolutional block to substitute Multi-Head Self-Attention as an inter-spatial processing unit and adapts Patch Embedding to generate meaningful latent representations suitable for the tasks at hand. The proposed model is lighter and exhibits a significant performance gain in the joint tasks of computational refocusing and holographic image reconstruction compared to ViT and other state-of-the-art approaches. We further introduce the use of large scale synthetic microscopic data exploiting generative models to train our network.

Keywords: In-line holography, computational microscopy, image reconstruction, image refocusing.

1 Introduction and Prior Art

Light fields are defined by an amplitude and a phase distributions, holography enables such information to be recorded and later reconstructed. It was first introduced by D. Gabor [7] who proposed an in-line imaging setup where the object and reference light fields travel on the same optical axis. Then computational phase retrieval is needed to recover the latent field, allowing the full potential of holography to be used in multiple applications such as medical imaging [9] and microscopy [15]. The biggest challenge hindering the wide spread application

*Corresponding author: mazen.mel@dei.unipd.it

of in-line holography is the twin image problem which is caused by the object wave's complex conjugate [14] recorded as a by-product by the image sensor and causes undesirable blur artifacts and deteriorates the quality of the reconstructed image. Adjustments on hardware and software parts can be used [16] to get rid of the twin image. In the former case, a typical solution is off-axis holography; By choosing appropriately the angle between the reference and object propagating waves, a displacement between the two images in Fourier space can be observed thus a simple filtering step in the Fourier domain can suppress the twin image [4]. However, in this work we are interested in in-line holography which can significantly reduce the hardware complexity resulting in compact imaging devices. The twin-image influence in this case can only be suppressed, to some degree, via computational approaches. Iterative error-reduction methods stand out as the main family of solvers used to tackle the reconstruction problem in in-line holography. Usually the field is computationally propagated in-between the detector and sample planes, while support constraints are enforced so that it gradually converges to the desired solution. Such constraints include positive absorption of the field in the object plane [13], filtering mask [5], denoising priors [18], and sparsity [8]. Learning-based approaches for holographic image reconstruction have also been investigated. Previous research has shown the success of several popular neural network architectures on such task among which are convolutional neural networks (CNNs) [21], recurrent neural networks (RNNs) [12], and generative adversarial models (GANs) [3]. In this paper, a novel learning-based approach Cross-Net is proposed that can perform joint computational refocusing and holographic image reconstruction; The network takes as input a hologram captured at an unknown object-detector distance z, and is then tasked to reconstruct both phase and amplitude distributions of the latent complex transmission field and estimate the distance z at which the hologram was recorded. Cross-Net is able to extract meaningful latent representations for patch-based image reconstruction and regression tasks than a standard Vision Transformer (ViT) [10] all while being significantly lighter. Our proposed model can also surpass state-of-the-art approaches including HRNet [17], Dense-U-Net [22], Fourier Imager Network (FIN) [1] and its subsequent variant Enhanced Fourier Imager Network (eFIN)[2] which are all designed for the task of holographic image reconstruction, notice also that eFIN performs refocusing in a joint framework. HRNet uses a deep residual network as an encoder and simple periodic shuffling. Dense-U-Net is an upgraded version of U-net, it adds specially designed dense blocks to the traditional U-net. The FIN architecture is based on spatial Fourier transform modules that process the spatial frequencies of its inputs using learnable filters and a global receptive field. Compared to FIN, eFIN uses a shallow U-net to extract filters of spatial frequencies and features needed for refocusing.

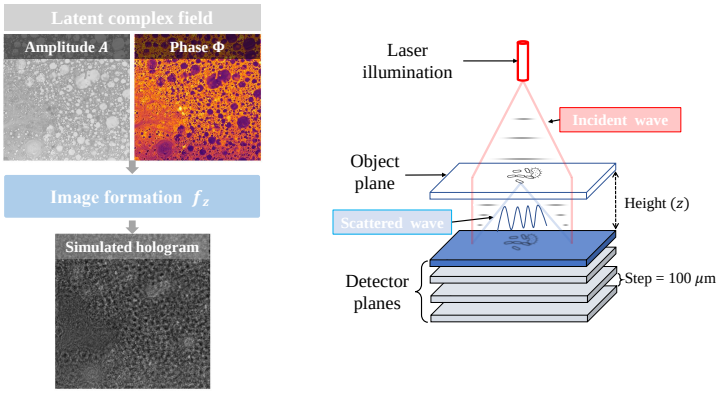

Fig. 1. Lens-free in-line holographic setup: Holographic measurement simulation from a latent complex field distribution (left). Schematics of our DIHM (right).

2 Methodology

In this section, the image formation model underlying in-line holography is discussed followed by the proposed model for joint complex field Reconstruction and refocusing.

2.1 Image Formation Model

The schematics in Fig. 1 illustrate the setup of a lens-free Digital In-line Holographic Microscope (DIHM). The goal is to recover a latent complex transmission field originating in the object plane with an amplitude A and a phase distribution Φ, $\mathbf{x} = A \cdot e^{j\Phi} \in \mathbb{C}^{h \times w}$, sampled at high resolution with spatial dimensions $h \times w$, as well as the actual refocusing distance or sensor/sample distance $z_i,\ i = 1, ..., N$ from a stack of input holographic measurements which can be simulated using the following forward model:

$$f_{z_i}(\mathbf{x}) = |P_{z_i}\mathbf{x}|^2 \tag{1}$$

Where $f_{\mathbf{z}} : \mathbb{C}^{hw} \mapsto \mathbb{R}^{hw}$ is the forward in-line holographic image formation model. The latent field $\mathbf{x}$ is first propagated to the detector plane using the complex near-field Fresnel propagation kernel P_{z_i} [14]. The real-valued hologram is obtained by calculating the square modulus of $P_{z_i}\mathbf{x}$. This image formation model is non-linear and encompasses different variety of microscopic specimens: the object of interest is assumed to have both absorption and phase shift properties even though absorption can sometimes be too weak in the case of thin and transparent cells. In order to incorporate measurement noise, sensor read and shot noise sources are simulated using the model in [6]:

$$\sigma(p) = \sqrt{\sigma_s^2 \cdot y(p) + \sigma_r^2} \tag{2}$$

Where σ is the pixel-dependent standard deviation of the noise level at pixel location p, σ_s^2 and σ_r^2 are the variances of shot and read noises, respectively, and $y(p)$ is the clean input pixel value.

2.2 Cross-Net: Learning for Joint Image Reconstruction and Refocusing

We propose a ViT-like architecture tailored for joint holographic image reconstruction and refocusing.

The workflow of the classic ViT [10] is shown in Fig. 2-(a-top); The output of patch embedding is fed to a Multi-Head Self-Attention (MHSA) layer that acts as an *inter-spatial* processing unit and then to a MLP that acts as an *inter-channel* processing unit, this module is usually repeated $\times l$ times. Finally, a simple decoder containing linear mapping and reshaping layers that could be used for reconstruction tasks or alternatively pooling and linear mapping that can be used for regression tasks. Which allows the model to not only reconstruct image data but also predict the refocusing distance.

Directly using ViT as is in holographic image reconstruction and refocusing poses two main challenges: First, the risk of over-fitting and high computational resources required to train the full model where the bulk of trainable parameters are those of the attention layer. Second, the loss of information in the Patch Embedding as the consistency of global image features spanning large areas might be affected by small patch sizes. Cross-Net is proposed to solve the

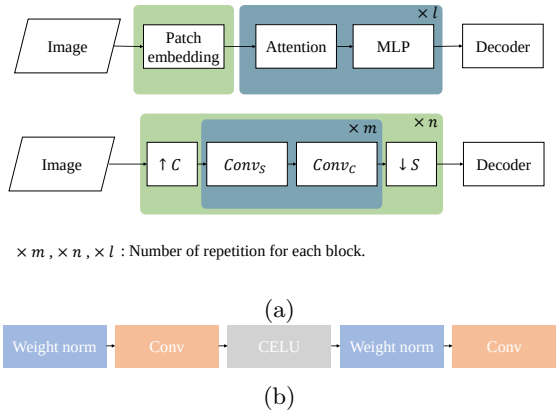

Fig. 2. (a): Original ViT workflow (top) and Cross-Net's (bottom). **(b):** The convolution block used to achieve $\uparrow C$, $Conv_S$, and $Conv_C$ units shown in (a-bottom).

aforementioned shortcomings of ViT. The model architecture is shown in Fig. 2-(a-bottom) and the building block for the different convolution layers is also depicted in Fig. 2-(b). MHSA is substituted with spatial convolution operation

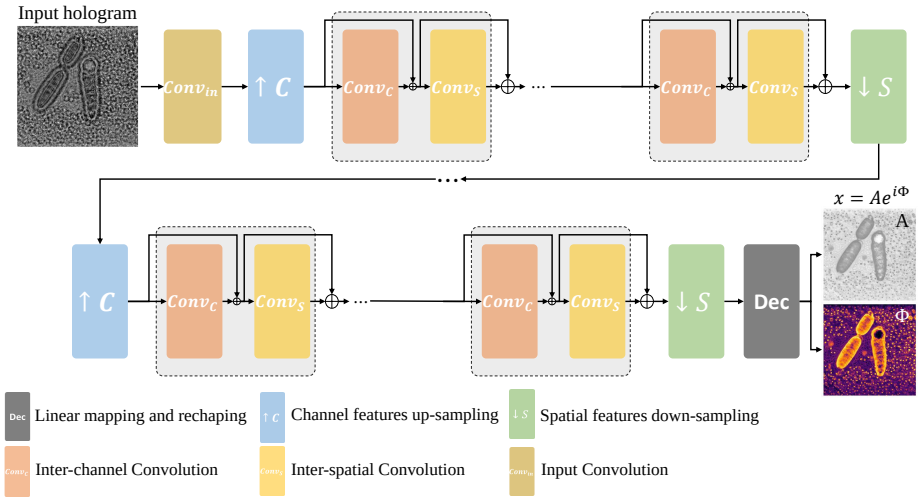

Fig. 3. Architecture of the proposed field reconstruction branch used in Cross-Net, regression branch uses the same architecture except for the decoder.

$Conv_S$ that acts as an *inter-spatial* layer with large kernel size, e.g., 17 for an image size of 128×128 pixels, worth noting that such an operation is carried out only on the spatial dimensions ignoring the depth dimension of feature maps. In this way $Conv_S$ tackles the over-fitting problem as the number of learned parameters can be significantly reduced with large kernel sizes compared to MHSA. Its effectiveness in substituting MHSA as an *inter-spatial* processing unit has been demonstrated in [20] for the simple task of image classification. $Conv_C$ is a channel-wise convolution operation with a kernel size of 1 and a similar depth to that of the input feature maps thus it captures inter-channel correlations. It has a similar function as a normal MLP. Patch embedding is decomposed into a series of operations that up-sample channel features $\uparrow C$ and down-sample spatial features $\downarrow S$. The up-sampling layer $\uparrow C$ uses the same settings as $Conv_C$ except that the number of output channels of the second convolution layer may be different from the number of input channels. The down-sampling $\downarrow S$ is a max-pooling layer, such decomposition addresses the issue of information loss in standard Patch Embedding with large patch sizes. Finally, in Cross-Net a simple decoder similar to that of ViT is used and can be adapted for reconstruction and regression tasks. Thus, the overall architecture has two main branches in order to perform the two tasks simultaneously. In any given iteration, simulated holograms of the same object captured at different heights are fed to Cross-Net, the reconstruction branch always reconstructs the latent complex field, while the regression branch estimates a different refocusing distance for each input.

3 Data and Training Details

We exploit state-of-the-art generative models to generate large scale synthetic data to train Cross-Net. A diffusion model [19] fine-tuned on microscopic images[†] was used to generate such data by prompting the model with the `"microscopic"` keyword. The diffusion model is able to generate training samples that have features similar to that of real microscopic samples as shown in Fig. 4. The

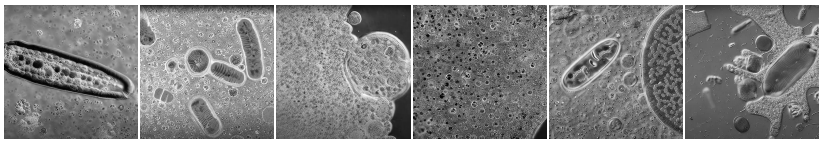

Fig. 4. Sample training images generated by stable diffusion model [19].

amplitude and phase distributions are obtained from a normalized gray-scale image I by: $A = e^{-1.6 \times I}$ $\Phi = I$. Noisy holograms are simulated using E.qs (1) and (2) described in 2.1 with a pixel pitch of 4.48 μm and a wavelength of 430 nm. For each sample, five different holograms are simulated using a height stack z in [300 μm $\longrightarrow$ 700 μm] with a step size of 100 μm. The train and test sets contain, respectively, 5000×5 and 1000×5 samples. Cross-Net is trained using Adam optimizer and a Mean Absolute Error (MAE) loss function between the predicted $\mathbf{x}_i$ and the ground truth latent complex field $\tilde{\mathbf{x}}_i$ for $i = 1, ..., T$:

$$\mathcal{L} = \frac{1}{T} \sum_{i=1}^{T} |\tilde{\mathbf{x}}_i - \mathbf{x}_i| \tag{3}$$

Where T is the number of samples per batch. Note that the regression branch is trained in a similar fashion.

4 Experimental Results

State-of-the-art approaches are reproduced using either publicly available source code or model description in the original papers: HRNet [17] is reproduced using the network description in [17] and is trained on amplitude-only data. Dense-U-Net [22] and ViT [10] are reproduced using a publicly available code. FIN [1] and its subsequent version eFIN [2] are reproduced using available code from Huang et al. [11]. We adapt ViT [10] to perform joint reconstruction and refocusing using two dedicated branches similar to Cross-Net and compare regression performance with that of eFIN [2] since it is the only approach from the literature that performs such tasks jointly. Quantitative results are shown in Tab. 1 where

[†]https://huggingface.co/Fictiverse/Stable_Diffusion_Microscopic_model

Structural Similarity Index (SSIM) and Peak Signal-to-Noise Ratio (PSNR) values are reported to assess the similarity between the ground truth complex field and the reconstructed one. The refocus distance estimation is assessed by the average L1 loss ($\delta\mathbf{z}$) from the ground truth height.

Table 1. Quantitative comparison results on synthetic holographic inner test data.

Method	$\delta\mathbf{z}\downarrow$	SSIM↑		PSNR↑	
		A	Φ	A	Φ
HRNet [17]	-	0.370	-	17.4	-
Dense-UNet [22]	-	0.527	0.501	20.2	17.6
FIN [1]	-	0.887	0.867	26.7	23.7
eFIN [2]	1.35	0.849	0.828	25.0	22.1
ViT [10]	120	0.779	0.753	23.4	20.6
Cross-Net	**0.04**	**0.911**	**0.896**	**27.8**	**24.7**

Table 2. Quantitative comparison results on synthetic holographic outer test data.

Method	$\delta\mathbf{z}\downarrow$	SSIM↑		PSNR↑	
		A	Φ	A	Φ
FIN [1]	-	0.917	0.908	33.2	31.2
eFIN [2]	41.1	0.752	0.742	26.9	25.0
Cross-Net	**0.821**	**0.935**	**0.927**	**34.2**	**32.2**

Quantitative metrics reported in Tab. 1 show that Cross-Net outperforms ViT and other state-of-the-art approaches on the reconstruction and regression tasks. Notice that such performance gain is achieved with lighter network architecture compared to ViT. In fact, The number of trainable parameters in the reconstruction branch is around 10.5 M and 7.5 M for regression branch. In contrast, ViT has 43.2 M and 19.1 M trainable parameters for the two branches. Loss curves during training and test phases are shown in Fig. 5 for both reconstruction and refocusing tasks. Cross-Net reduces the risk of over-fitting compared to ViT and Dense-UNet [22] and it is more stable compared to eFIN [2] in the reconstruction task, while ViT tends to predict an average distance value in an effort to minimize the regression loss, i.e., it tends to stuck in a bad local minima.

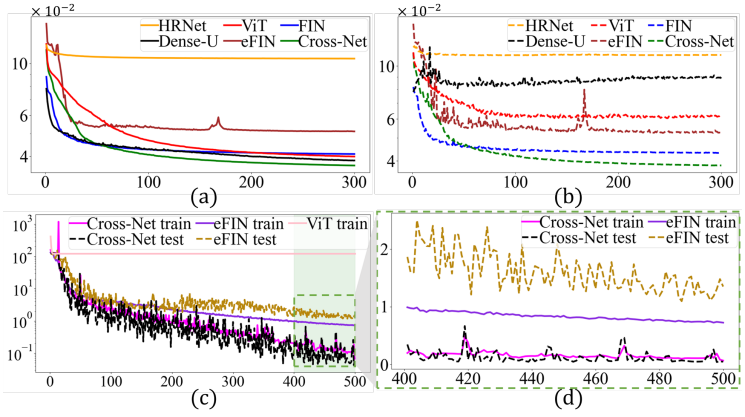

Fig. 5. (a): L1 loss of reconstruction on the train set. **(b):** L1 loss of reconstruction on the test set. **(c):** L1 loss of the regression task **(d):** Zoomed-in region from (c).

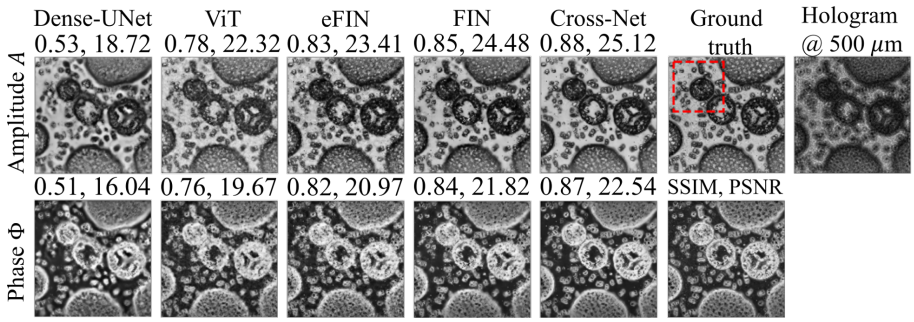

Fig. 6. Reconstructed phase and amplitude images from multiple approaches with an input hologram captured at 500 μm.

In order to assess model performance on test samples beyond the training domain, we tested Cross-Net as well as Fin and eFIN on other simulated test samples that exhibit completely different characteristics from the train set dubbed "outer" test set and reported metrics in Tab. 2. The images used contain a group of dense round cells with simple structures. Compared to eFIN, Cross-Net has better external generalization for both reconstruction and regression.

Qualitative reconstruction results from a sample hologram captured at 500 μm are shown in Fig. 6 and zoomed-in image region (red box) is shown in Fig. 7 highlights the refocus distance estimation using Cross-Net which has a deviation of less than 0.2 μm compared to the reference height.

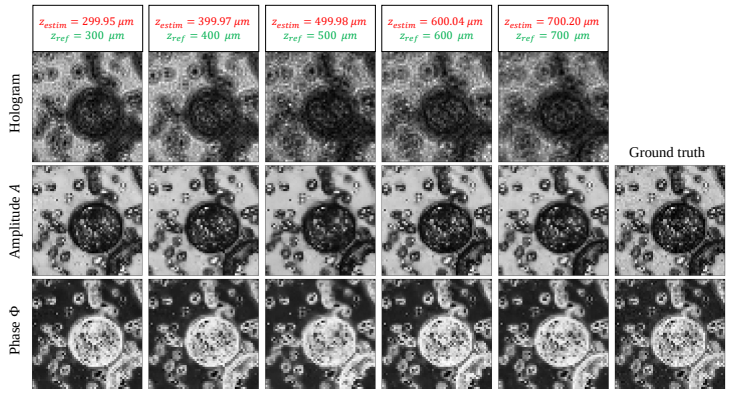

Fig. 7. Zoomed-In red square region from Fig. 6 with estimated refocus distances.

5 Ablation Studies

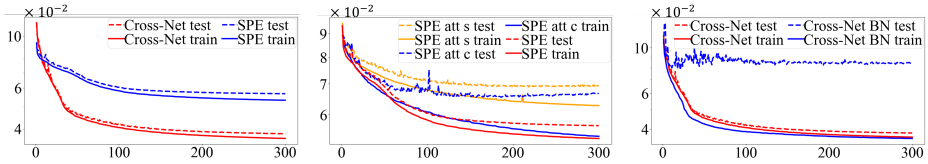

Fig. 8. L1 loss curves in the ablations experiments.

Ablations experiments have been carried out to assess the contributions of different modules within the reconstruction branch. Fig. 8 shows loss curves of multiple experiments highlighting the contributions of: (i) $Conv_S$ with respect to the standard multi-head attention unit, (ii) Patch embedding decomposition, and (iii) the use of weight normalization. Fig. 8 (left) shows loss curves of Cross-Net as well as a different variant which uses a single-step Patch Embedding rather than gradually up-sampling channel features and down-sampling spatial features. The model is dubbed as Single-step Patch Embedding (SPE) and has a similar size as Cross-Net. Fig. 8 (middle) shows loss curves obtained using the SPE model and another variant where $Conv_S$ is replaced with Multi-Head Self-Attention. Fig. 8 (right) shows L1 loss curves using Cross-Net trained with weight normalization and standard batch normalization separately.

6 Conclusions and Outlook

In this paper, Cross-Net is proposed. It can achieve joint refocusing and holographic image reconstruction. Compared to classic Vision Transformers, it can extract more meaningful latent features all while with smaller model size, achieved by decomposing standard Patch Embedding and using spatial-wise convolutions. Evaluation on synthetic holographic data highlight the proposed model's contribution and superior performance than the current state-of-the-art approaches. As a future work would focus on spatial super-resolution capability and model adaptation to real holographic measurements.

References

1. H. Chen, L. Huang, T. Liu, and A. Ozcan. Fourier imager network (fin): A deep neural network for hologram reconstruction with superior external generalization. *Light: Science & Applications*, 11(1):254, 2022.
2. H. Chen, L. Huang, T. Liu, and A. Ozcan. efin: Enhanced fourier imager network for generalizable autofocusing and pixel super-resolution in holographic imaging. *IEEE Journal of Selected Topics in Quantum Electronics*, 29(4: Biophotonics):1–10, 2023.

3. X. Chen, H. Wang, A. Razi, M. Kozicki, and C. Mann. Dh-gan: a physics-driven untrained generative adversarial network for holographic imaging. *Optics Express*, 31(6):10114–10135, 2023.
4. E. Cuche, P. Marquet, and C. Depeursinge. Spatial filtering for zero-order and twin-image elimination in digital off-axis holography. *Applied optics*, 39(23):4070–4075, 2000.
5. J. L. de Almeida, E. Comunello, A. Sobieranski, A. M. da Rocha Fernandes, and G. S. Cardoso. Twin-image suppression in digital in-line holography based on wave-front filtering. *Pattern Analysis and Applications*, 24(3):907–914, 2021.
6. A. Foi, M. Trimeche, V. Katkovnik, and K. Egiazarian. Practical poissonian-gaussian noise modeling and fitting for single-image raw-data. *IEEE transactions on image processing*, 17(10):1737–1754, 2008.
7. D. Gabor. Microscopy by reconstructed wave-fronts. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 197(1051):454–487, 1949.
8. B. D. Haeffele, R. Stahl, G. Vanmeerbeeck, and R. Vidal. Efficient reconstruction of holographic lens-free images by sparse phase recovery. In *MICCAI 2017: 20th International Conference*, pages 109–117. Springer, 2017.
9. A. Haleem, M. Javaid, and I. H. Khan. Holography applications toward medical field: An overview. *Indian Journal of Radiology and Imaging*, 30(03):354–361, 2020.
10. K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, pages 16000–16009, 2022.
11. L. Huang, H. Chen, T. Liu, and A. Ozcan. Self-supervised learning of hologram reconstruction using physics consistency. *Nature Machine Intelligence*, 5(8):895–907, 2023.
12. L. Huang, T. Liu, X. Yang, Y. Luo, Y. Rivenson, and A. Ozcan. Holographic image reconstruction with phase recovery and autofocusing using recurrent neural networks. *ACS Photonics*, 8(6):1763–1774, 2021.
13. T. Latychevskaia and H.-W. Fink. Solution to the twin image problem in holography. *Physical review letters*, 98(23):233901, 2007.
14. T. Latychevskaia and H.-W. Fink. Practical algorithms for simulation and reconstruction of digital in-line holograms. *Applied optics*, 54(9):2424–2434, 2015.
15. S.-H. Lee and D. G. Grier. Holographic microscopy of holographically trapped three-dimensional structures. *Optics Express*, 15(4):1505–1512, 2007.
16. A. E. G. Madsen, M. A. Panah, P. E. Larsen, F. Nielsen, and J. Glückstad. On-axis digital holographic microscopy: Current trends and algorithms. *Optics Communications*, page 129458, 2023.
17. Z. Ren, Z. Xu, and E. Y. Lam. End-to-end deep learning framework for digital holographic reconstruction. *Advanced Photonics*, 1(1):016004–016004, 2019.
18. Y. Romano, M. Elad, and P. Milanfar. The little engine that could: Regularization by denoising (red). *SIAM Journal on Imaging Sciences*, 10(4):1804–1844, 2017.
19. R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022.
20. A. Trockman and J. Z. Kolter. Patches are all you need? *arXiv preprint arXiv:2201.09792*, 2022.
21. H. Wang, M. Lyu, and G. Situ. eholonet: a learning-based end-to-end approach for in-line digital holographic reconstruction. *Optics express*, 26(18):22603–22614, 2018.
22. Y. Wu, J. Wu, S. Jin, L. Cao, and G. Jin. Dense-u-net: dense encoder–decoder network for holographic imaging of 3d particle fields. *Optics Communications*, 493:126970, 2021.