CROSS-NET: IN-LINE HOLOGRAPHIC IMAGE RECONSTRUCTION AND REFOCUSING

*A. Haitao Zhou, B. Mazen Mel**

EPFL A

ABSTRACT

Holography enables quantitative phase sensing through light wave interference allowing for interesting imaging modalities applied in computational microscopy where the light phase 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 requiring 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 it adapts Patch-Embedding to generate meaningful latent representation suitable for the tasks at hand. The proposed model is lighter and exhibits significant performance gain in the joint tasks of computational refocusing and holographic image reconstruction compared to ViT and other state-of-the-art approaches.

***Index Terms*—** In-line holography, Holographic reconstruction, computational image refocusing.XXX

1. INTRODUCTION

Holography is a technique that enables a light field, defined by both an amplitude and a phase distributions, to be recorded and later reconstructed. It is best known as a method to generate real three-dimensional images due to the fact that holography records the light's intensity and direction. Holographic image reconstruction is a critical step that allows to leverage the full potential of holography in a myriad of applications such as medical imaging [1], microscopy [2], information security [3], data storage [4], and industrial inspection [5]. The biggest challenge hindering the wide spread application of holography is the twin image problem which is caused by the object wave's complex conjugate [6] and causes undesirable blur artifacts and deteriorates the image quality. To get rid of twin image, both adjustments on hardware and software parts can be included [7]. 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 [8]. However, in this work we are interested in in-line holography wherein both reference and object waves propagate through the same optical axis thus significantly reducing the hardware complexity resulting in compact imaging devices. The twin-image influence in this case is 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. **MAZEN:** Better explain this part use clear and simple language (refer to other classic papers): The forward propagation and backward propagation of field between two points is a kind of natural iteration. Refer to Fig 1 When the field propagates to a plane, some constraints are performed on the field to refine it, so that it gradually converges to the desired solution. Such constraints include positive absorption profile of the field at the object plane [9], filtering mask [10], denoising prior [11], and sparsity [12]. Learning-based approaches for holographic image reconstruction have been also proposed. Previous research has shown the success of several popular neural network architectures on such task among which are convolutional neural networks (CNNs) [13], recurrent neural networks (RNNs) [14], as well as generative adversarial models (GANs) [15]. In this paper, a novel learning-based approach *Cross-Net* is proposed which 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 to estimate the distance z at which the hologram was recorded. Cross-Net is able to extract more meaningful latent vectors for patch-based image reconstruction and regression tasks than a standard Vision Transformer (ViT) [16, 17] all while being significantly lighter. Our proposed model can also surpass state-of-the-art approaches including HRNet [18], Dense-U-net [19], Fourier Imager Network (FIN) [20] and its subsequent variant Enhanced Fourier Imager Network (eFIN)[21] 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 encoder and simple periodic shuffling as a decoder. 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.

2. METHODOLOGY

In this section the image formation model underlying in-line holography is discussed followed by our proposed architecture.

2.1. Image Formation Model

The complex transmission of an object can be expressed as:

$$U_{\text{object}}(x,y) = A(x,y) \cdot e^{i\Phi(x,y)} \tag{1}$$

where U_{object} is the complex transmission of the object, A is the amplitude, and Φ is the phase. The propagation factor in angular spectrum method can be expressed as:

$$P = e^{\frac{2\pi i z}{\lambda}\sqrt{1-(\lambda u)^2-(\lambda v)^2}} \tag{2}$$

where P is the propagation factor in angular spectrum method, z is the distance between object and detector, λ is wavelength, and (u, v) is the Fourier domain coordinate. After propagation, the field at detector plane can be expressed as:

$$U_{\text{detector}}(X,Y) = \mathcal{F}^{-1}\left(\mathcal{F}\left(U_{\text{object}}(x,y)\right) \times P\right) \tag{3}$$

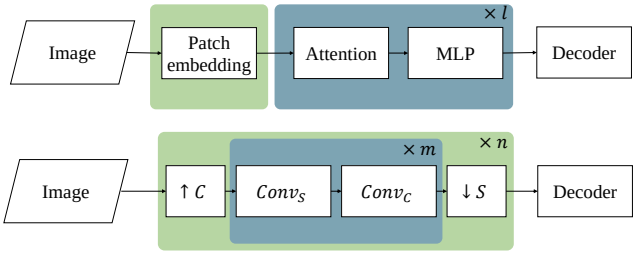
where U_{detector} is the propagated field at detector plane, $\mathcal{F}$ and $\mathcal{F}^{-1}$ are the Fourier and inverse Fourier Transforms. The detector is only sensitive to the intensity of the incoming field which corresponds to the square modulus of U_{detector} :

$$H(X,Y) = |U_{\text{detector}}(X,Y)|^2 \tag{4}$$

where H is the recorded hologram. The main task is to recover A and Φ given a hologram captured at a distance z.

2.2. Model Architecture

The workflow of the classic ViT model is shown on the top of Figure 1 (a); The output of patch embedding is fed to a Multi-Head Self-Attention layer that acts as an inter-spatial processing unit and then a MLP that acts as an inter-channel processing unit, this module is usually repeated $\times l$ times as shown in Figure 1 (a). 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. Directly applying the ViT model 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 in the attention layer. Second, the loss of information in the Patch Embedding as the consistency of image global features spanning large areas might be affected by small patches, furthermore fine image features could be lost due to the compression-like effect of Patch Embedding.



(a)

(b)

Fig. 1: **(a):** The basic concept compare of Vision Transformer(ViT) and Cross-Net. **(b):** The convolution block used to achieve Up C, Conv S and Conv C units shown in (a).

MAZEN: Please always mention the reason behind any architecture module (why convS and ConvC and so on... Cross-Net is proposed to solve the aforementioned shortcomings of ViT. The concept of our model is shown in the bottom of Figure 1 (a) and the workflow used to define the different convolution layers is also depicted in Figure 1 (b). Multi-Head Self-Attention is substituted with $Conv_S$ which acts as an inter-spatial layer defined by a convolution operation where the number of groups is equal to the number of XXX channels with large kernel size, e.g., 17 for an image size of 128$\times$ pixels, such settings prompt said layer to produce that XXXX EXPLAIN W.R.T FIGURE. In this way $Conv_S$ solves the over-fitting issue due to the fact XXXXXX EXPLAIN HERE XXXXX and consequently the high resource consumption. Its effectiveness in substituting Multi-Head Self-Attention as an inter-spatial processing unit has been shown in previous research [22] in classification we adapt it to reconstruction and regression tasks**MAZEN:** Highlight differences w.r.t trockman2022patches. $Conv_C$ indicates a channel-wise convolution operation with a kernel size of 1. It has similar function as a normal MLP. XXXX STOPED HERE XXXX **Haitao:** It is used to take in weight normalization and simplify the kinds of modules used in the

model. Patch embedding is decomposed into a series of operations to up-sample channel features $\uparrow C$ and down-sample space features $\downarrow S$. **Haitao:** The feasibility of decomposing Patch Embedding is ensured by the usage of $Conv_S$, because Multi-Head Self-Attention consumes quadratic runtime which makes it nearly unavailable on not totally embedded images. Furthermore, information processing units $Conv_S$ and $Conv_C$ **MAZEN:** What does this means: are emerged into this series.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. Finally, a simple decoder similar to that of Vit is used in Cross-Net. **MAZEN:** There is an issue here I thought the overall network is called cross-net and in the following paragraph there is another bigger architecture featuring two cross-nets. This is problematic because from the beginning the paper is based on the idea of a single cross-net as the overall architecture. Please reformulate the paragraph accordingly to the rest of the paper The overall architecture has two branches used to perform reconstruction and regression tasks at the same time. Every branch has an independent Cross-Net. The two Cross-Nets have similar architecture except the simple decoder at last. When inputting the holograms of the same object at different object-detector distances, the branch of reconstruction should always output the same complex reconstructions, and the branch of regression should output different distance estimates

3. DATA AND TRAINING DETAILS

MAZEN: emphasis the fact that we used a diffusion model finetuned on microscopic images which can generate large number of training samples and avoid too much detail about types of cells in the images: Notice that those are just prompts fed into the stable diffusion model which generates a random image not necessarily containing the type of cell used in the prompt. They are basically random images with statistics close to that of normal microscopy The original data set includes various microbial images. Based on this microscopic data set and the model described in 2.1. Image Formation Model, the data set used for this paper is simulated manually by the steps shown in Figure 2.

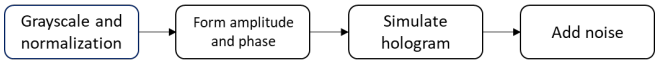


Fig. 2: Steps to generate the data set.

MAZEN: replace the following steps with a block diagram and reformulate those items into a coherent paragraph describing the block diagram

First, convert RGB image to grayscale image and normalize the grayscale image from 0-255 to 0-1. Second, form complex object transmission. The phase phi is directly set equal to the normalized image, and the amplitude is calculated as the exponential function with a minus factor applied to the normalized image. It can make both amplitude and phase range from 0 to 1. Third, Simulate hologram following principle described in 2.1.Image Formation Model. The details can also be found in previous research[6]. Fourth, Add shot shot noise and read noise to the formed hologram, and clip the noised hologram between zero and the maximum of hologram without noise.

The image size is 128×128. The pixel size and wavelength are 4.48um and 430nm. Use each basic image to generate five input-target sample pairs. Inputs of the five samples are holograms at 300um, 400um, 500um, 600um and 700um. Targets are the same complex object transmission with different ground truths of distances.

Train set contains roughly 5000×5 pairs, and test set contains 1000×5 pairs.

We train the network using the Adam optimizer with a learning rate of 0.0005, we use also a learning rate scheduler with a step size of 15 and gamma equals to 0.9. Four NVIDIA RTX A6000 are used for distributed learning, and the batch size on each GPU is 15. The loss function used in all training setups is a simple L1 loss.

4. EXPERIMENTAL RESULTS

The details of Cross-Net in the branch of patch-recover reconstruction are shown in Figure 3. Up C 1 and Up C 4 keeps the number of channel features rather than increase it as normal. Down S 1-4 use Max pool operations. Big kernel sizes used for Conv S are respectively 17, 9, 5, 3 for blocks at stage 1 to 4. Conv in also uses the convolution block shown in Figure 1 (b) and kernel size 17. Decoder here first uses a linear function to map the size of data from [B, 768, H//16, W//16] to [B, 512, H//16, W//16], then uses some basic reshape operations to transform it to the size [B, 2, H, W]. For Cross-Net in the branch of regression, some parameters and the last decoder are changed. The decoder is average pool and liner map.

	SSIM1	SSIM2	PSNR1	PSNR2	z_error
HRNet [18]	0.370	-	17.4	-	-
Den-U [19]	0.527	0.501	20.2	17.6	-
FIN [20]	0.887	0.867	26.7	23.7	-
eFIN [21]	0.849	0.828	25.0	22.1	1.35
ViT [16, 17]	0.779	0.753	23.4	20.6	120
Cross-Net	0.911	0.896	27.8	24.7	0.0485

Table 1: Metrics to show reconstruction and regression quality on test set. SSIM1, SSIM2, PSNR1, PSNR2 means SSIM, PSNR of amplitude and phase. z_error means the average L1 loss between the regression and ground truth distance. The empty cell indicates that the model lacks reconstruction or regression functionality corresponding to this parameter.

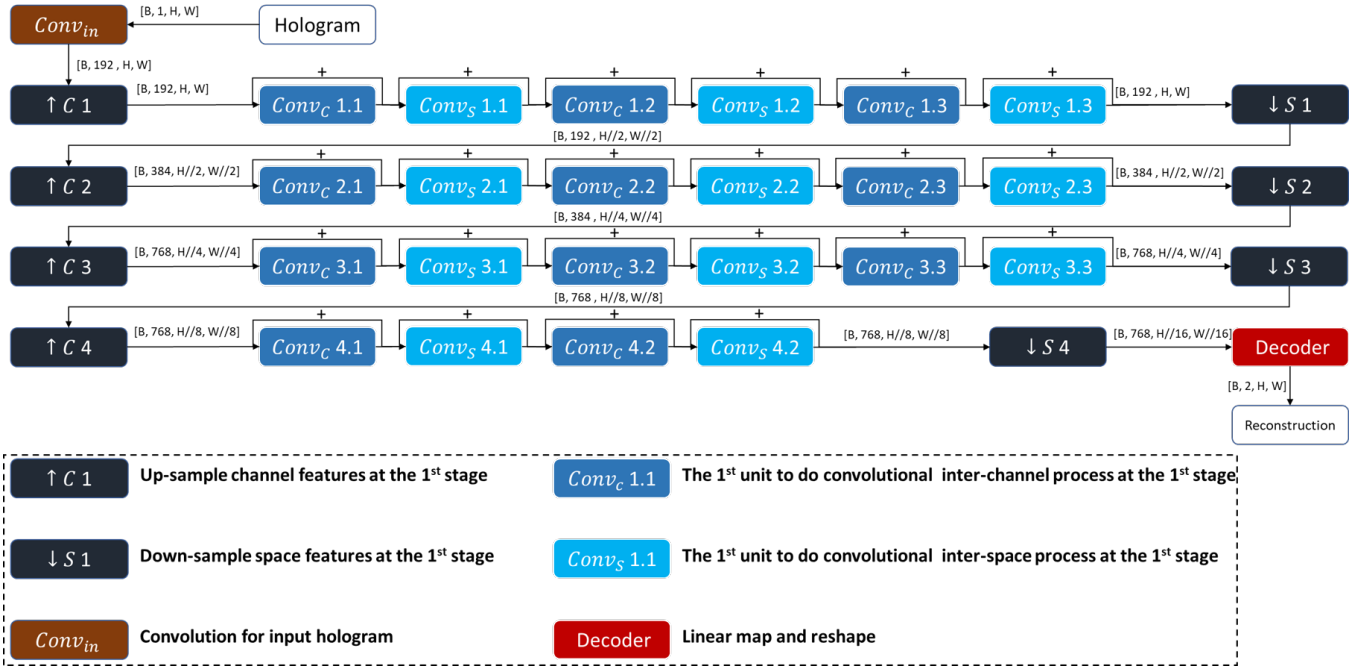


Fig. 3: Detailed architecture of Cross-Net used for reconstruction.

Models used for compare study follow the authors' original configuration. HRNet outputs reconstruction of amplitude. Dense-U-net and FIN outputs reconstruction of complex object transmission. eFIN, ViT and Cross-Net outputs complex object transmission and distance regression. ViT and Cross-Net uses two-branch architecture which substantially contains two similar networks. Keep the extracted latent vector before the last decoder of ViT and Cross-Net the same size for compare(like [15, 64 768] and [15, 768, 8, 8]).

The L1 loss curves of reconstruction on train set and test set are shown in Figure 4 (a) and Figure 4 (b). **Haitao:** The train loss and test loss of FIN, eFIN and Cross-Net at epoch 300 are 0.0409 and 0.0432, 0.0512 and 0.0528, 0.0366 and 0.0383. It shows that reconstruction branch of Cross-Net has both good capacity and anti-over-fit. The training of eFIN is very unstable whose loss usually explodes during the training procedure, while the convergence of Cross-Net is always very smooth. The L1 loss curves of z regression are shown in Figure 4 (c). The regression loss of ViT quickly converges to 120, because its capacity can not support it to distinguish holograms at different distances. Then, it reduces the loss by converging all regressions to 500um which is the middle of between 300um and 700um. All these three plots use log scale for y axis. Metrics on test set are shown in Table 1. The models used for reconstruction tasks are trained for 300 epochs, and those for regression tasks are trained for 500 epochs. The reconstruction images and regression values of an object in test set are shown in Figure 6.

Theses results show the superior performance of Cross-Net. It can surpass ViT and other state of the art for both reconstruction and regression tasks **Haitao:** which means Cross-Net can extract both the identity and difference of holograms under different distances. For Cross-Net, the size of reconstruction branch is around 10.5 million and that of regression branch is around 7.5 million. For ViT, the corresponding sizes are 43.2 million and 19.1 million. The size of Cross-Net is much smaller than that of ViT.

	SSIM1	SSIM2	PSNR1	PSNR2	z_error
FIN [20]	0.917	0.908	33.2	31.2	-
eFIN [21]	0.752	0.742	26.9	25.0	41.1
Cross-Net	0.935	0.927	34.2	32.2	0.821

Table 2: Metrics to show reconstruction and regression quality on valid set.

The images used for train set and test set are split from the same data set. It means they could have the similar statistic distribution. To valid the external generalization, a valid data set is created. The images used contain a group of dense round cells without complicated details, rather than one or several cells with complicated details as previous. The valid set has the same size as test set. The external generalization result of FIN, eFIN and Cross-Net are shown in Table 2. Compared to eFIN, Cross-Net has better external generalization for both reconstruction and regression.

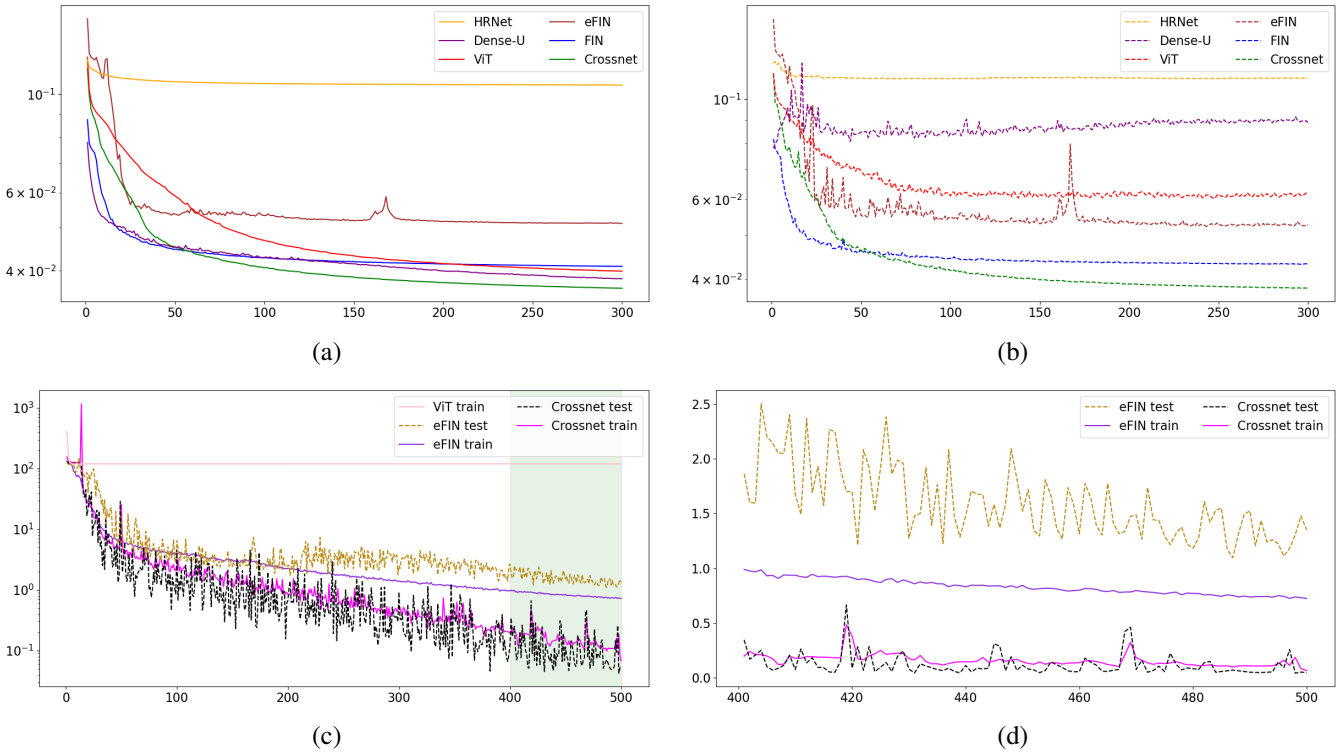


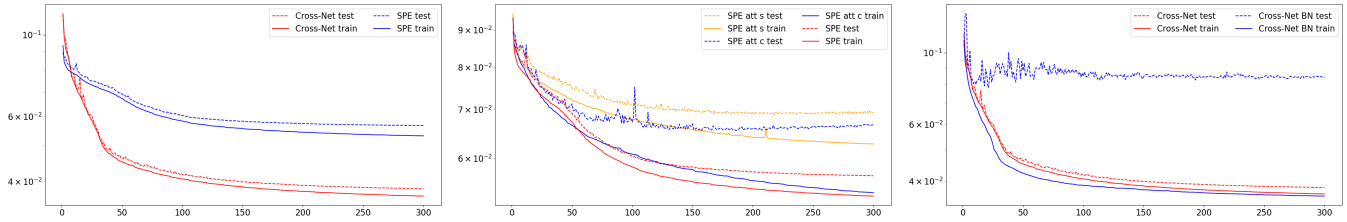
Fig. 4: The ticks on the x-axis mean the number of epochs. **(a):** L1 loss of reconstruction on the train set. **(b):** L1 loss of reconstruction on the test set. **(c):** L1 loss of distance z regression **(d):** Zoom in of the region with green shadow in (c). The plot (d) uses a linear scale instead of a logarithmic scale. **Haitao:** can't find a choice to make the plotting region same size

5. ABLATION STUDIES

Ablations experiments have been carried out to assess the contributions of different modules within the reconstruction branch, notice that the regression branch shares the same architecture. Figure 5 shows loss curves of multiple experiments highlighting the contributions of: (i) $Convs_S$ with respect to the standard multi-head attention unit, (ii) Patch embedding decomposition as illustrated in Figure 1, and (iii) the use of weight normalization. Figure 5 (a) shows loss curves of Cross-Net as well as a different variant which uses a single-step Patch Embedding (**Haitao:** directly from [B, 1, H, W] to [B, C, H/16, W/16]) rather than gradually up-sampling channel features and down-sampling spatial features, and uses single $Convs_S$ and $Convs_C$ units (those in the fourth stage of the reconstruction branch of Cross-Net) rather than four as in the original architecture. The model is dubbed as SPE (Single-step Patch Embedding) and has a similar size as Cross-Net. One can deduce that by decomposing Patch Embedding and processing the latent vector within multiple stages, **MAZEN:** explain this: the capacity of the model would significantly increase, **Haitao:** since the saturation losses of Cross-Net on both train set and test set are obviously smaller than those of SPE. Figure 5 (b) shows loss curves obtained using the SPE model and another variant where $Convs_S$ is replaced with Multi-Head Self-Attention. Notice that this gives rise to two different cases: (i) A model that has a similar size as the original SPE achieved by decreasing the number of inter-spatial and inter-channel processing units since Multi-Head Self-Attention is much bigger compared to $Convs_S$. (ii) A model that has the same cycles of processing units as SPE where $Convs_S$ are replaced with Multi-Head Self-Attention and the rest of the architecture is kept unchanged. These two cases are dubbed as *SPE att s* and *SPE att c* respectively. In the former case, both training and test losses are much worse than the original SPE. In the latter case severe over-fitting is observed as the test loss starts to diverge. Such behaviour It proves highlights the effectiveness of $Convs_S$ in not only keeping the model from over-fitting but also achieves faster and much better convergence. Figure 5 (c) shows L1 loss curves using Cross-Net trained with weight normalization and standard batch normalization separately, notice how the network severely over-fit the training data in the latter case.

6. CONCLUSIONS

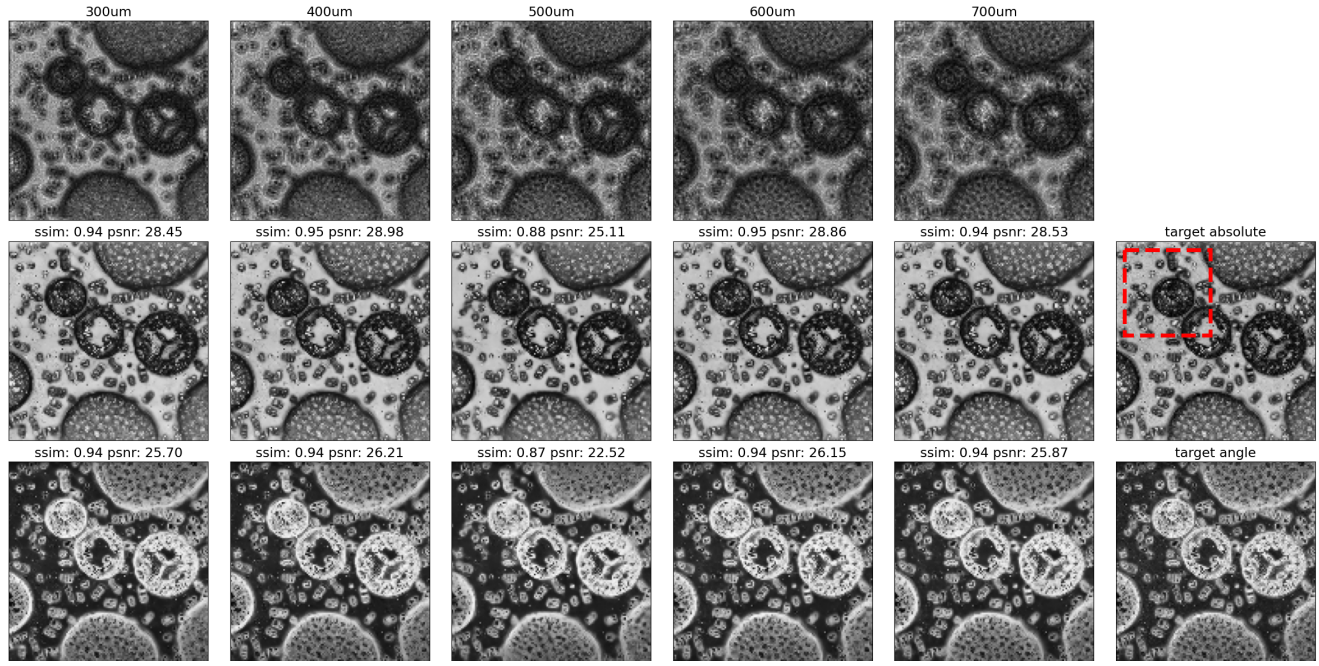
In this paper, Cross-Net is proposed. It can achieve joint refocusing and holographic image reconstruction. Compared



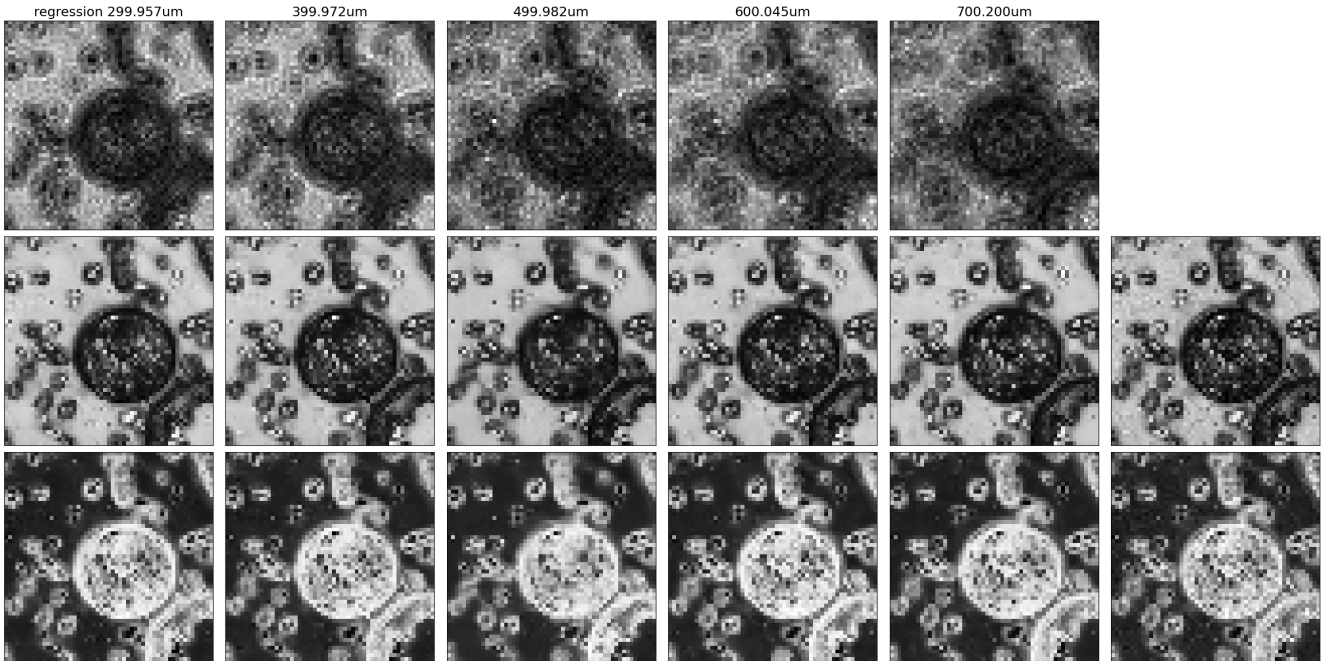
(a) (b) (c)

Fig. 5: The L1 loss curve for ablations study.

to Vision Transformer, it can extract a more meaningful latent vector with smaller model size. This effectiveness and tightness are achieved by substituting convolution block with space-wise convolution and large kernel size configuration for Multi-Head Self-Attention and decompose Patch Embedding. The performance of Cross-Net is also superior than other state-of-the-art approaches. As a future work spatial super-resolution capability can be incorporated as it may be achieved by simply changing the out channels of the linear map in the last simple decoder.

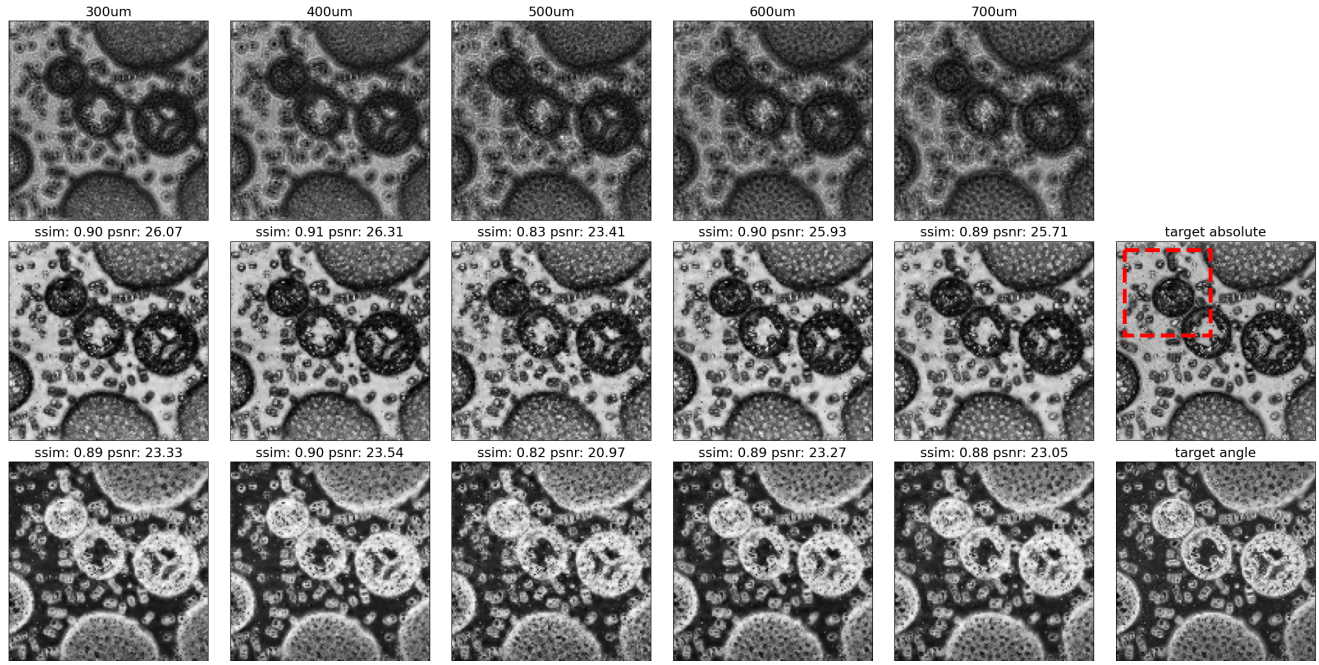


(a)

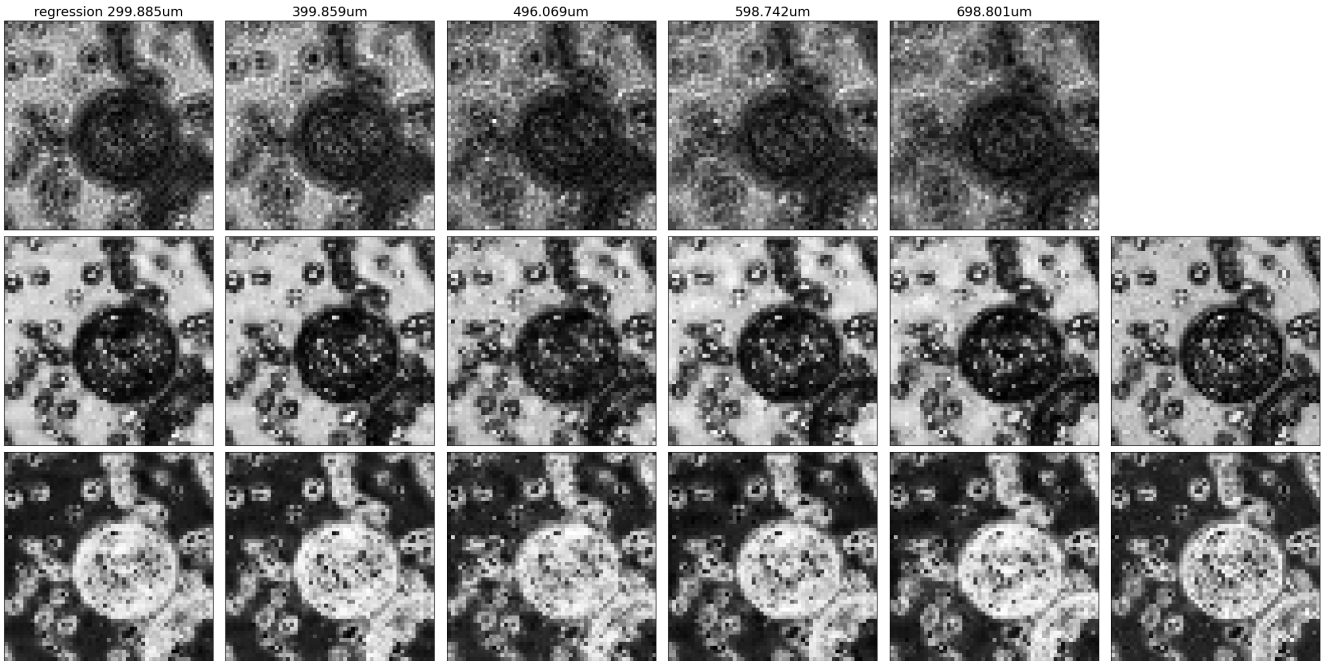


(b)

Fig. 6: **(a)** Auto-focusing reconstruction result of Cross-Net. The first row shows holograms at 5 different object-tensor distances 300um, 400um, 500um, 600um, 700um, the next two rows show corresponding reconstruction and target of amplitude and phase. **(b)** Zoom in of the region notified using red rectangle in target amplitude of figure(a). Subtitles of the first row are the regression prediction results of object-tensor distances.**MAZEN:** remove x,y axis, use more thick lines of the red square

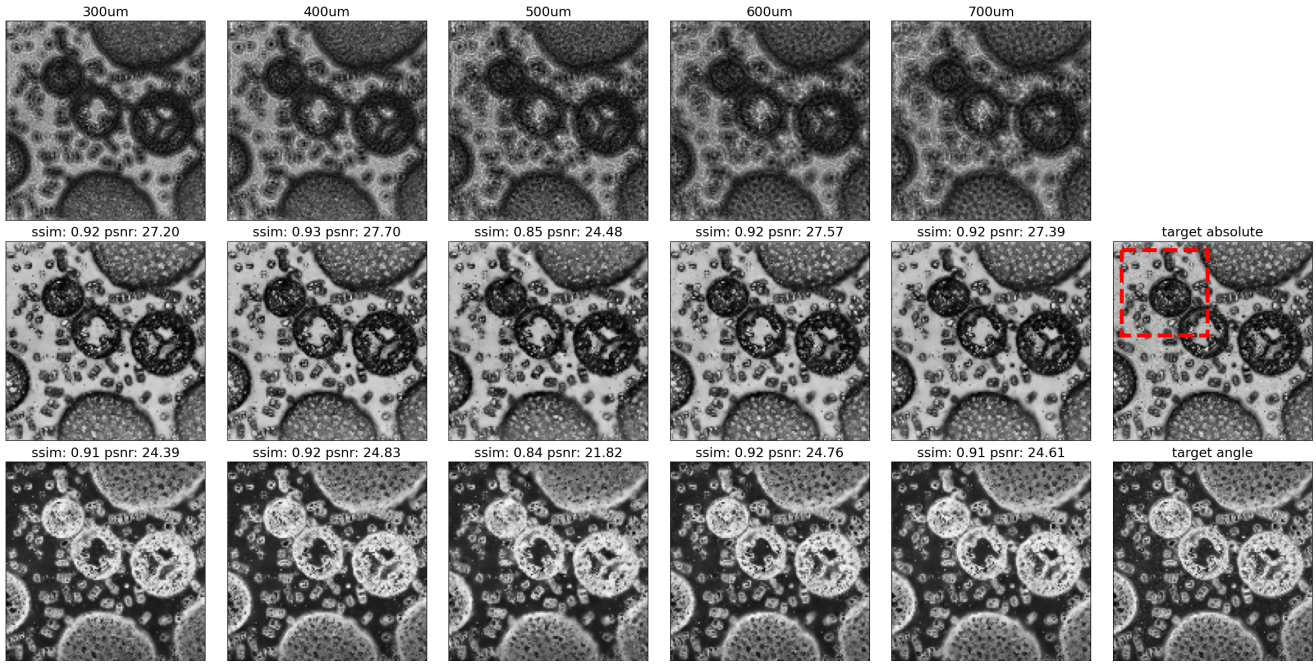


(a)

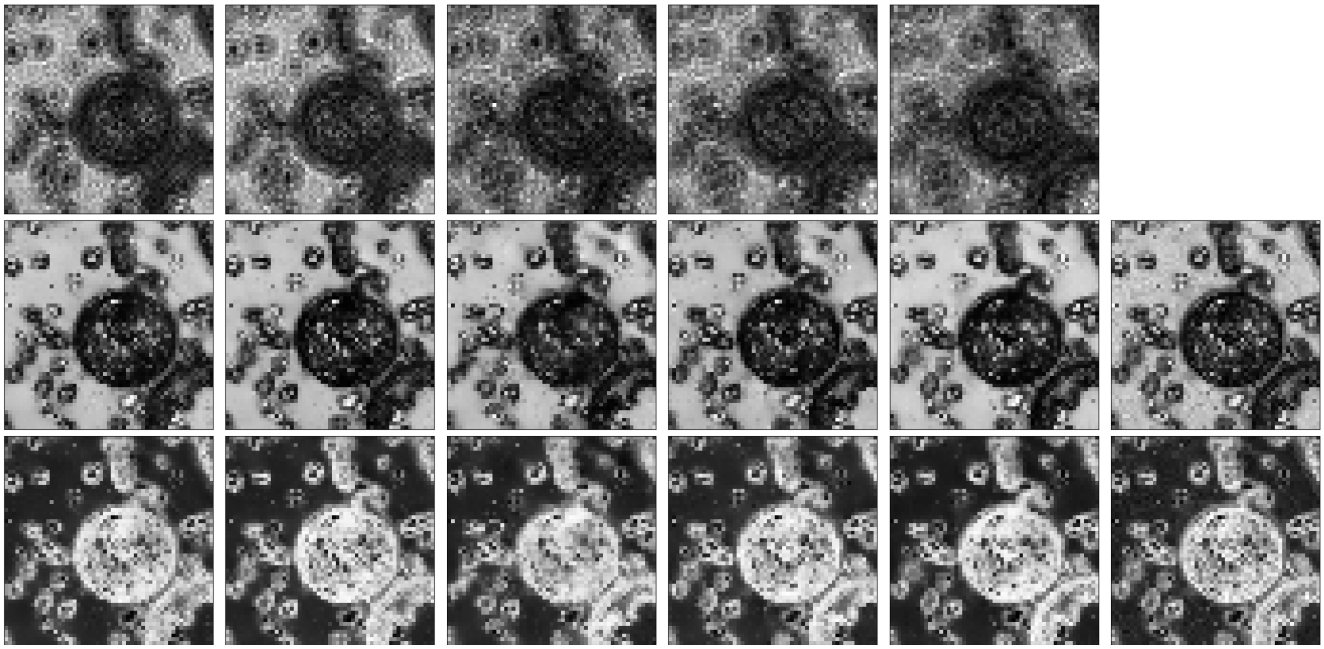


(b)

Fig. 7: The reconstruction and regression quality of eFIN. The text and figures correspond to those at the same location of Figure 6.

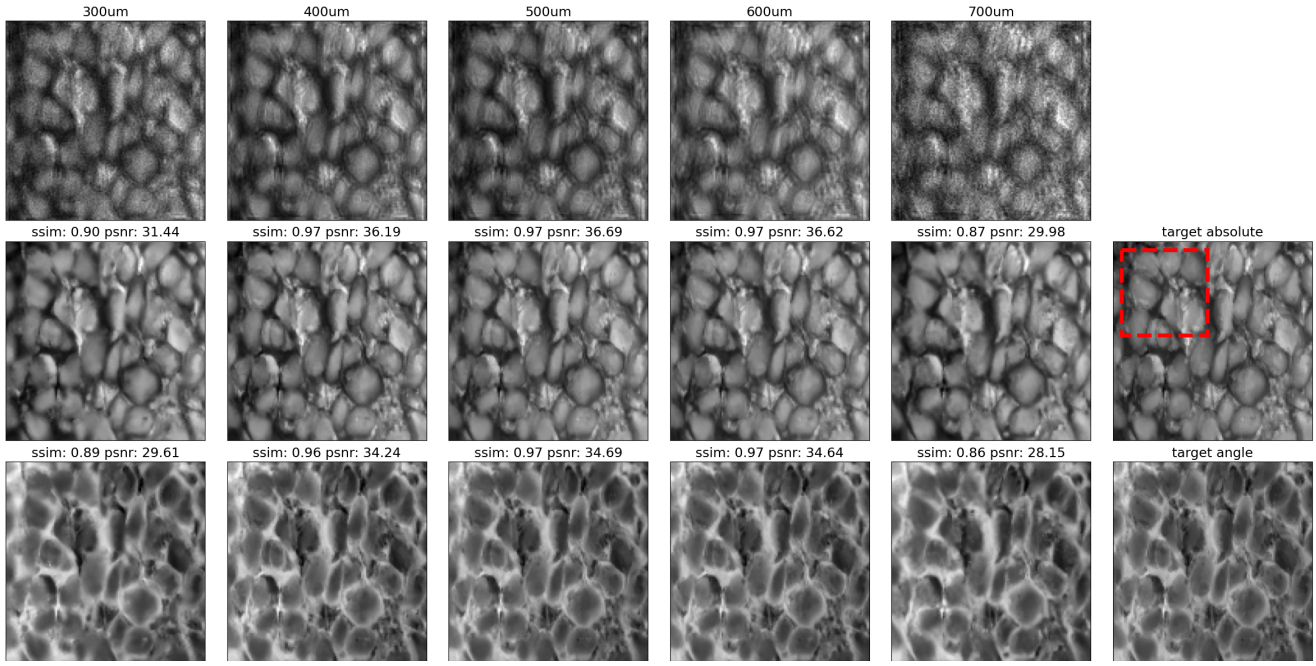

300um
400um
500um
600um
700um
ssim: 0.92 psnr: 27.20
ssim: 0.93 psnr: 27.70
ssim: 0.85 psnr: 24.48
ssim: 0.92 psnr: 27.57
ssim: 0.92 psnr: 27.39
target absolute
ssim: 0.91 psnr: 24.39
ssim: 0.92 psnr: 24.83
ssim: 0.84 psnr: 21.82
ssim: 0.92 psnr: 24.76
ssim: 0.91 psnr: 24.61
target angle

(a)

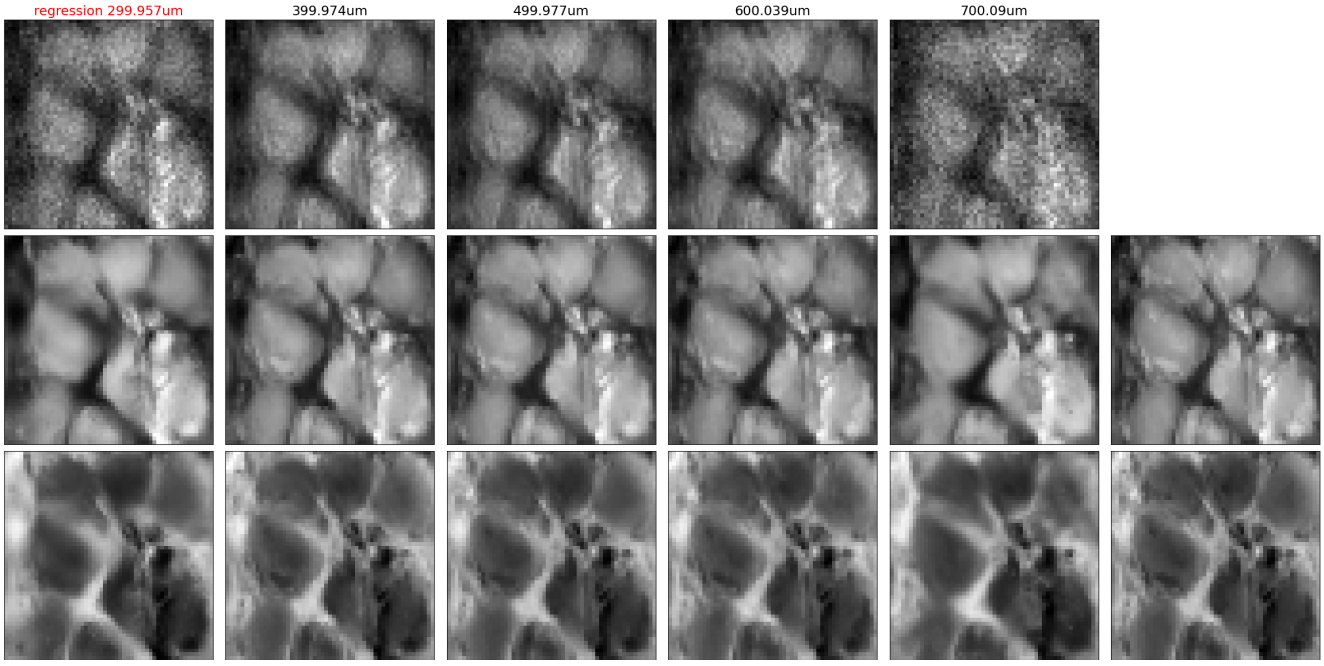


(b)

Fig. 8: FIN.

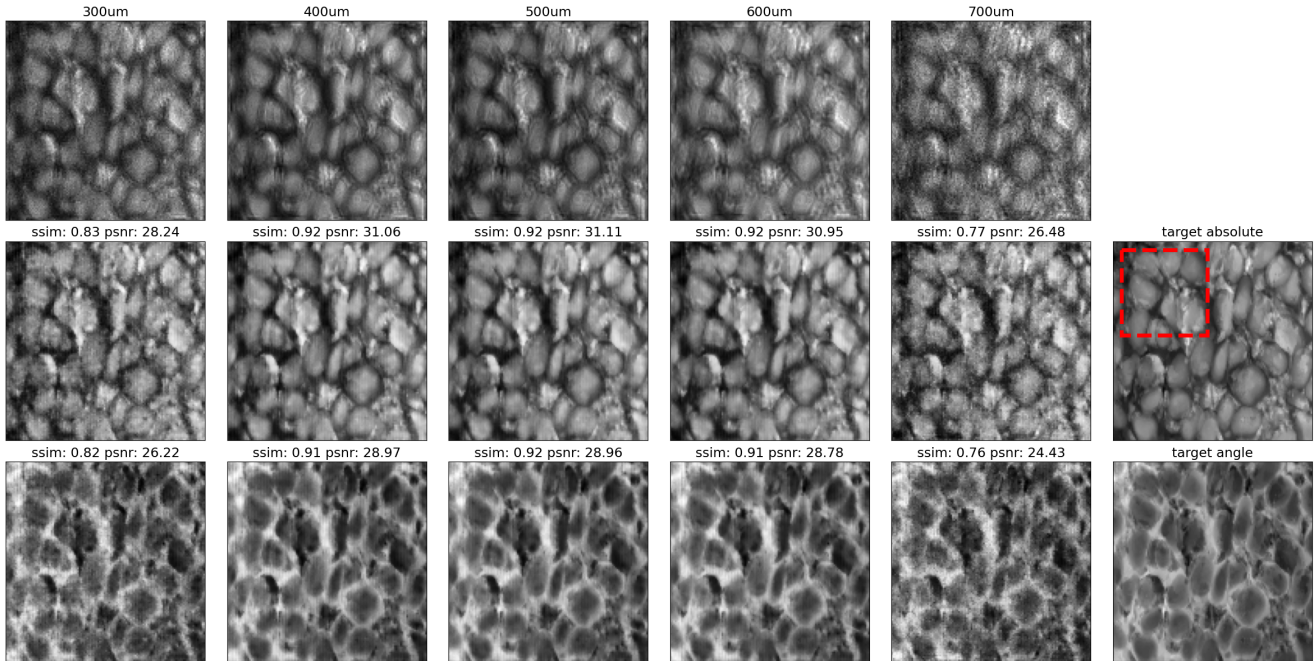


(a)

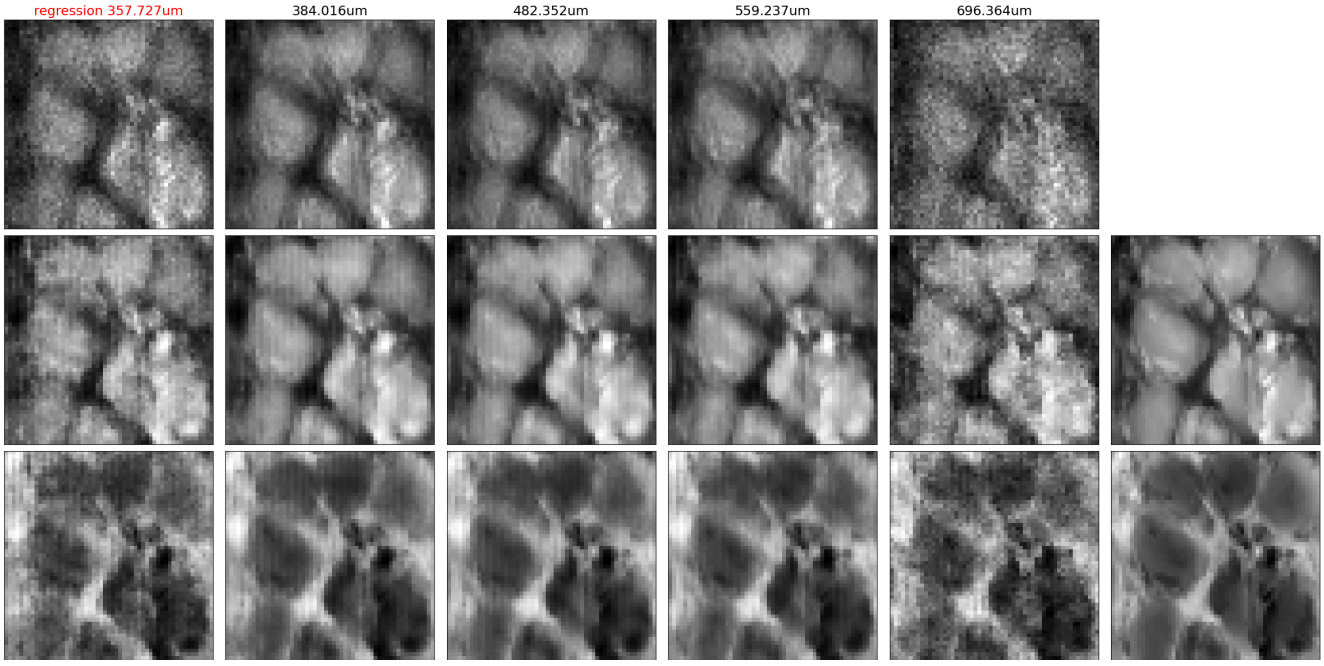


(b)

Fig. 9: The reconstruction and regression quality of Cross-net on external generalization valid set.



(a)



(b)

Fig. 10: The reconstruction and regression quality of eFIN on external generalization valid set.

7. REFERENCES

[1] Abid Haleem, Mohd Javaid, and Ibrahim Haleem Khan, "Holography applications toward medical field: An overview," *Indian Journal of Radiology and Imaging*, vol. 30, no. 03, pp. 354–361, 2020.

[2] Sang-Hyuk Lee and David G Grier, "Holographic microscopy of holographically trapped three-dimensional structures," *Optics Express*, vol. 15, no. 4, pp. 1505–1512, 2007.

[3] Takanori Nomura, Enrique Tajahuerce, Osamu Matoba, and Bahram Javidi, "Applications of digital holography for information security," *Optical and Digital Techniques for Information Security*, pp. 241–269, 2005.

[4] Alexander L Timofeev, Albert Kh Sultanov, and Pavel E Filatov, "Holographic method for storage of digital information," in *Optical Technologies for Telecommunications 2019*. SPIE, 2020, vol. 11516, pp. 14–20.

[5] Gaurav Dwivedi, Lavlesh Pensia, Viney Lohchab, and Raj Kumar, "Non-destructive inspection and quantification of soldering defects in pcb using an autofocusing digital holographic camera," *IEEE Transactions on Instrumentation and Measurement*, 2023.

[6] Tatiana Latychevskaia and Hans-Werner Fink, "Practical algorithms for simulation and reconstruction of digital in-line holograms," *Applied optics*, vol. 54, no. 9, pp. 2424–2434, 2015.

[7] Andreas Erik Gejl Madsen, Mohammad Aryaee Panah, Peter Emil Larsen, Frank Nielsen, and Jesper Glückstad, "On-axis digital holographic microscopy: Current trends and algorithms," *Optics Communications*, p. 129458, 2023.

[8] Etienne Cuche, Pierre Marquet, and Christian Depeursinge, "Spatial filtering for zero-order and twin-image elimination in digital off-axis holography," *Applied optics*, vol. 39, no. 23, pp. 4070–4075, 2000.

[9] Tatiana Latychevskaia and Hans-Werner Fink, "Solution to the twin image problem in holography," *Physical review letters*, vol. 98, no. 23, pp. 233901, 2007.

[10] Jhony Luiz de Almeida, Eros Comunello, Antonio Sobieranski, Anita Maria da Rocha Fernandes, and Gabriel Schade Cardoso, "Twin-image suppression in digital in-line holography based on wave-front filtering," *Pattern Analysis and Applications*, vol. 24, no. 3, pp. 907–914, 2021.

[11] Yaniv Romano, Michael Elad, and Peyman Milanfar, "The little engine that could: Regularization by denoising (red)," *SIAM Journal on Imaging Sciences*, vol. 10, no. 4, pp. 1804–1844, 2017.

[12] Benjamin D Haeffele, Richard Stahl, Geert Vanmeerbeeck, and René Vidal, "Efficient reconstruction of holographic lens-free images by sparse phase recovery," in *Medical Image Computing and Computer-Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part II 20*. Springer, 2017, pp. 109–117.

[13] Hao Wang, Meng Lyu, and Guohai Situ, "eholonet: a learning-based end-to-end approach for in-line digital holographic reconstruction," *Optics express*, vol. 26, no. 18, pp. 22603–22614, 2018.

[14] Luzhe Huang, Tairan Liu, Xilin Yang, Yi Luo, Yair Rivenson, and Aydogan Ozcan, "Holographic image reconstruction with phase recovery and autofocusing using recurrent neural networks," *ACS Photonics*, vol. 8, no. 6, pp. 1763–1774, 2021.

[15] Xiwen Chen, Hao Wang, Abolfazl Razi, Michael Kozicki, and Christopher Mann, "Dh-gan: a physics-driven untrained generative adversarial network for holographic imaging," *Optics Express*, vol. 31, no. 6, pp. 10114–10135, 2023.

[16] Ross Wightman, "Pytorch image models," `https://github.com/rwightman/pytorch-image-models`, 2019.

[17] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16000–16009.

[18] Zhenbo Ren, Zhimin Xu, and Edmund Y Lam, "End-to-end deep learning framework for digital holographic reconstruction," *Advanced Photonics*, vol. 1, no. 1, pp. 016004–016004, 2019.

[19] Yufeng Wu, Jiachen Wu, Shangzhong Jin, Liangcai Cao, and Guofan Jin, "Dense-u-net: dense encoder–decoder network for holographic imaging of 3d particle fields," *Optics Communications*, vol. 493, pp. 126970, 2021.

[20] Hanlong Chen, Luzhe Huang, Tairan Liu, and Aydogan Ozcan, "Fourier imager network (fin): A deep neural network for hologram reconstruction with superior external generalization," *Light: Science & Applications*, vol. 11, no. 1, pp. 254, 2022.

[21] Hanlong Chen, Luzhe Huang, Tairan Liu, and Aydogan Ozcan, "efin: Enhanced fourier imager network for generalizable autofocusing and pixel super-resolution in holographic imaging," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 29, no. 4: Biophotonics, pp. 1–10, 2023.

[22] Asher Trockman and J Zico Kolter, "Patches are all you need?," *arXiv preprint arXiv:2201.09792*, 2022.