
FAST AUTOFOCUSING USING TINY TRANSFORMER NETWORKS FOR DIGITAL HOLOGRAPHIC MICROSCOPY

**Stéphane Cuenat, Louis Andréoli, Antoine N.André, Patrick Sandoz,
Guillaume J. Laurent, Raphaël Couturier and Maxime Jacquot,**

Institut FEMTO-ST,
CNRS & Université Bourgogne Franche-Comté
France
stephane.cuenat@univ-fcomte.fr

ABSTRACT

The numerical wavefront backpropagation principle of digital holography confers unique extended focus capabilities, without mechanical displacements along z-axis. However, the determination of the correct focusing distance is a non-trivial and time consuming issue. A deep learning (DL) solution is proposed to cast the autofocusing as a regression problem and tested over both experimental and simulated holograms. Single wavelength digital holograms were recorded by a Digital Holographic Microscope (DHM) with a 10x microscope objective from a patterned target moving in 3D over an axial range of 92 μm . Tiny DL models are proposed and compared such as a tiny Vision Transformer (TViT), tiny VGG16 (TVGG) and a tiny Swin-Transformer (TSwinT). The proposed tiny networks are compared with their original versions (ViT/B16, VGG16 and Swin-Transformer Tiny) and the main neural networks used in digital holography such as LeNet and AlexNet. The experiments show that the predicted focusing distance Z_R^{Pred} is accurately inferred with an accuracy of 1.2 μm in average in comparison with the DHM depth of field of 15 μm . Numerical simulations show that all tiny models give the Z_R^{Pred} with an error below 0.3 μm . Such a prospect would significantly improve the current capabilities of computer vision position sensing in applications such as 3D microscopy for life sciences or micro-robotics. Moreover, all models reach an inference time on CPU, inferior to 25 ms per inference. In terms of occlusions, TViT based on its Transformer architecture is the most robust.

Keywords ViT · CNN · Tiny Networks · Digital Holographic Microscopy

1 Introduction

One major drawback when 3D moving samples are studied in microscopy is the balance between the focal range that limits out-of-plane measurements and the requirement of a high axial resolution, i.e. a short depth of field (DoF) (see for example [1, 2]). Several solutions have been proposed such as depth-from-focus imaging [3] and confocal microscopy [4] to reconstruct a topography of the scene. Scanning electron microscopy [5] can also be used to get large in-focus depths. In any case, all these methods require a scanning of the scene that slows down the image acquisition rate. Moreover, the working distances of these devices are very short and this reduces considerably the interest of a contactless measurement.

Coherent imaging approaches such as Digital Holography (DH) can be used instead of conventional microscopy to address the focusing issues [6, 7]. DH offers a means for recording the phase and amplitude of a propagating wavefront on a solid-state image sensor [8]. Then, by numerically propagating the recorded wavefront backward or forward at particular distances of interest, different characteristics can be extracted, typically three-dimensional surfaces, but also polarization states and intensity distributions. Several recording and processing schemes have been developed to assess diverse optical characteristics that make DH a highly powerful coherent imaging method for metrological applications [9, 10]. Significant progress, potential impact and challenging issues in the field of DH can be found in this recent roadmap article [11].

The targeted application aims to address the 3D position measurement needs encountered in small-scale mechatronics [12, 13]. At the microscale, automation involves centimeter-sized actuators necessary to perform diverse tasks with a high accuracy, down to the nanometer range. Contactless sensors are thus desired to control 3D motions with a high accuracy over large ranges [12]. Combined with optical microscopy, computer vision constitutes an efficient means to detect in-plane position and displacements. However, microscope objective (MO) lenses provide short in-focus depths and inherently rely on mechanical displacements along the optical axis. One way to extend computer vision capabilities to 3D microscopy is to harness the wave character of light by means of DH. DH is particularly suited to this aim because it requires a single hologram to reconstruct a 3D scene and because digital back-propagation computations allows image reconstruction in extended in-focus depths [14]. A key-point in DH is to note that, instead of an image of the object, it is the propagating wavefront incident on the image sensor that is recorded. The distance of the object does not impact the recording quality, it only changes the actually recorded wavefront in accordance with scalar diffraction theory. Therefore, blur does not exist at the recording stage of DH since it does not seek for any in-focus image. The object distance stands for a computation parameter that is numerically tunable over an extended range, but limited to coherent length of the light source used. This specificity makes the range of working distances allowed by DH incomparably larger than that allowed by conventional incoherent imaging methods. DH can be applied to micro-objects in microscopy with a Digital Holographic Microscope (DHM) setup. The reconstruction distance leading to the best-focused image has to be determined among the z-range explored by the object. There are various techniques for defining image-formation sharpness-criteria that apply to DH [15, 16].

Recently, many studies have proposed to study the capabilities of deep-learning Convolutional Neural Network (CNN) to determine in DH various unknown parameters such as focusing distance, or the phase recovery [17, 18, 19, 20, 21, 22]. These works have to be considered in the wider context of imaging techniques where Deep Learning (DL) approaches are applied to solve complex problems found in computer vision as well as in microscopy [23, 24, 25]. A recent work [17] even demonstrated that Deep CNN gives better results in terms of prediction of propagation distance in DH without knowing all the setup's physical parameters, than other learning-based algorithms such as Multi Layer Perceptron (MLP) [26], support vector machine [27], and k-nearest neighbor [28]. The hardware implementation of artificial neural networks has constituted a real challenge for many years [29, 30, 31], but the tasks that can be solved by such systems are limited to standard tests of classification and prediction, and they are still limited in terms of scalability for mega-pixel image processing.

This paper aims to illustrate a new high-profile application of machine learning by elevating DHM and autofocusing to a new level. Whereas many studies focus on life science microscopy [18, 32, 22], this work explores extended visual capabilities offered by combining DH and last generation of DL algorithms such as Vision Transformer (ViT)[33] and Swin-Transformer (SwinT)[34] networks for applications in micro-robotics [12, 13] or in real-time 3D microscopy [32]. This work introduces for the first time the neural network Transformer architectures applied to advanced coherent imaging field, such as digital holography. This is significant because these new generations of algorithms have already revolutionized the Natural Language Processing (NLP) and recent versions ViT [33] and SwinT[35] highly perform for image recognition thanks to their self-attention feature[36]. More specifically, our work deals with these new generation of deep learning approaches for autofocusing in digital holographic microscopy to obtain in-focus depth prediction with high accuracy. We developed new tiny ViT and tiny SwinT network architectures, and compared them with typical Convolutional Neural Network (CNN) ones used in optics and digital holography such as AlexNet[21], VGG[37] and LeNet[38]. Swin-Transformers propose a hierarchical Transformer whose representation is computed with Shifted windows. The shifted windowing scheme brings greater efficiency by limiting self-attention computation to non-overlapping local windows while also allowing cross-window connection. This hierarchical architecture has the flexibility to model at various scales and has linear computational complexity with respect to the size of images. Taking into account the demand for real time application and achievable training with a reasonable amount of data, tiny networks are developed. These first results pave the way to overcome in-focus depth limit[18] with a short DoF in DHM without any MO lens mechanical displacements[32].

2 New trends in deep learning for image processing: ViT and SwinT

Since the inception of DL neural networks, CNN occupies the field with architectures like VGG-16 [37], Densenet [39] or EfficientNet [40]. At the core of a CNN, there is a series of mixed convolution and pooling layers which extract a set of features from an input image. One of the main advantages of a CNN compared to MLP Network (first network proposed), is that they are translation invariant and less demanding in resource when it comes to large inputs. Later the ViT architecture was introduced. This architecture is based on the concept of attention [36]. The attention mechanism was born to help memorize long source sentences in NLP. Rather than building a single context vector out of the encoder's last hidden state, the attention creates shortcuts between the context vector and the entire source input. The weights of these shortcut connections are customizable for each output element.

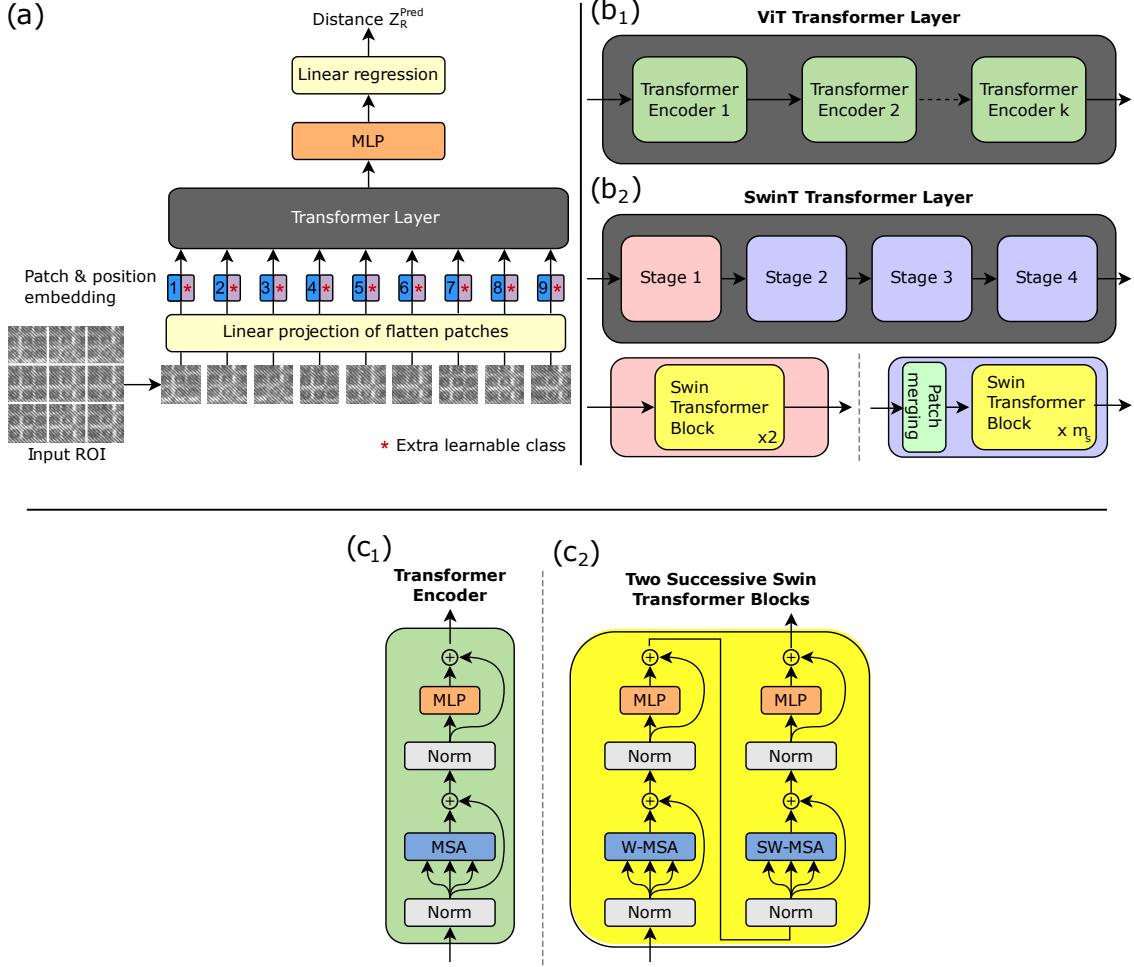


Figure 1: Schematic illustration of the ViT [33] and SwinT [34] architectures. (a) gives the general architecture of a ViT or a SwinT. The region of interest (ROI) is divided into patches which are linearly projected on an embedded dimension followed by a transformer layer, MLP and the linear regression. (b1) shows the successive layers for a ViT respectively k Transformer Encoders. (b2) gives a view of the SwinT Transfomer layer architecture which is formed by a series of 4 Stages. Each Stage encapsulates a Swin-Transformer Block which is repeated m_s times (with s representing the stage number) and patch merging layers (for the stages 2 to 4). Through the stages, the multi-head attention is computed taking different sizes of patch size and shifted windows. (c1) and (c2) represent the internal steps inside a Transformer Encoder and the two successive Swin-Transformer Block, where the Multi-Self Attention (MSA) is computed (MSA, W-MSA and SW-MSA). SwinT first computes a window multi-head attention (W-MSA) and then a shifted windows multi-head attention (SW-MSA).

ViT brings this concept to computer vision [33]. ViT is a pure Transformer architecture built from Transformer encoder layers to approach a classification or regression problem. ViT splits an input image in a series of patches which would be treated as word tokens by a Transformer Network. SwinT even surpasses the performance of a pure ViT network as shown in [35]. SwinT extends a ViT network by varying the patch dimension and computes the attention only for a given window (shifted-window over the space of the input image). Such a network is able to better assess the local and global information inside the input image.

Figure 1 shows the architecture of a ViT [33] and SwinT [34] network. Panel in Fig. 1(a) gives the global architecture of a ViT or SwinT network and in particular how the Region Of Interest (ROI) is processed. This input image is split into different patches and projected on an embedded dimension through the linear projection of the flatten patches (tokens). An additional position embedding and class token are added. The class token is the only token used to apply a regression. Each embedded patch is processed by the transformer layer which outputs the associated class token after the MLP block. The regression layer projects the class token to a scalar, the reconstruction distance Z in our case.

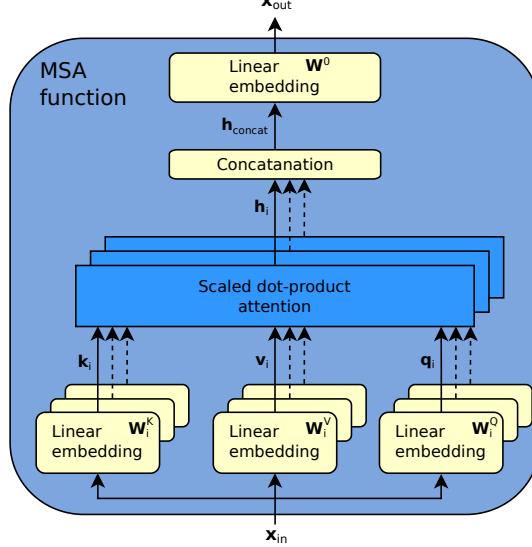


Figure 2: Multi-Headed Self-Attention implemented as a scaled dot-product attention. MSA, W-MSA and SW-MSA blocks on Fig. 1.

Panels in Fig. 1 (b_1) and (b_2) describe in more details the Transformer Layer (in gray) for ViT and SwinT models, respectively. For ViT, there is a total of k Transformer Encoder layers. In the SwinT architecture, the Transformer Layer is composed of a series of s stage layers with typically $s = 4$. The first stage layer contains a two Swin-Transformer block. The $s - 1$ other stages encapsulate a patch merging and m_s Swin-Transformer Blocks, where m_s can change through the stages s (typically $m_s \in [1, 4]$). The window size is set to a fix value (default: 7x7 patches). Moreover, the patch size of each stage is increased by a factor 2 through the patch merging layers [34]. This creates a hierarchy in comparison to a ViT which always considers the same patch size and a global window [33]. The Transformer Encoder (in green) of the ViT and the Swin Transformer block (in yellow) are explained in more details in panels of Fig. 1(c_1) and (c_2), respectively. The input of the transformer encoder is first normalized. The Multi-head Self-Attention (MSA) is first computed and then followed by a normalization and a MLP block. In the case of a SwinT, the architecture is similar but with two successive Swin-Transformer blocks where the MSA is first a Window Multi-head Self-Attention (W-MSA), then a Shifted Window Multi-head Self-Attention (SW-MSA). While the MSA (ViT case) is computed on the complete set of patches, the W-MSA (SW-MSA) uses a dedicated (shifted) window (SwinT case). SW-MSA allows inter-window interactions.

2.1 Multi-head self-attention

As schematically illustrated in Fig. 2, the MSA function is approached from a general perspective where the vector x_{in} is its input and x_{out} its output. For each $i \in [1, N]$, the head \mathbf{h}_i is implemented as a scaled dot-product attention. The N heads \mathbf{h} is called Multi-head Self-Attention. In case of ViT, N is fixed for all the Transformer Encoder layers. SwinT defines a N for each Stage. A key \mathbf{k}_i , value \mathbf{v}_i and query \mathbf{q}_i dimensional vectors are computed for each head \mathbf{h}_i by projecting the input \mathbf{x}_{in} using three learnable matrices ($\mathbf{W}_i^K, \mathbf{W}_i^V, \mathbf{W}_i^Q$):

$$\mathbf{k}_i = \mathbf{W}_i^K \mathbf{x}_{in}, \quad (1)$$

$$\mathbf{v}_i = \mathbf{W}_i^V \mathbf{x}_{in}, \quad (2)$$

$$\mathbf{q}_i = \mathbf{W}_i^Q \mathbf{x}_{in}. \quad (3)$$

For each head \mathbf{h}_i , the attention is computed by taking the key, value and query vectors.

$$\mathbf{h}_i = \text{Attention}(\mathbf{k}_i, \mathbf{v}_i, \mathbf{q}_i). \quad (4)$$

The attention is calculated by first applying a Softmax [41] used to normalize the dot product between a vector of keys \mathbf{k}_i and a vector of queries \mathbf{q}_i . Subsequently, this output acts as weights for the value vector \mathbf{v}_i , hence

$$\text{Attention}(\mathbf{k}_i, \mathbf{v}_i, \mathbf{q}_i) = \text{Softmax}\left(\frac{\mathbf{q}_i \mathbf{k}_i^T}{\sqrt{D}}\right) \mathbf{v}_i, \quad (5)$$

where D is the dimension of the key and query vectors (\mathbf{k}_i and \mathbf{q}_i). All the heads \mathbf{h}_i are concatenated as

$$\mathbf{h}_{\text{concat}} = \text{Concat}(\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_N). \quad (6)$$

The output vector x_{out} is obtained by the vector product of $\mathbf{h}_{\text{concat}}$ and a learnable matrix \mathbf{W}^0 as

$$\mathbf{x}_{\text{out}} = \mathbf{W}^0 \mathbf{h}_{\text{concat}}. \quad (7)$$

Multi-head attention is used since it allows the network to attend to different learned representations at different regions of input ROI as described by Fig. 2 and expressed as

$$\mathbf{x}_{\text{out}} = \text{MSA}(\mathbf{x}_{\text{in}}). \quad (8)$$

2.2 Tiny networks: TViT, TSwinT & TVGG

In this paper, tiny versions of the original ViT (TViT), Swin-Transformer (TSwinT) and VGG16 (TVGG) are proposed. Tiny networks allow to reduce the number of parameters without impacting much the accuracy of the models. Moreover, tiny models need less computation power and approach real-time processing. Figure 1 gives an overview of a ViT and SwinT architectures. TViT modifies a ViT as follows: is built with a total of 12 Transformer Encoder, 8 heads and a patch size of 16x16. The hidden-size of the Transformer encoder has been reduced from 768 (for a ViT/B16) to 128 (TViT) hidden neurons. The MLP dimension (Transformer Encoder) has been reduced to 1024 instead of 3072 hidden neurons (for a ViT/B16). TSwinT is a revisit of the SwinT architecture where several changes have been applied: the size of the embedding vector is set to 32, the number of Swin-Transformer blocks has been set as follows for each Stage: $m_1 = 2, m_2 = 2, m_3 = 4, m_4 = 2$. The number of heads for each Stage has also been modified: $N_1 = 2, N_2 = 4, N_3 = 8, N_4 = 8$. The window size has been fixed to 4 with an initial patch size of 4x4. TViT and TSwinT contrasts with canonical ViT architectures as these models are usually able to learn high-quality intermediate representations with large amounts of data as described in [22, 42]. TVGG is introduced to reduce the number of parameters of the original VGG16 [37] architecture for comparison purposes. All filters of each 2D convolution layer have been divided by 2 inside a TVGG architecture. These changes limit the width of the layers of the tiny networks by keeping their capacity to learn. The number of parameters for each model has drastically diminished as shown in table 1, by a factor between 5 and 20. Moreover, all models are trained from scratch only using experimental or simulated digital holograms of different patterns (pseudo-periodic pattern and USAF pattern) without any transfer learning from a pre-trained model on a dataset like ImageNet [43]. The tiny models take as input ROI of 128x128 of a single wavelength digital hologram.

Tiny model		Original model (pre-trained)		
Model	# parameters	Model	# parameters	Ref
TVGG	$3 \cdot 10^6$	VGG16	$14 \cdot 10^6$	[37]
TViT	$4 \cdot 10^6$	ViT-B16	$85 \cdot 10^6$	[33]
TSwinT	$2.7 \cdot 10^6$	SwinT Tiny	$28 \cdot 10^6$	[34]

Table 1: Number of parameters for each tiny neural network compared to the original version.

3 Applications to digital holographic microscopy

3.1 Experimental setup and targeted pose measurement application

Our final goal is to achieve 3D position and displacement measurements by means of DH combined with computer vision and thus to perform simultaneous high accuracy in-plane and out-of-plane measurements. For that purpose, a periodically micro-structured pattern is used in order to allow unambiguous in-plane position detection through absolute phase computations [13]. Using conventional computer vision, a 10^8 range-to-resolution ratio was demonstrated through robust phase-based decoding [13, 44]. However, this 2D measurement method also works with DH [45, 46]. In order to apply that kind of micro-structured pattern to out-of-plane motion, a DHM is used. This paper explores if DL, and more particularly tiny networks, are able to determine the correct focusing distance with high accuracy and robustness to speed-up the pattern intensity and phase reconstructions and to provide a more accurate Z-position estimation along an extended longitudinal direction close to $100 \mu\text{m}$.

In practice, experiments were carried out on an antivibration table with a DHM (by Lyncee Tec, Switzerland) equipped with a camera with a $5.86 \mu\text{m}$ pixel size (Basler acA1920-155um), a hexapode (Newport HXP50-meca) capable of precise motions along the six degrees of freedom and a micro-encoded pattern made in our clean room facility ($2 \times 2 \text{ cm}^2$, period $9 \mu\text{m}$, 12 bits encoding [13]) covered with a uniform 100 nm thick aluminium layer to obtain

a phase object. This pseudo periodic pattern was observed with a MO (Leica, mag 10x, NA=0.32) at wavelength $\lambda = 674.99$ nm. The light source consists of a superluminescent diode equipped with an interference filter whose width is of 5 nm at half maximum, leading to a coherence of about 100 μm . The sample was shifted along the Z direction by steps of $\sim 1 \mu\text{m}$ and on a total height of $\sim 92 \mu\text{m}$. At each Z step, a series of 400 holograms (1024×1024 px, 8 bits) was recorded with random displacements along the lateral X and Y directions and random planar angles between ± 8 degrees. In total the experimental dataset contains 40,040 holograms.

3.2 Autofocusing in digital holographic microscopy

The advantage of DH is to provide at different reconstruction distances Z_R the complex field diffracted over a distance Z_H from the hologram plane. The hologram propagation Z_H can be tuned over an extended range of up to 92 μm in our DHM setup (limited by the light source coherence length). Figures 3(a₁), (b₁) and (c₁) show an experimental hologram of 1024x1024 pixels propagated at three different distances, 65 μm , 115 μm and 157 μm , respectively. The insets are a zoom of the same sub-area of 128x128 pixels. Over this large propagation range, the holograms recorded are entirely different. Among the different reconstructed planes, the reconstruction distance $Z_R = Z_H$ corresponds to the image in focus. Panels in Fig. (a₂), (b₂) and (c₂) are respectively the amplitude image reconstruction of the holograms panels in Fig. (a₁), (b₁) and (c₁) with a back propagation distance $Z_R = Z_H$. The reconstruction is based on a plane waves angular spectrum method [47]. Except for illumination variation along the recording distance range of 92 μm , the three reconstructed images are highly similar.

To find the axial position of an object, the challenge is therefore to find this distance Z_R . Autofocusing techniques in DH are applied considering the modulus of the reconstructed complex field or the modulus of spatial spectrum of the propagated field [15]. The sharpness of the image can be determined from multiple focusing criteria such as the sum of the modulus of the complex field amplitude, the use of a logarithmically weighted Fourier spectral function, the variance of gray value distribution, focus measure based on autocorrelation, absolute gradient operator, Laplace filtering, Tamura coefficient estimation, or wavelet-based approaches [15]. A comparison of many focusing criteria in terms of computational time and accuracy in determining the focal plane have been already discussed [15, 16].

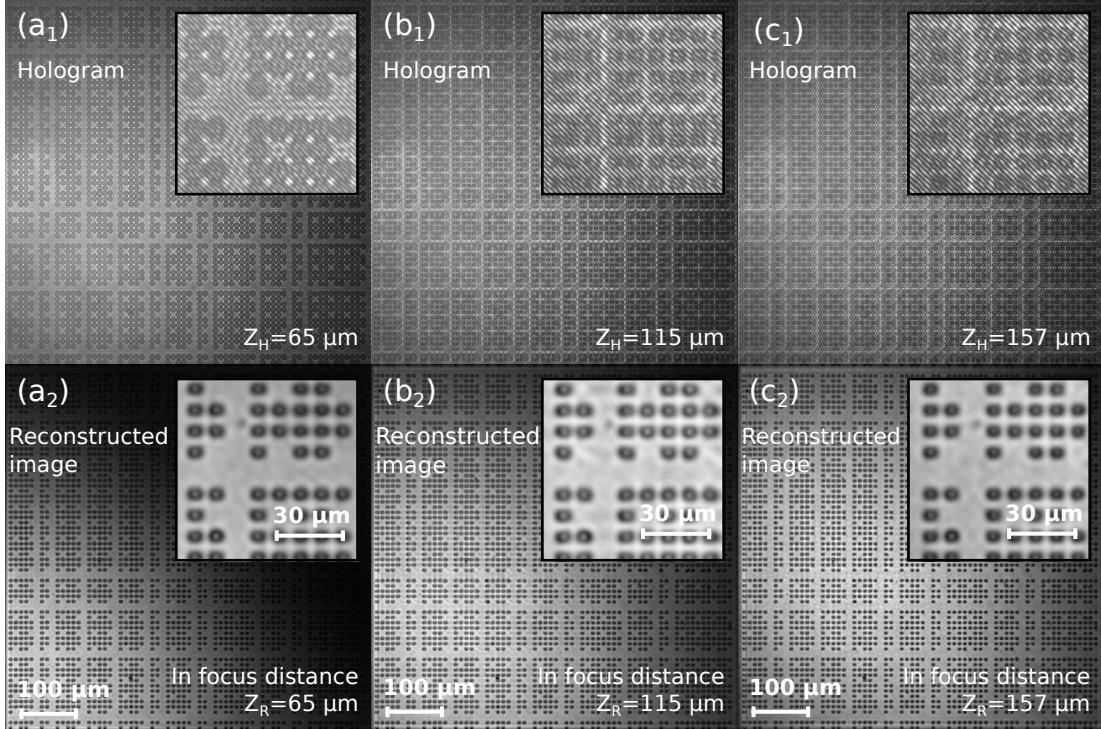


Figure 3: (a₁), (b₁) and (c₁), experimental holograms (1024x1024 pixels) of the same area of a pseudo-periodic pattern corresponding to a propagating distance Z_H of 65 μm , 115 μm and 157 μm , respectively. (a₂), (b₂) and (c₂), amplitude image reconstruction at a distance $Z_H = Z_R$, respectively. The insets are zooms of the same sub-area of 128x128 pixels.

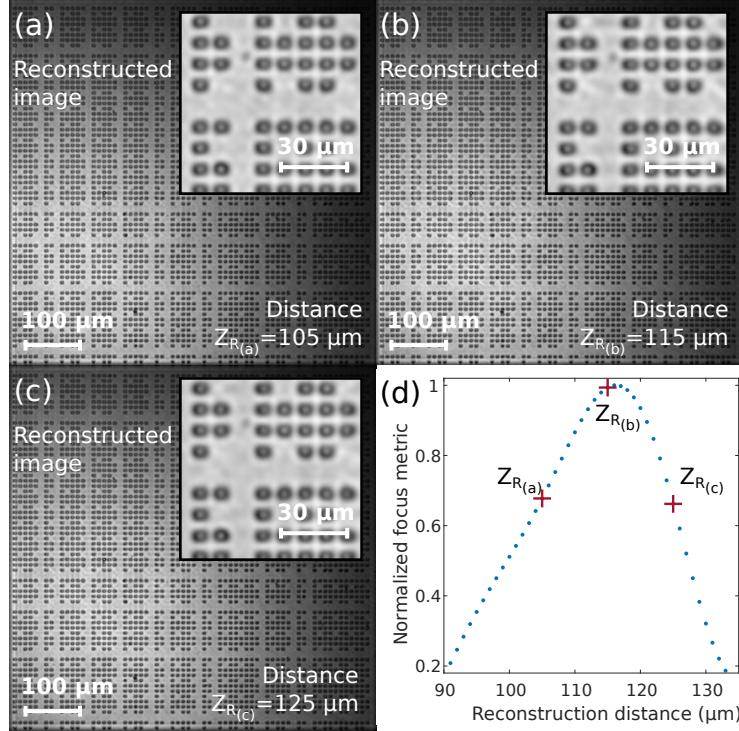


Figure 4: (a), (b) and (c), amplitude image reconstruction of the hologram of Fig. 3 (b₁) at distances Z_R of 105 μm , 115 μm and 125 μm , respectively. (d), focus estimation function calculated from the intensity image Laplacian (LAP) as described in [15]. The red crosses correspond to the distance reconstruction $Z_{R(a)}$, $Z_{R(b)}$ and $Z_{R(c)}$ of the panel (a), (b) and (c), respectively.

When the focus criterion is at an extremum, the focus of the reconstructed image is optimal. There are also various methods that would allow an automated determination of the optimal reconstruction distance [48]. However, all these approaches require the numerical reconstruction of a set of images within a given range of propagation distances. Then, the focus criteria is calculated from each reconstructed image, to determine the distance of focusing. Figures 4(a), (b) and (c) show three reconstructed images at different distance Z_R of the experimental hologram of the Fig. 3(b₁).

These three images spaced by 10 μm from each other illustrate the difficulties of obtaining a sharp focus criteria. Figure 4(d) shows the result of a focus metric based on the image Laplacian [15] where the red crosses correspond to the case of the panels (a), (b) and (c). The resolution of this normalized autofocusing method is close to the DoF provided by the Numerical Aperture (NA) of the MO and the wavelength λ used, which gives in our case $\text{DoF} = 2\lambda/\text{NA}^2 = 15 \mu\text{m}$. Even if these approaches could be efficient [48], they are computationally demanding and time consuming, especially if the size of the hologram is large.

4 Results

This section shows the results obtained by running a series of inferences on test sets of different holograms. All the test results have been generated using the proposed tiny models: TViT, TVGG and TSwinT. Four datasets are considered: experimental and simulated phase holograms of the pseudo-periodic pattern, and amplitude and phase holograms of a simulated USAF pattern. The simulated holograms are generated by using a plane-wave spectrum propagation algorithm. Although the simulation reproduces the experimental parameters of the sample and the imaging system, it is deliberately free of motion uncertainties, surface defects, optical aberration and noise. This approach allows to obtain the intrinsic performance limit of the neural networks proposed. Free of mechanical limitation, the simulated Z pitches are therefore less than 1 μm and over a total range of 100 μm .

For each set of holograms, TViT, TVGG and TSwinT have been trained from scratch. The neural networks have been configured to apply a regression on the input data. A total of 200 epochs have been executed and the learning curves have correctly converged. The learning rate was set to $1 \cdot 10^{-4}$ using the Adam optimizer [49]. During the training, a

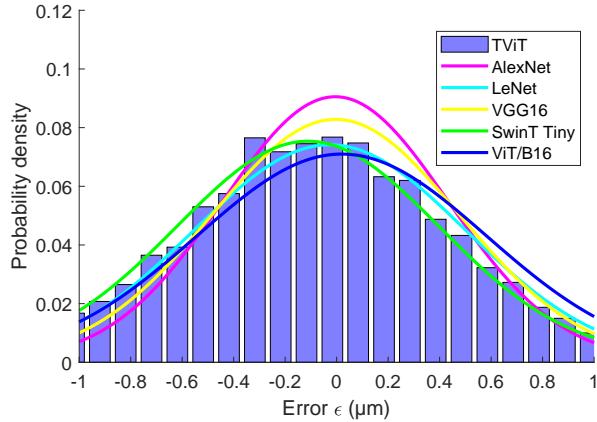


Figure 5: Error distribution comparison between TViT and other state-of-the-art models like AlexNet, LeNet and original models such as VGG16, SwinT Tiny and ViT/B16. $\bar{\epsilon}_{\text{TViT}} = -0.07 \pm 0.61 \mu\text{m}$, $\bar{\epsilon}_{\text{AlexNet}} = 0 \pm 0.51 \mu\text{m}$, $\bar{\epsilon}_{\text{LeNet}} = -0.02 \pm 0.61 \mu\text{m}$, $\bar{\epsilon}_{\text{VGG16}} = 0 \pm 0.56 \mu\text{m}$, $\bar{\epsilon}_{\text{SwinT Tiny}} = -0.17 \pm 0.59 \mu\text{m}$, $\bar{\epsilon}_{\text{ViT/B16}} = 0.01 \pm 0.66 \mu\text{m}$,

total of 64 (TViT) or 32 (TSwinT & TVGG) ROIs are selected randomly for each hologram. As showed in [50], the log(cosh) loss function can improve the result of Varational Auto-Encoder. This loss function

$$L(Z_H, Z^{\text{Pred}}) = \sum_{i=1}^n \log(\cosh(Z_i^{\text{Pred}} - Z_{H_i})), \quad (9)$$

is also less prone to outliers than the mean squared error (MSE) or the mean absolute error (MAE) where n is the number of training samples, Z and Z^{Pred} are the expected and predicted values, respectively.

Table 2 shows the performance of the validation loss functions on our experimental hologram dataset for all tiny networks. Table 3 shows the best validation loss L for each model and each training set. In the following sections, the error ϵ of one inference is measured by,

$$\epsilon = Z_R^{\text{Pred}} - Z_H. \quad (10)$$

All the codes to train the proposed tiny networks (TViT, TVGG and TSwinT) are accessible at this address <https://github.com/scuenat/DHMTinyNetworks> and the data is available in [51].

Model	MSE	MAE	log cosh
TVGG	0.25	0.39	0.11
TViT	0.42	0.50	0.13
TSwinT	0.30	0.43	0.12

Table 2: Comparison of the validation loss functions (log cosh, MSE or MAE) for each proposed tiny models trained on experimental holograms.

Model	Pseudo-periodic pattern		USAF (simulated)	
	Experimental	Simulated	Amplitude	Phase
TVGG	0.11	0.003	0.04	0.05
TViT	0.13	0.004	0.09	0.04
TSwinT	0.12	0.012	0.05	0.07

Table 3: Validation loss (log cosh) for each model and each set of holograms.

4.1 Experimental holograms: pseudo periodic pattern

Holograms of Fig. 3 (a₁), (b₁) and (c₁), recorded at different distances Z_H , are representative specimens of the set of experimental holograms used by the tiny networks during the training phase. The experimental dataset, which contains

a total of 40,400 holograms (400 holograms for each distance Z_H), was distributed between learning, validation and testing sets with a 70/20/10 ratio. The models have therefore been tested on a set of 4,040 holograms, 40 holograms for each Z_H spaced by 1.0 μm ranging on 92 μm . In Fig. 6, the results of the inferences testing of the TViT, TVGG and TSwinT are represented on the panels (a), (b) and (c), respectively. The average and the Full Width at Half Maximum (FWHM) of the error are represented by the color bold curves and areas, respectively. Comparable performances for the three neural networks with a high stability along the full range and small errors can be observed. Figure 9 gives another view allowing to appreciate the error distribution. The solid lines are Gaussian fits of the error for each network model. The average and standard deviation (half of the FWHM) are given for each case. Panel (a) illustrates the results of the experimental dataset. This graph shows an error bounded by 1 μm for all models. Figure 5 compares the error distribution of a TViT model with reference neural networks in digital holography as VGG16, LeNet (as presented in [38]) and AlexNet (as presented in [21]). The original versions, SwinTransformer Tiny (SwinT Tiny) and ViT/B16, are also represented. For the MO used, the 1 μm autofocusing accuracy achieved is 15 times smaller than the theoretical DoF.

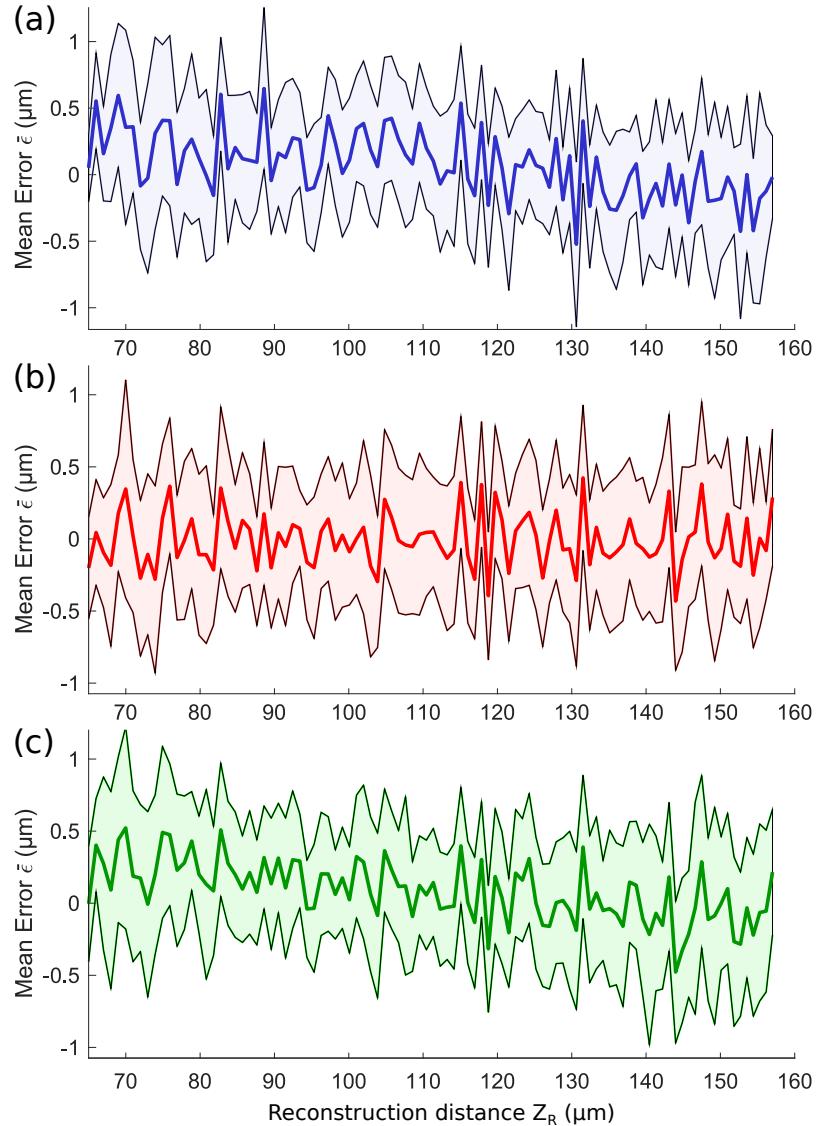


Figure 6: Prediction error on experimental holograms of pseudo-periodic patterns. The bold lines (area) are the average (standard deviation) of the error for each reconstruction distance over 92 μm . (a), (b) and (c) correspond to TViT, TVGG and TSwinT models, respectively.

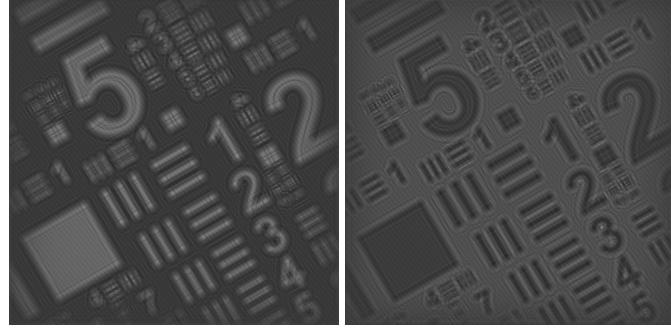


Figure 7: (a) and (b), two representative examples of simulated 1024X1024 USAF hologram in amplitude and phase, respectively.

4.2 Simulated holograms: pseudo periodic pattern

Simulated phase holograms of pseudo-periodic patterns are used to test the limit of the proposed models performances. 40,040 holograms, including 40 different sites (in-plane position and orientation) vertically scanned per steps of $1\text{ }\mu\text{m}$, ranging on $100\text{ }\mu\text{m}$, constitute the full training dataset. The testing dataset is composed of 6,819 holograms never viewed, spaced by $0.1\text{ }\mu\text{m}$ and ranging on $100\text{ }\mu\text{m}$. Figure 9(b) shows the error distribution of the reconstruction distance Z_R for the three models. Without noise and exact Z position labelling, the reconstruction error considerably decreases below $0.3\text{ }\mu\text{m}$. The best models are TViT and TVGG which have a FWHM of the error distribution of $0.160\text{ }\mu\text{m}$. These great performances for a testing with a step size ten time smaller than the training dataset prove the high regression quality of the three tiny networks.

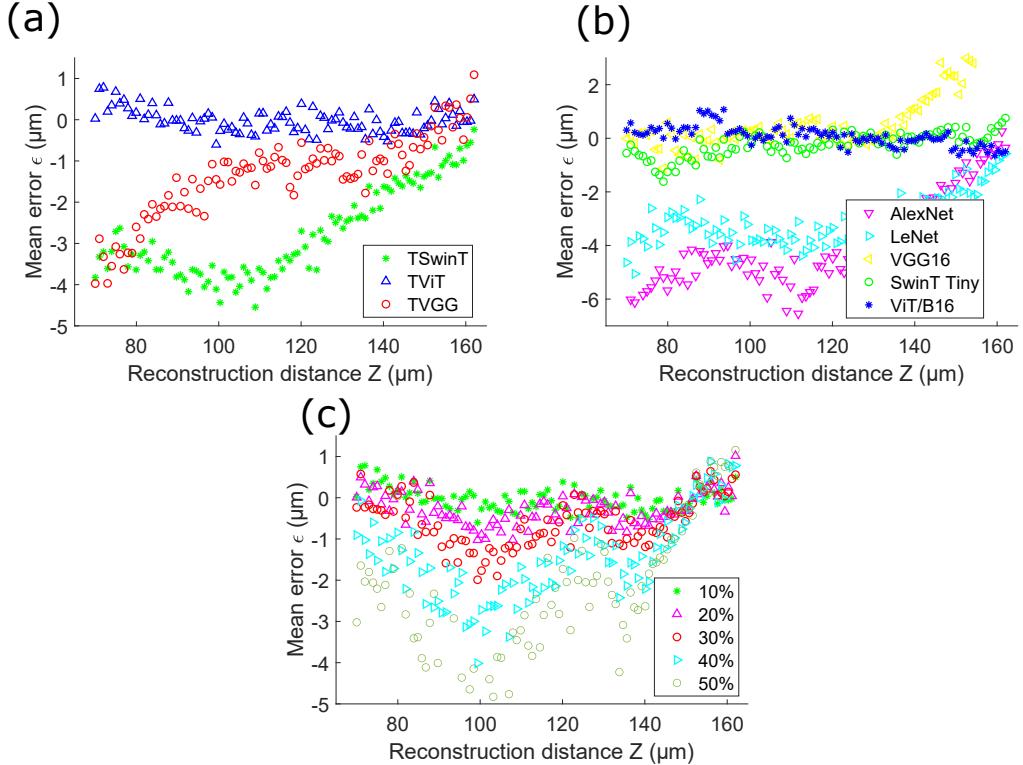


Figure 8: Experimental results for all models in the case where the network input ROI artificially undergoes a loss of information. (a) shows the average error over the entire Z reconstruction range for the proposed tiny models with a loss of 10%, (b) shows the same but for original and reference models and (c) the limit of occlusion for a TViT model

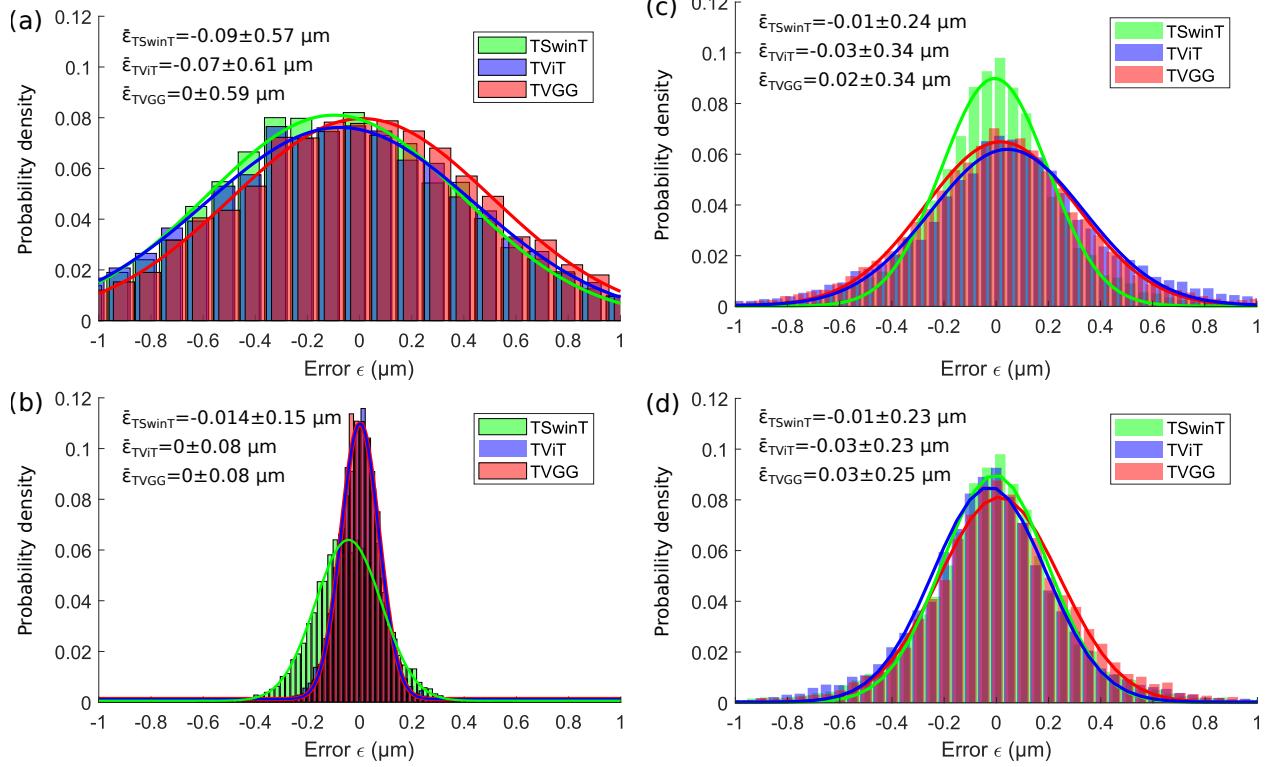


Figure 9: Error distribution of the three neural networks over the reconstruction distance Z . (a) and (b) are the results for the experimental and simulated pseudo-periodic patterns, respectively. (c) and (d) are the results for the simulated USAF patterns in amplitude and phase, respectively.

4.3 Amplitude and phase object: USAF 1951 resolution chart

In comparison with the pseudo-periodic sample, the USAF pattern is a more complex pattern due to a wider spatial frequencies bandwidth. To avoid building a dataset with empty regions without information, the pattern was simulated by filling the space more densely than a commercial target. In order to characterize the tiny networks performances in function of the amplitude or phase nature of the holograms, two different datasets were constituted. Figures 7(a) and (b) show two representative holograms in amplitude and phase, respectively. Both training (testing) datasets are constituted of 400 (50) different holograms at each step of $0.5 \mu\text{m}$ ranging on $130 \mu\text{m}$ ($100 \mu\text{m}$), for a total of 104,400 (10050) holograms. Figures 9(c) and (d) show the error distribution of the tiny networks in case of amplitude and phase objects, respectively. In comparison with the pseudo-periodic pattern, all tiny networks have worse performances but stays highly competitive with a error below $0.35 \mu\text{m}$.

In case of amplitude holograms, the TSwinT model shows better results than the other two. Contrary to the TSwinT model which has the same performances whatever the characteristic of the hologram, the TViT and the TVGG give more precise inferences for the phase holograms. In this case, the three neural networks show similar performances with an error smaller than $0.25 \mu\text{m}$.

4.4 Occlusion test

Another type of neural network performance is its robustness in a degraded configuration. Experimentally we have already seen that the three proposed models are resilient with respect to homogeneous sample illumination, as shown in Fig. 3 and Fig. 4. To go further, areas of the experimental testing dataset have been deleted. A random squared region of 10% of the ROI selected as input for the tiny neural networks is uniformly set to zero. This operation simulates a dust or a sample defect and evaluates the degree of locality of the neural networks. The results obtained are shown in Fig. 8, where panel (a) display the error average along the Z for the proposed tiny models, panel (b) display the error average along the Z for the reference and original models and panel (c) the limit of occlusion in case of a TViT model. It can be observed that the TSwinT and TVGG models are highly impacted by 10% of occlusion as the models, AlexNet and

LeNet. In contrast, TViT is clearly the most robust architecture against occlusion. Figure 8(a) shows that on average the TViT error remains stable on the full 92 μm range.

4.5 Inference speed

Whether for applications in microrobotics or in 3D microscopy for life sciences, there is great interest in being able to work in real time and with commercially accessible equipment. Therefore, Figure 10 shows the comparison of the median speed of 200 inferences with two different configurations for all the neural networks compared (AlexNet, LeNet, VGG16, SwinT Tiny, ViT/B16, TViT, TVGG and TSwinT). On the Intel i9-11900K @3.50GHz CPU the performance is comparable to using a GPU NVidia RTX 3090, 24Gb, with an inference speed below 25 ms for LeNet and TViT. As analyze in details [21], the reconstruction time of an hologram for twenty different distances takes a total of 318 ms on an Intel Core i5 processor. The image reconstruction knowing the predicted distance Z_R^{Pred} is therefore of ~ 15 ms. This value is also confirmed by the DHM which has a reconstruction rate of up to 60 frames per second. The tiny models proposed with low inference times, associated with an image reconstruction algorithm, therefore form a solution compatible with the constraints of real applications.

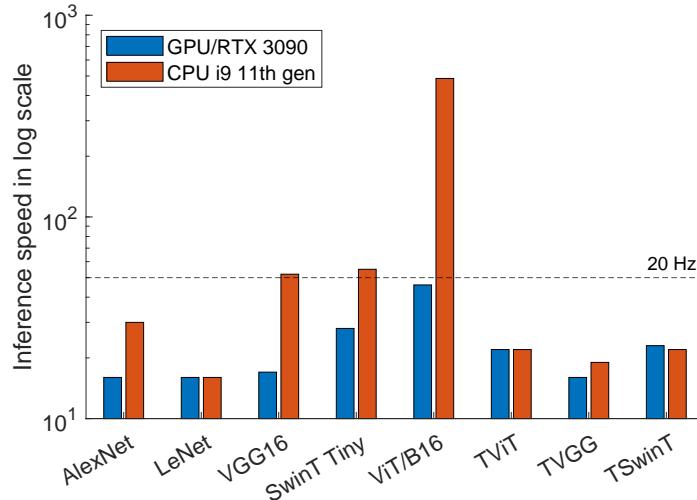


Figure 10: Comparison of the inference speed in log scale for AlexNet, LeNet, VGG16, SwinT Tiny, ViT/B16, TVGG, TSwinT and TViT on different architectures; GPU: RTX 3090 24Gb and CPU: Intel i9-11900K @3.50GHz. The dashed line represents the real-time limit in robotics.

5 Discussion & conclusion

TVGG, TViT and TSwinT give close results when taking the different sets of holograms (pseudo-periodic pattern experimental/simulated and simulated USAF phase/amplitude). In this paper, it has been shown that TViT is more robust in presence of occlusion (Fig. 8), considering that Z depth information is present on the entire hologram due to diffraction properties in coherent imaging. A TViT model benefits from the multi-head self-attention (Fig. 2) which takes the complete ROI at each layer in consideration (not a set of extracted features). A CNN like TVGG works a bit differently as it tries to build a set of features through its first Convolution/Pooling layers followed by full connected layers (regression). A CNN, by extracting at each layer a more complex representation of the features, explains why it focuses on dedicated regions which impacts the accuracy of the inference in presence of occlusion. According to the above, a TViT model seems more suited to the prediction of the in-focus distance in DHM, as it scans everywhere and is more robust in terms of occlusion. This goes in the same direction as presented in [52, 53] where it has been shown that a ViT model is a lot more robust than a CNN. Although, a TSwinT model is based on the derivative of a ViT model, it does not perform as well as a TViT in case of added occlusion. As TSwinT is only applying the self-attention on a set of windows (W-MSA) or shifted windows (SW-MSA) (through its Swin-Transformer Blocks, Fig. 1(c2)), it can be assumed that the occlusion has a bigger impact on the result of the multi-head attention than a pure ViT like TViT.

Figure 5 shows that the tiny models (TVGG, TSwinT and TViT) are as accurate as the original versions (VGG16, SwinT Tiny and ViT/B16). It also shows that a AlexNet or LeNet model reach similar performance. Considering the inference speed on CPU (Figure 10) and the robustness against an occlusion, TViT is the best model proposed.

As mentioned in [42], a ViT (TViT) would need a lot of data to be trained from scratch. This is not what has been experienced, as a TViT can be trained from scratch using our set of experimental holograms of pseudo-periodic pattern using a total of 2,327,040 ROIs (36,360 holograms \times 64 ROI). A huge amount of data is normally needed as a ViT (TViT) projects each patch on an embedded dimension (Fig. 1(a), Patch embedding). The reconstruction distance Z information is spread over the complete space of the hologram, which is most likely an argument to explain why a ViT-like network can learn from scratch without having a huge dataset at disposal.

Our experiments showed that the reconstruction distance Z can be predicted in DHM with a high accuracy using deep learning last generation techniques, especially regression models. An error bounded by $\sim 1 \mu\text{m}$ on the reconstruction distance Z has been reached for a dataset of experimental holograms on a range of $92 \mu\text{m}$. The regression approach allows experimentally to surpass the DoF of the MO by an order of magnitude. Moreover, this error can be lowered down to $\sim 0.3 \mu\text{m}$ when the models are trained on the simulated holograms of a pseudo-periodic pattern or USAF pattern (phase or amplitude). The discrepancy between results obtained from experimental and simulated datasets is partly due to the limited accuracy of the actuator used where bi-directional repeatability is of $0.3 \mu\text{m}$. Acquisition noise may also play a significant role in that reduction of performances obtained from experimental datasets. All proposed tiny models offer an alternative to expensive GPUs as the time for an inference is below the real-time limit in robotics of 20 Hz (Fig. 10), less than 25 ms on an Intel i9.

The ability of tiny networks to determine the in-focus depth with a FWHM of about one micron opens attractive application prospects. Indeed, if two wavelength DHM are considered, the ambiguity range is about twice the FWHM demonstrated in this paper (with our commercial DHM, $\lambda_1 = 674.99 \text{ nm}$ and $\lambda_2 = 793.63 \text{ nm}$; i.e. $\lambda_{eq}/2 = 2.25 \mu\text{m}$). This means that DL may bridge the gap between the MO DoF and the Z-information provided by the interferometric phase to achieve Z-position determination down to the interference sensitivity; i.e. around 1 nm over ranges of tens of microns. Such a prospect would significantly improve the current capabilities of computer vision position sensing applied to 3D microscopy.

6 Acknowledgments

This work was supported by HOLO-CONTROL (ANR-21-CE42-0009), TIRREX (ANR-21-ESRE-0015), SMARTLIGHT (ANR-21-ESRE-0040) by Cross-disciplinary Research (EIPHI) Graduate School (contract ANR-17-EURE-0002), Région Bourgogne Franche-Comté (HoloNET project). This work was performed using HPC resources from GENCI-IDRIS (Grant 20XX-AD011012913) and also the Mésocentre de Franche-Comté.

Disclosures

The authors declare no conflicts of interest.

Data availability

Data underlying the results presented in this paper are available in Zenodo, Ref. [51].

References

- [1] Yu Sun, Stefan Duthaler, and Bradley J Nelson. Autofocusing in computer microscopy: selecting the optimal focus algorithm. *Microscopy research and technique*, 65(3):139–149, 2004.
- [2] Jan-Mark Geusebroek, Frans Cornelissen, Arnold WM Smeulders, and Hugo Geerts. Robust autofocusing in microscopy. *Cytometry: The Journal of the International Society for Analytical Cytology*, 39(1):1–9, 2000.
- [3] Yalin Xiong and Steven A Shafer. Depth from focusing and defocusing. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 68–73. IEEE, 1993.
- [4] Gordon S Kino and Timothy R Corle. *Confocal scanning optical microscopy and related imaging systems*. Academic Press, 1996.
- [5] HJ Leamy. Charge collection scanning electron microscopy. *Journal of Applied Physics*, 53(6):R51–R80, 1982.

- [6] Frank Dubois, Cédric Schockaert, Natacha Callens, and Catherine Yourassowsky. Focus plane detection criteria in digital holography microscopy by amplitude analysis. *Opt. Express*, 14(13):5895–5908, Jun 2006.
- [7] Pietro Ferraro, Simonetta Grilli, Domenico Alfieri, Sergio De Nicola, Andrea Finizio, Giovanni Pierattini, Bahram Javidi, Giuseppe Coppola, and Valerio Striano. Extended focused image in microscopy by digital holography. *Opt. Express*, 13(18):6738–6749, Sep 2005.
- [8] U. Schnars and W. Jüptner. Direct recording of holograms by a ccd target and numerical reconstruction. *Appl. Opt.*, 33(2):179–181, Jan 1994.
- [9] Etienne Cuche, Frédéric Bevilacqua, and Christian Depeursinge. Digital holography for quantitative phase-contrast imaging. *Optics letters*, 24(5):291–293, 1999.
- [10] Maxime Jacquot, Patrick Sandoz, and Gilbert Tribillon. High resolution digital holography. *Optics communications*, 190(1-6):87–94, 2001.
- [11] Bahram Javidi, Artur Carnicer, Arun Anand, George Barbastathis, Wen Chen, Pietro Ferraro, J. W. Goodman, Ryoichi Horisaki, Kedar Khare, Małgorzata Kujawinska, Rainer A. Leitgeb, Pierre Marquet, Takanori Nomura, Aydogan Ozcan, YongKeun Park, Giancarlo Pedrini, Pascal Picart, Joseph Rosen, Genaro Saavedra, Natan T. Shaked, Adrian Stern, Enrique Tajahuerce, Lei Tian, Gordon Wetzstein, and Masahiro Yamaguchi. Roadmap on digital holography. *Opt. Express*, 29(22):35078–35118, Oct 2021.
- [12] Zhuoran Zhang, Xian Wang, Jun Liu, Changsheng Dai, and Yu Sun. Robotic micromanipulation: Fundamentals and applications. *Annual Review of Control, Robotics, and Autonomous Systems*, 2:181–203, 2019.
- [13] Antoine N André, Patrick Sandoz, Benjamin Mauzé, Maxime Jacquot, and Guillaume J Laurent. Sensing one nanometer over ten centimeters: A microencoded target for visual in-plane position measurement. *IEEE/ASME Transactions on Mechatronics*, 25(3):1193–1201, 2020.
- [14] Tristan Colomb, Nicolas Pavillon, Jonas Kühn, Etienne Cuche, Christian Depeursinge, and Yves Emery. Extended depth-of-focus by digital holographic microscopy. *Optics letters*, 35(11):1840–1842, 2010.
- [15] Elsa SR Fonseca, Paulo T Fiadeiro, Manuela Pereira, and António Pinheiro. Comparative analysis of autofocus functions in digital in-line phase-shifting holography. *Applied optics*, 55(27):7663–7674, 2016.
- [16] Patrik Langehanenberg, Björn Kemper, Dieter Dirksen, and Gert Von Bally. Autofocusing in digital holographic phase contrast microscopy on pure phase objects for live cell imaging. *Applied optics*, 47(19):D176–D182, 2008.
- [17] Zhenbo Ren, Zhimin Xu, and Edmund Y Lam. Learning-based nonparametric autofocusing for digital holography. *Optica*, 5(4):337–344, 2018.
- [18] Yair Rivenson, Yibo Zhang, Harun Günaydin, Da Teng, and Aydogan Ozcan. Phase recovery and holographic image reconstruction using deep learning in neural networks. *Light: Science & Applications*, 7(2):17141–17141, 2018.
- [19] Yichen Wu, Yair Rivenson, Yibo Zhang, Zhensong Wei, Harun Günaydin, Xing Lin, and Aydogan Ozcan. Extended depth-of-field in holographic imaging using deep-learning-based autofocusing and phase recovery. *Optica*, 5(6):704–710, 2018.
- [20] Gong Zhang, Tian Guan, Zhiyuan Shen, Xiangnan Wang, Tao Hu, Delai Wang, Yonghong He, and Ni Xie. Fast phase retrieval in off-axis digital holographic microscopy through deep learning. *Optics express*, 26(15):19388–19405, 2018.
- [21] Tomi Pitkäaho, Aki Manninen, and Thomas J Naughton. Focus prediction in digital holographic microscopy using deep convolutional neural networks. *Applied optics*, 58(5):A202–A208, 2019.
- [22] Tianjiao Zeng, Yanmin Zhu, and Edmund Y. Lam. Deep learning for digital holography: a review. *Opt. Express*, 29(24):40572–40593, Nov 2021.
- [23] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.
- [24] Ayan Sinha, Justin Lee, Shuai Li, and George Barbastathis. Lensless computational imaging through deep learning. *Optica*, 4(9):1117–1125, 2017.
- [25] Shaowei Jiang, Jun Liao, Zichao Bian, Kaikai Guo, Yongbing Zhang, and Guoan Zheng. Transform-and multi-domain deep learning for single-frame rapid autofocusing in whole slide imaging. *Biomedical optics express*, 9(4):1601–1612, 2018.
- [26] Simon Haykin. *Neural networks: a comprehensive foundation*. Prentice Hall PTR, 1994.
- [27] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.
- [28] Antonio Mucherino, Petraq J. Papajorgji, and Panos M. Pardalos. *k-Nearest Neighbor Classification*, pages 83–106. Springer New York, New York, NY, 2009.

- [29] Demetri Psaltis, David Brady, Xiang-Guang Gu, and Steven Lin. Holography in artificial neural networks. *Landmark Papers on Photorefractive Nonlinear Optics*, pages 541–546, 1995.
- [30] Laurent Larger, Antonio Baylón-Fuentes, Romain Martinenghi, Vladimir S Udal'tsov, Yanne K Chembo, and Maxime Jacquot. High-speed photonic reservoir computing using a time-delay-based architecture: Million words per second classification. *Physical Review X*, 7(1):011015, 2017.
- [31] Xing Lin, Yair Rivenson, Nezih T Yardimci, Muhammed Veli, Yi Luo, Mona Jarrahi, and Aydogan Ozcan. All-optical machine learning using diffractive deep neural networks. *Science*, 361(6406):1004–1008, 2018.
- [32] Henry Pinkard, Zachary Phillips, Arman Babakhani, Daniel A. Fletcher, and Laura Waller. Deep learning for single-shot autofocus microscopy. *Optica*, 6(6):794–797, Jun 2019.
- [33] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- [34] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*, 2021.
- [35] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. *arXiv preprint arXiv:2201.03545*, 2022.
- [36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [37] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [38] Keyvan Jaferzadeh, Seung-Hyeon Hwang, Inkyu Moon, and Bahram Javidi. No-search focus prediction at the single cell level in digital holographic imaging with deep convolutional neural network. *Biomed. Opt. Express*, 10(8):4276–4289, Aug 2019.
- [39] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.
- [40] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.
- [41] Lester James Miranda. Understanding softmax and the negative log-likelihood". *ljvmiranda921.github.io*, 2017.
- [42] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? *Advances in Neural Information Processing Systems*, 34, 2021.
- [43] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database.
- [44] Antoine N André, Patrick Sandoz, Maxime Jacquot, and Guillaume J Laurent. Robust, precise and scalable: A phase-encoded pattern for visual x , y , θ positioning. In *2020 International Conference on Manipulation, Automation and Robotics at Small Scales (MARSS)*, pages 1–5. IEEE, 2020.
- [45] Patrick Sandoz and Maxime Jacquot. Lensless vision system for in-plane positioning of a patterned plate with subpixel resolution. *JOSA A*, 28(12):2494–2500, 2011.
- [46] Miguel Asmad Vergara, Maxime Jacquot, Guillaume J Laurent, and Patrick Sandoz. Digital holography as computer vision position sensor with an extended range of working distances. *Sensors*, 18(7):2005, 2018.
- [47] J.W. Goodman. *Introduction to Fourier Optics*. Electrical Engineering Series. McGraw-Hill, 1996.
- [48] Mert Doğar, Hazar A İlhan, and Meriç Özcan. Real-time, auto-focusing digital holographic microscope using graphics processors. *Review of Scientific Instruments*, 84(8):083704, 2013.
- [49] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [50] Pengfei Chen, Guangyong Chen, and Shengyu Zhang. Log hyperbolic cosine loss improves variational auto-encoder. <https://openreview.net/forum?id=rkgIvsC9Ym>, 2019.
- [51] Experiment dataset (pseudo periodic pattern), 2022. Available at <https://zenodo.org/record/6337535>.

- [52] Muzammal Naseer, Kanchana Ranasinghe, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Intriguing properties of vision transformers. *Advances in Neural Information Processing Systems*, 34, 2021.
- [53] Stéphane Cuenat and Raphaël Couturier. Convolutional neural network (cnn) vs vision transformer (vit) for digital holography. *arXiv preprint arXiv:2108.09147*, 2021.