euspen's 25th International Conference & Exhibition, Zaragoza, ES, June 2025
www.euspen.eu

Video-Rate 3D pose measurement with sub-micrometer accuracy using deep learning and digital holography for robotic applications

Stéphane Cuenat[1], Jesús E. Brito Carcaño[2] , Patrick Sandoz[2] , Guillaume J. Laurent[2] , Raphaël Couturier[1] , Maxime Jacquot[2]

[1]*Université Marie et Louis Pasteur, CNRS, Institut FEMTO-ST, F-90000 Belfort, France*
[2]*Université Marie et Louis Pasteur, SUPMICROTECH, CNRS, Institut FEMTO-ST, F-25000 Besançon, France*

stephane.cuenat@univ-fcomte.fr

Abstract
This work implements a fast and efficient autofocusing method for 3D pose estimation using deep learning and digital holographic microscopy. The proposed approach achieves video-rate tracking of moving targets with sub-micrometere accuracy, with emphasis on determining the in-focus distance via a simplified Gedanken model, TinyGedanken, reaching inference times of ~1 ms using simulated digital holograms. By leveraging deep learning models, such as GendankenNet model, the method accelerates traditional holographic reconstruction to directly infer the in-focus distance Z from raw holographic dataset. This approach significantly improves processing speed, making it ideal for pose measurements for microrobotic applications, such as actuator characterization, micro-assembly and biomedical manipulation.

Digital Holography, Deep Learning, Autofocusing method, Object Tracking at the micro-scale, Video-rate 3D Localization, Convolutional Neural Networks (CNN), GedankenNet, Microrobotics, 3D Metrology, Object Trajectory Detection

1. Introduction

In computer micro-vision and micro-robotics, precise 3D positioning and trajectory determination are critical for various industrial and clinical applications [1]. Deep Neural networks (DNNs), particularly Convolutional Neural Networks (CNNs) and Vision Transformer (ViT) models, play a central role in processing visual data [2]. In microscopy, digital holography (DH) gives access to both wavefield diffracted by an object in both amplitude and phase, from a single image and over extended Z ranges without mechanical displacements. However, in-focus images are reconstructed numerically, from an off-axis or in-line configuration, through time-consuming digital image processing. By integrating DNNs, especially the GedankenNet model [3] alongside other CNNs models like UNet [4], with DH, this approach provides a fastest solution for accurately tracking object trajectories in automated microscopy under video-rate constraints [5]. The paper focuses on a tiny version of a GedankenNet model (TinyGedanken) and highlights its unique advantages in processing efficiency and inference speed in the realm of computer micro-vision and micro-robotics. Advanced micro-assembly platforms in robotics require precise translation and rotation stages (Fig. 1(a)) to handle tasks with nanoscale positioning accuracy and large-scale movements. This work aims to achieve 3D inference and video-rate pose estimation and automated microscopy for applications such as 3D MEMS micro-nano-assembly, alignment, 3D nanoprinting, and visual servoing for nanopositioning [1].

2. Context and Background

2.1. Deep neural networks

DNNs, inspired by biological neural systems, process and analyze complex data using multiple layers, after a training step based on input-output paired dataset. This data-driven approach allows to use non-linear transformations from input to output, enabling tasks like linearization in higher-dimensional spaces [4]. Training DNNs involves optimizing the network with input-output data pairs, and sufficient training data is essential for optimal performance. Notably, CNNs and ViTs have shown high effectiveness in tasks such as image classification, computer vision, and complex problems like autofocusing [2] and phase retrieval [6] in digital holography. This paper presents a modified version of the Gedanken neural network [3], utilizing spectral layers to accelerate inference for determining the autofocus distance Z in digital holographic microscopy.

2.2. Digital holographic microscopy and computer micro-vision for micro-robotics

Digital Holography (DH) captures both the amplitude and phase of an object's wavefield using a CMOS sensor. Combined with a 2D pseudo-periodic pattern (PPP) used as a phase object, it allows sub-voxel 3D pose measurement through micro-vision phase correlation computations [7, 8, 9]. Figure 1 summarizes major steps of the method; a Lyncee-Tec Digital Holographic Microscope (DHM) with a 10× microscope lens follows the displacements of a PPP placed on a precision robot (a), a typical recorded hologram (b), and the reconstructed intensity (c) and phase (d) of the object [2]. Digital hologram reconstruction uses the Angular Spectrum Method [10] and applies this approach to micro-objects (see [2] for details). DHM supports digital autofocusing, enabling automated microscopy and 3D pose estimation of micro-objects. Recent studies show that DNNs can accelerate autofocusing in DHM by treating it as a classification or regression task [5]. Key challenges include enhancing multiscale sensitivity for 6 degrees of freedom (DoF) pose estimation while maintaining a wide field of view and depth of field [1]. We explore this approach on our 2D pseudo-periodic pattern used as reference sample (Fig. 1(b), (c) and (d)). The latter is made of an altered periodic frame providing two complementary information sources. On the one hand, the pattern periodicity ensures redundancy and allows sharp spatial frequency filtering to remove noise and achieve high resolution. On the other hand, an absolute binary code is encrypted in the missing dot distribution to remove 2π ambiguities and allow the exact X, Y, α localization of the current view within the whole centimeter-sized encoded area.

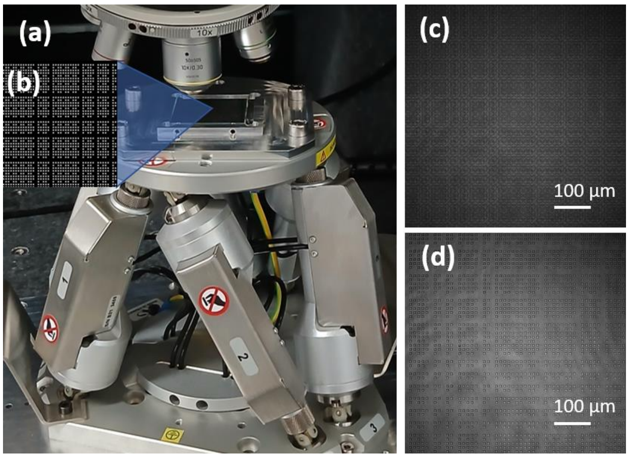

Figure 1. a) Lyncee-tec DHM observing a typical pseudo-periodical pattern (b), which is a micro-structured plate moved in 3D by a hexapod stage. (c) A typical experimental hologram of the pseudo-periodic pattern that allows 3D pose measurement [2]. (d) In-focus image in amplitude reconstructed by the Angular Spectrum Method at Z = 185 µm.

3. Autofocusing in digital holography

In automated microscopy and micro-robotics, precise 3D positioning, especially in the Z-direction, is critical for controlling micro-object trajectories. Traditional autofocusing methods rely on mechanical adjustments along the optical axis, which are slow and inefficient, especially for video-rate applications.

Digital holography (DH) offers a solution by capturing both amplitude and phase of the object's wavefront, allowing computational refocusing and 3D reconstruction without mechanical movement. However, determining the correct Z-position from the captured data remains a challenge, as traditional methods of numerical backpropagation and sharpness criteria can be slow, uncertain and computationally expensive [2].

To address this, deep learning models like CNNs or GedankenNet can be used to speed up autofocusing in DH. By leveraging spectral layers, GedankenNet can predict the optimal focusing distance directly from the holographic image, bypassing traditional multi-reconstruction methods. This enables faster, more accurate determination of the in focus distance, essential for video-rate 3D pose control in applications like micro-assembly and nano-positioning.

4. TinyGedanken Model

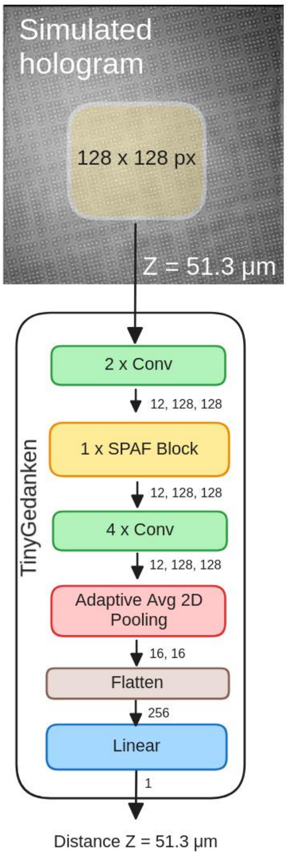

Figure 2. TinyGedanken model based on the original GedankenNet model [3]. The model takes 128x128 pixel images as input, with a single SPAF block. The SPAF block width is reduced to 12 instead of the original 48. All the changes results in a total of 440,436 parameters. The final layer is adapted for regression to predict the autofocus distance Z.

Figure 2 illustrates the TinyGedankenNet model based on the vanilla version proposed by [3]. In this version, the input images are reduced to 128x128 pixels, randomly cropped from the hologram space. The model features a single spatial Fourier transformation (SPAF) block, down from the original eight layers. The SPAF block has been adjusted to use Hadamard multiplication, replacing the complex multiplication. Additionally, the width of the SPAF block has been reduced from 48 to 12, bringing the total number of parameters from 39,476,004 to 440,436. The final layers have been modified to adapt for a regression. The output of the last convolution layer (12, 128, 128) is reduced to a 2D tensor (16, 16) by applying a 2D adaptive average pooling layer. Finally, the 2D tensor is flatten and linearized to predict the distance Z.

5. Results

The model has been trained only using 15,235 simulated holograms (taken from a complete set of 65,665 simulated holograms ranging along a distance Z of 185µm), with 13,713 used for training and 1,522 for validation. Each epoch processed 500 holograms from the training set and was validated on 16 holograms from the validation set (following a few-shot learning approach). A total of 5,000 epochs were run. The log cosh function was used as the loss function, and the Adam optimizer with a weight decay of 10^{-4} was employed.

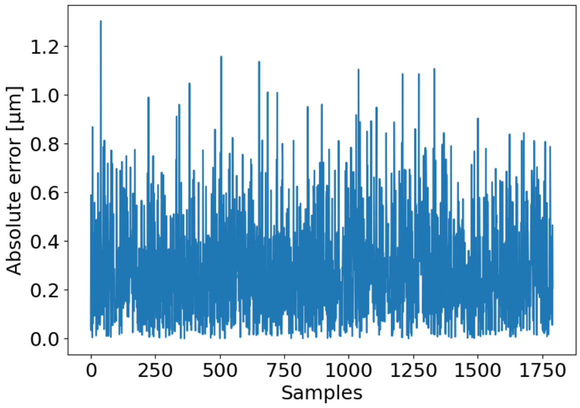

Figure 3. Absolute error for a set of 1791 test holograms unseen during training ranging on a total distance of 140 µm using a TinyGedanken model (single SPAF block). With a mean absolute error of 280 nm and a max absolute error of 1.3 µm.

Figure 3 shows the inference results for a set of 1791 holograms unseen during training ranging on a total distance of 140 µm along Z-axis. The absolute error is limited by 1.3 µm (with a mean error of 280 nm).

Table 1 shows the error on the detected distance Z varying the number of SPAF blocks inside the TinyGedanken model. The single SPAF block does not impact much the accuracy of the model, the mean absolute error is stable with a slight impact on the max absolute error.

Table 1 Comparison of the accuracy varying the number of SPAF blocks of the TinyGedanken.

Models	**Number of SPAF blocks**	**Z error**	
		Mean absolute error	**Max absolute error**
TinyGedanken	4	290 nm	970 nm
TinyGedanken	2	275 nm	718 nm
TinyGedanken	**1**	**280 nm**	**1.3 µm**

Figure 4 shows the inference speed on different GPUs like a Nvidia A100 and V100. The inference speed is below 2 ms and close to 1 ms for the A100 (considering that the transfer of an image to the GPU takes about 0.08 ms for a NVidia A100). This contrasts with the speeds presented in [2] and [11], respectively ~20 ms and 9 ms (using a CNN, MobileNetV3).

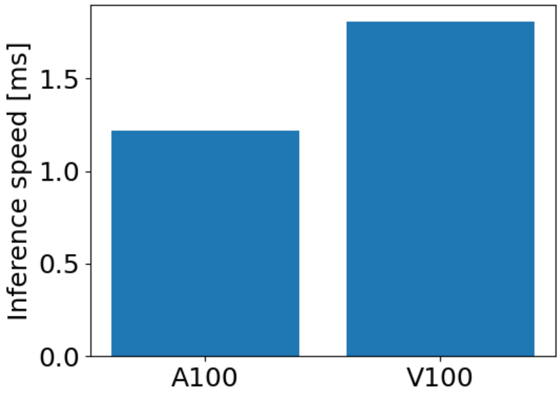

Figure 4. Inference speed on a Nvidia A100 or V100, ~1.25 ms per inference on a A100 using a TinyGedanken model (single SPAF block).

Table 2 provides a comparative analysis of inference speeds for the GedankenNet model and the variation of the TinyGedanken model, focusing on the impact of reducing the number of SPAF blocks. The table highlights inference times on two high-performance GPUs, the NVIDIA A100 and V100. The original GedankenNet model, which incorporates 8 SPAF blocks, serves as the baseline. It achieves inference speeds of 7.65 ms on the A100 GPU and 13.70 ms on the V100. This comparison underscores a clear trade-off between model complexity and computational efficiency. Notably, the A100 GPU consistently outperforms the V100 across all configurations, demonstrating its superior processing capabilities. These results highlight the potential of model simplification to achieve faster inference, particularly in resource-constrained or latency-sensitive applications.

Table 2 Comparison of inference speeds varying the number of SPAF blocks of the TinyGedanken with orginal GedankenNet model as baseline.

Models	**Number of SPAF blocks**	**Inference speed**	
		NVidia A100	**NVidia V100**
GedankenNet Model (512x512 input images)	8	7.65 ms	13.70 ms
TinyGedanken	4	3.68 ms	6.79 ms
TinyGedanken	2	1.93 ms	3.51 ms
TinyGedanken	**1**	**1.25 ms**	**1.81 ms**

6. Conclusion

The use of TinyGedanken model (with a single SPAF block) has proven to be highly effective in achieving fast autofocusing for 3D pose tracking in digital holography. With a significantly reduced number of layers and parameters, this modified model reaches an impressive inference time of just ~1.25 ms on an Nvidia A100 GPU. Despite the reduction in complexity, the accuracy of the in-focus prediction remains on par with results presented in [2], demonstrating that the streamlined architecture does not compromise performance. This rapid inference time and high accuracy make the TinyGedanken model a promising solution for video-rate applications in automated microscopy and micro-robotics, offering both efficiency and precision in 3D positioning tasks.

7. Prospects

Figure 5 demonstrates how the X and Y coordinates could be retrieved from the hologram space, with a focus on predicting and reconstructing a trajectory, for instance a Lissajous trajectory (Fig 5. (c)).

The input hologram, shown in Fig. 5(a), is a 768 x 768 pixels image processed by the model to produce a 64 x 64 pixels binary thumbnail (Fig. 5(b)) using a UNet-like architecture [5, 12]. A post-processing algorithm is then applied to the reconstructed thumbnail to extract binary vectors representing the X and Y positions. These binary vectors are part of a complete sequence of 4096 bits encoding the positions along X and Y [8].

An hologram represents a small area of the total encoded surface area of 11 x 11 cm², representing the full range of X and Y positions.

To derive the micron-scale coordinates, each vector's index in the sequence is identified. The position of each vector, multiplied by the distance per bit (27 µm), determines the final X and Y coordinates (Fig. 5(c)).

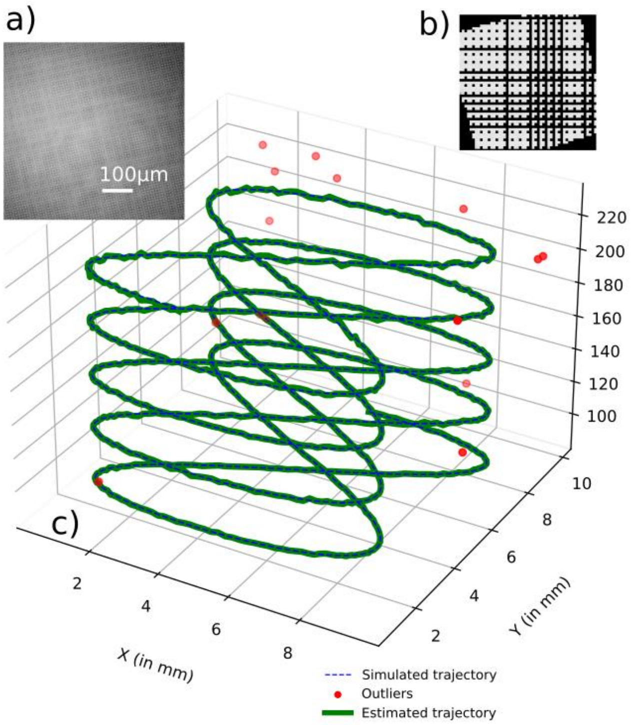

Figure 5. a) input hologram to the XY model as proposed in [5], (b) the output of the model, a binary thumbnail representing the target, (c) the computed trajectory from the thumbnails with outliers in red.

Funding

This work was supported by Agence Nationale de la Recherche, HOLO-CONTROL project (Contract No. ANR-21-CE42-0009), by the French Investissements d'Avenir program, the ROBOTEX Network and Its FEMTO-ST Technological Facility CMNR under Grant ANR-21-ESRE-0015, SMARTLIGHT project (ANR-21-ESRE-0040) and by Cross-disciplinary Research (EIPHI) Graduate School (contract No. ANR-17-EURE-0002). This work was performed using HPC resources from GENCIIDRIS (Grant 20XX-AD011012913R3) and the Mésocentre de Franche-Comté.

References

[1] Yao S., Li H., Pang S., Zhu B., Zhang X., Fatikow S. (2021) IEEE Trans. Instrum. Meas. 70, 1–28.
[2] Cuenat S., Andréoli L., André A.N., Sandoz P., Laurent G.J., Couturier R., Jacquot M. (2022) Opt. Express **30**, 14
[3] Huang L., Chen H., Liu T., et al. (2023), Nat. Mach. Intell. **5**, 895–907.
[4] Ronneberger O., Fischer P., Brox T. (2015) arXiv. 1505.04597.
[5] Stéphane Cuenat, Jesús E. Brito Carcaño, et al.J. Eur. Opt. Society-Rapid Publ., **20** 2 (2024) 31
[6] G. Zhang, T. Guan, Z. Shen, X. Wang, T. Hu, D. Wang, Y. He, and N. Xie, Opt. Express 26(15), 19388–19405 (2018).
[7] André A.N., Sandoz P., Mauzé B., Jacquot M., Laurent G.J. (2022) Int. J. Comput. Vis. **130**, 6.
[8] André A.N., Sandoz P., Mauzé B., Jacquot M., Laurent G.J. (2020) IEEE/ASME Trans. Mech. **25**, 1193–1201.
[9] Ahmad B., Sandoz P., Laurent G.J. (2024) *IEEE Trans. Instr. Meas.* doi: 10.1109/TIM.2024.3523360.
[10] J. W. Goodman, Introduction to Fourier Optics, 3rd ed. (Roberts & Company, 2004).[11] Liao J, Chen X, Ding G, Dong P, Ye H, Wang H, Zhang Y, Yao J., Biomed Opt Express. 2021 Dec 14;13(1):314-327.
[11] Liao J, Chen X, Ding G, Dong P, Ye H, Wang H, Zhang Y, Yao J., Biomed Opt Express. 2021 Dec 14;13(1):314-327.
[12] Ronneberger O., Fischer P., Brox T. (2015) arXiv. 1505.04597.