



Automatic multifocusing in digital holographic microscopy

DÁNIEL TERBE,¹  LÁSZLÓ ORZÓ,¹ AND ÁKOS ZARÁNDY^{1,2,*}

¹HUN-REN Institute for Computer Science and Control (HUN-REN SZTAKI), 1111 Budapest, Hungary

²Pázmány Péter Catholic University, 1083 Budapest, Hungary

*zarandy@sztaki.hu

Abstract: Holography captures the full three-dimensional optical field of a sample in a single intensity measurement. This enables numerical refocusing to arbitrary axial planes and, in principle, automatic multifocusing in digital holographic microscopy (DHM). Existing work on holographic autofocusing, however, largely targets the simpler single-plane setting and typically relies either on computationally expensive depth sweeps with hand-crafted sharpness metrics or on deep-learning models trained for a single focal distance per hologram. We address the more practical and challenging problem of automatic multifocusing for in-line flow-through DHM, where heterogeneous microscopic objects are distributed across multiple axial depths within a single hologram. We propose a two-stage pipeline for real-world DHM applications. First, a coarse, low-resolution 3D search detects object locations and provides approximate lateral and axial positions using a robust focus metric on downsampled holograms. Second, a deep-learning-based refinement stage processes the resulting object ROIs and predicts the residual defocus to obtain precise axial positions. Within this framework, we compare regression and classification formulations, introduce classification models with non-uniform depth bins that concentrate capacity near focus, and examine recursive refinement. Evaluated on a practical and diverse dataset of living micro-organisms, the proposed system achieves accurate, data-efficient, robust, and computationally efficient multifocusing, making it suitable for quasi-real-time DHM applications.

© 2026 Optica Publishing Group under the terms of the [Optica Open Access Publishing Agreement](#)

1. Introduction

Holography, originally proposed by Gabor [1], records the interference between a reference wave and an object wave originating from the same coherent source. In digital holography [2], this interference pattern is captured by an electronic sensor, enabling numerical reconstruction of the complex optical field using computational propagation models such as the angular spectrum (AS) method [3]. A widely used implementation—also adopted in this study—is the in-line Digital Holographic Microscope (DHM), in which unscattered light acts as the reference beam while the scattered component encodes object information. The simplicity and compactness of the in-line geometry make it highly suitable for field-based and deployable microscopy [4–9], although its design inherently produces spatially overlapping twin-image artifacts that degrade reconstruction quality.

Because image sensors measure only intensity, the recorded hologram does not directly contain the phase of the object wavefront. The resulting ambiguity—manifesting as real and virtual (twin) images—necessitates computational phase retrieval or twin-image removal. These inverse methods, extensively studied over several decades [10–38], aim to recover the lost phase and thereby reconstruct the full complex wavefield.

A hologram encodes the three-dimensional optical field of a sample, allowing the reconstruction of arbitrary axial planes by numerically propagating the complex wavefront [2]. When the position of the sample (or the positions of objects within a sparse liquid sample) relative to the image sensor is unknown, a practical challenge arises: determining the axial location of the sample (or

the locations of the micro-objects within the sample). This task is known as digital holographic automatic focusing or in short holographic autofocus. Accurately identifying the focal plane is crucial, as all downstream tasks—such as classification [39], segmentation [40], phase recovery, and domain transformation [41,42]—depend on precise focus localization. However, compared to phase-recovery research (discussed above), the literature on digital holographic autofocus is limited and largely focuses on scenarios where a single focal distance is to be determined [43–53]—typically using prepared or sectioned biological specimens confined to a single plane; microspheres or microbeads similarly restricted to one focal layer; standard optical test targets such as the USAF 1951 resolution chart; or synthetically generated holograms. Despite its high practical relevance, only a few studies [54–57] have addressed the more complex case of simultaneously focusing multiple objects within a single hologram. This setting is common in flow-through DHM, where microscopic biological objects are freely dispersed across multiple axial depths.

The main question this work addresses is: how can we automatically identify the focal plane of multiple objects of interest within the sample volume? We call this problem **automatic holographic multifocusing**.

Our approach is motivated by a practical application and evaluated on a representative real-world dataset. We introduce a two-stage multifocusing pipeline. First, coarse object-level localization is performed using a conventional search algorithm on downsampled holograms, producing approximately focused regions of interest (ROIs). Second, deep neural network models refine the focus estimation of each ROI, enabling accurate reconstruction of objects at their respective axial positions. We evaluate existing state-of-the-art hologram focusing networks and further propose models tailored specifically to this multifocusing setting, including iterative refinement strategies.

The main contributions of this paper can be summarized as follows:

1. **A two-stage, deployment-oriented autofocus pipeline for in-line flow-through DHM.** We combine a fast, conventional coarse search algorithm (which detects objects present in the volume) with proposed deep network focus refiners (FFocusRNet/FFocusCNet) that output relative (signed) defocus estimates. The design maintains accuracy and speed for freely dispersed micro-objects, covering six distinct morphologies, and generalizes well to unseen objects.
2. **Unified comparison of regression, classification, and recursive refinement.** In one framework, we benchmark (i) single-pass regression (FocusNet [43]/FocusNetV2 [53]/FFocusRNet), (ii) single-pass classification (FFocusCNet), and (iii) multi-pass recursive inference for each of the previous models.
3. **Class geometry for practical focusing.** We propose non-equidistant depth bins for classification to improve close-to-focus performance. We demonstrate that the proposed class geometry enables recursive refinement (in inference time), whereas recursion is less advantageous for pure regressors.
4. **Label-efficient training at realistic scale.** Our system achieves the above without requiring massive curated datasets; we rely on modest recordings and physics-consistent augmentation, enabling replication in typical lab settings.

The remainder of this paper is structured as follows. In the next section, we review related work. The Methods section then details the experimental setup, including the microscope parameters, the coarse prefocusing algorithm, the construction of the supervised learning dataset, the neural network models, and the evaluation protocols. This is followed by a presentation and discussion of our results. Finally, we conclude with a summary of our main findings.

2. Related works

The conventional solutions [47,54,55,58] reconstruct the hologram at a series of candidate depths and then selects the best-focused slice according to a hand-crafted sharpness (or focus) metric. Although conceptually simple, this brute-force strategy has three well-recognized limitations: 1. reconstructing the hologram at multiple different depths is computationally expensive and time-consuming; 2. focus precision can be improved with finer axial sampling, only at the expense of increasing the computational load; and 3. performance relies on the chosen heuristic focus metric, which may require fine-tuning to the specific specimen type, signal-to-noise ratio, or imaging conditions. These caveats motivated the development of more efficient and adaptive autofocus strategies, including optimization-guided searches [48] and, more recently, deep-learning models [43–45,53,59] that predict the in-focus depth directly from the raw hologram.

Langenhanen et al. [54] were among the first to address holographic multifocusing, i.e., the problem of automatically determining multiple focal planes within a single hologram containing several objects at different depths. Their method extended conventional single-plane autofocus principles to handle multiple depth layers simultaneously. Specifically, the hologram was numerically reconstructed across a range of axial distances using an angular spectrum propagation model, and local focus measures were evaluated on selected regions of interest (ROIs) across the field of view, allowing spatially separated regions to yield distinct focal depths. However, this approach necessitated computationally intensive reconstructions across numerous depths, with its accuracy highly dependent on the selected focus metric, and relied on a manual ROI selection process for the objects.

Building upon this direction, Trusiak et al. [55] introduced a more robust focus metric—referred to as DarkFocus—capable of detecting multiple focal planes by analyzing the statistical properties of the reconstructed intensity distribution. While the method improves robustness to noise and sample variations, it remains practically limited to detecting two focal planes and retains many of the inherent drawbacks of classical approaches, such as the need for repeated reconstructions at multiple depths and the trade-off between accuracy and computational cost.

Lang et al. [56] propose a segmentation-free automatic multifocus method based on connected-domain analysis (CDA) that reconstructs the hologram over a swept axial range and, at each depth, applies background removal and morphology-driven preprocessing before extracting connected components and linking them across z via centroid proximity; a per-object quality score Q (area plus a weighted squared semi-major axis) yields a focus curve whose peak defines each object's focal distance. On plankton and seawater holograms acquired with a 30 mm depth of field, CDA outperforms classical criteria (e.g., LAP, Tenengrad, DCT/FFT, wavelet) in SSIM and step-error analyses. A practical drawback is computational cost: the pipeline still requires dense z -stacks (e.g., 240 reconstructions at 0.25 mm steps per hologram) and per-depth connected-component analysis, which is time-consuming and may limit real-time, high-throughput deployment.

Learning-based methods cast holographic autofocus as either *classification* over discretized axial bins or *regression* to a continuous defocus value. Early convolutional neural networks (CNNs) by Pitkäaho et al. [44,45] demonstrated both formulations directly from raw (or lightly preprocessed) holograms, avoiding exhaustive depth sweeps at test time. Shimobaba et al. [51] likewise trained a compact CNN to predict the best-focus depth from a single hologram. The strongest single-object-to-estimate (SOTE) baselines are represented by the FocusNet family introduced by Montoya et al. [43]. *FocusNet* formulates focusing as a regression task to infer axial distance directly from raw holograms in a single pass. Its enhanced successor, *FocusNetV2* [53], incorporates architectural improvements and employs a hybrid loss function for greater robustness to outliers.

Conventional methods perform autofocus by reconstructing the entire hologram at a series of depths and evaluating hand-crafted sharpness metrics, resulting in a simple but computationally intensive process that is sensitive to the chosen metric and often limited in precision. In our

system, the initial 3D search is conducted on a downscaled version of the hologram, greatly increasing computational efficiency and enhancing robustness to metric choice, albeit at a cost of reduced axial resolution. This coarse stage rapidly and reliably localizes objects and estimates their approximate depths.

On the other hand, existing deep learning (DL) approaches typically target fast single-plane focusing, operating on full-resolution holograms and requiring large, curated datasets. These constraints make them less practical for real-world multifocusing, especially when samples contain many objects at different depths. Our two-stage system uniquely combines the strengths of both worlds: the conventional, efficient coarse search provides reliable object localization, while deep learning is selectively applied only in the second stage to small, coarsely prefocused ROIs. This approach sharply reduces data and compute demands, improves robustness, and enables accurate, efficient multifocusing for realistic DHM scenarios with diverse, sparsely distributed objects—representing a substantial advantage over previous single-stage strategies.

3. Methods

3.1. Microscope setup and measurement

Our setup consists of a microscope objective (20x, NA = 0.45) combined with additional lens forms an afocal optical system that ensures uniform magnification ($\approx 7\times$) and sub-micron (0.57 μm) resolution within the observed volume. Here we use a CMOS (Basler daA3840-45 μm) camera with 4 μm binned pixels, while a 405 nm diode laser provides the coherent illumination. The uniformity and parallelism of the illuminating laser wavefront is achieved through its free-space propagation. The size of the recorded holograms is 1930×1024 pixels.

We used an IBIDI (μ -Slide I Luer) flow-through cell with a channel height of 200 μm . The liquid sample is intermittently pumped through this cell. During acquisition, the pump (and flow) is stopped, and a short laser impulse ensures that we capture a still image. The depth of field of our DHM setup—which defines the axial step size—is 3.43 μm , and the observed volume spans the full thickness of 0–200 μm . The field of view is approximately 1 mm^2 , corresponding to an observed volume of approximately 0.12 μL . Throughout this study, unless stated otherwise, all axial distances are reported in the magnified space; accordingly, we use an **axial step size of 168 μm** and a total axial range of 10 mm.

To reconstruct the objects present in a measured hologram, their exact spatial positions and sizes must be determined, and for each object, the precise focus has to be identified. We solve this object-segmentation problem using a simple algorithm. We reconstruct a downsampled version of the hologram at multiple propagation distances and compare the resulting reconstructions. The sharpness of the actual reconstructions can be estimated by the distribution of some focus metric [58,47,60]. Using the local extrema of the focus metric distributions, the algorithm determines where and at what distance the objects are located. In the case of low resolutions—where the (effective) depth of field is larger—these focus metrics can provide a good estimate of the approximate axial position, spatial location, and size of the objects. Such descriptors have often been used previously [54], however at higher resolutions, due to the mixed phase and amplitude distributions of the objects, they can no longer determine the focus accurately [61]. For this reason, we introduce a two-stage approach: first, a conventional low-resolution 3D search identifies and localizes the objects; then, a dedicated deep learning method is applied to each selected object region (ROI) to accurately refine its focus.

An additional advantage of this two-step focus determination approach, besides enabling hologram segmentation, is that it makes the computations more efficient. First, the objects position and focus are searched for in a reduced-resolution image using relatively large steps. Then, by propagating the entire hologram to the estimated reconstruction distance, the information related to the object becomes concentrated in a much smaller area. At this stage, it is sufficient to search for the exact focus position by considering only a smaller neighborhood around the

object. In this way, any errors or inaccuracies in the initial estimation can be corrected during the subsequent deep-learning-based focus refinement process.

We note that downsampling confines the sample density at which the system can operate effectively. The full raw hologram should contain no more than 50 large objects (20–50 μm diameter); otherwise, they overlap heavily, and the system cannot distinguish individual objects.

Due to the reduced resolution, the depth of field increases, meaning that a larger axial step sizes can be employed during the search. The localization of individual objects and the estimation of their focus are carried out using a specific focus metric—similar to previously proposed focus measures [58,47,60]—derived from the distribution of a local sharpness descriptor.

In our case—where the goal is to detect small objects exhibiting both phase and amplitude modulation—the applied focus measure depends on the local edge density as well as the magnitude of the reconstructed amplitude. It is important to note that this focus measure contributes not only to the determination of the focal planes but also to locating the objects within the hologram. The smoothness of the resulting distribution is ensured by appropriately blurring the computed metric.

Formally, our focus metric is defined as follows for an image $x \in \mathbb{R}^2$:

$$\begin{aligned} B(x) &= \text{conv2}(x, \text{ones}(5, 5)) \\ V(x) &= (x - B(x))^2 \\ D(x) &= \text{abs}(x - \mu_x) \\ F(x) &= B(B(V(x) D(x) + D(x)^2)) \end{aligned} \tag{1}$$

where B is a two-dimensional convolution with a 5×5 unity matrix representing a blurring operator, V computes the squared difference between the image and its blurred version, D yields the deviation from the global mean (μ_x), and finally, F describes the focus metric comprised of the introduced operators.

Although our metric has proven effective in practice, other focus measures may also be used for this purpose, since at reduced resolution even simpler sharpness metrics can yield reasonably good segmentation and focus estimation.

We compute the distribution of the above-described focus metric for holograms reconstructed at each axial step throughout the observed volume. We assume that reconstructed objects exhibit the highest metric values near their true focal planes. Using the focus-metric distribution images obtained from reconstructions at different depths, we create a combined image containing only the metric maxima—the so-called focus-metric maxima distribution—along with a corresponding image storing the depth values associated with these maxima. By segmenting the maxima distribution image, we identify the spatial positions of the individual objects together with their estimated axial distances—each focused at different depths.

It should be emphasized that for high-resolution holograms this simple focus metric, as well as related approaches, is not universally reliable. In many cases, it relatively accurately determines the object focus, but in other cases (particularly for certain types of objects like phase objects) it systematically deviates from the correct value. Furthermore, the computation of such focus metrics is resource-intensive for high-resolution holograms.

3.2. Database construction

Our goal is to develop a system for automatic numerical focusing of volumetric water samples, where living micro-objects are dispersed throughout the entire volume and captured in a single in-line hologram. The aim is a solution that remains reliable under realistic conditions and is practical for deployment in real-world applications (e.g., functional for a wide range of specimen types). Therefore, we have built a dataset featuring a diverse range of living micro-objects (see panel A in Fig. 1): 1.) small point-like particle or bacterium (BAC) with around 1 μm diameter;

2.) *Yeast, Saccharomyces cerevisiae* (FU), alive; 3.) *Heamatococcus Pluvialis* (HC); 4.) larger and conglomerated *Heamatococcus Pluvialis* (HC2); 5.) a *Heamatococcus Pluvialis* (HC3) specimen at a different stage of its life cycle and with overlapping objects; and 6.) *Phaeodactylum tricornutum* (TRC), a versatile diatom. We prepared all samples to have a density of approximately 20-50 objects per 0.12 μL . Processing living micro-specimens is quite challenging because their morphology can vary throughout their lifespan. For example, *Haematococcus pluvialis* cell size varies significantly with its life stage, ranging from small motile zoospores (8–20 μm) to larger, spherical red cysts (30–60 μm) that form under stress for astaxanthin accumulation. Hence, for supervised learning models, we need to incorporate this diversity into the training dataset—which is why we include multiple types of HC samples.

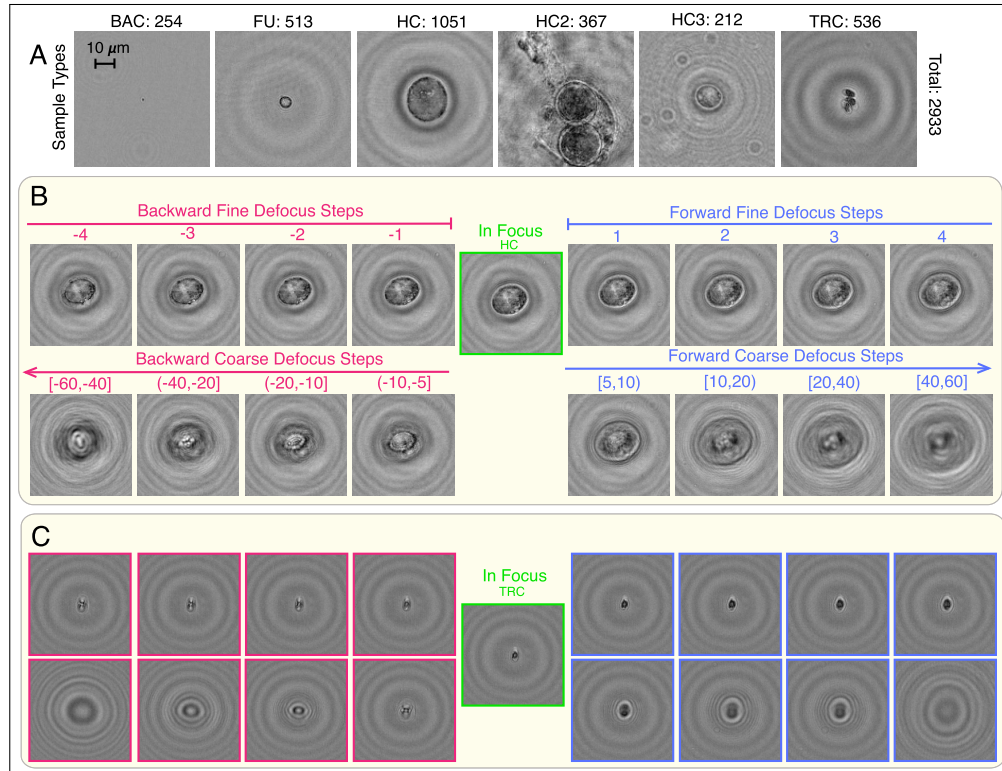


Fig. 1. **A:** Representative examples of different sample types and the number of samples in the database. **B:** Illustration of defocus classes where class labels are in step size units and a representative *HC* sample reconstructed at a corresponding depth. We have a total number of 17 classes: 1 in-focus class, 4+4 1-step fine defocus classes (4 steps in both directions), and 4+4 coarse defocus ranges (4 intervals in both directions). **C:** Same process for a representative *TRC* sample (without depicting the classes this time). Note that there is a significant difference between sample morphologies and defocus characteristics.

Another regular challenge when working with volumetric living samples occurs when other objects spatially overlap with the target object, as seen for HC3 (see Fig. 1). In this case, the model must learn to focus on the target object at the center of the image while disregarding other distractions, as these may be located in different axial planes. The spatial overlap wouldn't be problematic if the objects were situated in the same axial plane, but in our case, the objects are distributed through the full depth range, so spatially overlapping objects won't necessarily be in the same axial plane.

The measurement process (see previous section) detects objects present in the volume and extracts amplitude and phase images as 128×128 pixel regions of interest (ROIs) for each object from the full hologram. Each ROI is thus represented as a $2 \times 128 \times 128$ image, where the first dimension corresponds to the two channels (amplitude and phase). However, the objects are not focused accurately at the database construction stage, so a human annotator manually adjusts the focus by visual inspection. Note that for large objects, the selected focus plane may be somewhat arbitrary with the given step unit resolution, which can introduce labeling inconsistency and error. The smaller the object, the more easily the focus plane can be decided (see [Supplement 1 Figure S1 panel D](#) compared to other panels).

After the ground-truth focus plane distances are acquired for each object ROI, we reconstruct them at 17 distinct axial depths as shown in Fig. 1 panel B and C. Each object in the dataset is therefore represented by 17 distinct reconstructions: one at the focal plane, eight at increasing axial distances near the focal plane (four on each side), and another eight at larger distances (again, four on each side along the optical axis). The four reconstructions nearest on both sides of the focus plane use an axial increment of one step between successive planes; we will refer to this as the *fine-focus regime*. As we move further away in both directions, we define four increasingly broad intervals on each side of the focal plane; we refer to this as the *coarse-focus* or *far-focus regime*. On each of these intervals, we uniformly sample a focus distance value and reconstruct a hologram accordingly. For each object instance, its eight coarse-regime reconstruction distances will be different due to random sampling, and they will ultimately cover the entire far-focus range: $[-60, -5]$ steps and $[5, 60]$ steps. Thus, the total range spanned is $[-60, 60]$ steps centered around the focal plane of the object, corresponding to $\pm 205.8 \mu\text{m}$ in the original space or approx. $\pm 10 \text{ mm}$ in the magnified space.

It is important, that the reconstruction of an object is done by propagating the full size hologram, not the small ROIs, to prevent propagation artifacts. As reported by Langehanenberg et al. [54], diffraction artifacts from numerical propagation [3] corrupt the border regions of reconstructed holograms, which in turn degraded their focus-metric evaluation and were mitigated using a Hann filter. These artifacts become more prominent with increasing propagation distance and with decreasing size of the propagated hologram. Since the extent of the artifacts is a function of the propagated distances, neural networks bypass the task by learning the focus distance from these artifacts, rather than relying on features with true physical correspondence. This leads to excellent performance even on the test set, which is also corrupted by these artifacts, while the model remains completely clueless when deployed in the actual environment (where it can no longer rely on these spurious features). To prevent this, we reconstruct each object's ROI at 17 different axial depths by propagating the **full hologram**, ensuring that the model learns actual defocus characteristics instead of taking advantage of the data preparation artifacts.

The regression approach uses labels representing the relative distances to the focus plane, normalized to the $[-1, 1]$ range. For classification labels, we assign each instance to one of the corresponding 17 predefined categories. Thus, each instance is associated with a regression and a classification label and ready to be used by both approaches.

As shown in Fig. 1, the total number of (in-focus) object ROIs is only about 3 000, but the effective size—considering all unique axial depth realizations—becomes approximately 50 000. This approach efficiently increases dataset size without requiring the acquisition of thousands of holograms. For example, with an average of 30 objects per hologram, only 100 hologram recordings are needed (acquired with a fixed sample-to-sensor configuration) to construct a dataset with an effective sample size of approximately 50 000, making the process both practical and efficient. By comparison, for training the FocusNet [43] architecture, the authors acquired a total of 3 540 raw holograms at different axial positions spanning a reconstruction distance of 0.3 mm to 4.1 mm by mechanically adjusting the sample-to-camera distance.

3.3. Neural networks and training

This study examines and compares the effectiveness of regression and classification frameworks which are designed to share the same neural network backbone. The only difference is in the output layer. The regression network outputs a single signed scalar value (raw output without activation), representing the relative focus distance. In contrast, the classification network outputs a vector whose length equals the number of classes; this vector is normalized using the LogSoftmax function to yield class log-probabilities.

We adopt the main architecture elements from FocusNet [43] and FocusNetV2 [53]. Since our inputs are smaller and our dataset is more diverse we employ a different setup.

An alternative FocusNet and FocusNetV2 architecture is presented in Tables 1 and 2. The first three columns in Table 1 describe the convolutional backbone, while Table 2 lists the fully connected head. Together, these components constitute the architecture referred to in this study as FocusNet. FocusNetV2 [53] extends the original design by introducing a spatial attention layer [62] and a squeeze-and-excitation (SE) channel attention module [63]. In our implementation, the spatial attention layer is placed after the second convolutional block, and the SE block (reduction factor 16) follows the final convolutional block. This configuration, tailored for our task, was found to perform optimally and is consistently used throughout this work. Although the exact arrangement differs from the original publications, we retain the terms FocusNet and FocusNetV2 for clarity and continuity, as the core architectural principles remain unchanged.

Table 1. Convolutional backbone of the FocusNet [43] and FocusNetV2 [53]-like architectures. The first three columns describe the shared backbone; the last column lists extensions present only in FocusNetV2

Layer	Output shape	Operation / Details	FocusNetV2 additions
Input	$2 \times 128 \times 128$	Two channels (amplitude, phase)	—
ConvBlock1	$32 \times 128 \times 128$	7×7 Conv2D, BN, Swish	—
MaxPool1	$32 \times 64 \times 64$	2×2 MaxPool	—
ConvBlock2	$64 \times 64 \times 64$	5×5 Conv2D, BN, Swish	Spatial attention (CBAM)
MaxPool2	$64 \times 32 \times 32$	2×2 MaxPool	—
ConvBlock3	$128 \times 32 \times 32$	5×5 Conv2D, BN, Swish	—
MaxPool3	$128 \times 16 \times 16$	2×2 MaxPool	—
ConvBlock4	$256 \times 16 \times 16$	3×3 Conv2D, BN, Swish	—
MaxPool4	$256 \times 8 \times 8$	2×2 MaxPool	—
ConvBlock5	$128 \times 8 \times 8$	3×3 Conv2D, BN, Swish	SE channel attention ($r = 16$)
AvgPool	$128 \times 4 \times 4$	4×4 AvgPool	—

Table 2. Fully connected head structure shared by FocusNet and FocusNetV2 with different output configurations for regression and classification tasks.

Layer	Output shape	Operation / Details
Flatten	2048	Flatten
FC1	512	Dense, Dropout(0.3), Swish
FC2	256	Dense, Dropout(0.3), Swish
FC3	128	Dense, Dropout(0.3), Swish
FC4	64	Dense, Dropout(0.3), Swish
Output	1 or 17	Dense: Regression or Classification Identity or LogSoftmax output layer

The original loss function utilized by the FocusNet [43] paper was a simple Mean Squared Error (MSE) function; for an input tensor $\hat{z} \in \mathbb{R}^N$ and target tensor $z \in \mathbb{R}^N$, it is defined the following way:

$$\mathcal{L}_{\text{MSE}}(z, \hat{z}) = \frac{1}{N} \sum_{i=1}^N (z_i - \hat{z}_i)^2 \quad (2)$$

where z_i is the ground-truth target focus distance value, $\hat{z}_i$ is the model prediction, and N is the number of samples in the mini-batch. The drawback of this objective is that it disproportionately emphasizes outliers, making it highly sensitive to extreme values. To circumvent this issue, FocusNetV2 [53] introduces a Combined Loss (CMB) with a Mean Absolute Error term (MAE):

$$\mathcal{L}_{\text{MAE}}(z, \hat{z}) = \frac{1}{N} \sum_{i=1}^N |z_i - \hat{z}_i| \quad (3)$$

$$\mathcal{L}_{\text{CMB}}(z, \hat{z}) = \alpha \mathcal{L}_{\text{MSE}}(z, \hat{z}) + (1 - \alpha) \mathcal{L}_{\text{MAE}}(z, \hat{z}) \quad (4)$$

where $\mathcal{L}_{\text{CMB}}$ denotes the combined loss and $\alpha = 0.6$ as proposed by the paper [53].

We propose and examine an alternative loss function for regression, called Distance Weighted Smooth L1 Loss: The Smooth L1 loss, also known as the Huber loss, is a robust regression criterion that combines the advantages of the ℓ_1 and ℓ_2 losses by using a quadratic form for small residuals and a linear form for large residuals. Formally, it can be defined as follows:

$$\ell_{\text{SmoothL1}}(z_i, \hat{z}_i; \beta) = \begin{cases} \frac{1}{2\beta} (z_i - \hat{z}_i)^2, & \text{if } |z_i - \hat{z}_i| < \beta, \\ |z_i - \hat{z}_i| - \frac{\beta}{2}, & \text{otherwise,} \end{cases}$$

where $\beta > 0$ is a transition parameter that determines the width of the quadratic region. The mean Smooth L1 loss is then

$$\mathcal{L}_{\text{SmoothL1}}(z, \hat{z}; \beta) = \frac{1}{N} \sum_{i=1}^N \ell_{\text{SmoothL1}}(z_i, \hat{z}_i; \beta).$$

To account for depth-dependent reliability in our holographic training data, we introduce a distance-based weighting scheme into the Smooth L1 formulation. Let the per-element weight be defined as

$$w_i(\tau) = \exp\left(-\frac{|z_i|}{\tau}\right),$$

where $\tau > 0$ is a decay constant (aka half-life) controlling how rapidly the influence of samples decreases with increasing absolute target distance $|z_i|$. If $\tau \rightarrow \infty$ it defaults back to Smooth L1 Loss. To maintain stable learning dynamics, the weights are normalized in the mini-batch so that their mean value is approximately one:

$$\tilde{w}_i(\tau) = \frac{w_i(\tau)}{\frac{1}{N} \sum_{j=1}^N w_j(\tau)}.$$

The proposed Distance-Weighted Smooth L1 Loss (WSL1) is then given by

$$\mathcal{L}_{\text{WSL1}}(z, \hat{z}; \beta, \tau) = \frac{1}{N} \sum_{i=1}^N \tilde{w}_i(\tau) \ell_{\text{SmoothL1}}(z_i, \hat{z}_i; \beta). \quad (5)$$

This formulation down-weights the contribution of targets that lie farther from the focal plane (large $|z_i|$), and concentrates performance near focus. The rationale for this approach is that in the

fine-focus regime, there are more distinctive visual cues than in the *coarse-focus regime*, making it advantageous to emphasize focusing performance near the focal plane. This is further illustrated in Supplement 1 Figure S2. In this study, we set $\tau = 0.5$ and $\beta = 0.03$. The latter effectively means that we switch to ℓ_2 if the absolute error is within approximately 2 step increments.

For the classification task, the network is trained using the Negative Log-Likelihood Loss (NLLLoss). Given a mini-batch of N samples, each with a one-hot encoded target vector $y_i = (y_{i1}, \dots, y_{iC})$ and predicted log-probabilities $l_i = (l_{i1}, \dots, l_{iC})$, the per-sample loss is defined as

$$\ell_{\text{NLL}}(y_i, l_i) = - \sum_{c=1}^C y_{ic} l_{ic},$$

and the batch-averaged loss is

$$\mathcal{L}_{\text{NLL}}(y, l) = \frac{1}{N} \sum_{i=1}^N \ell_{\text{NLL}}(y_i, l_i) = - \frac{1}{N} \sum_{i=1}^N \sum_{c=1}^C y_{ic} l_{ic}. \quad (6)$$

Here, C denotes the total number of classes, $y_{ic} \in \{0, 1\}$ is the binary indicator of the correct class for sample i , and l_{ic} is the log-probability predicted for class c . The log-probabilities are obtained by applying the LogSoftMax transformation to the raw logits $r_i = (r_{i1}, \dots, r_{iC})$ of the final network layer (as shown in Table 2):

$$l_{ic} = \log \left(\frac{e^{r_{ic}}}{\sum_{j=1}^C e^{r_{ij}}} \right) = r_{ic} - \log \left(\sum_{j=1}^C e^{r_{ij}} \right).$$

This formulation is equivalent to the standard Cross-Entropy Loss computed on SoftMax-normalized outputs.

Based on the above definitions and specifications, we define four model configurations for comparison. The first baseline, *FocusNet*, reproduces the original implementation [43] and employs the FocusNet architecture trained with a mean squared error (MSE) loss. The second model, *FocusNetV2*, follows the updated design introduced by Liu et al. [53], utilizing the FocusNetV2 architecture in conjunction with the combined loss function (CMB). Our proposed regression model, denoted as *FFocusRNet* (**F**ine **F**ocus **R**egression **N**etwork), adopts the FocusNetV2 backbone but replaces the loss with the Distance-Weighted Smooth L1 (WSL1) regression loss. Finally, the proposed classification model, *FFocusCNet* (**F**ine **F**ocus **C**lassification **N**etwork), shares the same architectural design as FocusNetV2 while employing the Negative Log-Likelihood (NLL) loss for supervised focus classification. Table 3 summarizes the model definitions.

Table 3. Summary of the evaluated model configurations.

Model	Architecture	Learning type	Loss function
FocusNet [43]	FocusNet	Regression	Mean Squared Error (MSE)
FocusNetV2 [53]	FocusNetV2	Regression	Combined Loss (CMB)
FFocusRNet (ours)	FocusNetV2	Regression	Distance-Weighted Smooth L1 (WSL1)
FFocusCNet (ours)	FocusNetV2	Classification	Negative Log-Likelihood (NLL)

We developed and trained our neural networks using the widely adopted PyTorch deep learning framework. We used the AdamW optimizer [64] with an initial learning rate of 6×10^{-4} , and employed a learning rate scheduler that reduces the learning rate by a factor of 0.7 if the validation loss plateaus for 10 epochs. We stop training if there is no improvement in the validation loss for 30 epochs. The best model is used for testing.

The dataset was divided into training, validation, and test sets in proportions of 70%, 15%, and 15%, respectively. To ensure balanced representation and effective sampling, we group the samples by type. During each training iteration, we first uniformly select a sample type, then uniformly choose an individual sample from that subset, and finally randomly select one of the 17 available axial-depth reconstructions. As a result, each training epoch consists of a diverse set of samples and reconstruction depths, gradually exposing the network to the entire dataset over multiple epochs. We also applied random horizontal and vertical flips, random rotations in increments of 95 degrees, and hologram noise [65] transformations to enhance generalization and further augment the database. The training stops when the validation loss no longer improves, indicating that the network has sufficiently learned from the data and cannot extract any additional information (i.e., early stopping).

For evaluation, we employed a deterministic strategy distinct from the random sampling logic of the training phase. Specifically, for validation and testing, we systematically traverse each sample type group and its objects, evaluating all 17 axial-depth reconstructions per object in a batch. For example, if the test set comprises 500 unique objects, validation and testing would each consist of 500 mini-batches, each containing the object's reconstructions at each of the 17 distinct axial planes. This approach ensures robust and reproducible evaluation of model performance.

3.4. Evaluation

We evaluated all models within a unified framework, enabling direct comparison using both regression and classification metrics. To achieve this, we cast the regression outputs to classification labels and, conversely, derive regression estimates from the classification outputs. Specifically, for regression-to-classification, the predicted value is converted to step units, rounded, and assigned to the appropriate class bin. For classification-to-regression, we tested two strategies: (1) assigning the regression value as the center of the predicted bin, and (2) computing the expected value as the dot product between the class probabilities and their corresponding bin centers. We observed similar performance with both approaches and adopted the latter for its greater sophistication and interpretability. Notably, the use of class probability distributions in the classification setting also enables quantification of model confidence. For regression metrics, all results are presented using magnified space distances in micrometers (μm).

Classification performance was quantified using both accuracy and the F1-score, the latter being the harmonic mean of precision and recall. Accuracy (Acc) is defined as the ratio of correctly predicted labels to the total number of samples:

$$\text{Acc} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Samples}} \quad (7)$$

Precision (Pr) is the fraction of true positives among all positive predictions, and recall (Re) is the fraction of true positives among all actual positives:

$$\text{Pr} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \quad \text{Re} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

The F1-score (F1) is then given by:

$$\text{F1} = 2 \times \frac{\text{Pr} \times \text{Re}}{\text{Pr} + \text{Re}} \quad (8)$$

Both precision and recall are computed in a macro-averaged fashion across all classes.

For regression performance, we report both the Mean Absolute Error (MAE) and the normalized root mean squared error (NRMSE). The MAE quantifies the average of the absolute differences

between the predicted ($\hat{y}_i$) and true (y_i) values:

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i| \quad (9)$$

The NRMSE captures the root mean squared error, normalized by the range ($y_{\max} - y_{\min}$) of the ground-truth values, providing a dimensionless measure of relative error:

$$\text{NRMSE} = \frac{\sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2}}{y_{\max} - y_{\min}} \quad (10)$$

These are referred to as the *global regression metrics* throughout this study. We also introduce *fine-grained regression metrics*, which are evaluated only on a subset of samples where the ground-truth relative distance is within ± 1 mm of the focal plane (measured in the magnified space). This enables a targeted assessment of model performance in the *near-focus regime* (approximately ± 5 steps around the focal plane).

Formally, for a set of N_{fine} samples satisfying $|y_i| \leq 1$ mm (magnified space), the fine-grained MAE_{fine} and $\text{NRMSE}_{\text{fine}}$ are computed as:

$$\text{MAE}_{\text{fine}} = \frac{1}{N_{\text{fine}}} \sum_{i=1}^{N_{\text{fine}}} |y_i - \hat{y}_i| \quad (11)$$

$$\text{NRMSE}_{\text{fine}} = \frac{\sqrt{\frac{1}{N_{\text{fine}}} \sum_{i=1}^{N_{\text{fine}}} (y_i - \hat{y}_i)^2}}{y_{\max}^{\text{fine}} - y_{\min}^{\text{fine}}} \quad (12)$$

where $y_{\max}^{\text{fine}}$ and $y_{\min}^{\text{fine}}$ denote the maximum and minimum ground-truth distances considered within the fine-focus window, respectively. These are referred to as the *fine-grained regression metrics* or *near-focus regression metrics*.

Finally, we also evaluate model performance under recursive (multi-pass) inference, in which the model's predictions are used iteratively to update the input. Specifically, following an initial prediction, the input hologram is numerically propagated to the newly estimated axial position and resubmitted to the model for further refinement (3-times). For regression models, this process is direct, as the model outputs a relative axial distance to propagate the hologram. In the classification setting, the hologram is propagated by the expected relative distance inferred from the class probability distribution, as previously described. The final focus estimate is calculated as the sum of the incremental predictions across all iterations. All evaluation metrics described above are equally applicable in this recursive inference framework, enabling a direct comparison between single-pass and multi-pass results.

4. Results

Figure 2 summarizes the prefocusing performance of the low-resolution 3D volumetric detection algorithm that provides the initialization for the NN-based refinement stage. The error histogram reveals a non-uniform distribution with its peak around zero, meaning that most estimates are close to the true focus with occasional larger deviations. The scatter plot further reveals sample-dependent trends; for example, BAC samples (orange) show a consistent bias relative to the ideal 45° line.

Single-pass overall results Table 4 shows the overall test performance metrics across all sample types, and Fig. 3 displays the confusion matrices for different models, sample types, and for all samples.

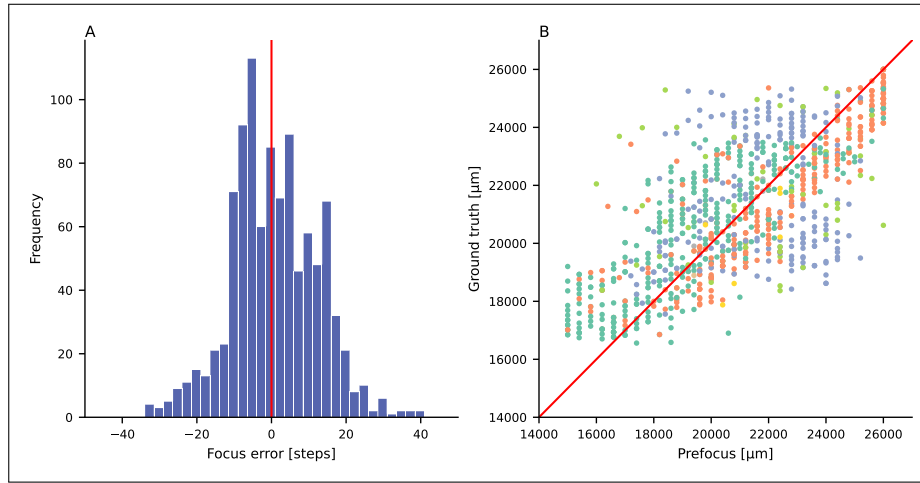


Fig. 2. Low-resolution prefocusing performance. **A:** Histogram of the prefocusing error (in axial step units); the red vertical line marks zero error. **B:** Ground-truth absolute axial distance versus the prefocus estimate (in μm); colors denote sample types, and the red 45° line indicates ideal agreement. Together, these plots show that the coarse stage provides a consistent initialization around the focal plane, supplying suitable inputs for the subsequent learning-based refinement.

Table 4. Overall test results. Best values in bold. For the MAE metrics, the μm unit is in the magnified space and top two ties within $30\mu\text{m}$ are co-highlighted.

Model	Acc	F1	MAE [μm]	MAE _{fine} [μm]	NRMSE	NRMSE _{fine}
FocusNet	0.512	0.504	393.8	227.3	0.022	0.191
FocusNetV2	0.601	0.593	324.7	173.4	0.020	0.170
FFocusRNet	0.641	0.636	337.0	152.4	0.025	0.162
FFocusCNet	0.726	0.725	578.4	126.2	0.041	0.123

Considering the classification metrics, *FFocusCNet* achieves the best overall classification ($\text{Acc}/\text{F1} \approx 0.726$), improving accuracy by +21.3 pp over the FocusNet baseline and by +12.1 pp compared to FocusNetV2. Regarding the global regression metrics, *FocusNetV2* attains the best MAE ($324.7\mu\text{m}$) and NRMSE (0.020). *FFocusRNet* is close on MAE ($337.0\mu\text{m}$; within $12.3\mu\text{m}$) and better than FocusNetV2 on MAE_{fine} ($152.4\mu\text{m}$). Regarding the near-focus regression metrics (where the fine window is $\pm 1\text{mm}$ around focus in the magnified space), *FFocusCNet* leads again with MAE_{fine} = $126.2\mu\text{m}$ and NRMSE_{fine} = 0.123.

Multi-pass overall results Table 5 shows the overall test performance metrics for the 3-step recursive evaluation method.

Table 5. Overall test results with 3-step recursive call of the same models. Best values in bold. For MAE metrics, the μm unit is in the magnified space.

Model	Acc	F1	MAE [μm]	MAE _{fine} [μm]	NRMSE	NRMSE _{fine}
FocusNetV2	0.633	0.621	491.4	236.3	0.052	0.364
FFocusRNet	0.666	0.658	414.5	195.7	0.043	0.279
FFocusCNet	0.717	0.713	316.4	136.8	0.033	0.154

FFocusCNet benefits substantially from recursion: MAE reduces from $578.4\mu\text{m}$ to $316.4\mu\text{m}$, with the best overall classification and regression after recursion, with a slight drop in classification

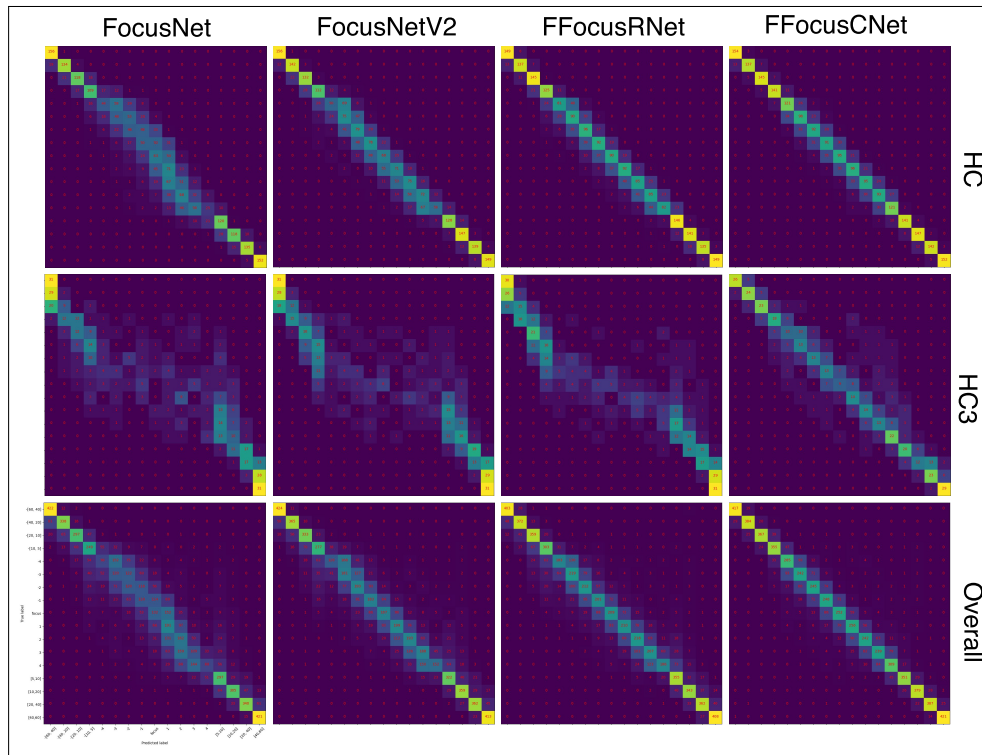


Fig. 3. Confusion matrices for different models and sample types where the x-axis is the prediction and the y-axis is the ground-truth label. The more diagonally concentrated the matrix, the better. The bottom row is the overall confusion matrix, including all samples. The regression networks were unable to effectively focus on the HC3 sample type; overlapping objects distracted the regression networks. On the other hand, *FFocusCNet* shows more robustness to these kinds of distractions.

and fine-regression metrics. By contrast, recursion degrades regression for *FocusNetV2* and *FFocusRNet* (e.g., *FocusNetV2* MAE = 491.4 μm ; *FFocusRNet* MAE = 414.5 μm), despite modest classification gains.

More detailed performance results, including evaluations by sample type, are published in the [Supplement 1](#) Table S1. Figure 3 displays the confusion matrices for different models, two selected sample types, and also for all samples (bottom row). Diagonal concentration reflects correct predictions, so a sharper diagonal corresponds to stronger classification performance. By this measure, *FFocusCNet* emerges as the most consistent and accurate model (in line with the results of Table 4). The extended figure version containing the confusion matrices for all sample types is published in the [Supplement 1](#) Figure S3.

See detailed interpretations of the results in the following discussion section.

5. Discussion

In the first stage, our system employs a conventional focusing heuristic to locate objects throughout the volume and select object ROIs that are reconstructed reasonably close to their actual axial positions. As illustrated in Fig. 2, this primal algorithm lacks robustness when applied to different types of objects within the sample. Traditional algorithms of this kind are computationally intensive and sensitive to variations in sample morphology and phase characteristics. To reduce

computational cost and improve robustness, we apply downsampling, which, while improving efficiency, comes at the expense of precision and introduces additional sources of bias. For example, a systematic bias emerges for the BAC sample type, as indicated by the orange dots in the scatter plot. This occurs because, at low spatial resolution, these objects are too small (around 1 μm in diameter) to be detected when sharply focused; they only become visible as their diffraction pattern expands. Thus, while such conventional algorithms are not suitable for general or precise focusing, they are sufficient for providing a preliminary, coarse localization of objects within the volume. As shown by the histogram in Fig. 2, the resulting prefocusing error distribution is centered near zero and remains well within the interval of ± 60 steps. This level of accuracy provides a suitable initialization for the subsequent deep learning-based refinement stage.

The test results shown in Table 4 align with our expectations: our proposed classification approach outperforms all other methods in terms of classification metrics and fine-grained regression metrics. This is reasonable because, first, we optimize *FFocusCNet* directly for a classification objective, and second, we designed the class geometry to maximize performance near the focus plane. Conversely, it underperforms in terms of global regression metrics due to the expanding distant class bins. Even if the network classifies the distant samples with high accuracy, the performance in terms of a regression metric is limited by the bin size. For example, suppose we have a sample 40 steps away from focus, and suppose we correctly classified it into the last class, which is defined as the $[40, 60]$ interval (in step units) with bin center 50. In this particular case, the classification is correct, yet the MAE is 10 steps. In general, the theoretical MAE we will have on average, assuming uniform distribution of samples within the bin, is $w/4$, where w is the bin width, which is 5 steps in the current example. In contrast, the regression network might miss the "correct" class with an estimate of 39 steps, yet the MAE is just 1 step. This exemplifies one reason why the regression networks perform better in terms of global regression metrics, despite the fact that confusion matrices indicate worse accuracy even for extreme classes with wide bins (see Fig. 3). Another apparent reason is that a pure regression model such as *FocusNetV2* is trained with a depth-agnostic loss, treating near- and far-from-focus sample errors equally across the full axial span. This objective naturally favors low global regression metrics (MAE, NRMSE) as it is directly optimized for that. However, because it does not emphasize the ± 1 mm band, it underperforms compared to *FFocusCNet* in the near-focus regime. To address this, we introduced *FFocusRNet*, a regression variant with an exponentially decaying sample weight $w(z) = \exp(-|z|/\tau)$ that downweights distant depths z and concentrates learning capacity near the focal plane, aimed to improve the fine-range error of *FocusNetV2* while maintaining a better global regression error than *FFocusCNet*. This interpretation is supported by the results in Table 4, where *FFocusRNet* achieves a global MAE comparable to that of the depth-agnostic (pure) regressor (*FocusNetV2*), while simultaneously outperforming it in terms of fine-range metrics such as MAE_{fine} and $\text{NRMSE}_{\text{fine}}$. However, both of our fine focus models underperform in terms of global NRMSE, as this metric tends to highlight outliers, resulting in a mean statistic biased towards the outliers, and in our particular case, indicating weak performance for distant samples in the far-focus regime.

With our expanding class bin geometry design, we increase performance near focus at the expense of losing precision far from focus. This trade-off is worth considering if we aim for optimal near-focus performance while not mind some inaccuracy farther away. For example, when the distribution of sample distances is **not uniform**, but rather concentrated closer around the focus plane—just like how our coarse prefocusing algorithm prepares the samples (see Fig. 2). Under these conditions, prioritizing near-focus performance at the cost of far-focus performance is justified because optimizing for the majority of samples results in a lower average error on the uneven sample distribution (than a pure regressor would).

An extra advantage of the non-uniform bin design is that it naturally supports **progressive refinement**: a sample initially assigned to a coarse distant bin can be propagated with the corresponding bin center and re-evaluated, allowing the network to converge into the fine-focus regime. This design can achieve fine-focus accuracy even for samples that are far from focus, with only three refinement steps proving sufficient in practice. To illustrate, consider a worst-case example of a hologram lying near the far edge of the outermost bin, e.g. at -40.5 steps in the interval $[-60, -40]$. In the first iteration, suppose the network assigns it to the correct bin, and we forward-propagate with its center ($+50$), which overshoots the true focus with 9.5 steps and places the sample in the opposite-side coarse bin $[5, 10]$. In the second iteration, we backpropagate with -7.5 steps and the network assigns the sample to the corresponding fine-focus bin 2 steps away from focus. In the third iteration, the focusing process reaches the focal plane. While not a formal guarantee, this reasoning explains why three steps are typically sufficient. This reasoning is corroborated by experiments where allowing more than three iterations produced no further improvement.

Empirical test results for the classification network in Table 5 confirms the effectiveness of progressive refinement: *FFocusCNet* emerges as the **top performer** across all classification, global regression, and fine-grained regression metrics. With iterative calling, *FFocusCNet* maintains its classification performance compared to the single-shot mode (albeit with slight degradation, which may be partly due to the rebinarization based on the predicted cumulative expected distance) and significantly improves its global MAE metric, surpassing all other models. In contrast, except for a slight improvement in classification, the iterative call had a negative impact on all regression networks. It was not expected that recursive calling would significantly improve *FocusNetV2*, since it has similar performance throughout the axial range; however, the regression metrics worsened, surprisingly even for *FFocusRNet*, which has increasing capacity approaching the focal plane. The reason is that the HC3 class represents cases where the target object's ROI contains overlapping objects (see panel A in Fig. 1). This overlap introduces ambiguous and conflicting features, which can disrupt the ability of regression networks to extract meaningful focus cues and often leads to incorrect predictions of focus direction and plane. In contrast, *FFocusCNet* handles these challenges more robustly, maintaining higher accuracy in the presence of overlapping objects. An iterative call can amplify this effect, producing significant errors that bias regression metrics. The binarization for the edge classes is conducted such that every estimate even larger than the extreme values (in an absolute sense) is assigned to the corresponding edge class (for example, -98 would be assigned to the class $[-60, -40]$); thus, "going off the map" in the wrong direction is not reflected in the classification results, although it is heavily penalized by the regression metrics—this explains why classification accuracy improves while regression metrics worsen at the same time. For *FFocusRNet*, the gains from converging cases are outweighed by this error-magnifying effect, leading to higher MAE than its single-shot counterpart, while the degradation is less severe than for *FocusNetV2*. The effect is especially pronounced in the NRMSE metrics, which emphasize large errors. The confusion matrices in Fig. 3 support this hypothesis, which shows that, unlike *FFocusCNet*, regression networks struggle to handle the HC3 samples efficiently. Furthermore, we verified that this effect is magnified in the case of iterative calls, particularly for the regression networks, which causes large outliers and results in worse statistics for HC3 samples—even though improvements are visible for the remaining sample types. This can be validated by inspecting Supplement 1 Table S1, which includes per-sample statistics, and by the per-sample error histogram presented in Supplement 1 Figure S4. Even when the challenging HC3 sample type is excluded, and thus object overlap within ROIs is ruled out, *FFocusCNet* continues to outperform all other models under recursive evaluation.

We should also mention another factor that may affect the effectiveness of the iterative approach: propagation artifacts. In our recursive setup, the small hologram ROIs are re-propagated to the

predicted distance, introducing distortions not present during training. The larger the propagation distance, the more prominent the artifacts will be, which may negatively impact the performance of the models upon iterative calls.

As previously discussed, although iterative refinement generally enhances focus estimation in most cases, it can also increase errors and outliers in the remaining instances. This trade-off can be effectively managed by monitoring the convergence of the refinement process, which allows for early stopping once the convergence criteria are met and helps identify cases where convergence does not occur.

Finally, we note that the dataset used in this study encompasses a broad spectrum of sample morphologies (see Fig. 1, panel A), and empirical observations indicate that this enabled the models to generalize robustly even to previously unseen objects, although a detailed quantitative analysis of this aspect is beyond the scope of the present study.

6. Conclusions

In the first stage in our 2-stage framework, the proposed conventional coarse focus search algorithm robustly and efficiently identifies object ROIs and provides focus estimates that are consistently distributed close around the ground truth. This ensures accurate and effective pre-initialization for the subsequent neural network refinement stage.

Overall, our proposed *FFocusCNet* is the most practical choice for deployment in our two-stage multifocusing pipeline. It delivers the strongest near-focus accuracy—where most ROIs lie after coarse prefocusing—and maintains robustness across diverse morphologies and challenging overlaps (e.g., HC3) that degrade regression-based models. In addition, the non-uniform binning scheme enables effective recursive refinement: far-from-focus ROIs can be iteratively brought into the fine-focus regime, yielding the best overall performance under multi-pass inference.

All models converged within roughly 4.5–7 hours of training using an NVIDIA RTX 2080 Ti GPU, with iteration counts between 18–24 k. The *FFocusCNet* classification network achieved the shortest training time (4.5 h), owing to the lighter computational cost of its NLL-based objective compared to regression losses. The three-step recursive inference sequence required only around 1 minute for 435 iterations, corresponding to 435 processed batches of object fields, containing 17 distinct reconstructions in each batch. This efficiency demonstrates that the proposed framework can perform rapid multi-step focus estimation and refinement and is well-suited for *quasi-real-time deployment* in flow-through inline digital holographic microscopy applications.

In future work, we aim to investigate recursive neural network approaches that maintain memory across iterations and are trained specifically for recursive operation.

Funding. Artificial Intelligence National Laboratory (RRF-2.3.1-21-2022-00004).

Acknowledgement. Support by the the European Union project RRF-2.3.1-21-2022-00004 within the framework of the Artificial Intelligence National Laboratory.

Disclosures. The authors declare no conflicts of interest.

Data Availability. The underlying data supporting the results presented in this paper are not publicly available due to commercial and intellectual property considerations. However, interested researchers may request access to the data directly from the authors.

Supplemental document. See [Supplement 1](#) for supporting content.

References

1. D. Gabor, “A new microscopic principle,” *Nature* **161**(4098), 777–778 (1948).
2. U. Schnars and W. P. Jüptner, “Digital recording and numerical reconstruction of holograms,” *Meas. Sci. Technol.* **13**(9), R85–R101 (2002).
3. K. Matsushima and T. Shimobaba, “Band-limited angular spectrum method for numerical simulation of free-space propagation in far and near fields,” *Opt. Express* **17**(22), 19662–19673 (2009).

4. J. Garcia-Sucerquia, W. Xu, S. K. Jericho, *et al.*, “Digital in-line holographic microscopy,” *Appl. Opt.* **45**(5), 836–850 (2006).
5. S. K. Jericho, J. Garcia-Sucerquia, W. Xu, *et al.*, “Submersible digital in-line holographic microscope,” *Rev. Sci. Instrum.* **77**(4), 043706 (2006).
6. A. B. Bochdansky, M. H. Jericho, and G. J. Herndl, “Development and deployment of a point-source digital inline holographic microscope for the study of plankton and particles to a depth of 6000 m,” *Limnol. Oceanogr.: Methods* **11**, 28–40 (2013).
7. O. Mudanyali, D. Tseng, C. Oh, *et al.*, “Compact, light-weight and cost-effective microscope based on lensless incoherent holography for telemedicine applications,” *Lab Chip* **10**(11), 1417–1428 (2010).
8. M. Lee, O. Yaglidere, and A. Ozcan, “Field-portable reflection and transmission microscopy based on lensless holography,” *Biomed. Opt. Express* **2**(9), 2721–2730 (2011).
9. A. Greenbaum, N. Akbari, A. Feizi, *et al.*, “Field-portable pixel super-resolution colour microscope,” *PLoS One* **8**(9), e76475 (2013).
10. F. Niknam, H. Qazvini, and H. Latifi, “Holographic optical field recovery using a regularized untrained deep decoder network,” *Sci. Rep.* **11**(1), 10903 (2021).
11. H. Chen, L. Huang, T. Liu, *et al.*, “Fourier Imager Network (FIN): A deep neural network for hologram reconstruction with superior external generalization,” *Light: Sci. Appl.* **11**(1), 254 (2022).
12. H. Chen, L. Huang, T. Liu, *et al.*, “eFIN: Enhanced Fourier Imager Network for Generalizable Autofocusing and Pixel Super-Resolution in Holographic Imaging,” *IEEE J. Sel. Top. Quantum Electron.* **29**(4), 1–10 (2023).
13. H. Li, X. Chen, Z. Chi, *et al.*, “Deep DIH: Single-Shot Digital In-Line Holography Reconstruction by Deep Learning,” *IEEE Access* **8**, 202648–202659 (2020).
14. K. Wang, J. Dou, Q. Kema, *et al.*, “Y-Net: a one-to-two deep learning framework for digital holographic reconstruction,” *Opt. Lett.* **44**(19), 4765 (2019).
15. Y. Rivenson, Y. Zhang, H. Günaydin, *et al.*, “Phase recovery and holographic image reconstruction using deep learning in neural networks,” *Light: Science Applications* **7**(2), 17141 (2017).
16. M. Lyu, H. Wang, G. Li, *et al.*, “eHoloNet: a learning-based end-to-end approach for in-line digital holographic reconstruction,” *Opt. Express* **26**(18), 22603–22614 (2017).
17. K. Wang, L. Song, C. Wang, *et al.*, “On the use of deep learning for phase recovery,” *Light: Sci. Appl.* **13**(1), 4 (2024).
18. G. Ongie, A. Jalal, C. A. Metzler, *et al.*, “Deep Learning Techniques for Inverse Problems in Imaging,” *arXiv* (2020).
19. P. Putzky and M. Welling, “Recurrent Inference Machines for Solving Inverse Problems,” *arXiv* (2017).
20. A. Qayyum, I. Ilahi, F. Shamshad, *et al.*, “Untrained Neural Network Priors for Inverse Imaging Problems: A Survey,” *IEEE Trans. Pattern Anal. Mach. Intell.* **45**, 1–20 (2022).
21. J. R. Fienup, “Phase retrieval algorithms: a comparison,” *Appl. Opt.* **21**(15), 2758 (1982).
22. C. A. Metzler, P. Schniter, A. Veeraraghavan, *et al.*, “prDeep: Robust Phase Retrieval with a Flexible Deep Network,” *arXiv* (2018).
23. P. Hand, O. Leong, and V. Voroninski, “Phase Retrieval Under a Generative Prior,” *NIPS’18: Proceedings of the 32nd International Conference on Neural Information Processing Systems* 9154–9164 (2018).
24. Ç. Işıl, F. S. Oktem, and A. Koç, “Deep Iterative Reconstruction for Phase Retrieval,” *Appl. Opt.* **58**(20), 5422 (2019).
25. T. Uelwer, A. Oberstraß, and S. Harmeling, “Phase Retrieval Using Conditional Generative Adversarial Networks,” in *2020 25th International Conference on Pattern Recognition (ICPR)*, (IEEE, Milan, Italy, 2021), pp. 731–738.
26. Y. Yang, Q. Lian, X. Zhang, *et al.*, “HIONet: Deep priors based deep unfolded network for phase retrieval,” *Digital Signal Processing* **132**, 103797 (2023).
27. M. Rogalski, P. Arcab, L. Stanaszek, *et al.*, “Physics-driven universal twin-image removal network for digital in-line holographic microscopy,” *Opt. Express* **32**(1), 742 (2024).
28. Y. Gao and L. Cao, “Iterative projection meets sparsity regularization: towards practical single-shot quantitative phase imaging with in-line holography,” *Light: Adv. Manuf.* **4**, 1 (2023).
29. X. Chen, H. Wang, A. Razi, *et al.*, “DH-GAN: a physics-driven untrained generative adversarial network for holographic imaging,” *Opt. Express* **31**(6), 10114–10135 (2023).
30. W. Zhang, L. Cao, D. J. Brady, *et al.*, “Twin-Image-Free Holography: A Compressive Sensing Approach,” *Phys. Rev. Lett.* **121**(9), 093902 (2018).
31. J. Bioucas-Dias and M. Figueiredo, “A New TwIST: Two-Step Iterative Shrinkage/Thresholding Algorithms for Image Restoration,” *IEEE Trans. on Image Process.* **16**(12), 2992–3004 (2007).
32. T. Lатычевскаиа, “Iterative phase retrieval for digital holography,” *J. Opt. Soc. Am. A* **36**(12), D31–D40 (2019).
33. G. Koren, F. Polack, and D. Joyeux, “Iterative algorithms for twin-image elimination in in-line holography using finite-support constraints,” *J. Opt. Soc. Am. A* **10**(3), 423 (1993).
34. T. Lатычевскаиа and H.-W. Fink, “Solution to the Twin Image Problem in Holography,” *Phys. Rev. Lett.* **98**(23), 233901 (2007).
35. H. Luo, J. Xu, L. Zhong, *et al.*, “Diffraction-Net: a robust single-shot holography for multi-distance lensless imaging,” *Opt. Express* **30**(23), 41724 (2022).
36. Y. Zhang, X. Liu, and E. Y. Lam, “Single-shot inline holography using a physics-aware diffusion model,” *Opt. Express* **32**(6), 10444 (2024).

37. H. Gan, X. Wang, L. He, *et al.*, “Learned Two-Step Iterative Shrinkage Thresholding Algorithm for Deep Compressive Sensing,” *IEEE Trans. Circuits Syst. Video Technol.* **34**(5), 3943–3956 (2024).
38. I. Daubechies, M. Defrise, and C. De Mol, “An iterative thresholding algorithm for linear inverse problems with a sparsity constraint,” *Comm Pure Appl Math* **57**(11), 1413–1457 (2003).
39. D. Terbe, L. Orzó, and Á. Zarándy, “Classification of holograms with 3d-cnn,” *Sensors* **22**(21), 8366 (2022).
40. M. Pachitariu, M. Rariden, and C. Stringer, “Cellpose-sam: superhuman generalization for cellular segmentation,” *bioRxiv* (2025).
41. Y. Wu, Y. Luo, G. Chaudhari, *et al.*, “Bright-field holography: cross-modality deep learning enables snapshot 3d imaging with bright-field contrast using a single hologram,” *Light: Sci. Appl.* **8**(1), 25 (2019).
42. D. Terbe, L. Orzó, and Á. Zarándy, “Deep-learning-based bright-field image generation from a single hologram using an unpaired dataset,” *Opt. Lett.* **46**(22), 5567–5570 (2021).
43. M. Montoya, M. J. Lopera, A. Gómez-Ramírez, *et al.*, “FocusNET: An autofocusing learning-based model for digital lensless holographic microscopy,” *Opt. Lasers Eng.* **165**, 107546 (2023).
44. T. Pitkäaho, A. Manninen, and T. J. Naughton, “Performance of Autofocus Capability of Deep Convolutional Neural Networks in Digital Holographic Microscopy,” in *Digital Holography and Three-Dimensional Imaging*, (OSA, JeJu Island, 2016), p. W2A.5.
45. T. Pitkäaho, A. Manninen, and T. J. Naughton, “Focus classification in digital holographic microscopy using deep convolutional neural networks,” in *SPIE Proceedings*, vol. 10414 E. Beaufort, F. S. Pavone, and P. T. C. So, eds. (SPIE, Munich, Germany, 2017), p. 104140K. ISSN: 0277-786X.
46. L. Huang, T. Liu, X. Yang, *et al.*, “Holographic Image Reconstruction with Phase Recovery and Autofocusing Using Recurrent Neural Networks,” *ACS Photonics* **8**(6), 1763–1774 (2021).
47. Y. Zhang, H. Wang, Y. Wu, *et al.*, “Edge sparsity criterion for robust holographic autofocusing,” *Opt. Lett.* **42**(19), 3824 (2017).
48. J. Winnik, D. Suski, P. Zdanowski, *et al.*, “Versatile optimization-based speed-up method for autofocusing in digital holographic microscopy,” *Opt. Express* **29**(21), 33297 (2021).
49. C. Vimala and A. Ajeena, “Auto focusing of in-Line Holography based on Stacked Auto Encoder with Sparse Bayesian Regression and Compressive Sensing,” *Multimed. Tools Appl.* **83**(30), 74593–74623 (2024).
50. Z. Ren, Z. Xu, and E. Y. Lam, “Learning-based nonparametric autofocusing for digital holography,” *Optica* **5**(4), 337 (2018).
51. T. Shimobaba, T. Kakue, and T. Ito, “Convolutional neural network-based regression for depth prediction in digital holography,” *2018 IEEE 27th International Symposium on Industrial Electronics (ISIE)* (2018).
52. X. Chen, H. Wang, Z. Zhang, *et al.*, “Enhancing Digital Hologram Reconstruction Using Reverse-Attention Loss for Untrained Physics-Driven Deep Learning Models with Uncertain Distance,” *arXiv* (2024).
53. K. Liu, J. Chen, H. Liao, *et al.*, “Digital holographic automatic focusing based on improved focusNET,” *Opt. Eng.* **64**(05), 056106 (2025).
54. P. Langehanenberg, G. Von Bally, and B. Kemper, “Autofocusing in digital holographic microscopy,” *3D Res.* **2**(1), 4 (2011).
55. M. Trusiak, J.-A. Picazo-Bueno, P. Zdanowski, *et al.*, “DarkFocus: numerical autofocusing in digital in-line holographic microscopy using variance of computational dark-field gradient,” *Opt. Lasers Eng.* **134**, 106195 (2020).
56. K. Lang, J. Qiang, Y. Qiu, *et al.*, “Autofocusing method for multifocal holograms based on connected domain analysis,” *Opt. Lasers Eng.* **184**, 108624 (2025).
57. Z. Göröcs, M. Tamamitsu, V. Bianco, *et al.*, “A deep learning-enabled portable imaging flow cytometer for cost-effective, high-throughput, and label-free analysis of natural water samples,” *Light: Sci. Appl.* **7**(1), 66 (2018).
58. M. Tamamitsu, Y. Zhang, H. Wang, *et al.*, “Comparison of gini index and tamura coefficient for holographic autofocusing based on the edge sparsity of the complex optical wavefront,” *arXiv* (2017).
59. T. Shimobaba, T. Kakue, and T. Ito, “Convolutional Neural Network-Based Regression for Depth Prediction in Digital Holography,” *2018 IEEE 27th International Symposium on Industrial Electronics (ISIE)* (2018).
60. H. A. İlhan, M. Doğan, and M. Özcan, “Digital holographic microscopy and focusing methods based on image sharpness,” *J. Microsc.* **255**(3), 138–149 (2014).
61. Z. Ren, Z. Xu, and E. Y. Lam, “Autofocusing in digital holography using deep learning,” in *Three-Dimensional and Multidimensional Microscopy: Image Acquisition and Processing XXV*, vol. 10499 (Spie, 2018), pp. 157–164.
62. S. Woo, J. Park, J.-Y. Lee, *et al.*, “Cbam: Convolutional block attention module,” in *Proceedings of the European conference on computer vision (ECCV)*, (2018), pp. 3–19.
63. J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, (2018), pp. 7132–7141.
64. I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” *arXiv* (2017).
65. D. Terbe, L. Orzó, B. Bicsák, *et al.*, “Hologram noise model for data augmentation and deep learning,” *Sensors* **24**(3), 948 (2024).