

Micron scale 3D imaging with a multi-camera array

Amey Chaware,^a Kevin C. Zhou,^a Vinayak Pathak,^a Clare B. Cook,^a Ramana Balla,^a Kanghyun Kim,^a Lucas Kreiss,^a Bryan Hilley,^b Julia McHugh,^b Praneeth Chakravarthula,^c and Roarke Horstmeyer^{a,*}

^aDepartment of Biomedical Engineering, Duke University, Durham, USA

^bNasher Museum of Art, Duke University, Durham, USA

^cDepartment of Computer Science, University of North Carolina, Chapel Hill, USA

Abstract. We present a novel approach for capturing gigapixel-resolution micron-scale three-dimensional (3D) images of large complex macroscopic objects using a 9×6 multi-camera array paired with a custom 3D reconstruction algorithm. Our system overcomes inherent trade-offs among resolution, field of view (FOV), and depth of field (DOF) by capturing stereoscopic focal stacks across multiple perspectives, enabling an effective FOV of approximately 135×128 degrees and capturing surface depth maps at lateral resolutions less than $40 \mu\text{m}$ and depth resolutions of about 0.5 mm . To achieve all-in-focus RGB (red, green, and blue) composites with precise depth, we employ a novel self-supervised neural network that integrates focus and stereo cues, resulting in highly accurate 3D reconstructions robust to variations in lighting and surface reflectance. We validate the proposed approach by scanning 3D objects, including those with known 3D geometries, and demonstrate sub-millimeter depth accuracy across a variety of scanned objects. This represents a powerful tool for digitizing large complex forms, allowing for near-microscopic details in both depth mapping and high-resolution image reconstruction.

Keywords: 3D imaging; camera array; depth imaging; imaging hardware; multi-view stereo; computational imaging.

Received Mar. 19, 2025; revised manuscript received May 30, 2025; accepted Jun. 19, 2025; published online Jul. 18, 2025.

© The Authors. Published by Chinese Laser Press under a Creative Commons Attribution 4.0 International License. Distribution or reproduction of this work in whole or in part requires full attribution of the original publication, including its DOI.

[DOI: [10.3788/AI.2025.10005](https://doi.org/10.3788/AI.2025.10005)]

1. Introduction

Capturing three-dimensional (3D) images of large complex objects with microscopic resolution is a challenging task due to inherent limitations in traditional imaging systems. Specifically, increasing a camera's resolution generally reduces its field of view (FOV) and depth of field (DOF), while expanding the FOV and DOF compromises resolution. While high-resolution microscopes are effective for small flat objects—such as biological specimens, silicon wafers, and electronic circuit boards—imaging large-scale objects at this level of detail remains challenging. Techniques like LiDAR, time-of-flight (ToF), and multi-view stereo imaging work well for capturing larger objects but lack the spatial resolution required for fine detailed 3D representation. Achieving microscopic resolution for macroscopic objects (on the order of $\sim 10 \text{ cm}$) could transform fields like photorealistic rendering, augmented reality and virtual reality

(AR/VR), and cultural heritage digitization. However, two core obstacles prevent high-resolution imaging of large-scale objects: 1) as the magnification and resolution of imaging lenses increase, the FOV is reduced significantly, limiting coverage for larger objects, and 2) the DOF decreases quadratically as resolution improves, making it difficult to capture curved surfaces with sharp detail. As outlined by Lohmann^[1], these scaling constraints make it particularly challenging to image large non-planar objects with the high spatial resolution required for representation in microscopic detail.

Several advanced imaging techniques have been developed over the years to address these limitations of traditional methods. Structured illumination approaches, for example, project encoded patterns onto the object to estimate the depth at each pixel, achieving precise depth information. However, these methods are limited by the resolution constraints of both the sensor and projection system, restricting lateral resolution. Additionally, they cannot typically capture registered RGB (red, green, and blue) images due to their reliance on monochromatic

*Address all correspondence to Roarke Horstmeyer, rwh4@duke.edu

light^[2]. Multi-view stereo^[3] and structure-from-motion (SfM)^[4] use multiple camera perspectives to infer depth, offering increased FOV. However, they rely on a long DOF to keep the entire scene in focus, which can compromise either resolution or depth range. Recently, multi-camera arrays^[5] have been proposed in microscopy for increasing effective FOV while maintaining high resolution. Yet, these systems are typically limited to small, flat objects with depths of only a few millimeters^[6,7]. While each of these methods excels in specific applications, none have fully overcome the dual challenge of achieving high-resolution 3D imaging across large non-planar surfaces. This leaves a significant gap in the ability to capture macroscopic objects with the fine microscopic precision required for detailed 3D representation.

In this work, we introduce a multi-camera array system designed to overcome traditional limitations in FOV and DOF for capturing large 3D objects with microscopic resolution; see Fig. 1.

Each camera in the 9×6 multi-camera array captures a high-resolution image of a small portion of the object, with overlapping FOVs between adjacent camera views, to produce a seamless imaging area covering about 135 cm^2 . This overlapping FOV also ensures that each point on the object is observed by at least four cameras (except at the boundary), enabling comprehensive multi-view data capture. To address the inherent DOF limitations of high-resolution imaging, we scan the object axially while capturing synchronized images from the entire camera array, creating a focal stack of up to 80 depth layers per camera, thereby capturing a total of $80 \times 9 \times 6$ images. This approach guarantees that every surface point on the object is sharply focused in at least one of the captured depth planes. Using knowledge of object depth at each point, we create an all-in-focus (AiF) composite image where all surface details are in the sharpest focus. Depth information is then extracted using a combination of multi-view stereo (leveraging the multiple

camera perspectives) and depth-from-focus methods to refine accuracy, resulting in a detailed depth map. Finally, we stitch together the RGB and depth data from all cameras, producing a high-resolution, wide-field RGBD (RGB-depth) representation of the object under $40 \mu\text{m}$ resolution.

Our contributions in this work can be summarized as:

- We introduce a rapid 3D scanning approach using a compact array of high-resolution synchronized cameras to capture detailed images of macroscopically curved objects across a large 3D FOV.
- We develop a self-supervised iterative algorithm that leverages stereoscopic and focus cues to jointly produce high-resolution ($<40 \mu\text{m}$ lateral resolution and 0.5 mm depth resolution) AiF RGB images along with precise depth maps of macroscopic objects extending up to 10 cm in depth.
- We qualitatively and quantitatively validate our approach on objects with known depth profiles. We further demonstrate the versatility of our method by digitizing a range of large ($\sim 10 \text{ cm}$) complex curved objects of diverse material types, underscoring the method's generalization and broad applicability.

2. Related Work

3D imaging is a broad and dynamic field of research, with diverse techniques for meeting different needs, depending on the specific application requirements^[8]. Most RGBD imaging techniques can be categorized into active and passive depth estimation techniques. With the emergence of deep learning, several neural network-based methods for monocular and multi-view depth estimation have also emerged. In this section, we provide an overview of the most widely used methods, highlighting the unique challenges involved in capturing large objects ($>100 \text{ cm}^2$) with significant 3D depth profiles ($\sim 10 \text{ cm}$).

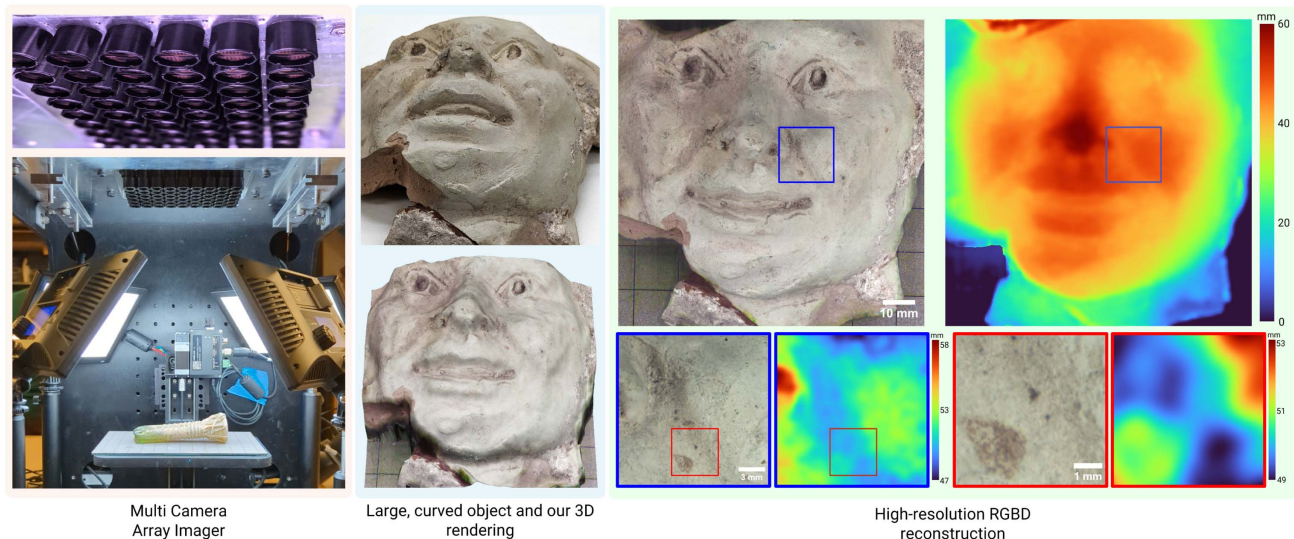


Fig. 1 We present a multi-camera array system that rapidly captures large focal stacks of macroscopically curved objects at microscopic resolutions over a wide FOV. A novel reconstruction algorithm then fuses the acquired imagery into large-area all-in-focus composites and recovers the associated depth maps to provide RGBD scans with uniquely high resolution. We demonstrate our method by obtaining high-resolution RGBD images of curved objects up to 10 cm in size.

Active depth estimation techniques use controlled illumination to recover depth information from objects. Techniques such as structured light and ToF are commonly employed for active depth reconstruction. In structured light, patterns of light are projected onto the scene, and depth is inferred from distortions in these patterns caused by the object's geometry. This approach has been used for precise depth mapping and inferring object geometry^[2,9,10] and to enhance stereo vision^[11]. In ToF sensing, light is emitted onto the scene, and the time taken for the light to return after reflecting off objects is measured^[12–14]. While these techniques offer high accuracy, they generally lack RGB imaging capability, are limited in resolution, and can be affected by factors like ambient light and surface reflectance. Profilometry techniques, such as microscopic fringe projection profilometry (MFPP)^[8,15,16], can provide high-resolution 3D images but are constrained by shallow DoF and have an FoV at least an order of magnitude smaller than our proposed system.

Passive depth estimation methods, on the other hand, use only the information in incoming light to estimate depth, relying on visual and geometric cues from the captured images. These techniques are typically based on stereo or multi-view stereo approaches^[3,17–19], where depth is inferred by analyzing perspective differences in multi-view images captured from slightly different viewpoints. SfM techniques^[4,20–22] reconstruct 3D scenes from image sequences captured by a single moving camera by leveraging the relationship between camera motion and object features. However, these methods typically assume an extended DOF, which can limit achievable resolution. Depth-from-focus^[23,24] offers an alternative by analyzing a focal stack of images to determine the point of sharpest focus for each pixel, thereby estimating depth based on focal depth. While classical approaches typically formulate depth estimation as an explicit photometric cost optimization problem, more recent neural network approaches learn visual features^[25–30] that incorporate non-local information and are more robust to variations in lighting and noise that distort classical depth estimation. Monocular depth estimation techniques have also gained recent attention by leveraging neural networks to learn complex relationships between visual features and depth values^[31–33]. Here, we introduce a multi-camera array system and a tailored 3D reconstruction

algorithm for high-resolution RGBD imaging of large decimeter-scale objects.

3. Multi-Camera Array Imager

Our multi-camera array imager is illustrated in Fig. 2. In the following, we describe our hardware prototype that overcomes the inherent limitations of traditional imaging systems and the associated gigapixel data acquisition process that we later use for high-quality 3D reconstruction.

Imaging Hardware. Our multi-camera array employs a parallelized imaging system comprising 54 compact microcamera modules arranged in a 9×6 grid with 13.5 mm spacing. Each microcamera module uses an Onsemi AR1335 CMOS sensor with a 3120 pixel \times 4208 pixel resolution and 1.1 μm pixel pitch. However, we use a 3072 pixel \times 3072 pixel crop per microcamera, each of which is paired with an $f = 26.23$ mm focal length lens from Supply Chain Optics. The lenses are axially arranged in a finite-conjugate configuration, achieving a magnification of $M \approx 0.11$. All hardware parameter values are provided in the Supplement 1, Table S1. This setup provides $\sim 63\%$ FOV overlap between neighboring cameras (see Fig. 2), enabling multi-view imaging where every point in the system's ~ 125 mm \times 108 mm effective FOV can be viewed by at least four cameras. Our multi-camera array system captures around 700 MP of data per snapshot, and these data are transferred via a field programmable gate array (FPGA) for fast data throughput.

Data Acquisition. Our imaging system features a Zaber X-LSM-E sample stage that moves along the z axis, equipped with a custom object mount to hold objects of interest and four LED panels with diffusers for illumination; see Fig. 2(c). Before data capture, the z position is manually set to focus on the object's top and then is incrementally shifted by 1 mm. At each step, all 54 cameras capture synchronized images until the entire object is out of focus. The largest scanned object on the system, for instance, required 80 z steps, producing an $80 \times 9 \times 6 \times 3072 \times 3072$ data hypervolume, yielding 4320 unique 9.4 MP images and about 40 GB of raw data. Typically, it takes approximately 30 s to acquire this z stack.

Impact of the Hardware Configuration. The lateral resolution of an imaging system is constrained either by the

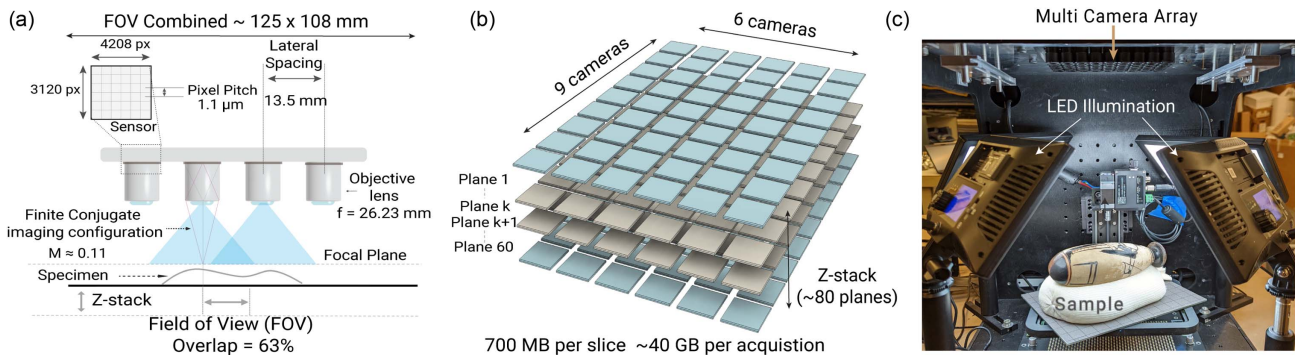


Fig. 2 Overview of the data acquisition process using the multi-camera array imager. (a) Hardware schematic of the multi-camera array, and more than 50% overlap in the FOVs of individual cameras allows imaging over a large continuous area. (b) The multi-camera array scans objects axially to create a focal stack for each camera with a total of $6 \times 9 = 54$ camera images per z slice. (c) Actual setup: The sample is mounted on a stage and illuminated using 4 LED panels placed symmetrically around the sample stage.

sensor pixel size or the diffraction limit and is, respectively, given as

$$r_{\text{pixel}} = \frac{2\mu}{M} \quad \text{or} \quad r_{\text{diff}} = \frac{\lambda}{M \cdot \text{NA}}, \quad (1)$$

where μ is the sensor pixel size, M is the system magnification, and NA is the numerical aperture of the lens. Changing the magnification is feasible in our design, and increasing it improves resolution proportionally. However, this improvement comes at the cost of a reduced DOF, given by

$$\text{DOF} = \frac{4f\mu}{D} \left(1 + \frac{1}{M}\right)^2, \quad (2)$$

where f is the focal length of the lens and D is the lens aperture diameter. See the [Supplement 1](#), Sec. S2 for derivations. Achieving micrometer-scale lateral resolution requires high magnifications, which in turn reduce the DOF to just a few millimeters or less. Consequently, imaging macroscopic objects with surface variations up to 10 cm demands capturing focal stacks to span the full depth range. While increased magnification improves lateral resolution, it also drastically limits the DOF, leading to a substantial rise in data capture requirements. Thus, selecting an optimal magnification involves the balance of the lateral resolution with storage and processing demands.

To leverage stereo depth, our system requires at least 50% overlap in the FOV of adjacent microcameras. This overlap requirement imposes an upper limit on system magnification, given by $M \leq s/(2p)$, where s is the sensor width and p is the inter-camera spacing. For our configuration, this results in a maximum theoretical magnification of $M = 0.17$, corresponding to a minimum lateral resolution of 13 μm and a DOF of 0.55 mm. However, to account for any potential lens distortions and camera positioning errors, we opted for a slightly lower magnification of $M = 0.11$, which increased the overlap in the FOV of adjacent cameras. This adjustment yields a minimum lateral resolution of 20 μm while significantly extending the DOF to 1.22 mm. We selected an axial step size of 1 mm based on this DOF to balance resolution and depth coverage.

The minimum error in stereo height estimation is closely related to the system hardware configuration, given by

$$\delta h = \frac{2\mu f(M+1)}{M(pM+2\mu)}. \quad (3)$$

This expression is derived in the [Supplement 1](#), Sec. S2. In our current setup, the minimum height estimation error is 0.392 mm. However, using feature-based methods could enable further refinement of depth estimations, potentially improving accuracy beyond this theoretical limit.

4. RGBD Reconstruction Method

4.1. Geometric and Photometric Calibration

Before data collection, we calibrate both the geometric and photometric properties of the camera array. Geometric calibration includes determining each camera's 6D pose (3D position and 3D orientation) and a radial distortion parameter shared across all cameras. Photometric calibration addresses intensity

variations within each camera caused by vignetting and pixel response differences, as well as inter-camera variations. To perform the calibration, we capture a focal stack of a patterned, flat target following the procedure outlined in this section. Although the calibration target is flat, a focal stack is necessary in practice to compensate for slight focal plane differences across cameras. For each camera, we select the sharpest plane from the stack by maximizing mean sharpness at each pixel. We then register these images by enforcing geometric and photometric consistency across overlapping areas. Using initial estimates of geometric and photometric properties, we dewarp and backproject the 54 target images onto a common object plane, reproject them into camera space, and apply gradient descent to minimize pixel-wise photometric error, refining both geometric and photometric parameters. A detailed description of the warping and dewarping as well the calibration process is given in the [Supplement 1](#), Sec. S4.

4.2. 3D Reconstruction and Stitching Algorithm

Fusing Stereo View and Depth-from-Focus. Depth-from-focus methods determine depth by locating areas of maximum sharpness within a focal stack. This approach is typically applied with telecentric optics, which maintain a constant, depth-independent magnification to ensure that features remain at the same xy pixel location throughout focal stack acquisition. This consistency allows for direct per-pixel focus or sharpness comparison throughout the stack. However, in our non-telecentric configuration, depth-from-focus imaging methods cannot be directly applied.

Fortunately, even in a non-telecentric configuration, individual features will still be sharpest in the focal stack image where they align with the focal plane, sharing a common magnification at that depth. Thus, in principle, the maximum sharpness method can still provide effective depth estimation for non-telecentric setups. The key difference in non-telecentric systems is that the sharpness of features must be detected among blurred features originating from different spatial locations. Therefore, we use an intensity-normalized sharpness metric (detailed in the [Supplement 1](#), Sec. S3) that is robust to variations in sample appearance, different illumination conditions, and textures. The resulting image stack, processed using this metric on the acquired focal stack, is henceforth denoted as $P(x, y, z)$.

Training Procedure. In each training iteration, we randomly select a 576×576 focal stack patch from one of the 54 cameras. We then identify all other cameras (typically four) that view the same area and extract the corresponding focal stack patches from them. These sets of focal stacks form a minibatch, which we feed into a convolutional neural network (CNN). By encoding the relationship between the object's photometric properties and depth, using a CNN allows us to defer generating a full depth map and RGB composite (with millions of pixels) until inference, resulting in a rapid, scalable algorithm adaptable to varying object sizes. Specifically, we use a U-Net^[34] to predict the depth map for the i th camera as

$$d_{\text{pred},i} = \text{CNN}_{\theta}[s_i(z)], \quad (4)$$

where $d_{\text{pred},i}$ is the predicted depth map, $s_i(z)$ is the captured focal stack, and θ represents the CNN parameters. We optimize the CNN with a set of loss functions that incorporate both

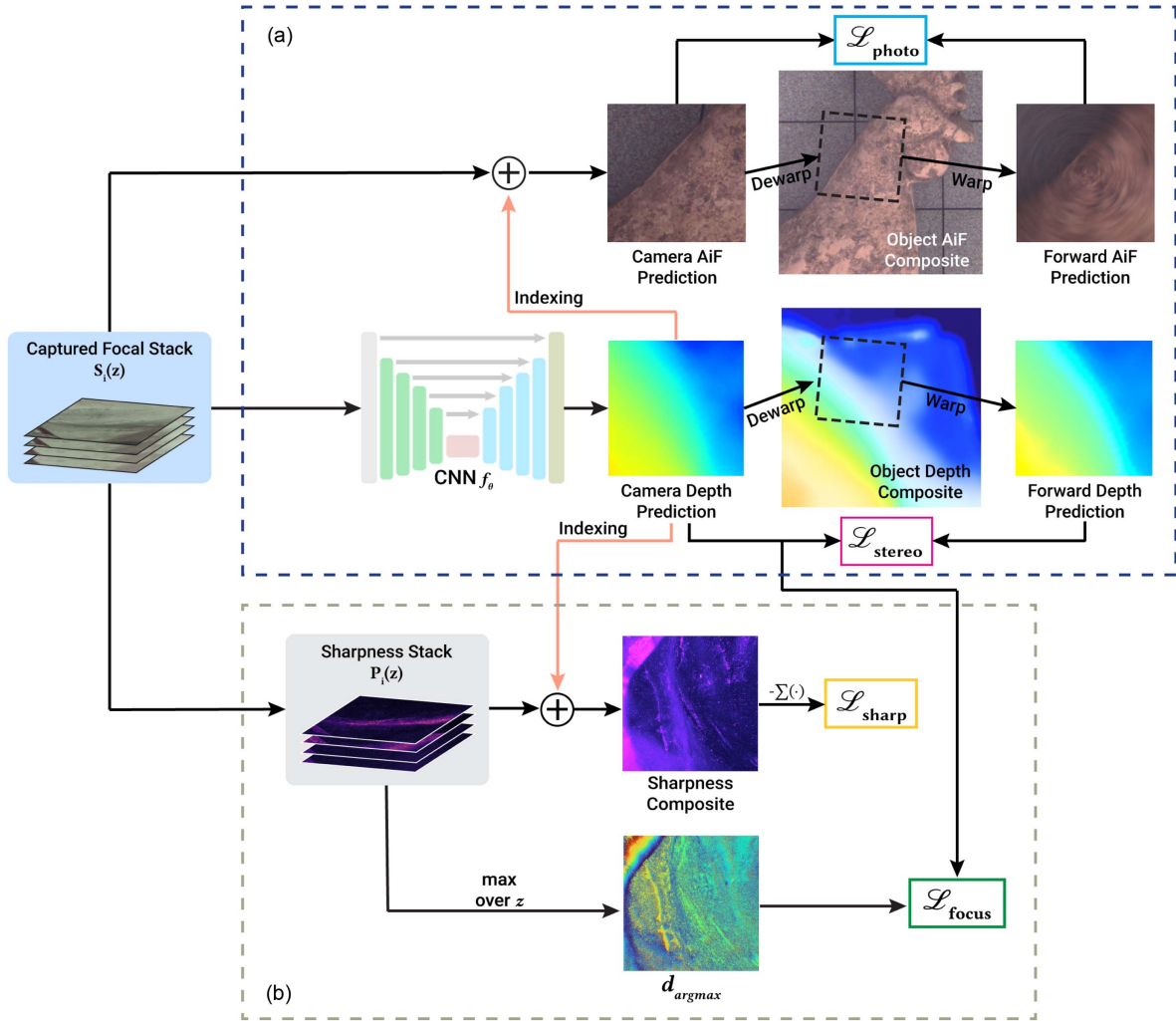


Fig. 3 Joint 3D reconstruction and stitching algorithm. There are two sets of losses to optimize the CNN. (a) Multi-view consistency losses. We use a CNN to convert captured focal stacks into camera-centric depth maps. These depth maps are used to create an AiF RGB image. Both the camera-centric depth map and RGB image are dewarped using calibrated camera parameters to create RGB and depth composites. The composites are rewarped to create forward predictions of camera-centric depth maps and AiF RGB images. The mean square error (MSE) between camera-centric and forward predictions is used to train the CNN. (b) Sharpness losses. The MSE between the depth map produced by the CNN and the depth from sharpest focus, and the negative sharpness of the composite created using the CNN-predicted depth map are also used to optimize the CNN.

sharpness and stereo cues, which we describe next and are depicted in Fig. 3.

Stereo Loss. After generating CNN-predicted depth maps from multiple camera views of the same area, we use the calibrated camera parameters to dewarp depth maps for the individual views and create a depth map composite. We then reproject this composite depth map to each camera view by rewarping it, generating a forward prediction of the camera-centric depth maps, denoted as $\hat{d}_{\text{pred},i}$. To ensure geometric consistency across neighboring camera views, we minimize the difference between CNN-predicted depths $d_{\text{pred},i}$ and $\hat{d}_{\text{pred},i}$. The stereo consistency loss is then given by

$$\mathcal{L}_{\text{stereo}} = \|(d_{\text{pred},i} - \hat{d}_{\text{pred},i})\|_2^2. \quad (5)$$

Photometric Loss. To ensure photometric consistency, we minimize the difference between predicted AiF photometric images. Using the CNN predicted depth map $d_{\text{pred},i}$ for each camera i , we sample from its focal stack $s_i(z)$ (via linear interpolation) to generate a predicted photometric AiF image:

$$C_{\text{AiF},i} = s_i[x, y, z = d_{\text{pred},i}(x, y)]. \quad (6)$$

We dewarp each $C_{\text{AiF},i}$ using the calibrated parameters to create a photometric composite of all cameras viewing the same area. With the composite depth map and camera parameters, we then reproject this photometric composite to obtain a forward prediction of the camera-centric AiF image for each view, denoted as $\hat{C}_{\text{AiF},i}$. We then define the photometric loss as

$$\mathcal{L}_{\text{photo}} = \|(C_{\text{AiF},i} - \hat{C}_{\text{AiF},i})\|_2^2 \quad (7)$$

Sharpness Loss. For each view focal stack $s_i(z)$, we compute a sharpness stack $P_i(x, y, z)$ using an intensity-normalized sharpness metric (see the [Supplement 1](#), Sec. S3). Using this sharpness stack and the CNN-predicted depth map d_{pred} , we create an AiF sharpness composite $P_{\text{AiF},i}$:

$$P_{\text{AiF},i}(x, y) = P_i[x, y, z = d_{\text{pred}}(x, y)]. \quad (8)$$

If the depth estimate $d_{\text{pred}}(x, y)$ is accurate, we assume that $P_{\text{AiF},i}$ is the maximum sharpness composite. To enforce this, we sum a weighted version of $P_{\text{AiF},i}$ across the xy plane and maximize it, defining the sharpness loss as

$$\mathcal{L}_{\text{sharp}} = -\sum_{x,y} P_{\text{AiF},i} \odot \max(P_{\text{AiF},i} - \delta, 0), \quad (9)$$

where \odot denotes the Hadamard product, $\max(\cdot, \cdot)$ denotes the element-wise maximum operator, and δ is a scalar constant number. This weighting excludes the areas of low-contrast regions from contributing to the loss, helping avoid reconstruction artifacts. Empirically, $\delta = 1$ was found to be a good value.

Focus Loss. As mentioned earlier, similar to depth-from-focus methods, our approach estimates depth by applying an arg-max operation over the sharpness volume. The resulting depth estimate, denoted d_{argmax} , is then compared to the CNN-predicted depth d_{pred} using an MSE loss:

$$\mathcal{L}_{\text{focus}} = \|d_{\text{pred}} - d_{\text{argmax}}\|_2^2. \quad (10)$$

While d_{argmax} is not a perfect depth estimate, this loss serves as long-range guidance for the CNN, especially when the predictions differ significantly from the ground truth (GT) depth. This is particularly helpful in reconstructing objects with large height variations. Thus, our final loss function is

$$\mathcal{L} = \alpha\mathcal{L}_{\text{stereo}} + \beta\mathcal{L}_{\text{sharp}} + \gamma\mathcal{L}_{\text{focus}} + \lambda\mathcal{L}_{\text{photo}}, \quad (11)$$

where α, β, γ , and λ are the hyperparameters. The importance and contribution of each term in this composite loss function were validated through an ablation study, demonstrating that the full model achieves superior performance (see the [Supplement 1](#), Sec. S6 for details).

Inference. After training, we feed captured focal stacks into the trained CNN to generate depth map patches. The depth maps are then used to index the focal stacks, creating AiF photometric images. Once this process is completed for all cameras, we apply the calibrated camera parameters to dewarp both the photometric images and depth maps into object space, resulting in the final stitched RGBD composite.

5. Assessment

Here, we describe our experimental evaluation of the prototype multi-camera imaging system and the gigapixel-scale micron-resolution 3D reconstruction algorithm. Specifically, we first quantitatively characterize our method by imaging objects with known depth profiles, and then we demonstrate the system's applicability by creating 3D renders of diverse historical artifacts from a museum.

Implementation Details. The reconstruction algorithm was implemented in TensorFlow v2.10^[35] and run on an NVidia RTX A5000 GPU for optimization and inference. A cosine decay learning schedule was used for all experiments, with an initial learning rate of 10^{-4} , running for a maximum of 160,000 iterations. On our hardware, each iteration took approximately 0.2 s on average. After the training process, generating a full, 150 MP stitched RGBD reconstruction took approximately 4.5 min. The hyperparameters were fine-tuned individually for every object and are provided in the [Supplement 1](#), Sec. S5. The focal stack data from cameras capturing only the background was excluded, and the reconstructed depth maps were smoothed using a weighted median filter.

System Characterization. To characterize and assess the depth prediction accuracy and lateral resolution of our system, we imaged two test objects (including one custom 3D printed) with known depth profiles. The depth estimates obtained from our reconstruction method were compared to the GT by analyzing the line profile of the predicted depth against the actual sample depth. Our method demonstrated submillimeter accuracy for depth predictions. The 3D rendering, recovered depth estimates and high-resolution AiF images of objects from this characterization experiment are shown in Fig. 4.

We used Fourier ring correlation (FRC)^[36] to evaluate the lateral resolution of our AiF images, a method proven effective for unbiased estimation of image quality and resolution in several applications including microscopy^[37–39], image reconstruction and restoration^[40,41], and X-ray imaging^[42]. Given two image views of a scene, FRC calculates the correlation between the two views in the frequency domain, along concentric “rings” of increasing spatial Fourier frequencies. For i th frequency bin r_i ,

$$\text{FRC}(r_i) = \frac{\sum_{r \in r_i} F_1(r) \cdot F_2(r)^*}{\sqrt{\sum_{r \in r_i} F_1^2(r) \cdot \sum_{r \in r_i} F_2^2(r)}}, \quad (12)$$

where F_1 and F_2 are the Fourier transforms of the two views. The spatial frequency where the FRC value drops below a fixed threshold, typically 1/7, is interpreted as the resolution of the full image^[36]. Across our object reconstructions, we achieved FRC resolution values consistently below 40 μm . FRC values across different objects are available in the [Supplement 1](#), Table S4. Next, we discuss the results from two characterization objects.

3D Printed Pyramid. We used a FormLabs 3+ resin printer to 3D print a pyramid with an 86.5 mm wide square base and a height of 24.97 mm, with surface features added via spray paint. The tolerance of this printer is reported to be 20 μm by the manufacturer. Figures 4(a)–4(e) illustrate the reconstruction results, resolution characterization, and line profile. FRC analysis showed a lateral resolution of 34 μm , and the line profile of the predicted sub-millimeter depth map showed a mean absolute error (MAE) of 0.536 mm and a root mean square error (RMSE) of 0.669 mm against the GT, with a standard deviation (STD) of 0.531 mm.

Packaging Foam Piece. We also imaged a cuboidal piece of packaging foam, a challenging object with rough material and naturally small features. The manufacturer-specified height of the piece is 24.60 mm, which we verified using precision calipers. FRC analysis indicated a lateral resolution of 27 μm , and the reconstruction and analysis results are presented in Figs. 4(f)–4(j). The line profile of the predicted depth map

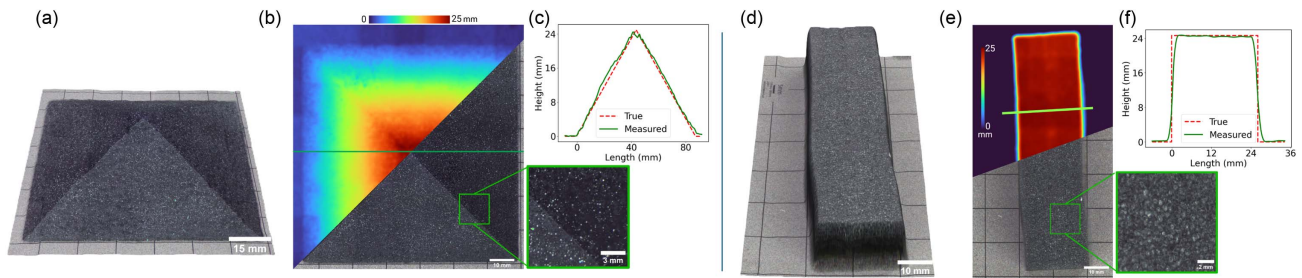


Fig. 4 Characterization of the imaging system. (a) 3D-printed pyramid rendering. (b) All-in-focus composite image and predicted depth map. (c) Line profile of the sample along the green line in (b). (d) Packaging foam piece rendering. (e) All-in-focus composite image and predicted depth map. (f) Line profile of the sample along the green line in (e).

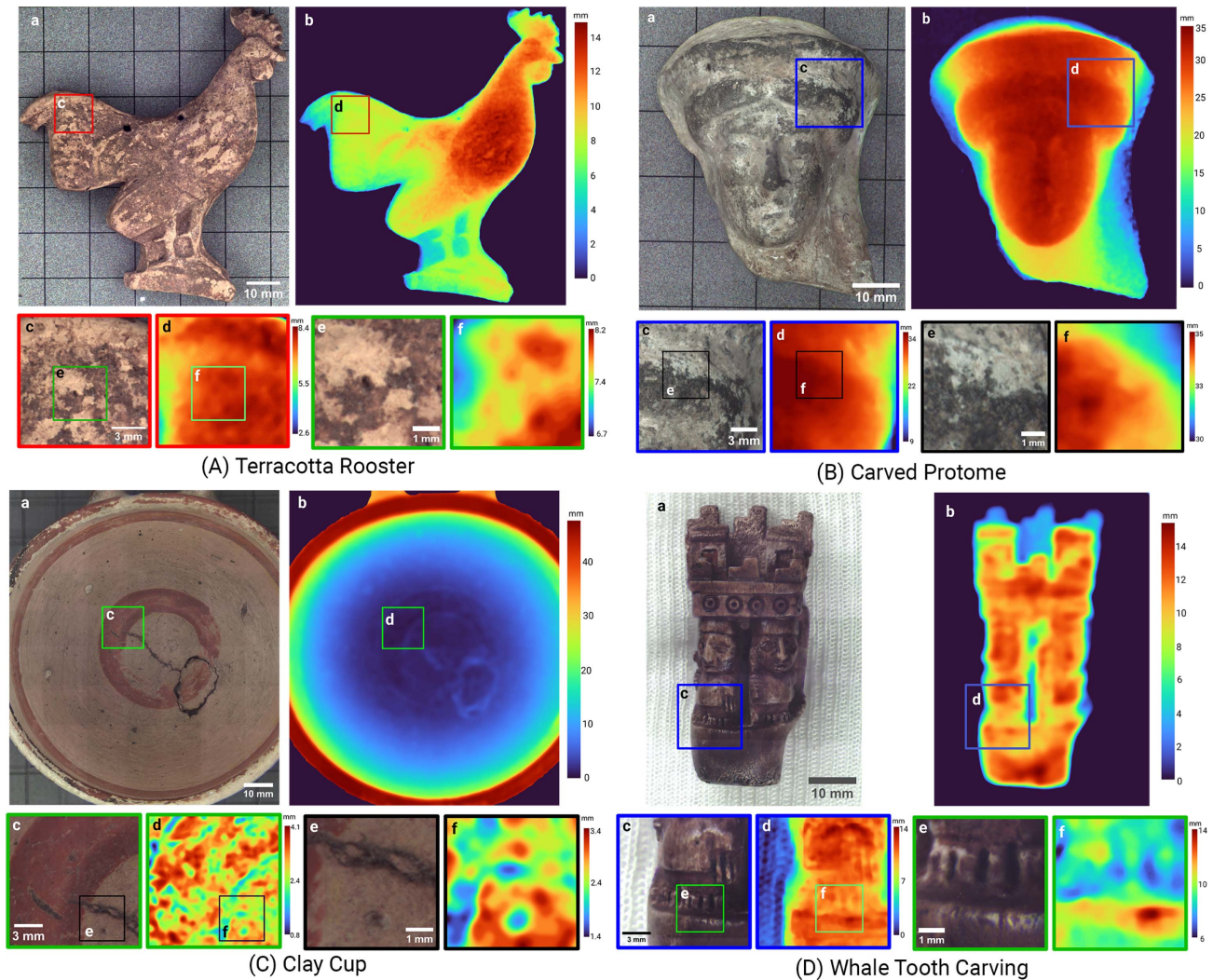


Fig. 5 High-resolution RGBD reconstructions of large objects. For each object, (a) shows the all-in-focus composite image, and (b) shows the depth map. Insets (c) and (e) are enlarged images of the all-in-focus composites, and insets (d) and (f) are enlarged images of the depth maps. Our method can identify small surface features like holes (A) and divots (C), effectively reconstruct objects with height variations much larger than the DOF (B), and is robust to variation in object material (D).

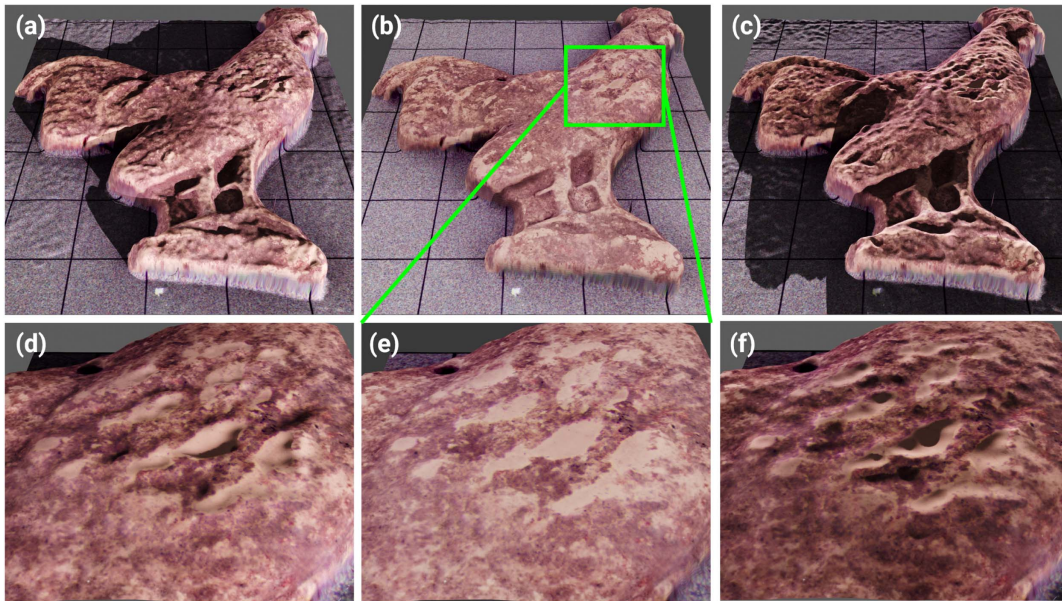


Fig. 6 3D rendering using reconstructed RGBD data. (a)–(c) show 3D renders of an object imaged by our system in different lighting conditions. (d)–(f) show enlarged renders of the marked region under the same lighting conditions. Due to the high spatial and depth resolution afforded by our method, small surface variations can be seen easily.

confirms sub-millimeter depth resolution, with an MAE of 0.919 mm and RMSE of 2.473 mm compared to the GT, with an STD of 2.456 mm. The lower accuracy stems from the discontinuous nature of the object, especially at the edges, where the algorithm’s continuous approximation caused boundary errors.

Digitization and Rendering of Historical Objects. To showcase our method’s ability to image and digitize complex, large-scale objects, we collaborated with a local museum to capture cultural heritage artifacts. These objects varied in size (extending several centimeters) and shape (often with significant 3D depth and curvature), and included diverse material types such as stone, clay, terracotta, whale tusk, and fabric. Several 3D reconstruction results are shown in Fig. 5. Please also refer to the [Visualization 1](#) for 3D renderings of these reconstructed objects. Our method accurately captured the 3D structure across both reflective and diffuse surfaces, offering valuable insights into fine details like pores and cracks. For example, in Fig. 5(A), insets (c) and (d) display a horizontal ridge near the top of the image, while insets (e) and (f) reveal a small pore in the object’s surface. Similarly, Fig. 5(C) showcases the uneven base [insets (c) and (d)] of a clay cup as well as small divots [insets (e) and (f)] in the surface. Figure 5(B) highlights our method’s ability to effectively image highly curved objects, with the photometric reconstruction in insets (c) and (e) remaining sharp despite substantial depth variations, far exceeding the system DOF. Figure 5(D) demonstrates our system’s versatility with different materials and textures. Here, the object—a smooth reflective whale tooth—is imaged accurately alongside a cloth background, with insets showing accurate depth reconstructions for both materials.

Our experimental results in 3D digitization of historical artifacts highlight the capabilities of our imaging prototype and reconstruction method to capture complex objects accurately,

with potential applications across fields requiring high-detail rendering.

For instance, Fig. 6 demonstrates small surface variations in 3D renderings, as revealed by shadows under varied lighting conditions. The renderings are derived from the reconstructed RGBD image of the object shown in Fig. 5(A). This high-resolution capture of surface details—such as fine textures and depth changes under different lighting conditions—supports applications in medical imaging, where precise surface topography aids diagnostics, and in AR, where realistic object rendering enhances immersion. Specific to artifact digitization, our method also contributes to preservation, conservation, and informed storage strategies, reducing risks from handling and transport. By capturing fine features like pores in sculpted materials or highlighting intricate surface textures, our setup not only enriches cultural heritage documentation but also supports their long-term digital preservation.

Overall, these experiments demonstrate that our imaging system and software offer a versatile solution for capturing complex, large-scale objects with sub-millimeter depth accuracy and micron-scale lateral resolution, effectively handling diverse materials and challenging surface geometries. Please see the [Supplement 1](#) for additional results.

6. Conclusion and Future Work

We developed a multi-camera array imaging system and a novel reconstruction algorithm for micron-scale RGBD imaging at gigapixel resolution. Our system achieves detailed 3D scans across a 135 cm² FOV, capturing complex objects with up to 8 cm height variations and spatial resolutions below 40 μm. Without modifying the optics in our current system, attaining higher resolutions is possible with additional scan steps, though this increases scanning and processing time. Additionally, the

system is scalable to larger FOVs, either through lateral scanning or by expanding the micro-camera array, allowing adaptability for larger-scale objects.

Our current system captures and renders detailed views of the top surface of the objects from a single macro-perspective. In future iterations, we aim to incorporate a rotational kinematic mount for multi-perspective scans at resolutions of tens of microns or finer, enabling comprehensive 3D modeling with neural rendering techniques like NeRF^[43] or Gaussian splatting^[44]. Additionally, our system demonstrated accurate height mapping across various shapes and materials, although highly reflective or transparent surfaces present challenges. Integrating structured illumination could effectively address these issues, improving depth accuracy for such complex surfaces. Overall, the proposed system serves as a powerful tool for precision capture of high-resolution intricate surface detail, making it invaluable for applications such as digital archiving and immersive AR/VR.

Disclosures

R. H. is a co-founder of Ramona Optics, Inc., which is commercializing multi-camera array microscopes. R. H. is also a co-founder of Microscopic Image Recognition Algorithms Inc., which is commercializing technology to fingerprint physical objects. K. C. Z. was a consultant for Ramona Optics, Inc. The remaining authors declare no competing interests.

Acknowledgments

P. C. was supported by the NSF (No. 2107454).

References

1. A. W. Lohmann, "Scaling laws for lens systems," *Appl. Opt.* **28**, 4996 (1989).
2. S. Zhang, "High-speed 3D shape measurement with structured light methods: a review," *Opt. Lasers Eng.* **106**, 119 (2018).
3. R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, (Cambridge University Press, 2003).
4. I. Chugunov, Y. Zhang, and F. Heide, "Shakes on a plane: unsupervised depth estimation from unstabilized photography," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), p. 13240.
5. B. Wilburn *et al.*, "High performance imaging using large camera arrays," *ACM Trans. Graph.* **24**, 765 (2005).
6. M. Harfouche *et al.*, "Imaging across multiple spatial scales with the multi-camera array microscope," *Optica* **10**, 471 (2023).
7. K. C. Zhou *et al.*, "Computational 3D topographic microscopy from terabytes of data per sample," *J. Big Data* **11**, 62 (2024).
8. L. Traxler *et al.*, "Experimental comparison of optical inline 3D measurement and inspection systems," *IEEE Access* **9**, 53952 (2021).
9. B. Freedman, "Depth mapping using projected patterns," US Application Publication, US 2010/0118123 A1 (2010).
10. S. R. Fanello *et al.*, "HyperDepth: learning depth from structured light without matching," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), p. 5441.
11. D. Scharstein and R. Szeliski, "High-accuracy stereo depth maps using structured light," in *Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (IEEE Computer Society, 2003), p. 195.
12. S.-H. Baek *et al.*, "Centimeter-wave free-space neural time-of-flight imaging," *ACM Trans. Graph.* **42**, 1 (2023).
13. C. Callenberg *et al.*, "Low-cost SPAD sensing for non-line-of-sight tracking, material classification and depth imaging," *ACM Trans. Graph.* **40**, 1 (2021).
14. A. Gupta *et al.*, "Asynchronous single-photon 3D imaging," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019), p. 7909.
15. L. Chen *et al.*, "Calibration method for a multi-focus microscopic 3D imaging system," *Opt. Lett.* **48**, 4348 (2023).
16. Y. Hu *et al.*, "Microscopic fringe projection profilometry: a review," *Opt. Lasers Eng.* **135**, 106192 (2020).
17. J. L. Schönberger *et al.*, "Pixelwise view selection for unstructured multi-view stereo," in *Computer Vision - ECCV 2016* (Springer International Publishing, 2016), p. 501.
18. Y. Hou *et al.*, "Multi-view stereo by temporal nonparametric fusion," in *IEEE/CVF International Conference on Computer Vision (ICCV)* (2019), p. 2651.
19. K. Wang and S. Shen, "MVDepthNet: real-time multiview depth estimation neural network," in *2018 International Conference on 3D Vision (3DV)* (2018), p. 248.
20. K. C. Zhou *et al.*, "Mesoscopic photogrammetry with an unstabilized phone camera," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), p. 7531.
21. H. Ha *et al.*, "High-quality depth from uncalibrated small motion clip," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), p. 5413.
22. J. L. Schönberger and J.-M. Frahm, "Structure-from-motion revisited," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), p. 4104.
23. Y. Xiong and S. A. Shafer, "Depth from focusing and defocusing," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition* (1993), p. 68.
24. M. Subbarao and G. Surya, "Depth from defocus: a spatial domain approach," *Int. J. Comput. Vis.* **13**, 271 (1994).
25. W. Luo, A. G. Schwing, and R. Urtasun, "Efficient deep learning for stereo matching," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2016), p. 5695.
26. B.-U. Lee *et al.*, "CNN-based simultaneous dehazing and depth estimation," in *2020 IEEE International Conference on Robotics and Automation (ICRA)* (2020), p. 9722.
27. F. Tosi *et al.*, "Learning monocular depth estimation infusing traditional stereo knowledge," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), p. 9799.
28. W. Yin *et al.*, "Enforcing geometric constraints of virtual normal for depth prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019), p. 5684.
29. C. Zhao *et al.*, "Monovit: self-supervised monocular depth estimation with a vision transformer," in *International Conference on 3D Vision (3DV)* (2022), p. 668.
30. H. Si *et al.*, "Fully self-supervised depth estimation from defocus clue," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), p. 9140.
31. L. Yang *et al.*, "Depth anything: unleashing the power of large-scale unlabeled data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2024), p. 10371.
32. I. Chugunov *et al.*, "The implicit values of a good hand shake: Handheld multi-frame neural depth refinement," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), p. 2852.
33. J. Watson *et al.*, "The temporal opportunist: Self-supervised multi-frame monocular depth," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), p. 1164.
34. O. Ronneberger *et al.*, "U-net: convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention-MICCAI 2015 18th International Conference* (2015), p. 234.

35. M. Abadi *et al.*, “TensorFlow: large-scale machine learning on heterogeneous systems software,” <https://www.tensorflow.org/> (2015).
36. R. P. J. Nieuwenhuizen *et al.*, “Measuring image resolution in optical nanoscopy,” *Nat. Methods* **10**, 557 (2013).
37. N. Banterle *et al.*, “Fourier ring correlation as a resolution criterion for super-resolution microscopy,” *J. Struct. Biol.* **183**, 363 (2013).
38. F. Preusser *et al.*, “FRC-QE: a robust and comparable 3D microscopy image quality metric for cleared organoids,” *Bioinformatics* **37**, 3088 (2021).
39. M. Kahnt *et al.*, “Multi-slice ptychography enables high-resolution measurements in extended chemical reactors,” *Sci. Rep.* **11**, 1500 (2021).
40. A. Berberich *et al.*, “Fourier ring correlation and anisotropic kernel density estimation improve deep learning based SMLM reconstruction of microtubules,” *Front. Bioinform.* **1**, 752788 (2021).
41. S. Koho *et al.*, “Fourier ring correlation simplifies image restoration in fluorescence microscopy,” *Nat. Commun.* **10**, 3103 (2019).
42. L. Loetgering *et al.*, “Generation and characterization of focused helical x-ray beams,” *Sci. Adv.* **6**, eaax8836 (2020).
43. B. Mildenhall *et al.*, “NeRF: representing scenes as neural radiance fields for view synthesis,” *Commun. ACM* **65**, 99 (2021).
44. B. Kerbl *et al.*, “3D Gaussian splatting for real-time radiance field rendering,” *ACM Trans. Graph.* **42**, 1 (2023).