

RESEARCH

Open Access



Computational 3D topographic microscopy from terabytes of data per sample

Kevin C. Zhou^{1,2,4*}, Mark Harfouche², Maxwell Zheng², Joakim Jönsson¹, Kyung Chul Lee^{1,3}, Kanghyun Kim¹, Ron Appel², Paul Reamey², Thomas Doman², Veton Saliu², Gregor Horstmeyer², Seung Ah Lee³ and Roarke Horstmeyer^{1,2*}

*Correspondence:
kevinczhou@berkeley.edu;
roarke.w.horstmeyer@duke.edu

¹ Department of Biomedical Engineering, Duke University, Durham, NC 27708, USA

² Ramona Optics Inc., 1000 W Main St., Durham, NC 27701, USA

³ School of Electrical and Electronic Engineering, Yonsei University, Seoul 03722, South Korea

⁴ Present Address: Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA 94720, USA

Abstract

We present a large-scale computational 3D topographic microscope that enables 6-gigapixel profilometric 3D imaging at micron-scale resolution across $>110\text{ cm}^2$ areas over multi-millimeter axial ranges. Our computational microscope, termed STARCAM (Scanning Topographic All-in-focus Reconstruction with a Computational Array Microscope), features a parallelized, 54-camera architecture with 3-axis translation to capture, for each sample of interest, a multi-dimensional, 2.1-terabyte (TB) dataset, consisting of a total of 224,640 9.4-megapixel images. We developed a self-supervised neural network-based algorithm for 3D reconstruction and stitching that jointly estimates an all-in-focus photometric composite and 3D height map across the entire field of view, using multi-view stereo information and image sharpness as a focal metric. The memory-efficient, compressed differentiable representation offered by the neural network effectively enables joint participation of the entire multi-TB dataset during the reconstruction process. Validation experiments on gauge blocks demonstrate a profilometric precision and accuracy of $10\text{ }\mu\text{m}$ or better. To demonstrate the broad utility of our new computational microscope, we applied STARCAM to a variety of decimeter-scale objects, with applications ranging from cultural heritage to industrial inspection.

Keywords: Computational imaging, Terabyte-scale, 3D reconstruction, Camera array, Parallelized

Introduction

All optical imaging systems operate within a trade-off space, in which resolution, field-of-view (FOV), and imaging speed must all be carefully selected for a given application of interest. For example, due to both practical and physics-based constraints, widely used commercial and high-end objective lenses can only resolve a limited number of points within their FOV [1]. This problem tends to get worse at higher spatial resolutions [2], as higher-order aberrations become practically more difficult and expensive to correct over wide FOVs. Furthermore, as lateral resolution increases, an imaging system's depth of field (DOF) becomes quadratically narrower, diminishing the axial FOV and therefore the number of axially resolvable points for 3D imaging applications. As a

result, many imaging techniques designed to capture 3D surface profiles, such as photogrammetry [3], active stereo [4], structured light imaging [5, 6], and line structured light imaging [7], have largely been applied in lower-resolution, low-magnification applications and offer macroscopic FOVs (centimeter and decimeter-scale). There are very few imaging methods that can acquire high, microscopic resolution 3D surface measurements across large areas, which is the central goal of this work.

The ability to jointly measure the 3D properties of large, macroscopic surfaces at high resolution can benefit a variety of applications. Such an instrument could, for example, fully digitize macroscopic 3D objects at microscopic resolution, which would prove valuable within cultural heritage, in particular for the inspection and digitization of artwork [8, 9]. Such a method would also find value within the industrial inspection of electronics components [10], including wafer defect detection [11], printed circuit boards (PCBs) [5, 12, 13], and chip-scale packages (CSPs) [14, 15]. In line with these diverse requirements, a variety of methods aimed at achieving higher-resolution micron-scale profilometry have been extensively explored [16]. Microscale optical interferometry [17–19] and microscopic digital fringe projection [20, 21] are a widely employed method for accurate measurements of microscale structures. Other established contenders in microscale 3D surface measurement are focus detection microscopy [22] and confocal microscopy [23]. More recently, computational imaging methods, previously applied in imaging biological specimens, such as optical coherence tomography [24] and Fourier ptychography [25], have found applications in the inspection and metrology field. However, even though these methods have increased the resolution of profilometry, they still generally operate over smaller FOVs (millimeter-scale). There is thus a need for a high-resolution topographic imaging system that can acquire data at reasonable speeds across large, macroscopic objects, along with new computational strategies that can scalably handle the associated orders-of-magnitude-increased dataset sizes, which could open up a wide range of exciting applications surrounding the topics outlined above.

To this end, to overcome the current 3D measurement throughput limitations of existing 3D profilometric techniques, we present a parallelized computational 3D topographic microscope that can perform 3D surface profilometric imaging at micron-scale resolution over a $13\text{ cm} \times 9\text{ cm} > 110\text{ cm}^2$ lateral FOV and multi-millimeter axial ranges (Fig. 1). Our method, termed STARCAM (Scanning Topographic All-in-focus Reconstruction with a Computational Array Microscope), uses a multi-camera array microscope (MCAM [26–29]) and 3-axis sample scanning to capture a 2.1-terabyte (TB) dataset for each sample of interest. Our MCAM contains a 9×6 array of cameras, in principle allowing for $54\times$ increased throughputs beyond conventional microscopes. Notably, the range of sample scanning is limited to the inter-camera spacing rather than the total extended FOV, and the total scan time is independent of the number of cameras. We then process the multidimensional, multi-TB dataset through a large-scale, self-supervised, neural-network-based reconstruction and stitching algorithm that estimates an all-in-focus (AiF) 6-gigapixel (GP) photometric composite along with a coregistered 3D height map, using the stereo cues from the overlapped lateral scanning and sharpness cues from the axial scanning. The neural network acts as a compressed differentiable representation of the reconstruction that enables a memory-efficient computational reconstruction process, while still

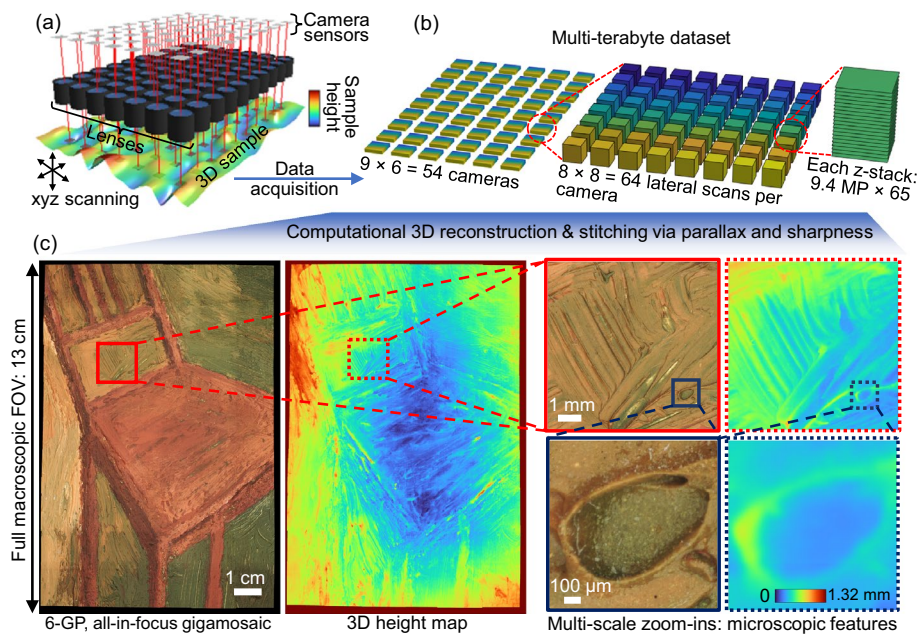


Fig. 1 Overview of STARCAM. **a, b** Data acquisition involves scanning a 3D object in three dimensions and synchronously capturing ~ 9.4 -MP, high-resolution images across an array of 9×6 cameras. Lateral scanning (8×8) is necessary to fill in the gaps between sensors, due to the high magnification, as well as provide stereo information for 3D estimation. z-stacking (65 steps) enables 3D topographic estimation of thicker samples using sharpness measures. **c** Our computational 3D reconstruction algorithm generates a 6-GP, all-in-focus gigamosaic along with a coregistered 3D height map

effectively allowing joint participation of the entire multi-TB dataset by loading and preprocessing random z-stack patches from storage to computer memory on the fly.

We applied STARCAM to a wide variety of decimeter-scale 3D objects, including an oil painting, PCB, and multiple CPU pin grid arrays (PGAs) and ball grid arrays (BGAs) in parallel. Our multi-GP 3D topographic microscopy technology paves the way to a solution to the high-throughput imaging demand of future industrial inspection applications that require in-line monitoring of complex parts at multiple stages of fabrication.

Multi-terabyte data acquisition for high-SBP 3D topographic microscopy

Multi-camera array microscope (MCAM) hardware design

The MCAM's highly parallelized design consists of an array of $9 \times 6 = 54$ micro-camera units, spaced by 1.35 cm in both lateral dimensions, each with a 13-MP Bayered CMOS sensor (Onsemi AR1335, 3120×4208 , pixel size = $1.1 \mu\text{m}$) [27]. For all experiments in this study, we used 3072×3072 square crops. Each micro-camera is equipped with a 25.05-mm effective focal length (EFL) lens (Edmund Optics), axially positioned to form a finite-conjugate, non-telecentric imaging configuration with a magnification of $\sim 0.8 \times$ (corresponding an object-side digital resolution of $1.378 \mu\text{m}$) and an object-side numerical aperture (NA) of ~ 0.088 . The working distance is about 39 mm. The sample is illuminated using white LEDs surrounding the micro-camera apertures. The image data from all 54 sensors are routed via a single FPGA to the computer memory via PCIe [28].

Multi-terabyte data acquisition

The inter-micro-camera spacing is larger than the per-micro-camera FOV (4.1 mm), leading to gaps in the extended FOV for a single synchronized snapshot. Furthermore, the sample height variation typically extends beyond the system depth of field (DOF). To fill in these lateral gaps and to cover an extended axial range, we placed the samples on a 3-axis translation stage (Zaber X-LSM) and translated them laterally in an 8 by 8 xy grid, and axially across 65 z planes, spanning up to 4.8 mm for the samples analyzed in this paper. Note that the stage only needs to travel laterally less than the inter-camera spacing (1.35 cm) and not the entire FOV, resulting in faster acquisition times compared to single-aperture systems covering the same area. In particular, we fixed xy translation step size to $13.5/8$ mm = 1.6875 mm, resulting in >50% overlap redundancy between adjacent scans to provide stereo information to facilitate 3D estimation.

In sum, for each sample, we capture synchronized snapshots across up to 54 cameras across $8 \times 8 \times 65$ scan positions, resulting in a 7D data hypervolume, with dimensions of $9 \times 6 \times 8 \times 8 \times 65 \times 3072 \times 3072$, corresponding to 224,640 9.4-MP images, or 2.1 TB of raw data per sample (saved as unsigned 8-bit integers, with a single bayered channel). The MCAM is able to stream data at >5 GB/sec, theoretically enabling capture of the entire 7D hypervolume in only 7 min. In practice, our data acquisition speeds are limited by the stage translation speed and settling time. The acquisition speed could also be further increased using sample-adaptive strategies, such as with tunable-focus lenses.

Mechanisms for 3D estimation: multi-view stereo and height from sharpest focus

Since we translated the sample axially and ensured >50% lateral overlap across the entire FOV, we were able to take advantage of two different methods for 3D estimation simultaneously, 3D from stereo vision [30] and sharpest focus [31], even though the experimental requirements of these two methods generally conflict. In particular, while stereo techniques require long DOFs, which determine the axial range of 3D estimation, sharpness-based approaches work best with narrow DOFs, which dictate the axial resolution of 3D estimation. Furthermore, while stereo techniques typically use a pinhole camera model, which implicitly assumes non-telecentric optics, sharpness-based approaches work best used with telecentric optics (or make a telecentric approximation over a restricted FOV). Despite these seemingly disparate requirements, we show that our method is able to take advantage of both physical mechanisms synergistically for robust 3D estimation at a few-micron lateral resolution over >10 cm FOVs.

Height from sharpest focus

Estimating depth from z -stacks using a sharpness metric applied slice-wise is perhaps most intuitive and straightforward with telecentric optics, which guarantee a constant, depth-independent magnification. This property ensures that every feature within the lateral FOV stays in the same place during z -stack acquisition, thereby

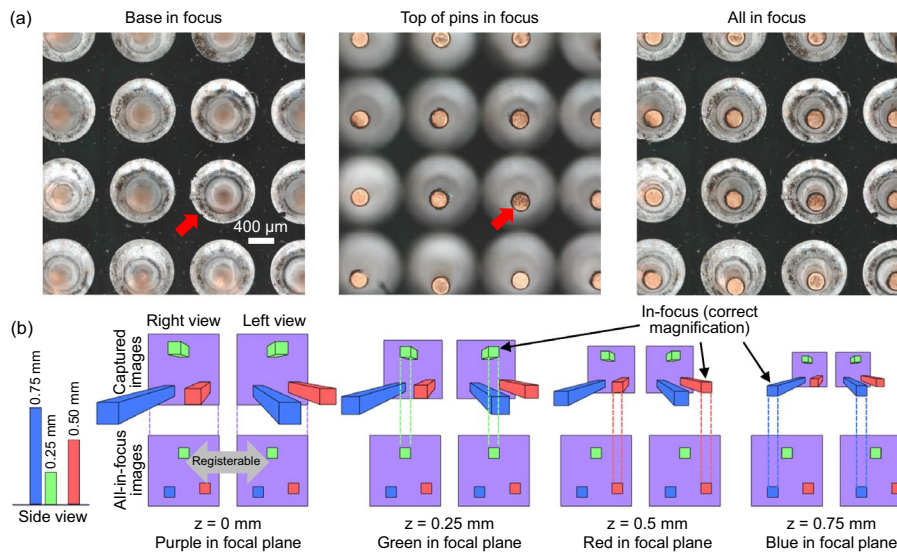


Fig. 2 The two mechanisms for 3D estimation. **a** Image sharpness can be used to estimate height. The left image shows the base of a pin grid array (PGA) in focus, while the middle image shows the tips of the pins of the PGA in focus. The right image shows the all-in-focus reconstruction. **b** Height can also be estimated by eliminating parallax. Only when a feature is at the nominal focal plane of the imaging system does it have the correct magnification and therefore in the correct location (no parallax). When the feature is out of focus, it moves away from or towards the optical axis (parallax). When all features are in focus, the images from different cameras are directly registerable using homographies

enabling direct sharpness comparisons. Thus, the argmax operation (i.e., the position of the largest value) across the z dimension applied pixel-wise would yield a good estimate of the sample height map. However, telecentric optics are disadvantageous in that they restrict the lateral FOV (i.e., to less than the diameter of the lens).

Perhaps counterintuitively, even for z -stacks captured with non-telecentric imaging configurations (as in our MCAM design), the same argmax operation would still in principle yield an accurate height estimate. That is, even though the z -stack exhibits depth-dependent magnification, the individual object features will be the sharpest when they intersect with the focal plane and thus will be governed by a common magnification (Fig. 2b). The difference from the telecentric case is that in the non-telecentric case, except at the very center of the FOV, the argmax operation must be able to identify the sharpest object features among blurred versions of other object features from different lateral positions. Thus, the sharpness metric should ideally be robust to changes in object appearance (see the weighted sharpness loss in “Self-supervised training”). Incidentally, correcting for depth-dependent magnification changes to enable more direct sharpness comparisons would essentially be tantamount to estimating the height map. We thus additionally use stereo cues to improve 3D estimation.

Height from stereo

As mentioned earlier, the data acquisition procedure ensured at least 50% overlap in both lateral dimensions. Thus, any point apart from those on the outer edges are viewed from at least four different perspectives. This multi-view information provides stereo parallax cues for 3D estimation [28, 30, 32]. Note that in combination with the z -stacking, we

overcome the implicit resolution limits of stereo techniques, imposed by their requirement for long DOFs. The way stereo information is incorporated is by enforcing consistency in the height values predicted based on the sharpness cues in overlapping regions. In particular, since each z-stack results in radial expansion about its respective center due to depth-dependent magnification, in overlapped regions of neighboring views the expansions occur in opposite directions. This is known as parallax. While our previous plane-plus-parallax implementations use orthorectification [28, 32] (i.e., undo these radial shifts due to depth-dependent magnification) by deforming a single image, in our case we retrieve the correct pixel from the corresponding depth slice of the z-stack.

3D topographic reconstruction and stitching algorithm

The STARCAM computational 3D topographic reconstruction algorithm extends our previous self-supervised 3D-RAPID algorithm [28] to much larger spatial scales, ranging from 40× to 650× larger SBPs per frame and 3–4 orders of magnitude more raw data per frame. Like 3D-RAPID, our algorithm jointly reconstructs across the entire FOV, requiring the simultaneous participation of z-stacks across all micro-cameras and all scan positions. This strategy promotes a globally-consistent reconstruction that would not be possible with a sequential algorithm. As such, we had to develop new data handling strategies, especially considering that the 8-bit images would not only need to be converted to 32-bit floats for computation, but also debayered into 3 color channels, which would otherwise cause the 2.1-TB dataset size to balloon up to 25.2 TB and therefore easily exceed the RAM capacity of even the highest-end of current consumer computers. Furthermore, it is non-ideal to make copies of the dataset refactored for batched training or debayer in advance, due to the increased data storage requirements. Instead, our algorithm trains a CNN on mini-batches of z-stack patches that are loaded from computer storage and debayered on the fly, enabling effectively joint optimization across the entire dataset using a single GPU.

Joint camera and sample scan calibration: 6D pose, distortion, intensity variation, and focal plane shift

To maximize registration, stitching, and 3D estimation accuracy, we pre-calibrated the microscope by capturing a 7D hypervolume of a flat patterned target. Note that a z-stack is still necessary for a flat target because the focal planes of the 54 micro-cameras in practice may not be coincident. For each z-stack, we pick the sharpest slice (using a Laplacian-based metric) and jointly registered the $9 \times 6 \times 8 \times 8 = 3456$ images by simultaneously optimizing the 6D poses (3D position + 3D orientation) for all 3456 images, along with a quadratic radial distortion parameter, assumed to be the same for all images, and inter-image and intra-image intensity variation (e.g., due to uneven illumination). The pixel-intensity-based joint registration algorithm extends our previously reported algorithm [32] by incorporating new strategies to support reconstructions with much higher SBPs, as required by the large datasets acquired for this study.

Before we discuss the new extensions, we briefly review the base algorithm [32], which takes in a collection of images to be registered according to an image deformation model using gradient descent. For the case of camera pose and distortion calibration, the deformation model is the 6D pose modeling homographic distortions, along with distortion

parameters. All the images are dewarped according to an initial guess of the deformation parameters and projected onto a blank canvas to form an estimate of the registered composite. When a pixel in the composite is visited multiple times (i.e., a collision), the values are averaged. To quantify how accurate the deformation parameters are, the values are reprojected from the composite back to the image space and compared to the original images via mean square error (MSE). This MSE is minimized with respect to the deformation parameters via gradient descent, reaching a minimum when all the pixel collisions are consistent. Note that the reconstruction is reset at every gradient descent iteration.

This algorithm has worked well for jointly calibrating tens of multi-MP cameras that form composites of up to a few hundreds of MP. To extend this algorithm to jointly register 3456 images to form a 6-GP composite, we introduced a multi-scale strategy along with pixel batching. In particular, first we optimize the deformation model parameters for the data captured by a single micro-camera (i.e., 8×8 images of size 3072×3072), using the previous method. An example registration is visualized in Additional file 1: Fig. S6a. The optimized parameters for the one micro-camera are then used to initialize the parameters for all micro-cameras, with a rough initial guess for the inter-camera spacing. From there, all 3456 camera images are coarsely registered with 53× linear downsampling, updating only the lateral camera positions during gradient descent. Next, the downsampling is decreased to 26× and the full 6D poses for all 3456 images are simultaneously updated.

Finally, the downsampling is decreased to 4×, which makes the multi-view image dataset too big for the previous approach. To overcome this challenge, instead of having all pixels from all images contributing to gradient descent, we select a random batch of pixels across all images at each gradient descent iterations. Here, the reconstruction is not reset after every iteration, but rather is updated as a moving average across sequential batches. This moving average image registration approach is similar to our previous algorithms [32, 33], except the batches are random pixels rather than random multi-view images or image columns. In this final step, all calibration parameters are optimized. An example registration of all 3456 images is visualized in Additional file 1: Fig. S6b.

Joint reconstruction of 3D topography across the entire FOV

Once the cameras are calibrated, the goal is to reconstruct an all-in-focus (AiF) photometric (RGB) gigamosaic along with a coregistered 3D height map. To do this in a compact and memory-efficient manner, we optimize a CNN in an end-to-end fashion to map from z-stacks to the 3D height map via self-supervised learning. Our training procedure does not require any data beyond the 7D hypervolume for the sample of interest, with the supervision coming from the stereo and height-from-sharpest-focus cues (See “[Mechanisms for 3D estimation: multi-view stereo and height from sharpest focus](#)”). In other words, the 3D reconstruction procedure itself is the training procedure of the CNN, which acts as a compact representation of the object’s photometric and height properties. This reparameterization leads to a memory-efficient reconstruction algorithm because the decompressed 6-GP photometric and 3D height map gigamosaics never materialize until after training, during CNN inference. Furthermore, the CNN confers regularizing effects due to their inductive biases [34] and compressive

representation (i.e., the CNN contains fewer parameters than the number of pixels in the gigamosaics). The CNN architecture is identical to the one used in 3D-RAPID [28] (Additional file 1: Sec. S4), except that the color channel of the CNN input is replaced with the z-stack dimension, and only the debayered green channel (the most sensitive channel) is used. The reason for choosing just the green channel was to reduce the extra computational overhead of either summing the three color channels or debayering to grayscale, given that our optimization procedure is bottlenecked by batch preparation for GPU consumption.

Self-supervised training

The self-supervised, patch-based CNN training algorithm of STARCAM is summarized in Fig. 3a, b. At each gradient descent iteration, random spatial patches of shape 576×576 are hierarchically sampled by first picking a random camera (out of 54), followed by picking a random xy coordinate from the FOV spanning the 8×8 scan of that camera. All the images that capture that point, according to the camera calibration (Additional file 1: Fig. S6) are identified, which can range from 2 to 9 images. Within each of these images, we identify the 576×576 patch centered around that selected point. If the point is too close to the edge of the image, then we shift the patch until its within the image. Further, the pixel coordinates are rounded to the nearest even number to ensure correct alignment of the bayer pattern. Finally, we load from storage (NVMe SSD: Sabrent 4TB Rocket 4 Plus or Kioxia 7.68TB CD6-R) the 2–9 cropped z-stacks, corresponding to these patches, each with shape $576 \times 576 \times 65$ – this is the first time any sample data has been loaded. We then debayer each image and arbitrarily select only the green channel. Let the i th z-stack patch be denoted as $g_i(z)$. This procedure constitutes the generation of a single batch element. In our experiments, we use a batch size of 2. Currently, our optimization is bottlenecked by batch generation, taking around $2.5 \times$ longer to generate a batch on the CPU (Intel Xeon CPU

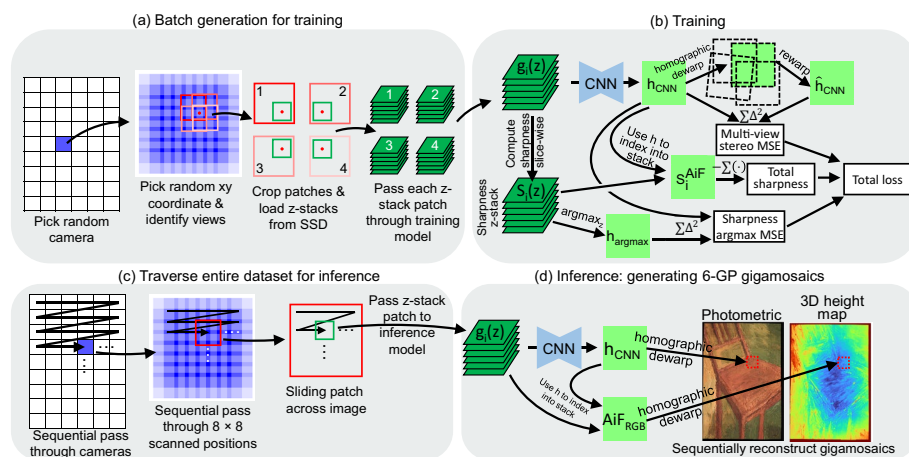


Fig. 3 Computational 3D topographic and all-in-focus (AiF) self-supervised reconstruction algorithm, using stereo and focus cues for supervision. **a** Procedure for generating a single batch element, consisting of a collection of 2–9 z-stack patches (the case of 4 illustrated here). **b** The batches of z-stack patches are passed through the computational model to generate a scalar loss to be minimized by gradient descent. **c**, **d** CNN inference for generating the 6-GP photometric and 3D height map gigamosaics

E5-1650 v4 or Intel Xeon W-2245) than for the GPU (NVIDIA GeForce RTX 3090) to consume it in a gradient update step. NVMe drives with faster read speeds will help bridge this gap.

The batches of collections of z-stack patches ($\{g_i(z)\}_i$) is generated on the CPU, after which they are transferred to the GPU for the gradient descent step. Each z-stack patch is passed through the CNN, which predicts the height map, h_{CNN} for that patch (Fig. 3b). We use three different loss functions to evaluate the fidelity of h_{CNN} that encapsulate the two physical mechanisms for 3D estimation described in "[Mechanisms for 3D estimation: multi-view stereo and height from sharpest focus](#)":

1. Stereo loss. The height maps, h_{CNN} , corresponding to patches from different views are dewarped according to the precalibrated camera parameters (the homographic dewarp step shown in Fig. 3b), and then superimposed and averaged in overlapped regions. The patches are then reprojected back to camera-centric coordinates to form \hat{h}_{CNN} for comparison with h_{CNN} via MSE. Thus, minimizing this loss promotes accurate registration of the h_{CNN} patches, enforcing stereo consistency between neighboring camera views.
2. Weighted sharpness. Each pixel of h_{CNN} can be converted into a depth index to retrieve values from the original z-stack, $g_i(z)$ (indeed, this is how the AiF image is generated). However, instead of indexing into the original z-stack, we first compute a sharpness z-stack. To this end, we first divide each image in $g_i(z)$ by a Gaussian-blurred version ($\sigma = 8$ pixels):

$$g_i^{hpf}(z) = \frac{g_i(z)}{g_i(z) \otimes Gauss_{2D}}. \tag{1}$$

This operation can be thought of as a normalized high-pass filter, to facilitate sharpness comparisons across different spatial regions due to depth-dependent magnification changes ("[Height from sharpest focus](#)"). We then compute the magnitude of discrete spatial gradients, which we blur with the same Gaussian kernel,

$$S_i(z) = \left| \nabla_{x,y} g_i^{hpf}(z) \right| \otimes Gauss_{2D}, \tag{2}$$

which is the final sharpness metric. Ignoring padding issues, $S_i(z)$ and $g_i(z)$ have the same shape (i.e., $576 \times 576 \times 65$). We then use h_{CNN} to index into $S_i(z)$, generating $S_i^{AiF}(z)$. Note that to preserve differentiability, the indexing process using the height map interpolates between the two closest depth slices. Finally, we sum a weighted version of S_i^{AiF} across lateral space to generate the final weighted sharpness loss,

$$loss_{sharpness} = \sum_{x,y} S_i^{AiF} \odot \max(S_i^{AiF} - \delta, 0), \tag{3}$$

where \odot denotes element-wise multiplication, $\max(\cdot, \cdot)$ is an element-wise maximum operation (equivalent to numpy's `maximum` function), and δ is a scalar constant hyperparameter. This loss gives higher weight to spatial regions that have high sharpness, and excludes low-sharpness regions from contributing to the loss (via δ), which we empirically found to avoid artifacts in low-contrast samples (Additional

file 1: Sec. S3). We can also generate a confidence map based on the max sharpness across the axial dimension (Additional file 1: Sec. S3).

3. Argmax. We also apply the argmax operation to $S_i(z)$ (Eq. 2) across z to generate a height map estimate, h_{argmax} , which we compare with h_{CNN} via MSE, weighted by the same factor as in Eq. (3). While h_{argmax} sometimes produces artifacts, it provides long-distance gradients (i.e., when $h_{CNN} - h_{argmax}$ is large, the argmax loss still provides guidance, unlike the weighted sharpness loss).

A weighted sum of these three loss terms constitutes the total loss that is minimized via gradient descent.

Inference: generating the 6-GP gigamosaics

Once the CNN that maps z-stacks to height maps is trained, we can apply the CNN to the entire 2.1-TB dataset to generate the 6-GP gigamosaics (Fig. 3c, d). Specifically, we use a sliding 576×576 window across each 3072×3072 image, looping through the each of the 8×8 lateral scans of each of the 9×6 cameras. We thus read 576×576 z-stacks (with 65 slices each) sequentially from storage, which we feed through the trained CNN to generate height map patches. These height maps are then used to index into the z-stacks to generate the AiF photometric image (AiF_{RGB}). Note that only the green channel is fed into the CNN while all three color channels are used for creating the AiF image patch. Finally, these height map and AiF image patches are homographically dewarped according to the precalibrated camera parameters to enable correct placement of the patches within the gigamosaic. After looping through the entire dataset, we obtain the 6-GP photometric gigamosaic with the coregistered 3D height map.

As we're accumulating patches, in regions where patches overlap, we have the option to blend, that is, average overlapping pixels (see Additional file 1: Sec. S2 for discussion on blending). In Figs. 5, 6, 7 and 8, we show the global photometric views using blending and zoom-ins without blending, but we show all 3D height maps with blending.

Results

System characterization

We characterized the lateral resolution of STARCAM by imaging a USAF test chart at the center and edge of the FOV of a single camera view (Additional file 1: Fig. S1a). Our system can resolve group 7 element 6, corresponding to a bar width of $\sim 2 \mu\text{m}$ (or a full-pitch resolution of $4 \mu\text{m}$). The DOF of our system is ~ 0.37 mm, full-width at half maximum (FWHM) (Additional file 1: Fig. S1b).

To characterize the height accuracy and precision, we applied our method to image four precisely machined (sub-micron accuracy) gauge blocks (Mitutoyo), with heights of 1.000, 1.010, 1.030, and 1.060 mm (Fig. 4). The axial step size of the z-stack was $2.5 \mu\text{m}$. The distributions of the height estimates of the gauge blocks across the imaged area are shown in Fig. 4e. The mean height estimates of the gauge blocks were, respectively, 0.998, 1.007, 1.024, and 1.072 mm. Note that since the height map has an arbitrary offset, we set the offset to that which minimizes the mean square error between the means and the ground truths. The accuracy is thus sub- $10 \mu\text{m}$ (mean absolute error = $5.8 \mu\text{m}$; root mean square error = $6.8 \mu\text{m}$). The

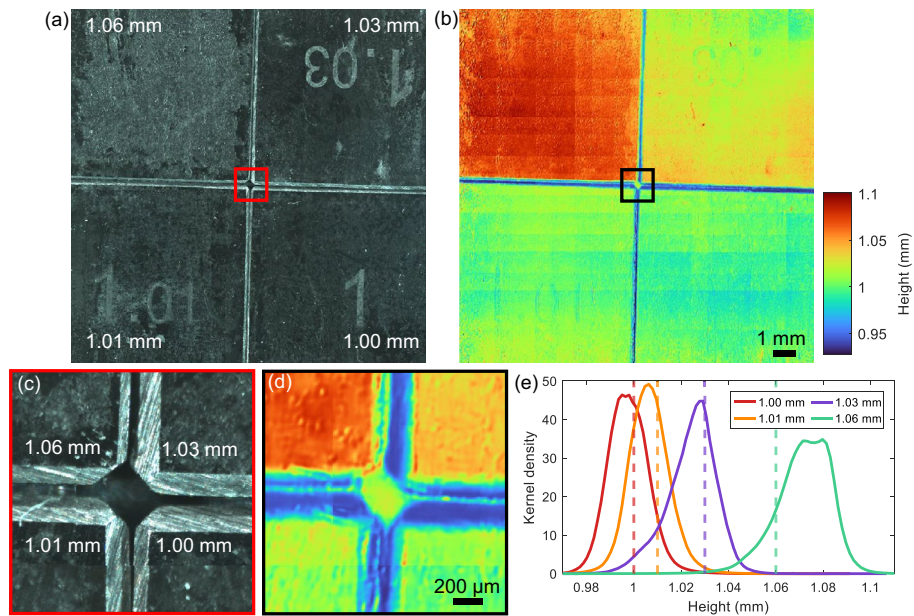


Fig. 4 Axial accuracy and resolution characterization. **a** All-in-focus photometric composite of four gauge blocks with the denoted heights. **b** 3D height reconstruction of the four gauge blocks. **c, d** Zoom-ins of **(a)** and **(b)**. **e** Kernel density estimates of the distribution of the height values of the four gauge blocks. Dotted vertical lines are the ground truth heights. The accuracy (root mean square error) and precision (standard deviation of the height estimates) are $\leq 10 \mu\text{m}$

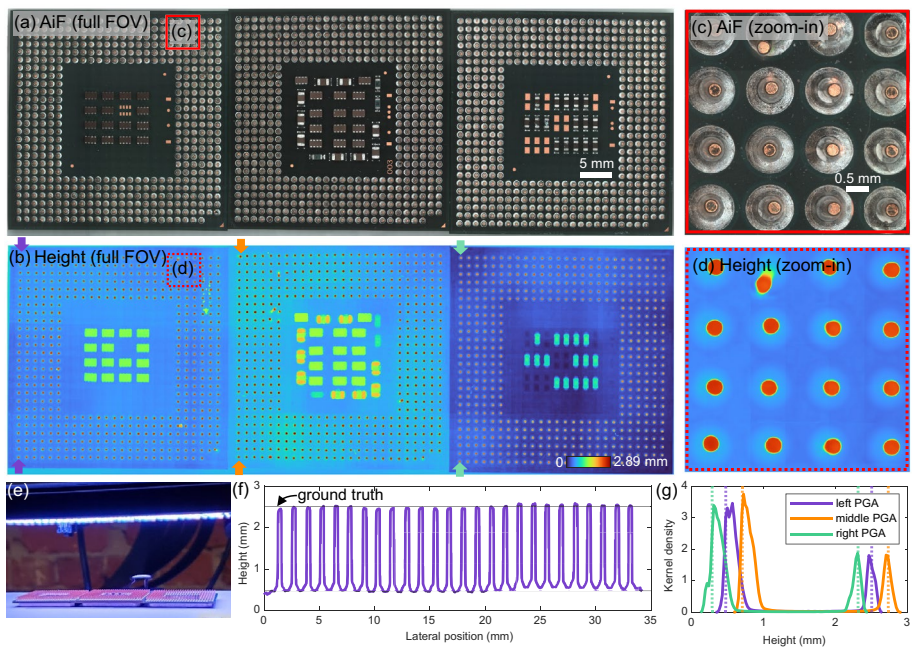


Fig. 5 Three Intel Socket 478 pin grid arrays (PGAs) imaged in parallel. **a** All-in-focus (AiF) photometric gigamosaic. **b** 3D height map. **c** Zoom-in of the region denoted in **a** and the corresponding height map **(d)**. **e** Photograph of the three PGA samples imaged by STARCAM. **f** 1D cross section of a row of pins denoted with purple arrows in **b**. Horizontal lines are separated by 2.03 mm, the nominal pin height specification (ground truth). **g** Kernel density estimates of the height distribution of the pins in the rows denoted by the arrows in **b**. Vertical dotted lines are separated by 2.03 mm for each PGA, the ground truth

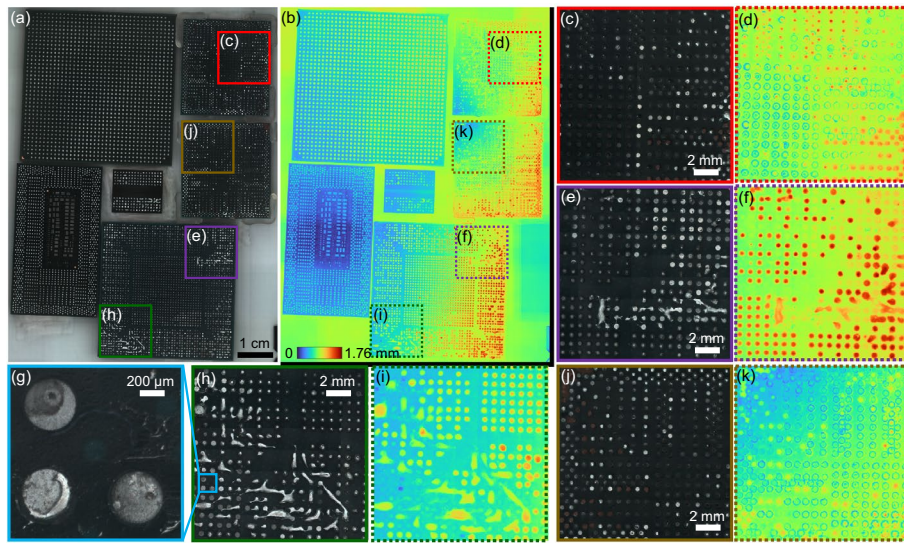


Fig. 6 An assortment of six ball grid arrays (BGAs) imaged in parallel. **a** All-in-focus photometric gigamosaic. **b** 3D height map. **c, e, h, j** Zoom-ins of **a**. **d, f, i, k** Zoom-ins of **b**. **g** Zoom-in of **h**

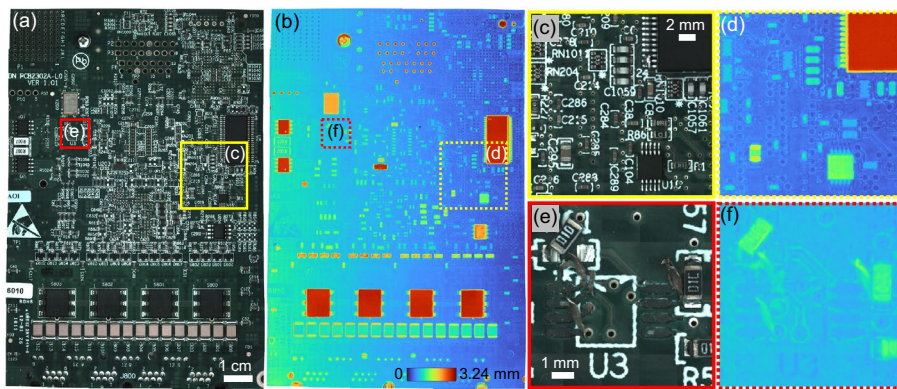


Fig. 7 Printed circuit board (PCB). **a** All-in-focus photometric gigamosaic. **b** 3D height map. **c–f** Zoom-ins of **a, b**

precision, quantified by the standard deviation of the height estimates of the gauge block, is approximately 10 μm . Thus, the height estimation accuracy and precision are better than the system DOF. Finally, note that the gauge blocks have beveled edges, whose lower heights are correctly predicted. Our 3D height map assigns an arbitrary value to the gaps in between the gauge blocks, as nothing comes in focus in the z-stack (Fig. 4d).

3D topographic reconstructions of a variety of samples

We next applied STARCAM to four different extended samples: a pin grid array (PGA) (Fig. 5), ball grid array (BGA) (Fig. 6), printed circuit board (PCB) (Fig. 7), and oil painting (Fig. 8).

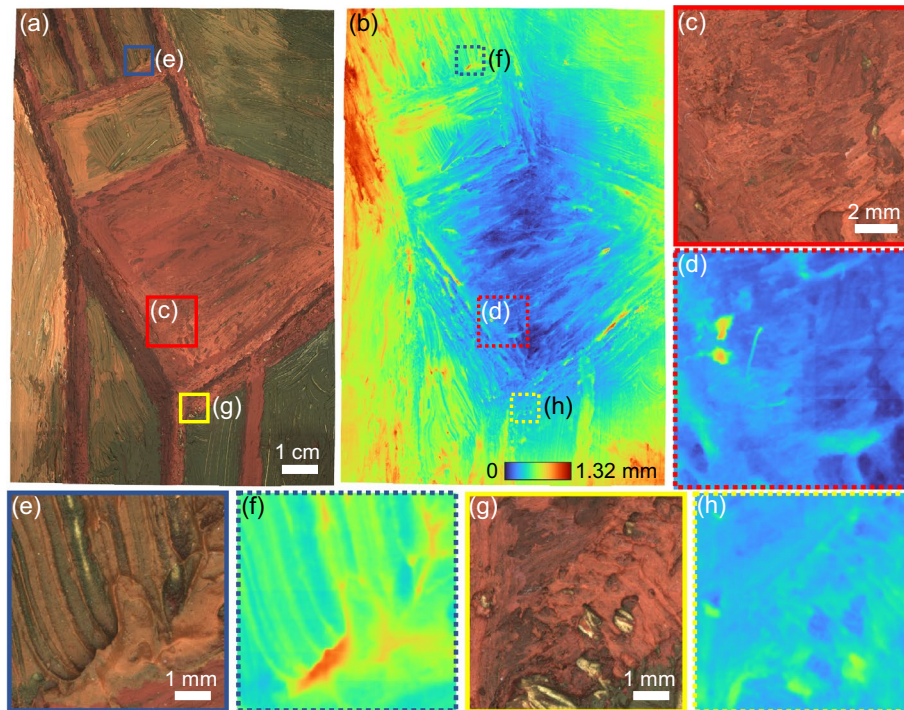


Fig. 8 Oil painting of a chair. **a** All-in-focus photometric gigamosaic. **b** 3D height map. **c–h** Zoom-ins of **a**, **b** reveal the topographies of different brush strokes

Pin grid array

We imaged three PGAs in parallel, each based on Socket 478 used in Intel's Pentium 4 CPU processors. They each contain an array of 478 pins covering a $35 \times 35 \text{ mm}^2$ area: 26×26 array of pins with a 14×14 gap in the center, as well as two fewer pins in an outer row (bottom row in Fig. 5a). According to the manufacturing specifications, the pins have a nominal height of $2.03 \pm 0.08 \text{ mm}$. Our $>110 \text{ cm}^2$ FOV in principle could have supported at least another three PGAs, with some extra area to spare. Figure 5a–d show the AiF photometric composite and 3D height map. The height profile for a single row of pins is plotted in Fig. 5f, whose heights match the nominal specification of 2.03 mm (horizontal lines). In Fig. 5g, kernel density estimates of the height distributions are shown for a row of pins from each PGA (denoted by arrows in Fig. 5b). The two peaks of the bimodal distributions represent the tip of the pins and the base of the chip. Although each PGA has a different height offset, the relative heights of the pins still match the nominal value of 2.03 mm (vertical dotted lines). Note also that some pins are bent away from their normal orientation, which can be seen in the height map.

Ball grid array

We also imaged an assortment of BGA chips (Fig. 6), which consist of arrays of pads with solder balls acting as the interface between integrated circuit chips and PCBs. Some of the chips were forcibly detached, resulting in regions with various heights, depending on whether parts of or the entire solder balls detached. Figure 6 shows the six BGA chips imaged simultaneously, along with several zoom-ins, highlighting the heterogeneity of

the damaged solder balls and pads, both in the photometric color images and the 3D height maps. The 3D height map indicates that some of the chips are tilted. The high resolution of our system also allows capture of sub-pad features that would be missed by conventional wide-FOV inspection devices (Fig. 6g).

Printed circuit board

We imaged a PCB covering the entire $>110 \text{ cm}^2$ FOV of our computational microscope (Fig. 7). The AiF photometric composite reveals fine features at micron-scale resolution of various capacitors, resistors, microchips, and other electronic components across the whole FOV, while the 3D reconstruction reveals their heights. For example, Fig. 7e, f shows an example of a detached resistor as well as evidence of a removed microchip (component “U3”). We note that of all the objects to which we applied our method, this PCB sample posed the most challenges due to the surface texture properties. In Additional file 1: Sec. S3, we discuss the use of a weighted sharpness (discussed in “Self-supervised training”) and confidence maps to address these challenges. Validation experiments comparing STARCAM with other measurement techniques are described in Additional file 1: Sec. S5.

Oil painting

Finally, we applied our method to image a large oil painting, covering the entire $>110 \text{ cm}^2$ FOV (Fig. 8). The AiF photometric and 3D height map reconstructions reveal not only the textures of different brushstrokes (Fig. 8c–h), but also dust particles and fibers that have accumulated above the dried paint (Fig. 8c–d).

Discussion

We have presented a new large-scale computational 3D topographic microscope, STARCAM, that enables high-SBP imaging (6 GP) of 3D samples at micron-scale resolution over a $>110 \text{ cm}^2$ synthetic FOV. We achieve this using a highly parallelized hardware design [27], consisting of 54 cameras along with 3-axis sample scanning to generate a multi-dimensional, multi-TB dataset for each sample, which is then distilled into a 6 GP photometric gigamosaic and a coregistered 3D height map using a CNN-based, self-supervised joint 3D reconstruction and stitching algorithm. The CNN offers a compressed differentiable representation that allows for memory-efficient computational reconstruction across the entire 2.1-TB-per-sample datasets. The self-supervision comes from the stereo overlap redundancy from lateral sample scanning and from image sharpness cues from axial sample scanning. Note that although our method uses sample scanning, we only need to scan laterally across the inter-camera spacing (1.35 cm), and can in principle operate $54\times$ faster than a conventional single-camera system. We applied our method to a variety of samples, including a PCB, integrated circuit components, and a painting, exemplifying the broad applicability of our method.

While we have demonstrated the potential of STARCAM, there are several avenues for future direction. Currently, besides the diffraction limit and aberrations of the imaging optics of our MCAM, the resolution is further limited by the registration accuracy and the axial step size of the z-stacks. The registration accuracy in turn is currently bottlenecked not by the registration algorithm but by the sample scanning repeatability and

accuracy (see Additional file 1: Sec. S2), which can be improved with a more accurate stage and secure sample mounting (or scanning the MCAM instead of the sample). Alternatively or additionally, we could optimize the 6 degree-of-freedom camera poses (and radial distortion parameters) jointly with the 3D height maps, as we did previously when the camera poses could not be precalibrated [32]. Here, since the resolution is much higher than in previous works, we could improve our model by modeling field curvature. The axial step size also affects not only the axial resolution and accuracy, but also the lateral resolution. Currently, we are taking 65 equal steps axially, regardless of the axial scan range, meaning our step sizes in practice were coarser than the empirical results described in “System characterization” and Fig. 4. Furthermore, if the axial sampling is too coarse, it can result in reduced lateral resolution due to the sample missing the DOF of the objectives. Thus, to facilitate high-speed z-stack acquisition with more axial steps, we could use tunable lenses.

On the computation side, our current implementation of our reconstruction algorithm is bottlenecked by the batch preparation step, which consists of loading and debayering the z-stack patches from the SSD and takes about 2.5× longer than the GPU takes to process the batch. As the sustained read speeds of NVMe SSDs improve, we expect this gap to close. The reconstruction times could also be substantially improved with pretraining on multiple types of samples, to ensure that the CNN learns to map z-stacks to height primarily based on physical cues rather than semantic cues. In the case of a well-trained CNN, only the inference step would be necessary. Learning primarily from physical cues could also reduce prediction errors stemming from rare 3D structures whose height value may fall far from the height distribution of the rest of the extended object.

Another possible direction would be to incorporate active patterned illumination [4–7, 21] to improve height estimation, especially for low-contrast samples (Additional file 1: Sec. S3). While it could be a challenge to generate a high-SBP pattern covering the entire >110 cm² FOV, a low-resolution structured pattern covering the whole FOV could still be beneficial in combination with the intrinsic sample features. Alternatively, partially coherent off-axis illumination could be used to emphasize edges and small features for stereo-based height estimation.

In sum, we have developed a high-SBP computational 3D topographic microscope based on a parallelized camera array design and a computational reconstruction and stitching algorithm. Our method can be employed to characterize the 3D topographies of a broad range of extended samples at high resolution, or multiple smaller samples in parallel, with promising applications in accelerating in-line industrial inspection and cultural heritage digitization. Furthermore, our work demonstrates the feasibility of computational imaging with joint, end-to-end optimization across large, multi-TB-per-sample datasets, opening the door to upscaling of other types of computational imaging problems beyond high-SBP topographic microscopy.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s40537-024-00901-0>.

Additional file 1: Figure S1. Lateral resolution (a) and DOF (b) characterization of our imaging system. In (a), the left USAF target was placed at the center of the FOV, while the right USAF target was placed at the edge of the FOV. **Figure S2.** PCB analysis. (a) All-in-focus photometric gigamosaic. (b) Confidence map, based on max sharpness across the z-stack dimension. (c) 3D height map reconstruction. (d) 3D height map reconstructed using only argmax

across the z-stack dimension. **Figure S3.** Architecture of the CNN mapping from z-stacks to height maps. **Figure S4.** Comparison of STARCAM with laser confocal microscopy (Keyence), focal stacking (Keyence), and calipers at six locations (a–f) across the PCB sample (Fig. 7). These locations are indicated in Additional file 1: Fig. S5. The first column shows the photometric (RGB) image. The middle three columns show the height map comparisons, with the same color range in each row. The fifth column shows the 1D height profiles indicated by the dotted lines in the middle three columns. The 1D height profiles are also accompanied by the caliper estimates (based on the mean of 10 independent measurements). **Figure S5.** The locations of the crops analyzed in Additional file 1: Fig. S4. **Figure S6.** Overlap maps, showing how many times (0–9) each point in the FOV. (a) The overlap map for all 3456 camera views, based on the joint camera calibration from imaging a flat reference target. (b) The overlap map for one 8×8 scan from a single camera. (c) Each view contains over 9 MP.

Acknowledgements

We would like to thank Dr. Aurélien Bègue for providing helpful feedback, Prof. Navid Asadi for providing the PCB, and Margaret Aery for painting and providing the oil painting of a chair.

Author contributions

Conceptualization: KCZ, MH, RH; Methodology: KCZ, MH, RH; Software: KCZ, MH, MZ, PR, TD, VS; Formal Analysis: KCZ; Investigation: KCZ, MH, MZ, GH; Data Curation: KCZ, MH; Writing: KCZ, JJ, KCL, RA, RH; Visualization: KCZ, JJ; Supervision: RH; Funding Acquisition: MH, RH.

Funding

Research reported in this publication was supported by the Office of Research Infrastructure Programs (ORIP), Office Of The Director, National Institutes Of Health of the National Institutes Of Health and the National Institute Of Environmental Health Sciences (NIEHS) of the National Institutes of Health under Award Number R44OD024879, the National Cancer Institute (NCI) of the National Institutes of Health under Award Number R44CA250877, the National Institute of Biomedical Imaging and Bioengineering (NIBIB) of the National Institutes of Health under Award Number R43EB030979, the National Science Foundation under Award Number 2036439, and the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. 2021R1C1C101290013).

Availability of data and materials

Data underlying the results presented in this paper are not publicly available at this time, due to the exceedingly large size (~ 10 TB), but may be obtained from the authors upon reasonable request. The Python code used to generate the gigamosaics and 3D height maps is available at <https://github.com/kevinczhou/starcam>.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

MH and RH are cofounders of Ramona Optics Inc., which is commercializing the MCAM. MH, MZ, RA, PR, TD, VS, and GH are or were employed by Ramona Optics Inc. when conducting this research. KCZ is a consultant for Ramona Optics Inc. RH is a founder of MIRA Inc., which is applying MCAM technology to art. RA is employed by MIRA Inc. The remaining authors declare no Conflict of interest.

Received: 4 June 2023 Accepted: 14 March 2024

Published online: 02 May 2024

References

- Park J, Brady DJ, Zheng G, Tian L, Gao L. Review of bio-optical imaging systems with a high space-bandwidth product. *Adv Photon*. 2021;3(4): 044001.
- Zheng G, Ou X, Horstmeyer R, Chung J, Yang C. Fourier ptychographic microscopy: a gigapixel superscope for biomedicine. *Optics Photon News*. 2014;25(4):26–33.
- Luhmann T. Close range photogrammetry for industrial applications. *ISPRS J Photogram Remote Sens*. 2010;65(6):558–69.
- Scharstein D, Szeliski R. High-accuracy stereo depth maps using structured light. In: 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings., 2003; vol. 1, p. IEEE.
- Yen H-N, Tsai D-M, Yang J-Y. Full-field 3-D measurement of solder pastes using LCD-based phase shifting techniques. *IEEE Trans Electron Packaging Manuf*. 2006;29(1):50–7.
- Geng J. Structured-light 3D surface imaging: a tutorial. *Adv Opt Photon*. 2011;3(2):128–60.
- Xu X, Fei Z, Yang J, Tan Z, Luo M. Line structured light calibration method and centerline extraction: a review. *Results Phys*. 2020;19: 103637.
- Spagnolo GS, Guattari G, Sapia C, Ambrosini D, Paoletti D, Accardo G. Three-dimensional optical profilometry for artwork inspection. *J Opt A Pure Appl Opt*. 2000;2(5):353.
- Pieraccini M, Guidi G, Atzeni C. 3d digitizing of cultural heritage. *J Cult Herit*. 2001;2(1):63–70.

10. Traxler L, Ginner L, Breuss S, Blaschitz B. Experimental comparison of optical inline 3d measurement and inspection systems. *IEEE Access*. 2021;9:53952–63.
11. Wu M-J, Jang J-SR, Chen J-L. Wafer map failure pattern recognition and similarity ranking for large-scale data sets. *IEEE Trans Semicond Manuf*. 2014;28(1):1–12.
12. Acciani G, Brunetti G, Fornarelli G. Application of neural networks in optical inspection and classification of solder joints in surface mount technology. *IEEE Trans Industr Informat*. 2006;2(3):200–9.
13. Hong D, Lee H, Kim MY, Cho H, Moon JI. Sensor fusion of phase measuring profilometry and stereo vision for three-dimensional inspection of electronic components assembled on printed circuit boards. *Appl Opt*. 2009;48(21):4158–69.
14. Xue K, Wu J, Chen H, Gai J, Lam A. Warpage prediction of fine pitch BGA by finite element analysis and shadow moiré technique. In: 2009 International Conference on Electronic Packaging Technology & High Density Packaging, 2009;pp. 317–321. IEEE.
15. Li D, Liu C, Tian J. Telecentric 3d profilometry based on phase-shifting fringe projection. *Opt Exp*. 2014;22(26):31826–35.
16. Marrugo AG, Gao F, Zhang S. State-of-the-art active optical techniques for three-dimensional surface metrology: a review. *JOSA A*. 2020;37(9):60–77.
17. Thomas M, Su R, Nikolaev N, Coupland J, Leach R. Modeling of interference microscopy beyond the linear regime. *Opt Eng*. 2020;59(3):034110–034110.
18. Hibino K, Oreb BF, Fairman PS, Burke J. Simultaneous measurement of surface shape and variation in optical thickness of a transparent parallel plate in wavelength-scanning fizeau interferometer. *Appl Opt*. 2004;43(6):1241–9.
19. Chen L-C, Yeh S-L, Tapilouw AM, Chang J-C. 3-d surface profilometry using simultaneous phase-shifting interferometry. *Opt Commun*. 2010;283(18):3376–82.
20. Yin Y, Wang M, Gao BZ, Liu X, Peng X. Fringe projection 3d microscopy with the general imaging model. *Opt Exp*. 2015;23(5):6846–57.
21. Hu Y, Chen Q, Feng S, Zuo C. Microscopic fringe projection profilometry: a review. *Opt Lasers Eng*. 2020;135: 106192.
22. Kagami J, Hatazawa T, Koike K. Measurement of surface profiles by the focusing method. *Wear*. 1989;134(2):221–9.
23. Jordan H-J, Wegner M, Tiziani H. Highly accurate non-contact characterization of engineering surfaces using confocal microscopy. *Measure Sci Technol*. 1998;9(7):1142.
24. Czajkowski J, Prykäri T, Alarousu E, Palosaari J, Myllylä R. Optical coherence tomography as a method of quality inspection for printed electronics products. *Opt Rev*. 2010;17:257–62.
25. Wang H, Zhu J, Sung J, Hu G, Greene J, Li Y, Park S, Kim W, Lee M, Yang Y, et al. Fourier ptychographic topography. *Opt Exp*. 2023;31(7):11007–18.
26. Thomson E, Harfouche M, Kim K, Konda P, Seitz CW, Cooke C, Xu S, Jacobs WS, Blazing R, Chen Y, et al. Gigapixel imaging with a novel multi-camera array microscope. *Elife*. 2022;11:74988.
27. Harfouche M, Kim K, Zhou KC, Konda PC, Sharma S, Thomson EE, Cooke C, Xu S, Kreiss L, Chaware A, et al. Imaging across multiple spatial scales with the multi-camera array microscope. *Optica*. 2023;10(4):471–80.
28. Zhou KC, Harfouche M, Cooke CL, Park J, Konda PC, Kreiss L, Kim K, Jönsson J, Doman T, Reamey P, et al. Parallelized computational 3d video microscopy of freely moving organisms at multiple gigapixels per second. *Nat Photon*. 2023;1–9.
29. Yang X, Harfouche M, Zhou KC, Kreiss L, Xu S, Kim K, Horstmeyer R. Multi-modal imaging using a cascaded microscope design. *arXiv preprint arXiv:2208.08875* 2022.
30. Furukawa Y, Hernández C, et al. Multi-view stereo: a tutorial. *Found Trends Comput Graph Vis*. 2015;9(1–2):1–148.
31. Krotkov E. Focusing. *Int J Comput Vis*. 1988;1(3):223–37.
32. Zhou KC, Cooke C, Park J, Qian R, Horstmeyer R, Izatt JA, Farsiu S. Mesoscopic photogrammetry with an unstabilized phone camera. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021. pp. 7535–7545.
33. Zhou KC, McNabb RP, Qian R, Degan S, Dhalla A-H, Farsiu S, Izatt JA. Computational 3d microscopy with optical coherence refraction tomography. *Optica*. 2022;9(6):593–601.
34. Ulyanov D, Vedaldi A, Lempitsky V. Deep image prior. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018. pp. 9446–9454.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.