Micron Scale 3D Imaging with Multi-Camera Array
Supplementary Information

S1. Table of Optical and Hardware Parameters

Parameter	Value
Camera Spacing	13.5 mm
Pixel Size	1.1 µm
Sensor Width	4.63 mm
Effective Focal Length	26.23 mm
Image Space F/#	2.6
Working F/#	3.013
Image Space NA	0.164
Object Space NA	0.0357
Entrance Pupil Diameter	9.636 mm
Primary Wavelength	0.5875618 µm

Table S1: Summary of optical and hardware Parameters

S2. Impact of Hardware Design on Height Prediction

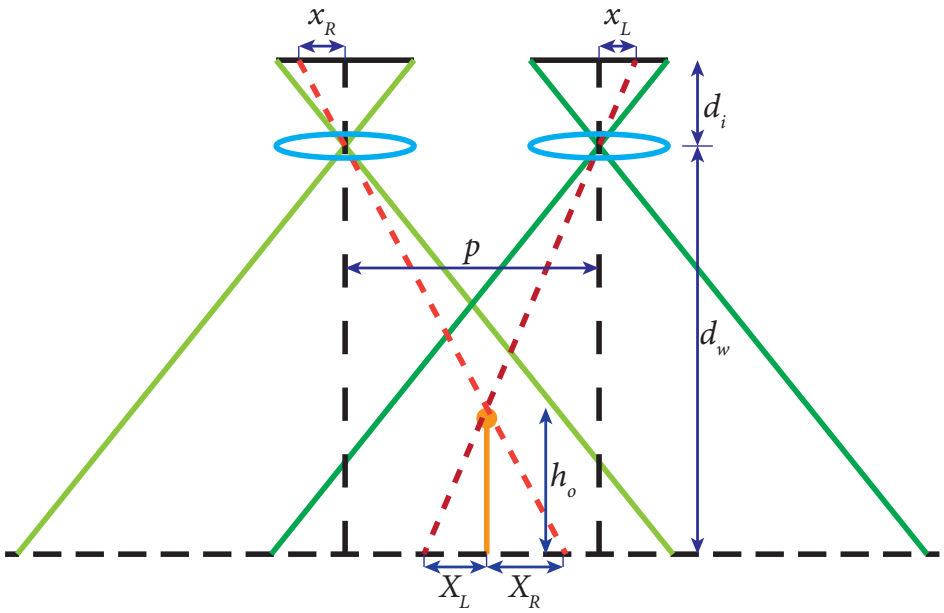


Figure S1: Simplified view of the array showing two cameras looking at an object with height h_o. Both cameras are assumed to be identical.

List of variables:

Focal Length	f	Magnification	M
Working Distance	d_w	Numerical Aperture	NA
Image Distance	d_i	Lateral Resolution	r
Object Height	h_o	Pixel Size	μ
Camera Spacing	p	Sensor Width	s

Let us consider a simplified view of the system with two cameras viewing an object with height h_o as shown in Figure S1. We assume both cameras are identical and ignore the effect of camera distortion, lens aberration, and

misalignment. We also assume that the object lies within the depth-of-field of the system, and ignore occlusions. The center-to-center distance between the cameras is p. From the lens equation,

$$\frac{1}{d_i} + \frac{1}{d_w} = \frac{1}{f} \tag{S1}$$

and the magnification of the system is given as:

$$M = \frac{d_i}{d_w} \tag{S2}$$

Due to parallax, the object's apparent position is shifted X_R and X_L in the right and left cameras, respectively. The accuracy of height estimation via stereo depends on how well the total parallax $\Delta X = X_R + X_L$ can be measured.

Using similar triangles,

$$\frac{h_o}{\Delta X} = \frac{d_w - h_o}{p}$$

Using Eqns S1 and S2 we get $d_w = f(1 + \frac{1}{M})$. Substituting and rearranging,

$$h_o = \frac{\Delta X f(M+1)}{pM + \Delta X M} \tag{S3}$$

To obtain the uncertainty in the object height δh, we can set the parallax error ΔX to the object-side lateral resolution of the system r.

The object-side lateral resolution can be either diffraction-limited or pixel-limited and is respectively given as:

$$r_{pixel} = \frac{2\mu}{M} \tag{S4}$$

$$r_{diff} = \frac{\lambda}{M \cdot NA} \approx \frac{\lambda \cdot 2d_i}{M \cdot D} = \frac{2\lambda f(M+1)}{M \cdot D} \tag{S5}$$

where λ is the wavelength and D is the diameter of the lens aperture.

Substituting these in Eqn S3 we get,

$$\delta h_{pixel} = \frac{2\mu f(M+1)}{M(pM + 2\mu)} \tag{S6}$$

and,

$$\delta h_{diff} = \frac{2\lambda f^2 (M+1)^2}{M(pDM + 2\lambda f(M+1))} \tag{S7}$$

So we can see that having a lower f, higher M, and p will improve the height accuracy in both cases. Since at least 50% overlap in the FOVs of adjacent cameras is needed to ensure that at least two cameras are looking at each point in the scene, increasing the sensor pitch p would mean decreasing the magnification by the same factor. From above we can see that while increasing the pitch can improve the height accuracy, the corresponding decrease in magnification is more effective, resulting in decreased height accuracy.

One of our assumptions was that the object lies within the system's depth-of-field (DOF). However our choice of f, M, and p also affects the DOF. DOF is usually given as:

$$DOF = \frac{2d_w^2 Nc}{f^2} \tag{S8}$$

where N is the f-number of the lens, and c is the acceptable circle of confusion. We can replace $N = f/D$ and set $c = 2\mu$ so that the circle of confusion is limited to two pixels.

$$DOF = \frac{2d_w^2 Nc}{f^2} = \frac{4f\mu}{D}(1 + \frac{1}{M})^2 \tag{S9}$$

So we can see that decreasing the focal length and increasing the magnification (which improves the height accuracy) reduces the system DOF. A shallower depth of field will necessitate a denser focal stack.

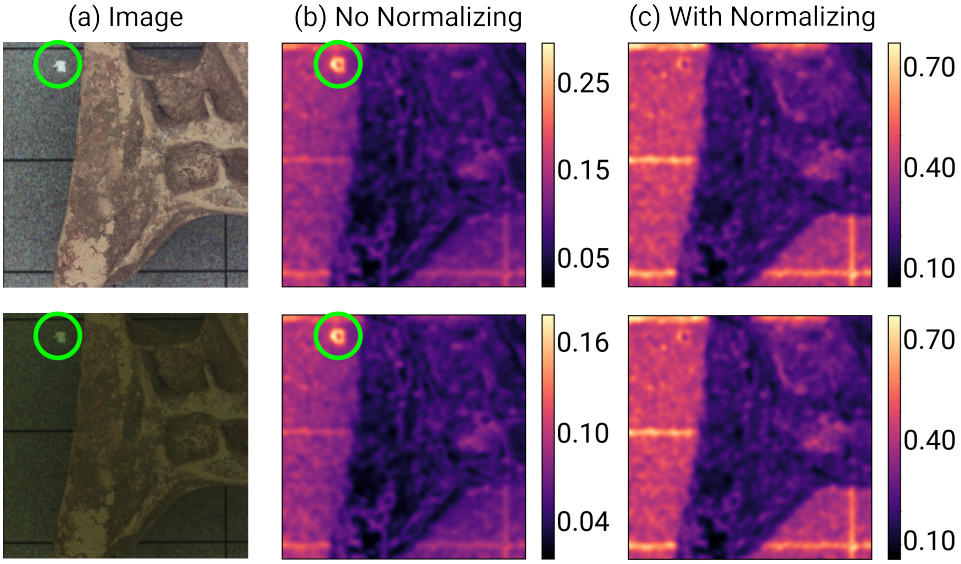


Figure S2: **Robust sharpness metric.** As the illumination changes (a), sharpness calculated without normalization (b) also changes. Stray reflections (marked in green) also affect the metric calculation. Our method (c) is robust against both.

S3. Robust Sharpness Metric

Since we wish our sharpness metric to be robust to illumination and magnification changes, we first create a normalized version of our image by dividing it by its own Gaussian blurred version. The output of this operation is akin to a normalized high-pass filtered image. We then take the magnitude of the gradients of this image as our sharpness metric. Let us say that $s(x, y, z)$ denotes a slice from our z-stack. Then,

$$s_{norm}(x, y, z) = \frac{s(x, y, z)}{s(x, y, z)G(x, y)} \tag{S10}$$

where $G(x, y)$ denotes a 2D gaussian kernel. And we obtain the sharpness metric as:

$$P(x, y, z) = \left|\nabla_{x,y}\ s_{norm}(x, y, z)\right| G(x, y) \tag{S11}$$

Gaussian filtering helps obtain a smooth sharpness metric.

This process is demonstrated in Figure S2. Normalizing the images using our method yields a consistent metric even as the illumination changes, as well as provides robustness against stray reflections.

Finally, we can also obtain a depth-from-focus estimate using this metric as:

$$d_{argmax}(x, y) = \arg\max_{z} P(x, y, z) \tag{S12}$$

S4. Geometric and Photometric Calibration

We calibrate the Geometric and Photometric Properties of the camera array before capturing data. 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, properties include the variation in the pixel intensity within the individual cameras due to vignetting and variations in pixel response, as well as inter-camera variations. These arise due to multiple factors, including uneven illumination, differences in pixel response, and stray reflections.

We capture a focal stack of a patterned, flat target to perform the calibration. For each camera, we select the sharpest plane from the stack by selecting the per-pixel location z that maximizes the mean sharpness given as $P_m = \text{mean}_{x,y} P(x, y, z = z)$. Sharpness is calculated using the metric described above.

Starting with initial estimates of the geometric and photometric properties, we dewarp and backproject the 54 target images onto a shared object plane. These images are then reprojected into camera space. Using gradient descent, we iteratively minimize pixel-wise photometric error, refining both the geometric and photometric parameters in the process. Specifically, let the geometric properties for the ith camera be parameterized by θ_i, and let x_0 and y_0 be two

vectors representing the spatial coordinates of the camera pixels. We will use an image deformation operation, $D_\theta\{\cdot,\cdot\}$, to map these coordinates into a common coordinate space, specifically the object plane. i.e.

$$x_i, y_i = D_{\theta_i}\{x_0, y_0\} \tag{S13}$$

where x_i, y_i are the dewarped coordinates of the ith camera on the object plane. Specific implementation of D_{θ_i} follows [1]. Next, let the photometric properties for the i th camera be parametrized by ϕ_i, and let $I_{i,0}$ be a vector of the same length as x_0 and y_0, which gives the measured intensity at every pixel coordinate in the i th camera. We can then use a photometric correction operator $C_{\phi,x_0,y_0}\{\cdot\}$ to photometrically adjust the intensity values for the i th camera, so that

$$I_i = C_{\phi,x_0,y_0}\{I_{i,0}\} \tag{S14}$$

where I_i represents the adjusted intensity values. Note that the operation depends on x_0 and y_0 due to the spatially varying nature of this photometric correction. Once the deformation and photometric adjustments have been completed for all 54 cameras, we can initialize a blank matrix $R[\cdot,\cdot]$ to hold the stitched reconstruction. We will then use the corrected coordinate vectors, x_i and y_i, to back-project the corrected intensity values I_i for each camera into R. When specific coordinates in R are visited more than once, the values are averaged. The end result of this step is a single-stitched prediction of the full calibration target. Finally, we will again use the estimated geometric and photometric parameters to reproject R back into camera space, providing estimated images I_{pred}. We minimize an error metric given by

$$\mathcal{L} = \|I_{pred} - I\|^2 + \sum_i stdev(I_i) \tag{S15}$$

with respect to $\{\theta_i, \phi_i\}_{i=1}^{54}$ via gradient descent. These steps are then repeated until the loss plateaus (reduces by less than 0.001 for 5 consecutive iterations).

S5. Network Architecture and Training Details

Downsample Block	**Upsample Block**
	2x Bilinear Upsampling
3x3 Conv layer, k filters, stride=2	3x3 Conv layer, k filters, stride=1
Batch Norm	Batch Norm
Leaky ReLU	Leaky ReLU
3x3 Conv layer, k filters, stride=1	1x1 Conv layer, k filters, stride=1
Batch Norm	Batch Norm
Leaky ReLU	Leaky ReLU

Table S2: Network Architecture

Our CNN consists of Downsample and Upsample blocks as summarized in Table S2. In all our experiments, we used 4 sequential downsample blocks with $k = 16, 16, 32, 32$ filters, followed by 4 sequential upsample blocks with $k = 32, 32, 16, 16$ filters.

Hyperparameter values used for various objects were:

Object	α	β	γ	λ
3D-printed Pyramid	10	5000	5	8
Packaging Foam Piece	10	5000	5	8
Terracotta Rooster	10	5000	1	5
Carved Protome	20	5000	10	5
Clay Cup	10	5000	10	5
Whale Tooth Carving	0.5	5000	1	10

S6. Ablation Study

To evaluate the contribution of individual loss components to the overall performance of our reconstruction algorithm, we conducted an ablation study using the 3D printed pyramid object. Each of the four loss terms i.e. stereo ($\mathcal{L}_{stereo}$),

sharpness ($\mathcal{L}_{sharp}$), focus ($\mathcal{L}_{focus}$), and photometric ($\mathcal{L}_{photo}$), were systematically removed during the training phase by setting its corresponding hyperparameter (α, β, γ, or λ) to zero. All other hyperparameters were kept consistent with those used for the 3D printed pyramid results reported previously (Supplementary Material S5). Each ablated configuration was trained on the same hardware for 160,000 iterations, using a learning rate of 10^{-4}. The resulting Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) in depth estimation are summarized in Table S6.

Configuration	**MAE (mm)**	**RMSE (mm)**
Full Model (All losses)	**0.536**	**0.669**
No stereo loss ($\alpha = 0$)	0.658	0.848
No sharpness loss ($\beta = 0$)	0.828	1.063
No focus loss ($\gamma = 0$)	0.719	0.956
No photometric loss ($\lambda = 0$)	0.718	0.843

Table S3: Quantitative results of the ablation study on the 3D printed pyramid. MAE and RMSE are reported in mm.

These results clearly demonstrate that the full model, incorporating all four loss terms, achieves the lowest error. The removal of any single loss component resulted in a degradation of performance. Notably, excluding the sharpness loss ($\beta = 0$) led to the most significant increase in error, with MAE rising to 0.828 mm and RMSE to 1.063 mm. This underscores the critical role of the sharpness cue in refining the depth predictions. The absence of the stereo loss (where $\alpha = 0$) also substantially impacted accuracy, yielding an MAE of 0.658 mm and RMSE of 0.848 mm. Removing the focus loss ($\gamma = 0$) or the photometric loss ($\lambda = 0$) also resulted in increased errors compared to the full model, highlighting their respective contributions to the final reconstruction quality.

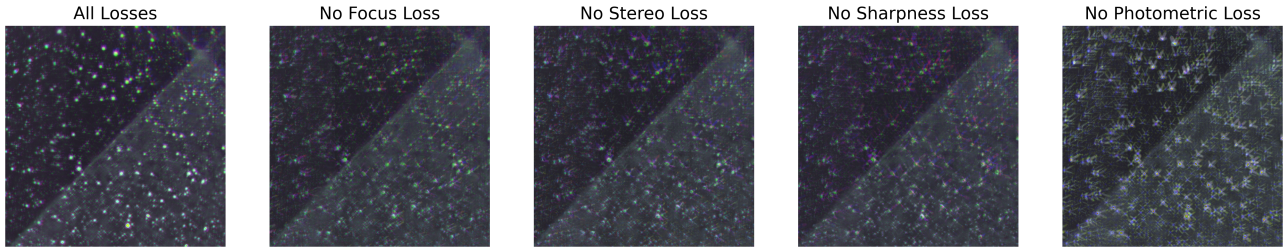


Figure S3: Section of the RGB stitched image obtained for different ablations

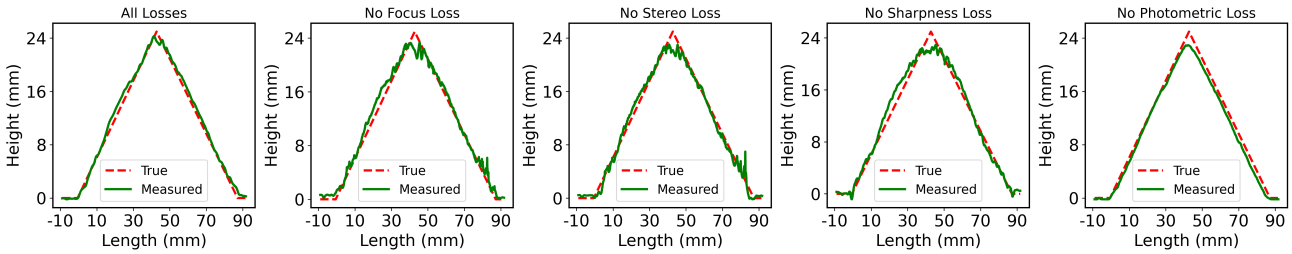


Figure S4: Surface Profile Plots obtained for different ablations

Qualitatively, these findings are supported by the visual results. Figure S3 presents sections of the stitched RGB images, where the reconstruction from the full model exhibits enhanced clarity and fewer artifacts compared to the ablated configurations. Furthermore, Figure S4 displays the corresponding surface profile plots, visually confirming that the depth map generated by the full model aligns more closely with the object's true geometry.

S7. Resolution via Fourier Ring Correlation

The FRC values obtained for different objects are given below.

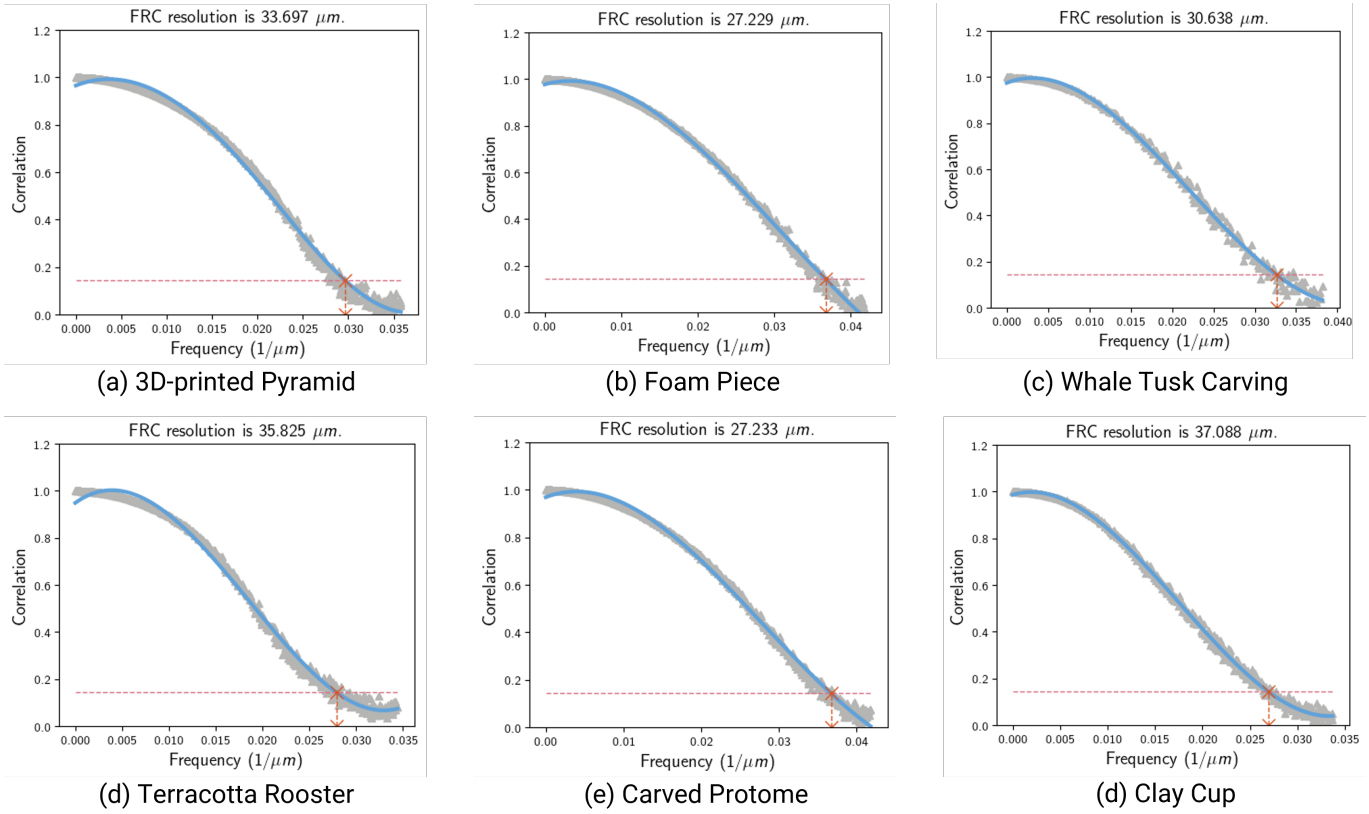
Figure S5 shows the FRC plots for all these objects.

Object Name	FRC Resolution
3D-printed Pyramid	34 μm
Foam Piece	27 μm
Terracotta Rooster	36 μm
Carved Protome	27 μm
Clay Cup	37 μm
Whale Tooth Carving	30 μm

Table S4: FRC values obtained for various objects

S8. Supplementary Results

Please look at Figure S6 and Figure S7 for additional reconstructions of the 3D-printed Pyramid and Packaging Foam Piece. Figure S8 shows all-in-focus and depth reconstruction of a small painting.



(a) 3D-printed Pyramid (b) Foam Piece (c) Whale Tusk Carving

(d) Terracotta Rooster (e) Carved Protome (d) Clay Cup

Figure S5: Fourier Ring Correlation plots for various objects.

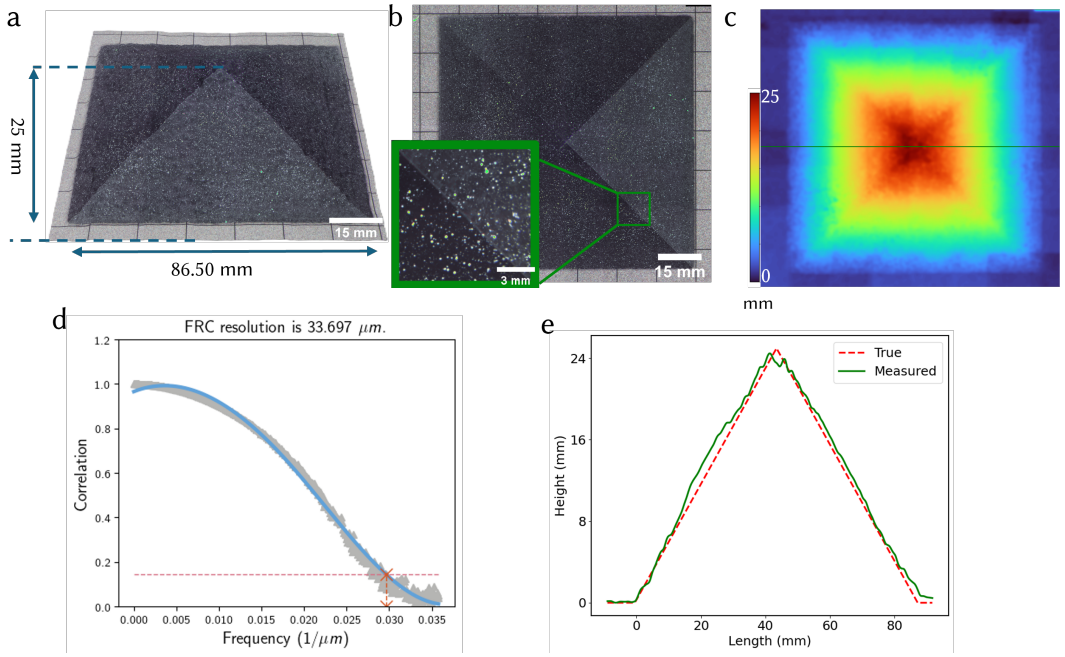


Figure S6: (a) 3D printed pyramid, (b) All-in-focus composite image and zoom-in, (c) Predicted depth map, (d) Resolution Characterization with Fourier Ring Correlation, (e) Line profile

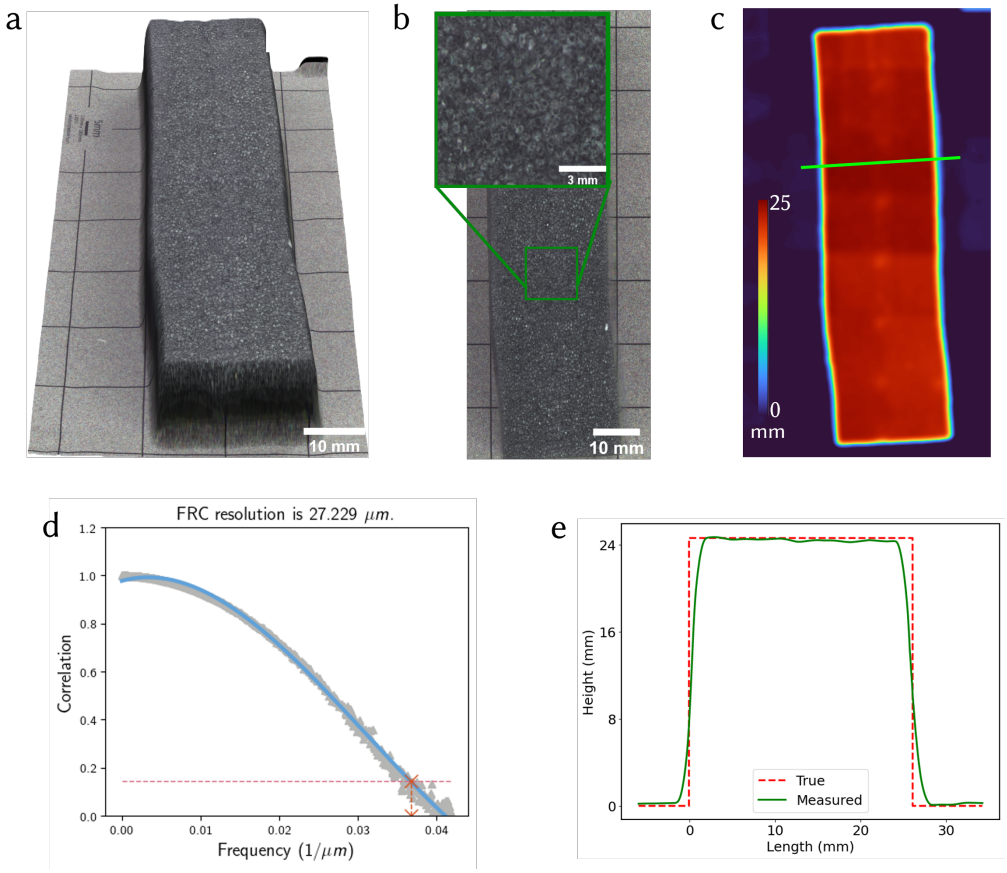


Figure S7: (a) Packaging Foam Piece, (b) All-in-focus composite image and zoom-in, (c) Predicted depth map, (d) Resolution Characterization with Fourier Ring Correlation, (e) Line profile

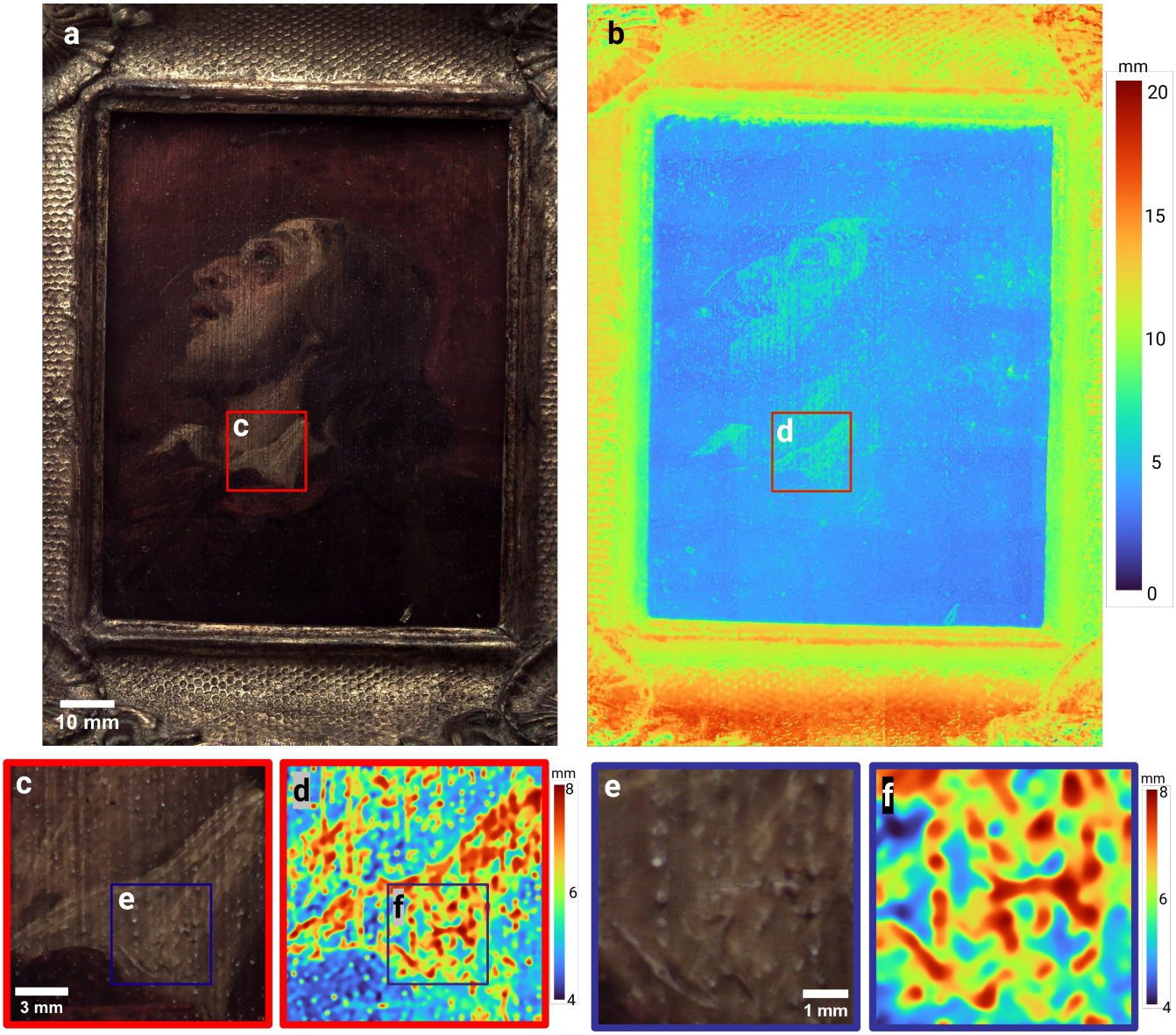


Figure S8: (a) shows the all-in-focus composite image and (b) shows the depth map. Insets (c) and (e) are zoom-ins of the all-in-focus composites, and insets (d) and (f) are zoom-ins of the depth maps. The thick brush strokes are clearly visible in both the photometric and depth composites.