Collaborative On-Sensor Array Cameras: Supplementary Information

JIPENG SUN, Princeton University, USA
KAIXUAN WEI, KAUST, Saudi Arabia
THOMAS EBOLI, Université Paris-Saclay, France
CONGLI WANG and CHENG ZHENG, Princeton University, USA
ZHIHAO ZHOU and ARKA MAJUMDAR, University of Washington, USA
WOLFGANG HEIDRICH, KAUST, Saudi Arabia
FELIX HEIDE, Princeton University, USA

CCS Concepts: • **Computing methodologies → Computational photography, Massively parallel algorithms**.

Additional Key Words and Phrases: Computational Optics

ACM Reference Format:
Jipeng Sun, Kaixuan Wei, Thomas Eboli, Congli Wang, Cheng Zheng, Zhihao Zhou, Arka Majumdar, Wolfgang Heidrich, and Felix Heide. 2025. Collaborative On-Sensor Array Cameras: Supplementary Information. *ACM Trans. Graph.* 44, 4 (August 2025), 20 pages. https://doi.org/10.1145/3731200

Contents

Authors' addresses: Jipeng Sun, jipeng.sun@princeton.edu, Princeton University, 35 Olden St, Princeton, New Jersey, USA, 08540; Kaixuan Wei, kaixuan.wei@kaust.edu.sa, KAUST, Saudi Arabia; Thomas Eboli, thomas.eboli12@gmail.com, Université Paris-Saclay, Paris, France; Congli Wang, congli.wang@princeton.edu; Cheng Zheng, chengzh@princeton.edu, Princeton University, USA; Zhihao Zhou, zzhou99@uw.edu; Arka Majumdar, arka@uw.edu, University of Washington, 1410 NE Campus Pkwy, Seattle, Washington, USA, 98195; Wolfgang Heidrich, wolfgang.heidrich@kaust.edu.sa, KAUST, Thuwal, Saudi Arabia; Felix Heide, fheide@princeton.edu, Princeton University, 35 Olden St, Princeton, New Jersey, USA, 08540.

0730-0301/2025/8-ART
https://doi.org/10.1145/3731200

A ADDITIONAL DETAILS ON IMAGE FORMATION MODEL AND RECONSTRUCTION

In this section, we provide additional details on the proposed nanophotonic image formation model and the collaborative image deconvolution method for the on-sensor array camera.

A.1 Image Formation Model (Section 3.1 in main)

A camera will look at a scene U in the visible spectrum. Our camera has six apertures, with relative 3D-2D projection matrices $P^1, \dots, P^6$. The focal images are $u^1, \dots, u^6$ with $u^k = P^k U$. It captures the fact that there is parallax between images. Each image is defined locally on the visual spectrum Λ: $u^k : \lambda \in \Lambda \mapsto u^k(\lambda)$. We denote in what follows $u^k_\lambda = u^k(\lambda)$.

Each aperture comes with a metalens, and thus a point-spread function (PSF) p^k also defined on the visual spectrum Λ. The PSF of a metalens is typically greatly varying on Λ. After passing through each of the K apertures, we actually have blurry variants of the u^k. This is commonly modeled with a convolutional forward process per wavelength λ in Λ and for each aperture k:

$$v^k_\lambda = p^k_\lambda * u^k_\lambda, \tag{S1}$$

where $*$ is the convolution operator.

The process of converting a hyperspectral image u_λ to an RGB image is given by operator I:

$$u^c = I(u_\lambda) = \int_{\lambda \in \Lambda} u_\lambda \kappa^c_\lambda d\lambda. \tag{S2}$$

In this model, κ^c_λ is the camera spectral response function (CSRF) for color channel c and wavelength λ. These coefficients mix the different bands of the input hyperspectral signal into three distinct discrete channels R, G and B.

The recorded RGB images $v^1, \dots, v^6$ (what we alternatively call *measurements* in the main paper) are obtained with:

$$v^{k,c} = I\left(v^k_\lambda\right) + \varepsilon^{k,c} = I\left(p^k_\lambda * u^k_\lambda\right) + \varepsilon^{k,c}, \tag{S3}$$

where $\varepsilon^{k,c}$ is the additive noise. Additional mosaicking further leads to a final recorded image. We assume that an in-body demosaicking algorithm is applied under the hood. We thus work with the measurements $v^{k,c}$ in what follows and neglect demosaicking artifacts.

In practice, one barely has access to hyperspectral measurements, but instead RGB ones. As a result, a common approximation of the formation model above is for all channels, c

$$v^{k,c} = p^{k,c} * u^{k,c} + \varepsilon^{k,c} = I\left(p^k_\lambda\right) * I\left(u^k_\lambda\right) + \varepsilon^{k,c}, \tag{S4}$$

which swaps the operators I and $*$, or in other words, swapping integration on the sensor, and passing through an optical system.

A.2 Reconstruction (Section 3.2 and 3.3 in main)

Derivaton of Collaborative Joint Wiener Deconvolution. Assume for the moment we have access to the v^k_λ and p^k_λ values. Assume further that we have compensated for the parallax in the measurements v^k_λ. As a result, we have $u^1 = \dots = u^6 = u$, such that the formation model becomes:

$$v^k_\lambda = p^k_\lambda * u_\lambda. \tag{S5}$$

The problem has now become how can we combine the K measurements and the PSFs to obtain an estimate of u_λ. Since the formation model is independent of λ, we can also posit the solution should not blend several

wavelengths but instead be also independent with respect to λ. We approach the recovery of u_λ as a MAP estimate, solving the following inverse problem

$$\min_{u_\lambda} \frac{1}{K} \sum_{k=1}^{K} \|v_\lambda^k - p_\lambda^k * u_\lambda\|_2^2 + \beta\Omega(u_\lambda), \tag{S6}$$

where Ω is an image prior that compensates for the ill-posedness of the data-fidelity term. A simple choice for Ω is using the squared norm, which favors solution with smaller energy, *i.e,* no large peaks in the signal

$$\min_{u_\lambda} \frac{1}{K} \sum_{k=1}^{K} \|v_\lambda^k - p_\lambda^k * u_\lambda\|_2^2 + \beta\|u_\lambda\|_2^2, \tag{S7}$$

Via the Parceval theorem, this least-squares problem is equivalent in the Fourier domain to

$$\min_{U_\lambda} \sum_{k=1}^{K} \|V_\lambda^k - P_\lambda^k \odot U_\lambda\|_2^2 + \beta\|U_\lambda\|_2^2, \tag{S8}$$

where the capital letters are the Fourier transforms of the spatial quantities and $\odot$ is the pointwise product. Each entry in the Fourier transform arrays corresponds to a distinct spatial frequency. Taking the gradient of this energy and setting it to 0 yields at the solution $\widehat{U}_\lambda$

$$0 = \frac{1}{K} \sum_{k=1}^{K} \overline{P_\lambda^k} \odot V_\lambda^k - \frac{1}{K} \sum_{k=1}^{K} \overline{P_\lambda^k} \odot P_\lambda^k \odot \widehat{U}_\lambda + \beta J \odot \widehat{U}_\lambda, \tag{S9}$$

where J is an array of the same size as $\widehat{U}_\lambda$ full of ones (with the convention that $J \odot \widehat{U}_\lambda = \widehat{U}_\lambda$). Remarking that $\overline{P_\lambda^k} \odot P_\lambda^k = |P_\lambda^k|^2$ and isolating $\widehat{U}_\lambda$ yields:

$$\widehat{U}_\lambda = \frac{\sum_{k=1}^{K} \overline{P_\lambda^k} \odot V_\lambda^k}{\sum_{k=1}^{K} |P_\lambda^k|^2 + K\beta J}. \tag{S10}$$

The fraction bar is elementwise. In the main paper, we dropped J with slight abuse of notation but the additive term $K\beta J$ means we add $K\beta$ to each entry of $|P_\lambda^k|^2$. The solution is thus the weighted average of the measurements in the Fourier domain, with respect to the sharpness of the filters at each spatial frequency.

Analyis of Derived Joint Wiener Filter. Let us dissect the joint Wiener filter. First, let us understand the difference between the case $K = 1$, *i.e.,* the original Wiener filter, and the joint Wiener filter for $K > 1$. When $K = 1$, the solution reads:

$$\widehat{U}_\lambda^{K=1} = \frac{\overline{P_\lambda^1} \odot V_\lambda^1}{|P_\lambda^1|^2 + \beta J} \tag{S11}$$

In the original Wiener formulation, the filter multiplies by the inverse of the Fourier coefficients of the PSF, plus β to not divide by 0 or multiply of $+\infty$, the Fourier coefficients of the blurry image. Since the PSF is a low-pass filter the Fourier coefficients of the higher frequencies are close to 0 so β is crucial to not create large peaks in the solution, and thus ensuring the validity of the inverse filter in the Fourier domain. However setting a value of β that is large enough to prevent such peaks in the solution biases the coefficients that are not 0's, and leads to blurry solutions. There is thus a trade-off between retain too much blur in the reconstruction, and a sharp image but at the cost of several overshoots in the reconstruction.

In the joint Wiener filter instead, the denominator is instead $\sum_{k=1}^{K} |P_\lambda^k|^2 + K\beta J$. In the case where $K = 2$, we can see that at a given frequency ω if the first PSF has a coefficient that is 0, we could still have a valid denominator without setting a large value of β if the second PSF's coefficient at the same frequency is not 0. As a result, if

one selects the kernels P_λ^1 and P_λ^2 such that they are *complementary* over the whole Fourier domain, *i,e,* for all spatial frequency ω, we have $\min(|P_\lambda^1(\omega)|, |P_\lambda^2(\omega)|) > 0$, one could get a sharp result without the overshoots of the single-frame case and without needing to tune β to prevent blurring of the reconstructed image. When now we have more than two PSFs, the probability that at least one PSF has not a 0 at frequency ω is larger, thus the joint Wiener filter is more likely be invertible everywhere. Since the K PSFs are all low-pass filters, we still need β for the higher frequencies.

Converting HS Back to RGB Space. Although a hyperspectral image formation model is required to capture the wavelength-dependent phase response of meta-optical elements, our metalens array is designed for general RGB sensors. Consequently, the final reconstruction of interest also resides in the RGB domain.

Concretely, in the hyperspectral domain, converting a reconstructed volume $\widehat{U}_\lambda$ into an RGB image $\widehat{U}^c$ is straightforward, involving an integration over the camera's spectral response:

$$\widehat{U}^c = I(\widehat{U}_\lambda) = \sum_{\lambda \in \Lambda_{\text{disc}}} \kappa_\lambda^c \widehat{U}_\lambda = \sum_{\lambda \in \Lambda_{\text{disc}}} \kappa_\lambda^c \left(\frac{\sum_{k=1}^K \overline{P_\lambda^k}}{\sum_{k=1}^K |P_\lambda^k|^2 + K\beta} \odot V_\lambda^k \right). \tag{S12}$$

However, real physical experiments do not provide hyperspectral measurements V_λ^k but instead record RGB images $V^{c,k}$. To incorporate $V^{c,k}$, we need to make the assumption that the PSF P_λ^k is wavelength-invariant when $\sum_{\lambda \in \Lambda_{\text{disc}}} \kappa_\lambda^c = 1$. Under this assumption, the equation becomes (we drop Λ_{disk} for conciseness):

$$\begin{aligned}
\widehat{U}^c &= \sum_{k=1}^K \frac{\sum_\lambda \kappa_\lambda^c \overline{P_\lambda^k}}{\sum_{k=1}^K |\sum_\lambda \kappa_\lambda^c P_\lambda^k|^2 + K\beta} \odot \sum_\lambda \kappa_\lambda^c V_\lambda^k \\
&= \sum_{k=1}^K \frac{\sum_\lambda \kappa_\lambda^c \overline{P_\lambda^k}}{\sum_{k=1}^K |\sum_\lambda \kappa_\lambda^c P_\lambda^k|^2 + K\beta} \odot V^{c,k} \\
&= \sum_{k=1}^K \frac{\overline{P^{c,k}}}{\sum_{k=1}^K |P^{c,k}|^2 + K\beta} \odot V^{c,k} \\
&= \sum_{k=1}^K \frac{\overline{I(P_\lambda^k)}}{\sum_{k=1}^K |I(P_\lambda^k)|^2 + K\beta} \odot I(V_k^c)
\end{aligned}$$

This reformulation maps the hyperspectral reconstruction into RGB space to directly leverage measured $V_{(\alpha,\gamma)}^{c,k}$. One may then perform deconvolution using either the integrated broadband PSFs $P_{\lambda,(\alpha,\gamma)}^k$ obtained by simulation or the empirically measured RGB PSFs $P_{(\alpha,\gamma)}^{c,k}$. In practice, however, wavelength-invariant PSFs rarely hold, prompting previous literature to incorporate explicit spectral consistency losses [**?**] in broadband deconvolution. By contrast, we adopt an end-to-end training paradigm that implicitly enforces spectral consistency, resulting in superior broadband imaging performance.

Note that this form of $\widehat{U}^c$ is straightforward if we derive the joint Wiener filter for each color channel c starting from the approximate formation model in Eq. (S4). This suggests that the assumption that $\sum_{\lambda \in \Lambda_{\text{disc}}} \kappa_\lambda^c = 1$ is what allows to permute I and $*$. We could thus measure how far away from the real model we are by simply measuring how far $\sum_{\lambda \in \Lambda_{\text{disc}}} \kappa_\lambda^c$ is from 1.

A.3 Noise Variance After Joint Wiener Deconvolution (Section 3.3 in main)

We assume that each sensor noise ϵ^k (we drop here the color upscript c) is i.i.d. white in the spatial domain with variance σ_k^2. Let us denote by η^k the Fourier transform of the noise ϵ^k. Evaluating the joint Wiener filter applied to K measurements reads:

$$\begin{aligned}\widehat{U} &= \frac{\sum_{k=1}^K \overline{P^k} \odot V^k}{\sum_{k=1}^K |P^k|^2 + K\beta} \\ &= \frac{\sum_{k=1}^K \overline{P^k} \odot (P^k \odot U + \eta^k))}{\sum_{k=1}^K |P^k|^2 + K\beta} \\ &= \frac{\sum_{k=1}^K |P^k|^2 \odot U}{\sum_{k=1}^K |P^k|^2 + K\beta} + \frac{\sum_{k=1}^K \overline{P^k} \odot \eta^k}{\sum_{k=1}^K |P^k|^2 + K\beta}.\end{aligned}$$

The term on the left corresponds to the restored image and the term on the right corresponds to the boosted correlated noise in the final image. We call η the correlated noise:

$$\eta = \frac{\sum_{k=1}^K \overline{P^k} \odot \eta^k}{\sum_{k=1}^K |P^k|^2 + K\beta}. \tag{S13}$$

One can note that the distribution of the noise in the prediction $\widehat{U}$ depends on the distribution of ϵ^k and the K MTF of our camera array.

Since we assume that ϵ^k is zero-mean, we can compute the variance of this correlated noise via the power spectrum of η, we call S_η. First, the expected square modulus of η reads for each spatial frequency ω (we drop the dependency on ω everywhere for conciseness):

$$\begin{aligned}\mathbb{E}[|\eta|^2] &= \frac{\mathbb{E}[|\sum_{k=1}^K \overline{P^k} \odot \eta^k|^2]}{(\sum_{k=1}^K |P^k|^2 + K\beta)^2} \\ &= \frac{\mathbb{E}[\sum_{k=1}^K |\overline{P^k}\eta^k|^2]}{(\sum_{k=1}^K |P^k|^2 + K\beta)^2} \\ &= \frac{\sum_{k=1}^K |P^k|^2 \mathbb{E}[|\eta^k|^2]}{(\sum_{k=1}^K |P^k|^2 + K\beta)^2} \\ &= \frac{\sum_{k=1}^K |P^k|^2 \sigma_k^2}{(\sum_{k=1}^K |P^k|^2 + K\beta)^2}.\end{aligned}$$

Passing from the second to third line uses the fact that the different instances of noise are independent. Passing from the third to the fourth line uses the fact that the expected energy of white noise at a given frequency ω is its variance.

Summing over all the spatial frequencies ω gives S_η, which normalizes the contribution of all the frequencies on the Fourier domain and gives the total energy of the noise, which is the variance of the correlated noise in the reconstructed image $\widehat{U}$ (up to a normalization factor):

$$S_\eta \;=\; \sum_\omega \frac{\sum_{k=1}^K |P^k(\omega)|^2 \sigma_k^2}{(\sum_{k=1}^K |P^k(\omega)|^2 + K\beta)^2}. \tag{S14}$$

This final expression can be evaluated for each wavelength (or wavelength band) in a broadband meta-lens array design. By combining these wavelength-specific noise variances appropriately, we obtain an overall noise estimate for the multi-lens, multi-wavelength joint Wiener deconvolution result.

B ADDITIONAL DETAILS ON COMPUTATIONAL MODELING OF METASURFACE

B.1 Proxy FDTD Simulation Details (Section 3.1 in main)

The phase and amplitude response of the nanopillar for the proxy estimation were computed through simulations using the finite-difference time-domain (FDTD) method in Lumerical software. The unit cell comprised a silicon nitride pillar deposited on a fused silica substrate, with a fixed thickness of 1000 nm and widths ranging from 80 nm to 280 nm. The array featured a pitch of 350 nm. Periodic boundary conditions were applied along the x- and y-axes, with perfectly matched layers (PML) at the top and bottom boundaries. A linearly polarized plane wave (x-polarized) was incident from the substrate side along the positive z-axis. The phase and amplitude data were collected using a power monitor positioned several wavelengths above the silicon nitride layer.

B.2 Shifted ASM Kernel (Section 3.1 in main)

Wave propagation in free space from one plane to another can be calculated by Rayleigh-Sommerfeld scalar diffraction integral [Goodman 2005], numerically implemented by a band-limited (shifted) angular spectrum propagation model [Matsushima 2010; Matsushima and Shimobaba 2009] as

$$\begin{aligned} \mathbf{U}_d^{(\alpha,\gamma)} &= \mathcal{F}^{-1}\{\mathcal{F}\{\mathbf{U}_s^{(\alpha,\gamma)}\} \otimes \mathbf{H}\} \\ \mathbf{H}(u,v) &= e^{j2\pi(\Delta x_{p^{(\alpha,\gamma)}} u + \Delta y_{p^{(\alpha,\gamma)}} v + f w)} \operatorname{rect}\left(\frac{u-u_0}{u_{\text{width}}}\right) \operatorname{rect}\left(\frac{v-v_0}{v_{\text{width}}}\right) \\ \boldsymbol{w}(u,v) &= \begin{cases} e^{j2\pi f\sqrt{\frac{1}{\lambda^2}-u^2-v^2}} & \text{if } \sqrt{u^2+v^2} \le \frac{1}{\lambda} \\ 0 & \text{otherwise} \end{cases} \end{aligned} \tag{S15}$$

where $\mathbf{U}_s^{(\alpha,\gamma)}$ and $\mathbf{U}_d^{(\alpha,\gamma)}$ denote source and destination fields of interest. $\mathcal{F}\{\cdot\}$ represents the fast Fourier transform, and u, v are the corresponding spatial frequencies. $\operatorname{rect}(\cdot)$ is the window function that truncates the high-frequency parts of the transfer function to suppress aliasing, see [Matsushima 2010] for the detailed derivation. f is the focal length of the designed camera. $(\Delta x_{p^{(\alpha,\gamma)}}, \Delta y_{p^{(\alpha,\gamma)}}) = (-f \tan\alpha, -f \tan\gamma)$ indicate the spatial coordinate shifts — the origin of the destination coordinates $(\hat{x}, \hat{y})$ is shifted from the source coordinates as $\hat{x} = x + \Delta x_{p^{(\alpha,\gamma)}}$, $\hat{y} = y + \Delta y_{p^{(\alpha,\gamma)}}$. We utilize this algorithm for PSF computation. Given the wavefront modulated by metalens defined in Equation (7) of the main paper, we can derive the destination field in the sensor plane with Eq. (S15) and finally compute its intensity (the square of the field amplitude) to get the corresponding PSF. Note the transfer function $\mathbf{H}(u,v)$, *i.e.*, the ASM kernel, also depends on the incident angle (α, γ).

B.3 Distributed Learning Pipeline Details (Section 3.4 in m qain)

Our distributed framework addresses three core issues: **data distribution**, **model sharding**, and **communication synchronization**. To meet these demands, our pipeline features three principal components: a *distributed incident wavefield sampler*, *forward and backward model sharding*, and *global information synchronization*.

For illustration, we will use our optical training task as an example where six 1.5 mm-aperture sublenses are jointly optimized for broadband, full-angle illumination. Each of the following components plays a distinct role in this pipeline:

Distributed Incident Wavefield Sampler. In our framework, the incident wavefield—characterized by wavelength (λ) and incident angles (α, γ)—serves as the input data. A naive parallelization strategy, assigning each GPU rank a static subset of wavelengths or angles, would be prohibitively memory-intensive. For instance, sampling wavelengths at 0.5 nm resolution from 400 nm to 700 nm, along with 1° angle increments from 0° to 30°, demands approximately 360 TB of GPU memory for our computational model. To circumvent this, we employ a *stochastic sampling* approach. Rather than allocating fixed spectral or angular ranges, each iteration randomly selects a

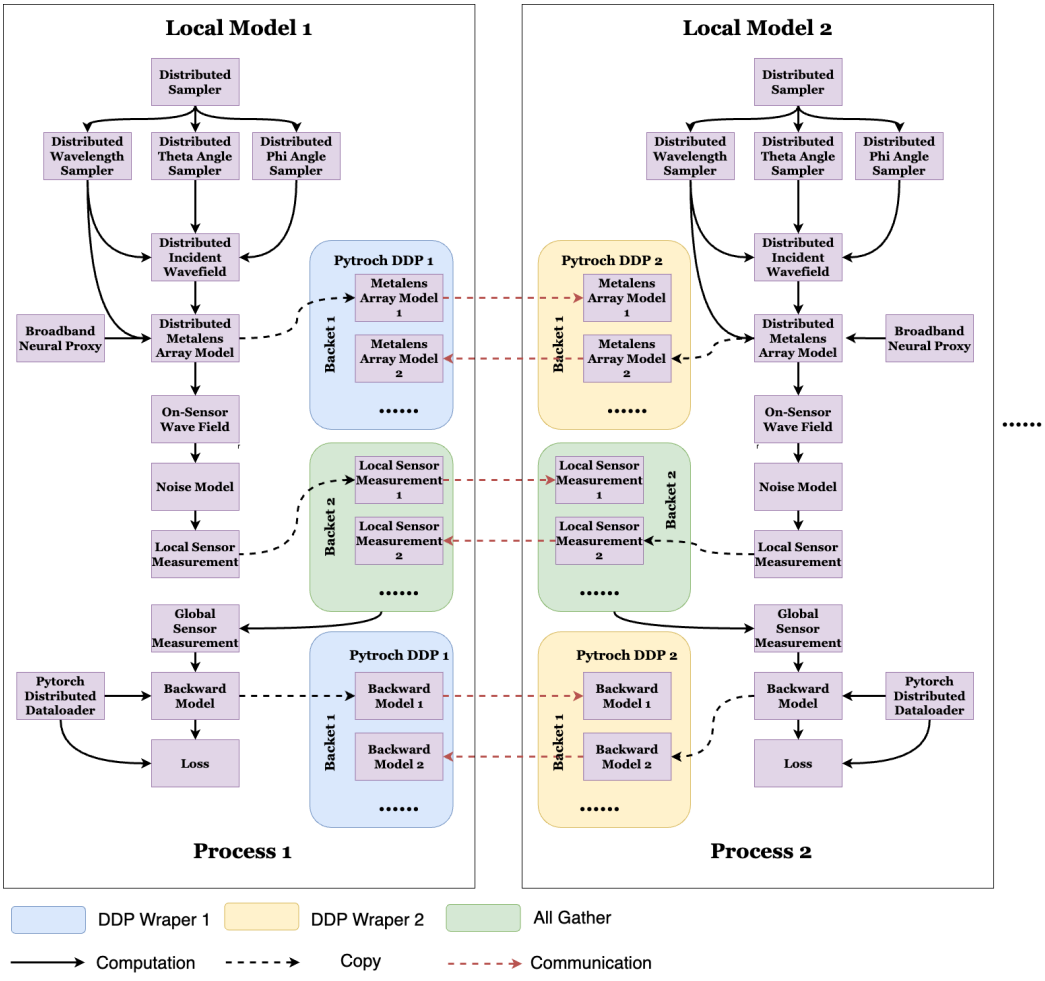

Fig. S1: Proposed Distributed Large Meta-Optics Training Framework.

subset of wavefields covering diverse wavelengths and angles. This ensures thorough exploration of the high-dimensional sampling space over multiple iterations and keeps per-GPU memory requirements within practical limits. After this sampling stage, each GPU rank receives a distinct set of three wavefields for further processing.

Forward and Backward Model Sharding. Although the parameter count of a single meta-optics model can typically fit within a single GPU, the intermediate tensors from phase-response calculations over multiple wavefields can exceed the GPU memory by up to two orders of magnitude. To address this, we split the model across several GPU ranks and partition the wavefield inputs accordingly. Leveraging PyTorch's Distributed Data Parallel (DDP) [Li et al. 2020], our pipeline divides both the model and data, ensuring each GPU stores only a fraction of the intermediate results. This design lowers memory usage by approximately a factor of N (the world size), while maintaining an end-to-end training flow. Furthermore, our pipeline supports Fully Sharded Data Parallel (FSDP) [Zhao et al. 2023], which fragments the meta-optics model itself across multiple ranks,

dynamically reassembling the shards at runtime. This flexibility is particularly advantageous for multilayer meta-optics or cases in which backward computations incur a large memory footprint.

In our example, each rank initializes its own instance of the metalens array model. The model is wrapped via DDP or FSDP so that gradient synchronization among intermediate results is handled automatically once the sharding strategy is defined. Incident wavefields are modulated by the metasurface on a per-rank basis and propagated to the sensor plane to form the PSF, all computed locally on each GPU.

Global Information Synchronization. Beyond parameter and gradient synchronization, which is managed by DDP or FSDP, large-scale meta-optics models require additional coordination of intermediate computational results across ranks. For example, the broadband PSF calculation during training (see Eq. (16)) necessitates an `all-gather` operation, aggregating each rank's partial PSFs into a unified dataset.

Our framework features several custom utility functions to facilitate cross-rank tensor gathering *with* gradient tracking. Once the broadband PSFs are collected from all ranks, the model continues with its forward pass and subsequent backward pass. Rank-specific losses are computed and then backpropagated, with synchronization performed by the DDP/FSDP module. Since the entire pipeline remains differentiable and global synchronization retains gradients, the resulting loss function can effectively backpropagate through all metalens parameters, enabling an end-to-end update of the metasurface design to achieve the desired task objectives.

C ADDITIONAL EXPERIMENTAL PROTOTYPE DETAILS

Here, we describe additional details on the implemented hardware prototype.

C.1 Meta-Optic Fabrication

To fabricate the device, a 1000 nm silicon nitride (SiN) thin film was deposited on a 500 µm fused silica wafer using plasma-enhanced chemical vapor deposition (PECVD) in an SPTS chamber. The wafer was diced into 1.5 cm × 1.5 cm chips with a Disco wafer dicer. A positive-tone resist (ZEP 520A) was spin-coated at 4000 rpm for 60 s, baked at 180 °C for 3 min, then coated with a conductive polymer (DisCharge H2O) to mitigate charging.

Resist patterning was performed on a JEOL JBX-6300FS 100 kV electron-beam lithography system at a dose of 275 µC cm^{-2} and developed in amyl acetate for 2 min. A 67 nm aluminum oxide layer was deposited by electron-beam evaporation, followed by overnight liftoff in N-methyl-2-pyrrolidone (NMP). The exposed SiN was etched in an Oxford PlasmaLab 100 ICP-RIE using a C_4F_8/SF_6 gas mixture.

For aperture fabrication, a negative-tone resist (NR9G3000PY) was spin-coated, patterned via optical direct laser writing (Heidelberg DWL66), and developed in AZ 726 MIF for 1 min. A 200 nm chromium layer was deposited by e-beam evaporation, and liftoff was carried out in acetone for 5 min. The final metasurface array measures 6.57 mm × 10.64 mm. A customized inner baffle was then designed to mount the array above the bare-board sensor.

C.2 Optical Mounting (Section 4.2 in main)

We designed a specialized internal baffle for metasurface on-sensor mounting to mitigate optical cross-talk among the sublenses. We first obtained a 3D model of the board-level sensor (Allied Vision, Alvium 1800 U-2050c) from the official website [AlliedVision 2025]. Due to the presence of the cover glass, we experimentally measured the distance between the sensor cover glass and the best lens focusing plane to be 1.3 mm. This implies that the total thickness between the outer cover glass plane and the sensor plane is 2.3 mm.

Using the 1.3 mm distance, we calculated the aperture size of each lens opening on the inner baffle based on the 52° field of view (FOV) of the sublens. We also incorporated a lofting operation into the baffle design to reduce internal reflections. Subsequently, we 3D-printed the inner baffle using stereolithography with Accura 7820 material. After precisely aligning the metalens with the baffle apertures, we used tape to affix the metasurface onto the baffle. Finally, we fastened the assembly to the board-level sensor using M2 screws. Notably, by employing an internal baffle rather than an external one, we did not introduce any additional thickness to the imaging system.

C.3 Experimental Acquisition with A Dual Camera Setup (Section 4.2 in main)

We constructed a dual-camera characterization system for reference captures and comparison. Figure S3 illustrates the setup, along with the hyperspectral PSF calibration system: (a) a precision optical breadboard assembly with kinematic mounts supporting our metalens prototype and a FLIR reference camera, and a beam splitter at calculated angles, utilizing industry-standard mounting hardware for precise and robust alignment; and (b) a hyperspectral PSF calibration module incorporating a linear translation stage (Thorlabs, XF100) with millimeter precision for positioning the spectral filter (Edmund Optics, 88-365) during bandpass PSF characterization.

Figure S4 shows our custom-developed graphical user interface (GUI) for the metalens/ground truth (GT) camera capture system provides comprehensive control over both imaging systems. The interface features three main control sections: Metalens Settings, GT Settings, and Capture Settings. The Metalens Settings panel allows adjustment of exposure time (in milliseconds) and white balance parameters (red and blue). The GT Settings section enables configuration of the ground truth camera with an exposure ratio (relative to the metalens camera) and white balance values. The Capture Settings section includes options for single image capture and video

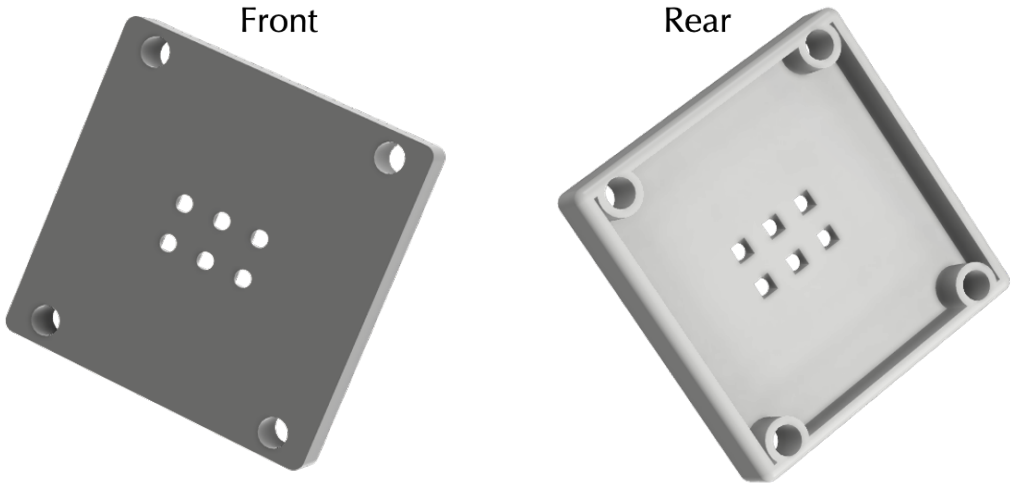

Fig. S2: We 3D-printed a customized optical baffle that mounts the metalens together with the sensor and serves as an optical stop to prevent crosstalk between adjacent lenses.

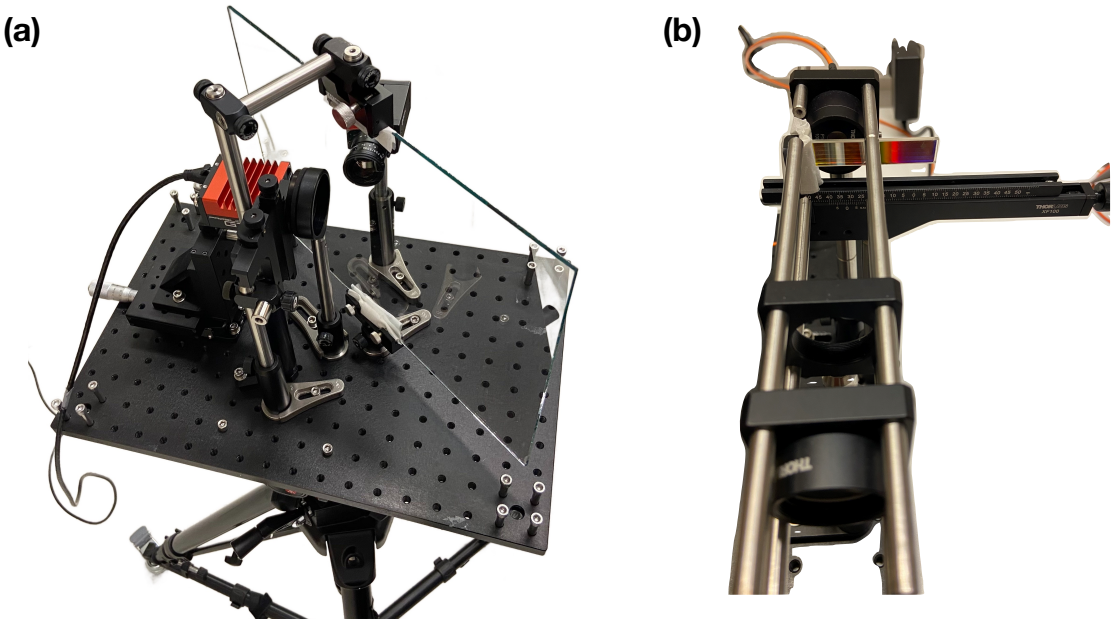

Fig. S3: Experimental setup for dual-camera system and PSF characterization: (a) optical breadboard configuration showing kinematic mounts, two cameras, and precision beam splitter arrangement; (b) hyperspectral PSF calibration module with Thorlabs XF100 translation stage and Edmund Optics 88-365 spectral filter.

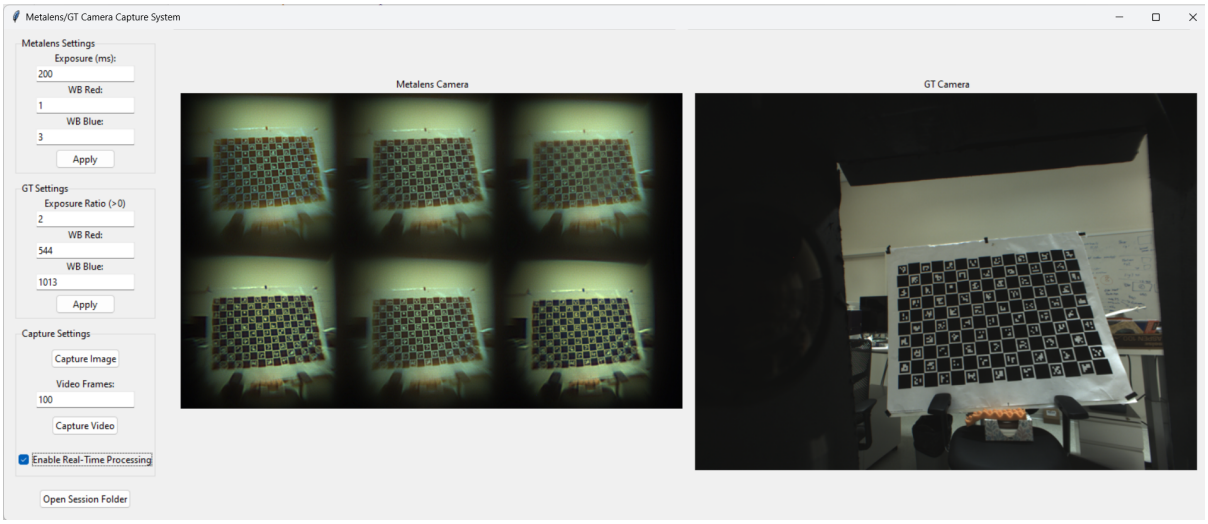

Fig. S4: Metalens/GT camera control interface showing dual-camera preview and parameter adjustment panels. Control panels for exposure, white balance, capture settings, and real-time preview of metalens camera output (2×3 grid). The real-time processing and visualization here are preliminary; not the paper's full method. The scene is a ChArUco broad pattern for homography calibration.

recording with an adjustable frame count. The interface displays real-time previews from both cameras side-by-side: the metalens camera showing a 2×3 grid of captures with a ChArUco broad pattern, and the GT camera displaying a single view of the same ChArUco broad pattern. A checkbox for 'Enable Real-Time Processing' and an 'Open Session Folder' button provide additional functionality for processing, real-time reconstruction, and data-saving management. The GUI allows us to capture images and videos for indoor and outdoor scenes and visualizes preliminary reconstruction results in real-time. Exposure times were adjusted differently for scenes under various illumination conditions.

C.4 Experimental PSF Measurements

The optical hyperspectral PSFs were measured to validate the fabrication quality of the prototype. A fiber-coupled broadband LED source (Thorlabs, MBB1F1) was spatially filtered through a 100 µm pinhole (Thorlabs, P100K) and spectrally filtered using a linear variable bandpass filter (Edmund Optics, 88-365). The filtered light was collimated by an achromatic doublet lens (Thorlabs, AC254-150-A-ML) with an iris (Thorlabs, ID20). The calibration setup was positioned at ~2 m from the prototype to ensure proper PSF formation on the sensor plane. For spectral sampling, the variable bandpass filter was mounted on a manual translational stage (Thorlabs, XF100) and laterally shifted in 1 mm increments, spanning 400-700 nm. This yielded 37 distinct spectral bands with 8.1 nm bandwidth per PSF measurement. At each spectral band, PSFs were captured across multiple exposure times (20-1500 ms), with exposure time adjusted per spectral window to compensate for the source's spectral power distribution and prototype image sensor spectrum response. Figure S5 shows the measured PSFs for each sublens across the spectrum range, normalized to each spectrum band for visualization purposes. While all sub-lenses exhibit spectrally uniform PSFs, different elements focus in different bands of the spectrum – matching the trend in simulation – and this allows us to collaboratively recover full-color images.

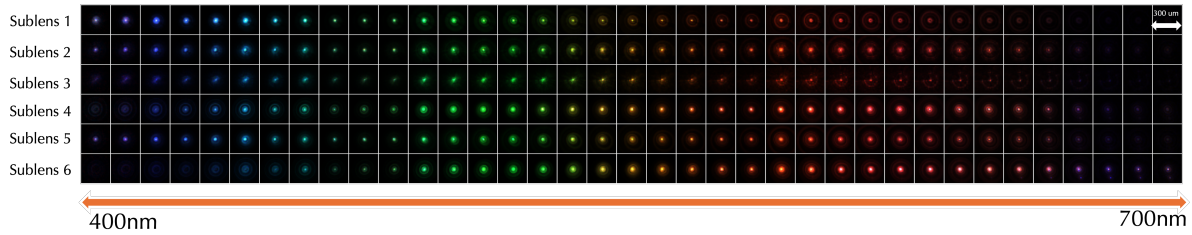

Fig. S5: Measured hyperspectral PSFs from the 2×3 metalens array demonstrate sharp focusing with minimal aberrations across 400 nm-700 nm over 37 spectral bands, validating our broadband design.

D PSF SIMULATION AND ADDITIONAL SYNTHETIC RESULTS

We present additional synthetic PSF simulations that visualize the full-angle response of the proposed lens array design, along with ablation results, to further support the effectiveness of the proposed reconstruction method discussed in the main manuscript.

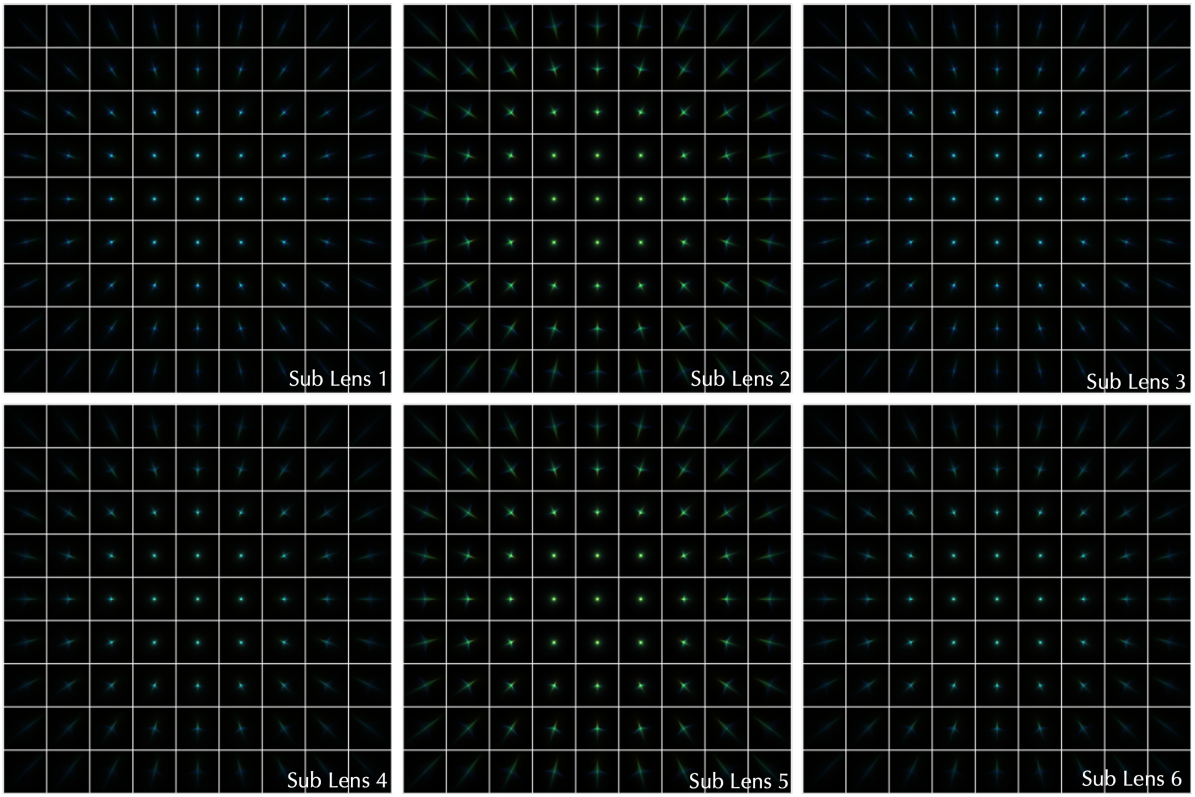

Fig. S6: Simulated Full-Angle PSF Response of the Proposed Metalens Array. We simulated the 9×9 RGB spatially varying PSFs based on the incident angle of the center pixel of each image patch. The PSFs were calculated using the shifted ASM kernel in hyperspectral space and subsequently integrated into RGB space based on the camera spectral response function. For improved visualization, the PSFs were center-cropped. Notably, the collaborative lens array exhibits different spectral focusing behaviors for each sublens, as evidenced by the varying RGB colors of the PSFs under identical broadband plane waves.

Full-Angle PSF Response Simulation. The PSFs were calculated based on the center pixel coordinates of the image patches and the focal length of the proposed lens array design. In our experiments, we divided the 576×576 sub-image measurements into 9×9 patches, each with a size of 64×64. The incident angles (α, γ) were derived from the center pixel coordinates (x, y) of the patches and the focal length $f = 3.6\,\text{mm}$, as follows:

$$\alpha = \arctan\left(\frac{x}{f}\right), \quad \gamma = \arctan\left(\frac{y}{f}\right). \tag{S16}$$

This calculation resulted in the following incident angles for both α, γ:

$-22.6220°, -17.3558°, -11.7695°, -5.9475°, 0.0000°, 5.9475°, 11.7695°, 17.3558°, 22.6220°$.

For each incident angle combination, we utilized the proposed distributed meta-optics framework to infer the hyperspectral response for wavelengths sampled every 2 nm across the visible spectrum. These hyperspectral responses were then integrated into RGB space using the camera spectral response function. The broadband array PSF tensors were saved as intermediate results, which were subsequently used to train the reconstruction network to accelerate the training speed.

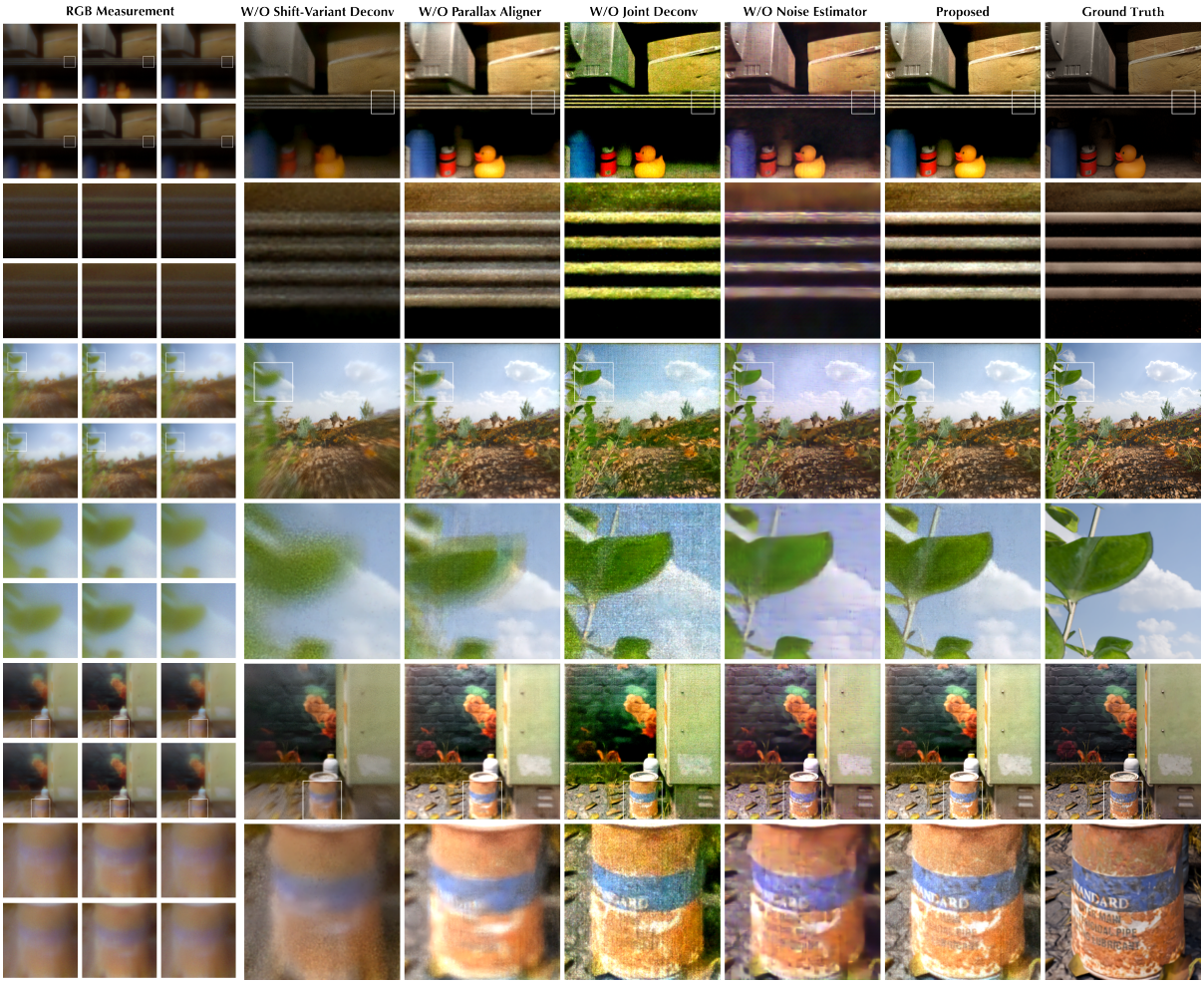

Fig. S7: Additional Ablation Experiments on the Proposed Reconstruction Method. In line with the ablation findings presented in the main text, the method's performance degrades distinctly when each submodule is disabled, underscoring the necessity of each design component.

Additional Ablation Study Results on the Proposed Synthetic Hyperspectral Dataset. We conducted further ablation experiments on the synthetic hyperspectral array dataset exhibiting parallax. The results corroborate the observations made in the main manuscript. Specifically, deactivating the shift-variant deconvolution module

reduces resolution in the peripheral regions, while omitting the parallax aligner adversely affects the resolution of nearby objects. Replacing joint deconvolution with six individual Wiener deconvolutions introduces color-aberration artifacts, and removing the noise estimator prevents compensation for wavelength-dependent noise. Collectively, these findings reinforce the efficacy of our reconstruction method and its various submodules.

E ADDITIONAL EXPERIMENTAL RESULTS

Fig. S8: Additional Reconstruction Results on the Experimental Capture Dataset.

To further corroborate the real-world reconstruction performance, we include additional results from the experimental capture dataset, covering both indoor and outdoor scenes. The reconstruction results are consistent with those presented in the main paper, maintaining high-quality reconstructions despite varying illumination conditions.

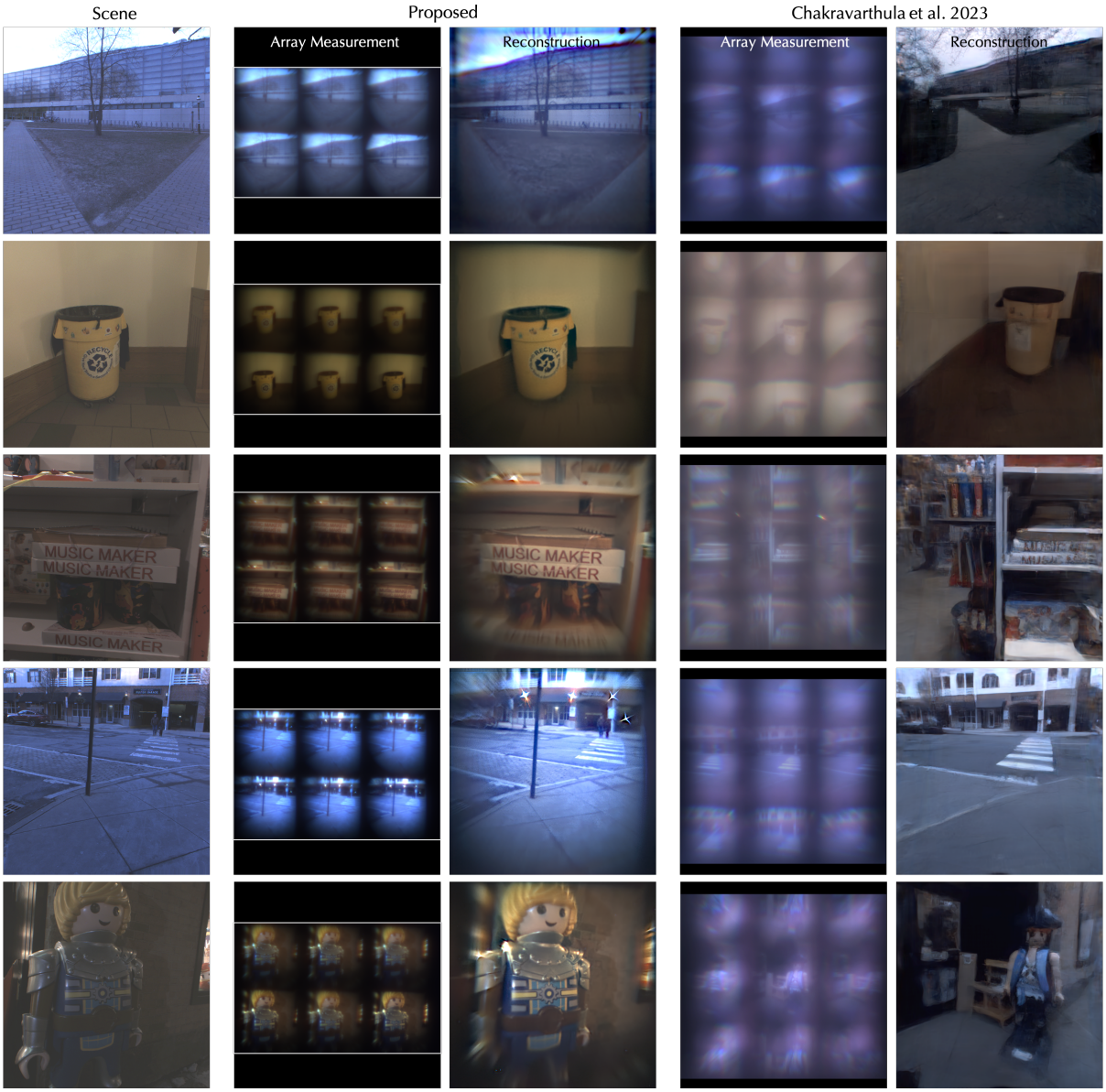

Fig. S9: Additional results on comparison with [Chakravarthula et al. 2023].

Under similar scenes and illumination conditions, the collaborative metalens array provides hallucination-free reconstruction results. Here, we present more results of different indoor and outdoor scenes under different illumination conditions. From top to bottom, the collaborative metalens array shows details like building wall textures, text on the trash bin, text on the book, bricks on the ground, and details on the toy model, which are invisible or not distinguishable in the corresponding [Chakravarthula et al. 2023] results.

REFERENCES

AlliedVision. 2025. Allied Vision Offical Website. (2025). https://www.alliedvision.com/en/products/alvium-configurator/alvium-1800-c/2050/

Praneeth Chakravarthula, Jipeng Sun, Xiao Li, Chenyang Lei, Gene Chou, Mario Bijelic, Johannes Froesch, Arka Majumdar, and Felix Heide. 2023. Thin on-sensor nanophotonic array cameras. *ACM Transactions on Graphics (TOG)* 42, 6 (2023), 1–18.

Joseph W Goodman. 2005. *Introduction to Fourier optics*. Roberts and Company publishers.

Shen Li, Yanli Zhao, Rohan Varma, Omkar Salpekar, Pieter Noordhuis, Teng Li, Adam Paszke, Jeff Smith, Brian Vaughan, Pritam Damania, et al. 2020. Pytorch distributed: Experiences on accelerating data parallel training. *arXiv preprint arXiv:2006.15704* (2020).

Kyoji Matsushima. 2010. Shifted angular spectrum method for off-axis numerical propagation. *Optics Express* 18, 17 (2010), 18453–18463.

Kyoji Matsushima and Tomoyoshi Shimobaba. 2009. Band-limited angular spectrum method for numerical simulation of free-space propagation in far and near fields. *Optics Express* 17, 22 (2009), 19662–19673.

Yanli Zhao, Andrew Gu, Rohan Varma, Liang Luo, Chien-Chin Huang, Min Xu, Less Wright, Hamid Shojanazeri, Myle Ott, Sam Shleifer, et al. 2023. Pytorch fsdp: experiences on scaling fully sharded data parallel. *arXiv preprint arXiv:2304.11277* (2023).