Thin On-Sensor Nanophotonic Array Cameras Supplementary Information

PRANEETH CHAKRAVARTHULA*, JIPENG SUN*, XIAO LI, CHENYANG LEI, GENE CHOU, and MARIO BIJELIC, Princeton University, USA
JOHANNES FROESCH and ARKA MAJUMDAR, University of Washington, USA
FELIX HEIDE, Princeton University, USA

CCS Concepts: • **Computing methodologies → Computational photography**.

Additional Key Words and Phrases: Computational Optics

ACM Reference Format:
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 Supplementary Information. *ACM Trans. Graph.* 42, 6, Article 252 (December 2023), 21 pages. https://doi.org/10.1145/3618398

OUTLINE

This supplementary document provides further description and additional results to support the findings from the main manuscript. The document is organized as follows.

Section A: Additional details on the reconstruction algorithm presented in the main manuscript.
Section B: Further details on the fabrication of the prototype metasurface optics in this section.
Section C: To facilitate the reproducibility of the results presented in this work, this section lists additional details on the setup and experimental capture protocols.
Section D: Additional details on the generation of the synthetic dataset. We also list additional samples of the dataset used to train and evaluate our approach.
Section E: This section provides additional synthetic results in support of the findings in the main manuscript.
Section F: This section provides additional experimental results that further validate the findings of the main document. Additional experimental in-the-wild captures and captures of a monitor with a sparse spectral response are presented as results in addition to the results from the main paper.

A ADDITIONAL DETAILS ON THE IMAGE RECONSTRUCTION METHOD

In this section, we provide additional details on the proposed nanophotonic phase optimization and probabilistic image deconvolution method for the on-sensor array camera.

*Indicates equal contribution.

0730-0301/2023/12-ART252 $15.00
https://doi.org/10.1145/3618398

A.1 Closed-Form Solution for Data Fidelity Term

As described in Sec. 4.1 of the main manuscript, the linear data fidelity term of the alternating optimization objective

$$\mathbf{I}^{t+1} = \arg\min_{\mathbf{I}} \frac{1}{2}\left\|\mathbf{I} \otimes \mathbf{k} - \mathbf{S}\right\|^2 + \frac{\mu^t}{2}\left\|\mathbf{I} - \mathbf{z}^t\right\|^2, \tag{1}$$

can be solved in closed form assuming a circular convolution, with the following inverse filter update

$$\mathbf{I}^{t+1} = \mathcal{F}^{\dagger}\left(\frac{\mathcal{F}^*(\mathbf{k})\mathcal{F}(\mathbf{S}) + \mu^t\mathcal{F}(\mathbf{I}^t)}{\mathcal{F}^*(\mathbf{k})\mathcal{F}(\mathbf{k}) + \mu^t}\right), \tag{2}$$

where $\mathcal{F}(\cdot)$ denotes the Fast Fourier Transform (FFT), $\mathcal{F}^*(\cdot)$ denotes the complex conjugate of FFT, and $\mathcal{F}^{\dagger}(\cdot)$ denotes the inverse FFT. We derive the solution here.

Primer. Typical filtering in image processing is represented in three forms:

- Matrix form $\mathbf{Ax} = \mathbf{b}$ denoted by the matrix multiplication of a filter $\mathbf{A}$ and vectorized images $\mathbf{x}$ and $\mathbf{b}$,
- Filter form $A \otimes X = B$ denoted by convolution of a filter kernel A and images X and B, and
- Using a fast Fourier transform (FFT) $\mathcal{F}(A) \circ \mathcal{F}(X) = \mathcal{F}(B)$ where $\mathcal{F}(\cdot)$ is the FFT and $\circ$ denotes the Hadamard pixel-wise product.

In the case of convolution with circular boundary conditions, the above three representations are equivalent, *i.e.*

$$\mathbf{Ax} = \mathbf{b} \leftrightarrow A \otimes X = B \leftrightarrow \mathcal{F}(A) \circ \mathcal{F}(X) = \mathcal{F}(B) \tag{3}$$

Similarly,

$$\mathbf{A}^T\mathbf{x} = \mathbf{b} \leftrightarrow A^{'} \otimes X = B \leftrightarrow \mathcal{F}^*(A) \circ \mathcal{F}(X) = \mathcal{F}(B) \tag{4}$$

where $A^{'}$ is the 180° rotated mirror kernel of A.

Now, the ℓ_2 objective in Eq. (1) can be expressed in matrix form as

$$\mathbf{I}^{t+1} = \arg\min_{\mathbf{I}} \frac{1}{2}\left\|\mathbf{Ki} - \mathbf{s}\right\|^2 + \frac{\mu^t}{2}\left\|\mathbf{i} - \mathbf{z}^t\right\|^2, \tag{5}$$

whose solution can be obtained by setting the first-order derivate w.r.t. $\mathbf{i}$ to zero. Therefore,

$$\mathbf{K}^T(\mathbf{Ki} - \mathbf{s}) + \mu^t(\mathbf{i} - \mathbf{i}^t) = 0 \tag{6}$$

Given the circular boundary condition, following Eq. (3) and Eq. (4), the above Eq. (7) is equivalent to

$$\mathcal{F}^*(\mathbf{k})\left\{\mathcal{F}(\mathbf{I}^{t+1})\mathcal{F}(\mathbf{k}) - \mathcal{F}(\mathbf{S})\right\} + \mu^t(\mathcal{F}(\mathbf{I}^{t+1}) - \mathcal{F}(\mathbf{I}^t)) = 0 \tag{7}$$

which can be rearranged to Eq. (2) for calculating $\mathbf{I}^{t+1}$ as a closed-form solution.

A.2 Diffusion Models

Background on Diffusion Probabilistic Models. While different variations of diffusion models exist, we implement a canonical one [Ho et al. 2020; Sohl-Dickstein et al. 2015]. From a data distribution $q(\mathbf{x})$, we denote a sampled datapoint as x_0, and iteratively add small Gaussian noise to obtain $x_1, x_2 ... x_T$, until x_T approximates an isotropic Gaussian. This forward step is a Markovian fixed process [Ho et al. 2020; Song and Ermon 2019] and can be defined as

$$q(x_{1:T}|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1}), q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I) \tag{8}$$

where β_t is a variance schedule. In practice, we sample x_t using a closed-form parameterization

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon \tag{9}$$

where $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$, and $\epsilon \sim \mathcal{N}(0, I)$.

The goal of each training iteration is to train a model p_θ, often represented by a neural network, that inverts the forward diffusion (i.e., learns the *reverse* diffusion process):

$$p_\theta(x_{0:T}) = p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t) \tag{10}$$

and

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \tag{11}$$

The reverse process is also Markovian, and we fix the variances Σ_θ. The reverse conditional probability is tractable when conditioned on x_0:

$$q(x_{t-1}|x_t, x_0) = \mathcal{N}(x_{t-1}; \tilde{\mu}(x_t, x_0), \tilde{\beta}_t I) \tag{12}$$

We apply Bayes' rule to rearrange the terms and obtain

$$\tilde{\mu}(x_t, x_0) = \frac{\sqrt{\alpha_t}(1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t} x_t + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1 - \bar{\alpha}_t} x_0 \tag{13}$$

The closed form parameterization of x_t yields

$$\tilde{\mu}_t = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_t \right) \tag{14}$$

when we represent x_0 as

$$x_0 = \frac{1}{\sqrt{\bar{\alpha}_t}} \left(x_t - \sqrt{1 - \bar{\alpha}_t}\epsilon_t \right) \tag{15}$$

by rearranging Eq. (9). Thus, we can train our model to predict $\tilde{\mu}_t$, or alternatively, ϵ_t by rearranging the terms. This work predicts $\tilde{\mu}_t$ for generating samples.

During test time, our diffusion model performs generation iteratively. In the vanilla DDPM [Ho et al. 2020], generation is performed as follows

$$z_0 = (f \circ ... \circ f)(z_T, T), \;\; f(x_t, t) = \Omega(x_t) + \sigma_t \epsilon, \tag{16}$$

where $z_T \sim \mathcal{N}(0, I)$, σ_t is the fixed standard deviation at the given timestep, Ω is the model, and $\epsilon \sim \mathcal{N}(0, I)$. However, this results in long sampling times. Instead, we follow DDIM [Song et al. 2021], which proposes a non-Markovian diffusion process to reduce the number of sampling steps. Furthermore, DDIM has a "consistency" property that allows us to manipulate the initial latent variable to guide the generated output. As a result, $f(x_t, t)$ from Eq. (16) can be defined as

$$f(x_t, t) = \sqrt{\alpha_{t-1}} \left(\frac{x_t - \sqrt{1 - \alpha_t}\Omega(x_t)}{\sqrt{\alpha_t}} \right) + \sqrt{1 - \alpha_{t-1} - \sigma_t^2} \cdot \Omega(x_t) + \sigma_t \epsilon. \tag{17}$$

Importantly, from Eq. (12), we have

$$\tilde{\beta}_t = \sigma_t^2 = \frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t} \cdot \beta_t. \tag{18}$$

We let $\sigma_t^2 = \eta \tilde{\beta}_t$ so that we can adjust η to control sampling stochasticity. In a special case when $\eta = 0$, the forward process becomes deterministic except for $t = 1$, and the added random noise during generation becomes zero.

Implementation. For the architecture of our diffusion model, we follow DDPM [Ho et al. 2020] and use a UNet [Ronneberger et al. 2015]. The first input layer takes as input a tensor with 15 channels, where each input condition in Eq. (20) of the main paper has 3 channels. There are 4 downsampling and 4 upsampling layers, with dimensions 64, 128, 256, 512. Each layer contains two ResNet blocks with the corresponding dimensions and self-attention. For the timestep, we employ a sinusoidal positional embedding followed by a 2-layer MLP.

We fix our forward variances β using a cosine schedule, following [Nichol and Dhariwal 2021]. [Ho et al. 2020] set their variances to be a sequence of linearly increasing constants between $[0.0001, 0.02]$, but the authors of [Nichol and Dhariwal 2021] found that the end of the forward noising process was too noisy and could not contribute much to sample quality. Thus, they employed a cosine schedule so that there is a near-linear drop in the middle of the training timesteps and small changes around $t = 0$ and $t = T$.

$$\beta_t = \text{clip}\left(1 - \frac{\bar{\alpha}_t}{\bar{\alpha}_{t-1}}, 0.999\right), \quad \bar{\alpha}_t = \frac{f(t)}{f(0)}, \quad f(t) = \cos\left(\frac{t/T + s}{1 + s} \cdot \frac{\pi}{2}\right)^2. \tag{19}$$

β_t is clipped at 0.999 to prevent singularities near $t = T$ and the offset s prevents excessively small values of β_t near $t = 0$.

B ADDITIONAL DETAILS ON NANOPHOTONIC OPTICS FABRICATION

Here we describe the fabrication details of our on-chip metalens array. In detail, for the fabrication of the meta-optic we first deposited a thin film of 700 nm SiN via plasma-enhanced chemical vapor deposition on top of a 300 µm quartz wafer (purchased from University Wafer). This deposition was performed in a SPTS DELTA LPX using a mix of Silane and Ammonia. After deposition, the wafer was diced into square pieces with a side width of 2 cm. Then, we performed ultrasonication in Acetone and IPA to remove residues from previous steps, followed by a short (30 s) plasma cleaning step using a barrel etcher in O2. Then, a layer of ZEP 520-A was spin-coated on top of the sample with a thickness of 400 nm and baked at 180 C. This was followed by spin-coating a thin layer (DisCharge H2O) of discharging polymer to reduce charging effects during the following lithography step. The optimized phase profiles were transferred into GDS file formats (Figure 1) using custom python scripts and subsequently converted into a specialized file format through GeniSys beamer software, including large scale proximity correction to account for varying pillar sizes and varying doses.

Electron beam lithography was then performed in a JEOL-JBX6300FS EBL system using a 100 kV, 8 nA electron beam. After patterning, the discharging layer was removed in IPA and the polymer was developed in Amyl Acetate for 2 min under gentle agitation. After developing the resist, we performed a short descum step (10 s) using a barrel etcher with O2 plasma. Then, a 65 nm thick layer of Alumina was deposited on top using a home-built e-beam evaporation system with Al2O3 crystals as the evaporation source. The layer was then lifted off in heated (110 C) NMP overnight. After copious rinsing with water and IPA, the sample was further plasma-cleaned in a barrel etcher for 30 s to remove organic residues. We then used inductively coupled plasma reactive ion etching (Oxford Instruments, PlasmaLab100) with SF6 and C4F8 in a 1:2 ratio to etch the SiN layer. An aperture layer was then fabricated on the sample by first using laser direct writing in a Heidelberg-DWL66 with a negative resist. Subsequently, a layer of Cr (5 nm) and gold (150 nm) was deposited and subsequently lifted off in acetone overnight. Optical microscope images after the etching process of the 3 by 3 MO array are shown in Figure 2.

Table 1. Fitted polynomial coefficients for the inverse, phase-to-structure mapping.

Coefficient	b_0	b_1	b_2
Value	-0.1484	0.6809	0.2923

Table 2. Fitted polynomial coefficients for the forward, structure-to-phase mapping. Note that c_{12}, c_{21}, c_{22} are all zero.

Coefficient	c_{00}	c_{01} (nm^{-1})	c_{10}	c_{02} (nm^{-2})	c_{11} (nm^{-1})	c_{20}
Value	6.051	-2.03×10^{-2}	2.26	1.37×10^{-5}	-2.95×10^{-3}	0.797

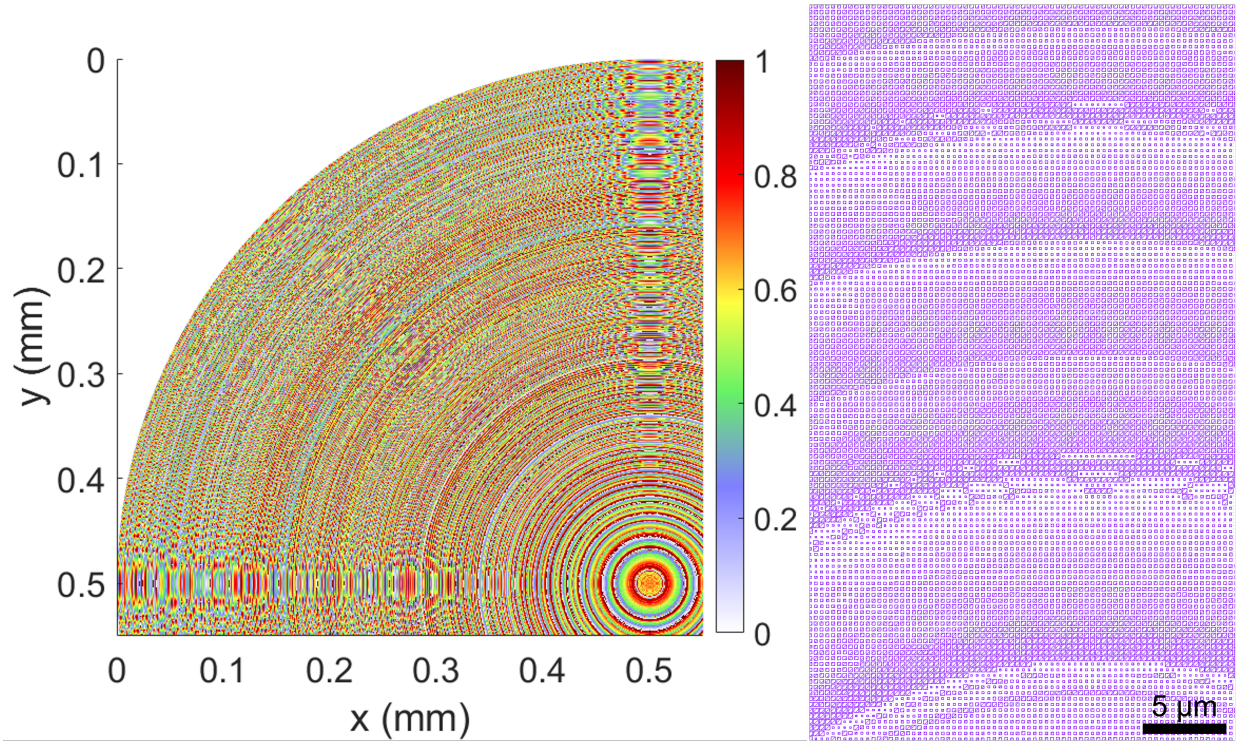

Fig. 1. A quarter of the phase profile for the optimized meta-optic in units of 2π. From the phase profile, the corresponding scatterer was calculated based on the equations mentioned in the main manuscript.

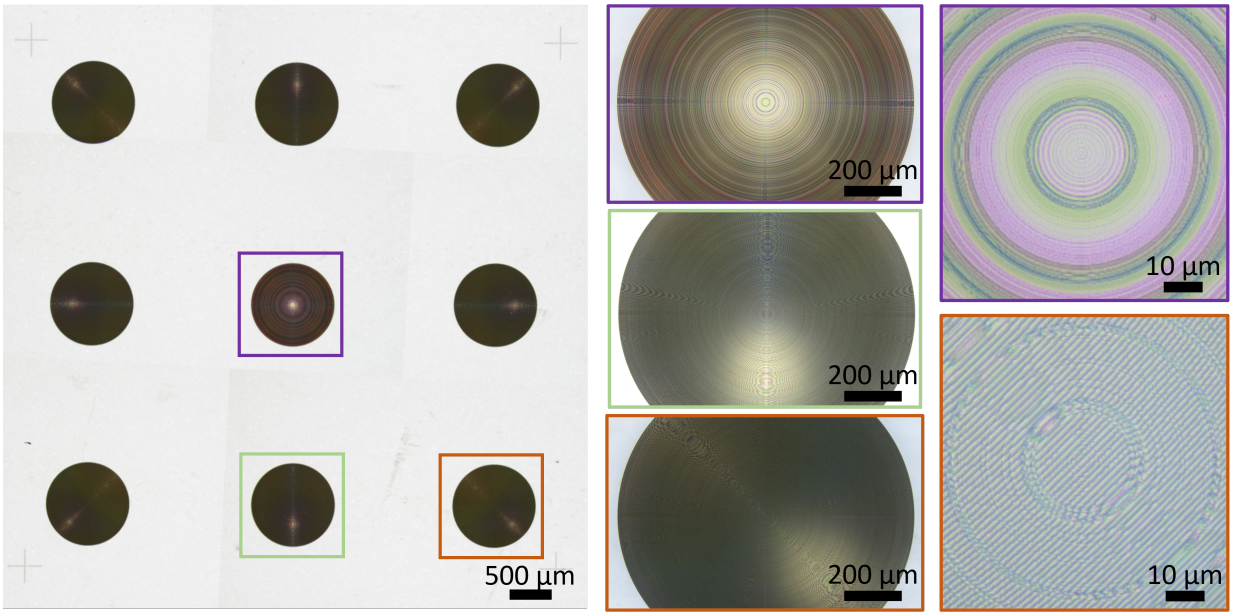

Fig. 2. Optical microscope images of the fabricated array optics. Images at higher magnification of the center, bottom, and bottom right MO are shown in the second column. Further magnified images of the center MO and the bottom right MO are shown in the very right column.

C ADDITIONAL EXPERIMENTAL SETUP DETAILS

In this section, we provide additional details on the experiment prototype, including the hardware design and build, PSF measurements and in-the-wild dataset capture.

C.1 Metalens Array Layout

For the metalens array camera, we employ an Allied Vision GT1930 C sensor, which has 11.34 mm sensor width and 7.13 mm sensor height with 5.86 micron pixel pitch. We design the layout and prismatic wedge angles of the metalens array elements based on the sensor size and the target FoV such that the effective FoV from all the metalens elements in the array can be captured in the same frame. We target a per-element FoV of 40° width and 40° height, which translates to images of 1.82 mm width and 1.82 mm height on the sensor plane. Additional spacing between two adjacent metalens elements is introduced to account for potential FoV overlapping due to the difference between the designed and fabricated device. In the final design, the metalens elements have 1 mm diameter and 2.5 mm focal length, and the center-to-center spacing between adjacent metalens elements is 2.42 mm.

C.2 Experimental Camera Prototype

In the experiment prototype, we employ a plate beam splitter that splits world light into two optical paths such that paired experimental data (metalens array camera captures and ground truth captures) can be acquired. However, this approach comes with many design trade-offs. Assuming that the camera optical center is exactly on the plate beam splitter, the theoretical maximum FoV (Field of View) would be 90° horizontally and vertically. Given the total mechanical length of the conventional off-the-shelf lens used for the reference camera, the distance between the camera optical centers and the plate beam splitter needs to be sufficiently large in order to align the optical centers of both cameras and ensure that the FoV of both cameras is not impacted by the edge of the beam splitter or the mechanical structure of the off-the-shelf lens. However, increasing the distance between the optical centers and the beam splitter would reduce the maximum achievable FoV. While a larger beam splitter would help achieve a larger FoV, it would also make the prototype bulkier and increase the difficulty of mounting it securely. With all those constraints, the experiment prototype that we design and build achieves a maximum FoV around 70° horizontally and vertically for both cameras, which is similar to the FoV of the metalens array camera. We employ an off-the-shelf 10.75” x 8.75” plate beam splitter with 70% transmission and 30% reflection coating on one side and anti-reflective coating on the other side to prevent ghosting.

We employ an Allied Vision GT1930 C sensor for the metalens array camera such that the effective FoV (Field-of-View) from all the metalens elements in the array can be captured in the same frame. The same sensor is employed for the reference camera, which has a wide FoV low-distortion lens with a 3.5 mm focal length from Edmund Optics (stock number #68-669) such that we can achieve a FoV larger than the full FoV of the metalens array camera in the “ground truth” captures. We use Precision Time Protocol (PTP) to synchronize the two cameras such that the captures are taken at the same timestamps with sub-millisecond precision.

To allow for fine alignment, the metalens array camera sensor is mounted on a 3D translation stage. After we align the sensor parallel to the fabricated metalens array, we use the 3D translation stage to precisely shift the sensor position such that the sensor captures the effective FoV of all the metalens array elements and the images are focused on the sensor plane. Next, we align the optical center and optical axis of the central element from the metalens array camera to those of the reference camera. We use a collimated laser and pinhole apertures to make sure the beam splitter is positioned at a 45° tilting angle. Then, we set up the position of the metalens array camera and adjust the laser beam height such that the transmission path is incident on the center metalens element. The center of the reference camera is positioned in the reflection beam path and the distance between the beam splitter and the reference camera sensor is adjusted to the same as that between the beam splitter

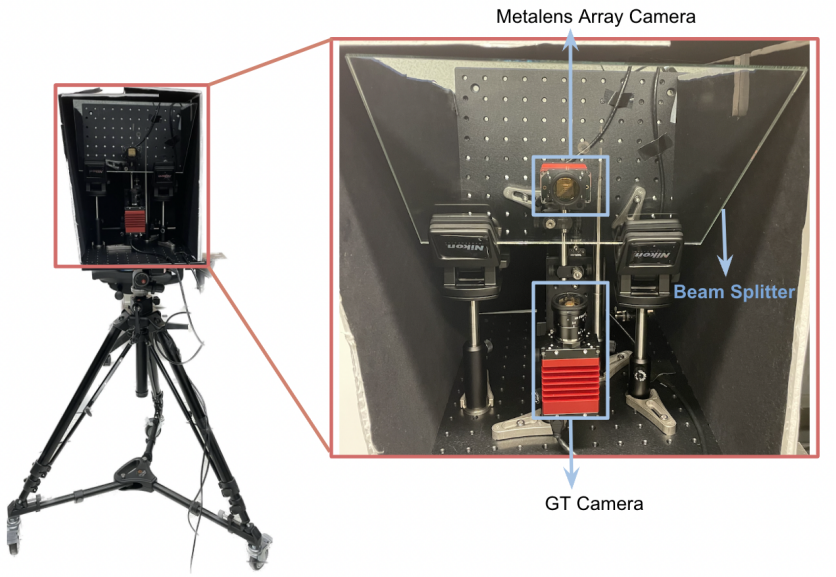

Fig. 3. Experiment data acquisition setup. In our capture setup, we employ a plate beam splitter, which splits world light into two optical paths by 70% transmission and 30% reflection such that the setup can simultaneously capture real-world scenes with one camera in the transmission path that employs the designed metalens array and another camera in the reflection path that employs a conventional off-the-shelf lens (GT camera). The setup is mounted on a tripod with rollers, so that it can be moved around indoors and outdoors for acquiring a diverse dataset.

and the metalens array camera. We achieve more accurate alignment by observing a reference target with both cameras simultaneously until both cameras are aligned. After the alignment is completed, the setup is mounted on a tripod with rollers, as shown in Figure 3, so that it can be moved around indoors and outdoors for acquiring a diverse dataset. The entire setup is put in an enclosure to make sure the reference camera only captures scenes from the desired light path but not the other transmission light path from the ceiling.

C.3 Metalens Array Camera PSF Calibration

After the alignment, we conduct PSF measurements of the individual metalens elements in the array. The light sources that we use are red, green, and blue fiber-coupled LEDs from Thorlabs (M455F3, M530F2, and M660FP1). The fiber has a core diameter of 800 microns and the fiber tip is placed 340 mm away from the metalens array so that it can be seen as a point source with the same angular resolution as that of a pixel in the captured metalens images (~8 arc-min, given a 2.5 mm focal length and a 5.86 micron pixel pitch). The PSFs of all the metalens elements are captured in the same frame. By turning on and off each individual color LED, we can acquire the PSFs of different colors. When alternating between colors, we change the input of the fiber without introducing mechanical shifts to the output of the fiber such that the position of the point light source is fixed. Figure 4 shows the setup for PSF measurement and Figure 5 shows the red, green, and blue PSFs for all the metalens array elements.

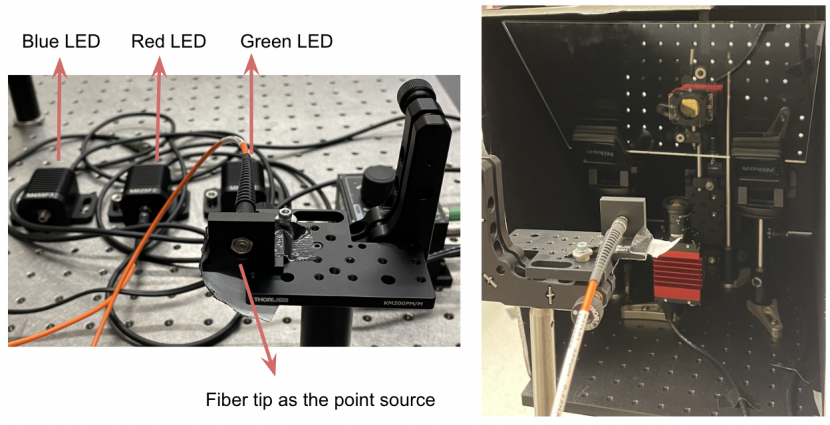

Fig. 4. PSF measurement setup. The light sources that we use are red, green, and blue fiber-coupled LEDs from Thorlabs (M455F3, M530F2, and M660FP1). The fiber has a core diameter of 800 microns and the fiber tip is placed 340 mm away from the metalens array so that it can be seen as a point source with the same angular resolution as that of a pixel in the captured metalens images (~ 8 arc-min, given a 2.5 mm focal length and a 5.86 micron pixel pitch). The PSFs of all the metalens elements are captured in the same frame.

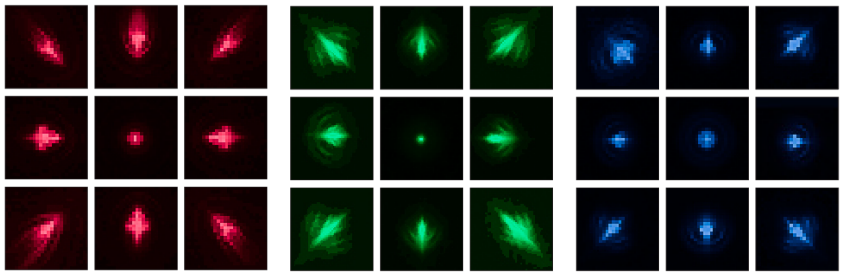

Fig. 5. Red, green, and blue PSFs for all the metalens array elements.

C.4 Camera Calibration and Homography Alignment

To find the per-pixel mapping between the reference camera and metalens array camera, we have both cameras capture red, green and blue checkerboard patterns shown on a large LCD screen and then calibrate the distortion coefficients of the two cameras per color channel. Figure 6 shows some of the captured images of blue, green, and red checkerboards displayed on a LCD screen viewed by the reference camera from different poses.

After the image acquisition, we perform image rectification for the captures from both cameras. Then, to account for the difference in camera FoV and the difference in viewing perspectives between each metalens array element and the reference camera, we perform homography-based alignment to map the reference camera captures to the captures from all the metalens array elements.

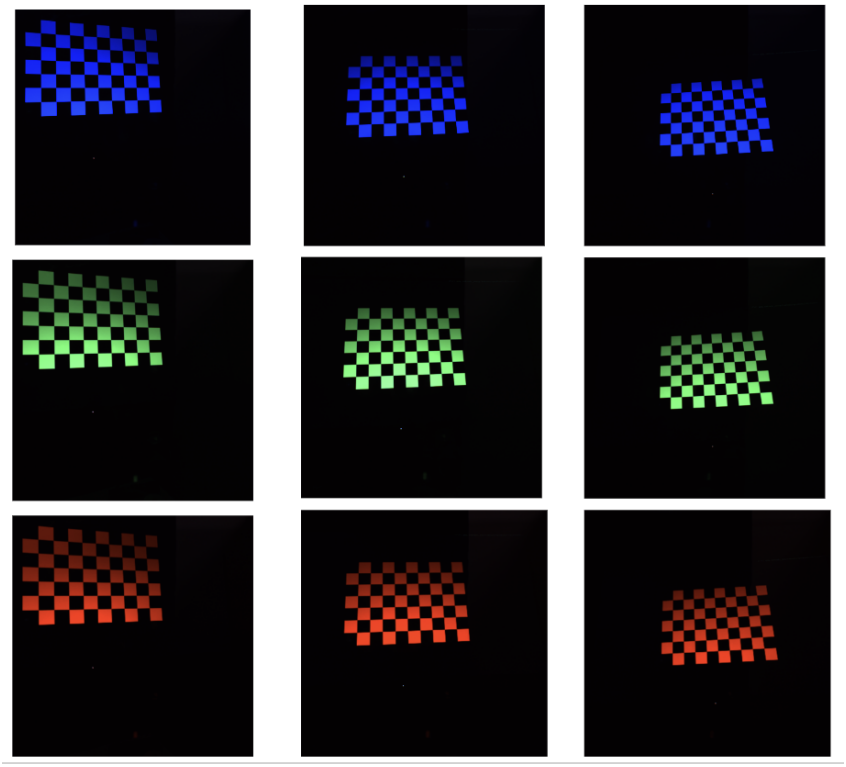

Fig. 6. Calibration patterns of the reference camera. Blue, green, and red checkerboards are displayed on a LCD screen and captured by the reference camera from different poses.

D SIMULATION

Training the probabilistic image recovery network described in the main paper requires a large and diverse set of paired data, which is challenging to acquire in the wild. Therefore, we simulate the nanophotonic array camera with the corresponding metalens design parameters to create a large synthetic dataset of paired on-sensor and groundtruth measurements. We use this large synthetic dataset for training alongside a smaller real-world dataset for fine-tuning. Each metalens in the array camera has a focal length of 2 mm and covers an FoV of 60° for a broadband illumination, with the center-to-center distance between the on-chip metalenses being 2.42 mm. Due to the circular aperture of each meta-optic, the sensor measurements suffer from vignetting at higher eccentricities. We describe the detailed simulated process in the following.

We first crop an image into smaller images to match the number of metalens array camera measurement. Specifically, for a given groundtruth image with a 1080×1080 spatial resolution, we first crop 9 images that correspond to the final 3×3 metalens array camera measurement, with each metalens measurement corresponding to 60° FoV and the groundtruth image corresponding to a total of 90° FoV. Specifically, we sweep the full image from top to bottom with a step size of 135 pixels. The resolution of cropped image is 810×810. There are overlapping areas between all cropped images.

We then resize and arrange each cropped image into a 3×3 array to simulate the nanophotonic sensor capture. To this end, we first compute homographies between the 9 local image patches as measured by the real nanophotonic array camera and the ground truth compound optic camera, such as described in Sec. 5.2 of the main paper, in order to transform the ground truth image to map that of the sensor capture. The resolution of the local image patches on the sensor is 360×360. We then utilize these homography transforms to project each of the 9 simulated metalens measurements onto the appropriate local patch on the sensor

$$\hat{\mathbf{p}}^{gt}_{mn} \rightarrow \mathbf{H}_{mn}\mathbf{p}^{s}_{mn}, \tag{20}$$

where $\hat{\mathbf{p}}^{gt}_{mn}$ denotes the coordinates in the ground truth image corresponding to the FoV as captured by the (m, n)-th metalens in the array camera, $\mathbf{p}^{s}_{mn}$ denotes the sensor coordinate corresponding to the (m, n)-th metalens measurement and $\mathbf{H}_{mn}$ denotes the corresponding homography.

We then simulate the vignetting of real-world measurements. Each of the 9 images are subjected to vignetting where we model the vignetting mask as a fourth-order Butterworth filter with a linear intensity fall-off, given by

$$\mathbf{V} = \left(1 + \left(\frac{||w||^2}{f_c^2}\right)^4\right)^{-1} \tag{21}$$

where $||\cdot||^2$ denotes the squared magnitude, w is the spatial frequency and f_c is the cutoff frequency of the filter. All parameters are matched to the experimental setting. Note that we apply this filter on each individual metalens measurement only as an intensity mask to the sensor image, with a cutoff frequency that corresponds to 45° of the metalens FoV. The vignetted images are convolved with the simulated PSFs on the sensor by Eq. (7) of the main manuscript.

Finally, we corrupt the vignetted images by adding simulated sensor noise. The sensor noise added is as determined by the parameters $C_{\text{SENSOR}} = \{\sigma_g, a_p\}$ which we set to $\sigma_g = 1 \times 10^{-5}$ and $a_p = 4 \times 10^{-5}$ using the calibration method as described in Foi et al. [Foi et al. 2008]

$$\eta_{\text{SENSOR}}(\mathbf{X}) = \eta_p(\mathbf{X}, a_p) + \eta_g(\sigma_g), \tag{22}$$

where $\mathbf{X}$ is the clean image, η_p a Poissonian signal-dependent component, and η_g a Gaussian signal-independent component. The detailed information for the two function can be found in [Foi et al. 2008]. The final sensor measurement is computed as

$$\mathbf{S}_{mn} = (\mathbf{H}^{-1}_{mn}\mathbf{I} * \mathbf{V}) \otimes \mathbf{k}_{mn} + \eta_{\text{SENSOR}}, \tag{23}$$

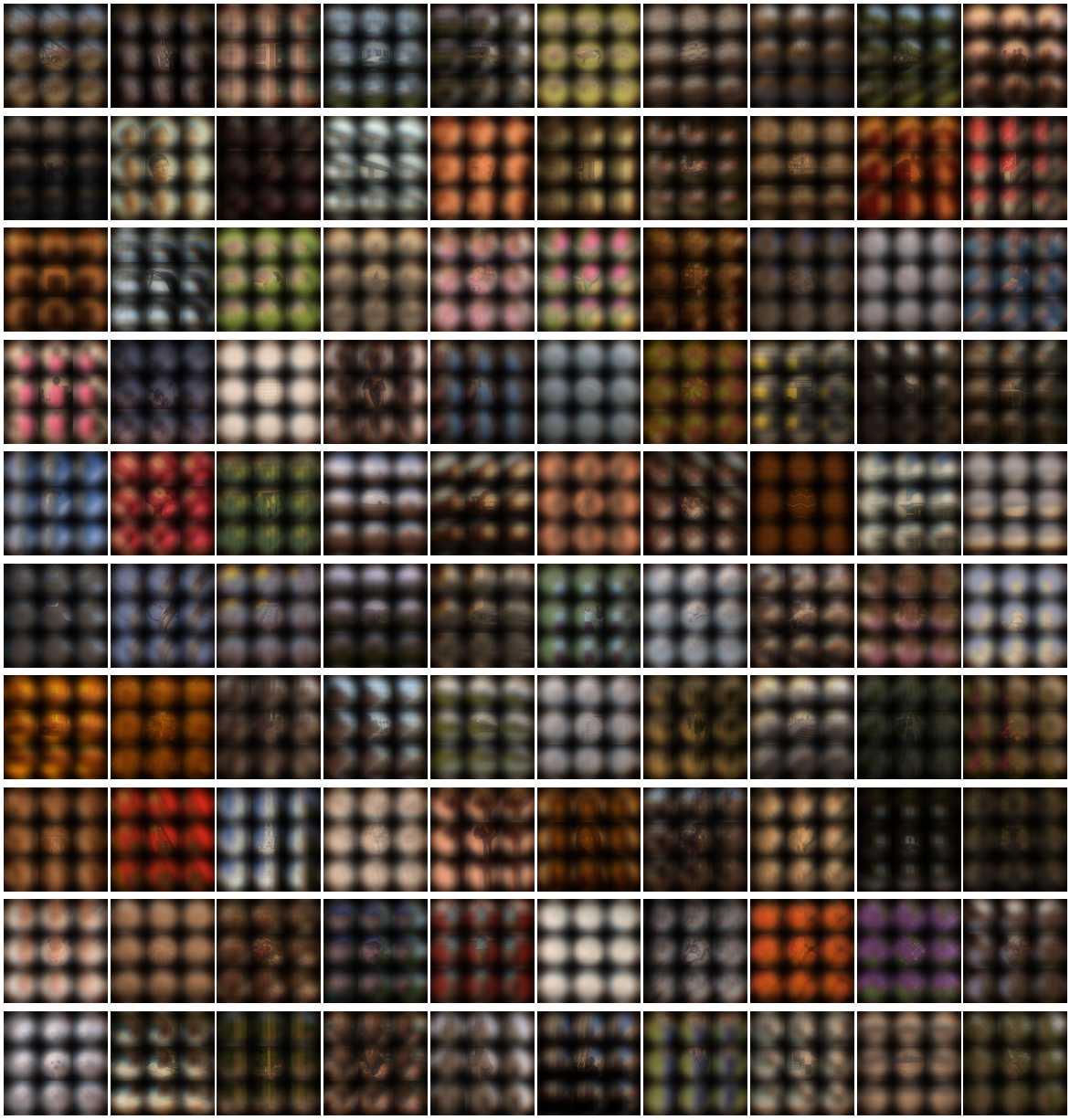

Fig. 7. Synthetic data samples. We present 100 synthesized images (10 columns, 10 rows) obtained from our simulated pipeline. Zoom in on the electronic version of this document for details.

$$\mathrm{S} = \sum_{m,n} \mathrm{S}_{mn} \quad s.t. \quad (m,n) \in \{0,1,2\}; \quad m+n \leq 2, \tag{24}$$

where $\mathbf{S}_{mn}$ denotes the (m, n)-th array measurement on the sensor, $\mathbf{S}$ is the final sensor measurement, and $\mathbf{H}_{mn}^{-1}$ and $\mathbf{k}_{mn}$ are the corresponding inverse homography and PSF, respectively.

To generate the full synthetic dataset, we randomly sample 10,000 images from a combination of ImageNet [Deng et al. 2009] and MIT 5K [Bychkovsky et al. 2011] datasets for groundtruth images. For the ImageNet dataset, our training, validation, and test splits respectively contain 8000, 1000, and 1000 images. For the MIT 5K dataset, our training, validation, and test splits contain 4000, 500, and 500 images. Figure 7 presents 100 examples of our simulated data on MIT 5K dataset.

E ADDITIONAL SYNTHETIC RESULTS

We present additional synthetic evaluation results that validate the proposed deconvolution method and camera design described in the main manuscript.

Additional Validation of Probabilistic Deconvolution Method. In Figure 8, we provide additional results in support of the probabilistic deconvolution method presented in the main manuscript. Similar to Sec. 6.1 in the main manuscript, instead of considering all 9 sub-apertures of the proposed meta-optic, we consider only the central portion. Doing so allows us to compare the proposed reconstruction method with a single PSF and image. The additional results confirm the trend from the findings in the main paper: the conventional deconvolution approaches (Wiener [1949] and Richardson-Lucy [1972]) suffer from severe reconstruction artifacts while the learned predictions from Flatnet [Khan et al. 2020] and Multi-Wiener-Net [Yanny et al. 2022] are overly smooth with fine details missing. The proposed probabilistic method recovers fine details in the reconstructions, in line with the quantitative evaluations from the main manuscript.

Validation of Thin Imager Design. In Figures 9 and 10, we provide additional results that further validate the proposed thin camera design in simulation. Similar to Sec. 6.1 in the main manuscript, we use the synthetic dataset with all 9 sub-apertures on the sensor and the full proposed reconstruction method, including the blending network. The additional supplementary results confirm the findings from the main manuscript. While FlatCam [2017] and DiffuserCam [2017] sensing allows the capture of rays from a large cone of angles, spatial and color information are entangled in PSFs with support of the entire sensor, making the recovery of high-frequency content challenging independently of the FoV. The proposed metasurface array imager is able to image fine details across almost the entire field of view. The learned reconstruction methods FlatNet [Khan et al. 2020] (for FlatCam observations) and Kingshott et al. [2022] (for DiffuserCam measurements) are improving on conventional reconstruction algorithms in both cases but cannot match the quality of the proposed camera design.

F ADDITIONAL EXPERIMENTAL RESULTS

We present additional experimental evaluation results that validate the proposed method.

Additional Validation on Images in the Wild. Figures 11 and 13 provide additional experimental results that validate the proposed camera design for in-the-wild captures. Specifically, we acquire scenes in typical indoor and outdoor scenarios. We provide sensor measurements, image reconstructions and corresponding compound lens captures captures for a variety of scenes. The proposed thin imaging system is capable of recovering the scene with accurate color reproduction. The center region of the recovered images has high image quality and preserves fine detail. The reconstructed images suffer from no apparent chromatic aberrations. We also provide sensor measurements and reconstruction using the metasurface optic proposed by Tseng et al. [Tseng et al. 2021]. This method fails to accurately reconstruct the captured in-the-wild scenes, and especially for outdoor captures, the recovered images remain hazy and do not match the quality of our results.

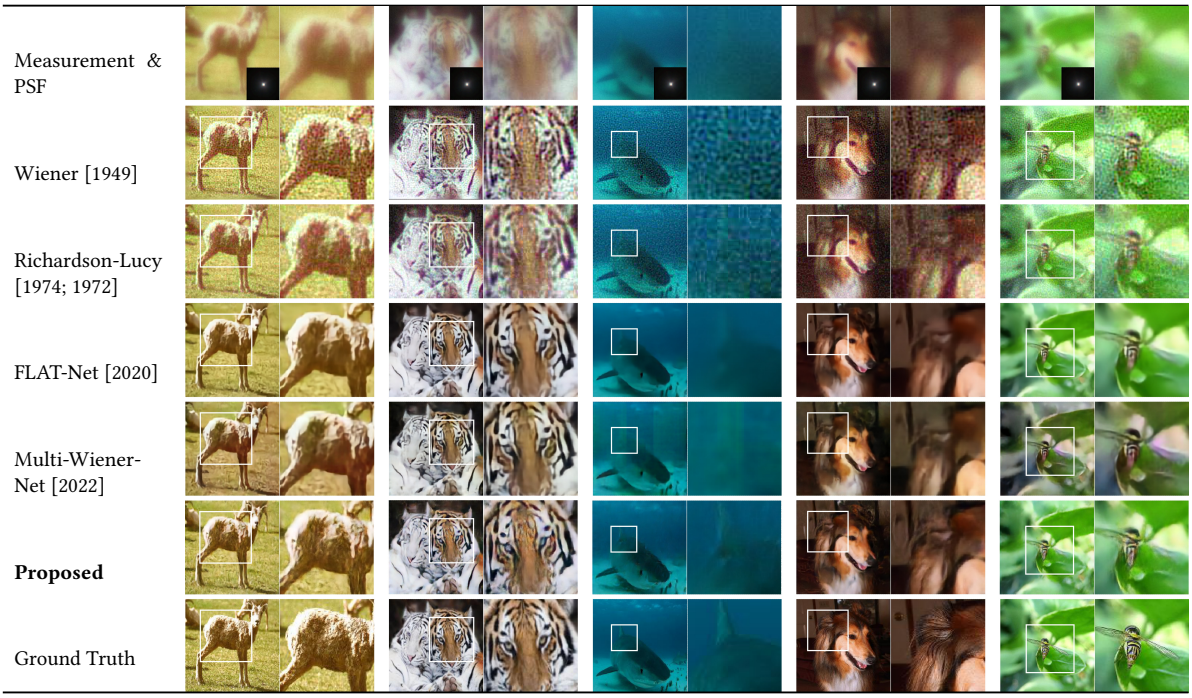

Fig. 8. Additional qualitative assessment of diffusion-based deconvolution. These additional results confirm the findings from the main manuscript. The two conventional deconvolution approaches (Wiener [Wiener et al. 1949] and Richardson-Lucy [Richardson 1972]) suffer from apparent reconstruction noise, and the predictions from Flatnet [Khan et al. 2020] and Multi-Wiener-Net [Yanny et al. 2022] are overly smooth with high-frequency details missing. The proposed probabilistic reconstruction method is capable of recovering fine details.

Additional Validation with Screen Captures. We also assess the proposed imager in a controlled setting where the scene is displayed on an LCD monitor. Specifically, we capture images on the screen with black lab surrounding, as shown in Figure 13. Different from scenes captured in the wild that require recovery over the full visible spectrum, the monitor has a sparse spectral response. Hence, it is easier for designs optimized for these sparse spectral responses to reconstruct the scene, that is both for proposed method and the method from Tseng et al. [2021]. Following Tseng et al. [2021], for this experiment, we capture images and randomly split into training set, validation set and test set. We retrain our reconstruction network and the method from Tseng et al. on the training dataset of screen images. Both methods perform well for this controlled setting, preserving color fidelity and spatial detail. This validates the challenge of generalizing imaging in lab-controlled environments to real-world outdoor/indoor scenes outside of the lab as demonstrated in the main manuscript and in the experimental findings in the previous section.

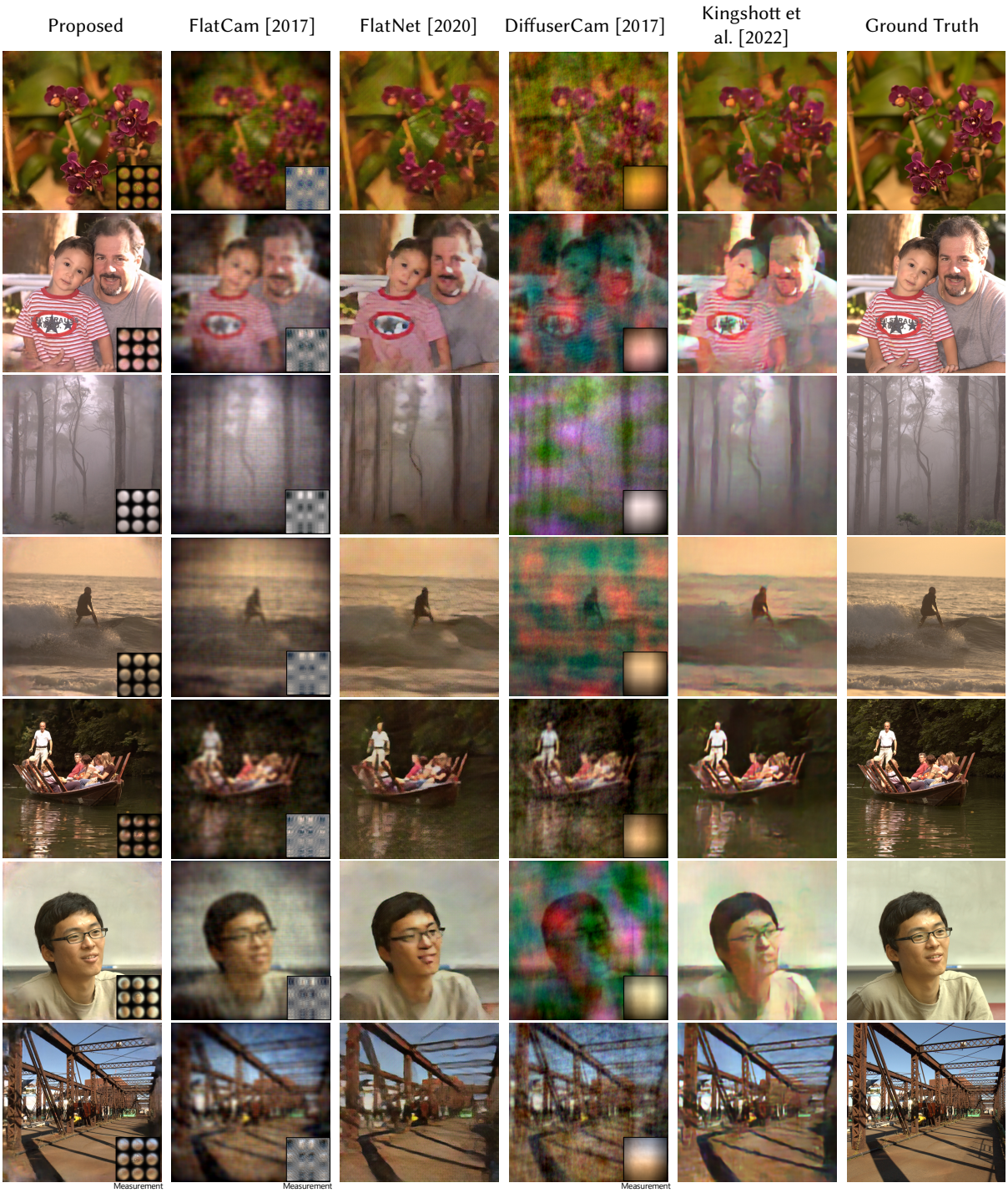

Fig. 9. Additional synthetic assessment of thin cameras. Alternative thin sensing approaches in FlatCam [2017] and DiffuserCam [Kuo et al. 2017] mix spatial and color information in PSFs with support of the entire sensor, see insets. This makes the recovery of fine detail challenging, even for learning-based methods [Khan et al. 2020; Kingshott et al. 2022].

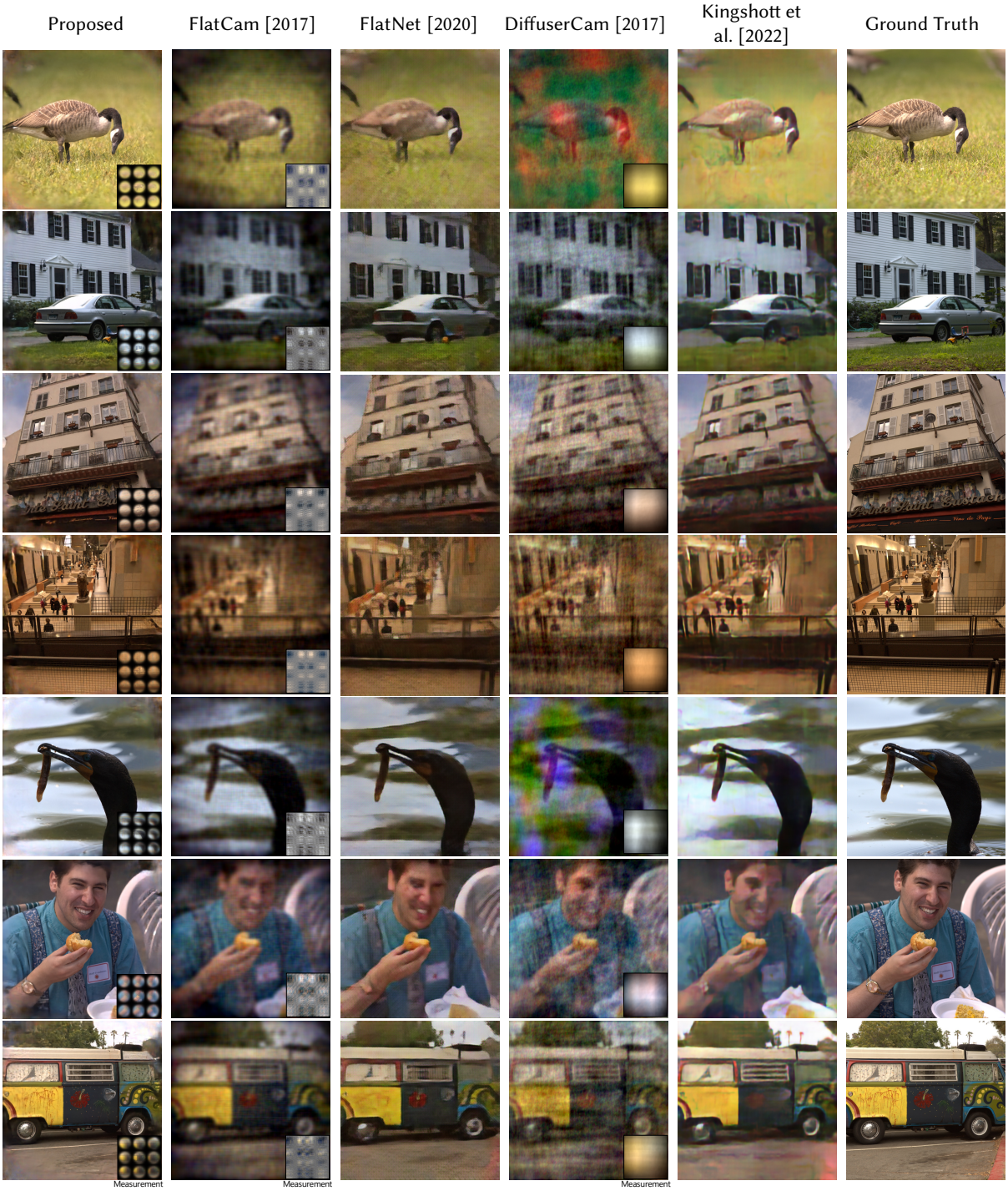

Fig. 10. Additional synthetic assessment of thin cameras. Alternative thin sensing approaches in FlatCam [2017] and DiffuserCam [Kuo et al. 2017] mix spatial and color information in PSFs with support of the entire sensor, see insets. This makes the recovery of fine detail challenging, even for learning-based methods [Khan et al. 2020; Kingshott et al. 2022].

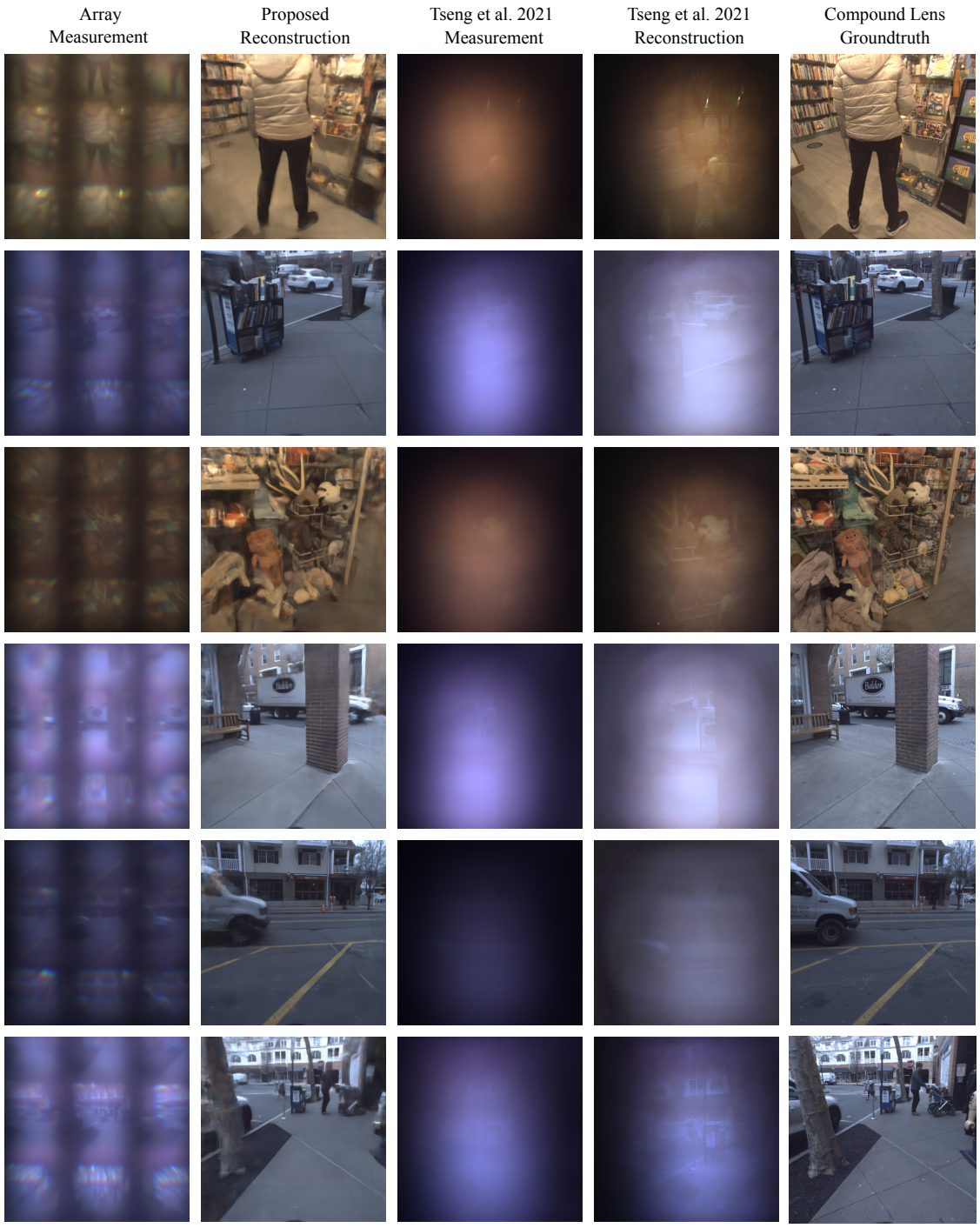

Fig. 11. Additional Real-world Assessment of Thin Cameras on In-the-wild Captures. The proposed nanophotonic array optic with the probabilistic deconvolution method reconstructs the underlying latent image robustly in broadband lit environments, outperforming Tseng et al. [2021] especially in outdoor scenes.

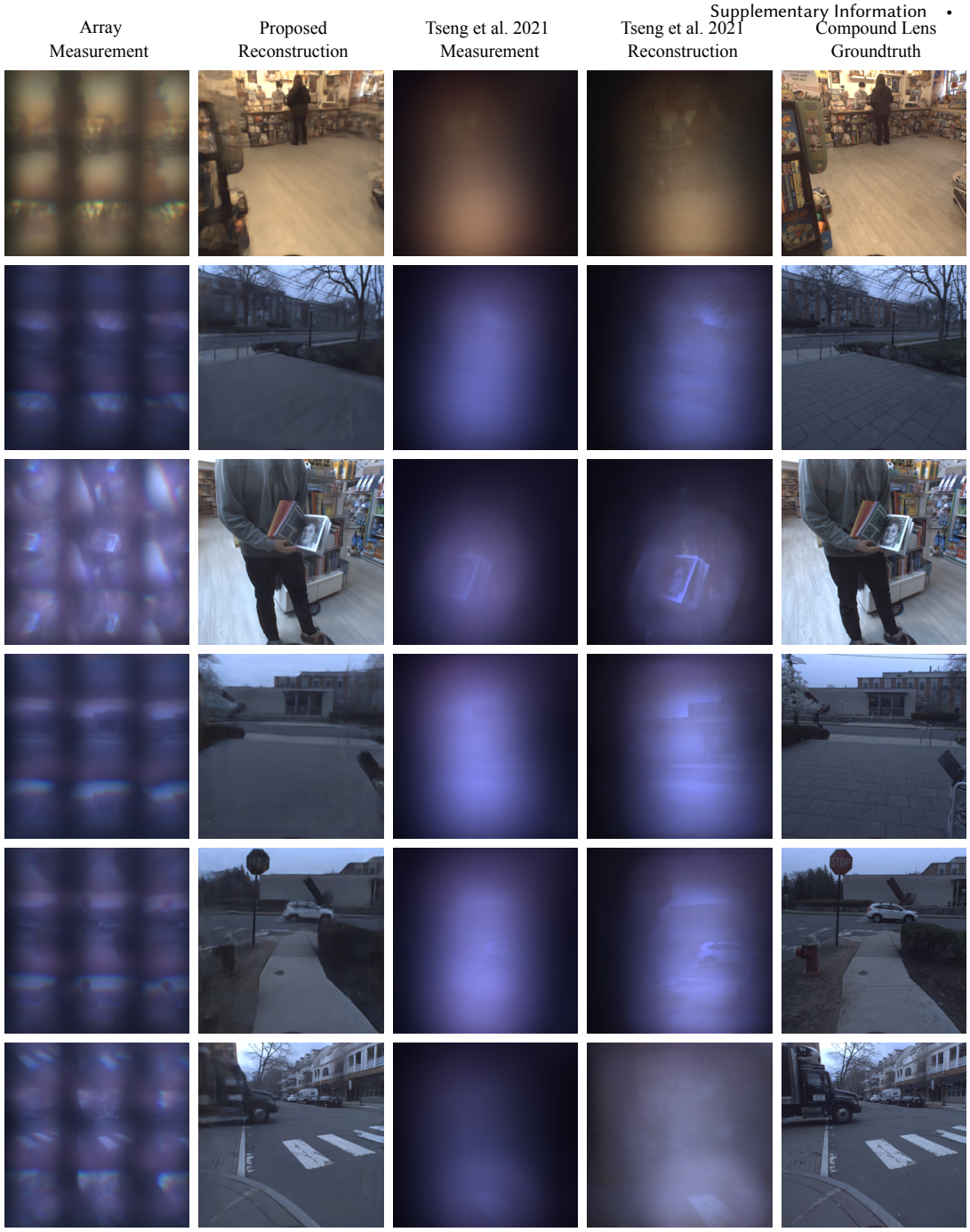

Fig. 12. Additional real-world assessment of thin cameras on in-the-wild captures. The proposed nanophotonic array optic with the probabilistic deconvolution method reconstructs the underlying latent image robustly in broadband lit environments, outperforming Tseng et al. [2021] especially in outdoor scenes.

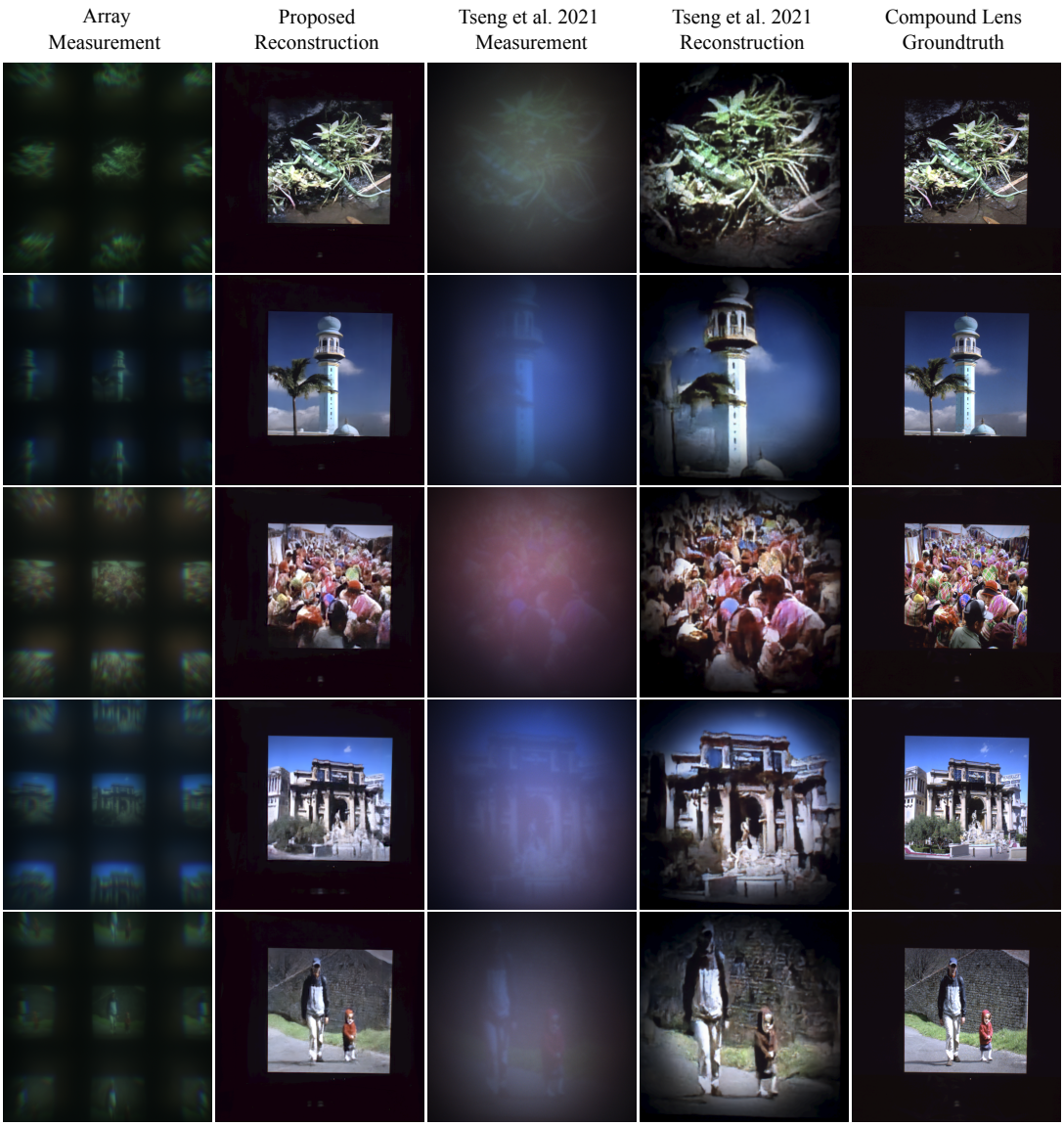

Fig. 13. Additional experimental assessment for narrow-band screen captures. The proposed design and Tseng et al. [2021] both perform well in this controlled setting. The reconstructions from Tseng et al. were measured on a sensor of smaller size compared to the compound lens ground truth and our thin metalens camera (see Section 5 of the main manuscript).

REFERENCES

M. Salman Asif, Ali Ayremlou, Aswin C. Sankaranarayanan, Ashok Veeraraghavan, and Richard G. Baraniuk. 2017. FlatCam: Thin, Lensless Cameras Using Coded Aperture and Computation. *IEEE Transactions on Computational Imaging* 3, 3 (2017), 384–397.

Vladimir Bychkovsky, Sylvain Paris, Eric Chan, and Frédo Durand. 2011. Learning Photographic Global Tonal Adjustment with a Database of Input / Output Image Pairs. In *The Twenty-Fourth IEEE Conference on Computer Vision and Pattern Recognition*.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

A. Foi, Mejdi Trimeche, V. Katkovnik, and K. Egiazarian. 2008. Practical Poissonian-Gaussian Noise Modeling and Fitting for Single-Image Raw-Data. *IEEE Transactions on Image Processing* 17 (2008), 1737–1754.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems* 33 (2020), 6840–6851.

Salman Siddique Khan, Varun Sundar, Vivek Boominathan, Ashok Veeraraghavan, and Kaushik Mitra. 2020. Flatnet: Towards photorealistic scene reconstruction from lensless measurements. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020).

Oliver Kingshott, Nick Antipa, Emrah Bostan, and Kaan Akşit. 2022. Unrolled primal-dual networks for lensless cameras. *Opt. Express* 30, 26 (Dec 2022), 46324–46335. https://doi.org/10.1364/OE.475521

Grace Kuo, Nick Antipa, Ren Ng, and Laura Waller. 2017. DiffuserCam: diffuser-based lensless cameras. In *Computational Optical Sensing and Imaging*. Optical Society of America, CTu3B–2.

Leon B Lucy. 1974. An iterative technique for the rectification of observed distributions. *The astronomical journal* 79 (1974), 745.

Alexander Quinn Nichol and Prafulla Dhariwal. 2021. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*. PMLR, 8162–8171.

William Hadley Richardson. 1972. Bayesian-based iterative method of image restoration. *JoSA* 62, 1 (1972), 55–59.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Proc. Medical Image Computing and Computer-Assisted Intervention (MICCAI)*.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*. PMLR, 2256–2265.

Jiaming Song, Chenlin Meng, and Stefano Ermon. 2021. Denoising Diffusion Implicit Models. In *International Conference on Learning Representations*. https://openreview.net/forum?id=St1giarCHLP

Yang Song and Stefano Ermon. 2019. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems* 32 (2019).

Ethan Tseng, Shane Colburn, James Whitehead, Luocheng Huang, Seung-Hwan Baek, Arka Majumdar, and Felix Heide. 2021. Neural Nano-Optics for High-quality Thin Lens Imaging. *Nature Communications* (December 2021).

Norbert Wiener, Norbert Wiener, Cyberneticist Mathematician, Norbert Wiener, Norbert Wiener, and Cybernéticien Mathématicien. 1949. *Extrapolation, interpolation, and smoothing of stationary time series: with engineering applications*. Vol. 113. MIT press Cambridge, MA.

Kyrollos Yanny, Kristina Monakhova, Richard W Shuai, and Laura Waller. 2022. Deep learning for fast spatially varying deconvolution. *Optica* 9, 1 (2022), 96–99.