Supplementary Information

Compact snapshot meta-optic HDR camera for handheld, edge-deployable imaging

Contents

1 Supplementary Note 1: Metalens Array Design and Fabrication 3
- 1.1 Hardware architecture . . . 3
- 1.2 Metalens phase profile design and optimization . . . 3
  - 1.2.1 EM proxy phase function . . . 4
- 1.3 Loss functions . . . 4
- 1.4 Optimization . . . 5
- 1.5 Compute constraints . . . 6
- 1.6 Fabrication and assembly . . . 6
- 1.7 Meta-atom design . . . 7

2 Supplementary Note 2: Network Architecture and Training Details 7
- 2.1 Problem formulation and architecture overview . . . 7
- 2.2 Stage 1a: Per-lens DDF restoration front-end . . . 8
- 2.3 Stage 1b: Feature pyramid and FPN top-down denoising . . . 9
- 2.4 Stage 1c: PCD deformable align-to-reference . . . 9
- 2.5 Stage 2: Adaptive exposure/confidence fusion . . . 10
- 2.6 Stage 3: Static residual reconstruction . . . 12
- 2.7 Field-coordinate conditioning (CoordConv) . . . 12
- 2.8 Decoupled Dynamic Filter (DDF) . . . 12
- 2.9 Parameter and compute budget . . . 14
- 2.10 Loss function . . . 14
- 2.11 Training configuration . . . 15
- 2.12 Edge deployment . . . 15

3 Supplementary Note 3: Real-Scene Training Data Pipeline 15
- 3.1 Controlled data acquisition with OLED display . . . 15
  - 3.1.1 Calibration . . . 16
  - 3.1.2 Paired data generation . . . 17
  - 3.1.3 In-domain validation . . . 17
- 3.2 Domain gap between lab and real scenes . . . 18
- 3.3 Challenges in cross-device registration . . . 19
- 3.4 Pseudo-reference construction via reconstruction-domain alignment . . . 20
  - 3.4.1 Core idea: alignment in the reconstruction domain . . . 20
  - 3.4.2 Alignment method comparison . . . 20
  - 3.4.3 Results . . . 20
  - 3.4.4 Analysis of alignment method behavior . . . 20

4 Supplementary Note 4: Extended Real-World Results 22
- 4.1 Comparison with acquisition modes and existing HDR methods . . . 22
- 4.2 Dynamic scenes: snapshot acquisition versus temporal burst . . . 23

5 Supplementary Note 5: Extended Ablation Analysis **24**
5.1 Evaluation protocol 24
5.2 Master ablation table 25
5.3 Alignment: removing it washes out color and fine detail 27
5.4 Field conditioning and the field-varying PSF 27

1 Supplementary Note 1: Metalens Array Design and Fabrication

1.1 Hardware architecture

The aperture array is positioned between the lens array and the sensor (Fig. S1). The sensor cover glass is polished until the aperture array lies at a distance $b = a - c$ above the sensor surface, yielding minimal inter-channel crosstalk x given by:

$$x = \max\left(\frac{bd - cy}{2c}, 0\right) \tag{1}$$

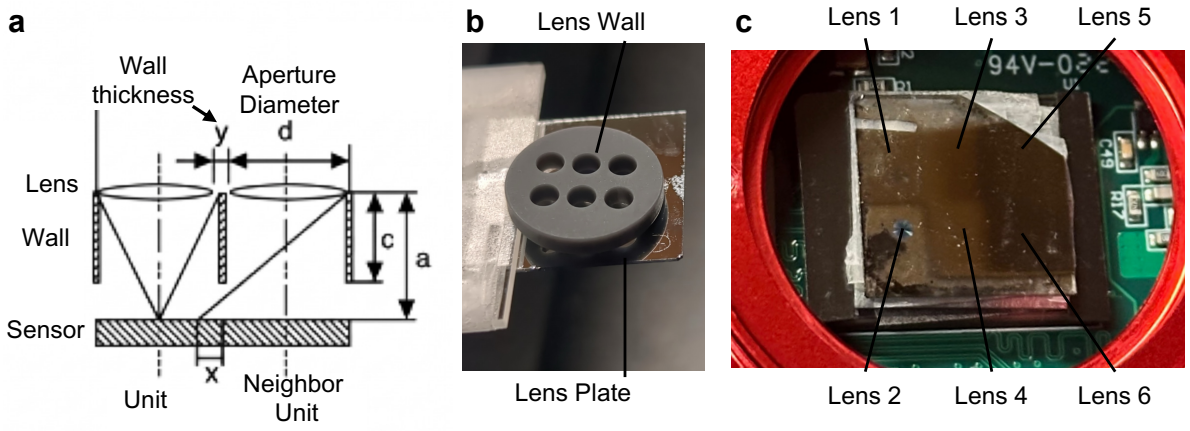


Figure S1: **Hardware architecture of the metalens array camera. a**, Schematic cross-section of the lens-frame assembly. Walls between adjacent lens units limit the cross-talk signal from a neighboring lens to a region of width x on the sensor (PD array); a is the distance of the lens array from the sensor, c the height of the aperture walls (the thickness of the aperture array), y the wall thickness, and d the aperture diameter. **b**, Photograph of the aperture array, showing the lens walls and the lens plate. **c**, Photograph of the metalens array mounted above the sensor, with lens positions labeled; lenses are numbered 1–6 in order of decreasing transmittance (AP1–AP6 in main-text Fig. 2).

The system employs a Sony IMX183 1-inch sensor for image capture. In our system, the aperture diameter $d = 1.5\,\text{mm}$ (1 mm for lens and 0.5 mm for lens placement padding), lens-to-sensor distance $a = 2.5\,\text{mm}$, wall height $c = 1.5\,\text{mm}$ and its thickness $y = 0.5\,\text{mm}$, giving a cross-talk signal width of x ≈ 0.25 mm on both sides of the aperture walls. The field of view of the resulting lens-camera assembly is obtained as $2\arctan(d/2a) \approx 33°$.

1.2 Metalens phase profile design and optimization

A metalens focuses light with an array of subwavelength scatterers rather than a curved refractive surface: each scatterer imparts a prescribed phase delay, and the desired optical function emerges from the controlled interference of the transmitted wavefront. Three constraints prevent the ideal phase function of an equivalent refractive element from being realized directly. First, a finite meta-atom library provides only a discrete set of achievable phase values. Second, the phase and amplitude response of each meta-atom varies with wavelength and with angle of incidence, so a profile that is correct at one design wavelength degrades across the visible band and toward the field edge. Third, lithographic resolution and aspect-ratio limits restrict the attainable geometries.

The phase profile must therefore be treated as a constrained optimization target rather than an analytic prescription.

We address this with a differentiable design pipeline. Free-space propagation from the metalens plane to the image plane is implemented with differentiable operators, so the phase parameters can be updated by gradient descent against optical figures of merit evaluated directly on the simulated field.

For the array reported here the parameterization has two levels: a rotationally symmetric aspheric base profile for the central lens (main-text Methods, Eq. 1), and a freeform optimized radial phase profile, which must retain focusing efficiency at multiple depths of field. The figures of merit combine spot size, modulation transfer, and depth-of-focus terms (Section 1.3), rather than a single focal-spot criterion.

1.2.1 EM proxy phase function

The optimization above produces a continuous phase profile, but the fabricated device realizes that profile through a discrete set of meta-atom geometries. The mapping between the two directions is not analytic: the transmitted phase of a meta-atom depends nonlinearly on its geometry, on wavelength, and on angle of incidence, and evaluating it rigorously requires a full-wave electromagnetic solve. Placing such a solve inside the optimization loop is computationally prohibitive, since the loop evaluates the field for every wavelength, incidence angle, and defocus sample at every iteration.

1.3 Loss functions

To characterize the performance of the differentiable phase function $\Phi(r)$ we employ several figures of merit (FOM) used as loss functions L_{FOM_i} during optimization,

$$\underset{\xi}{\operatorname{argmin}} \sum_{i\in\{FOM\}} \sum_{\lambda\in\Lambda} \sum_{\theta\in\Theta} \sum_{|r|<R} L_i(\xi, \lambda, \theta; \zeta). \tag{2}$$

The lens parameters ξ are optimized over a discrete set of wavelengths λ. The parameters ζ are fixed during the optimization and correspond to system parameters like the back focal length, lens pixel pitch, sensor pixel pitch, and sensor magnification. The loss function is a collection of various merit functions characterizing the imaging performance of the system as listed below.

Focusing loss: The point spread function (PSF), i.e., the response of the lens to a parallel incident beam, measures its focusing efficiency. The PSF is obtained by propagation of the light field from metalens surface to the sensor through the Angular Spectrum Method (ASM) and its shifted variant [1] for beams incident at a non-zero angle (θ). By minimizing the focusing loss (or maximizing the focal spot intensity), we optimize for the most common case of imaging through the lens,

$$L_{PSF} = -PSF_{\lambda,\theta}(r) \tag{3}$$

since a sharper PSF leads to an overall sharper sensor image with a higher signal to noise ratio. The sum is taken over a circular aperture of radius $r_f(\theta) < R$, different for different angles of incidence. We determine $r_f(\theta)$ by finding the circular aperture that contains at least 80% of the initial point spread function energy.

MTF loss: The modulation transfer function (MTF) measures the number of resolvable line pairs per mm (lp/mm) and is typically used to balance contrast versus maximum achievable resolution. It is obtained from the Fourier transform (FT) of the PSF i.e. $MTF(\cdot) = \texttt{FT}(PSF(\cdot))$. The loss is evaluated over the entire aperture over the broadband range[1],

$$L_{MTF} = \texttt{ReLU}[C - MTF_{\lambda,\theta}(r)] \tag{4}$$

where $C = 0.005$ is the minimum threshold contrast enforced at each spatial location r. The `ReLU` makes this a one-sided hinge penalty: it is inactive wherever the MTF already clears the threshold, so the term constrains the worst-case resolvable contrast rather than pushing the MTF up uniformly.

EDOF loss: To produce a continuous depth of focus along the optical axis we minimize the focal spot radius at multiple planes around the focus plane (F) i.e. $F + \delta$ for $|\delta| < D$ as,

$$L_{EDOF} = \sum_{\delta \in [-D,D]} w_{\lambda,\delta}(F + \delta) \tag{5}$$

where $w_{\lambda,\delta}(F + \delta)$ is the focal spot radius for depth of focus $F + \delta$ obtained by propagating the field from the lens to the sensor plane via a 1D Rayleigh–Sommerfeld integral [2], and $D = 0.5\,\text{mm}$ sets the extended depth-of-focus range about the nominal focal plane.

1.4 Optimization

The phase coefficients are optimized by gradient descent through the differentiable propagation model, with the field at the image plane re-evaluated at every iteration over the discrete design set of wavelengths Λ spanning 410–690 nm (intervals of 10 nm) and angle-of-incidence offsets $\Theta = \{0°, \pm 5°, \pm 7.5°\}$ taken about the central angle (0°), i.e., the phase profile is optimized over a 15° full field of view. Because the three merit functions differ by orders of magnitude in scale, each L_{FOM_i} is normalized by its value at initialization, so that Eq. 2 is minimized in the form

$$\mathcal{L}_{\text{total}} = \sum_{i \in \{FOM\}} \frac{1}{|L_i^{(0)}|} \sum_{\lambda \in \Lambda} \sum_{\theta \in \Theta} L_i(\xi, \lambda, \theta; \zeta), \tag{6}$$

with no further relative weighting between terms. Fabrication imperfections can lead to the lens deviating from this optimized phase; to mitigate this we inject noise during optimization by simulating the phase duty cycles (d_r) with random Gaussian fluctuations [3] centered on the desired values, $\tilde{d}_r \sim \mathcal{N}(d_r, \sigma^2)$ with $\sigma = 0.1 \min(\{d_r\})$. We use the Adam optimizer on the phase, with the learning rate decayed cosine-wise over $T_{\max} = 200$ epochs.

The conic coefficients ρ and κ of the aspheric phase are not optimized by gradient descent but obtained by a grid search around an initial estimate (ρ_0, κ_0) fitted to a quadratic phase at $\lambda_{nominal}$. The search penalizes both the chromatic spread of the focal spot and its mean radius,

$$FOM_{\{\rho,\kappa\}} = Var\big(w_{\lambda,\delta=0}(F)\big) + \omega \sum_{\lambda} \frac{w_{\lambda,\delta=0}(F)}{|\Lambda|}, \tag{7}$$

where ω is a weighting factor balancing achromaticity (measured as the variance Var over the set Λ) against absolute spot size and $w_{\lambda,\delta=0}(F)$ gives the 1-dimensional spot radius at the focal

[1] For the radially symmetric case, the MTF is the same along each direction (θ) and hence can be reduced to a 1D function.

plane ($\delta = 0$) computed efficiently via the Rayleigh–Sommerfeld integral operator. The resulting distribution of optimized meta-atom phases across the array is shown in Fig. S2.

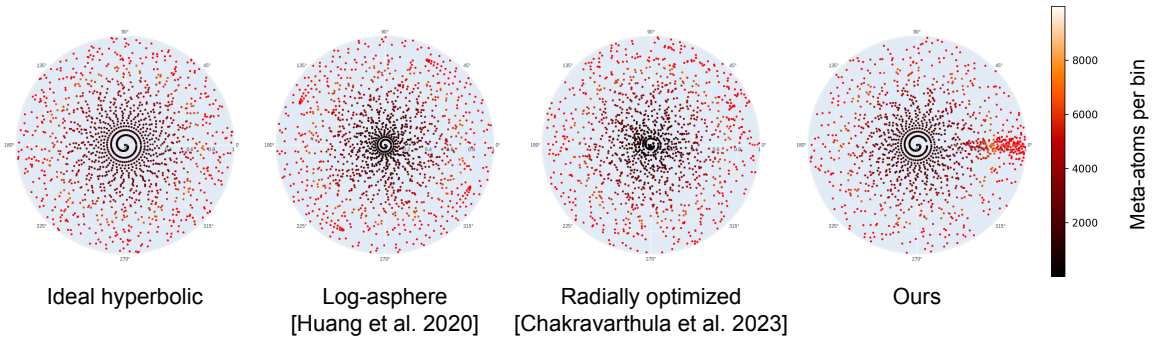


Figure S2: **Joint distribution of realized meta-atom phase and radial position.** Each panel is a 2D histogram over the lens aperture: radius gives the normalized radial position of a meta-atom, angle gives its realized phase modulo 2π, and color denotes the number of meta-atoms falling in each bin. Panels correspond to four phase parameterizations: the ideal hyperbolic profile, the log-asphere of Huang et al. [4], the radially optimized phase of Chakravarthula et al. [5], and ours. All four share the dense central spiral produced by a slowly varying paraxial phase, which unwinds into azimuthal spread at larger radii as the phase gradient steepens and wrapping accelerates. Our design tracks the ideal hyperbolic profile most closely near the optical axis, a consequence of initializing the conic coefficients (ρ_0, κ_0) by fitting to a quadratic phase at $\lambda_{nominal}$ (Eq. 7), while departing from it at the aperture edge comparably to the two baselines. Distinct from all three, our design shows a pronounced concentration of near-zero-phase atoms extending across the full radial range (the streak at 0°), which we attribute to truncation of the meta-atom library: target phases outside the realizable coverage of the proxy phase function (Section 1.2.1) are clipped to the nearest fabricable duty cycle, producing a pile-up at the library boundary.

1.5 Compute constraints

The computation of L_{EDOF} requires a large number of forward simulations even for relatively small depth of focus $D = 0.5\,\text{mm}$ since it needs to be sampled discretely around the focal plane with a very fine sampling interval. This is to avoid overfitting the phase to discrete bins along the optical axis. Optimizing over the entire range of wavelengths and focus points simultaneously is time consuming and quickly runs out of available memory during optimization. Hence, we optimize the overall loss functions in alternating phases: first along the depth of focus axis (L^t_{EDOF}) using 1D Rayleigh–Sommerfeld operator and then along spatial 2D axes (L^t_{PSF} and L^t_{MTF}) using Angular Spectrum Method (ASM). We iterate between 1D and 2D every T'th iteration,

$$L^t = \begin{cases} L^t_{EDOF} & \text{if } t = T, 2T, ... \\ L^t_{PSF} + L^t_{MTF} & \text{else} \end{cases} \tag{8}$$

1.6 Fabrication and assembly

The metalens array is fabricated in an ISO Class 5–7 cleanroom. Quartz wafers (500 µm thick) are cleaned in successive ultrasonic baths of acetone, isopropanol (IPA), and DI water, followed by a short oxygen descum in a barrel etcher. An 800 nm thick silicon nitride film is then deposited by plasma-enhanced chemical vapor deposition (PECVD, SPTS chamber). The wafer is diced into 1.5 cm square pieces and cleaned again with the same ultrasonic and barrel-etch sequence.

A 400 nm layer of positive resist (ZEP 520A) is spin-coated onto the sample, followed by a conductive polymer overcoat (DisCharge H_2O) to mitigate charging during patterning. The resist is exposed with a 100 keV electron beam (JEOL JBX6300FS) at a dose of 275 µC cm^{-2}. After exposure the conductive polymer is removed in a short IPA bath and the resist is developed in amyl acetate for 120 s at room temperature. Residual resist is cleared in a brief barrel-etch step, after which a 75 nm alumina hard mask is deposited by electron-beam evaporation and lifted off overnight in NMP at 100 °C. The silicon nitride layer is then etched in a fluorine-based chemistry in an inductively coupled reactive-ion etcher (Oxford PlasmaLab System 100).

The transmittance ladder is defined after etching by sequential masked depositions of Au/Pd on the back side of the chip. Before each deposition one additional metalens is masked off, so that every film adds to all still-exposed apertures. This cumulative scheme produces six apertures with monotonically increasing total Au/Pd thickness, and hence a monotonically decreasing transmittance ladder, from a single film thickness per deposition step. Finally, the chip is integrated with a 3D-printed baffle and mounted on the sensor cover glass with double-sided tape applied at the edge of the cover glass.

1.7 Meta-atom design

Each meta-atom is a subwavelength silicon nitride pillar on the quartz substrate. Its lateral dimensions set the local optical path length and therefore the phase delay imparted to the transmitted field, while the film thickness is fixed at 800 nm by the PECVD deposition (Section 1.6). Sweeping the pillar dimension over the fabricable range yields the meta-atom library from which the optimized phase profile is discretized, using the bidirectional proxy described above.

2 Supplementary Note 2: Network Architecture and Training Details

2.1 Problem formulation and architecture overview

A conventional high-dynamic-range (HDR) pipeline merges a temporal exposure bracket; our metalens system collapses that bracket into a *single snapshot*: a planar 2×3 metalens (lenslet) array images the scene through six independent apertures onto one sensor read-out, producing six RGB sub-images with no temporal gap and no moving parts. The sub-images are therefore not clean, co-registered exposures but physically distinct measurements that the reconstruction network must jointly invert.

Let $L(\mathbf{u})$ be the latent display-referred radiance to recover, sampled on pixel grid $\mathbf{u}$; for lab data it is the OLED-displayed ground truth, which supervises the controlled ablations of Supplementary Note 5 (1101 scenes; Table S2). The i-th sub-image I_i is formed as

$$I_i(\mathbf{u}) = \Gamma\Big(g_i \cdot \big[\, k_i(\mathbf{u}) * L \,\big]\big(\mathcal{W}_i(\mathbf{u}; d) \big) \Big) + n_i, \qquad i = 1, \ldots, 6, \tag{9}$$

which couples four degradations:

1. **Per-lens exposure ladder g_i with saturating response Γ.** The apertures span a 14× transmission range, so bright regions clip in the high-transmission lenses while dark regions are buried in read noise in the low-transmission lenses; well-exposedness is therefore *spatially varying* and migrates from frame to frame.

2. **Bounded inter-aperture parallax $\mathcal{W}_i(\mathbf{u}; d)$.** The apertures sit at slightly different positions on a short shared baseline, so a scene point shifts across the sub-images by a depth-dependent disparity *bounded* by the short baseline, and fabrication and mounting tolerances add fixed geometric offsets between the sub-images; naive channel stacking therefore misaligns and doubles edges (Fig. S3).

3. **Field-varying PSF $k_i(\mathbf{u})$.** A metalens is sharp on-axis, but its point-spread function broadens and aberrates toward the field edge, so the blur kernel varies with aperture and field position, unlike the spatially uniform blur that ordinary deblurring assumes.

4. **Read noise and quantization n_i.** The low-transmission apertures collect the least light, so their shadows sit near the sensor read-noise floor while quantization truncates the darkest recoverable stops.

A standard merge would first estimate dense correspondence (optical flow such as RAFT [6], or parametric ECC [7]) and then warp and blend. On this data the first step is *unusable*: the frames differ simultaneously in exposure (14×, with clipping), field-varying blur and noise floor, so the photometric constancy that flow and ECC assume is violated at every pixel, most severely in saturated highlights, dark shadows and aberrated edges. Alignment must therefore be *learned*, in a feature space the network shapes to be exposure- and blur-invariant.

The capture is *field-consistent by construction*: the six sub-images are always read in a fixed order and center-cropped, with *no flips or rotations* in augmentation, so aperture 3 (AP3) is always the geometric reference and every input position always carries the same aperture, transmission, field region and PSF. DAFNet (DDF-Align-Fuse; main-text Fig. 3a) demultiplexes the snapshot into the six sub-images in this order and inverts Eq. (9) in three stages, aligning before fusing. **ALIGN** addresses the field-varying PSF, read noise and parallax in turn: each input position owns a dedicated decoupled dynamic filter (DDF) restoration stack as its PSF prior (Section 2.2), with normalized field coordinates $[r, r^2]$ appended to the front-end input to supply absolute field position (Section 2.7); a feature pyramid with an FPN top-down path denoises the features (Section 2.3); and learned pyramid, cascading and deformable (PCD) alignment [8] registers each neighbor to the reference in feature space, estimating offsets coarse-to-fine (Section 2.4). **FUSE** addresses the exposure ladder and read noise by selecting, rather than averaging, the better-exposed apertures per pixel through a softmax over frames driven by raw sub-aperture well-exposedness (Section 2.5). **RECON** is a static residual U-Net whose output is added to the raw reference frame (Section 2.6). Content-adaptive capacity thus sits entirely in the small front-end, where the field-varying optics act, while the reconstruction backbone, which holds most of the parameters, uses plain static convolution (Section 2.9; Table S1). Input spatial dimensions must be divisible by 16 because the reconstruction U-Net downsamples four times.

2.2 Stage 1a: Per-lens DDF restoration front-end

Each of the six fixed slots has its own restoration stack of three DDF residual blocks (width 28), applied before any cross-frame interaction; these stacks are the network's explicit per-aperture PSF prior. The front-end first lifts every sub-image to feature width 28 with a shared weight-normalized first convolution (3→28, or 5→28 once the $[r, r^2]$ field coordinates of Section 2.7 are concatenated; 3×3), producing $L1 \in (B{\cdot}N, 28, H, W)$. After a leaky-ReLU activation, the features of slot i pass through the i-th lens-specific stack, and the six restored feature maps are restacked to form the restored $L1$. Each per-lens stack has ∼0.009 M parameters (≈0.054 M for all six). Because the PSF is field- and aperture-varying, a single shared extractor must average over six optical regimes,

whereas a per-slot extractor specializes to its aperture, with milder filtering for sharper apertures and stronger filtering for more aberrated ones. Field consistency (fixed slot order) makes this specialization well-posed.

2.3 Stage 1b: Feature pyramid and FPN top-down denoising

From the restored $L1$, three stride-2 convolutions build a four-level pyramid, doubling channels each step (28→56→112→224): $L2$ at $H/2$, $L3$ at $H/4$, $L4$ at $H/8$. An FPN-style top-down path [9] then re-injects coarse context: a 1×1 projection maps $L4$ back to width 28, and lateral 1×1 projections plus upsample-add plus 3×3 smoothing convolutions propagate it down to $L1$, bringing all levels to width 28. The pyramid convolutions dominate this sub-block (pyramid levels $L2$ to $L4$: $\approx 0.014/0.057/0.226$ M); the entire FPN head is only $\sim$0.033 M. The pyramid's top-down FPN mainly supplies *denoised* multi-scale features so that fine-level offset estimation stays robust to cross-exposure noise. The deformable warp is applied only at the finest scale, guided by the coarse-to-fine offset estimation; the coarser warp modules are unused by the trained model and are removed at inference at no cost (Section 2.4).

2.4 Stage 1c: PCD deformable align-to-reference

Alignment is solved in feature space rather than by explicit flow, following our previous work on reconstruction from aberrated metalens bursts [10]: a pyramid, cascading, deformable (PCD) alignment module [8] registers each of the five neighbor frames to the reference using learned deformable offsets [11]. Neighbor and reference features are concatenated, passed through offset convolutions, and used to drive a separable deformable convolution that warps the neighbor toward the reference. Offsets are estimated coarse-to-fine: a coarser-level offset feature is bilinearly upsampled, scaled $\times 2$, and fused into the finer estimate, so the *offset estimator* is multi-scale. The reference aperture (AP3) is aligned to itself (near-identity warp); the deformable modules additionally produce a **learned modulation mask** that softly down-weights unreliable matches, so alignment degrades gracefully rather than hallucinating.

Although the offset convolutions at all four pyramid levels are trained through this coarse-to-fine cascade, the deformable *warp* is applied only at the finest scale, because the reconstruction consumes only the finest aligned feature. The coarse-level deformable warp modules therefore receive no gradient and their warp-offset generators remain at zero initialization (warp-offset magnitude $< 10^{-3}$ px at levels 2–4), producing identity (discarded) outputs. Removing these three coarse warp modules is *exactly* output-preserving (max pixel difference 0.0 over validation) while eliminating 0.29 M parameters (over half of the 0.512 M PCD module, the single largest front-end block) and cutting end-to-end inference latency by **17.6%** on a workstation GPU (244.8 ms → 201.7 ms at 512^2, batch 1) and ≈7–9% on the Jetson Orin Nano; the multi-scale offset *estimation* cascade is retained.

The per-lens DDF restoration and its $[r, r^2]$ field conditioning act *before* alignment: together they denoise and deblur each severely degraded, field-varying sub-image, so that alignment and reconstruction operate on clean, position-corrected features rather than on the raw metalens output. This order matters because a classical aligner on raw lenslet sub-images must rely on coarse-scale pooling to raise the signal-to-noise ratio of the features it matches, whereas an explicit restoration front-end supplies that signal directly and at full resolution. Two observations support the design: the converged finest-scale offsets are smooth and strongly scene-structured rather than noise-like (Fig. S4), and a single deformable warp on the restored features already reaches competitive quality.

Learned feature-space deformable alignment avoids the photometric- and gradient-constancy assumptions of flow and ECC (Section 2.1) by matching restored, denoised, exposure-normalized

features, and the modulation mask limits the influence of residual mismatches. Alignment is therefore ghost-*bounded* rather than ghost-*free* where exact correspondence is physically unrecoverable.

Fig. S3 illustrates the effect of alignment on a real capture. Corresponding crops of the reference and a neighboring aperture are displaced relative to each other (panel b). When the same trained model is run with its deformable offsets set to zero at inference, the fluorescent tube is distorted and its edge ghosted; with alignment enabled, the edge is reconstructed as a single, straight structure. In the planar OLED ablation, where the model is retrained with the offsets fixed to zero, fusing unregistered sub-images instead washes out color and contrast and softens detail (Section 5.3).

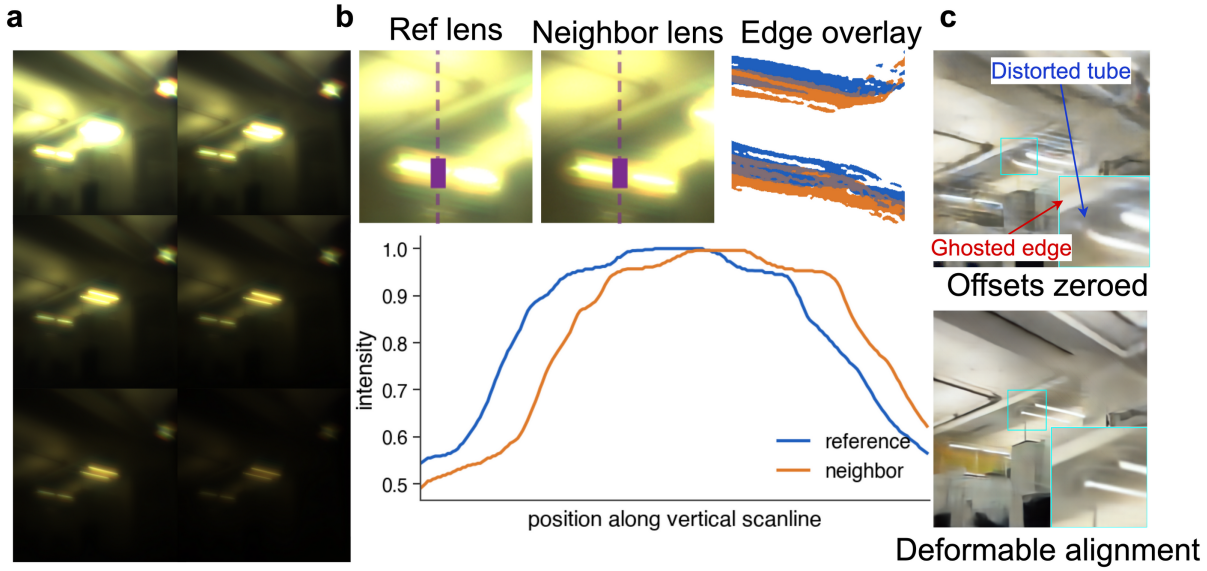


Figure S3: **Inter-aperture parallax in a real capture and its compensation by deformable alignment.** **a**, Raw six-aperture capture of an indoor ceiling scene. **b**, Crops of the reference aperture (AP3) and a neighboring aperture around the same fluorescent tube; the dashed line marks the vertical scanline and the purple bar the sampled segment. The overlaid edge maps (blue, reference; orange, neighbor; upper and lower bands trace the top and bottom edges of the tube) and the normalized intensity profiles along the scanline show that the neighboring view is displaced relative to the reference. **c**, Reconstructions by the same trained DAFNet with the deformable offsets set to zero at inference (top) and with alignment enabled (bottom). Without alignment, the tube is distorted and its edge is ghosted (insets); deformable alignment reconstructs a single, straight edge.

To test whether the deformable branch learns content-dependent offsets, we visualize the magnitude of the finest-level offset field across OLED validation scenes (Fig. S4). The displacements are substantial (mean $\sim$8 px) and **scene-structured**: checkerboard cells, building façades, and window frames are legible in the displacement field, showing content-dependent sampling rather than a uniform global shift.

2.5 Stage 2: Adaptive exposure/confidence fusion

The adaptive fusion head collapses the six PCD-aligned per-frame features into one fused feature with a per-pixel, content- and exposure-aware frame selection. From the raw sub-aperture frames it computes a per-frame, per-pixel **well-exposedness** map with a Mertens-style Gaussian

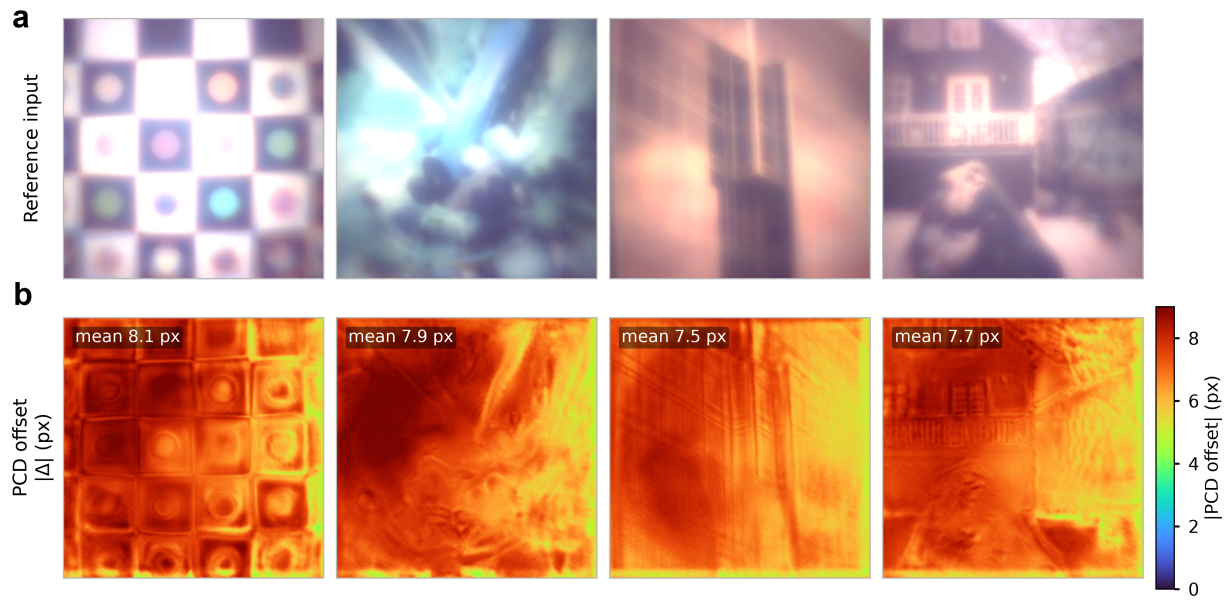


Figure S4: **The learned PCD deformable offsets are scene-structured and concentrated at a single scale.** **a**, Reference sub-aperture (AP3) input for four OLED validation scenes (a checkerboard, sky, a building façade, and a street scene). **b**, Magnitude of the finest-level PCD offset field for the same scenes, on a shared color scale; the per-scene mean is inset. The field averages ∼8 px and is strongly modulated by image content: the scene structure (checkerboard cells, building edges, and window frames) is imprinted in the offset map, indicating content-dependent rather than uniform sampling. The coarser-level warp modules remain at their zero initialization (mean $|\text{offset}| < 10^{-3}$ px; not shown) because only the finest aligned feature is used downstream.

on luminance [12], $\text{wexp} = \exp(-(\ell - 0.5)^2/2\sigma^2)$ (σ=0.2); a shared scoring convolution ingests [aligned feature ‖ wexp] per frame and a **softmax over the six frames** yields per-pixel soft selection weights; the aligned features are weighted-summed, and the result, the fused well-exposedness, and the reference raw are concatenated through a small convolutional head to 32 fused channels:

$$w = \text{softmax}_{\text{frames}}\big(\text{score}([\text{aligned} \,\|\, \text{wexp}])\big); \quad \text{out} = \text{head}\big([\textstyle\sum_f w \cdot \text{aligned} \,\|\, \sum_f w \cdot \text{wexp} \,\|\, \text{ref raw}]\big).$$

The adaptive fusion head contains ∼0.091 M parameters and preserves the same compact 32-channel interface to the reconstruction backbone. With a 14× exposure ladder, the most informative aperture varies across the image: short-exposure views preserve highlights, whereas high-transmission views retain shadow detail. The frame-wise softmax, which uses the raw-image well-exposedness maps as a direct exposure cue, therefore provides a lightweight per-pixel selection of this complementary exposure information. As an approximation, the well-exposedness maps are computed in each frame's own raw geometry, whereas the aligned features are in reference geometry; the resulting parallax-sized offset is accepted, with the maps serving as a global exposure cue and the softmax making the per-pixel decision. Multi-aperture fusion is the dominant quality lever in the design: a single aperture loses highlight and shadow detail that the six-aperture array captures, as the single-shot baseline in the main comparison shows (Supplementary Note 4). The two controlled component ablations in Supplementary Note 5 therefore evaluate second-order architectural refinements on top of multi-aperture fusion.

2.6 Stage 3: Static residual reconstruction

The reconstruction U-Net [13] (32 base channels) has a 4-stage encoder, bottleneck, and 4-stage decoder with skip connections, but its residual base is the *reference frame* rather than the network input:

$$\text{out} = \text{reference} + \text{Conv}_{\text{final}}\big(\text{dec}(\text{enc}(\text{fused}))\big). \tag{10}$$

Encoder channels follow 32→64→128→256→512 with block counts $(2, 2, 2, 4)$ and decoder $(2, 2, 2, 2)$. In the proposed model the reconstruction backbone uses *static* (not dynamic) convolutions by design: content-adaptive filtering is reserved for the front-end, where the field-varying optics act. The backbone holds the large majority of DAFNet's parameters (Table S1) yet uses no dynamic filtering, since the residual left after restoration, registration, and fusion is well-conditioned. Anchoring the output as a residual to the reference (Eq. 10) further eases the task.

2.7 Field-coordinate conditioning (CoordConv)

Translation-equivariant convolutions are blind to absolute position, yet the PSF blur depends on distance from the optical axis. DAFNet concatenates normalized field coordinates $[r, r^2]$ [14] (r = distance from the crop center, i.e., the optical axis; 0 at center → 1 at corner) to the front-end input (expanding the first convolution's input 3→5 channels). The added coordinate channels are **zero-initialized**, so the conditioning is an exact no-op at initialization, including when the remaining weights are warm-started from a pretrained model; the network then learns to use field position. The cost is negligible (a few hundred weights). A radial loss weight further pushes optimization effort onto the aberrated periphery (Section 2.10).

The coordinates thus complement the per-lens priors (Section 2.2), which encode aperture identity but not field position.

2.8 Decoupled Dynamic Filter (DDF)

The per-lens restoration stacks (Section 2.2) are built from decoupled dynamic filters [15]. A fully general dynamic convolution emits a distinct $k \times k$ kernel for every pixel *and* every channel, an $H \times W \times C \times k^2$ tensor whose materialization dominates memory traffic. DDF instead predicts two cheap components in parallel: a *spatial branch* (1×1 convolution) produces a spatially varying attention map of shape $G \times k^2 \times H \times W$, shared across the channels of each of G groups, and a *channel branch* (global average pooling followed by a squeeze-and-excitation bottleneck [16]) produces a channel-specific, spatially global filter of shape $C \times k \times k$. Their product, $F_{i,j,c} = F^{sp}_{i,j,g(c)} \odot F^{ch}_{c}$ for pixel (i, j) and channel c in group $g(c)$, is a kernel that is both spatially varying and channel-adaptive; filter normalization rescales the predicted filters to keep training well-conditioned. The predicted parameters thus scale as $\mathcal{O}(HWk^2 + Ck^2)$ rather than $\mathcal{O}(HWCk^2)$, a footprint of the order of a depthwise convolution. Each DDF residual block applies a DDF operator (stride 1) and channel attention, with a residual connection and a scaled identity shortcut, $\mathbf{y} = \text{ReLU}(\mathbf{x} + \text{CA}(\text{DDF}(\mathbf{x}))) + 0.1 \cdot \mathbf{x}$.

DDF fits this system for two reasons. First, the metalens PSF varies strongly with field position, wavelength, and exposure channel, and differs between apertures; a fixed convolution can only learn an average inverse of such a degradation, whereas DDF applies a locally tuned kernel at every pixel, and each aperture has its own DDF stack. Second, avoiding the full filter tensor keeps peak activation memory bounded and the operator bandwidth-friendly, which is what governs throughput on a Jetson-class accelerator, where DRAM bandwidth and on-chip memory, not raw FLOPs, are the binding constraints. The operator reduces to standard depthwise primitives, so it

is realizable as an efficient CUDA kernel for on-device inference (Section 2.12), and the six per-lens stacks together add only a small fraction of the parameter budget (Table S1).

K-means clustering (K=6, after ℓ_2 normalization) of the spatial branch's per-pixel 3×3 kernels, taken from the per-lens restoration stack of the converged model, confirms that the branch exploits its spatially adaptive capacity (Fig. S5). The clusters form spatially coherent regions that track image content (checkerboard cells, object boundaries, exposure structure); the kernels vary substantially across pixels (mean per-tap standard deviation ≈0.14 after normalization); and the six prototype kernels of each scene differ in the strength and sign of their center and surround weights, so different image regions receive different filters. The network therefore applies structurally different convolutions at different image locations rather than one position-invariant filter. The DDF front-end is not independently ablated in the controlled study of Supplementary Note 5; its inclusion rests on this per-aperture, content-adaptive formulation, its low parameter cost, and these learned spatially varying filters.

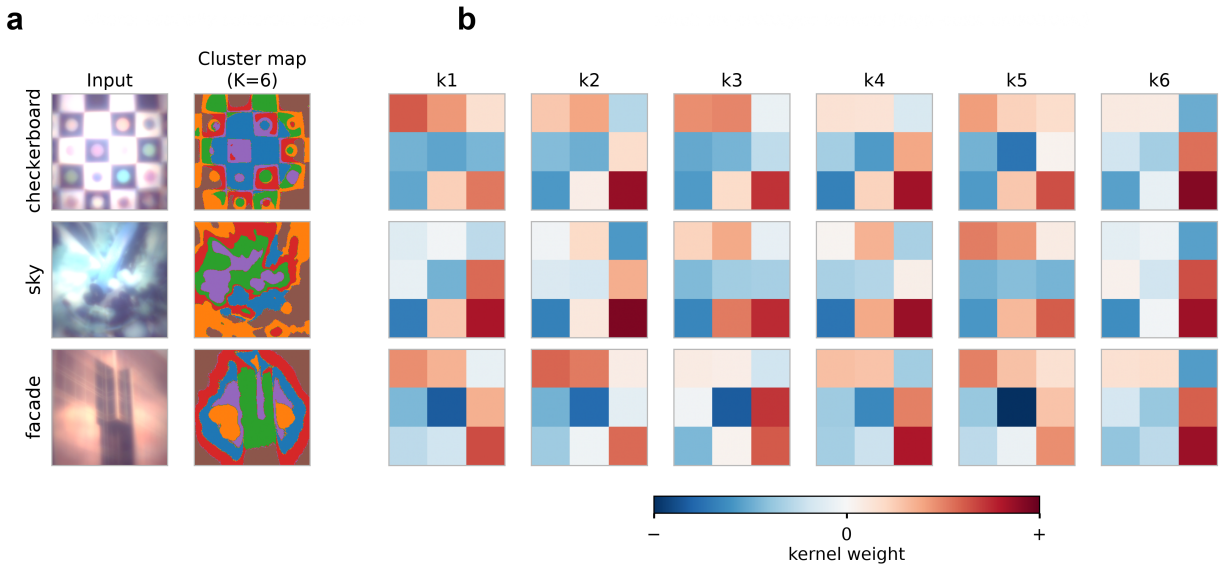


Figure S5: **The DDF spatial branch learns content-adaptive, spatially varying filters, not a single position-invariant kernel.** Per-pixel 3×3 kernels predicted by the decoupled dynamic filter (DDF) spatial branch in the per-lens restoration front-end, analyzed for three representative scenes (a checkerboard, sky, and a building façade). **a**, Reference sub-aperture input and a K-means clustering ($K = 6$) of the per-pixel kernels. The cluster regions are spatially coherent and track image structure, including checkerboard cells, object boundaries, and exposure regions. **b**, The six prototype kernels of each scene (k_1 to k_6, one per cluster) as 3×3 heatmaps on a shared diverging scale (blue, negative; red, positive). The prototypes differ in the strength and sign of their center and surround weights, indicating that the network applies structurally different convolutions at different image locations rather than one position-invariant filter. The per-tap standard deviation across pixels is ≈0.14, confirming that the branch is not collapsed to a uniform kernel. Kernels are taken from the most spatially diverse spatial-branch group of the per-lens restoration stack.

2.9 Parameter and compute budget

DAFNet totals 15.62 M parameters, or 15.33 M as deployed after inference-time PCD pruning; Table S1 gives the per-module breakdown. Content-adaptive filtering is confined to the front-end (per-lens DDF restoration, feature pyramid, and PCD alignment; under 1 M parameters in total), while the 14.63 M reconstruction backbone is plain static convolution (Section 2.6).

Table S1: DAFNet parameter budget (deployed model). Components are rounded independently; subtotals are computed at full precision.

Module	Params (M)
Restoration and alignment front-end	**0.61**
per-lens DDF restoration (6 lenses $\times$ 3 DDF residual blocks, width 28)	0.054
feature pyramid + FPN fusion (incl. input convolution, 5 $\rightarrow$ 28 channels)	0.331
PCD alignment (single-scale)	0.224
Adaptive fusion head	**0.09**
Reconstruction backbone (static residual U-Net)	**14.63**
encoder / bottleneck / decoder	7.13 / 3.54 / 3.94
Total	**15.33**

2.10 Loss function

Ablation arms are trained from scratch (the deployed model is instead fine-tuned from the lab-pretrained model, Section 3) with the perceptual-gradient composite loss:

$$\mathcal{L} = \mathrm{Charb}(\hat{y}, y) + 0.1\,\mathcal{L}_{\mathrm{over}} + 0.3\,(1 - \mathrm{SSIM}) + 0.5\,\mathcal{L}_{\mathrm{LPIPS}} + 1.0\,\mathcal{L}_{\mathrm{grad}}\ (+\,w_{\mathrm{edge}}\,\mathcal{L}_{\text{radial-L1}}), \quad (11)$$

where $\mathrm{Charb} = \sqrt{(\hat{y} - y)^2 + 10^{-6}}$ is the Charbonnier penalty [17], $\mathcal{L}_{\mathrm{over}} = \mathrm{ReLU}(\hat{y}-1)+\mathrm{ReLU}(-\hat{y})$ is a $[0, 1]$ range guard, SSIM [18] promotes structure, $\mathrm{LPIPS}_{\mathrm{vgg}}$ [19] penalizes perceptual distortion, and $\mathcal{L}_{\mathrm{grad}}$ is an ∇_x, ∇_y $L1$ term for edge preservation. The optional radial-weighted $L1$ (w_{edge} increasing $0 \rightarrow 1$ from optical axis to corner) is the loss-side complement of the field-coordinate input (Section 2.7). It is used as a *training-time* aid for peripheral detail and is enabled in the deployed model; because it is a loss term rather than an architectural component, it is not ablated separately. The 1.0-weighted gradient and 0.5 LPIPS terms are the dominant drivers of sharp, edge-faithful reconstructions.

2.11 Training configuration

Table S2: Key training settings.

Setting	Value
Optimizer	AdamW [20], weight decay 1×10^{-4}
Schedule	cosine annealing [21]
Learning rate	2×10^{-4} (from scratch) / 5×10^{-5} (warm-started)
Batch / crop	1 / 512×512
Selection	validate every 2 epochs; best by validation LPIPS
Stopping	patience-based ($\approx$160–180 epochs)
Dataset	1101 OLED ground-truth scenes; 80/20 split
Augmentation	photometric only; **no** flips, rotations, or random crops (to preserve the field-dependent PSF geometry)

All ablation arms share the split and settings of Table S2, so each change is evaluated in isolation. Training additionally guards against *reference-only collapse*, in which the network reconstructs from the reference aperture alone and ignores the neighboring views.

2.12 Edge deployment

DAFNet targets on-device inference on a Jetson Orin Nano (8 GB; JetPack 6 / CUDA 12.6). Two implementation choices keep the full six-frame model within the device's latency and memory envelope: a grouped per-lens front-end that lays the six lens-specific stacks out as a single grouped convolution (numerically equivalent to evaluating the six stacks separately, but computed in a single pass), and CUDA-graph capture to remove the kernel-launch-overhead floor. In eager mode, single-forward latency (pure fp16, median of three runs) is $\approx$0.37/0.87/1.52/3.42 s at 256/384/512/768 px, with peak memory $\approx$0.32–2.35 GB (Fig. S6). CUDA-graph capture reduces latency by a further 13–16% at 256–512 px (e.g. $\approx$0.31 s at 256 px), at the cost of higher memory use and a 512 px ceiling.

3 Supplementary Note 3: Real-Scene Training Data Pipeline

This note details the real-scene training pipeline summarized in the main-text Methods: controlled data acquisition under laboratory conditions (Section 3.1), the domain gap that prevents lab-only training from generalizing (Section 3.2), the difficulty of constructing real-scene supervision (Section 3.3), and the pseudo-reference construction and fine-tuning strategy that addresses it (Section 3.4).

3.1 Controlled data acquisition with OLED display

To obtain initial paired training data under controlled conditions, the metalens array camera is placed facing an OLED television at a fixed distance (Fig. S7). The display presents two types of content: geometric calibration patterns and natural target images drawn from the DIV2K [22], Inter4K [23], and Liu4K [24] datasets. All target images are center-cropped to a square aspect ratio and scaled to fill the OLED screen.

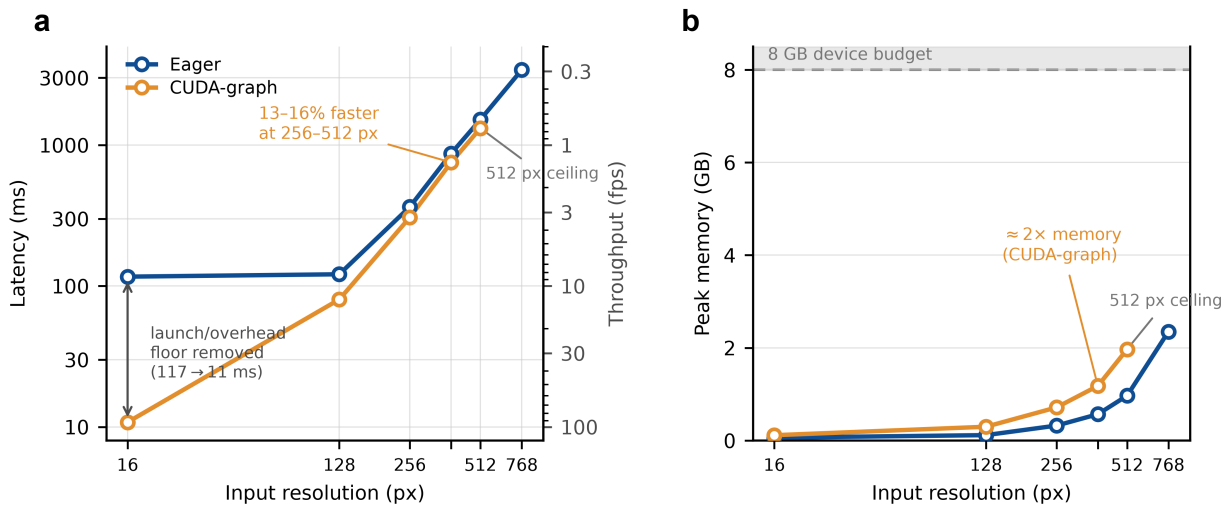


Figure S6: **On-device execution: eager mode versus CUDA-graph capture.** Single-forward benchmark of the DAFNet reconstruction network on an NVIDIA Jetson Orin Nano (8 GB), across input resolution (pure fp16, network-only inference, median of three runs). **a**, Reconstruction latency (left axis, log scale) and the equivalent throughput (right axis). In eager mode, latency is bounded below by a fixed kernel-launch and overhead floor of ≈117 ms that dominates at small resolutions. CUDA-graph capture removes this floor, reducing latency from 117 ms to 11 ms at 16 px, and is a further 13–16% faster where computation dominates (256–512 px); its capture is, however, limited to a 512 px ceiling. **b**, Peak memory versus input resolution. CUDA-graph capture approximately doubles the memory footprint, yet both modes remain far below the 8 GB device budget (dashed line) at every resolution. Reported latencies exclude sensor readout, preprocessing, and display.

3.1.1 Calibration

A checkerboard pattern displayed on the OLED screen is captured by the metalens array. Because the six sub-apertures observe the display from slightly different viewpoints, each produces a geometrically distinct image of the checkerboard. To estimate a per-sub-aperture homography $\mathbf{H}_i$ ($i = 1, \ldots, 6$), corresponding points are manually identified between each sub-aperture view and the displayed ground-truth image. The resulting homographies map each sub-aperture observation into a common reference frame aligned with the displayed content.

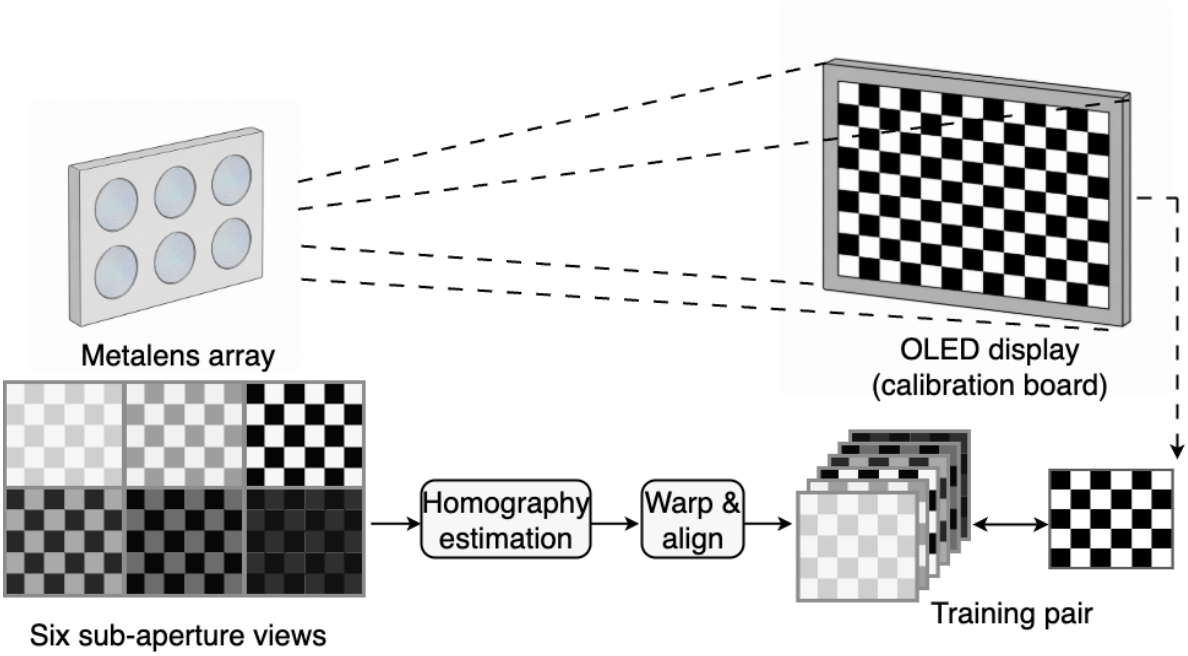


Figure S7: Controlled laboratory pipeline for paired data generation. The metalens array faces an OLED display showing calibration patterns and target scenes. A per-sub-aperture homography estimated from manually identified correspondences on a displayed checkerboard registers the display image to the reference sub-aperture.

3.1.2 Paired data generation

With the per-sub-aperture homographies established, natural images are displayed on the OLED and captured by the metalens array. Each capture produces six sub-aperture observations at different effective exposure levels (due to the per-lens transmittance coatings). Because the network fuses the six views in the coordinate frame of the reference sub-aperture (AP3), the display image is mapped into that frame with the inverse of the reference homography and serves as the ground truth for the six raw sub-aperture crops. The inputs themselves are not warped, so the inter-aperture misregistration is left for the network to resolve, exactly as for real captures, whose pseudo-references are likewise expressed in the reference-aperture frame by aligning the smartphone image to the reconstruction (Section 3.4). Warped inputs (all six views mapped to the display frame) were used only for a short initial warm-up, which gives the network a coarse initialization that already outputs the rough scene structure; all subsequent training uses unwarped inputs. A fixed margin of 100 pixels is cropped from all borders of both input and target images to discard invalid regions left by the warp. The resulting 1101 pairs are split 80/20 into training and held-out validation sets; the ablations of Supplementary Note 5 use the same split.

3.1.3 In-domain validation

Using the OLED capture pairs, the reconstruction network is trained following the configuration in Supplementary Note 2 and evaluated on the held-out split. A DAFNet trained on these pairs alone reaches 24.04 dB PSNR and 0.794 SSIM on this in-domain OLED split (Table S3), producing visually plausible HDR reconstructions with faithful recovery of structure and dynamic range content (Fig. S8). These results confirm that the six sub-aperture measurements contain sufficient

complementary information for HDR reconstruction when training and testing distributions are matched and scene geometry is well approximated by a planar surface.

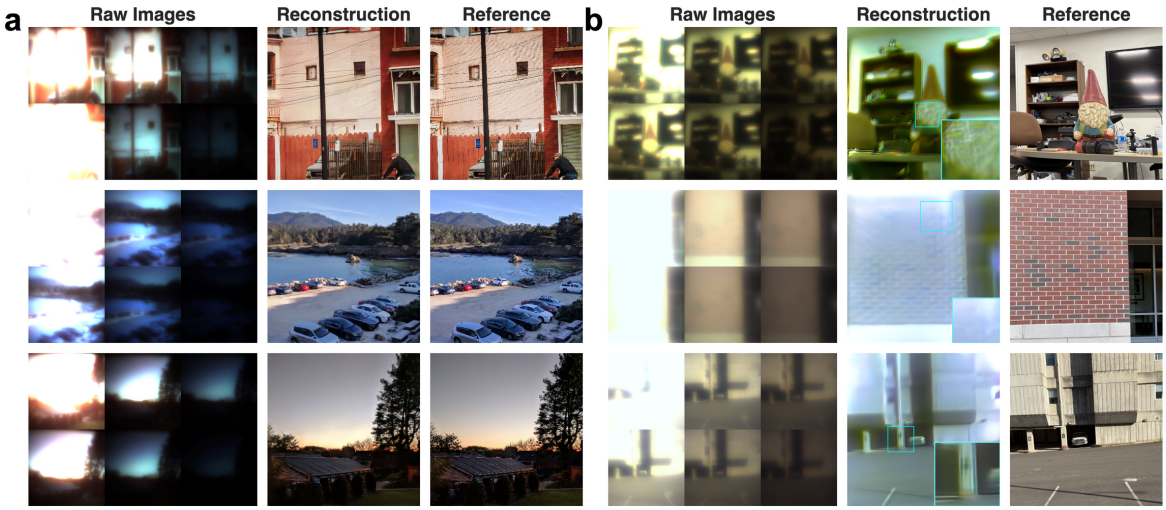


Figure S8: **In-domain validation and real-world generalization of the lab-trained model.** Each row shows one scene; columns show the six sub-aperture raw inputs (left, 2×3 grid), the network reconstruction (center) and a reference image (right). **a**, Held-out lab samples (in-domain), where the reference is the noise-free OLED-displayed image used as ground truth. With matched training and test distributions and a planar scene, the model recovers structure and dynamic-range content with high fidelity. **b**, The same lab-trained model applied directly to real indoor and outdoor scenes (out-of-domain), where the reference is a conventional capture of the scene. Cyan boxes mark regions where the reconstruction fails. Under this distribution shift the reconstructions exhibit pronounced color casts, residual blur and structural artifacts that are absent in (a), revealing the gap between the planar OLED training domain and real-world scenes.

3.2 Domain gap between lab and real scenes

Despite this in-domain performance, the lab-trained model degrades substantially when applied to real indoor and outdoor scenes. Two systematic failure modes are observed (Fig. S8b).

In scenes with depth variation, the spatial separation of the six sub-apertures introduces depth-dependent cross-lens parallax. A single planar homography, estimated from the flat OLED calibration surface, cannot simultaneously align foreground and background structures at different depths. This mismatch causes local distortions, ghosting, and edge duplication around depth boundaries. In the depth-varied scene of Fig. S8b (bottom row), for example, the vertical structures in front of the building are visibly distorted.

Even in scenes where geometric misalignment is less dominant, the lab-trained model still produces color shifts, highlight tone inconsistencies, and unnatural rendering around bright light sources (Fig. S8b, top and middle rows). These artifacts indicate that the performance gap is not purely geometric. OLED-displayed content does not faithfully reproduce the illumination statistics, surface reflectance distribution, spectral characteristics, or color gamut encountered in natural imaging. A model trained only on OLED-displayed targets therefore overfits to the photometric characteristics of the lab domain.

These two gaps are coupled: real scenes violate both the planar geometry and the photometric statistics assumed by the lab pipeline. Closing them requires training on real-scene data, for which

aligned supervision is itself difficult to obtain (Section 3.3).

3.3 Challenges in cross-device registration

Forming real-scene training pairs by directly registering co-captured metalens and smartphone images is impractical for three reasons. **First**, the six metalens views come from spatially separated sub-apertures rather than a single center of projection, so cross-lens parallax in real 3D scenes produces depth-dependent disparities that no single global transformation can describe, particularly near occlusion boundaries and thin structures (Fig. S9a). **Second**, raw metalens observations are heavily blurred, low in contrast, and poor in high-frequency texture, whereas smartphone images are sharp, well exposed, and computationally processed; this appearance gap prevents standard feature matchers from establishing reliable correspondences (Fig. S9b). **Third**, uncontrolled illumination, specular highlights, cast shadows, surface reflectance variation, and dynamic ranges far broader than an OLED display can represent (Fig. S8b) further degrade registration accuracy.

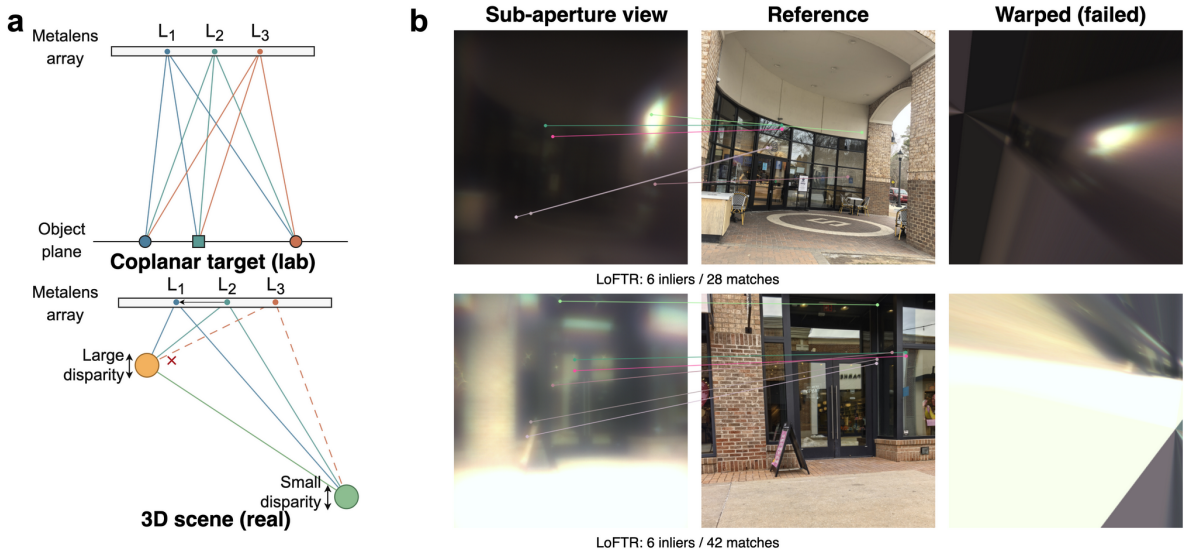


Figure S9: **Challenges in constructing real-scene supervision. a**, Depth-dependent parallax. For a coplanar target (top) a single homography per sub-aperture suffices, since all scene points share one plane. For a real 3D scene (bottom) disparity varies with depth (large for near points, small for far points), so no single homography can align the sub-aperture views, and near-field occlusions further break the planar assumption. **b**, Cross-device registration on raw observations. Each row shows a metalens sub-aperture raw observation (left), the corresponding smartphone reference (center) and the result of warping the raw view onto the reference (right). Even LoFTR [25], a state-of-the-art detector-free dense matcher, yields few geometrically consistent correspondences on the severely degraded raw observations (6 inliers from 28 and 42 putative matches, respectively), and the resulting alignment fails. This failure stems from the optical degradation of the raw observations rather than the matcher itself; it is resolved in Section 3.4 by preprocessing the inputs prior to matching.

This difficulty concerns only the external training targets, not the internal alignment of the sub-aperture observations, which the network performs with its deformable convolution layers [11] (Supplementary Note 2).

3.4 Pseudo-reference construction via reconstruction-domain alignment

Real-scene dataset. We acquired 628 captures of static indoor and outdoor scenes with the metalens camera. For each capture, the same scene was also recorded with a smartphone using its built-in multi-exposure burst HDR mode. We used the HDR image from the mobile phone to build the pseudo-reference described below. The captures are split 80/20 into fine-tuning and held-out sets; the held-out captures are used only for the real-scene evaluation in main-text Fig. 3c. Fine-tuning starts from the OLED-pretrained weights and otherwise follows Table S2, with the warm-start learning rate of 5×10^{-5}.

3.4.1 Core idea: alignment in the reconstruction domain

The key observation is that the lab-pretrained model, although inaccurate on real data, fails *gracefully*: its reconstructions retain the pixel-level errors of lab-only training (color shifts, residual blur, and structural distortions) but preserve the principal object contours, boundaries, and overall spatial layout of the scene. Because alignment requires correspondence of scene geometry rather than pixel-accurate appearance, these reconstructions can be registered to the smartphone references even though they are not yet usable as final outputs, whereas the raw measurements lose even this coarse structure (Fig. S9b).

This is the bootstrapped training strategy summarized in main-text Fig. 3b: the OLED-pretrained model is applied to the real metalens measurements, its reconstructions (rather than the raw measurements) are aligned to the co-captured smartphone references, and the aligned references serve as pseudo-references for fine-tuning. The real-scene fine-tuning uses the same network architecture and loss function as the OLED pretraining stage (Supplementary Note 2, Sections 2.10–2.11).

3.4.2 Alignment method comparison

To evaluate how the pseudo-reference construction strategy affects final reconstruction quality, we compare four alignment methods: SIFT [26], ECC [7], RAFT [6], and LoFTR [25]. All methods operate on the same initial reconstructions produced by the same pretrained model, so the comparison isolates the effect of the alignment strategy on downstream reconstruction quality.

For each method, alignment is performed independently for every scene in the real-scene training set: the initial reconstruction produced by the pretrained model is aligned to the co-captured smartphone image, and the aligned smartphone image serves as the pseudo-reference target for that scene. The real captures themselves are not pre-warped by the calibration homographies: for fine-tuning and inference, the network receives the raw sub-aperture crops and resolves their misregistration with its deformable alignment. The complete set of per-scene pseudo-reference targets then forms the fine-tuning training set.

3.4.3 Results

Fig. S10 shows representative pseudo-reference targets and the corresponding final reconstructions after fine-tuning with each method's output.

3.4.4 Analysis of alignment method behavior

Residual blur in the initial reconstructions flattens SIFT's Difference-of-Gaussians (DoG) response and makes its gradient-histogram descriptors weakly distinctive, so few keypoints are detected and the ratio test rejects most candidate matches. The survivors cluster in isolated high-contrast regions

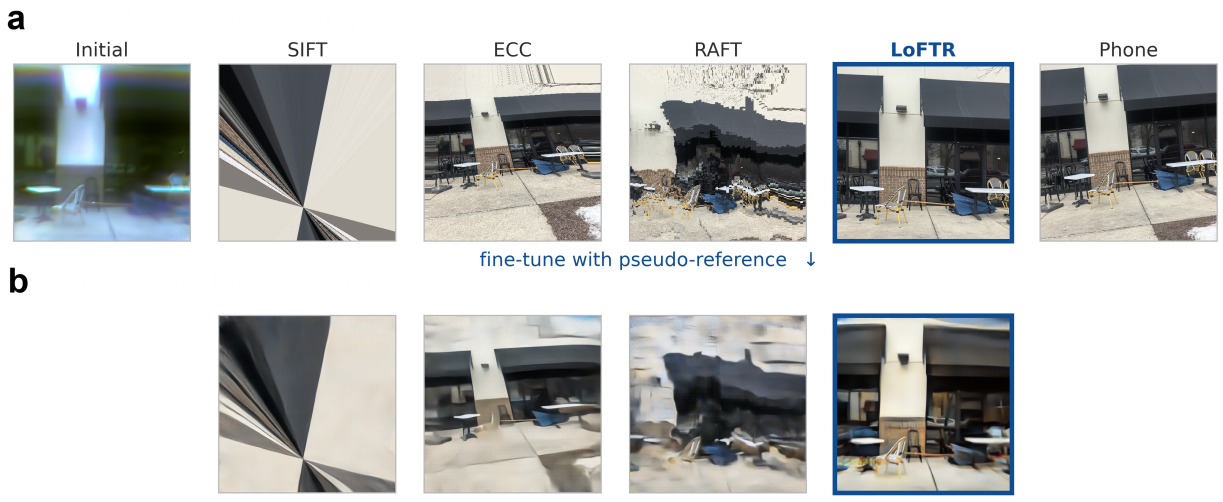


Figure S10: **LoFTR-aligned pseudo-references yield the best downstream reconstruction.** Comparison of pseudo-reference construction strategies on a real scene from the fine-tuning set. **a**, The initial reconstruction of the lab-pretrained model (Initial), the smartphone image aligned to it by SIFT, ECC, RAFT and LoFTR, and the unaligned smartphone image (Phone). **b**, Reconstructions after fine-tuning on each method's pseudo-references, placed under the corresponding alignment method. SIFT finds too few reliable correspondences, so its homography collapses the target into radial streaks, which the fine-tuned model reproduces. ECC converges to a tilted target with stretched borders, and the fine-tuned model blurs and distorts the scene. RAFT's dense flow tears the awnings and storefront, and the fine-tuned model inherits these tears. LoFTR (highlighted) best preserves the global scene structure, and the model fine-tuned on its pseudo-references recovers the awnings, pillar, chairs and tables; LoFTR is therefore used to build the pseudo-references throughout.

(e.g., strong edges against sky), leaving the estimated homography unreliable over the full image extent.

The ECC objective (maximized over an affine warp in our configuration) is invariant only to global linear intensity changes, but the reconstruction and the smartphone reference differ nonlinearly in tone mapping, local contrast, gamma, and color rendering, creating numerous local optima. Near-zero gradients in blurred, low-texture regions leave the Gauss-Newton update without signal, so it stalls or drifts, and even the best available initialization (from SIFT, when it finds enough matches) is unreliable.

RAFT's learned features give partially usable correspondences in textured regions, but its official pretrained weights were trained on synthetic optical-flow datasets (FlyingChairs [27], FlyingThings3D [28], Sintel [29]) with only moderate appearance variation, so its correlation volume becomes ambiguous in low-texture or photometrically mismatched regions. Its dense warp also lacks the RANSAC homography regularization [30] used for SIFT and LoFTR, so per-pixel flow errors propagate directly into local distortions of the pseudo-reference.

LoFTR first matches at 1/8 of the input resolution, where blur and noise are suppressed but the structural cues preserved by the initial reconstruction (object boundaries, region-level contrast, spatial layout) remain, and its self- and cross-attention transformers [31] use global scene context to disambiguate regions that defeat the local mechanisms of SIFT, ECC, and RAFT. Its training data, MegaDepth [32] and ScanNet [33], span large viewpoint, lighting, and camera-model changes

and are thus much closer to this cross-domain problem than RAFT's synthetic flow data. Finally, its correspondences are more numerous and more evenly distributed than SIFT's, giving a well-conditioned RANSAC homography that regularizes residual matching errors.

Although the LoFTR alignment is geometrically imperfect, fine-tuning on its pseudo-references yielded the best downstream reconstruction among the four alignment methods in the scenes we inspected (Fig. S10). These pseudo-references provide supervision in the correct target domain (real illumination, surface reflectance, and color statistics), which addresses the dominant error source of the lab-pretrained model: the photometric and structural gap between OLED-displayed content and real scenes (Section 3.2). Moreover, LoFTR's residual misalignment is spatially incoherent, varying in direction and magnitude across the image, so we expect it to average out over training iterations and spatial locations, whereas OLED-only training imposes systematic, globally correlated biases (consistent color and tone-mapping shifts) that reinforce rather than cancel.

4 Supplementary Note 4: Extended Real-World Results

This note reports two extended real-world evaluations that complement the representative scenes of the main text (Fig. 4): a static-scene comparison of our reconstruction against alternative acquisition modes and existing HDR methods on the same metalens array (Section 4.1), and a dynamic-scene comparison that isolates the advantage of snapshot acquisition over a temporal burst under motion (Section 4.2). All results are produced by the same reconstruction pipeline without scene-specific tuning.

4.1 Comparison with acquisition modes and existing HDR methods

To place the snapshot array in context, Fig. S11 compares our reconstruction against two *acquisition modes* on the same optics (a single-aperture single shot and a six-frame temporal burst) and four *existing HDR reconstruction methods* applied to the same metalens captures: the classical multi-exposure fusions of Mertens [12] and Debevec [34], and the learned HDR networks AHDRNet [35] and HDRUNet [36]. The learned networks are trained on the same paired data that we construct for our model rather than used with their released weights. For the three-frame networks (AHDRNet here, and SAFNet [37] in the quantitative comparison of main-text Fig. 3c), the six sub-aperture images are sorted by brightness and every other one is used as input; the single-image network HDRUNet takes the reference sub-aperture image (AP3). Both acquisition modes are captured through the reference aperture of the array (AP3): the single shot uses one exposure, and the temporal burst records six sequential frames of that aperture at different exposure times, set manually under the same illumination so that the six frames reproduce the exposure levels of the snapshot sub-apertures, including the extent and brightness of saturated regions. For each acquisition mode, our reconstruction network (DAFNet) is retrained on the corresponding input. All methods are evaluated on the same held-out real captures against the same pseudo-references. The single-shot column shows that a single aperture loses highlight and shadow detail that the array captures; the classical fusions break down on the blurred, cross-exposure sub-images; and our network resolves cleaner structure and tone than the learned baselines, which were not designed for the metalens array's coupled blur-parallax-exposure degradation.

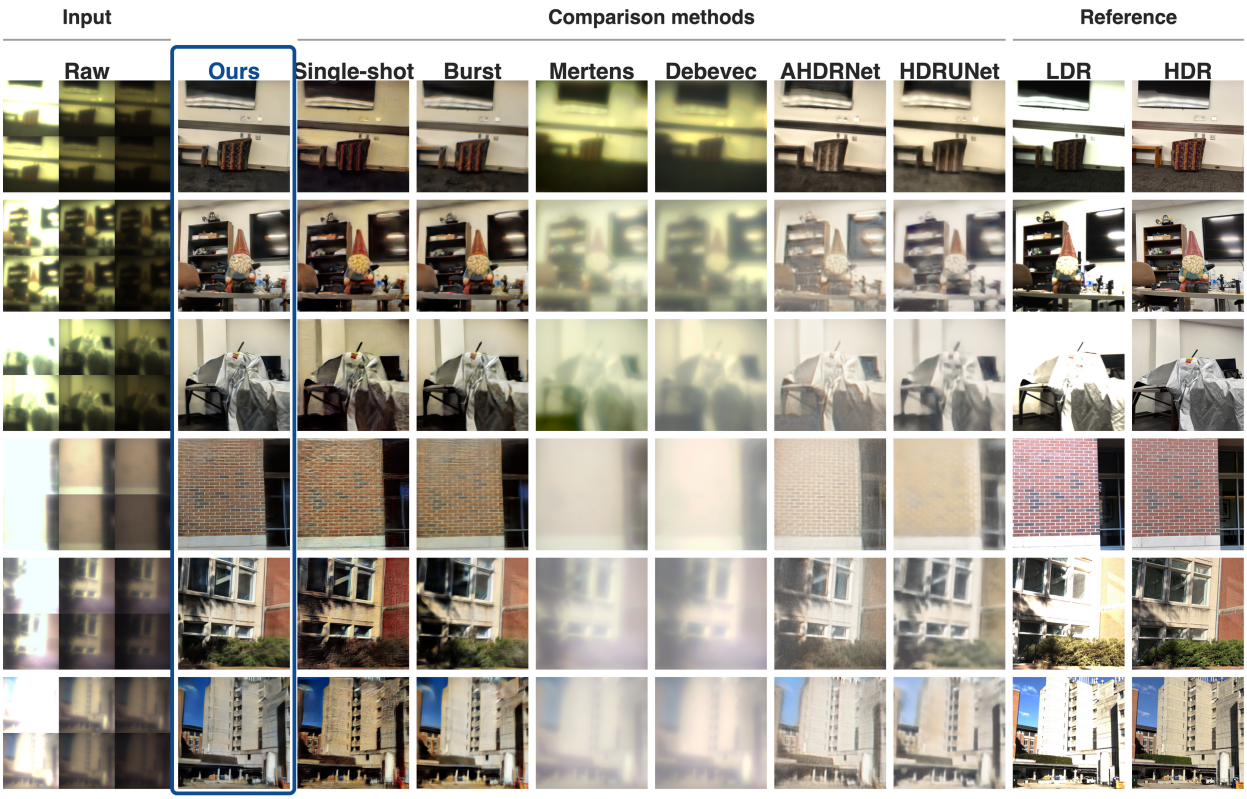


Figure S11: **Our reconstruction resolves cleaner structure and tone than existing acquisition modes and HDR methods on the metalens array.** Reconstruction comparison over six indoor and outdoor scenes. From left to right: the raw six-aperture array capture (Raw); our reconstruction (Ours, highlighted) together with two acquisition-mode baselines on the same optics (single-aperture single-shot and a six-frame temporal burst) and four existing HDR reconstruction methods applied to the same metalens captures, namely the classical multi-exposure fusions of Mertens [12] and Debevec [34] and the learned networks AHDRNet [35] and HDRUNet [36]; and the LDR and HDR references. The single-aperture shot loses highlight and shadow detail relative to ours, the classical fusions break down on the blurred, cross-exposure sub-images, and the learned baselines, which were not designed for the array's coupled blur, parallax, and exposure degradation, leave residual artifacts, whereas our network recovers the cleanest geometry and tone across scenes.

4.2 Dynamic scenes: snapshot acquisition versus temporal burst

The comparisons above are restricted to static scenes because the LDR and HDR references, and the pseudo-reference protocol used throughout, require a stationary scene (main text). To assess robustness beyond this regime, we compare the three acquisition modes on captures for which no aligned reference can be co-captured, grouped by capture condition: a static camera imaging moving content under low light and under strong (high-dynamic-range) light, and a handheld camera imaging scenes with and without a moving object (Fig. S12). The comparison is therefore relative across acquisition modes rather than against a ground truth. Across every condition the temporal burst ghosts wherever the scene or the camera moves, and the single-shot baseline loses highlight detail and stays noisy in shadow, whereas our reconstruction stays stable.

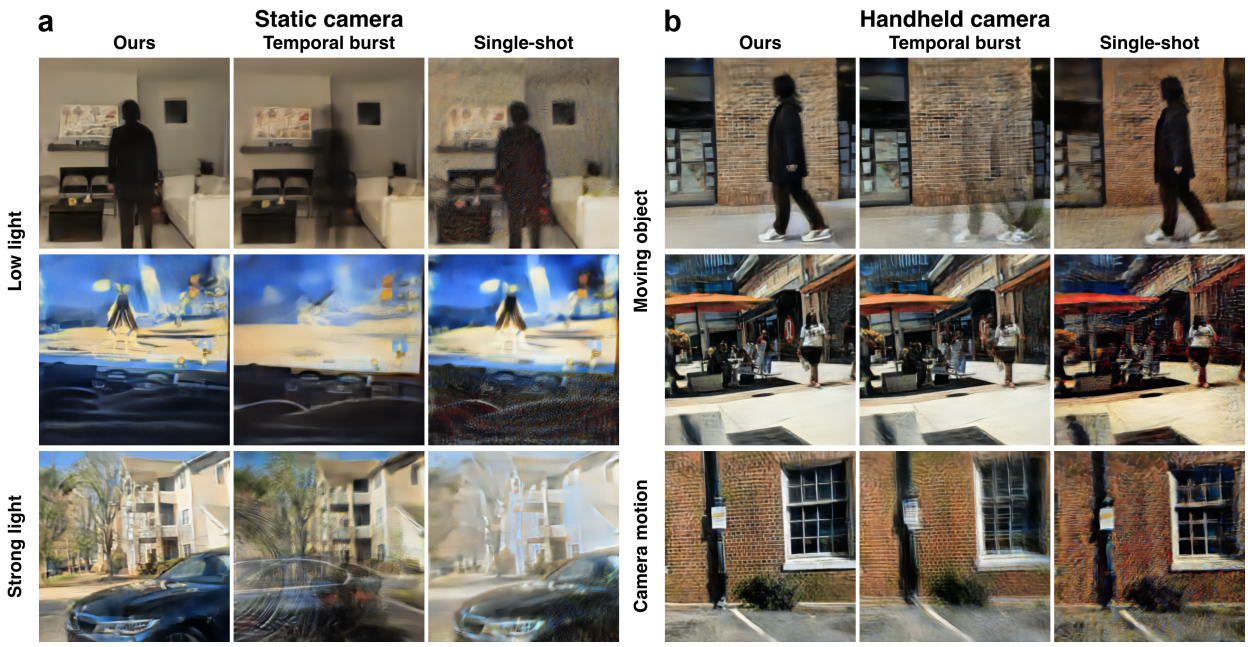


Figure S12: **Snapshot reconstruction stays stable across lighting extremes and camera motion, whereas the temporal-burst and single-shot baselines each fail in a characteristic way.** Robustness comparison across capture conditions for which no aligned reference can be co-captured, since dynamic content and handheld capture admit no pixel-registered ground truth; the comparison is therefore relative across acquisition modes rather than against a ground truth. For every scene, three reconstructions of the same scene, one per acquisition mode, are shown: our reconstruction (Ours), a six-frame temporal exposure burst (Temporal burst), and a single-shot baseline (Single-shot). **a**, Static-camera capture across lighting regimes. Two low-light scenes, an indoor dimly lit scene and a projected-display scene captured under low light, probe shadow recovery; a high-dynamic-range daylight scene containing a moving vehicle probes highlight retention. **b**, Handheld capture. Two scenes contain a moving object, a walking pedestrian and a street crowd, and one is an otherwise static building facade imaged under camera motion alone. Across all conditions the temporal burst leaves ghosting and edge doubling wherever the scene or the camera moves, and the single-shot baseline loses highlight detail and stays noisy in shadow, whereas our reconstruction retains sharp boundaries and balanced tone. No ground-truth or smartphone reference is shown, as none can be acquired for moving or handheld capture.

5 Supplementary Note 5: Extended Ablation Analysis

5.1 Evaluation protocol

We evaluate one system-level input baseline and two controlled component ablations. The reference-only baseline tests the value of multi-aperture capture by removing the five neighboring aperture measurements. The remaining ablations independently disable learned deformable alignment or field-coordinate conditioning, with each model retrained from scratch under otherwise identical settings.

All ablations are evaluated on the controlled OLED dataset, which provides exact, pixel-registered ground truth (the displayed image). As in pretraining (Section 3.1), the six sub-aperture crops are not warped, so the inter-aperture misalignment is preserved and must be resolved by the

network; only the displayed target is registered to the reference aperture. This choice is deliberate: isolating a single component's contribution requires ground-truth error smaller than the effect being measured, whereas the LoFTR-aligned smartphone pseudo-references used for real scenes carry residual alignment noise that would confound per-component deltas. The degradations these components address (the field-varying PSF, inter-aperture misregistration, and the 14× exposure ladder) are properties of the optics and are present in both lab and real captures. On the planar OLED targets, however, the misregistration consists mainly of fixed offsets between the sub-images, whereas real scenes add depth-dependent parallax; the alignment ablation should be read with this difference in mind. Two consequences follow for the interpretation of Table S3. First, its absolute values are not comparable with the real-scene metrics of the main text (Fig. 3c): the two are computed against different references, the pixel-registered OLED image here, versus LoFTR-aligned smartphone pseudo-references there, so only within-table differences are meaningful. Second, every arm in Table S3 is retrained from scratch on the OLED split, whereas the deployed model reported in the main text is fine-tuned from the lab-pretrained model on real-scene pseudo-references (Section 3.4). The full system's real-scene performance is reported separately (main text; Supplementary Note 4).

5.2 Master ablation table

Table S3: DAFNet evaluation on the controlled OLED dataset, comprising one system-level reference-only baseline and two controlled component ablations. All variants are retrained from scratch under otherwise identical settings. Values are referenced to the pixel-registered OLED ground truth and are not directly comparable with the main-text metrics (Section 5.1).

Configuration	PSNR↑	SSIM↑	LPIPS↓	GMSD↓
DAFNet (proposed)	**24.04**	**0.794**	**0.408**	**0.134**
reference-only (single aperture)	11.34	0.326	0.710	0.276
w/o alignment	23.14	0.764	0.443	0.147
w/o field coordinates	23.59	0.780	0.435	0.139

The proposed model is best on all four metrics. **Multi-aperture fusion** has by far the largest effect: reconstructing from the reference aperture alone lowers PSNR by ∼12.7 dB and SSIM by 0.47, confirming that the six-aperture array is essential rather than incidental. Among the learned components added to this multi-aperture base, **alignment** has the largest effect (removing it lowers SSIM the most, consistent with the loss of contrast and detail when unregistered sub-images are averaged), followed by field conditioning.

Fig. S13 shows the corresponding reconstructions for six OLED scenes together with the pixel-registered ground truth.

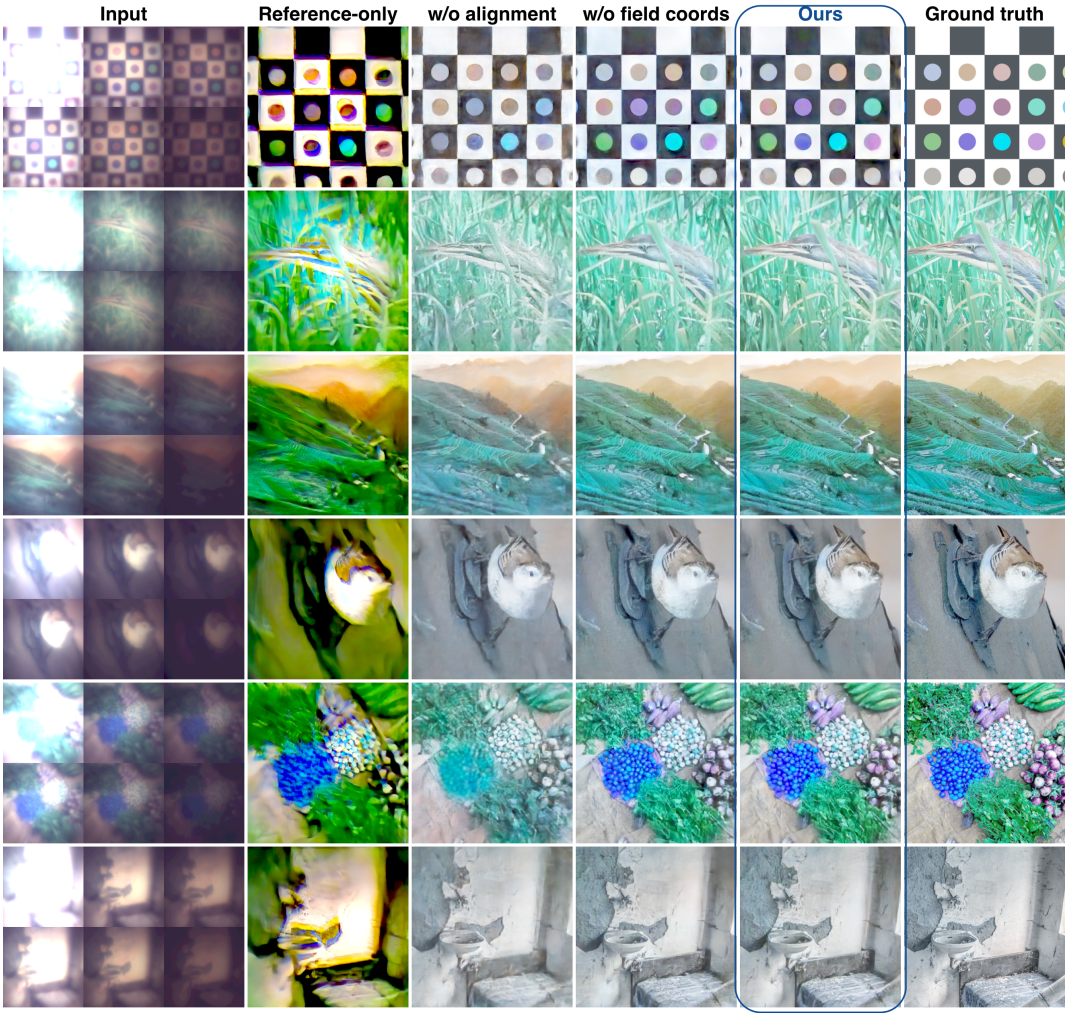


Figure S13: **Each removed component leaves a distinct, visible failure signature, and only the full model recovers structure, color, and dynamic range that match the ground truth.** Qualitative ablation on the controlled OLED dataset, which provides pixel-registered ground truth (Section 5.1). Each row is one scene; from left to right the columns show the six sub-aperture inputs (Inputs; a 2×3 montage ordered column-wise from the brightest exposure at top-left to the darkest at bottom-right), the reference-only baseline, the two component ablations, our full model (Ours, boxed), and the ground truth. **Reference-only** reconstructs from a single aperture and cannot recover the captured dynamic range: highlights saturate, shadows are clipped to black, and strong color casts, chromatic fringing, and hallucinated texture appear, the largest degradation in Table S3. **w/o alignment** fuses the six apertures without registration; averaging mismatched exposures washes out color and contrast and softens fine detail, most visibly in the checkerboard (top row) and the berries (fifth row). **w/o field coordinates** is visually close to ours at this scale; its differences are subtle and concentrated toward the image corners, consistent with the smaller quantitative drop in Table S3. **Ours** restores sharp single edges, faithful color, and the full tonal range, and is the closest to the ground truth in every scene.

5.3 Alignment: removing it washes out color and fine detail

The w/o-alignment arm is retrained from scratch with the deformable offsets fixed to zero; the offset-prediction layers and the modulation mask are kept, so the parameter count is unchanged and the five neighbors are fused *unregistered*. On these planar OLED scenes, averaging mismatched exposures **washes out color and contrast** and softens fine detail. The learned offsets are characterized in Section 2.4; in real captures, disabling the offsets also distorts and ghosts edges (Fig. S3). Removing alignment costs LPIPS 0.408 → 0.443, PSNR 24.04 → 23.14, and SSIM 0.794 → 0.764, the **largest single-component drop** in the table. The checkerboard scene of Fig. S13 (top row) shows the signature directly: the colored circles lose saturation and shift in hue, whereas the full model reproduces them faithfully.

5.4 Field conditioning and the field-varying PSF

Removing the field-coordinate conditioning (Section 2.7) costs LPIPS 0.408 → 0.435 and PSNR 24.04 → 23.59, and leaves a field-dependent color error and peripheral softness toward the image corners (Fig. S13). This signature is consistent with the convolutions exploiting absolute field position to invert the position-dependent metalens aberration, and the conditioning is among the cheapest of the learned components to add.

References

[1] Matsushima, K. Shifted angular spectrum method for off-axis numerical propagation. *Optics Express* **18**, 18453–18463 (2010).

[2] Goodman, J. W. *Introduction to Fourier Optics* (Roberts & Company Publishers, Englewood, CO, 2005), 3 edn.

[3] Fröch, J. E. *et al.* Beating spectral bandwidth limits for large aperture broadband nano-optics. *Nature communications* **16**, 3025 (2025).

[4] Huang, L., Whitehead, J., Colburn, S. & Majumdar, A. Design and analysis of extended depth of focus metalenses for achromatic computational imaging. *Photonics Research* **8**, 1613–1623 (2020).

[5] Chakravarthula, P. *et al.* Thin on-sensor nanophotonic array cameras. *ACM Transactions on Graphics* **42**, 1–18, 10.1145/3618398 (2023).

[6] Teed, Z. & Deng, J. RAFT: Recurrent all-pairs field transforms for optical flow. In *Computer Vision – ECCV 2020*, vol. 12347 of *Lecture Notes in Computer Science*, 402–419 (Springer, 2020).

[7] Evangelidis, G. D. & Psarakis, E. Z. Parametric image alignment using enhanced correlation coefficient maximization. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **30**, 1858–1865 (2008).

[8] Wang, X., Chan, K. C. K., Yu, K., Dong, C. & Loy, C. C. EDVR: Video restoration with enhanced deformable convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 1954–1963, 10.1109/CVPRW.2019.00247 (2019).

[9] Lin, T.-Y. *et al.* Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 936–944, 10.1109/CVPR.2017.106 (2017).

[10] Mandal, D., Peng, Z., Wang, Y. & Chakravarthula, P. Enabling high-quality in-the-wild imaging from severely aberrated metalens bursts. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 849–859, 10.1109/WACV61042.2026.00089 (2026).

[11] Dai, J. *et al.* Deformable convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 764–773 (2017).

[12] Mertens, T., Kautz, J. & Van Reeth, F. Exposure fusion. In *15th Pacific Conference on Computer Graphics and Applications (PG)*, 382–390, 10.1109/PG.2007.17 (IEEE Computer Society, 2007).

[13] Ronneberger, O., Fischer, P. & Brox, T. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, vol. 9351 of *Lecture Notes in Computer Science*, 234–241 (Springer, 2015).

[14] Liu, R. *et al.* An intriguing failing of convolutional neural networks and the CoordConv solution. In *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 31, 9605–9616 (2018).

[15] Zhou, J., Jampani, V., Pi, Z., Liu, Q. & Yang, M.-H. Decoupled dynamic filter networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6647–6656, 10.1109/CVPR46437.2021.00658 (2021).

[16] Hu, J., Shen, L. & Sun, G. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 7132–7141 (2018).

[17] Charbonnier, P., Blanc-Féraud, L., Aubert, G. & Barlaud, M. Two deterministic half-quadratic regularization algorithms for computed imaging. In *Proceedings of the 1st International Conference on Image Processing (ICIP)*, vol. 2, 168–172, 10.1109/ICIP.1994.413553 (IEEE, 1994).

[18] Wang, Z., Bovik, A. C., Sheikh, H. R. & Simoncelli, E. P. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing* **13**, 600–612 (2004).

[19] Zhang, R., Isola, P., Efros, A. A., Shechtman, E. & Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2018).

[20] Loshchilov, I. & Hutter, F. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations* (2019).

[21] Loshchilov, I. & Hutter, F. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations (ICLR)* (2017).

[22] Agustsson, E. & Timofte, R. NTIRE 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 1122–1131, 10.1109/CVPRW.2017.150 (2017).

[23] Stergiou, A. & Poppe, R. AdaPool: Exponential adaptive pooling for information-retaining downsampling. *IEEE Transactions on Image Processing* **32**, 251–266 (2023).

[24] Liu, J. *et al.* A comprehensive benchmark for single image compression artifact reduction. *IEEE Transactions on Image Processing* **29**, 7845–7860, 10.1109/TIP.2020.3007828 (2020).

[25] Sun, J., Shen, Z., Wang, Y., Bao, H. & Zhou, X. LoFTR: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8922–8931 (2021).

[26] Lowe, D. G. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision* **60**, 91–110 (2004).

[27] Dosovitskiy, A. *et al.* FlowNet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2758–2766 (2015).

[28] Mayer, N. *et al.* A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4040–4048 (2016).

[29] Butler, D. J., Wulff, J., Stanley, G. B. & Black, M. J. A naturalistic open source movie for optical flow evaluation. In *Computer Vision – ECCV 2012*, vol. 7577 of *Lecture Notes in Computer Science*, 611–625 (Springer, 2012).

[30] Fischler, M. A. & Bolles, R. C. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM* **24**, 381–395, 10.1145/358669.358692 (1981).

[31] Vaswani, A. *et al.* Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30 (2017).

[32] Li, Z. & Snavely, N. MegaDepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2041–2050 (2018).

[33] Dai, A. *et al.* ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2432–2443, 10.1109/CVPR.2017.261 (2017).

[34] Debevec, P. E. & Malik, J. Recovering high dynamic range radiance maps from photographs. In *Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, 369–378, 10.1145/258734.258884 (1997).

[35] Yan, Q. *et al.* Attention-guided network for ghost-free high dynamic range imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1751–1760, 10.1109/CVPR.2019.00185 (2019).

[36] Chen, X., Liu, Y., Zhang, Z., Qiao, Y. & Dong, C. HDRUNet: Single image HDR reconstruction with denoising and dequantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 354–363, 10.1109/CVPRW53098.2021.00045 (2021).

[37] Kong, L. *et al.* SAFNet: Selective alignment fusion network for efficient HDR imaging. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 256–273, 10.1007/978-3-031-73347-5_15 (2024).