Efficient Nano-Optical Eye Tracking for Smart Glasses

JIPENG SUN*, Google Inc., Princeton University, USA
LEKANG YUAN*, Princeton University, USA
RUSLAN GUSEINOV, Google Inc., Austria
BERND BICKEL and THABO BEELER, Google Inc., Switzerland
THOMAS AUZINGER, Google Inc., Austria
FELIX HEIDE, Princeton University, USA

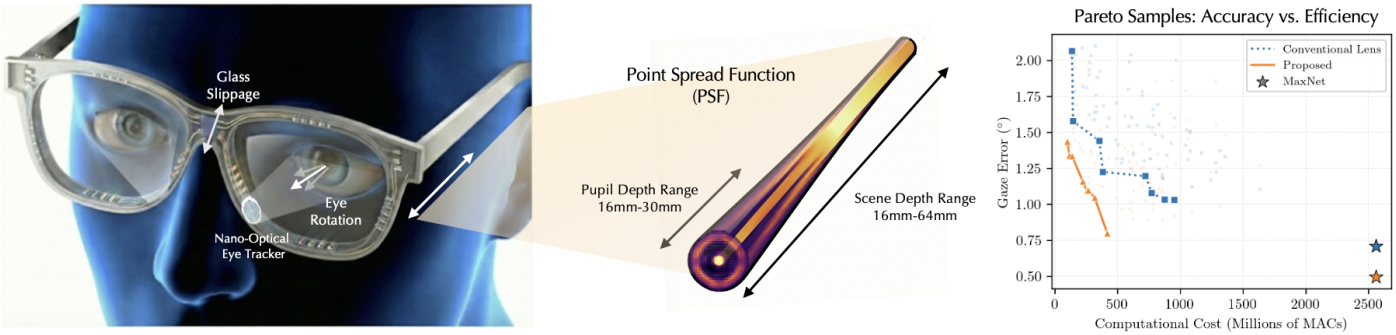


Fig. 1. Eye tracking in smart glasses is constrained by strict battery limits and device slippage. Existing trackers project glints onto the cornea to resolve the ambiguity between slippage and eye rotation. Left: We present a method that resolves this ambiguity optically using an end-to-end designed camera. Middle: We co-engineer a Point Spread Function (PSF) that varies rapidly across the slippage depth range, jointly optimizing it with the sensor and neural network. Right: Our approach matches the 0.75° average gaze error of the best glint-free refractive baseline but uses more than 6× fewer floating-point operations.

Accurate gaze tracking is essential for interaction in VR and AR devices; however, integrating it into the compact form factor of smart glasses remains a challenge. Existing approaches typically rely on active glint-based pupil-center corneal reflections to decouple eye rotation from devices slippage. While effective, the required geometric arrangement of illumination sources is challenging to integrate into compact eyewear. Without the reference features provided by glints, passive tracking requires computationally expensive processing to resolve the ambiguity between slippage and rotation.

In this work, we introduce a glint-free nano-optical eye-tracking method with the lens, sensor operation, and reconstruction network tailored to this task. We design a point spread function that provides an optical encoding that physically disentangles slippage from rotation. We validate our method both in simulation and with an experimental prototype, finding that our optical encoding with end-to-end design allows for an approximately 6× reduction in floating-point operations (FLOPs) compared to conventional glint-free methods relying on mobile-optimized networks. This creates a viable path for always-on, glint-free eye tracking in compact smart glasses.

*Equal contribution

Authors' Contact Information: Jipeng Sun, jipeng.sun@princeton.edu, Google Inc., Princeton University, USA; Lekang Yuan, ly7150@princeton.edu, Princeton University, USA; Ruslan Guseinov, ruslanguseinov@google.com, Google Inc., Austria; Bernd Bickel, berndbickel@google.com; Thabo Beeler, tbeeler@google.com, Google Inc., Switzerland; Thomas Auzinger, thomasauzinger@google.com, Google Inc., Austria; Felix Heide, fheide@princeton.edu, Princeton University, USA.

SA Conference Papers '26, Kuala Lumpur, Malaysia

ACM ISBN 979-8-4007-2842-6/2026/12
https://doi.org/10.1145/3829340.3842311

CCS Concepts: • **Computing methodologies** → **Computational photography**; *Neural networks*; • **Hardware** → **Emerging optical and photonic technologies**; • **Human-centered computing** → *Mixed / augmented reality*.

Additional Key Words and Phrases: Eye Tracking, Computational Optics, Computational Imaging, Metasurface, End-to-End Optimization, Smart Glasses, Point Spread Function Engineering

ACM Reference Format:
Jipeng Sun, Lekang Yuan, Ruslan Guseinov, Bernd Bickel, Thabo Beeler, Thomas Auzinger, and Felix Heide. 2026. Efficient Nano-Optical Eye Tracking for Smart Glasses. In *SIGGRAPH Asia 2026 Conference Papers (SA Conference Papers '26), December 01–04, 2026, Kuala Lumpur, Malaysia.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3829340.3842311

1 Introduction

The transition from handheld devices to always-on smart glasses is the next frontier in personal computing, and in this persistent Augmented Reality (AR) paradigm eye tracking plays the role of the desktop mouse cursor: the primary, continuous input for intent recognition and interface control. Unlike tethered Virtual Reality (VR) systems backed by powerful GPUs, smart glasses rely on constrained mobile processors that must simultaneously handle SLAM, display rendering, and hand tracking, so eye tracking runs as an "always-on" background service that must be robust at a minimal computational footprint without monopolizing the inference capacity of the primary AR applications.

Today's eye tracking devices largely rely on Pupil Center Corneal Reflection (PCCR) [Hansen and Ji 2009; Hutchinson et al. 1989], using active near-infrared (NIR) glints to disentangle eye rotation

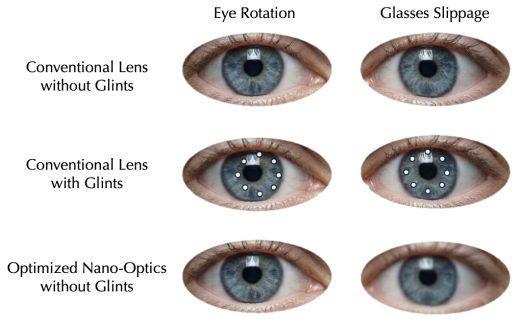


Fig. 2. **Eye Rotation vs. Glasses Slippage Ambiguity.** (Row 1) With a conventional lens, a rightward eye rotation without glasses movement looks similar to the glasses slipping sideways without eye movement. (Row 2) Glint-based trackers triangulate the true state from corneal reflections of device-mounted IR LEDs, which stay surface-locked under rotation but shift under slippage. (Row 3) Our depth-coded metalens resolves the ambiguity without glints through optical signatures that vary rapidly with depth, letting the network disentangle rotation from slippage in a single image.

from headset slippage. In smart glasses, however, frame-mounted LEDs impose steep illumination angles that are easily occluded by eyelashes or missed during rotation, and outdoor solar NIR further degrades the signal [Guestrin and Eizenman 2006]. Appearance-based alternatives remove the need for glints but inherit an ill-posed ambiguity between rotation and slippage, as Fig. 2 illustrates. This ambiguity can be mitigated either with multi-camera setups [Tonsen et al. 2017], at the cost of large footprint and power, or with over-parameterized networks that exceed the inference budget of always-on wearables [Cheng et al. 2024; Yu et al. 2019]. In both regimes, optics, sensor, and downstream network are designed in isolation: lenses are optimized for sharpness and discard depth cues, sensors are optimized for pixel count and flood the processor with redundant data, and networks are asked to resolve ambiguities that the optics could have removed physically. We instead treat these three stages as a single system.

We introduce a depth-sensitive, compute-efficient eye tracker built around a custom metasurface whose Point Spread Function (PSF) optically encodes 3D depth cues into the sensor image. This physical encoding disentangles eye pose from glasses slippage without a heavy inference network. We jointly optimize the metasurface, effective pixel pitch, and backbone architecture against a task-error + FLOPs objective, and validate the design in simulation and on a hardware prototype, matching the accuracy of the best glint-free refractive baseline with approximately 6× fewer floating-point operations. Depth-coded PSFs have precedent in microscopy and computational imaging [Guo et al. 2019; Pavani et al. 2009]; our contribution is their use as a task-specific physical encoder of the rotation–slippage ambiguity (Eq. 6), co-optimized with sensor pitch and network architecture under a compute budget (Eq. 10).

We make the following contributions in this paper:

- We present a new approach to disentangling eye rotation from glasses slippage for glint-free eye tracking using nano-optical depth encoding.
- We implement an end-to-end differentiable method that jointly optimizes meta-optics and network, balancing tracking accuracy against computational cost.
- We validate the method in simulation and on a physical prototype: the co-designed system reaches sub-degree gaze error relative to refractive glint-free baselines with 6× fewer floating-point operations.

2 Related Work

Eye Tracking. Eye tracking has evolved from invasive scleral coils [Young and Sheena 1975] to Video-Oculography, where Pupil Center Corneal Reflection (PCCR) remains the glint-based gold standard [Hansen and Ji 2009; Hutchinson et al. 1989]. Wearable trackers must resolve the ambiguity between eye rotation and sensor slippage [Guestrin and Eizenman 2006]: classical systems use active corneal glints or depth-based cues from RGB-D rigs or deflectometry [Funes Mora et al. 2014; Sun et al. 2015; Wang et al. 2025b], but these impose severe geometric constraints [Tonsen et al. 2017] and are easily disrupted by frame flex [Niehorster et al. 2020]. Appearance-based gaze estimation has matured from CNNs to Transformer backbones [Cheng et al. 2024; Kim et al. 2019; Krafka et al. 2016; Xiao et al. 2025; Zhang et al. 2015], but without explicit depth or glints the rotation–slippage ambiguity is paid for in FLOPs or in continuous domain adaptation [Yu et al. 2019]. Event sensors [Angelopoulos et al. 2020; Gallego et al. 2020] and MEMS raster-scan trackers [Meyer et al. 2020; Sarkar et al. 2017] cut bandwidth but sacrifice static fixation, mechanical robustness, or the image fidelity needed for tasks like iris authentication [Daugman 2009]. We pursue a passive, image-based method that offloads the spatial encoding of slippage and rotation to a depth-coded metalens.

Computational Optics and End-to-end Design. End-to-end optics couple a learnable optical element to a reconstruction network, encoding high-dimensional visual signal into a 2D measurement [Heide et al. 2016, 2013; Peng et al. 2015, 2019; Sitzmann et al. 2018; Tseng et al. 2021a,b; Wetzstein et al. 2020], and have enabled achromatic imaging, depth, HDR, and snapshot hyperspectral capture [Baek et al. 2021; Chakravarthula et al. 2023; Chang and Wetzstein 2019; Dun et al. 2020; Fröch et al. 2025; Guo et al. 2019; Ikoma et al. 2021; Jeon et al. 2019; Metzler et al. 2020; Shi et al. 2024a,b; Su et al. 2018; Sun et al. 2025, 2020], with fabrication-aware extensions [Wei et al. 2025; Zheng et al. 2023]. Recent near-eye nano-optics fold the optical path or trim camera protrusion [Kim et al. 2024; Yun et al. 2025], but focus on miniaturization rather than task-specific signal encoding and — importantly — do not treat compute cost as an objective.

Efficient Vision. On-sensor computing, sparse sampling, and NAS reduce inference cost in isolation [Baek et al. 2025; Cai et al. 2019; Feng et al. 2024; Wan et al. 2020; Wang et al. 2025a], while optical neural networks [Lin et al. 2018; Mennel et al. 2020; Wei et al. 2024] offload work into the optics but leave the sensor readout bottleneck untouched. Hardware-software co-designs such as *EyeCoD* [You et al. 2023] and *Minimalist Vision* [Klotz and Nayar 2025] couple frontend and backend but do not optimize the optics. We instead

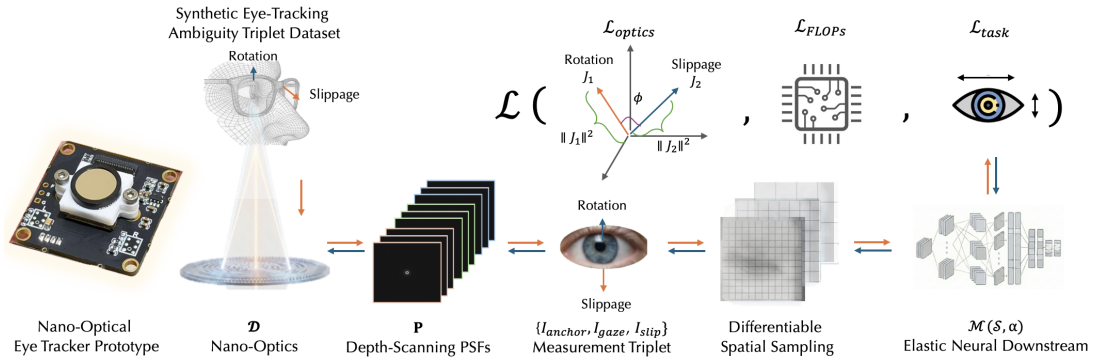


Fig. 3. **Efficient Nano-optical Eye Tracking.** We co-design the full eye tracking pipeline -— optics, sensor, and neural inference -— for eye tracking accuracy and compute cost. Specifically, we minimize a joint objective of task loss $\mathcal{L}_{\text{task}}$ and energy cost $\mathcal{L}_{\text{energy}}$. To offload computation to the optical domain with $\mathcal{L}_{\text{optics}}$, we use the Fisher Information Matrix (FIM) to penalize similarity of eye rotation and slippage after optical encoding. We adopt an elastic neural backbone to search for the minimum viable architecture subject to the compute budget. (Left) We built a hardware prototype to confirm the effectiveness of the method.

co-optimize metasurface phase, effective pixel pitch, and model architecture under a FLOPs budget.

3 Efficient End-to-End Eye Tracking

We describe a fully differentiable model for jointly designing the optics, sensor readout, and downstream processing of the proposed eye tracker (Fig. 3): the depth-aware image formation model and its metasurface depth encoding, the processing model and its compute cost, and the training strategy.

3.1 Depth-Aware Image Formation

We model light propagation from a point source of amplitude A_0 at lateral position (x, y) and depth z to the sensor plane; under the paraxial approximation, the incident field U_{in} at the metasurface plane is the spherical wave

$$U_{in}(x, y; z, \lambda) = A_0 \exp\left(-i\frac{2\pi}{\lambda}\sqrt{x^2 + y^2 + z^2}\right). \tag{1}$$

We design a single nanophotonic element for simultaneous focusing and encoding, composed of sub-wavelength nano-pillars parameterized by a diameter map $\mathbf{D} \in \mathbb{R}^{H \times W}$ [2021a]. We map these diameters to phase modulation ϕ_{meta} and transmission magnitude T via a proxy function

$$\phi_{meta}(x, y, \lambda), T(x, y, \lambda) = \mathcal{F}_{proxy}(\mathbf{D}(x, y); \lambda). \tag{2}$$

Here, $\mathcal{F}_{proxy}$ is a polynomial fitted to a library of achromatic nano-structures simulated via Rigorous Coupled-Wave Analysis (RCWA) [Liu and Fan 2012], see *Supplementary Material* for details.

The complex field leaving the metasurface is $U_{out} = U_{in} \cdot T \cdot e^{i\phi_{meta}}$. We propagate this field to the sensor plane using the Angular Spectrum Method (ASM) to obtain the Point Spread Function (PSF) for any depth z, that is

$$\text{PSF}(z, \lambda) = \left|\mathcal{F}^{-1}\left\{\mathcal{F}\{U_{out}\} \cdot H_{prop}(d_{sensor}, \lambda)\right\}\right|^2, \tag{3}$$

where d_{sensor} is the lens-sensor distance and H_{prop} is the free-space transfer function. To compensates for depth-dependent radiometric fall-off, we normalize the PSFs to maintain consistent signal energy $\widehat{\text{PSF}}(z) = \frac{\text{PSF}(z)}{\sum_{x,y}\text{PSF}(z)}$.

Image Synthesis with Distributed Field Sampling. We sample a dense PSF stack $\mathbf{P} = \{\widehat{\text{PSF}}(z_k)\}_{k=1}^{K}$ at ≤0.5 mm depth intervals with a data-parallel `AllGather` scheme [Sun et al. 2025], alpha-composite the scene radiance layers with $\mathbf{P}$ following an occlusion-aware model [Ikoma et al. 2021], and apply a differentiable sensor model that includes a learnable box filter $\mathcal{W}_{\hat{p}}$ (effective pixel pitch $\hat{p}$), energy-preserving factor $\kappa = (\hat{p}/p_{base})^2$, and Gaussian + Poisson noise. Full derivation in *Supplementary Material* Sec. A.

3.2 Separable Optical Encodings via Fisher Information

Eye rotation (*gaze*) and glasses slippage (*slip*) often induce linearly dependent intensity variations on the sensor, and resolving this ambiguity with priors alone requires complex, power-hungry networks. Instead, we optimize the optical frontend as a physical encoder that yields an inherently separable representation: the metalens phase profile is a learnable parameter that maximizes the *Fisher Information* with respect to the motion parameters $\Theta = (\theta_1, \theta_2)$, with θ_1 the gaze rotation and θ_2 the device slippage.

We analyze the sensitivity of the sensor image $\mathbf{I}$ to changes in these parameters. We define the sensitivity field (or Jacobian image) for the i-th parameter as $\mathbf{J}_i = \frac{\partial \mathbf{I}}{\partial \theta_i}$. Since the physical rendering process is non-differentiable with respect to scene coordinates, we approximate these fields using a finite difference triplet strategy. In each iteration, we render a triplet of images $\{I_{anchor}, I_{gaze+}, I_{slip+}\}$ to compute the discrete gradients, which we describe in detail in Sec 3.6. The elements of the Fisher Information Matrix (FIM), $\mathbf{F} \in \mathbb{R}^{2\times 2}$, are then computed as the inner product of these sensitivity fields over the sensor pixel domain Ω

$$\mathbf{F}_{ij}(\Theta) \approx \langle \mathbf{J}_i, \mathbf{J}_j \rangle = \sum_{p \in \Omega}\left(\frac{\partial \mathbf{I}_p}{\partial \theta_i} \cdot \frac{\partial \mathbf{I}_p}{\partial \theta_j}\right), \tag{4}$$

where indices $i, j \in \{1, 2\}$ refer to the gaze and slip parameters, respectively.

Geometrically, $\mathbf{J}_1, \mathbf{J}_2 \in \mathbb{R}^{|\Omega|}$ are vectors in pixel space: the diagonals $\mathbf{F}_{ii} = \|\mathbf{J}_i\|^2$ measure the sensitivity to each motion and the off-diagonal $\mathbf{F}_{12}$ their alignment, with span angle

$$\cos\phi = \frac{\mathbf{F}_{12}}{\|\mathbf{J}_1\|\|\mathbf{J}_2\|} = \frac{\mathbf{F}_{12}}{\sqrt{\mathbf{F}_{11}\mathbf{F}_{22}}}. \tag{5}$$

As illustrated in Fig. 3, we seek large $\|\mathbf{J}_i\|$ (high sensitivity) and $\phi \to 90°$, so that rotation and slippage are encoded along orthogonal directions in pixel space (*Supplementary Material* Sec. A.5).

Optical Embedding Loss. As an optical embedding loss, we use a compound loss function designed to maximize information content while favoring disentanglement

$$\mathcal{L}_{optics}(\phi_{opt}) = -\sum_i \log(\mathbf{F}_{ii}) + \lambda_{orth}\sum_{i\neq j}(\mathbf{F}_{ij})^2 + \lambda_e\,(1 - E_{\text{encircled}}). \tag{6}$$

The three terms encode *sensitivity*, *orthogonality*, and *compactness*: $-\sum_i \log \mathbf{F}_{ii}$ maximizes the gradient magnitudes so that minute motions yield intensity changes well above the sensor noise floor; $\sum_{i\neq j}\mathbf{F}_{ij}^2$ drives the off-diagonal Fisher entries to zero, decorrelating the encoded motions; and $(1 - E_{\text{encircled}})$ penalizes PSF energy dispersion to preserve the per-pixel SNR for the downstream detector. See *Supplementary Material* Sec. A for details.

3.3 Neural Reconstruction Model

To recover gaze efficiently, we search over sub-networks of a unified supernet [Cai et al. 2019], each trading network capacity against compute.

Architecture. We construct an elastic backbone $\mathcal{B}$ based on the edge-friendly MobileNetV3 [Howard et al. 2019]. To resolve the ambiguity between sensor slippage and eye rotation, we bifurcate the architecture into two functionally orthogonal heads. Given an input sensor image $\mathbf{I}$, the backbone extracts a shared feature map $\mathbf{F} = \mathcal{B}(\mathbf{I}; \theta_{shared})$.

The first branch, *Perception Head* ($\mathcal{H}_{perc}$), employs a lightweight decoder to predict a binary segmentation mask $\mathbf{M} \in \mathbb{R}^{H\times W}$, isolating the 2D pupil centroid $\mathbf{u} \in \mathbb{R}^2$ independent of scene depth. The second branch, *Calibration Head* ($\mathcal{H}_{calib}$), flattens the feature map to retain coordinate awareness, employing a regression layer to decode the camera displacement vector $\Delta\mathbf{t} \in \mathbb{R}^3$ (slippage).

$$\Delta\mathbf{t} = \mathcal{H}_{calib}(\text{Flatten}(\mathbf{F})). \tag{7}$$

This regression is physically grounded in two cues: (1) lateral slippage $(\Delta x, \Delta y)$ is inferred from the translation of static periorbital anchors (e.g., the medial canthus) relative to the sensor frame [Sesma et al. 2012], and (2) axial slippage (Δz) is decoded from the depth-dependent Point Spread Function (PSF) of the metasurface, which modulates the texture sharpness of these anchors as a function of camera-to-eye distance.

The pupil centroid $\mathbf{u}$ and slippage $\Delta\mathbf{t}$ are fused into gaze via a deterministic spherical-eye solver (*Supplementary Material* Sec. A.4) — a slippage-corrected ray from $\mathbf{u}$ is intersected with an eye sphere of radius R to obtain the optical axis $\mathbf{g}_{opt}$ — followed by a lightweight residual regressor $\mathcal{R}(\mathbf{g}_{opt}, \Delta\mathbf{t}; \theta_{reg})$ that predicts the subject-specific angle-kappa offset to the visual axis $\mathbf{g}_{vis}$. We supervise with a Dice segmentation loss on the predicted pupil mask and an ℓ_2 loss on the predicted slippage vector, $\mathcal{L}_{task} = \mathcal{L}_{seg} + \lambda_{slip}\mathcal{L}_{calib}$, see *Supplementary Material* Sec. A for the details.

We share weights across three elastic axes: kernel size $k \in \{3, 5, 7\}$ (quadratic in depthwise cost), depth d (linear in layer count), and width w (quadratic in pointwise cost). This spans a family of subnets trading accuracy against compute.

Sandwich Rule Training. Training each sub-network independently is infeasible, so we train the supernet as a stochastic weight-sharing graph with the "sandwich rule" [Cai et al. 2019]: at every step we optimize the largest sub-network ($\mathcal{N}_{max}$), the smallest ($\mathcal{N}_{min}$), and sampled intermediate ones ($\mathcal{N}_{sample}$), accumulating gradients into the shared weights θ and the optical parameters ϕ as

$$\mathcal{L}_{train} = \mathcal{L}_{task}(\mathcal{N}_{max}) + \mathcal{L}_{task}(\mathcal{N}_{min}) + \sum_{i=1}^{n}\mathcal{L}_{task}(\mathcal{N}_{sample}^{(i)}). \tag{8}$$

$\mathcal{N}_{max}$ and $\mathcal{N}_{min}$ anchor the upper and lower accuracy bounds; since the metasurface receives gradients from all sub-networks at once, it becomes a "generalist" encoder that preserves the high-frequency detail $\mathcal{N}_{max}$ exploits while keeping the contrast $\mathcal{N}_{min}$ needs—a robust warm start for the specializations in Sec. 3.5 (*Supplementary Material* Sec. B.4).

3.4 Computational Cost Model

We model the per-inference cost C_{total} as a differentiable function of two coupled variables: the effective pixel pitch $\hat{p}$, which sets the input resolution $H_sW_s \propto \hat{p}^{-2}$ via physical readout or digital resizing [Sze et al. 2017], and the active subnet $\alpha = \{w, d, k\}$, which sets the network capacity [Tan and Le 2019]. Summing the floating-point operations of the MobileNetV3 inverted-residual blocks [Howard et al. 2019; Ma et al. 2018],

$$C_{\text{total}}(\hat{p}, \alpha) = \sum_{\ell=1}^{d}\left[\text{Ops}_\ell^{\text{pw}}(\cdot) + \text{Ops}_\ell^{\text{dw}}(\cdot) + \text{Ops}_\ell^{\text{se}}(\cdot)\right], \tag{9}$$

where the terms are the pointwise expansion/projection, depthwise convolution, and squeeze-and-excitation operations, respectively. This lets the co-design trade pixel density against depth and width (*Supplementary Material* Sec. B.5).

3.5 Bi-Level Optimization

We formulate the co-design as a bi-level optimization (Fig. 4) that decouples the discrete architecture $\alpha \in \mathcal{A}$ from the continuous parameters $\theta = \{\phi_{opt}, \hat{p}, \mathbf{w}_{net}\}$. For a fixed α, the inner loop solves a budget-penalized compound loss that couples the task loss with the Fisher-information optics loss of Eq. 6,

$$\theta^*(\alpha) = \underset{\theta}{\text{argmin}}\left[\,\mathcal{L}_{task} + \lambda_o\,\mathcal{L}_{optics} + \lambda_c\left[\,C_{total}(\alpha, \hat{p}) - C_{budget}\,\right]_+\right], \tag{10}$$

where $[\,\cdot\,]_+ = \max(0, \cdot)$ is the budget penalty, $\mathcal{L}_{optics}(\phi_{opt})$ shapes the learned PSF directly through the Fisher information at the sensor (Eq. 6), and $\mathcal{L}_{task}(\alpha, \theta)$ supervises downstream gaze accuracy. We warm-start θ from the pre-trained supernet $\mathcal{S}$ and fine-tune for 3 epochs via differentiable wave propagation and sensor integration [Cai et al. 2019; Pham et al. 2018; Sitzmann et al. 2018]. The outer loop then maximizes $\mathcal{V}(\alpha, \theta^*(\alpha))$ over $\mathcal{A}$ with a Regularized Evolutionary Algorithm (REA) [Real et al. 2019], and sweeping

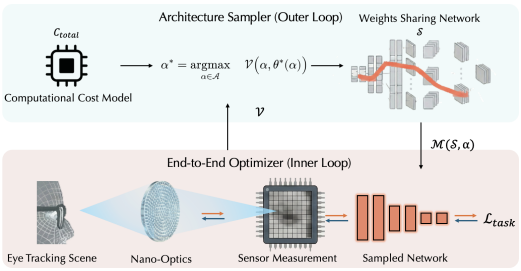


Fig. 4. **Bi-level Optimization.** Outer loop: a Regularized Evolutionary Algorithm (REA) samples architectures from a weight-sharing Supernet $\mathcal{S}$ under a compute budget and takes validation feedback from the inner loop. Inner loop: the sampled architecture inherits Supernet weights and is fine-tuned end-to-end to jointly optimize the optics, sensor pitch, and network weights.

C_{budget} yields a Pareto set of accuracy–efficiency configurations. Full outer objective, weight-inheritance projection $\mathcal{M}(\mathcal{S}, \alpha)$, and REA hyperparameters are provided in *Supplementary Material* Sec. B.

3.6 Synthetic Eye-Tracking Data

Training and evaluating our co-designed optics requires precise supervision of both gaze vectors and sensor displacements across a diverse user population, which we obtain from synthetic data. We render subjects via physically-based path tracing (Blender Cycles [Blender Online Community 2025]) of parametric face and eye models registered against One-Light-At-A-Time (OLAT) scans of a diverse population [Sarkar et al. 2023]. Each frame is produced from a random subject and a sampled 6-DoF state vector $\mathbf{s} = (\phi, \theta, \omega, t_x, t_y, t_z)$ that jointly encodes eye rotation (ϕ, θ with Listing's torsion ω [Straumann et al. 1996; Tweed and Vilis 1990]) and glasses slippage (translation along the nasal bridge plus a small mounting perturbation). The pipeline emits the discretized radiance layers I_k consumed by our forward model (Sec. 3.2) together with per-frame ground-truth landmarks, gaze directions, and pupil-segmentation masks; representative samples are shown in Fig. 6.

From this pipeline we construct two task-specific subsets. A *variational training set* emits Fisher-information triplets $(I_{\text{anchor}}, I_{\text{gaze+}}, I_{\text{slip+}})$ whose small gaze and slippage perturbations stay in the local linear regime of the intensity function, so that finite differences approximate the sensitivity fields $\nabla_{\text{gaze}} I$ and $\nabla_{\text{slip}} I$ used by the optical objective (Eq. 4). A *projection-ambiguity* evaluation set then enforces pairs $(I_{\text{slip}}, I_{\text{gaze}})$ in which distinct slippage and rotation states project to the *same* 2D pupil centroid on the sensor, so that disentanglement must rely on the depth-encoded PSF rather than on lateral pupil position. Full rendering pipeline, triplet equations, projection-equivalence construction, and per-parameter sampling distributions are provided in *Supplementary Material* Sec. C.

4 Analysis

We evaluate the proposed design in simulation along four axes: optical encoding separability (Sec. 4.1), PSF depth encoding (Sec. 4.2), the accuracy–compute Pareto frontier (Sec. 4.3), and co-design ablations (Sec. 4.4), using the projection-ambiguity dataset (*Supplementary Material* Sec. C).

Table 1. **Quantitative Encoding Sensitivity and Separability.** Average absolute differential sensitivity and separability energies across the dataset. The proposed lens achieves the highest signal energy across all metrics, indicating robust encoding.

Lens Design	Gaze Sensitivity $\|I_{\text{gaze}} - I_{\text{anc}}\|^2$	Slip Sensitivity $\|I_{\text{slip}} - I_{\text{anc}}\|^2$	Separability $\|I_{\text{gaze}} - I_{\text{slip}}\|^2$
Hyperbolic	3.25	17.82	16.98
Double Helix	4.99	15.75	14.45
Proposed	**6.00**	**28.47**	**26.60**

4.1 Evaluation of Optical Encoding

We benchmark our lens against two fixed baselines at matched working distance (details in *Supplementary Material* Secs. D and E): a *hyperbolic* lens that encodes depth only by defocus blur, and a *double-helix* lens [Pavani et al. 2009] whose PSF rotates across the 16–26 mm depth range. We measure two quantities on the anchor/gaze/slip triplet set (*Supplementary Material* Sec. C): *sensitivity* $\|I_{\text{pert}} - I_{\text{anc}}\|^2$ and *separability* $\|I_{\text{gaze}} - I_{\text{slip}}\|^2$ — high separability means rotation and slippage leave distinct optical signatures at the sensor.

Fig. 5 (rows 1–2) shows that the hyperbolic and double-helix baselines produce low-energy differential maps, i.e., nearly colinear measurements for gaze and slippage, whereas the proposed lens (row 3) reshapes the PSF into distinct patterns for each motion, decoupling the parameters before digitization. Table 1 confirms this quantitatively: our design nearly doubles the separability of the double-helix baseline (26.60 vs. 14.45).

4.2 Evaluation of PSF Depth Encoding

We next characterize the depth encoding directly on the PSF, which must be sensitive to axial displacement (for slippage) while remaining laterally compact (for gaze).

We quantify axial sensitivity with depth-specific Fisher Information, $\mathcal{I}(z)$, computed via the finite difference of the PSF intensity P with respect to axial depth z

$$\mathcal{I}(z) = \sum_{u,v} \left(\frac{\partial P(u,v;z)}{\partial z} \right)^2 \approx \sum_{u,v} \left(\frac{P(u,v;z+\Delta z) - P(u,v;z)}{\Delta z} \right)^2, \tag{11}$$

where $\Delta z = 0.2$ mm. We compare the proposed method against the *Hyperbolic* and *Double Helix* lens across the 16–26 mm depth range. This window encompasses the 3-σ pupil-depth distribution derived from large-scale demographic scans ([Sarkar et al. 2023], Sec. 4.1) mapped to a smart-glasses rig, so operating over this range naturally accommodates pupil-depth variation across demographics.

Fig. 7a visualizes the three encodings: the hyperbolic lens spreads energy by defocus and degrades lateral resolution, the double helix rotates and aliases for pixel binning, while our design combines shape aberration (16–21 mm) with intensity gradients (21–26 mm) and stays laterally compact. This is reflected in Fig. 7b: our axial Fisher information $\mathcal{I}(z)$ stays high across 1×1, 2×2, 3×3 binning, whereas the double-helix information collapses under binning and the hyperbolic is limited away from focus — so the learned optics

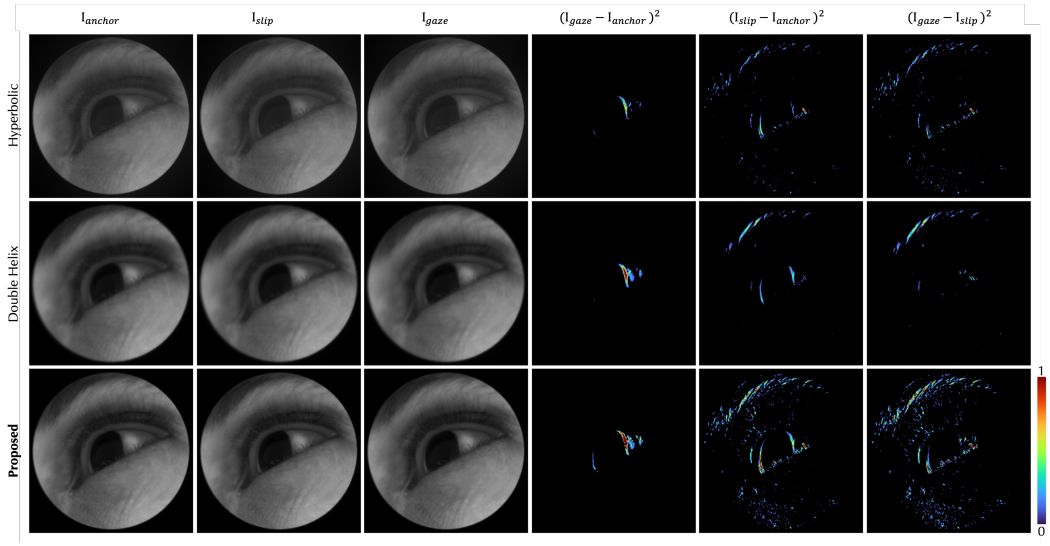


Fig. 5. **Effect of Optical Encoding.** Cols. 1–3 show sensor measurements for our design vs. Hyperbolic and Double Helix baselines on an ambiguous scene (16.7–38.5 mm); cols. 4–6 report differential sensitivity $(I_{\text{gaze}} - I_{\text{anchor}})^2$ and separability $(I_{\text{gaze}} - I_{\text{slip}})^2$. Baseline maps are dim (poor separability), whereas ours produces bright, distinct patterns that decouple rotation from slippage at the sensor level.

keep depth information in features that survive coarse sensor readout. X–Y PSF projections at more depths and resolutions are in the *Supplementary Material*.

Since a smart-glasses camera views the eye obliquely and the pupil traverses the sensor field, Tab. 3 reports $\mathcal{I}(z)$ averaged over 16–26 mm for simulated PSF stacks at field angles of 5–20°: the learned design retains 2.5–5× the axial information of both baselines at every angle, so the depth encoding does not collapse off-axis.

4.3 Prediction Gaze Error vs. Computational Cost

We benchmark accuracy vs. compute on the ambiguity dataset (Sec. 3.6) against fixed hyperbolic and double-helix baselines, using the shared supernet "sandwich" scheme over 100 REA-selected subnetworks (5 stages; kernels $\{3, 5, 7\}$, expansion $\{3, 4, 6\}$, depths $\{2, 3, 4\}$) with continuous resolution scaling in $[0.3\times, 1.0\times]$ and per-configuration batch-norm recalibration. Compute is reported in multiply-accumulate operations (MACs) [Howard et al. 2017] (*Supplementary Material* Sec. B.6).

Our co-designed optic produces a uniformly superior accuracy–compute Pareto frontier (Fig. 8a): matching peak baseline accuracy with roughly half the compute, reducing error by $\sim 0.5°$ at a small 120-MMAC budget, and still improving over the MobileNetV3-MaxNet baseline at maximum capacity. Jointly optimizing lens, sensor resolution, and network therefore yields both efficiency and accuracy gains. The advantage persists at low SNR: with attenuated illumination and an intense ambient background injected before Poisson shot noise (15 dB SNR, emulating outdoor NIR flooding or low light), our 480-MMAC subnet degrades gracefully from 0.75° to 1.12°, whereas the matched hyperbolic and double-helix baselines reach 2.58° and 2.84°, as the nearly doubled signal energy (Tab. 1) keeps the encoded contrast above the shot-noise floor.

Comparison to Deep Eye Tracking. To contextualize the co-design, we benchmark three learning-based baselines on our 3D-slippage ambiguity dataset, reporting 3D gaze error jointly with compute (Tab. 2). NVGaze [Kim et al. 2019] is a lightweight, anatomically-informed CNN for near-eye gaze trained on synthetic and real captures under conventional optics; it is fast (70 MMACs) but still maps a single camera image to gaze without optical disambiguation of device motion, so camera slippage entangles with eye rotation and error rises to 3.81°. Popovic et al. [Popovic et al. 2023] introduce *model-aware* 3D gaze from weak and few-shot supervision; the differentiable 3D eyeball still explains appearance through *fixed* imaging, so depth–slippage ambiguities in the conventional projection remain and their estimator reaches 1.23° only at ∼13.2 GMACs. De^{2}Gaze [Xiao et al. 2025] is a recent representation-learning method (deformable sparse attention + decoupled latent factors for 3D gaze from eye crops) at 0.69° and 6.78 GMACs. At 480 MMACs our pipeline reaches 0.75° — comparable error at ∼14× less compute than De^{2}Gaze — and at 2.5 GMACs we reach 0.50°, well below De^{2}Gaze at under half its MAC count. Jointly optimizing optics, resolution, and decoding moves part of the slippage–rotation disambiguation into the measurement, which is why we improve both axes relative to these purely computational pipelines. To separate optical from architectural gains, we retrained the De^{2}Gaze backbone on images formed with our metalens: its error drops from 0.69° to 0.53° with the optics alone and to 0.47° when optics and sensor resolution are additionally unfrozen for end-to-end co-optimization, so the optical and co-design gains compound.

4.4 Ablation Experiments

We ablate in three co-design axes by freezing one at a time under the same bi-level search (Fig. 8b, Sec. 3.5). Freezing the optics to a hyperbolic lens collapses the Pareto frontier — the network cannot

Table 2. **Comparison to Deep Eye Tracking Baselines.** 3D angular gaze error and compute (MMACs) on our 3D-slippage ambiguity dataset; NVGaze [Kim et al. 2019], Popovic et al. [Popovic et al. 2023], and De^{2}Gaze [Xiao et al. 2025] follow their published formulations.

Method	NVGaze [2019]	Popovic et al. [2023]	De^{2}Gaze [2025]	Ours (eff.)	Ours (acc.)
Gaze Error (°)	3.81	1.23	0.69	0.75	**0.50**
Compute (MMACs)	70	13220	6780	**480**	2500

Table 3. **Off-Axis Depth Encoding.** Axial Fisher information $\mathcal{I}(z)$ (Eq. 11) of simulated PSF stacks averaged over 16–26 mm at off-axis field angles; the learned encoding retains the highest value at every angle.

Field angle	5°	10°	15°	20°
Hyperbolic	0.24	0.24	0.31	0.33
Double Helix	0.31	0.31	0.33	0.30
Proposed	**1.12**	**1.25**	**0.92**	**0.81**

resolve the ambiguity within 1500 MMACs. Fixing the pixel pitch to $\hat{p} = 3\,\mu$m (OVM6211 [OMNI 2015]) inflates early-layer MACs by ≈3× with no accuracy gain. Freezing the architecture to max-capacity MobileNetV3 pushes the optimal resolution upward [Tan and Le 2019] and closes off the low-compute region. Jointly designing over all three axes reaches the most favorable system.

5 Experimental Assessment

5.1 Experimental Setup and Data Acquisition

Experimental Prototype. We fabricate the proposed co-optimized metasurface and a hyperbolic baseline on the same 0.5 mm fused-silica / 700 nm amorphous-silicon stack with a 1 mm circular aperture, using two-step electron-beam lithography (aperture + nanopillar etch) on a pseudo-Bosch recipe. Full fabrication protocol in *Supplementary Material* Sec. D.

Acquisition. We captured a controlled-slippage dataset using the experimental setup in Fig. 9. Each subject's head was fixed while viewing a 7×7 stimulus grid spanning a 25° × 35° field of view on a calibrated display [Chai et al. 2025]. The prototype was mounted ≈2 cm below the eye and tilted ≈45° upward toward it, emulating the oblique camera placement of a smart-glasses frame, and translated by motion stages to emulate seven camera slippage states sharing the same origin: the origin state, two lateral shifts along d_x, and four axial shifts along d_z. This yields 7×49 samples per prototype. In total, we collected data from 20 subjects. The study protocol was approved by the Princeton University Institutional Review Board, and all participants gave informed consent. Full display geometry and per-trial protocol are provided in *Supplementary Material* Sec. D.

5.2 Experimental Eye Tracking Evaluation

Evaluation of Pupil Segmentation and Slippage Prediction. We evaluate the co-optimized network on the experimental dataset (8,219 frames; pupil masks from SAM-3 [Carion et al. 2025] with manual inspection; slippage from the controlled camera translations). The perception head reaches IoU = 0.94 with sub-pixel pupil-center localization across subjects and slippage states (Fig. 9). Projected into the camera frame, the same on-screen stimulus maps to distinct pupil locations under each slippage state (Fig. 10), confirming that pupil-only estimation is ambiguous without depth cues; our calibration head resolves this, predicting slippage with 0.52 mm MAE vs. 0.95 mm for the hyperbolic baseline.

Evaluation of Eye Tracking Accuracy. We next evaluate eye-tracking accuracy under controlled camera slippage. A regressor (a fully connected network with one hidden layer of 10 neurons) maps pupil parameters and the estimated slippage vector to the on-screen stimulus location; it is trained on 20% of the points and evaluated on the remaining 80%, aggregated over all evaluated points. As shown in Fig. 11, under axial (d_z) slippage, our method achieves a mean gaze error of 1.17°, outperforming the hyperbolic baseline of 2.02°. Under lateral (d_x) slippage, which is easier to predict for both designs, our method still achieves lower error than the baseline, with 1.26° versus 1.49°. These results show that the proposed nano-optical design improves gaze prediction accuracy under both axial and lateral camera slippage. The *Supplementary Video* shows the prototype tracking gaze in real time during a gaze-controlled game under induced camera slippage. Sec. 6 discusses the gap to simulation and the lateral-slippage margin.

6 Discussion

Passive vs. active tracking. Unlike active PCCR, the passive encoding needs no glint sources and their geometric arrangement, and it is not bound by the eye-safety illumination budget that washes out glints under outdoor solar NIR (*Supplementary Material* Sec. D.5); the rotation–slippage ambiguity is resolved once in the optics and decoded at 480 MMACs. In exchange, the depth cue lives in a learned PSF: the tracker still needs flood NIR illumination, a one-time per-device calibration absorbing fabrication and assembly deviations of the realized PSF, and, like any tracker, a per-user calibration to the visual axis, which here fits the gaze regressor on only 20% of the stimulus points (Sec. 5.2). Drift-free operation under continuous wear remains to be shown.

Integration in smart glasses. The prototype PCB is a prototyping platform, not an integrated module: a 0.5 mm NIR filter, 0.3 mm metalens substrate, ≈2 mm lens–sensor spacing, ≲0.5 mm OG0TB-class sensor package, and ≈0.2 mm flexible circuit stack to ≈3.5 mm thickness with a sub-2 mm lateral footprint, within a 4–6 mm frame cross-section, with sub-millimeter NIR LEDs on the flex. Our captures already use the oblique geometry of such a mount (Sec. 5.1) and Tab. 3 shows that the depth encoding is retained across the field, but validating the tracker in a worn device under free head motion and variable lighting remains future work.

Trade-offs. The metalens has NA ≈ 0.24, set by the 2×2×4 mm module volume and the 16–26 mm pupil-depth range. A higher NA strengthens defocus and axial Fisher information (helping z-slippage) but enlarges the defocused PSF and off-axis aberrations, reducing lateral Fisher information and the usable field of view; Eq. 6 therefore targets 3D separability rather than axial sensitivity alone. Lateral slippage is rarer in well-fitted frames, where nose pads and temples constrain sliding, and appears mostly as a global translation predictable from periorbital features (Sec. 3.3), so both designs handle it better than axial slippage; at our 5 mm lateral shift,

larger than expected in practice, the margin narrows to 1.26° vs. 1.49°.

Limitations. Our evaluation is a head-fixed indoor benchtop study with motion-stage slippage, chosen because repeatable millimeter-level slippage and reliable gaze labels are hard to obtain on a worn device. The experimental error (1.17°) exceeds the simulated one (0.75°) mainly because the ground truth comes from instructed fixations, so fixation uncertainty and residual saccades count as tracking error, and because fabrication and assembly tolerances blur and shift the realized PSF; the hyperbolic baseline degrades alike. Finally, electron-beam lithography does not scale to volume production, although nanoimprint lithography was recently shown feasible for smart-glasses optics [Choi et al. 2025].

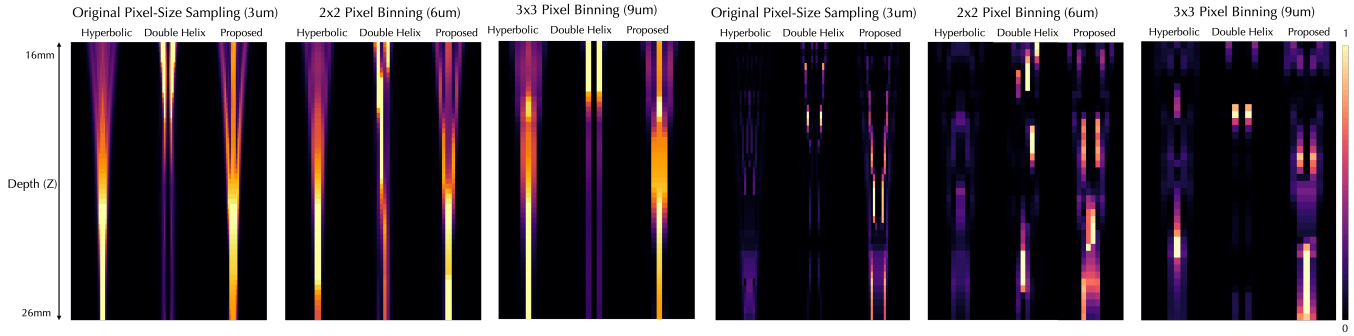
7 Conclusion

We propose an efficient eye tracker that co-designs optics, sensor, and neural processing: a nano-photonic metasurface optically decouples gaze from glasses slippage, resolving the tracking ambiguity in the optics instead of in heavy computation. Simulation and experiments confirm a 6× lower computational cost than existing methods at robust accuracy under camera slippage. Future work will extend the co-design to neuromorphic and event-based sensing for ultra-fast, robust eye tracking.

Acknowledgments

We are grateful to the team of the Nanofabrication Facility at the Institute of Science and Technology Austria, especially Rodolfo Previdi and Salvatore Bagiante, for their indispensable help in developing the nanofabrication recipe and fabricating the samples. We thank Congli Wang for organizing the eye-safety evaluation materials for Institutional Review Board (IRB) approval. We also thank Shreyas Spotnis, Ivana Tosic Rodgers, Jay Chen, Yousef Ibrahim, Andrew Logan, Swati Jindal, and Gordon Wan for helpful discussions.

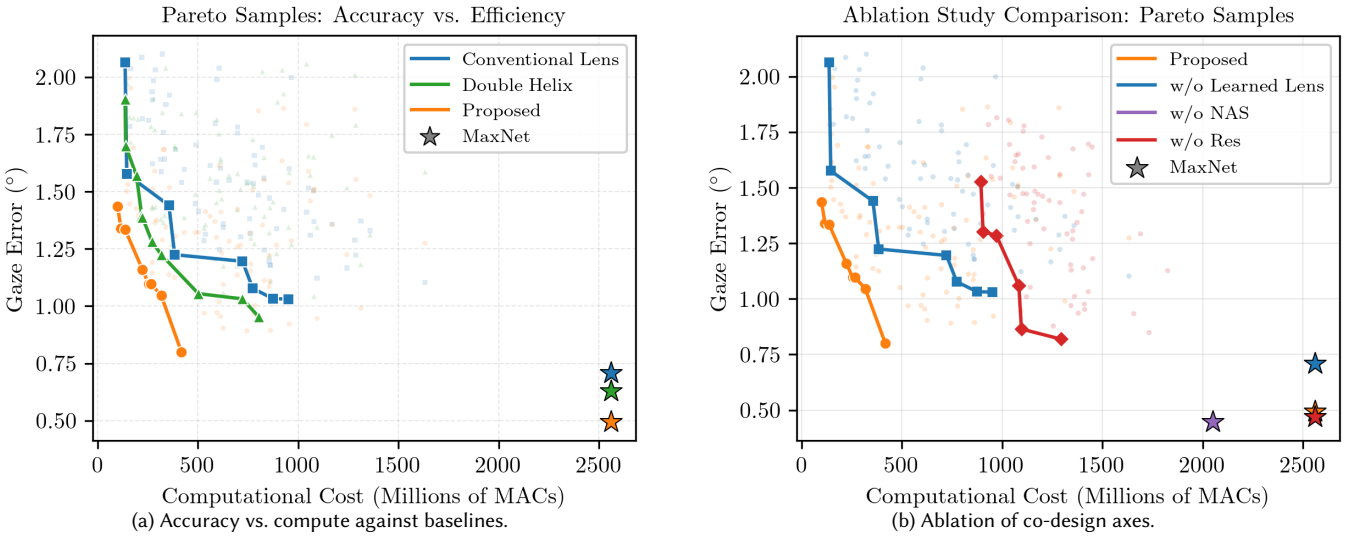
Fig. 6. **Synthetic Eye Tracking Dataset Samples.** 60 representative samples from our synthetic dataset, rendered by physically-based path tracing of parametric face/eye models from random state vectors **s**; the resulting variation in subject identity, gaze, and slippage is essential for training the optical encoder to disentangle eye pose from device motion.



(a) X-Z PSF slices across depths and pixel binnings

(b) X-Z PSF Fisher information across depths and pixel binnings

Fig. 7. **PSF Analysis of the Learned Optical Encoding.** (a) X–Z PSF slices for hyperbolic, double-helix, and our design across 16–26 mm and three pixel binnings (1×1, 2×2, 3×3): hyperbolic encodes depth by defocus (lateral spread), double-helix by rotation (which breaks under binning), while ours combines shape aberration (16–21 mm) and intensity gradients (21–26 mm) and stays laterally compact at every binning. (b) The corresponding per-depth-normalized Z-axis Fisher information $\bar{I}(z)$ (Eq. 11) confirms that our design retains high information density across all binnings, whereas the hyperbolic (focal-plane-only) and double-helix (resolution-dependent) baselines do not.



(a) Accuracy vs. compute against baselines.

(b) Ablation of co-design axes.

Fig. 8. **Accuracy–Compute Pareto Frontier.** (a) Angular gaze MAE vs. MMACs for hyperbolic, double-helix, and our co-designed lens; dots are individual architecture–resolution configurations and lines are Pareto frontiers. Our design matches peak baseline accuracy at ∼50% compute and reduces error by ∼40% at the 120-MMAC budget. (b) Pareto frontiers of our full pipeline vs. fixed-optics, fixed-pixel-pitch, and fixed-architecture ablations: the full co-design dominates at every budget (7× fewer MMACs than fixed-optics at matched accuracy and 3× fewer than fixed-pitch), confirming the benefit of jointly learning over all three axes.

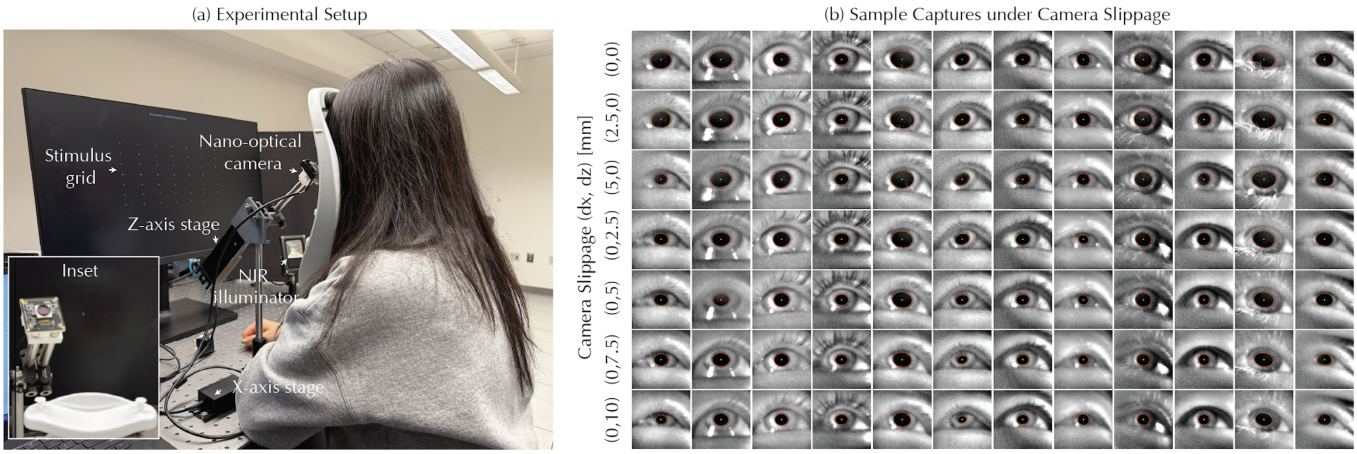


Fig. 9. **Experimental Setup and Capture Samples.** (a) Head-stage-fixed subject views stimuli on a monitor; a nano-optical prototype below the eye emulates near-eye AR. Controlled slippage is induced by translating the camera along its optical axis (z slippage) and along a direction perpendicular to the optical axis (x slippage) in 2.5 mm steps. Left inset: first-person view. (b) Captured eyes across subjects and different x-z slippage states; ground truth (green) vs. prediction (red) show that the energy-efficient network segments pupils reliably across conditions.

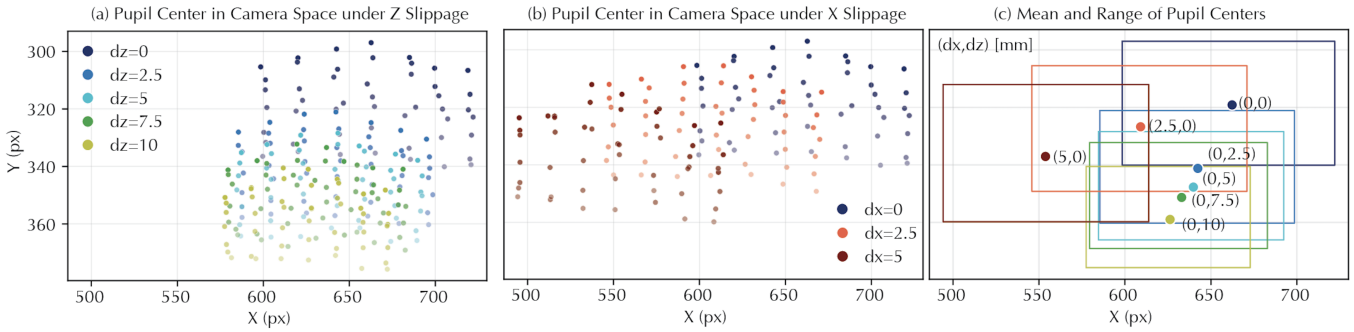


Fig. 10. **Slippage-Induced Ambiguity.** Slippage-induced pupil-center ambiguity. (a) Pupil-center projections under axial camera slippage, color-coded by d_z. Different opacity levels indicate different stimulus rows on the display. (b) Pupil-center projections under lateral camera slippage, color-coded by d_x. (c) Mean and range of pupil centers for each slippage state (d_x, d_z) in millimeters. The same gaze angle can lead to different pupil-center locations under different camera slippage states, causing slippage-induced ambiguity.

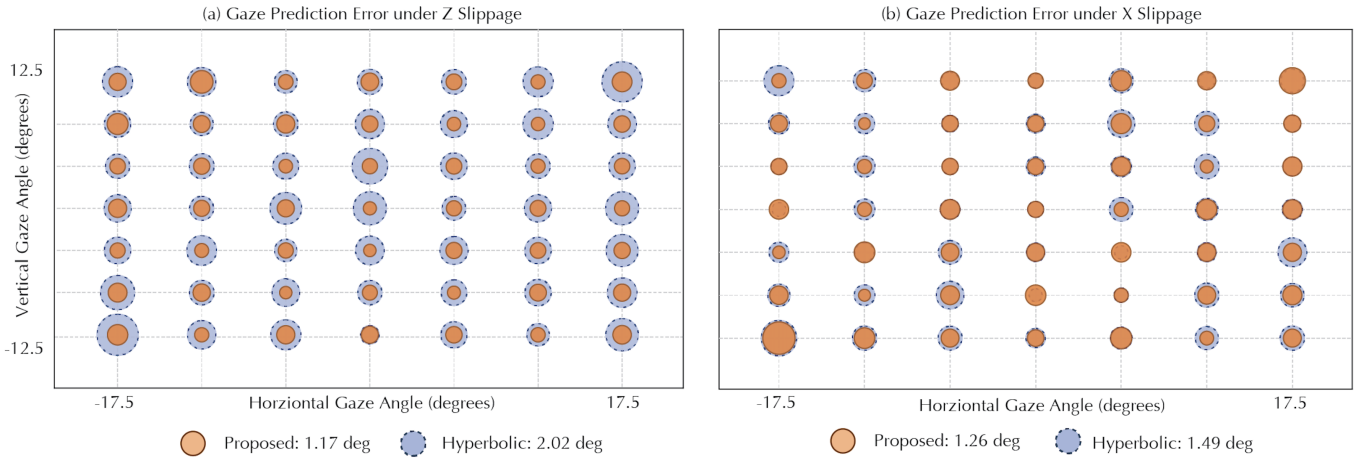


Fig. 11. **Gaze prediction accuracy under camera slippage.** Per-target angular gaze error for the co-optimized metasurface design (orange) and hyperbolic baseline (blue), with marker size indicating error magnitude. (a) Under axial (d_z) slippage, our method substantially reduces the mean gaze error from 2.02° to 1.17°. (b) Under lateral (d_x) slippage, which is easier to predict for both designs, our method still achieves lower error than the baseline.

References

Anastasios N Angelopoulos, Julien NP Martel, Amit PS Kohli, Jorg Conradt, and Gordon Wetzstein. 2020. Event based, near eye gaze tracking beyond 10,000 hz. *arXiv preprint arXiv:2004.03577* (2020).

Seung-Hwan Baek, Hayato Ikoma, Daniel S Jeon, Yuqi Li, Wolfgang Heidrich, Gordon Wetzstein, and Min H Kim. 2021. Single-shot Hyperspectral-Depth Imaging with Learned Diffractive Optics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2651–2660.

Yongmin Baek et al. 2025. Edge Intelligence through In-Sensor and near-Sensor Computing for the Artificial Intelligence of Things. *Npj Unconventional Computing* 2, 1 (2025), 25.

Blender Online Community. 2025. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam. https://www.blender.org

Han Cai, Chuang Gan, Tianzhe Wang, Zhekai Zhang, and Song Han. 2019. Once-for-all: Train one network and specialize it for efficient deployment. *arXiv preprint arXiv:1908.09791* (2019).

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. 2025. Sam 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719* (2025).

Ruixiang Chai, Jipeng Sun, and Jack Tumblin. 2025. Camera Against Display: An Automated and Precise Mutual Photometric Calibration Method. In *Computational Optical Sensing and Imaging*. Optica Publishing Group, CW4B–5.

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

Julie Chang and Gordon Wetzstein. 2019. Deep optics for monocular depth estimation and 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10193–10202.

Yihua Cheng, Haofei Wang, Yiwei Bao, and Feng Lu. 2024. Appearance-based gaze estimation with deep learning: A review and benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46, 12 (2024), 7509–7528.

Minseok Choi, Joohoon Kim, Seokil Moon, Kilsoo Shin, Seung-Woo Nam, Yujin Park, Dohyun Kang, Gyoseon Jeon, Kyung-il Lee, Dong Hyun Yoon, Yoonchan Jeong, Chang-Kun Lee, and Junsuk Rho. 2025. Roll-to-plate printable RGB achromatic metalens for wide-field-of-view holographic near-eye displays. *Nature Materials* 24, 4 (apr 2025), 535–543. https://doi.org/10.1038/s41563-025-02121-0

John Daugman. 2009. How iris recognition works. In *The essential guide to image processing*. Elsevier, 715–739.

Xiong Dun, Hayato Ikoma, Gordon Wetzstein, Zhanshan Wang, Xinbin Cheng, and Yifan Peng. 2020. Learned rotationally symmetric diffractive achromat for full-spectrum computational imaging. *Optica* 7, 8 (Aug 2020), 913–922.

Yu Feng et al. 2024. BlissCam: Boosting Eye Tracking Efficiency with Learned In-Sensor Sparse Sampling. In *ISCA*. 1262–1277.

Johannes E Fröch, Praneeth Chakravarthula, Jipeng Sun, Ethan Tseng, Shane Colburn, Alan Zhan, Forrest Miller, Anna Wirth-Singh, Quentin AA Tanguy, Zheyi Han, et al. 2025. Beating spectral bandwidth limits for large aperture broadband nano-optics. *Nature communications* 16, 1 (2025), 3025.

Kenneth Alberto Funes Mora, Florent Monay, and Jean-Marc Odobez. 2014. Eyediap: A database for the development and evaluation of gaze estimation algorithms from rgb and rgb-d cameras. In *Proceedings of the symposium on eye tracking research and applications*. 255–258.

Guillermo Gallego, Tobi Delbrück, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J Davison, Jörg Conradt, Kostas Daniilidis, et al. 2020. Event-based vision: A survey. *IEEE transactions on pattern analysis and machine intelligence* 44, 1 (2020), 154–180.

Elias Daniel Guestrin and Moshe Eizenman. 2006. General theory of remote gaze estimation using the pupil center and corneal reflections. *IEEE Transactions on biomedical engineering* 53, 6 (2006), 1124–1133.

Qi Guo, Zhujun Shi, Yao-Wei Huang, Emma Alexander, Cheng-Wei Qiu, Federico Capasso, and Todd Zickler. 2019. Compact single-shot metalens depth sensors inspired by eyes of jumping spiders. *Proceedings of the National Academy of Sciences* 116, 46 (2019), 22959–22965.

Dan Witzner Hansen and Qiang Ji. 2009. In the eye of the beholder: A survey of models for eyes and gaze. *IEEE transactions on pattern analysis and machine intelligence* 32, 3 (2009), 478–500.

Felix Heide, Qiang Fu, Yifan Peng, and Wolfgang Heidrich. 2016. Encoded diffractive optics for full-spectrum computational imaging. *Scientific reports* 6, 1 (2016), 33543.

Felix Heide, Mushfiqur Rouf, Matthias B Hullin, Bjorn Labitzke, Wolfgang Heidrich, and Andreas Kolb. 2013. High-quality computational imaging through simple lenses. *ACM Transactions on Graphics (ToG)* 32, 5 (2013), 1–14.

Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. 2019. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*. 1314–1324.

Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861* (2017).

Thomas E Hutchinson, K Preston White, Worthy N Martin, Kelly C Reichert, and Lisa A Frey. 1989. Human-computer interaction using eye-gaze input. *IEEE Transactions on systems, man, and cybernetics* 19, 6 (1989), 1527–1534.

Hayato Ikoma, Cindy M Nguyen, Christopher A Metzler, Yifan Peng, and Gordon Wetzstein. 2021. Depth from defocus with learned optics for imaging and occlusion-aware depth estimation. In *2021 IEEE International Conference on Computational Photography (ICCP)*. IEEE, 1–12.

Daniel S Jeon, Seung-Hwan Baek, Shinyoung Yi, Qiang Fu, Xiong Dun, Wolfgang Heidrich, and Min H Kim. 2019. Compact snapshot hyperspectral imaging with diffracted rotation. (2019).

Youngjin Kim, Taewon Choi, Gun-Yeal Lee, Changhyun Kim, Junseo Bang, Junhyeok Jang, Yoonchan Jeong, and Byoungho Lee. 2024. Metasurface folded lens system for ultrathin cameras. *Science Advances* 10, 44 (2024), eadr2319.

Joohwan Kim, Michael Stengel, Alexander Majercik, Shalini De Mello, David Dunn, Samuli Laine, Morgan McGuire, and David Luebke. 2019. Nvgaze: An anatomically-informed dataset for low-latency, near-eye gaze estimation. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–12.

Jeremy Klotz and Shree K. Nayar. 2025. Minimalist Vision with Freeform Pixels. In *ECCV*.

Kyle Krafka, Aditya Khosla, Petr Kellnhofer, Harini Kannan, Suchendra Bhandarkar, Wojciech Matusik, and Antonio Torralba. 2016. Eye tracking for everyone. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2176–2184.

Xing Lin, Yair Rivenson, Nezih T Yardimci, Muhammed Veli, Yi Luo, Mona Jarrahi, and Aydogan Ozcan. 2018. All-optical machine learning using diffractive deep neural networks. *Science* 361, 6406 (2018), 1004–1008.

Victor Liu and Shanhui Fan. 2012. S4: A free electromagnetic solver for layered periodic structures. *Computer Physics Communications* 183, 10 (2012), 2233–2244.

Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. 2018. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European conference on computer vision (ECCV)*. 116–131.

Lukas Mennel, Joanna Symonowicz, Stefan Wachter, Dmitry K Polyushkin, Aday J Molina-Mendoza, and Thomas Mueller. 2020. Ultrafast machine vision with 2D material neural network image sensors. *Nature* 579, 7797 (2020), 62–66.

Christopher A Metzler, Hayato Ikoma, Yifan Peng, and Gordon Wetzstein. 2020. Deep optics for single-shot high-dynamic-range imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1375–1385.

Johannes Meyer, Thomas Schlebusch, Wolfgang Fuhl, and Enkelejda Kasneci. 2020. A novel camera-free eye tracking sensor for augmented reality based on laser scanning. *IEEE Sensors Journal* 20, 24 (2020), 15204–15212.

Diederick C Niehorster, Thiago Santini, Roy S Hessels, Ignace TC Hooge, Enkelejda Kasneci, and Marcus Nyström. 2020. The impact of slippage on the data quality of head-worn eye trackers. *Behavior research methods* 52, 3 (2020), 1140–1160.

OMNI. 2015. OVM6211. (2015). https://www.ovt.com/products/ovm6211/ Accessed: 2026-01-20.

Sri Rama Prasanna Pavani, Michael A Thompson, Julie S Biteen, Samuel J Lord, Na Liu, Robert J Twieg, Rafael Piestun, and William E Moerner. 2009. Three-dimensional, single-molecule fluorescence imaging beyond the diffraction limit by using a double-helix point spread function. *Proceedings of the National Academy of Sciences* 106, 9 (2009), 2995–2999.

Yifan Peng, Qiang Fu, Hadi Amata, Shuochen Su, Felix Heide, and Wolfgang Heidrich. 2015. Computational imaging using lightweight diffractive-refractive optics. *Optics express* 23, 24 (2015), 31393–31407.

Yifan Peng, Qilin Sun, Xiong Dun, Gordon Wetzstein, Wolfgang Heidrich, and Felix Heide. 2019. Learned large field-of-view imaging with thin-plate optics. *ACM Trans. Graph.* 38, 6 (2019), 219–1.

Hieu Pham, Melody Guan, Barret Zoph, Quoc Le, and Jeff Dean. 2018. Efficient neural architecture search via parameters sharing. In *International conference on machine learning*. PMLR, 4095–4104.

Nikola Popovic, Dimitrios Christodoulou, Danda Pani Paudel, Xi Wang, and Luc Van Gool. 2023. Model-aware 3D eye gaze from weak and few-shot supervisions. In *2023 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*. IEEE, 746–751.

Esteban Real, Alok Aggarwal, Yanping Huang, and Quoc V Le. 2019. Regularized evolution for image classifier architecture search. In *Proceedings of the aaai conference on artificial intelligence*, Vol. 33. 4780–4789.

Kripasindhu Sarkar, Marcel C. Bühler, Gengyan Li, Daoye Wang, Delio Vicini, Jérémy Riviere, Yinda Zhang, Sergio Orts-Escolano, Paulo Gotardo, Thabo Beeler, and Abhimitra Meka. 2023. LitNeRF: Intrinsic Radiance Decomposition for High-Quality View Synthesis and Relighting of Faces. In *SIGGRAPH Asia 2023 Conference Papers (SA '23)*. Association for Computing Machinery, New York, NY, USA, Article 42,

11 pages. https://doi.org/10.1145/3610548.3618210

N Sarkar, B O'Hanlon, A Rohani, D Strathearn, G Lee, M Olfat, and RR Mansour. 2017. A resonant eye-tracking microsystem for velocity estimation of saccades and foveated rendering. In *2017 IEEE 30th International Conference on Micro Electro Mechanical Systems (MEMS)*. IEEE, 304–307.

Laura Sesma, Arantxa Villanueva, and Rafael Cabeza. 2012. Evaluation of pupil center-eye corner vector for gaze estimation using a web cam. In *Proceedings of the symposium on eye tracking research and applications*. 217–220.

Zheng Shi, Ilya Chugunov, Mario Bijelic, Geoffroi Côté, Jiwoon Yeom, Qiang Fu, Hadi Amata, Wolfgang Heidrich, and Felix Heide. 2024a. Split-Aperture 2-in-1 Computational Cameras. *ACM Transactions on Graphics (TOG)* 43, 4 (2024), 1–19.

Zheng Shi, Xiong Dun, Haoyu Wei, Siyu Dong, Zhanshan Wang, Xinbin Cheng, Felix Heide, and Yifan Peng. 2024b. Learned multi-aperture color-coded optics for snapshot hyperspectral imaging. *ACM Transactions on Graphics (TOG)* 43, 6 (2024), 1–11.

Vincent Sitzmann, Steven Diamond, Yifan Peng, Xiong Dun, Stephen P. Boyd, Wolfgang Heidrich, Felix Heide, and Gordon Wetzstein. 2018. End-to-end optimization of optics and image processing for achromatic extended depth of field and super-resolution imaging. *ACM Transactions on Graphics (SIGGRAPH)* 37, 4 (2018), 1–13.

D Straumann, DS Zee, D Solomon, and PD Kramer. 1996. Validity of Listing's law during fixations, saccades, smooth pursuit eye movements, and blinks. *Experimental brain research* 112, 1 (1996), 135–146.

Shuochen Su, Felix Heide, Gordon Wetzstein, and Wolfgang Heidrich. 2018. Deep end-to-end time-of-flight imaging. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6383–6392.

Jipeng Sun, Kaixuan Wei, Thomas Eboli, Congli Wang, Cheng Zheng, Zhihao Zhou, Arka Majumdar, Wolfgang Heidrich, and Felix Heide. 2025. Collaborative On-Sensor Array Cameras. *ACM Transactions on Graphics (TOG)* 44, 4 (2025), 1–18.

Li Sun, Zicheng Liu, and Ming-Ting Sun. 2015. Real time gaze estimation with a consumer depth camera. *Information Sciences* 320 (2015), 346–360.

Qilin Sun, Ethan Tseng, Qiang Fu, Wolfgang Heidrich, and Felix Heide. 2020. Learning rank-1 diffractive optics for single-shot high dynamic range imaging. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 1386–1396.

Vivienne Sze, Yu-Hsin Chen, Tien-Ju Yang, and Joel S Emer. 2017. Efficient processing of deep neural networks: A tutorial and survey. *Proc. IEEE* 105, 12 (2017), 2295–2329.

Mingxing Tan and Quoc Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*. PMLR, 6105–6114.

Marc Tonsen, Julian Steil, Yusuke Sugano, and Andreas Bulling. 2017. Invisibleeye: Mobile eye tracking using multiple low-resolution cameras and learning-based gaze estimation. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 1–21.

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

Ethan Tseng, Ali Mosleh, Fahim Mannan, Karl St-Arnaud, Avinash Sharma, Yifan Peng, Alexander Braun, Derek Nowrouzezahrai, Jean-Francois Lalonde, and Felix Heide. 2021b. Differentiable compound optics and processing pipeline optimization for end-to-end camera design. *ACM Transactions on Graphics (TOG)* 40, 2 (2021), 1–19.

Douglas Tweed and Tutis Vilis. 1990. Geometric relations of eye position and velocity vectors during saccades. *Vision research* 30, 1 (1990), 111–127.

Alvin Wan et al. 2020. FBNetV2: Differentiable Neural Architecture Search for Spatial and Channel Dimensions. In *CVPR*. 12962–12971.

Haiyu Wang et al. 2025a. Process Only Where You Look: Hardware and Algorithm Co-Optimization for Efficient Gaze-Tracked Foveated Rendering in Virtual Reality. In *ISCA*. 344–358.

Jiazhang Wang, Tianfu Wang, Bingjie Xu, Oliver Cossairt, and Florian Willomitzer. 2025b. Accurate eye tracking from dense 3D surface reconstructions using single-shot deflectometry. *Nature Communications* 16, 1 (2025), 2902.

Kaixuan Wei, Xiao Li, Johannes Froech, Praneeth Chakravarthula, James Whitehead, Ethan Tseng, Arka Majumdar, and Felix Heide. 2024. Spatially varying nanophotonic neural networks. *Science Advances* 10, 45 (2024), eadp0391.

Kaixuan Wei, Hector Romero, Hadi Amata, Jipeng Sun, Qiang Fu, Felix Heide, and Wolfgang Heidrich. 2025. Large-Area Fabrication-aware Computational Diffractive Optics. *ACM Transactions on Graphics (TOG)* 44, 6 (2025), 1–17.

G. Wetzstein, A. Ozcan, S. Gigan, S. Fan, D. Englund, M. Soljačić, C. Denz, D. Miller, and D. Psaltis. 2020. Inference in artificial intelligence with deep optics and photonics. *Nature* 588 7836 (2020), 39–47.

Yunfeng Xiao, Xiaowei Bai, Baojun Chen, Hao Su, Hao He, Liang Xie, and Erwei Yin. 2025. Deˆ 2Gaze: Deformable and Decoupled Representation Learning for 3D Gaze Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3091–3100.

Haoran You et al. 2023. EyeCoD: Eye Tracking System Acceleration via FlatCam-Based Algorithm and Hardware Co-Design. *IEEE Micro* 43, 4 (2023), 88–97.

Laurence R Young and David Sheena. 1975. Survey of eye movement recording methods. *Behavior research methods & instrumentation* 7, 5 (1975), 397–429.

Yu Yu, Gang Liu, and Jean-Marc Odobez. 2019. Improving few-shot user-specific gaze adaptation via gaze redirection synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 11937–11946.

Jeong-Geun Yun, Hyunjung Kang, Kyookeun Lee, Youngmo Jeong, Eunji Lee, Joohoon Kim, Minseok Choi, Bonkon Koo, Doyoun Kim, Jongchul Choi, et al. 2025. Compact eye camera with two-third wavelength phase-delay metalens. *Nature Communications* 16, 1 (2025), 7299.

Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. 2015. Appearance-based gaze estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4511–4520.

Cheng Zheng, Guangyuan Zhao, and Peter So. 2023. Close the Design-to-Manufacturing Gap in Computational Optics with a'Real2Sim'Learned Two-Photon Neural Lithography Simulator. In *SIGGRAPH Asia 2023 Conference Papers*. 1–9.