Distortion-Free Mid-Air Image Inside a Transparent Object via Viewpoint-Adaptive Neural Correction

Shunji Kiuchi
kiuchi@media.lab.uec.ac.jp
The University of Electro-Communications
Chofu, Tokyo, Japan

Yuichi Hiroi
y.hiroi@cluster.mu
Cluster Metaverse Lab
Tokyo, Japan

Naoya Koizumi
koizumi.naoya@uec.ac.jp
The University of Electro-Communications
Chofu, Tokyo, Japan

Abstract

Mid-air imaging optics display a digital image at a physical location in real space. Placing a transparent object there lets the image share its place with real matter, and puts the object—not the optics—in front of the observer. The light must then traverse the object's refractive surfaces, which distort the image differently at every viewpoint; and because the head moves continuously, the correction has to be regenerated as the head moves, far faster than ray tracing the whole optical path allows. We formulate view-dependent distortion correction as learning the inverse mapping from a light-source pixel and the observer's pose to the point at which that pixel is perceived, and represent it implicitly with a multilayer perceptron distilled from the ray-traced correction. The network also predicts a validity mask, suppressing light-source pixels that no valid light path reaches and that would otherwise appear as grid-like ghosting. In a ray-tracing simulation of a refractive sphere and a micro-mirror array plate, one 51.8 ms forward pass reproduces the ray-traced correction at unseen viewpoints.

CCS Concepts

• **Computing methodologies** → **Mixed / augmented reality**.

Keywords

mid-air imaging, distortion correction, implicit neural representation, view-dependent rendering, ray tracing

ACM Reference Format:
Shunji Kiuchi, Yuichi Hiroi, and Naoya Koizumi. 2026. Distortion-Free Mid-Air Image Inside a Transparent Object via Viewpoint-Adaptive Neural Correction. In *32nd ACM Symposium on Virtual Reality Software and Technology (VRST '26), November 16–18, 2026, Sendai, Japan.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3822517.3849931

1 Introduction

Mid-air imaging optics such as micro-mirror array plates (MMAPs) display a digital image at a physical location in real space. Putting a transparent object in that optical path opens up imagery that shares its place with real matter: a character can drift inside a glass ornament, or an image can spill out of it. The object also stands in the foreground, so what the observer looks into is the object, not the optics behind it.

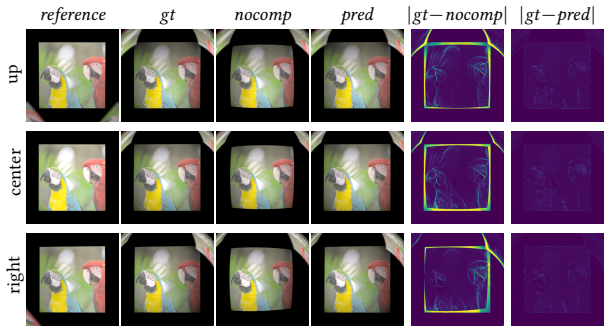


Figure 1: Three viewpoints of the 330-viewpoint test set, on the $z = 100$ cm plane. *reference* is the image the observer should see; *gt* is that same geometry after it has passed through the transparent object, which absorbs light that no geometric correction can restore. The last two columns are absolute differences from *gt*, on one shared color scale.

The object's refractive surfaces, however, bend the light on its way to the eye, and two requirements follow from the observer being free to move. First, the correction is view-dependent: each ray is bent according to where it crosses the refractive body, so a correction that undistorts the image at one viewpoint is wrong at the next (Fig. 1). Second, the head moves continuously, so the correction has to be regenerated while the head is moving, within a per-frame budget of milliseconds.

Existing corrections differ in where they cancel the distortion. Fujii et al. [1] cancel it optically, placing a mirror-image copy of the target object on the light-source side of an aerial imaging by retro-reflection (AIRR) system, so every target object needs a physical counterpart fabricated for it. Kiuchi and Koizumi [3] cancel it in software and need no counterpart, but they obtain the correction by tracing rays through the whole optical path, which costs on the order of a second per viewpoint and grows with the geometric complexity of the optics.

Implicit neural representations supply the missing speed, encoding a continuous function in network weights and reading it out in a single forward pass. NeRF [4] represents a scene as a radiance field, and Neural Distortion Fields [2] applies an MLP to near-eye display distortion. NDF treats the optics as a black box observable only through captured images, so it can learn only the *forward* map—from a spatial position and gaze direction to a display pixel—and has to sample points along each ray and sum them with weights to

This work is licensed under a Creative Commons Attribution 4.0 International License.
VRST '26, Sendai, Japan
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2811-2/2026/11
https://doi.org/10.1145/3822517.3849931

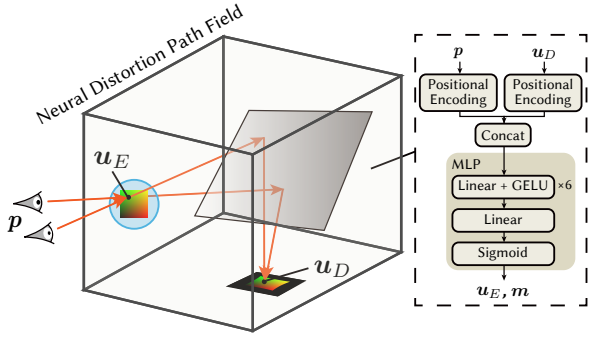


Figure 2: Neural distortion path field. An MLP maps the encoded pixel and pose (u_D, p) to the perceived-plane coordinate and mask $(\hat{u}_E, \hat{m})$.

synthesize a distortion map. Our optical system is known and can be ray traced, so the *inverse* map can be supervised directly.

We therefore distill the static ray-traced correction [3] into a viewpoint-conditioned field of that inverse map, which we call a *neural distortion path field* (NDPF). Because the field is the inverse map, rendering it is a gather: one query per light-source pixel, with no ray sampling and nothing to interpolate. Its cost is set by the network, not by the optics, so it does not grow with optical complexity.

2 Method

Let $\boldsymbol{u}_D \in \mathbb{R}^2$ be a pixel on the light-source display, $\boldsymbol{p}$ the observer's pose, and $\boldsymbol{u}_E \in \mathbb{R}^2$ a coordinate on the image the observer perceives in mid-air (Fig. 2). Tracing rays from the viewpoint through the MMAP and the transparent object connects each perceived coordinate to the display pixel that feeds it; this forward mapping $f : (\boldsymbol{u}_E, \boldsymbol{p}) \to \boldsymbol{u}_D$ *is* the distortion the object introduces, and it differs at every $\boldsymbol{p}$. Pushing the target image along f [3] scatters samples irregularly over the display, so a smooth light-source image needs dense sampling and interpolation. We instead solve the *inverse mapping* $g : (\boldsymbol{u}_D, \boldsymbol{p}) \to \boldsymbol{u}_E$, which tells each display pixel where it will be seen: reading the target at $g(\boldsymbol{u}_D, \boldsymbol{p})$ fills every pixel exactly once, so one gather pass gives a complete light-source image with nothing to interpolate.

We realize g as an NDPF, an implicit representation held in the weights of an MLP; pre-rendering it into a look-up table instead would need dense viewpoint sampling and a large amount of memory. Fig. 2 outlines the network, which takes $\boldsymbol{u}_D$ and the pose, the latter flattened from its 3×4 transform into 12 dimensions. Each passes through a Fourier positional encoding (6 bands for the coordinate, 4 for the pose) so the low-frequency-biased MLP can represent the distortion's high-frequency variation; the encoded inputs are concatenated and passed through six width-256 GELU layers, and a final sigmoid outputs $\hat{\boldsymbol{u}}_E \in [0, 1]^2$ and a validity mask $\hat{m} \in [0, 1]$, without which the smooth MLP returns plausible but wrong coordinates wherever no valid light path exists.

We train with a mean-squared-error loss on $\hat{\boldsymbol{u}}_E$ over valid samples plus an equally weighted binary cross-entropy loss on the mask. Targets come from the static correction [3], run once per training viewpoint: back-propagating the undistorted view image through the optics to the light source yields the ground-truth $\boldsymbol{u}_D$–$\boldsymbol{u}_E$ correspondence and the mask. We optimize with AdamW for 10 epochs at batch size 262,144 and learning rate 1×10^{-3}, which takes four hours on an NVIDIA RTX 3090. For inference the network is converted through ONNX to a TensorRT engine, which maps a 1080×1080 frame in 51.8 ms on the same GPU (FP32, transfers included; s.d. 0.28 ms over 200 runs).

3 Evaluation

We evaluate the NDPF's correction under viewpoint change in a ray-tracing simulation. An MMAP 48.8 cm on a side is tilted at 45° to the light source, with the source and a glass sphere of refractive index 1.5 and radius 5 cm each 20.0 cm from its center. Viewpoints lie on a grid over $x, y \in [-10, 10]$ cm and $z \in [90, 110]$ cm: $5 \times 5 \times 3 = 75$ for training, and 330 held-out viewpoints from a denser grid for testing. The displayed image comes from the LIVE database [5].

Table 1: Mean over the 330 test viewpoints, against the ideal *reference*. *gt* bounds what any correction can reach here.

	SSIM ↑	LPIPS ↓
nocomp	0.697	0.294
pred (ours)	0.736	0.229
gt (ceiling)	0.743	0.227

We measure against the image the observer should see, rendered with the transparent object absent from the path. Nothing can reach it: the object absorbs light that a geometric correction cannot restore. Table 1 therefore also reports the static correction [3] ray traced at the test viewpoint itself (*gt*), which bounds what is attainable. Our correction (*pred*) closes 85% of the SSIM gap and 97% of the LPIPS gap between the uncorrected image (*nocomp*) and that bound, from one 51.8 ms forward pass at a viewpoint never seen in training, a cost over an order of magnitude below that of ray tracing the same correction. Qualitatively (Fig. 1), the bright band that traces the displaced image border in $|gt - nocomp|$ is absent from $|gt - pred|$, whose mean absolute difference is 4.5× smaller.

Acknowledgments

This work was supported by JST FOREST Program, Grant Number JPMJFR216L.

References

[1] Kengo Fujii, Masaki Yasugi, Satoshi Maekawa, and Hirotsugu Yamamoto. 2021. Reduction of retro-reflector and expansion of the viewpoint of an aerial image by the use of AIRR with transparent spheres. *OSA Continuum* 4, 4 (2021), 1207–1214. doi:10.1364/OSAC.418116

[2] Yuichi Hiroi, Kiyosato Someya, and Yuta Itoh. 2022. Neural Distortion Fields for Spatial Calibration of Wide Field-of-View Near-Eye Displays. *Optics Express* 30, 22 (2022), 40628–40644. doi:10.1364/OE.471370

[3] Shunji Kiuchi and Naoya Koizumi. 2022. Distortion-free Mid-air Image Inside Refractive Surface and on Reflective Surface. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. 606–614. doi:10.1109/VR51125.2022.00081

[4] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 405–421.

[5] Hamid R. Sheikh, Zhou Wang, Lawrence Cormack, and Alan C. Bovik. 2005. LIVE Image Quality Assessment Database. https://live.ece.utexas.edu/research/quality.