

Reconstructing Realistic and Relightable Eyes - Supplementary Material

Wesley Khademi
Oregon State University
Corvallis, OR, USA

khademiw@oregonstate.edu

Jogendra Kundu
Meta
Menlo Park, CA, USA

jogendrak@meta.com

Yatong An
Meta
Menlo Park, CA, USA

yatong@meta.com

Alexander Fix
Meta
Menlo Park, CA, USA
alexander.fix@meta.com

David Colmenares
Meta
Menlo Park, CA, USA
dcol@meta.com

1. Overview

An overview of our supplemental material is presented as follows:

- Dataset (Section 2): we provide more details about our dataset
- Model details (Section 3): we provide more details of our model
- Training settings (Section 4): we describe our training setup in more detail
- Results (Section 5): we provide more results of our method

2. Dataset

We share an example of a subject captured by our light stage in Figure 1. The subject is captured under 18 different viewpoints with each viewpoint capturing an image of the eye and periocular region under 35 different lighting conditions.

2.1. Cameras

In our light stage, we use IR cameras with a 850nm infrared wavelength. Capturing at 850nm blocks out most of the external environment illumination, and therefore the illumination info that is captured in our images is only from our point light sources. We capture images at a resolution of 2048×2048 and directly use the raw images without any processing (e.g., tone mapping) in our pipeline.

2.2. Lighting conditions

With our 16 point light sources, we capture images under 35 different lighting conditions. Our lighting conditions can be broken down into 4 settings:

- **One-on** (16 lighting conditions): Only 1 of the 16 lights is turned on during the capture. We cycle through turning

on each light to produce 16 one-light-at-a-time (OLAT) images per viewpoint.

- **One-off** (16 lighting conditions): Starting with all lights on, we turn off 1 of the 16 lights and capture an image. We repeat this process for each of the 16 lights.
- **Half-on** (2 lighting conditions): We turn on every other point light to capture the face under half our illumination sources. The process is then repeated for the other 8 lights.
- **All-on** (1 lighting condition): We capture an image with all our point lights turned on, which provides uniform illumination of the face.

Throughout our entire capture, we keep the intensity of each point light fixed. Since we do not process the raw images, a linear relationship exists between the exposure length and brightness in our final image. We make use of this linear relationship and adjust exposure length during our capture to roughly match the perceived brightness across the 4 settings. For example, to match the brightness of a "one-on" image with the "all-on" image, we use an exposure length that is $16\times$ the exposure length used to capture the "all-on" image.

3. Model details

3.1. Light intensity

While we calibrate the position of our point lights, we do not calibrate for the intensity of each point light. However, the intensity of each point light is needed to model our outgoing radiance $L_o(\mathbf{x}, \omega_i)$. To address this, we jointly learn the intensity of each point light along with our NeRF directly from our captures.

Our capture rig keeps the intensity of each point light fixed while varying exposure length to adjust brightness in the image; however, the outgoing radiance in our model is

defined in terms of point light intensity not exposure length. Since there exists a linear relationship between the perceived brightness in our images and exposure length, we choose to model the intensity of a point light l as a linear function of exposure:

$$i_l(e) = \alpha_l * e + \beta_l \quad (1)$$

where e is the exposure length and α_l and β_l are learned scale and offset parameters, respectively.

3.2. Camera ray refraction

To model camera ray refraction, we explicitly bend rays using Snell’s law. Snell’s law describes the relationship between the angles of incidence and refraction for a light ray which crosses over from one media to another as:

$$\eta_1 \sin \theta_1 = \eta_2 \sin \theta_2 \quad (2)$$

where η_1 and η_2 are the index of refraction of each media, and θ_1 and θ_2 are the angle of incidence and refraction, respectively. In our experiments, we are interested in the transition from air to cornea, where we define the index of refraction for air to be $\eta_1 = 1.0$ and index of refraction for the cornea to be $\eta_2 = 1.337$.

Given a camera ray $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$, we first compute where the ray intersects our eye mesh. We denote the ray-mesh intersection point as \mathbf{x}_{mesh} which has a corresponding surface normal \mathbf{n} . Then the refracted ray direction \mathbf{d}_d can be computed as:

$$\mathbf{d}_d = \frac{\eta_1}{\eta_2} \mathbf{d} + \left(\frac{\eta_1}{\eta_2} \cos \theta_1 - \cos \theta_2 \right) \mathbf{n} \quad (3)$$

where $\cos \theta_1 = -\mathbf{d} \cdot \mathbf{n}$ and $\cos \theta_2$ can be derived from rearranging Snell’s law (Equation 2) and using trigonometric identities:

$$\begin{aligned} \cos \theta_2 &= \sqrt{1 - (\sin \theta_2)^2} \\ &= \sqrt{1 - \left(\frac{\eta_1}{\eta_2} \sin \theta_1 \right)^2} \\ &= \sqrt{1 - \left(\frac{\eta_1}{\eta_2} \sqrt{1 - (\cos \theta_1)^2} \right)^2} \\ &= \sqrt{1 - \left(\frac{\eta_1}{\eta_2} \right)^2 (1 - (\cos \theta_1)^2)} \end{aligned} \quad (4)$$

Having computed the refracted ray direction \mathbf{d}_d , the refracted ray can then be defined as $\mathbf{r}_d(t) = \mathbf{x}_{mesh} + t\mathbf{d}_d$.

3.3. Light ray refraction

We train our Direct Illumination Ray network per subject by generating subject specific training data based on their eye mesh reconstruction. In particular, from a subject’s eye mesh, we generate data pairs $\{(\mathbf{p}_l, \mathbf{x}_{eye}), \omega_{gt}\}$ where \mathbf{p}_l is

the position of a point light, \mathbf{x}_{eye} is a point inside the eye, and ω_{gt} is the direction vector which is emitted from \mathbf{p}_l and intersects point \mathbf{x}_{eye} after refracting at the cornea.

We begin by sampling a set of point light positions along hemispheres of varying radii which surround the eye mesh. We center the hemispheres about the axis that the cornea is facing and set the smallest hemisphere radius to be as close as possible to the eye without any samples falling inside the eye. Hemispheres are uniformly sampled between the near and far radius and we fix the far radius to be $r_{far} = 0.4\text{m}$. Point light positions \mathbf{p}_l are then uniformly sampled from each hemisphere. Sampling light positions between our near and far radii ensure that our model can predict light directions for the point lights in our light stage during NeRF training while generalizing at inference time to near-field illumination settings that we expect in AR/VR devices.

For each sampled light position \mathbf{p}_l , we randomly sample a light ray direction ω_{gt} which intersects with our eye mesh. This direction vector will serve as the ground truth direct illumination ray that we use for training and we will use it to generate a corresponding point inside the eye that falls along its refracted ray. In practice, we generate these directions on the fly at every iteration during training to ensure we are covering all the possible direct illumination rays.

From a point light position \mathbf{p}_l and a light ray direction ω_{gt} , we can generate a corresponding point inside the eye. Starting from point \mathbf{p}_l we can trace a ray in the direction ω_{gt} until we intersect with our eye mesh. At the point of intersection, we compute the refracted ray using Equation 3 and then randomly sample a point \mathbf{x}_{eye} along the refracted ray. From this process, we can generate training data pairs $\{(\mathbf{p}_l, \mathbf{x}_{eye}), \omega_{gt}\}$ per subject to use for training our Direct Illumination Ray network.

4. Training settings

In this section, we describe the differences in training data used for comparing our method to Instant NGP [1].

4.1. Instant NGP

Instant NGP directly models the outgoing radiance at a point in space, ignoring the fact that outgoing radiance is really a result of how incoming light interacts with the object’s geometry and material properties. As lighting information is ignored when modeling the outgoing radiance, Instant NGP can only be trained from images of a scene which were captured under the same illumination. Moreover, once a scene has been trained, it cannot be relit under novel illumination conditions. Because of this, it is not possible for us to leverage all the data from our light stage to train Instant NGP on for our comparisons. Instead, we choose to show that Instant NGP struggles to learn the geometry and appearance of the eye and periorbital region when trained on data captured with fixed illumination, if that illumina-

tion setting is non-ideal. In particular, we train Instant NGP on one of our "one-on" frames, which consists of 18 different viewpoints of the face captured under a single point light source. In this setting, the face is non-uniformly lit and there is greater potential for large shadows due to self-occlusion. We choose to use 17 of our camera viewpoints for training and holdout a single viewpoint to perform evaluation on.

4.2. Relightable Instant NGP

Instead of directly comparing against Instant NGP in the same training setting, we aim to demonstrate the benefits of training with more data. Since our model explicitly accounts for incoming light information, we are able to train on images of the face that were captured from the same viewpoint but under varying illumination. To demonstrate the effectiveness of a relightable model which can leverage more training data, we choose to hold out any images containing the point light source that was used for training Instant NGP, and instead train on the images containing all the other point light sources in our light stage. In particular, our training data consists of 15 different "one-on" lighting conditions, 1 "one-off" lighting condition, and 1 "half-off" lighting condition. For each of the selected lighting conditions, we once again use 17 of the camera viewpoints for training and hold one out for evaluation. In this way, we are not only evaluating our model on novel view synthesis but also relighting capabilities, while Instant NGP is only evaluated on novel view synthesis. Therefore, while our method has more training data in our comparisons against Instant NGP, we are performing a harder task in our evaluation.

5. Results

5.1. Novel view synthesis

In Figure 2, we present more examples of rendering subjects from novel viewpoints. Despite our light stage's cameras being positioned 30cm from the face with most of them being more centrally focused, our method is able to faithfully render appearance from viewpoints much closer to the face (e.g., 6cm) and from more oblique viewpoints. Moreover, our method is able to recover detailed iris texture, represent thin structures such as eyelashes, and represent the highly specular glints on the cornea.

5.2. Relighting

We share more results on relighting a subject in Figure 3. Our method is able to model shadowing effects due to self-occlusion as can be seen on the sclera, upper eyelid, and left side of the face as the point light moves from left to right. Furthermore, we see that the general appearance of the face, iris, and sclera varies with light source position and incident illumination direction. Finally, it can be observed how the

size and shape of glints on the cornea change as a function of the light position.

In Figure 4, we present more comparisons between the three variants of our method (NR, CRR, CRR+LRR) for relighting. We find that not modeling refraction (NR) can sometimes lead the textured iris to appear smoothed out under certain lighting conditions. When modeling only camera ray refraction (CRR), we observe that from certain lighting directions the iris appears overly dark, almost turning completely black, despite the eye not being fully occluded from the light source. However, when modeling both camera and light ray refraction, the iris texture is better maintained across different lighting conditions and does not suffer from overly dark appearance similar to the CRR case.

Finally, in Figure 5, we demonstrate that our method can render arbitrary configurations of point light sources, allowing us to simulate LED rings commonly used in eye tracking. When lit only by a few point lights (e.g. 1-2) our model produces realistic shadowing on the face due to self-occlusion. Moreover, we observe that these shadows begin to disappear as we introduce more point lights which illuminate the face in a more uniform manner.

5.3. Gaze estimation

In Figure 6, we visualize our NeRF's iris reconstruction along with its estimated gaze direction (red arrow) obtained from fitting a plane to the reconstructed iris. When measuring the angular error of our estimated gaze direction against the reference optic axis we have from fitting our eye mesh's pose (blue arrow), we find that accounting for refraction leads to significant improvements in gaze estimation. Since the iris is roughly a plane whose normal is aligned with the optic axis, this large improvement in recovering the subject's gaze is an indication that the iris reconstruction when modeling refraction is more accurately representing the geometry of the iris.

5.4. Structured light

We present more qualitative comparisons of simulating structured light in Figures 7 and 8. Looking at the zoomed insets in Figure 7, we see that not modeling refraction (NR) leads to fringe patterns which are blurred (first row) or distorted (second and third row). Moreover, adding in camera ray refraction (CRR) does not correct for the distortion of the fringe pattern, producing warped fringes that are closer together on the left and right outer regions of the iris. However, adding in light ray refraction (CRR+LRR), helps straighten out the warped fringes, producing more evenly spaced and parallel fringes on the iris. This can also be seen in the cross sections of the eye which we visualize in Figure 8. The projected sinusoidal fringe pattern exhibits significant inconsistencies between the left and right side of the iris in both the no refraction (NR) and camera ray refrac-

tion (CRR) case. On the other hand, when modeling both camera and light ray refraction (CRR+LRR), we find that the projected sinusoidal fringe pattern is much better maintained.

5.5. Light ray prediction

In Figure 9, we justify our hyperparameter choices for our Direct Illumination Ray network. We find that using a greater number of frequencies in our positional encoding leads to a more accurate prediction of the direct illumination ray that passes through a sampled point inside the eye after undergoing refraction. In regards to the depth of the network, we find that 3-layers is the ideal setting. Shallow networks (e.g., 1-layer) do not have enough capacity and end up underfitting, while deeper networks (5 or 7-layers) perform slightly worse, likely due to requiring longer training times to converge to the same performance. Finally, we find that wider networks tend to perform better. For our experiments, we opted to use 128 hidden dimensions as performance does not improve by much after increasing past this point.

References

- [1] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM transactions on graphics (TOG)*, 41(4):1–15, 2022. [2](#)

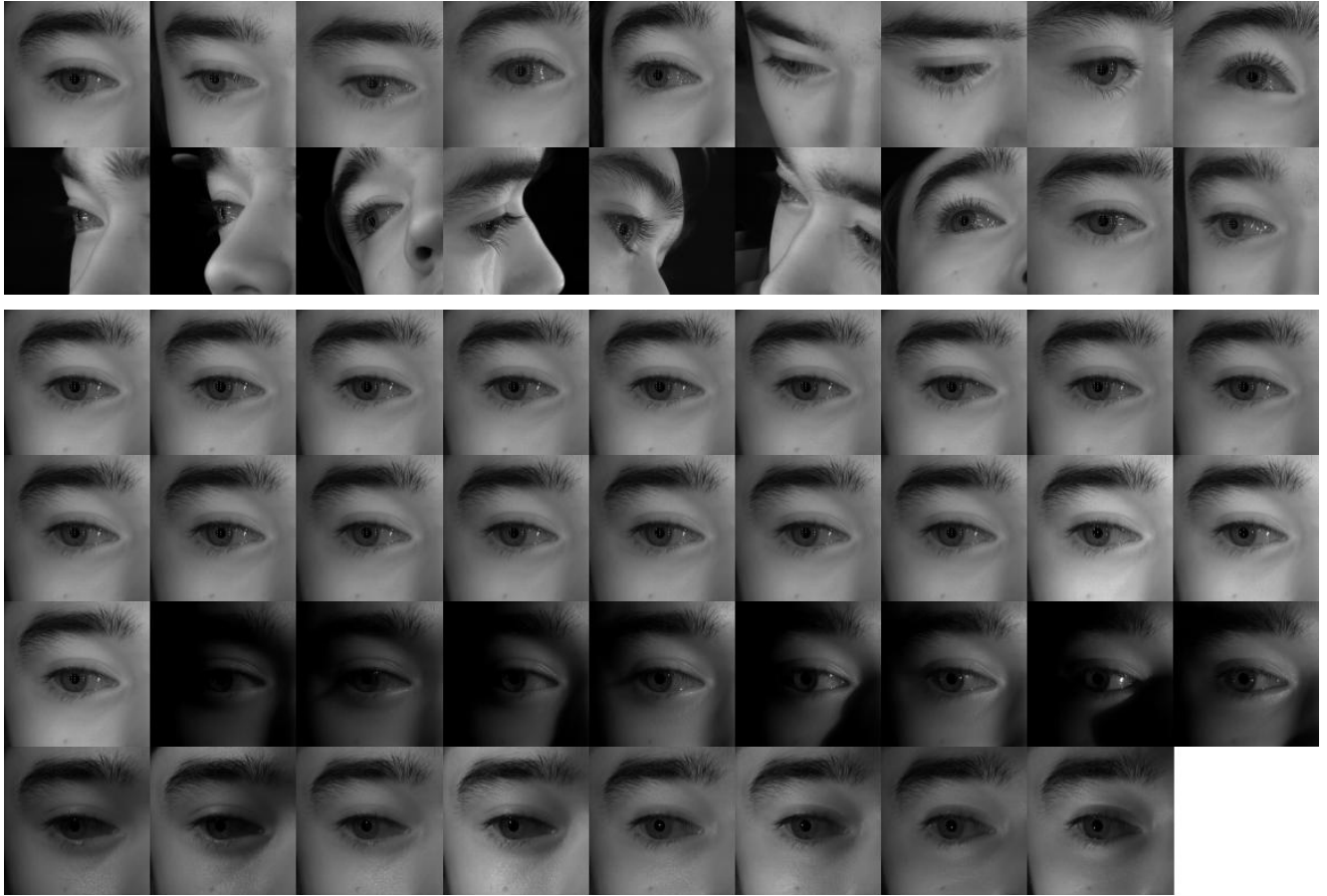


Figure 1. Example of a subject captured by our light stage. We capture a subject from 18 different camera viewpoints (top 2 rows) and under 35 different illumination combinations (bottom 4 rows).



Figure 2. Examples of subjects rendered under novel viewpoint.

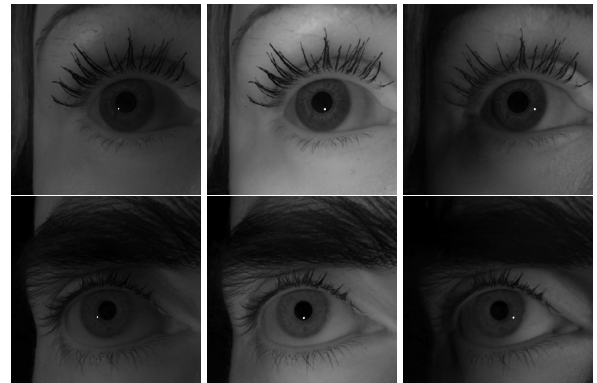


Figure 3. Examples of subjects rendered under novel illumination.



Figure 4. Qualitative comparison of variants of our model when relighting the eye.

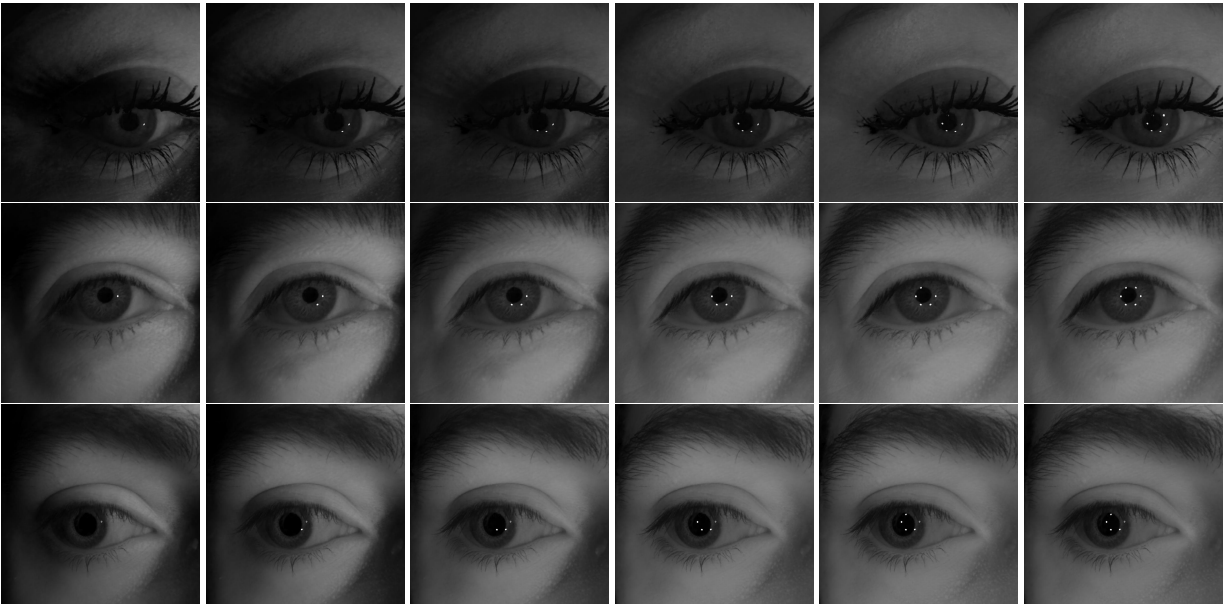


Figure 5. Examples of subjects rendered under multiple point light sources.

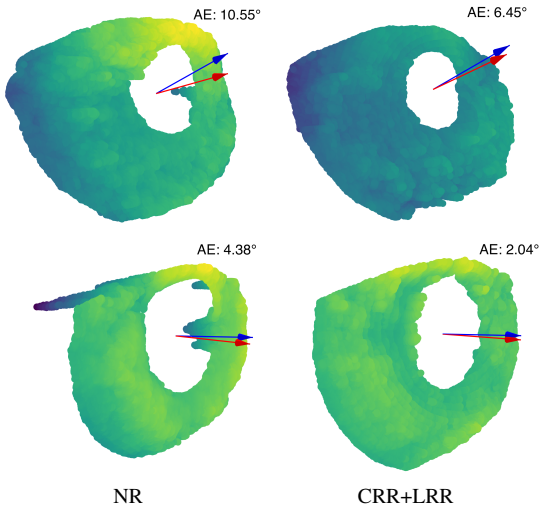


Figure 6. Visualization of our reconstructed iris when ignoring refraction (NR) vs. accounting for both camera and light ray refraction (CRR+LRR). We measure the angular error (AE) between a reference optic axis produced by our eye mesh pose fitting (blue arrow) and an estimated optic axis produced by fitting a plane to our reconstructed iris (red arrow).

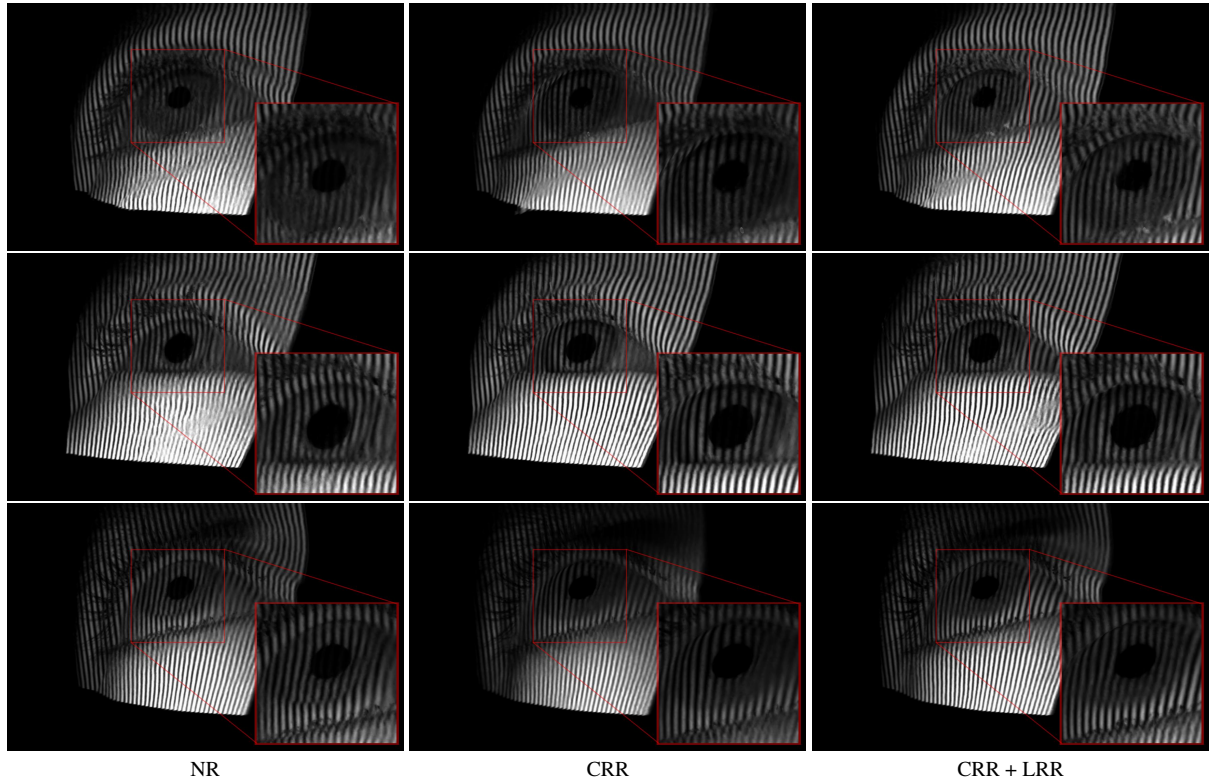


Figure 7. Qualitative comparison of simulating structured light.

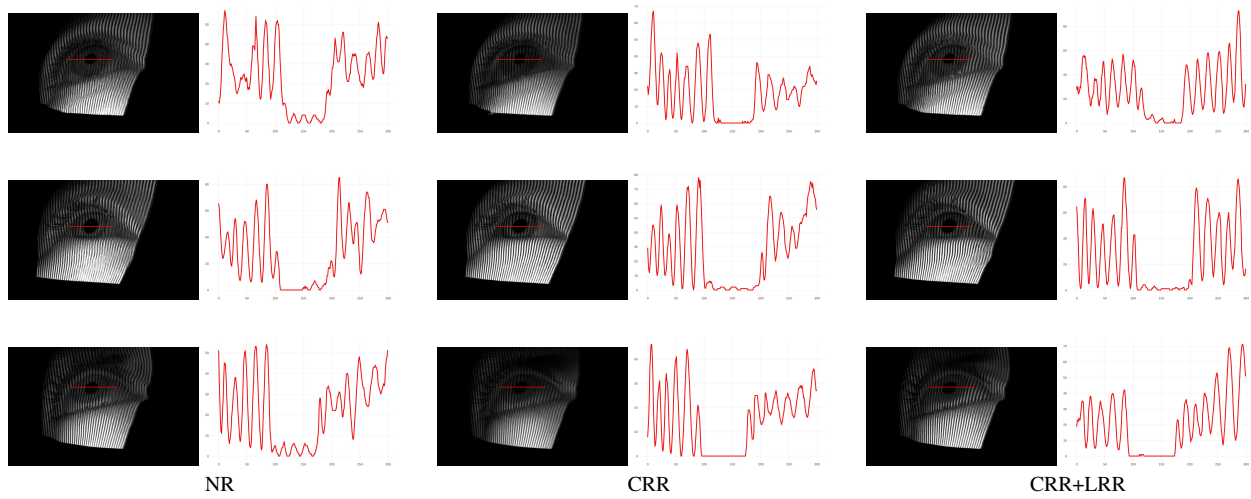


Figure 8. Visualizing a cross section of the eye when simulating structured light.

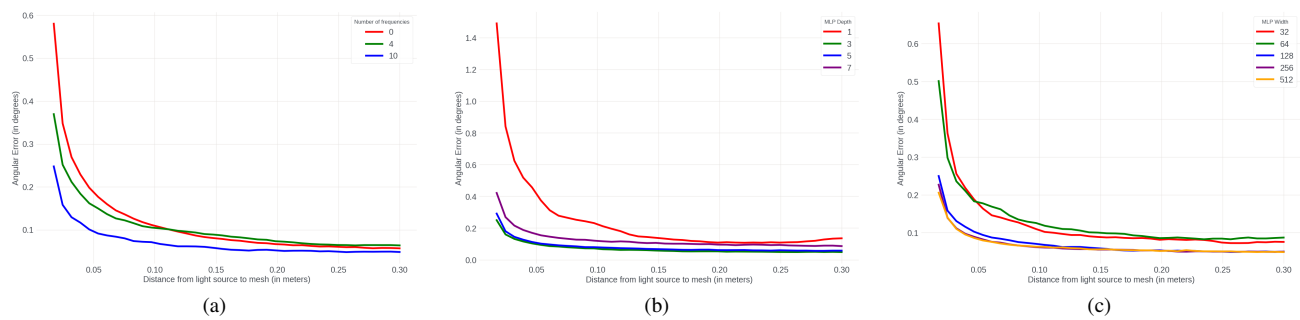


Figure 9. Average angular error of our Direct Illumination Ray network at varying distances from the eye with different settings of positional encoding frequencies (a), MLP depths (b), and MLP widths (c).