

# EyeNeRF: A Hybrid Representation for Photorealistic Synthesis, Animation and Relighting of Human Eyes

GENGYAN LI, ETH Zurich and Google Inc., Switzerland

ABHIMITRA MEKA, Google Inc., USA

FRANZISKA MUELLER, Google Inc., Switzerland

MARCEL C. BUEHLER and OTMAR HILLIGES, ETH Zurich, Switzerland

THABO BEELER, Google Inc., Switzerland

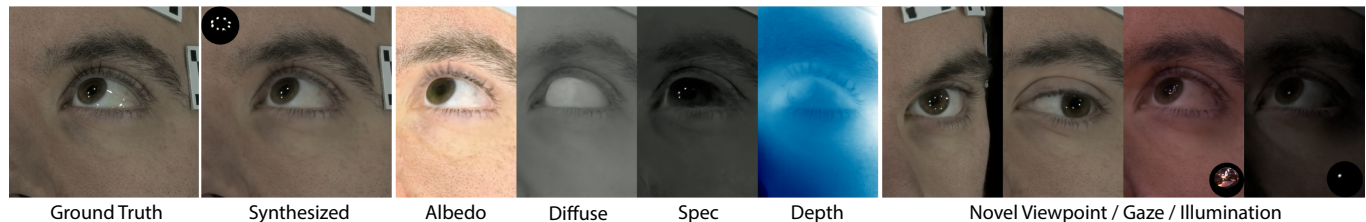


Fig. 1. We introduce EyeNeRF, a novel hybrid representation for photorealistic modeling and synthesis of human eyes. Our technique combines an explicit mesh based representation for the eyeball surface with an implicit neural representation for the periocular region (skin, eyelashes, eyelids etc.) and allows to synthesize photoreal images with novel gaze from a novel viewpoint and under novel illumination.

A unique challenge in creating high-quality animatable and relightable 3D avatars of real people is modeling human eyes, particularly in conjunction with the surrounding periocular face region. The challenge of synthesizing eyes is multifold as it requires 1) appropriate representations for the various components of the eye and the periocular region for coherent viewpoint synthesis, capable of representing diffuse, refractive and highly reflective surfaces, 2) disentangling skin and eye appearance from environmental illumination such that it may be rendered under novel lighting conditions, and 3) capturing eyeball motion and the deformation of the surrounding skin to enable re-gazing.

These challenges have traditionally necessitated the use of expensive and cumbersome capture setups to obtain high-quality results, and even then, modeling of the full eye region holistically has remained elusive. We present a novel geometry and appearance representation that enables high-fidelity capture and photorealistic *animation, view synthesis and relighting of the eye region* using only a sparse set of lights and cameras. Our hybrid representation combines an explicit parametric surface model for the eyeball surface with implicit deformable volumetric representations for the periocular region and the interior of the eye. This novel hybrid model has been designed specifically to address the various parts of that exceptionally challenging facial area - the explicit eyeball surface allows modeling refraction and high frequency specular reflection at the cornea, whereas the implicit representation is well suited to model lower frequency skin

reflection via spherical harmonics and can represent non-surface structures such as hair (i.e. eyebrows) or highly diffuse volumetric bodies (i.e. sclera), both of which are a challenge for explicit surface models. Tightly integrating the two representations in a joint framework allows controlled photoreal image synthesis and joint optimization of both the geometry parameters of the eyeball and the implicit neural network in continuous 3D space. We show that for high-resolution close-ups of the human eye, our model can synthesize high-fidelity animated gaze from novel views under unseen illumination conditions, allowing to generate visually rich eye imagery.

CCS Concepts: • **Computing methodologies** → **Rendering**.

Additional Key Words and Phrases: Relighting, NeRF, Eye Modeling, Regazing, Neural Rendering, Novel View Synthesis, Pose Optimization, Model Fitting, Volumetric Rendering, Differentiable Rendering, Refraction, HDR rendering, specular synthesis, geometry deformation modeling

## ACM Reference Format:

Gengyan Li, Abhimitra Meka, Franziska Mueller, Marcel C. Buehler, Otmar Hilliges, and Thabo Beeler. 2022. EyeNeRF: A Hybrid Representation for Photorealistic Synthesis, Animation and Relighting of Human Eyes. *ACM Trans. Graph.* 41, 4, Article 166 (July 2022), 16 pages. <https://doi.org/10.1145/3528223.3530130>

## 1 INTRODUCTION

“Eyes are windows to the soul” – while an old adage, it captures what many media creators believe is an essential medium of human expression. There is no dearth of close up shots of actor’s eyes on screen - from expressing horror in Hitchcock’s *Psycho* (1960) to showing excitement in Scorsese’s *Goodfellas* (1990) - such shots are a favorite tool in a director’s toolbox. Recent research backs the art; [Lee and Anderson \[2017\]](#) show that eyes provide not only important social cues, but are overwhelmingly interpreted as diagnostic of the subject’s emotional state even in presence of competing signals from the lower face. Thus, it is only natural that as the graphics

Authors’ addresses: Gengyan Li, [gengyan.li@inf.ethz.ch](mailto:gengyan.li@inf.ethz.ch), ETH Zurich and Google Inc., Switzerland; Abhimitra Meka, [abhim@google.com](mailto:abhim@google.com), Google Inc., USA; Franziska Mueller, [franziskamu@google.com](mailto:franziskamu@google.com), Google Inc., Switzerland; Marcel C. Buehler, [buehlmar@ethz.ch](mailto:buehlmar@ethz.ch); Otmar Hilliges, [otmar.hilliges@inf.ethz.ch](mailto:otmar.hilliges@inf.ethz.ch), ETH Zurich, Switzerland; Thabo Beeler, [tbeeler@google.com](mailto:tbeeler@google.com), Google Inc., Switzerland.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

© 2022 Copyright held by the owner/author(s).

0730-0301/2022/7-ART166

<https://doi.org/10.1145/3528223.3530130>

community embarks on another wave of photorealistic 3D avatar technologies, sufficient attention is paid to the development of methods and systems that allow for fine-grained photorealistic control over human eye imagery.

Modeling the eye region is challenging due to its complex anatomy (see Fig. 2). The eyeball is a nearly smooth and rigid spheroid, which experiences negligible deformation. Its outer surface consists of a thin clear layer which is highly reflective. The corneal bulge is the protruding center that allows light to enter the eye. The cornea refracts light rays into the eyeball, which are further concentrated by the iris sphincter into the pupil. On the other hand, the surrounding region of the eye consists of multiple non-rigid and deforming materials like skin, eyelashes and eyebrows with fine geometry that exhibit light scattering effects. The motion of the eyeball is controlled by extraocular muscles that allow for fast rotation, while the surrounding skin exhibits smooth non-linear deformation.

Modern eye reconstruction methods such as Bérard et al. [2016, 2019, 2014] achieve impressive results; they synthesize close to photorealistic imagery of the eye, depending on the amount and quality of data captured. Such methods however generally do not reconstruct the periocular region including the eyelids, eyelashes, eyebrows and periorbital skin, which are key to capturing eye expressions such as squinting, drooping, widening etc. An important step towards holistic modeling of the eye region is the work of Schwartz et al. [2020], which proposes a technique to drive eye gaze in mesh based 3D avatars for VR communication. They achieve an impressive level of control over animation of the eye gaze and consistent deformation of the surrounding skin. While their application warrants real-time performance, it suffers from other drawbacks that come with mesh based 3D reconstruction techniques – it cannot model fine structures of eyebrows and lashes and it requires a dense multi-view capture setup to reconstruct the mesh in a pre-processing step. Their method does not disentangle the complex reflectance of the eye and skin from the scene illumination, which makes it unsuitable for the applications of high-quality image synthesis under desired lighting environments.

In the more general setting, Park et al. [2021a] introduce Deformable Neural Radiance Fields, also known as Nerfies, a generalization of the popular Neural Radiance Fields (NeRF) [Mildenhall et al. 2020], which can be used to reconstruct and synthesize the entire face and upper body through casual capture using a single hand-held moving camera. This and other similar techniques including [Park et al. 2021b; Pumarola et al. 2021; Tretschk et al. 2021] use volumetric rendering in continuous 3D space using a multi-layer perceptron to model the canonical shape of an object and learn a warp field on top of it to model dynamic deformations in the video sequence on a per-frame basis. While these techniques excel at modeling both dynamic deformations of skin and thin structures such as hair due to the underlying volumetric representation, they are not animatable and do not provide a solution to modeling surface reflectance and high-frequency light-transport effects such as corneal light reflection and refraction associated with the eye, which prohibits relighting. Works such as Bi et al. [2020]; Srinivasan et al. [2021]; Zhang et al. [2021a,b] have extended volumetric rendering models to further disentangle reflectance and scene lighting, initially

using single point light sources co-located with the camera to illuminate the scene and later generalizing to unconstrained lighting, but only under rigid and static settings.

Inspired by these techniques, we propose a novel hybrid representation that combines the best of mesh based and volumetric reconstruction to achieve animatable synthesis of the eye region under desired environmental lighting. We use a light-weight capture system consisting of a small number of static cameras and lights along with a single hand-held camera with a co-located light source. We model the eyeball surface as an explicit mesh and the canonical shape of the periocular skin region and the interior eye volume using an implicit volumetric representation. The eyeball mesh is used to explicitly compute specular reflections of light rays as well as refraction of the camera rays at the cornea surface. The deformations of the surrounding skin and hair is computed using a learnt warp field over the canonical volume. In order to achieve relightability, we learn the underlying reflectance represented by spherical harmonics coefficients. The outgoing radiance is then computed as a product of the reflectance and environmental illumination in the 3D frequency domain. We jointly optimize for the shape and pose of the eyeball mesh and the density and reflectance in the implicit volume, supervised to minimize the photometric loss between the modeled outgoing radiance and pixel values in the captured video.

We show that our method is able to successfully reconstruct the canonical geometry of the eye region and model its appearance by accurately disentangling shading from diffuse and specular albedos (see Fig. 1). This enables photorealistic view synthesis and relighting by recomputing the shading under novel environmental illumination. Moreover, by interpolating between the learnt warp field of the captured frames and rotating the explicit eyeball mesh, we achieve fine-grained control over the subject's gaze. We demonstrate these capabilities of our method on several subjects with variation in face and eye appearance. We summarize our contributions as follows:

- (1) We propose a hybrid mesh + implicit volumetric representation that allows for modeling of complex reflectance and fine scale geometry of the eye region.
- (2) We design a capture system using only off-the-shelf hardware that allows capturing data to disentangle appearance from scene illumination to achieve high-frequency relighting.
- (3) We demonstrate exciting animated and relit results on several real subjects with varying facial and ocular characteristics.

## 2 RELATED WORK

Automatic reconstruction of human eye and face involves complex modeling of geometry, appearance and deformation. Our work builds on several lines of work that we discuss in greater detail here.

*Eye Modeling and Synthesis.* Photorealistic control of real eye imagery has been important to various application domains. Image-based techniques such as [Buehler et al. 2019; Ganin et al. 2016; He et al. 2019; Kaur and Manduchi 2020, 2021; Kononenko and Lempitsky 2015; Lee et al. 2018; Shrivastava et al. 2017; Zheng et al. 2020] have used eye keypoints, gaze direction or underlying segmentation masks as control signals to synthesize eye images for a given subject. Such works are primarily focused on gaze redirection or gaze

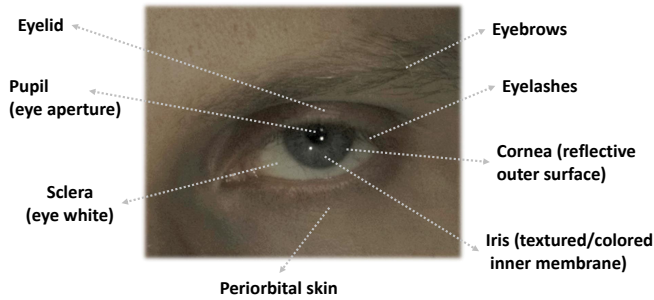


Fig. 2. The eyeball and the periocular region exhibit very different geometry, motion and appearance characteristics. The eyeball is rigid, rotating and reflective/refractive, whereas the surrounding region is non-rigid, smoothly deforming and light-scattering. Hence, we use different representations to model each of these parts.

correction application for portrait images and videos. A complementary line of work [Lu et al. 2015; Sugano et al. 2014; Wang et al. 2018; Wu et al. 2019; Zhang et al. 2015] explores appearance based gaze estimation using analysis by synthesis techniques. Real-time control of gaze is also crucial to avatar based communication in AR/VR settings. Thies et al. [2018, 2016] propose facial reenactment systems using a head-mounted display (HMD) that performs gaze estimation using IR cameras inside the device and photorealistically resynthesizes the subject’s face from an external camera with desired gaze direction. Schwartz et al. [2020] extend this application by fully modeling the eyeball using a parametric eye model. They also model view-dependent texture and periocular deformation using a neural rendering network. While these technique excel at controlling gaze, they do not model the complex reflectance characteristics of the eye and hence cannot resynthesize it under novel lighting.

Bérard et al. [2014] propose a capture system to model subject-specific eyeball characteristics such as the white sclera, the transparent cornea, and the non-rigidly deforming colored iris. Using multi-view close-up images of the eye, they reconstruct the geometry, appearance and deformation of individual parts of the eye to generate a high-quality graphical model for computer graphics applications. Bérard et al. [2016] extend this work by leveraging a database of eye images to generate a prior to constrain the eyeball shape and iris texture and deformation. This enables a lightweight technique to generate graphics models of eyes even from single in-the-wild close up photographs. Bérard et al. [2019] focus on rigging a parametric eye model by reconstructing accurate eye poses from a multi-view capture. They estimate subject-specific geometry properties such as eyeball shape, rotation center, inter-ocular distance and visual axis by optimizing the reprojection error of manually annotated 2D eye keypoints and contours. These techniques aim at modeling the appearance and geometry of the eyeball in great detail, but do not model the fine scale structure and deformations of the periocular face region.

**Volumetric Deformable Surfaces.** Neural Volumes [Lombardi et al. 2019] and NeRF [Mildenhall et al. 2020] introduced implicit neural network based volumetric reconstruction from multi-view imagery

that enables photorealistic novel view synthesis of scenes. Such a representation is particularly well suited for capturing view-dependent appearance, such as on skin, and volumetric effects, including hair, and hence is a suitable representation for capturing digital versions of real people. But in order to realistically animate such a model, it is important to also learn the modes of deformation of skin and hair. [Gafni et al. 2021] densely tracked a face in a monocular video sequence using a parametric face model and reprojected it into 3D space to learn an implicit canonical volume. This allowed them to control face pose and also perform dynamic animations with it by modifying the face model parameters at test time. But this method relies on a differential face model to model the deformation.

Park et al. [2021a] introduced an extension to NeRF, called Nerfies, that learns a canonical volume of a deforming object such as the upper torso of a person, and also generates a learnt per-frame deformation field that warps the volume in order to synthesize the corresponding ground-truth image. [Li et al. 2021; Pumarola et al. 2021; Treitsch et al. 2021; Xian et al. 2021] concurrently proposed methods with similar underlying principles. These methods showed that it is possible to not only extract neutral volume density of a non-rigid scene, but also learn the modes of deformation of such a volume without any priors, using only monocular video capture from a freely moving camera.

**Reflectance Modeling and Relighting.** Methods such as [Shu et al. 2017; Zhou et al. 2019] have achieved environmental relighting of portrait images from casual monocular capture, but produce only diffuse lighting effects and hence do not achieve photorealism. Amongst image-based techniques, Debevec et al. [2000a] was the first to achieve accurate high-frequency relighting of a static face by acquiring the full reflectance field using a Light Stage, a dome containing densely distributed calibrated lights and cameras. More recently, Light Stage data has been used to train neural rendering techniques to relight in-the-wild images [B R et al. 2021; Mallikarjun et al. 2021; Pandey et al. 2021], dynamic sequences [Meka et al. 2019], animatable avatars [Bi et al. 2021] and denser portrait reflectance fields [Sun et al. 2020]. Differentiable ray-tracers have also been used to inverse render portrait images to disentangle reflectance and scene lighting [Dib et al. 2021; Zhang et al. 2021a].

Amongst implicit volumetric reconstruction techniques, Bi et al. [2020] were one of the first to estimate higher-order surface reflectance and visibility in the scene. They use unstructured flash images from a camera with a collocated point light source to learn volume density and reflectance (parameterized as Disney BRDF parameters of diffuse albedo and specular roughness), enabling them to relight the scene under any point light source. Sun et al. [2021] train their model on synthetic data to estimate the light transport field in 3D from multi-view portrait images. Their method achieves convincing reflectance decomposition and relighting, but is limited by the photorealism of the synthetic training data. Boss et al. [2021]; Srinivasan et al. [2021]; Zhang et al. [2021c,b] extend volumetric scene representation to further factorize the radiance field in dense lighting and reflectance and other components such as visibility and indirect illumination, enabling full environmental relighting of the scene. These techniques are constrained to rigid and static settings.



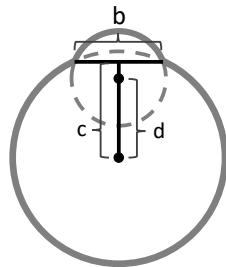
Our method takes inspiration from explicit eyeball modeling methods of Bérard et al. [2016, 2019, 2014] for the eyeball surface, the deformable volumetric reconstruction of Park et al. [2021a] for the periocular region and the volumetric relighting technique of Bi et al. [2020] to disentangle reflectance from environmental lighting using a sparse multi-view capture setup including a single handheld freely-moving camera with a co-located light source, and the lighting visibility model from [Srinivasan et al. 2021] which avoids expensive secondary reflection rays by approximating self occlusion using a neural network.

### 3 HYBRID MODEL

The eye region is comprised of components with vastly different visual properties. The surface of the eye is so specular that it mirrors the environment [Nishino and Nayar 2006] overlaid over the underlying iris, which is heavily distorted through the optical refraction at the corneal surface, especially for side-views. The white sclera is highly scattering, exhibiting veins at different depths inside. The eye is embedded in the periocular region, further adding to the challenge as it combines highly deforming skin and hair from lashes and brows. To address this large diversity, we propose a novel hybrid model that combines the strengths of explicit and implicit representations. In this section we describe the individual parts of the model and how they fit together.

#### 3.1 Explicit Eyeball Surface Model

To model the highly reflective and refractive eyeball surface we represent it with an explicit parametric shape model—concretely we employ a variant of the LeGrand eye model [Le Grand 1957], which consists of two overlapping spheres, but other parametric models could be used as well. The model is fully parameterized by 3 parameters: iris radius  $b$ , which specifies the width of the intersecting circle of the eyeball and cornea spheres, the iris offset  $c$  which specifies the distance of the aforementioned circle from the eyeball center, and the cornea offset  $d$ , which specifies the relative distance of the two sphere centers. These values can then be used to derive the main eyeball radius as well as the cornea radius. Akin to Schwartz et al. [2020] we smoothly blend the eyeball and corneal sphere at the transition (limbus). This blending is controlled by two additional learnable parameters, determining where the transitions on the eyeball and corneal spheres start. The model is discretized as a triangular mesh with 10242 vertices, and enriched with per-vertex displacements, which enable the surface to represent shapes that lie outside the model’s subspace. We then compute the shading normals at each vertex (used for refraction and reflection) by interpolating between the neighboring face normals weighted by their vertex angle. For the index of refraction (IOR), we use the value of 1.4, which is a reasonable value for the human cornea [De Freitas et al. 2013; Patel et al. 1995]. While we assume the eyeball surface to remain static for a subject, its pose changes as a function of gaze. Though eye gaze is often modelled as a 2-DoF rotation only, it is actually more



complex than that [Bérard et al. 2019] and hence we model its motion by a 6-DoF transformation per frame, encoded in axis angle representation, with translation being applied after rotation.

#### 3.2 Implicit Eye Interior and Periocular Model

The highly scattering sclera, the volumetric iris, the transition between eyelid and eye, and especially hair fibers present a formidable challenge for explicit models, and so we argue that an implicit representation is better suited to represent the periocular region and to seamlessly integrate with the eye. Since the skin deforms during acquisition we base our representation on Nerfies [Park et al. 2021a], which employs warp fields to transform frames into a common canonical space, where an MLP network encodes opacity and appearance values as proposed by the seminal NeRF paper [Mildenhall et al. 2020]. The warp field is defined by a secondary MLP which predicts a rotation quaternion and translation vector. We combine the smooth warp field proposed by Nerfies with the rigid transformation from the eyeball surface to explicitly transport rays to a canonical eye volume once they intersect the eyeball surface, where the NeRF encodes the interior.

**3.2.1 NeRF-SHL.** To enable relighting we propose NeRF-SHL (NeRF with Spherical Harmonics Lighting), which extends the traditional NeRF network as proposed by Mildenhall et al. [2020] to additionally predict spherical harmonics coefficients alongside opacity and albedo. These coefficients are used to predict the exiting radiance given an environment map and implicitly approximate various light-transport effects, including reflectance, subsurface scattering, occlusion, and indirect illumination. To better model the combination of diffuse and specular reflectance, we predict two sets of SH coefficients, using 5th order SH for the diffuse and 8th order SH for the specular reflectance. As diffuse reflectance is constant with regards to the outgoing light direction, only the specular SH coefficients are conditioned on the view direction. The network takes as input a 3D World-Space point and its corresponding 3D NeRF-Space point alongside their positional encodings as well as the view direction and outputs diffuse and specular spherical harmonics coefficients for the query point (Fig. 3). These are integrated with the environment illumination and combined to produce the RGB value. As in the original NeRF network, we constrain the opacity to be positive using a relu activation, and we apply a sigmoid to the albedo, similar to how the RGB is constrained in the original network. Additionally, we apply a softplus activation function to the 0th degree spherical harmonics function to force it to be positive, as that particular spherical harmonic corresponds to the uniform function which needs to be positive for any physically correct light transport function. A schematic of the network architecture is shown in Fig. 4.

#### 3.3 Illumination Model

Environmental illumination is represented as a lat-long environment map  $E(\omega_i)$ , which determines the amount of radiance entering the scene from a direction  $\omega_i$ . While relighting the explicit eyeball surface can consume the environment map directly, it needs to be converted to a spherical harmonics representation for relighting the implicit parts of the model to be compatible with our network architecture as introduced in Sec. 3.2.1.



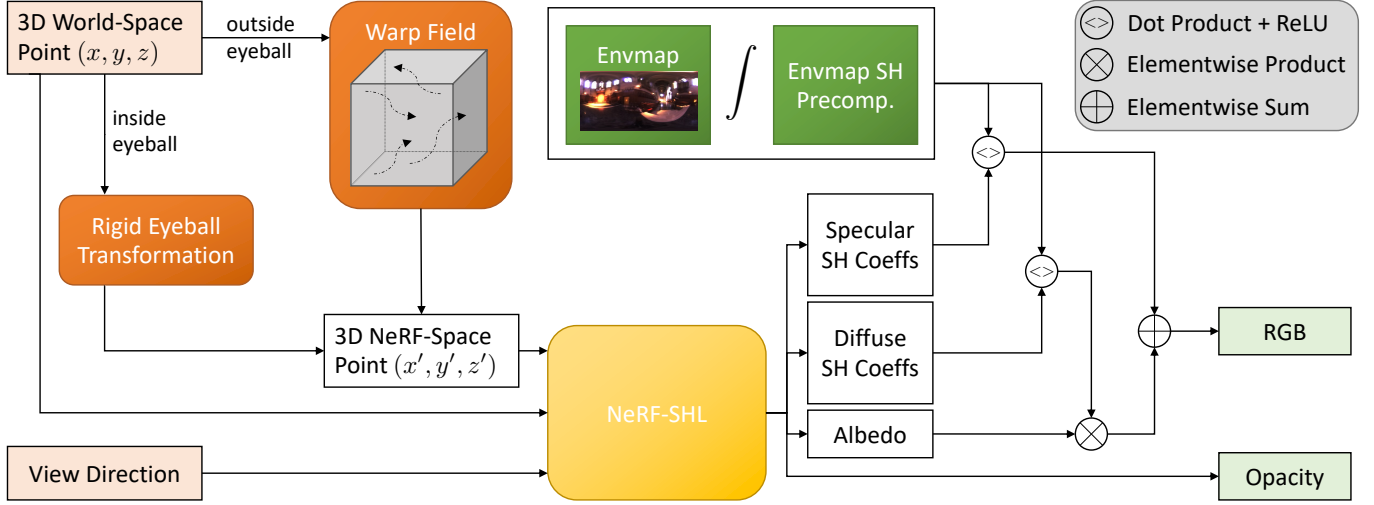


Fig. 3. **Hybrid Model Evaluation.** For every 3D point in world space and a view direction, our hybrid model is trained to output the corresponding RGB color and opacity value. We first transform the 3D world-space point to the canonical NeRF space by using the learned per-frame rigid eyeball transformation or warp field, depending on whether the point lies inside the eyeball. Next, we evaluate our NeRF-SHL to obtain the opacity, the albedo, and specular and diffuse spherical harmonics (SH) coefficients. The SH coefficients are multiplied with the pre-computed SH representation of the environment map and composited with the albedo to obtain the final RGB color for the 3D point.

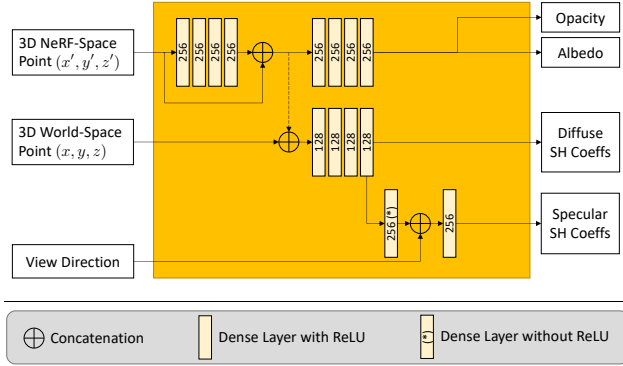


Fig. 4. The architecture for NeRF-SHL can be divided into three branches. The first branch predicts opacity and albedo from the 3D point in canonical NeRF space. For the second branch, we additionally feed the 3D world-space point as input to better model shadowing. Lastly, we add the view direction input for the branch that predicts specular SH coefficients.

**Spherical Harmonics Precomputation.** We start with the standard light transport equation (without emission), at any given point  $\mathbf{x}$  and outgoing light direction (or camera direction)  $\omega_o$ :

$$L_o(\mathbf{x}, \omega_o) = \int_{\Omega} f(\mathbf{x}, \omega_o, \omega_i) L_i(\mathbf{x}, \omega_i) d\omega_i. \quad (1)$$

In our model,  $L_i$  includes all incident light at position  $\mathbf{x}$ , i.e. light from both direct and indirect light as well as incident illumination from subsurface scattering, and hence  $f$  approximates the full light transport at  $\mathbf{x}$  for the entire sphere  $\Omega$ . We can therefore reformulate this equation to instead determine the amount of light transported

from the environment map  $E$  via a point  $\mathbf{x}$  towards  $\omega_o$  as

$$L_o(\mathbf{x}, \omega_o) = \int_{\Omega} f_{\text{tot}}(\mathbf{x}, \omega_o, \omega_i) E(\omega_i) d\omega_i. \quad (2)$$

While intractable with traditional computer graphics methods, it is possible to approximate  $f_{\text{tot}}$  using machine learning techniques. We can approximate  $f_{\text{tot}}$  using the spherical harmonics basis functions  $Y_{lm}$  and their corresponding coefficients  $c_{lm}$

$$f_{\text{tot}}(\mathbf{x}, \omega_o, \omega_i) \approx \sum_{l=0}^{\text{order}} \sum_{m=-l}^l c_{lm}(\mathbf{x}, \omega_o) Y_{lm}(\omega_i), \quad (3)$$

resulting in our approximation

$$L_o(\mathbf{x}, \omega_o) \approx \int_{\Omega} \sum_{l=0}^{\text{order}} \sum_{m=-l}^l c_{lm}(\mathbf{x}, \omega_o) Y_{lm}(\omega_i) E(\omega_i) d\omega_i. \quad (4)$$

By using associativity of sums and integrals as well as distributivity of sums and products, we can reorder Eq. 4 such that the integral can be precomputed independently of the coefficients

$$L_o(\mathbf{x}, \omega_o) \approx \sum_{l=0}^{\text{order}} \sum_{m=-l}^l c_{lm}(\mathbf{x}, \omega_o) \left( \int_{\Omega} Y_{lm}(\omega_i) E(\omega_i) d\omega_i \right). \quad (5)$$

Most notably, the integral is now independent of the conditioning variables,  $\mathbf{x}$  and  $\omega_o$ , which allows to compute the integral once for each spherical harmonics order and degree for a given environment map, making model training in Sec. 5.4 tractable. As described in Sec. 5 we capture our subjects under a mixture of static environment illumination and a moving point light, and precompute a spherical harmonics representation for both. We rotate these representations appropriately to compensate for head-rotation, which can be done efficiently as outlined in Appendix A. Lastly, the SH coefficients of the moving light are scaled as a function of the distance to the

subject to account for squared intensity falloff and the two sets of coefficients are summed to allow relighting of our implicit model.

## 4 MODEL EVALUATION

The presented model can be evaluated for a desired gaze direction and rendered from novel viewpoints under novel illumination.

### 4.1 Gaze Animation

Trained on a diverse set of discrete gaze directions, EyeNeRF allows continuous re-animation of the eye and surrounding region by interpolating the deformation fields and eyeball poses of the discrete training data. Given a novel gaze direction  $\gamma$  the first step is to identify three gaze directions  $\hat{\gamma}_i$  which form a convex hull that contains the target gaze direction. For this, we project all training gaze directions onto the unit sphere by applying their respective rigid eyeball transforms to the unit vector  $(0, 0, 1)^T$  and triangulate those points using the Ball-pivoting Algorithm [Bernardini et al. 1999; Zhou et al. 2018] to obtain a discrete mesh. We compute the intersection of the target gaze direction  $\gamma$  with this mesh. The vertices of the intersected triangle are the desired discrete gazes  $\hat{\gamma}_i$  and the barycentric weights of the intersection point serve as interpolation weights which are then used to blend the warp fields of the three sample gaze directions by linearly interpolating between the warped points. This method limits novel gazes to be within the gaze distribution from the training set, which is reasonable since we capture the full range of motion of the eye during acquisition.

For the eyeball transformation, we consider the translation and rotation components separately. The translation is interpolated using barycentric weights as done for the warp field. The rotation on the other hand is more challenging, since the basic spherical linear interpolation (slerp) leads to unnatural eye motion. Instead, we independently rotate the eyeball for each sample gaze to the target gaze, and slerp the resulting poses sequentially. We found that this yields satisfactory eye motion, but more advanced eye rigging methods, such as the Listings model used in [Bérard et al. 2019], could also be integrated with the proposed hybrid model.

### 4.2 Rendering

Once the interpolated warp field and rigid transformation have been computed, the goal is to render an image for a given camera and environmental illumination. Fig. 5 provides a schematic of the approach presented below.

**4.2.1 Ray Computation and Intersection.** For each pixel in the image we compute the camera ray  $r_c$  and trace it through the scene within appropriate clipping planes. We start by testing if the ray intersects with the explicit eyeball surface using Trimesh [Dawson-Haggerty et al. [n.d.]], a raytracer implemented using Embree [Wald et al. 2014]. In case there is an intersection, we calculate appropriate reflection and refraction rays at the corneal surface using Snell's law, splitting the original ray into three parts: The pre-intersect ray  $r_c$ , the refracted ray  $r_e$ , and the reflected ray  $r_r$ . If there is no intersection, we only need to consider the pre-intersect ray  $r_c$ .

**4.2.2 Point Sampling and Transformation.** As depicted in Fig. 3 and Fig. 5, we sample points along the three rays and transform them to

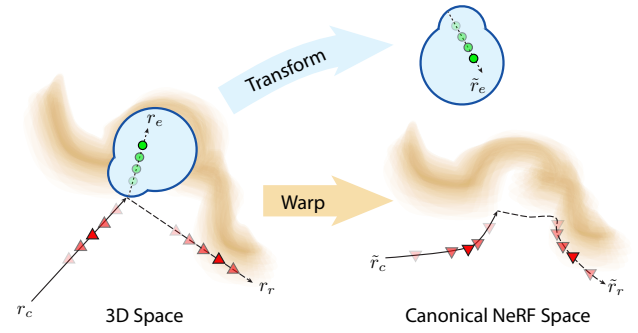


Fig. 5. EyeNeRF uses raytracing to compute reflection  $r_r$  and refraction  $r_e$  rays by intersecting with the explicit eyeball surface. Those rays are then used to raymarch the implicit representation. Points are sampled in 3D Space, and then transformed to the canonical NeRF Space. Points sampled from the refraction ray  $r_e$  are transformed rigidly by the inverse estimated eyeball pose (green points) where points sampled from the other rays are warped non-rigidly by the learned warp field (red triangles).

the canonical NeRF volume. Following [Mildenhall et al. 2020] we employ a combination of equidistant and importance sampling in 3D world-space to determine the sample points. For each sample point of the pre-intersect and reflected rays, we evaluate the warp field neural network and warp the points accordingly, akin to Park et al. [2021a], leading to sample points along the distorted rays  $\tilde{r}_c$  and  $\tilde{r}_r$  in the canonical 3D NeRF-space (red triangles in Fig. 5). For each sample point on the refracted ray, we apply the inverse of the rigid eyeball transform leading to sample points in the canonical 3D NeRF-space of the eye (green circles in Fig. 5).

**4.2.3 Point Shading.** Next, we calculate the contribution of the illumination from the environment map for each sample point. We then query NeRF-SHL to obtain albedo and opacity, as well as specular and diffuse SH coefficients that determine the amount of light transported from the environment map at the queried volume point towards the queried camera ray direction. Incident illumination and transfer function are integrated by multiplying the SH coefficients with the precomputed SH environment coefficients, which is very efficient. The final color value for the sample point is then obtained by multiplying the diffuse lighting with the albedo and adding the specular lighting. For the end point of each ray, i.e. when it is leaving the captured volume, we add one sample point with infinite opacity and a color value. For reflection rays the color is retrieved from the environment map and for the other rays it is set to black. To account for ambiguities in scale when reconstructing the environment map using the mirror ball, we also learn a scale factor which is multiplied with the environment map radiance sample.

**4.2.4 Color Accumulation.** The previous step provided opacity and color for each of the  $N_S$  sample points independently. To obtain the contribution weight of each sample to the final color value, we employ traditional volume rendering techniques [Kajiya and Von Herzen 1984]. To this end, we calculate the accumulated transmittance for each sample point based on the opacity of all previous samples on the ray. In the traditional NeRF setting, the color value

of a single ray  $C(\mathbf{r})$  is computed using the following approximation of the continuous integral along the ray

$$\alpha_i = e^{-\sigma_i \delta_i} \quad T_t = \prod_{i=0}^{t-1} \alpha_i \quad C(\mathbf{r}) = \sum_{t=0}^{N_S} T_t (1 - \alpha_t) \mathbf{c}_t \quad (6)$$

where  $\sigma_i$  and  $\mathbf{c}_i$  are the predicted opacity and color for the  $i$ -th sample, respectively, and  $\delta_i$  represents the distance between the  $i$ -th and  $i + 1$ -th sample. We refer the reader to [Mildenhall et al. \[2020\]](#) for more details. In our implementation, we first compute the color value of the reflected ray separately, and then merge the refracted and pre-intersect ray using the Fresnel Equations [[Fresnel 1868](#)] (assuming unpolarized light), by combining the radiance of the reflected ray with the last sample prior to the intersection using standard alpha compositing rules, effectively placing it behind that sample. We can then treat the resulting combined samples as one ray, resulting in the following system of equations

$$\begin{aligned} \alpha'_i &= e^{-\sigma'_i \delta'_i} \quad T'_t = \prod_{i=0}^{t-1} \alpha'_i \quad C(\mathbf{r}') = \sum_{t=0}^{N'_S} T'_t (1 - \alpha'_t) \mathbf{c}'_t \\ \alpha_{\text{comb}} &= \alpha_k (1 - f) \quad c_{\text{comb}} = \frac{(1 - \alpha_k) \mathbf{c}_k + f \alpha_k C(\mathbf{r}')}{\alpha_{\text{comb}}} \\ \mathbf{c}''_i &= \begin{cases} \mathbf{c}_i & i \neq k \\ \mathbf{c}_{\text{comb}} & i = k \end{cases} \quad \alpha''_i = \begin{cases} \alpha_i & i \neq k \\ \alpha_{\text{comb}} & i = k \end{cases} \\ \alpha_i &= e^{-\sigma_i \delta_i} \quad T_t = \prod_{i=0}^{t-1} \alpha_i \quad C(\mathbf{r}) = \sum_{t=0}^{N_S} T_t (1 - \alpha'_t) \mathbf{c}''_t \end{aligned} \quad (7)$$

where  $f$  is the Fresnel factor,  $k$  is the index of the last sample prior to the intersection,  $\alpha'$ , etc. refer to values sampled along the reflected ray, and  $\mathbf{c}''_i$  refers to the samples of the combined ray (which are the same as the ones along the pre-intersect and refracted ray, other than the sample prior to the eyeball intersection).

Rays that do not intersect the eye model are computed using the original NeRF ray marching method (see Eq. 6).

## 5 MODEL TRAINING

High quality synthesis requires high-quality data. While there are several publicly available eye image datasets, they are unfortunately not directly suited for our purposes of view, gaze, and illumination synthesis. The majority of these datasets are tailored for the task of gaze-tracking [[Fuhl et al. 2021](#); [Fusek 2018](#); [Kim et al. 2019](#); [Tonsen et al. 2017](#); [Wood et al. 2015](#); [Wu et al. 2020](#); [Zhang et al. 2015](#)] while others cater to different problems such as pupil detection [[Tonsen et al. 2016](#)], eye closure detection [[Song et al. 2014](#)] or eyelash segmentation [[Xiao et al. 2021](#)]. While there are datasets that aim at modeling high-quality eyes and periocular region [[Bérard et al. 2019, 2014](#)], these are not suited for relighting purposes.

Hence we built our own capture system that provides sufficient signal for the task of gaze reanimation, view synthesis and relighting of the periocular region. We detail our hardware setup and the capture protocol below.

### 5.1 Capture System

We aim to minimize the complexity of our hardware to make our solution cost and space effective. The subject sits on a chair in the

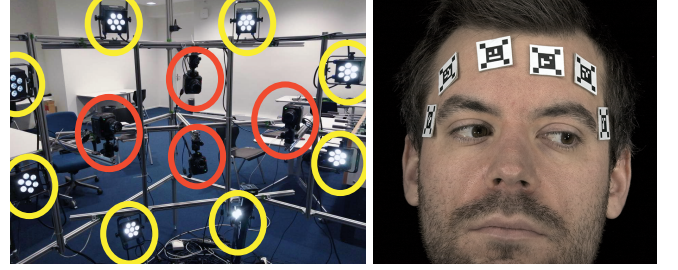


Fig. 6. **left:** Our static setup consists of 4 high-quality cameras (red) arranged in a diamond-shape and surrounded by 8 illuminators (yellow). **right:** A set of AR markers is attached to the forehead of the subject to track head movement as well as relative camera motion.

center of the setup shown in Fig. 6. Our multi-view setup consists of 4 high quality, hardware synchronized cameras (Z-Cam E2 4K) fixed in the frontal hemisphere of the subject, as well as a small, lower quality mobile GoPro® Hero9 camera which is moved freely by hand by the operator. The moving camera contains a co-located LED light, which is crucial for relighting, as we demonstrate in the next section. The subject is lit with 8 white LED point lights that are nearly uniformly located over a  $100^\circ$  field-of-view (FOV) in front of the subject. The cameras span roughly  $75^\circ$  horizontal FOV and  $25^\circ$  vertical FOV in front of the subject, arranged in a diamond shape.

In order to account for the free head motion of the subject, we affix a set of small calibration markers on their forehead. These calibration markers are used to localize both the subject's head and the moving camera with respect to the static setup. While we assume that the eyeball is rigidly attached to the head, our method makes no such assumption for the periocular region which can experience strong deformations due to facial muscles and skin.

### 5.2 Capture Protocol

We capture subjects under four different conditions. In the first condition, the subject is instructed to follow the mobile camera with their gaze while keeping their head static and forward facing. The mobile camera is translated freely, while orienting it towards the subject's head, covering about  $60^\circ$  horizontally and  $30^\circ$  vertically. In this setting, we only use the 8 static lights. Since the eye gaze follows the mobile camera, we know the gaze direction and get a good multi-view coverage of the eye from the 4 static cameras.

In the other three conditions, the subject is instructed to keep their gaze focused on one of the four static cameras, switching their gaze between them when instructed. Instead of moving their gaze, they are instructed to rotate their head around, while attempting to keep one eye roughly stationary, to keep that eye in frame and at the same distance from the cameras. In order to more easily keep track of where that eye was, we place a tripod below the subject's head for reference (not as chin rest). In these settings we get good viewpoint coverage of the periocular region from the 4 cameras.

Each of these three conditions models a different illumination scenario; static lights, mobile light, and both. This provides a good mixture of frames where we have relatively flat lighting which is useful for reconstructing the geometry, and very high frequency lighting which can be used to learn shading and shadowing.



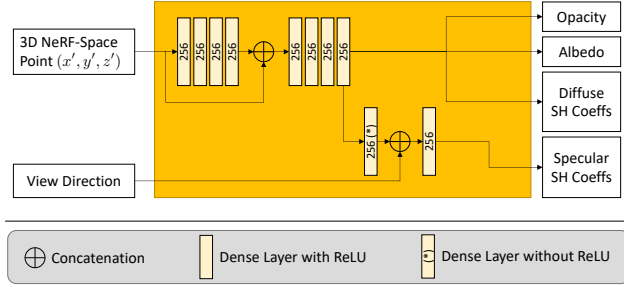


Fig. 7. To improve initial network convergence, we start training with the simplified architecture shown above, and later on continue to train the full architecture as depicted in Fig. 4. The main difference is that here the diffuse SH coefficients only depend on the 3D NeRF-Space points, where in the full model they also depend on the 3D World-Space points, in order to better model shading.

For the demonstrations in this paper, we capture the performances of 3 subjects with a variety of face shapes and eye colors.

### 5.3 Data Preprocessing

**5.3.1 Cameras.** Intrinsics for all cameras are calibrated using a checkerboard pattern. Extrinsics of the static cameras are recovered from the same calibration process. The extrinsics of the mobile camera are estimated from the marker tags on the subjects forehead using OpenCV. By also estimating the rigid transformation between the static cameras and the subject we can relate all cameras into the same world frame, registered to the subjects head.

**5.3.2 Environment Illumination.** We compute an HDR environment map from a series of images from a mirror sphere with varying exposure, similar to [Debevec et al. 2000b]. We capture an environment map for the static illumination and the mobile light separately. Both the mobile and static lights are captured only once; we model the light motion using rotation and falloff as described in Section 3.3.

**5.3.3 Model Initialization.** Our method relies on an initial estimate of the eyeball pose and shape. As the subject is instructed to look at either the mobile camera or one of the static cameras, the initial eye pose can be estimated from the line-of-sight that connects the eyeball and camera centers. We manually initialize eyeball pose and shape from three frames, roughly placing the eyeball in the correct position and shape in a 3D modeling software (Blender). Note that the initialization can be approximate, as it will be refined throughout the training process as shown in Fig. 13.

### 5.4 Network Training

As the main training loss, we use the mean squared error in sRGB space between the equidistant and importance sampled RGB values and the target pixel in the training image. The loss is computed on the outputs of the coarse and fine network

$$l_{\text{im}} = \|\text{srgb}(x') - \text{srgb}(x_f)\|_2^2 + \|\text{srgb}(x'), \text{srgb}(x_c)\|_2^2. \quad (8)$$

As the sRGB transformation is not meaningful above 1, we use a linear transformation for such values. To encourage our hybrid model to represent specular reflection on the sclera by the explicit

eyeball model, we ignore sclera pixels above a defined threshold when training the implicit volume.

In addition, we employ the non-negative SH loss for regularization. Each iteration, we randomly sample ten random directions, check if there are any negative predicted SH response functions, and then apply an  $l_2$  loss on negative values. This avoids dead zones if the SH becomes negative. Furthermore, we observe that the diffuse shading is almost never non-positive anyway, and only apply this loss to the specular part of the shading. In order to reduce the uncertainty between diffuse and specular shading, we furthermore apply an  $l_2$  loss to the specular coefficients. These losses are applied to the mean across all sample points.

$$l_{\text{noneg}} = \mathbb{E}_{\omega_i \sim \Omega} [-\min(0, \sum_{l=0}^{\text{order}} \sum_{m=-l}^l c_{lm}(\mathbf{x}, \omega_o) Y_{lm}(\omega_i))] \quad (9)$$

$$l_{\text{spec}} = \frac{1}{(\text{order} + 1)^2} \sum_{l=0}^{\text{order}} \sum_{m=-l}^l \|c_{lm}(\mathbf{x}, \omega_o)\|_2^2$$

We then apply an L2 loss on the per-vertex offsets ( $N_V = 10242$ ), to avoid strong deviations from the underlying analytical model

$$l_{\text{off}} = \sum_{i=0}^{N_V} \text{offset}_i^2. \quad (10)$$

Finally, we apply the same elastic regularization as used by Nerfies [Park et al. 2021a] onto the warp field. We refer the reader to their paper on more details on how this loss term is computed. This results in our final loss function

$$l_{\text{tot}} = \lambda_{\text{im}} l_{\text{im}} + \lambda_{\text{noneg}} l_{\text{noneg}} + \lambda_{\text{spec}} l_{\text{spec}} + \lambda_{\text{elastic}} l_{\text{elastic}} + \lambda_{\text{off}} l_{\text{off}} \quad (11)$$

where we use the empirically chosen weights  $\lambda_{\text{im}} = 1$ ,  $\lambda_{\text{noneg}} = 1\text{e-}2$ ,  $\lambda_{\text{spec}} = 5\text{e-}4$ ,  $\lambda_{\text{elastic}} = 1\text{e-}3$ ,  $\lambda_{\text{off}} = 1\text{e-}6$ .

We train EyeNeRF in three stages, using slightly different architectures, on a total of 16 Nvidia V100s GPUs. First, we use the simplified architecture for our NeRF-SHL as shown in Fig. 7 to focus on learning the eyeball model parameters and per-frame rigid transformations. During this stage, every 50000 iterations, we additionally reset and reinitialize all learnable parameters other than the ones for our parametric model. We observe that this re-initialization greatly improves the final quality. This is due to two reasons. On the one hand, every time we reinitialize the volume, we remove all possible biases that may have been baked into the eyeball volume, for example due to view dependent effects. On the other hand, by resetting the volume we effectively blur it out, making the spatial gradients much smoother and therefore easier to learn with. For better signal when optimizing the eyeball pose and shape, we do not ignore the sclera pixels during this stage. This takes a total of approximately three days when trained on 8 of the 16 GPUs. Second, we start training with our main NeRF-SHL architecture as shown in Fig. 4 to obtain initial network weights (for the warp field network and the NeRF-SHL network) helping robustness of the third step. Note that this second step can take place in parallel with the first as they do not depend on each other, and also takes approximately three days, using the other 8 GPUs. Third, once the eyeball model parameters and transformations have converged, we load them into

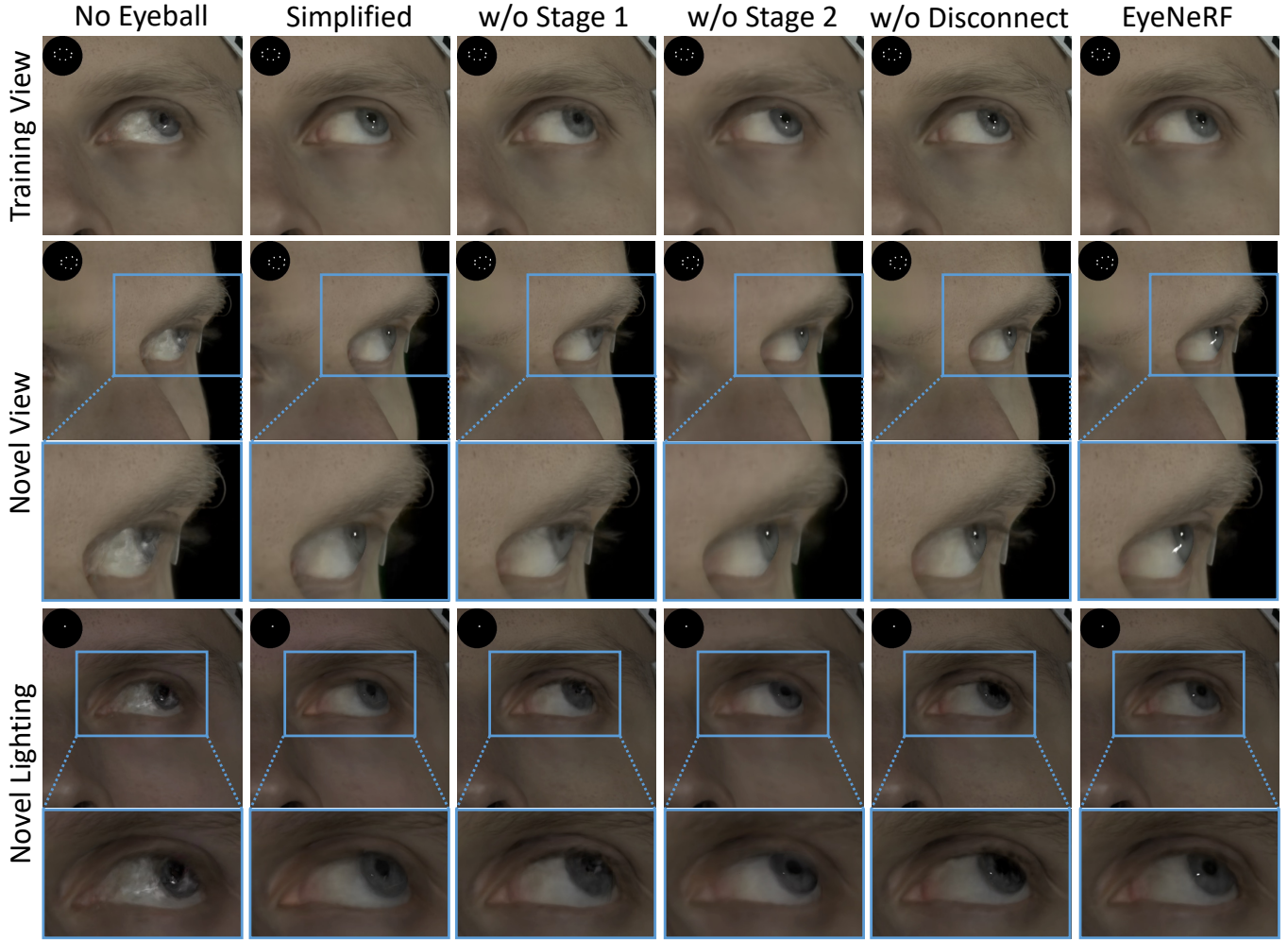


Fig. 8. **Ablation Study.** We evaluate different design choices for EyeNeRF. Without the explicit eyeball model, the volumetric model fails to represent reflection and refraction, leading to strong artifacts. Without the first training stage, the eyeball pose and shape is inaccurate (see novel side view) hence many specular effects on the eyeball are not reproduced. Without the second training stage, the quality of the periorbital region is reduced, yielding significant blur in the images. Results from the simplified architecture or without disconnect look similar to EyeNeRF in the training view. However, they exhibit artifacts on the iris and sclera for novel views or novel illumination conditions.

our main architecture and continue training for roughly 100,000 iterations. During this final training stage, we disconnect the dotted connection in Fig. 4 for points inside the eyeball by conditionally zeroing those values, to stronger condition the estimated SH lighting contribution on the world space points. We also disable the contribution from the specular shading, in order to encourage our network to model as much as possible using direct reflections. During this stage, we only use 8 GPUs, and train for roughly 24 hours. As the first two stages are trained in parallel, our total pipeline can be trained in around four days.

## 6 RESULTS AND EVALUATION

In this section, we discuss various results and evaluations of our method. We opted to suppress rendering of the specular highlights on the sclera for most results in this paper by using an IoR of zero

for the sclera surface, because even though the gross position of the highlights on the optimized sclera are correct, their appearance is visually off since the spatial resolution of our parametric eye model is not high enough to represent the high frequency surface variation caused by the lacrimal liquid and conjunctiva (see Fig. 9). We start with an ablation study (Sec. 6.1) to evaluate our design choices for EyeNeRF. Next, we compare to different baselines for gaze redirection and relighting (Sec. 6.2). Section 6.3 provides additional qualitative results such as eye animation, view synthesis, and relighting for different subjects as well as intrinsic decompositions into albedo and shading.

### 6.1 Ablation Study

We evaluate the following settings in our ablation study. For network architectures and training stages, please refer to Section 5.4.



Fig. 9. While the gross position of the highlights on the optimized sclera are correct, their appearance is visually off since the spatial resolution of our parametric eye model is not high enough to represent the high frequency surface variation caused by the lacrimal liquid and conjunctiva.

Table 1. Results from our quantitative ablation study. Our full method outperforms or achieves comparable performance with the ablation settings on most metrics for all 3 tasks.

	Novel View		Regazing		Relighting	
	MSE ↓	SSIM ↑	MSE ↓	SSIM ↑	MSE ↓	SSIM ↑
No Eyeball	1.10e-3	0.843	1.24e-3	0.838	9.62e-4	0.813
Simplified	9.74e-4	0.843	7.49e-4	0.853	7.53e-4	0.825
w/o Stage 1	1.03e-3	0.847	1.07e-3	0.844	8.68e-4	0.822
w/o Stage 2	1.78e-3	0.814	8.1e-4	0.835	7.58e-4	0.816
w/o Disc.	7.78e-4	0.854	<b>7.22e-4</b>	0.855	<b>7.19e-4</b>	0.828
EyeNeRF	<b>7.67e-4</b>	<b>0.863</b>	7.23e-4	<b>0.857</b>	7.23e-4	<b>0.829</b>

- **No Eye Model:** We do not use the explicit eye model. There is no explicit refraction or reflection and all points are transformed using the warp field.
- **Simplified Architecture:** We only use the simplified architecture for NeRF-SHL (Fig. 7) and do not switch to the main branched architecture (Fig. 4). Everything is trained in a single training stage.
- **Without Stage 1:** We skip the first training stage in which the eyeball model parameters and rigid transforms are pre-trained. All parameters are directly trained using the main branched architecture.
- **Without Stage 2:** We skip the second training stage in which the NeRF-SHL and warp field weights are pretrained. Instead, the weights are trained from scratch using the already refined eyeball poses from the first training stage.
- **Without Disconnect:** Same as the final version but during the final training stage, the dotted connection in Fig. 4 is not disconnected.

Fig. 8 shows results of the ablation study. Please refer to our supplemental video for more visuals. Without the eyeball model, the specular highlights on the eye need to be modelled by the volumetric representation. This leads to severe artifacts on the eyeball, especially for novel views or novel lighting scenarios. Without the first training stage, the eyeball pose and shape is inaccurate as shown in the novel side view. Due to the incorrectly placed eyeball, the specular reflections are missing in many cases. Without the second training stage, the quality of the periocular region degrades significantly which is especially visible in areas with high-frequency details such as the eyebrows. Whereas the results with the simplified

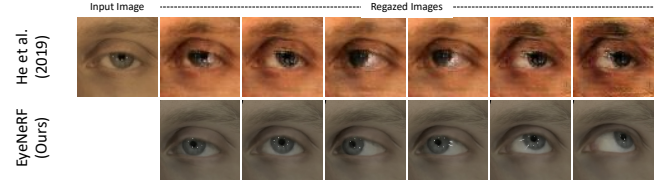


Fig. 10. We compare with [He et al. 2019] for regazing. EyeNeRF controls the gaze more accurately and preserves the identity better. Please note that [He et al. 2019] operates on a single input image only, where ours is a much more involved setting.

architecture or without disconnect look similar to EyeNeRF from the training views, the iris and sclera quality reduces significantly for novel views or novel lighting.

We show results from the quantitative ablation study for novel view synthesis, regazing and relighting in Table 1. For novel view synthesis, we select a number of frames, from which we exclude one of the four camera views during training as a held-out test view.

For both regazing and relighting, we define a set of held-out test frames, for which all camera views are excluded from training. For regazing, we use frames from the sequence where the captured subject holds their head still and looks at the moving camera. For relighting, we use the sequence where only a moving light is used for illumination. However, we note that our initial estimates for gaze (see Section 5.3.3) are not completely accurate. For the test frames, we optimize the eyeball pose using the same strategy as for training frames. We use the so-computed eyeball pose for synthesizing the test frames using our trained model and perform quantitative comparison with the ground-truth to obtain the results reported in Table 1.

We compute SSIM and MSE as quantitative image comparison metrics for models trained for all settings in the ablation study. As our ablation study conditions primarily differ in how the eye is modeled, we use a smaller 300x300 crop around the eye while computing the metric, in order to more strongly focus on the eye itself. Please note that these image metrics have their limitations, e.g. over-penalizing misalignments or under-penalizing blur. Nonetheless, the results indicate that our full method outperforms or is on par with the other ablation study conditions.

## 6.2 Comparisons

We compare EyeNeRF to existing methods for regazing [He et al. 2019] and relighting [Pandey et al. 2021].

**Regazing.** We perform regazing on subject 1 and compare with He et al. [2019] in Fig. 10. EyeNeRF can synthesize more extreme gaze angles, it preserves the identity better, and shows less artifacts. Note that their method is defined for a different setting, as it can be applied for a single image, and works across multiple identities. Our setting is much closer to that of Schwartz et al. [2020], who also use multi-view footage of the same subject for regazing. Unfortunately a direct comparison with Schwartz et al. [2020] is difficult because their data capture setting uses a dense multi-view setup, and their implementation is not available publicly or on request. Nonetheless, we show a demonstrative comparison in Fig. 11. Our method enables



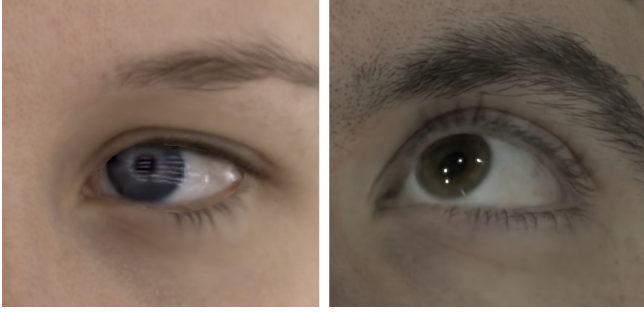


Fig. 11. We compare an image synthesized by our method (right) with an image generated by Schwartz et al. [2020] [©Schwartz et al.] (left). While Schwartz et al. [2020] show that they can perform view synthesis and regazing, their method does not allow for capturing high frequency skin/eye reflectance, perform relighting or capturing and synthesizing thin structures such as eyelashes and eyebrows in 3D.

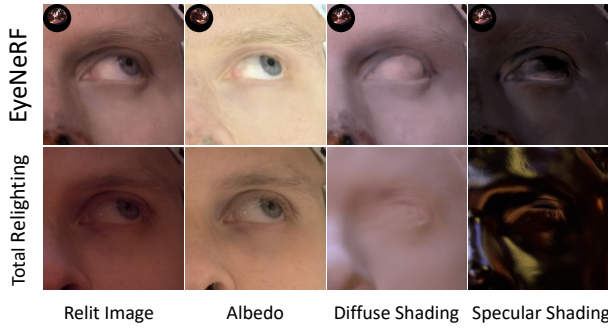


Fig. 12. We compare with Pandey et al. [2021] for relighting. EyeNeRF produces higher quality relighting, particularly the environmental reflections on the eye.

additional capabilities such as high-quality relighting and 3D capture of thin structures associated with the eye region.

**Relighting.** We compare environmental relighting on a single frame against the state-of-the-art portrait relighting method of Pandey et al. [2021] in Fig. 12. They use high-quality dense Light Stage data to train their technique. While they do estimate surface normals, they do not estimate the full 3D geometry of the face, hence they are unable to model the full light transport. EyeNeRF can synthesize very high quality relighting, including self-shadowing, particularly in the eye region.

### 6.3 Additional Results

**Eyeball Model Refinement.** In Fig. 13, we demonstrate how EyeNeRF successfully refines the eyeball geometry during the course of training. This leads to more accurately synthesized highlights and a better aligned eyeball contour.

**Synthesizing View, Lighting, and Gaze.** In Fig. 15, we compare our synthesized result to an unseen ground-truth image. We show additional results in Fig. 14. For each subject, we show results for deviating from a training frame by changing a single property (view,

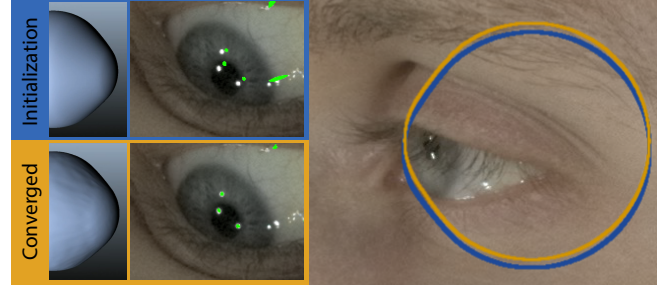


Fig. 13. The initial estimate (blue) for the eyeball pose and shape is refined (orange) during training (Sec. 5.4). **left row:** the shape changes and the sclera becomes bumpy while the cornea remains smooth. **middle row:** the synthesized highlights (green) are initially off but coincide with the ground-truth highlights for the converged shape. **right:** eyeball contours of subject 1 at the beginning (blue) and at the end of training (orange).

lighting condition, or gaze directions). Furthermore, EyeNeRF also synthesizing compelling results for combinations of unseen scene properties (view, lighting, and gaze) or relighting using an environment map.

**Intrinsic Decomposition.** In Fig. 16, we show the final rendered image for different subjects as well as decompositions into albedo, diffuse and specular shading.

## 7 DISCUSSION & LIMITATIONS

Although we demonstrate exceptional overall quality, our method does still have a number of shortcomings and limitations.

One major drawback is the requirement of multiview, multi-lighting, multi-gaze data. As previously mentioned, very few (if any) datasets exist which contain all of the above. Furthermore, capturing this kind of data, although lightweight compared to many professional capture setups such as a full light stage, still requires very high quality cameras and good lighting. Additionally, as we capture the subjects under a variety of conditions, we require around 40 minutes per subject. To reduce complexity we rely on AR tags to track the rigid pose, which requires attaching small tags to the subjects face. Relying instead on a face tracker would be a more convenient alternative.

Another drawback is the long training and evaluation time. Our full pipeline takes around 4 days to train on 8 Nvidia V100s, and currently requires about thirty seconds to render an 800x800 sized image. The majority of the computational cost lies in the NeRF-SHL and NeRFies networks and recently there have been a surge of papers that greatly accelerate training such networks (e.g. [Müller et al. 2022]).

A more fundamental limitation of our method is the inability to model high frequency effects other than the cornea reflection. NeRF and NeRF-like methods often already struggle with exceptionally high frequency details, such as with the eyelashes, especially when time-varying due to deformation. A lot of the detail loss can be attributed to the deformation field, which struggles to accurately warp the frames to the canonical volume, e.g. for the upper eyelid.

By approximating the light transport equation using spherical harmonics, we are inherently limited by the frequency of the SH



Fig. 14. Here we show synthesis results for different subjects. In the three left columns, we vary only one attribute (view, lighting, or gaze) from a training image. However, EyeNeRF can also generate combinations of unseen attributes (4th column) or use arbitrary environment maps for relighting (5th column).



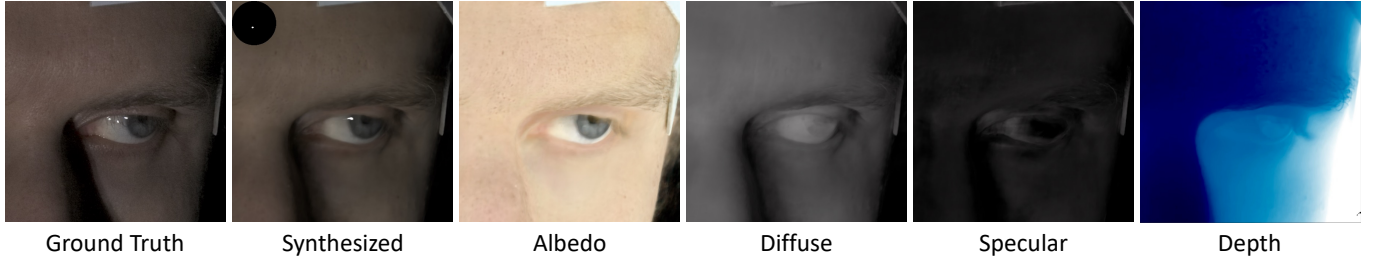


Fig. 15. We show a side-by-side comparison with a ground-truth image that has not been seen during training (first two columns). The third to sixth columns show the corresponding intrinsic decomposition.



Fig. 16. Intrinsic decomposition results for different subjects. EyeNeRF is able to disentangle albedo, diffuse and specular shading.

degree we choose to use. As such, we are unable to model very high frequency effects, such as strongly specular reflections on the skin. As we model the entire light transport function using these functions, this further implies that we have similar issues with high frequency shadows for example. Since we designed our method for human faces these problems are less of an issue, but we do not expect EyeNeRF to work nearly as well in a more general setting. Going to higher order SH will quickly become intractable as the number of coefficients grows in a squared manner. A potential solution could be to pool information spatially across a patch of skin and thus

provide more constraints to estimate the coefficients, or to explore alternative reflectance representations (i.e. data driven priors).

Furthermore, we are currently unable to reconstruct the high-frequency sclera reflections since the surface resolution of our eye-ball model is not high enough. Increasing the resolution however will only be beneficial up to a certain amount since learning high-frequency spatial variation in the current setting is also limited by the image resolution. At the current image resolution pixels are either specular or not, which does not provide good gradients for



learning. In addition we are currently not modeling the relative deformation of the conjunctiva and the sclera, which leads to blurred veins, as well as the pupil dilation. Exploring deformation fields within the eye itself could potentially address this.

Finally, we currently do not model expressions (lid closing, squinting, smiling, etc), and assume the periocular shape is entirely correlated to eye gaze. Additionally capturing expression and properly disentangling gaze and expression would be an exciting avenue for future work.

## 8 CONCLUSION

We presented EyeNeRF, a novel hybrid model combining the strengths of explicit parametric surface models with implicit neural representations, specifically designed to represent the greatly varying visual properties of the periocular region. Our model enables explicit control over the gaze and to render realistic images from novel viewpoints under novel illumination. EyeNeRF is well suited to generate visually rich training data of the periocular area to train ML models, i.e. for gaze prediction, or to create visually compelling content for VFX productions. While our method has several limitations (see Sec. 7) which inspire future work, it is a leap forward towards realistic reproduction of one of the most challenging and most important areas of the human face.

## ACKNOWLEDGMENTS

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme grant agreement No 717054.



## REFERENCES

- Mallikarjun B R, Ayush Tewari, Tae-Hyun Oh, Tim Weyrich, Bernd Bickel, Hans-Peter Seidel, Hanspeter Pfister, Wojciech Matusik, Mohamed Elgharib, and Christian Theobalt. 2021. Monocular Reconstruction of Neural Face Reflectance Fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Pascal Bérard, Derek Bradley, Markus Gross, and Thabo Beeler. 2016. Lightweight Eye Capture Using a Parametric Model. *ACM Trans. Graph.* 35, 4 (2016). <https://doi.org/2897824.2925962>
- P. Bérard, D. Bradley, M. Gross, and Thabo Beeler. 2019. Practical Person-Specific Eye Rigging. *Computer Graphics Forum* 38 (05 2019), 441–454. <https://doi.org/10.1111/cgf.13650>
- Pascal Bérard, Derek Bradley, Maurizio Nitti, Thabo Beeler, and Markus Gross. 2014. High-Quality Capture of Eyes. *ACM Trans. Graph.* 33, 6, Article 223 (nov 2014), 12 pages. <https://doi.org/10.1145/2661229.2661285>
- Fausto Bernardini, Joshua Mittleman, Holly Rushmeier, Cláudio Silva, and Gabriel Taubin. 1999. The Ball-Pivoting Algorithm for Surface Reconstruction. *IEEE Transactions on Visualization and Computer Graphics* 5, 4 (oct 1999), 349–359. <https://doi.org/10.1109/2945.817351>
- Sai Bi, Stephen Lombardi, Shunsuke Saito, Tomas Simon, Shih-En Wei, Kevyn Mcphail, Ravi Ramamoorthi, Yaser Sheikh, and Jason Saragih. 2021. Deep relightable appearance models for animatable faces. *ACM Transactions on Graphics (TOG)* 40, 4 (August 2021).
- Sai Bi, Zexiang Xu, Kalyan Sunkavalli, Miloš Hašan, Yannick Hold-Geoffroy, David Kriegman, and Ravi Ramamoorthi. 2020. Deep Reflectance Volumes: Relightable Reconstructions from Multi-View Photometric Images. [arXiv:2007.09892](https://arxiv.org/abs/2007.09892) [cs.CV]
- Mark Boss, Raphael Braun, Varun Jampani, Jonathan T. Barron, Ce Liu, and Hendrik P.A. Lensch. 2021. NeRD: Neural Reflectance Decomposition from Image Collections. In *IEEE International Conference on Computer Vision (ICCV)*.
- Marcel C. Buehler, Seonwook Park, Shalini De Mello, Xucong Zhang, and Otmar Hilliges. 2019. Content-Consistent Generation of Realistic Eyes with Style. In *International Conference on Computer Vision Workshops (ICCVW)* (Seoul, Korea).
- Dawson-Haggerty et al. [n.d.]. *trimesh*. <https://trimesh.org/>
- Carolina De Freitas, Marco Ruggeri, Fabrice Manns, Arthur Ho, and Jean-Marie Parel. 2013. In vivo measurement of the average refractive index of the human crystalline lens using optical coherence tomography. *Optics letters* 38, 2 (2013), 85–87.
- Paul Debevec, Tim Hawkins, Chris Tchou, Haarm-Pieter Duiker, and Westley Sarokin. 2000a. Acquiring the Reflectance Field of a Human Face. In *SIGGRAPH*. New Orleans, LA. <http://ict.usc.edu/pubs/Acquiring%20the%20Re%EF%AC%82ectance%20Field%20of%20a%20Human%20Face.pdf>
- Paul Debevec, Tim Hawkins, Chris Tchou, Haarm-Pieter Duiker, Westley Sarokin, and Mark Sagar. 2000b. Acquiring the Reflectance Field of a Human Face. In *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '00)*. ACM Press/Addison-Wesley Publishing Co., USA, 145–156. <https://doi.org/10.1145/344779.344855>
- Abdallah Dib, Cedric Thebault, Junghyun Ahn, Philippe-Henri Gosselin, Christian Theobalt, and Louis Chevallier. 2021. Towards High Fidelity Monocular Face Reconstruction with Rich Reflectance using Self-supervised Learning and Ray Tracing. [arXiv:2103.15432](https://arxiv.org/abs/2103.15432) [cs.CV]
- Augustin Fresnel. 1868. *Oeuvres complètes d'augustin Fresnel*. Vol. 2. Imprimerie impériale.
- Wolfgang Fuhl, Gjergji Kasneci, and Enkelejd Kasneci. 2021. TEyeD: Over 20 million real-world eye images with Pupil, Eyelid, and Iris 2D and 3D Segmentations, 2D and 3D Landmarks, 3D Eyeball, Gaze Vector, and Eye Movement Types. [arXiv:2012.02115](https://arxiv.org/abs/2012.02115) [eess.IV]
- R. Fusek. 2018. Pupil localization using geodesic distance. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 11241 LNCS (2018), 433–444. [https://doi.org/10.1007/978-3-030-03801-4\\_38](https://doi.org/10.1007/978-3-030-03801-4_38)
- Guy Gafni, Justus Thies, Michael Zollhöfer, and Matthias Nießner. 2021. Dynamic Neural Radiance Fields for Monocular 4D Facial Avatar Reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 8649–8658.
- Yaroslav Ganin, Daniil Kononenko, Diana Sungatullina, and Victor Lempitsky. 2016. DeepWarp: Photorealistic Image Resynthesis for Gaze Manipulation. [arXiv:1607.07215](https://arxiv.org/abs/1607.07215) [cs.CV]
- Zhe He, Adrian Spurr, Xucong Zhang, and Otmar Hilliges. 2019. Photo-Realistic Monocular Gaze Redirection Using Generative Adversarial Networks. In *IEEE International Conference on Computer Vision (ICCV)* (Seoul, Korea). IEEE.
- Joseph Ivanic and Klaus Ruedenberg. 1996. Rotation matrices for real spherical harmonics. Direct determination by recursion. *The Journal of Physical Chemistry* 100, 15 (1996), 6342–6347.
- James T Kajiya and Brian P Von Herzen. 1984. Ray tracing volume densities. *ACM SIGGRAPH computer graphics* 18, 3 (1984), 165–174.
- Harsimran Kaur and Roberto Manduchi. 2020. EyeGAN: Gaze-Preserving, Mask-Mediated Eye Image Synthesis. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*.
- Harsimran Kaur and Roberto Manduchi. 2021. Subject Guided Eye Image Synthesis With Application to Gaze Redirection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 11–20.
- 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 SIGCHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland UK) (CHI '19). ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3290605.3300780>
- Daniil Kononenko and Victor Lempitsky. 2015. Learning to look up: Realtime monocular gaze correction using machine learning. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4667–4675. <https://doi.org/10.1109/CVPR.2015.7299098>
- Yves Le Grand. 1957. *Light, Color and Vision*. Wiley, New York, NY, USA.
- Daniel H. Lee and Adam K. Anderson. 2017. Reading What the Mind Thinks From How the Eye Sees. *Psychological Science* 28, 4 (2017), 494–503. <https://doi.org/10.1177/0956797616687364> [arXiv:https://doi.org/10.1177/0956797616687364](https://arxiv.org/abs/https://doi.org/10.1177/0956797616687364) PMID: 28406382.
- Kangwook Lee, Hoon Kim, and Changho Suh. 2018. Simulated+Unsupervised Learning With Adaptive Data Generation and Bidirectional Mappings. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net. <https://openreview.net/forum?id=SkHDoG-Cb>
- Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. 2021. Neural Scene Flow Fields for Space-Time View Synthesis of Dynamic Scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. 2019. Neural Volumes: Learning Dynamic Renderable Volumes from Images. *ACM Trans. Graph.* 38, 4, Article 65 (July 2019), 14 pages.
- Feng Lu, Yusuke Sugano, Takahiro Okabe, and Yoichi Sato. 2015. Gaze Estimation From Eye Appearance: A Head Pose-Free Method via Eye Image Synthesis. *IEEE Transactions on Image Processing* 24, 11 (2015), 3680–3693. <https://doi.org/10.1109/TIP.2015.2445295>

- B R Mallikarjun, Ayush Tewari, Abdallah Dib, Tim Weyrich, Bernd Bickel, Hans-Peter Seidel, Hanspeter Pfister, Wojciech Matusik, Louis Chevallier, Mohamed Elgharib, et al. 2021. PhotoApp: Photorealistic Appearance Editing of Head Portraits. *ACM Transactions on Graphics* 40, 4 (2021), 1–16.
- Abhimitra Meka, Christian Haene, Rohit Pandey, Michael Zollhofer, Sean Fanello, Graham Fyfe, Adarsh Kowdle, Xueming Yu, Jay Busch, Jason Dourgarian, Peter Denny, Sofien Bouaziz, Peter Lincoln, Matt Whalen, Geoff Harvey, Jonathan Taylor, Shahram Izadi, Andrea Tagliasacchi, Paul Debevec, Christian Theobalt, Julien Valentin, and Christoph Rhemann. 2019. Deep Reflectance Fields - High-Quality Facial Reflectance Field Inference From Color Gradient Illumination. *ACM Transactions on Graphics (Proceedings SIGGRAPH)* 38, 4. <https://doi.org/10.1145/3306346.3323027>
- 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 *ECCV*.
- Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. 2022. Instant Neural Graphics Primitives with a Multiresolution Hash Encoding. *arXiv:2201.05989* (January 2022).
- Ko Nishino and Shree Nayar. 2006. Corneal Imaging System: Environment from Eyes. *International Journal of Computer Vision* 70 (10 2006), 23–40. <https://doi.org/10.1007/s11263-006-6274-9>
- Rohit Pandey, Sergio Orts-Escolano, Chloe LeGendre, Christian Haene, Sofien Bouaziz, Christoph Rhemann, Paul Debevec, and Sean Fanello. 2021. Total Relighting: Learning to Relight Portraits for Background Replacement. *ACM Transactions on Graphics (Proceedings SIGGRAPH)* 40, 4. <https://doi.org/10.1145/3450626.3459872>
- Keunhong Park, Utkarsh Sinha, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Steven M. Seitz, and Ricardo Martin-Brualla. 2021a. Nerfies: Deformable Neural Radiance Fields. *ICCV* (2021).
- Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M. Seitz. 2021b. HyperNeRF: A Higher-Dimensional Representation for Topologically Varying Neural Radiance Fields. *ACM Trans. Graph.* 40, 6, Article 238 (dec 2021).
- Sudi Patel, John Marshall, and Frederick W Fitzke. 1995. Refractive index of the human corneal epithelium and stroma. , 100–141 pages.
- Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. 2021. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10318–10327.
- Gabriel Schwartz, Shih-En Wei, Te-Li Wang, Stephen Lombardi, Tomas Simon, Jason Saragih, and Yaser Sheikh. 2020. The eyes have it: An integrated eye and face model for photorealistic facial animation. *ACM Transactions on Graphics (TOG)* 39, 4 (2020), 91–1.
- Ashish Srivastava, Tomas Pfister, Oncel Tuzel, Joshua Susskind, Wenda Wang, and Russell Webb. 2017. Learning from Simulated and Unsupervised Images through Adversarial Training. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2242–2251. <https://doi.org/10.1109/CVPR.2017.241>
- Zhixin Shu, Sunil Hadap, Eli Shechtman, Kalyan Sunkavalli, Sylvain Paris, and Dimitris Samaras. 2017. Portrait Lighting Transfer Using a Mass Transport Approach. *ACM Trans. Graph.* 37, 1, Article 2 (oct 2017), 15 pages. <https://doi.org/10.1145/3095816>
- Fengyi Song, Xiaoyang Tan, Xue Liu, and Songcan Chen. 2014. Eyes closeness detection from still images with multi-scale histograms of principal oriented gradients. *Pattern Recognition* 47, 9 (2014), 2825–2838. <https://doi.org/10.1016/j.patcog.2014.03.024>
- Pratul P. Srinivasan, Boyang Deng, Xiuming Zhang, Matthew Tancik, Ben Mildenhall, and Jonathan T. Barron. 2021. NeRF: Neural Reflectance and Visibility Fields for Relighting and View Synthesis. In *CVPR*.
- Yusuke Sugano, Yasuyuki Matsushita, and Yoichi Sato. 2014. Learning-by-Synthesis for Appearance-Based 3D Gaze Estimation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*. 1821–1828. <https://doi.org/10.1109/CVPR.2014.235>
- Tiancheng Sun, Kai-En Lin, Sai Bi, Zexiang Xu, and Ravi Ramamoorthi. 2021. NeLF: Neural Light-transport Field for Portrait View Synthesis and Relighting. In *Eurographics Symposium on Rendering*.
- Tiancheng Sun, Zexiang Xu, Xiuming Zhang, Sean Fanello, Christoph Rhemann, Paul Debevec, Yun-Ta Tsai, Jonathan T Barron, and Ravi Ramamoorthi. 2020. Light stage super-resolution: continuous high-frequency relighting. *ACM Transactions on Graphics (TOG)* 39, 6 (2020), 1–12.
- J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner. 2018. HeadOn: Real-time Reenactment of Human Portrait Videos. *ACM Transactions on Graphics* 2018 (TOG) (2018).
- Justus Thies, Michael Zollhöfer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. 2016. FaceVR: Real-Time Facial Reenactment and Eye Gaze Control in Virtual Reality. *arXiv preprint arXiv:1610.03151* (2016).
- Marc Tonsen, Julian Steil, Yusuke Sugano, and Andreas Bulling. 2017. InvisibleEye: Mobile Eye Tracking Using Multiple Low-Resolution Cameras and Learning-Based Gaze Estimation. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 3, Article 106 (sep 2017), 21 pages. <https://doi.org/10.1145/3130971>
- Marc Tonsen, Xucong Zhang, Yusuke Sugano, and Andreas Bulling. 2016. Labelled Pupils in the Wild: A Dataset for Studying Pupil Detection in Unconstrained Environments. In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research Applications* (Charleston, South Carolina) (ETRA '16). Association for Computing Machinery, New York, NY, USA, 139–142. <https://doi.org/10.1145/2857491.2857520>
- Edgar Tretschk, Ayush Tewari, Vladislav Golyanik, Michael Zollhöfer, Christoph Lassner, and Christian Theobalt. 2021. Non-Rigid Neural Radiance Fields: Reconstruction and Novel View Synthesis of a Dynamic Scene From Monocular Video. In *IEEE International Conference on Computer Vision (ICCV)*. IEEE.
- Ingo Wald, Sven Woop, Carsten Benthin, Gregory S Johnson, and Manfred Ernst. 2014. Embree: a kernel framework for efficient CPU ray tracing. *ACM Transactions on Graphics (TOG)* 33, 4 (2014), 1–8.
- Kang Wang, Rui Zhao, and Qiang Ji. 2018. A Hierarchical Generative Model for Eye Image Synthesis and Eye Gaze Estimation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 440–448. <https://doi.org/10.1109/CVPR.2018.00053>
- Erroll Wood, Tadas Baltrusaitis, Xucong Zhang, Yusuke Sugano, Peter Robinson, and Andreas Bulling. 2015. Rendering of Eyes for Eye-Shape Registration and Gaze Estimation. In *Proc. of the IEEE International Conference on Computer Vision (ICCV 2015)* (2015-12-12).
- Zhengyang Wu, Srivignesh Rajendran, Tarrence Van As, Vijay Badrinarayanan, and Andrew Rabinovich. 2019. EyeNet: A Multi-Task Deep Network for Off-Axis Eye Gaze Estimation. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*. 3683–3687. <https://doi.org/10.1109/ICCVW.2019.00455>
- Zhengyang Wu, Srivignesh Rajendran, Tarrence van As, Joelle Zimmermann, Vijay Badrinarayanan, and Andrew Rabinovich. 2020. MagicEyes: A Large Scale Eye Gaze Estimation Dataset for Mixed Reality. *arXiv:2003.08806* [cs.CV]
- Wenqi Xian, Jia-Bin Huang, Johannes Kopf, and Changil Kim. 2021. Space-time Neural Irradiance Fields for Free-Viewpoint Video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 9421–9431.
- Qinjie Xiao, Hanyuan Zhang, Zhaorui Zhang, Yiqian Wu, Luyuan Wang, Xiaogang Jin, Xinwei Jiang, Yong-Liang Yang, Tianjia Shao, and Kun Zhou. 2021. EyelashNet: A Dataset and a Baseline Method for Eyelash Matting. *ACM Trans. Graph.* 40, 6, Article 217 (dec 2021), 17 pages. <https://doi.org/10.1145/3478513.3480540>
- Jason Y. Zhang, Gengshan Yang, Shubham Tulsiani, and Deva Ramanan. 2021c. NeRS: Neural Reflectance Surfaces for Sparse-view 3D Reconstruction in the Wild. In *Conference on Neural Information Processing Systems*.
- Kai Zhang, Fujun Luan, Qianqian Wang, Kavita Bala, and Noah Snavely. 2021a. PhysSG: Inverse Rendering with Spherical Gaussians for Physics-based Material Editing and Relighting. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Xiuming Zhang, Pratul P. Srinivasan, Boyang Deng, Paul Debevec, William T. Freeman, and Jonathan T. Barron. 2021b. NeRFactor: Neural Factorization of Shape and Reflectance under an Unknown Illumination. *ACM Trans. Graph.* 40, 6, Article 237 (dec 2021), 18 pages. <https://doi.org/10.1145/3478513.3480496>
- Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. 2015. Appearance-based Gaze Estimation in the Wild. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4511–4520.
- Yufeng Zheng, Seonwook Park, Xucong Zhang, Shalini De Mello, and Otmar Hilliges. 2020. Self-Learning Transformations for Improving Gaze and Head Redirection. In *Neural Information Processing Systems (NeurIPS)*.
- Hao Zhou, Sunil Hadap, Kalyan Sunkavalli, and David Jacobs. 2019. Deep Single-Image Portrait Relighting. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. 7193–7201. <https://doi.org/10.1109/ICCV.2019.00729>
- Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. 2018. Open3D: A Modern Library for 3D Data Processing. *arXiv:1801.09847* (2018).

## A SH ENVIRONMENT MAP ROTATION

Rotating the environment map can be done in the spherical harmonics domain extremely efficiently. In particular, for any given set of SH coefficients  $c$  (as a vector), we can compute  $R'$  as a block diagonal with blocks of size  $(1 \times 1, 3 \times 3, 5 \times 5, \text{etc.})$  equal to the corresponding SH order), from any given rotation matrix  $R$ . We compute  $R'$  using the method shown by [Ivanic and Ruedenberg \[1996\]](#).

For the purpose of simplicity, we assume that the spherical harmonics coefficients and functions of order  $l$  and degree  $m$  are stacked into a vector. Given the matrix  $R'$  computed from  $R$ , we have the following property

$$Y(R(\omega)) = R'(Y(\omega)) . \quad (12)$$

We note that to rotate an environment map, we need to query it at the inverse rotated locations

$$R(E)(\omega) = E(R^{-1}(\omega)) , \quad (13)$$

resulting in the following equation

$$L_o(x, \omega_o) \approx \int_{\Omega} \sum_{l=0}^{\text{order}} \sum_{m=-l}^l c_{lm}(x, \omega_o) Y_{lm}(\omega_i) E(R^{-1}(\omega_i)) d\omega_i. \quad (14)$$

Then, we rotate both  $\omega_i$ 's simultaneously by  $R$  by change of variables. In particular, rotating the integral domain does nothing, and the determinant of the Jacobian is 1 as it is a rotation.

$$L_o(x, \omega_o) \approx \int_{\Omega} \sum_{l=0}^{\text{order}} \sum_{m=-l}^l c_{lm}(x, \omega_o) Y_{lm}(R(\omega_i)) E(\omega_i) d\omega_i \quad (15)$$

We do the same reordering as in Sec. 3.3, and reinterpret the spherical harmonics functions as the vector  $\mathbf{Y}$

$$L_o(x, \omega_o) \approx \sum_{l=0}^{\text{order}} \sum_{m=-l}^l c_{lm}(x, \omega_o) \left( \int_{S\Omega} \mathbf{Y}(R(\omega_i)) E(\omega_i) d\omega_i \right)_{lm}. \quad (16)$$

We can then use our previously stated SH rotation property

$$L_o(x, \omega_o) \approx \sum_{l=0}^{\text{order}} \sum_{m=-l}^l c_{lm}(x, \omega_o) \left( \int_{S\Omega} R'(\mathbf{Y}(\omega_i)) E(\omega_i) d\omega_i \right)_{lm}. \quad (17)$$

Using the fact that  $E(\omega_i)$  is a scalar, we can then reorder the equation into the following form

$$L_o(x, \omega_o) \approx \sum_{l=0}^{\text{order}} \sum_{m=-l}^l c_{lm}(x, \omega_o) R' \left( \int_{S\Omega} \mathbf{Y}(\omega_i) E(\omega_i) d\omega_i \right)_{lm}. \quad (18)$$

We observe that this is only a slight deviation from the original formula. Now we merely need to apply a single matrix multiplication to the previously precomputed integral.

## B EYE MODEL PARAMETRIZATION

As mentioned in Section 3.1, we use a model very similar to [Schwartz et al. \[2020\]](#), with only one difference. For their eye model, they warp a sphere into their desired shape, by using a smoothstep function conditioned on the angle of the original sphere relative to the eye gaze direction. In particular, they define the linear blending parameter as:

$$\alpha = \text{smoothstep}(2\theta_{\text{diff}} + 0.5) \quad (19)$$

We replace 2 and 0.5 using learnable coefficients  $\theta_{\text{mod}}$  and  $\theta_{\text{offset}}$ , which are constrained to lie between 1 and 3, and -0.5 and 1.5 respectively, using a tanh function. For further details, we refer the reader to the appendix of [Schwartz et al. \[2020\]](#) where they describe their model in full detail.

## C OPTIMIZATION DETAILS

In order to train the warp field more efficiently, we use annealing on the positional encoding as proposed by Nerfies. We add a constant offset to the sample points inside the eyeball after the rigid transformation to ensure that the rigidly transformed eyeball interior does not overlap with the deformed periorcular region volume. Furthermore, in order to allow the network to correct for cases where the eyeball volume is misaligned with the eyeball model surface, we learn a global 6-DoF transformation on top of the transformation per frame, which does not affect the volume. This greatly improves training performance when there is some initial bias.

As with Nerfies and NeRF, we use a separate coarse and fine MLP with approximately 700k optimizable parameters each (for the full NeRF-SHL network). The deformation MLP has approximately 100k optimizable parameters, plus an 8-dimensional encoding for each frame. As explained in Section B, our eye model has 5 shape parameters. We then have 6 parameters which encode a global rigid transformation, plus 6 parameters per frame for the per-frame transformation. Finally, we learn a scalar environment map scale factor. We use the Adam optimizer to train our models, with a learning rate of 0.001 which decays exponentially by a factor of 10 every 100'000 timesteps. We use a batch size of 2816 rays, with 192 coarse and 192 fine samples each. In order to improve training performance, we learn the parameters at a different scale, by premultiplying the underlying values with a constant factor. As Adam is invariant to diagonal rescaling of gradients, this effectively corresponds to different learning rates for each parameter. We use the following empirically derived premultiplication factors:

glob. trans	glob. rot	trans	rot	v. offsets	eye model
0.0005	0.05	0.0001	0.01	0.00005	0.0005

## D CAPTURE DETAILS AND IMAGE PREPROCESSING

For each capture condition, we record between 1.5 to 3 minutes of footage at 4k resolution with 24 fps. In order to have tolerance for head motion, we use an f-stop value of 9, an ISO of 1000 and an exposure time of 1/48 seconds, and focus the camera to the subjects' right eyes in the neutral pose. We use ProRes 422 compression for improved quality. In order to temporally synchronize the static cameras and mobile camera, we turn a cell phone light on and off briefly, and sync on the frame where it turns off. We then extract only every 8-th frame to reduce redundancy and the amount of data that needs to be handled. We only aim to reproduce the eye and periorcular region. As such, we crop the original image to 800x800 centered around the eye. To do so, we use the projected eyeball center estimate as the center of the image crop. This also allows to estimate a near and far culling plane, which we define as 10 cm in front of and behind the eyeball center. We then discard the frames with bad or nonexistent AR tag detection, for example due to motion blur or poor illumination quality. Finally, to minimize the data handling, we randomly sample a set of frames from each capture condition for training. We use 100 frames from the sequence where the subject follows the mobile camera with their gaze. For the three other conditions, we sample 40 frames each from the sequences where they look at each of the four static cameras, resulting in a total of 160 frames each. This results in a total of 580 training frames for each of the 4 camera views.