MEGANE: Morphable Eyeglass and Avatar Network

Junxuan Li[1,2]*, Shunsuke Saito[2], Tomas Simon[2], Stephen Lombardi[2], Hongdong Li[1], Jason Saragih[2]
[1]Australian National University, [2]Meta Reality Labs Research

junxuan-li.github.io/megane

Eyeglasses Swapping | Directional Light Rendering | Environment Map Rendering | Lens Insertion | Few-shot Reconstruction

Figure 1. Our morphable eyeglasses supports the exchange of eyeglasses on face. Our relightable appearance correctly models glasses with different materials, and interactions between face and eyeglasses. In addition, our model enables lens insertion with appealing lens reflection and refraction effects. Once trained, our model can reconstruct and re-light an unseen eyeglasses with only a few inputs.

Abstract

Eyeglasses play an important role in the perception of identity. Authentic virtual representations of faces can benefit greatly from their inclusion. However, modeling the geometric and appearance interactions of glasses and the face of virtual representations of humans is challenging. Glasses and faces affect each other's geometry at their contact points, and also induce appearance changes due to light transport. Most existing approaches do not capture these physical interactions since they model eyeglasses and faces independently. Others attempt to resolve interactions as a 2D image synthesis problem and suffer from view and temporal inconsistencies. In this work, we propose a 3D compositional morphable model of eyeglasses that accurately incorporates high-fidelity geometric and photometric interaction effects. To support the large variation in eyeglass topology efficiently, we employ a hybrid representation that combines surface geometry and a volumetric representation. Unlike volumetric approaches, our model naturally retains correspondences across glasses, and hence explicit modification of geometry, such as lens insertion and frame deformation, is greatly simplified. In addition, our model is relightable under point lights and natural illumination, supporting high-fidelity rendering of various frame materials, including translucent plastic and metal within a single morphable model. Importantly, our approach models global light transport effects, such as casting shadows between faces and glasses. Our morphable model for eyeglasses can also be fit to novel glasses via inverse rendering. We compare our approach to state-of-the-art methods and demonstrate significant quality improvements.

1. Introduction

Humans are social animals. How we dress and accessorize is a key mode of self-expression and communication in daily life [11]. As social media and gaming has expanded social life into the online medium, virtual presentations of users have become increasingly focal to social presence, and with it, the demand for the digitization of clothes and accessories. In this paper, we focus on modeling eyeglasses, an everyday accessory for billions of people worldwide.

In particular, we argue that to achieve realism it is not sufficient to model eyeglasses in isolation: their interactions with the face have to be considered. Geometrically, glasses and faces are not rigid, and they mutually deform one another at the contact points. Thus, the shapes of eyeglasses and faces cannot be determined independently. Similarly, their appearance is coupled via global light transport, and shadows as well as inter-reflections may appear and affect the radiance. A computational approach to model these interactions is therefore necessary to achieve photorealism.

* Work done while Junxuan Li was an intern at Reality Labs Research.

Photorealistic rendering of humans has been a focus of computer graphics for over 50 years, and yet the realism of avatars created by classical authoring tools still requires extensive manual refinement to cross the uncanny valley. Modern realtime graphics engines [10] support the composition of individual components (e.g., hair, clothing), but the interaction between the face and other objects is by necessity approximated with overly simplified physically-inspired constraints or heuristics (e.g., "no interpenetrations"). Thus, they do not faithfully reconstruct all geometric and photometric interactions present in the real world.

Another group of approaches aims to synthesize the composition of glasses in the image domain [28, 66, 69] by leveraging powerful 2D generative models [25]. While these approaches can produce photorealistic images, animation results typically suffer from view and temporal inconsistencies due to the lack of 3D information.

Recently, neural rendering approaches [56] achieve photorealistic rendering of human heads [14, 17, 35, 36, 48] and general objects [40, 44, 60, 70] in a 3D consistent manner. These approaches are further extended to generative modeling for faces [6] and glasses [39, 64], such that a single morphable model can span the shape and appearance variation of each object category. However, in these approaches [6, 39, 64] interactions between objects are not considered, leading to implausible object compositions. While a recent work shows that unsupervised learning of a 3D compositional generative model from an image collection is possible [43], we observe that the lack of structural prior about faces or glasses leads to suboptimal fidelity. In addition, the aforementioned approaches are not relightable, thus not allowing us to render glasses on faces in a novel illumination.

In contrast to existing approaches, we aim at modeling the geometric and photometric interactions between eyeglasses frames and faces in a data-driven manner from image observations. To this end, we present MEGANE (**M**orphable **E**ye**g**lass and **A**vatar **Ne**twork), a morphable and relightable eyeglass model that represents the shape and appearance of eyeglasses frames and its interaction with faces. To support variations in topology and rendering efficiency, we employ a hybrid representation combining surface geometry and a volumetric representation [37]. As our hybrid representation offers explicit correspondences across glasses, we can trivially deform its structure based on head shapes. Most importantly, our model is conditioned by a high-fidelity generative human head model [6], allowing it to specialize deformation and appearance changes to the wearer. Similarly, we propose glasses-conditioned deformation and appearance networks for the morphable face model to incorporate the interaction effects caused by wearing glasses. We also propose an analytical lens model that produces photorealistic reflections and refractions for any prescription and simplifies the capture task, enabling lens insertion in a post-hoc manner.

To jointly render glasses and faces in novel illuminations, we incorporate physics-inspired neural relighting into our proposed generative modeling. The method infers output radiance given view, point-light positions, visibility, and specular reflection with multiple lobe sizes. The proposed approach significantly improves generalization and supports subsurface scattering and reflections of various materials including translucent plastic and metal within a single model. Parametric BRDF representations can not handle such diverse materials, which exhibit significant transmissive effects, and inferring their parameters for photorealistic relighting remains challenging [41, 74, 77].

To evaluate our approach, we captured 25 subjects using a multi-view light-stage capture system similar to Bi *et al*. [3]. Each subject was captured three times; once without glasses, and another two times wearing a random selection out of a set of 43 glasses. All glasses were captured without lenses. As a preprocess, we separately reconstruct glasses geometry using a differentiable neural SDF from multi-view images [60]. Our study shows that carefully designed regularization terms based on this precomputed glasses geometry significantly improves the fidelity of the proposed model. We also compare our approach with state-of-the-art generative eyeglasses models, demonstrating the efficacy of our representation as well as the proposed joint modeling of interactions. We further show that our morphable model can be fit to novel glasses via inverse rendering and relight them in new illumination conditions.

In summary, the contributions of this work are:

- the first work that tackles the joint modeling of geometric and photometric interactions of glasses and faces from dynamic multi-view image collections.
- a compositional generative model of eyeglasses that represents topology varying shape and complex appearance of eyeglasses using a hybrid mesh-volumetric representation.
- a physics-inspired neural relighting approach that supports global light transport effects of diverse materials in a single model.

2. Related Work

We discuss related work in facial avatar modeling, eyeglasses modeling, and image-based editing.

Facial Avatar Modeling. Modeling photorealistic human faces is a long standing problem in computer graphics and vision. Early works leverage multi-view capture systems to obtain high-fidelity human faces [2, 4, 5, 12, 13, 22, 49, 75]. While these approaches provide accurate facial reflectance and geometry, photorealistic rendering requires significant manual effort [50] and typically not real-time with physics-based rendering. Later, the prerequisites of facial avatar

modeling are reduced to monocular videos [7, 16, 23, 57], RGB-D inputs [58] or a single image [19, 42]. However, these approaches do not provides authentic reconstruction of avatars. Lombardi *et al*. [35] demonstrate photorealistic rendering of dynamic human faces in a data-driven manner using neural networks. The learning-based avatar modeling is later extended to volumetric representations [36], a mesh-volume hybrid representation [37], and a tetrahedron-volume hybrid representation [15]. Bi *et al*. [3] enable high-fidelity relighting of photorealistic avatars in real-time. While the aforementioned approaches require multi-view capture systems, recent works show that modeling of photorealistic avatars from monocular video inputs is also possible [1, 14, 17]. Cao *et al*. [6] recently extend these person-specific neural rendering approaches to a multi-identity model, and demonstrates the personalized adaptation of the learned universal morphable model from a mobile phone scan. Notably, these learning-based photorealistic avatars neither study nor demonstrate the accurate composition of accessories including eyeglasses.

Eyeglasses Modeling. Eyeglasses are one of the most commonly used accessories in our daily life, and virtual try-on has been extensively studied [20, 21, 30, 45, 54, 73, 76]. An image-based eyeglasses try-on is possible by composing a glasses image onto a face using Poisson blending [30]. 3D-based solutions have been also proposed for virtual reality [45] or mixed reality [73] by leveraging predefined 3D eyeglasses models. Zhang *et al*. [76] enable lens refraction and reflection in their proposed try-on system. However, these approaches rely on predefined 3D glass models, and cannot represent novel glasses. In addition, supported frames are limited to non-transparent reflective materials and the fidelity is limited by real-time graphics engines.

Recent progress in neural rendering [40, 56, 60] enables photorealistic modeling of general 3D objects. Several works extend the neural rendering techniques to generative models to represent various shapes and materials of objects in the same category using a single model [39, 64]. GeLaTO [39] presents a billboard-based neural rendering method to represent different glasses. Fig-NeRF [64] extends neural radiance fields (NeRF) [40] to generative modeling. However, these methods individually model glasses and are not conditioned by the information of the wearers. Thus, the complex geometric and photometric interactions are not incorporated in the composition. More recent approaches learn to decompose multiple 3D objects in an unsupervised manner, allowing us to compose them with different combination [43, 62, 68]. GIRAFFE [43] models the scene as composition of multiple NeRFs using adversarial training. While these approaches are promising, we observe that lack of explicit structural prior leads to suboptimal decomposition, failing to model photorealistic interactions.

Generative Models. Generative models have demonstrated remarkable ability in synthesizing photorealistic images, including human faces [25]. Recent work has extended these models to add intuitive semantic editing, such as synthesis of glasses on faces [18, 28, 33, 66, 69]. Fader Networks [28] disentangle the salient image information, and then generate different images by varying attribute values, including glasses on faces. Subsequent work has proposed two decoders for modeling latent representations and facial attributes [18], selective transfer units [33], and geometry-aware flow [72] to further improve editing fidelity. Yao *et al*. [69] extend facial attribute editing to video sequences via latent transformation and a identity preservation loss, which is further improved by Xu *et al*. [66], incorporating flow-based consistency. More recent works propose 3D-aware generative models to achieve view-consistent synthesis [8, 9, 46, 51, 59, 63, 67]. In particular, IDE-3D [52] proposes a 3D-aware semantic manipulation. However, the precise modeling and relighting of interactions between glasses and faces has been neither studied nor demonstrated.

Image-based Relighting. Various image-based solutions have been proposed to enable human face relighting [47, 53, 55, 61, 71]. Sun *et al*. [53] enables image-based relighting using an encoder-decoder network. StyleRig [55] proposes a method to invert StyleGAN [25] with explicit face prior, allowing the synthesizing pose or illumination changes for an input portrait. Wang *et al*. [61] and Total Relighting [47] infer skin reflectances such as surface normal and albedo in the image space, and use them to generate shading and reflection, which are fed into network for better generalization. Lumos [71] trains a relighting network on large-scale synthesized data and proposes several regularization terms to enable domain transfer to real portraits.

While these image-based approaches successfully synthesize photorealistic interaction and relighting of glasses and faces, lack of 3D information including contact and occlusion leads to limited fidelity and incoherent results in motion and changing views.

3. Method

Our method consists of two components, morphable geometry and relightable appearance, as shown in Fig. 2.

3.1. Morphable Geometry

Our approach is based on Mixture of Volumetric Primitives (MVP) [37], a distinct volumetric neural rendering approach that achieves high-fidelity renderings in real-time. Compared to neural fields approaches [65], it contains explicit volumetric primitives that move and deform to efficiently allow expressive animation with semantic correspondences across frames. Also unlike mesh-based approaches [35], it supports topological changes in geometry.

To model faces without glasses, we adopt the pretrained face encoder $\mathcal{E}_f$ and decoder $\mathcal{G}_f$ from Cao *et al*. [6]. Given

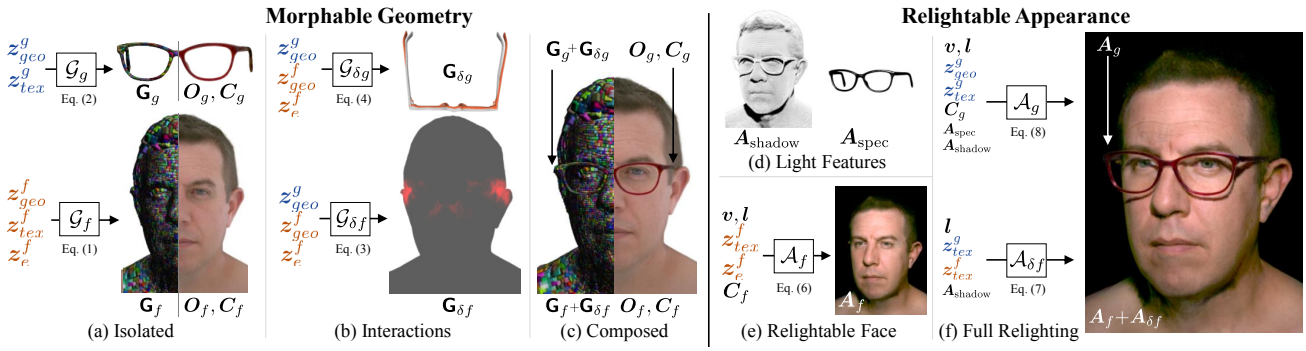

Figure 2. **Overview**. Our approach learns (a) separate latent spaces to model variations in faces and eyeglasses, as well as (b) their geometric interactions such that the models can be (c) composed together. Additionally, to accurately render relightable appearance, we compute features (d) that represent light interactions with (e) a relightable face model to allow for (f) joint face and eyeglass relighting.

an encoding of the facial expression $\boldsymbol{z}_e^f$, and face identity encoding of geometry $\boldsymbol{z}_{geo}^f$ and textures $\boldsymbol{z}_{tex}^f$, the face primitive geometry and appearance are decoded as:

$$\mathbf{G}_f, \boldsymbol{O}_f, \boldsymbol{C}_f = \mathcal{G}_f(\boldsymbol{z}_e^f, \boldsymbol{z}_{geo}^f, \boldsymbol{z}_{tex}^f), \tag{1}$$

where $\mathbf{G}_f = \{\boldsymbol{t}, \boldsymbol{R}, \boldsymbol{s}\}$ is the tuple of the position $\boldsymbol{t} \in \mathbb{R}^{3\times N_{\text{fprim}}}$, rotation $\boldsymbol{R} \in \mathbb{R}^{3\times 3\times N_{\text{fprim}}}$ and scale $\boldsymbol{s} \in \mathbb{R}^{3\times N_{\text{fprim}}}$ of face primitives; $\boldsymbol{O}_f \in \mathbb{R}^{M^3\times N_{\text{fprim}}}$ is the opacity of face primitives; $\boldsymbol{C}_f \in \mathbb{R}^{3\times M^3\times N_{\text{fprim}}}$ is the RGB color of face primitives in fully-lit images. N_{fprim} denotes the number of face primitives and M denote the resolution of each primitives. We follow previous work [6] and use $N_{\text{fprim}} = 128 \times 128$ and $M = 8$.

To model glasses, we propose a generative morphable eyeglass network that consists of a variational auto-encoder architecture: $\boldsymbol{z}_{geo}^g, \boldsymbol{z}_{tex}^g = \mathcal{E}_g(\boldsymbol{w}_{\text{id}}^g)$, where $\mathcal{E}_g$ is a glasses encoder that takes a one-hot-vector $\boldsymbol{w}_{\text{id}}^g$ of glasses at input, and generates both geometry and appearance latent codes for the glasses $\boldsymbol{z}_{geo}^g, \boldsymbol{z}_{tex}^g$ as output. We then use the latent codes for a morphable glasses geometry decoder:

$$\mathbf{G}_g, \boldsymbol{O}_g, \boldsymbol{C}_g = \mathcal{G}_g(\boldsymbol{z}_{geo}^g, \boldsymbol{z}_{tex}^g), \tag{2}$$

where $\mathbf{G}_g = \{\boldsymbol{t}_g, \boldsymbol{R}_g, \boldsymbol{s}_g\}$ is the tuple of the position, rotation and scale of the eyeglasses primitives, with position $\boldsymbol{t}_g \in \mathbb{R}^{3\times N_{\text{gprim}}}$, rotation $\boldsymbol{R}_g \in \mathbb{R}^{3\times 3\times N_{\text{gprim}}}$ and scale $\boldsymbol{s}_g \in \mathbb{R}^{3\times N_{\text{gprim}}}$; $\boldsymbol{O}_g \in \mathbb{R}^{M^3\times N_{\text{gprim}}}$ the opacity of glasses primitives; $\boldsymbol{C}_{\text{g}} \in \mathbb{R}^{3\times M^3\times N_{\text{gprim}}}$ is the RGB color of glasses primitives in fully-lit images. N_{gprim} denotes the number of glasses primitives; we use $N_{\text{gprim}} = 32 \times 32$.

We model the deformation caused by the interaction as residual deformation of the primitives:

$$\mathbf{G}_{\delta f} = \mathcal{G}_{\delta f}(\boldsymbol{z}_e^f, \boldsymbol{z}_{geo}^g, \boldsymbol{z}_{geo}^f), \tag{3}$$

$$\mathbf{G}_{\delta g} = \mathcal{G}_{\delta g}(\boldsymbol{z}_e^f, \boldsymbol{z}_{geo}^g, \boldsymbol{z}_{geo}^f), \tag{4}$$

where $\mathbf{G}_{\delta f} = \{\delta_{\boldsymbol{t}}, \delta_{\boldsymbol{R}}, \delta_{\boldsymbol{s}}\}, \mathbf{G}_{\delta g} = \{\delta_{\boldsymbol{t}_g}, \delta_{\boldsymbol{R}_g}, \delta_{\boldsymbol{s}_g}\}$ are the residuals in position, rotation and scale from their values in the canonical space. Specifically, the interaction influences the eyeglasses in two different ways: non-rigid deformations caused by fitting to the head, and rigid deformations caused by facial expressions. We found that individually modeling these two effects better generalize to a novel combination of glasses and an identity. Therefore, we model the deformation residuals as

$$\mathcal{G}_{\delta g}(\cdot) = \mathcal{G}_{\text{deform}}(\boldsymbol{z}_{geo}^g, \boldsymbol{z}_{geo}^f) + \mathcal{G}_{\text{transf}}(\boldsymbol{z}_e^f, \boldsymbol{z}_{geo}^g) \tag{5}$$

where $\mathcal{G}_{\text{deform}}$ takes facial identity information to deform the eyeglasses to the target head, and $\mathcal{G}_{\text{transf}}$ takes expression encoding as input to model the relative rigid motion of eyeglasses on face caused by different facial expressions (e.g., sliding up when wrinkling the nose).

3.2. Relightable Appearance

The appearance model in previous works based on volumetric primitives [6, 37] integrates the captured lighting environment as part of appearance, and cannot relight the avatar to novel illuminations. The appearance values of primitives under the uniform tracking illumination in Sec. 3.1, $\boldsymbol{C}_f$ and $\boldsymbol{C}_g$ are only used for learning geometry and the deformation by interactions. To enable relighting of the generative face model, we train a relightable appearance decoder that is additionally conditioned on view direction $\boldsymbol{v}$ and light direction $\boldsymbol{l}$ following [3]:

$$\boldsymbol{A}_f = \mathcal{A}_f(\boldsymbol{z}_e^f, \boldsymbol{v}, \boldsymbol{l}, \boldsymbol{z}_{tex}^f, \boldsymbol{C}_f), \tag{6}$$

where $\boldsymbol{A}_f \in \mathbb{R}^{3\times M^3\times N_{\text{fprim}}}$ is the appearance slab consists of RGB colors under a single point-light.

To model the photometric interaction of eyeglasses on faces, we consider it as residuals conditioned by a eyeglasses latent code, similarly to the deformation residuals. Additionally, we observed that the most noticeable appearance interactions of eyeglasses on the face are from cast shadows. We explicitly provide shadow feature as an input to facilitate shadow modeling:

$$\boldsymbol{A}_{\delta f} = \mathcal{A}_{\delta f}(\boldsymbol{l}, \boldsymbol{z}_{tex}^g, \boldsymbol{z}_{tex}^f, \boldsymbol{A}_{\text{shadow}}) \tag{7}$$

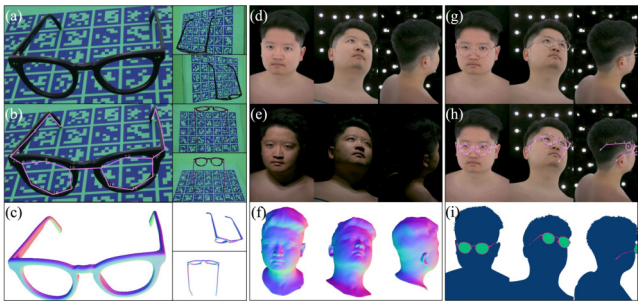

Figure 3. Datasets for *Eyeglasses* (a-c), *Faces* (d-f), and *Faces with Eyeglasses* (g-i). See text for description.

where $\boldsymbol{A}_{\delta f} \in \mathbb{R}^{3\times M^3\times N_{\text{fprim}}}$ is the appearance residual for the face; and $\boldsymbol{A}_{\text{shadow}} \in \mathbb{R}^{M^3\times N_{\text{fprim}}}$ is the shadow feature computed by accumulating opacity while ray-marching from each of the light sources to the primitives, representing light visibility [34]. Thus, the shadow feature represents the first bounce of light transport on both the face and glasses.

We model the relightalble glasses appearance similarly to the relightable face. Since this work focuses on modeling eyeglasses on faces, we define it as a conditional model with face so that occlusion and multiple bounces of lights by an avatar's head is already incorporated in the appearance:

$$\boldsymbol{A}_g = \mathcal{A}_g(\boldsymbol{v}, \boldsymbol{l}, \boldsymbol{z}^g_{tex}, \boldsymbol{z}^g_{geo}, \boldsymbol{A}_{\text{shadow}}, \boldsymbol{A}_{\text{spec}}, \boldsymbol{C}_g). \quad (8)$$

where $\boldsymbol{A}_g \in \mathbb{R}^{3\times M^3\times N_{\text{gprim}}}$ is the glasses appearance slab, and $\boldsymbol{A}_{\text{spec}} \in \mathbb{R}^{3\times M^3\times N_{\text{gprim}}}$ is the specular feature; $\boldsymbol{A}_{\text{shadow}}$ is the shadow feature computed in the same way as in Eq. (7), which encodes face information. We compute specular feature $\boldsymbol{A}_{\text{spec}}$ at every point on primitives based on normal, light and view directions with a specular BRDF parameterized as Spherical Gaussians [29] with three different lobes. We observe that explicitly conditioning specular reflection significantly improves fidelity of relighting and generalization to various frame materials. Similar observations have been made for recent portrait relighting approaches [47,71].

3.3. Differentiable Volumetric Rendering

We render the predicted volumetric primitives following previous work [37]. Denote the position of all primitives in the space as $\mathbf{G}$, when only render the face without wearing any eyeglasses, $\mathbf{G} = \mathbf{G}_f$; and when wearing glasses $\mathbf{G} = \{\mathbf{G}_f + \mathbf{G}_{\delta f}, \mathbf{G}_g + \mathbf{G}_{\delta g}\}$. Denote the opacity of all primitives as $\boldsymbol{O}$, it takes form $\boldsymbol{O} = \boldsymbol{O}_f$ or $\boldsymbol{O} = \{\boldsymbol{O}_f, \boldsymbol{O}_g\}$ for without and with glasses. Denote the color of all primitives as $\boldsymbol{C}$, $\boldsymbol{C} = \boldsymbol{C}_f$ and $\boldsymbol{C} = \{\boldsymbol{C}_f, \boldsymbol{C}_g\}$ in fully-lit images, while $\boldsymbol{C} = \boldsymbol{A}_f$ and $\boldsymbol{C} = \{\boldsymbol{A}_f + \boldsymbol{A}_{\delta f}, \boldsymbol{A}_g\}$ in relighting frames. We then use volumetric aggregation [37] to render images.

3.4. Data Acquisition

We aim to learn a generative model of eyeglasses and faces as well as the interactions between them. Therefore, we capture three types of data: *Eyeglasses*, *Faces*, and *Faces with Eyeglasses*. To decouple learning frame style from lens effects (which vary across prescriptions), we remove the lenses from the eyeglasses for all datasets.

Eyeglasses. We selected a set of 43 eyeglasses to cover a wide range of sizes, styles, and materials, including metal and translucent plastics of various colors. For each eyeglasses instance, we capture approximately 70 multi-view images using a hand-held DSLR camera (Fig. 3(a)). We apply a surface reconstruction method [60] to extract 3D meshes of the eyeglasses (Fig. 3(c)). These 3D meshes will later provide supervision for the eyeglasses MVP geometry. However, because the glasses will change geometrically once they are worn, we use Bounded Biharmonic Weights (BBW) [24] to define a coarse deformation model that will be used to fit these meshes to the *Face With Eyeglasses* dataset using keypoint detections (Fig. 3(b)). Please see our supplementary material for details of eyeglasses mesh reconstruction and registration.

Faces and **Faces with Eyeglasses**. We capture a dataset of faces without eyeglasses and the same set of faces with eyeglasses. This dataset consists of 25 subjects captured using a multi-view light-stage capture system with 110 cameras. Participants are instructed to perform various facial expressions, yielding recordings with changing expressions and head pose ((Fig. 3(d)). Each subject was captured three times: once without glasses, and another two times wearing a random selection out of the set of 43 glasses (Fig. 3(g)).

To allow for relighting, this data is captured under different illumination conditions. Similar to Bi *et al*. [3], the capture system uses time-multiplexed illuminations. In particular, fully-lit frames, *i.e*. frames for which all lights on the lightstage are turned on, are interleaved every third frame to allow for tracking, and the remaining two thirds of the frames are used to observe the subject under changing lighting conditions where only a subset of lights ("group" lights) are turned on (Fig. 3(e)).

Similar to prior work [6,37], we first pre-process the data using a multiview face tracker to generate a coarse but topologically consistent face mesh for each frame (Fig. 3(f)). Tracking and detections are performed on fully lit frames and interpolated to partially lit frames when necessary. Additionally, for the *Faces with Eyeglasses* portion, we detect a set of 20 keypoints on the eyeglasses [31] (Fig. 3(h)) as well as face and glasses segmentation masks [27] (Fig. 3(i)), which are used to fit the eyeglasses BBW mesh deformation model to match the observed glasses.

3.5. Training and Losses

We train the networks in two stages. In the first stage we use the fully-lit images to train the geometry of faces and glasses. Then, we use the images under group lights to train the relightable appearance model.

Morphable Geometry Training We denote the parameters of the expression encoder in $\mathcal{E}_f$, glasses encoder $\mathcal{E}_g$, and decoders $\mathcal{G}_f, \mathcal{G}_g, \mathcal{G}_{\delta f}, \mathcal{G}_{\delta g}$ as Φ_g, and optimize them using:

$$\Phi_g' = \arg\min_{\Phi_g} \sum_{N_I} \sum_{N_{F_i}} \sum_{N_C} \mathcal{L}_{\text{fully-lit}}(\Phi_g, \boldsymbol{I}^{i,r}), \quad (9)$$

over N_I different subjects; N_{F_i} different fully-lit frames including with and without glasses; and N_C different camera view points; and $\boldsymbol{I}^i$ denotes all the ground truth camera images and associated processed assets for a frame, including face geometry, glasses geometry, face segmentation, and glasses segmentation; likewise, $\boldsymbol{I}^r$ denotes the reconstructed images from volumetric rendering and the corresponding assets. Our fully-lit loss function consists of three main components:

$$\mathcal{L}_{\text{fully-lit}}(\cdot) = \mathcal{L}_{\text{rec}}(\boldsymbol{I}^{i,r}) + \mathcal{L}_{\text{gls}}(\boldsymbol{I}^{i,r}) + \mathcal{L}_{\text{reg}}(\Phi_g, \boldsymbol{I}^{i,r}), \quad (10)$$

where the $\mathcal{L}_{\text{rec}}$ are photometric reconstruction losses:

$$\mathcal{L}_{\text{rec}}(\cdot) = \mathcal{L}_{\text{L1}}(\boldsymbol{I}^{i,r}) + \mathcal{L}_{\text{vgg}}(\boldsymbol{I}^{i,r}) + \mathcal{L}_{\text{gan}}(\boldsymbol{I}^{i,r}), \quad (11)$$

where $\mathcal{L}_{\text{L1}}$ is the l_1 loss between observed images and reconstruction; $\mathcal{L}_{\text{vgg}}, \mathcal{L}_{\text{gan}}$ are the VGG and GAN loss in [6].

We also propose a geometry guidance loss $\mathcal{L}_{\text{gls}}$ using the separately reconstructed glasses (Sec. 3.4) to improve the geometric accuracy of glasses, leading to better separations of faces and glasses in the joint training:

$$\mathcal{L}_{\text{gls}}(\cdot) = \mathcal{L}_{\text{c}}(\boldsymbol{I}^{i,r}) + \mathcal{L}_{\text{m}}(\boldsymbol{I}^{i,r}) + \mathcal{L}_{\text{s}}(\boldsymbol{I}^{i,r}) \quad (12)$$

including chamfer distance loss $\mathcal{L}_{\text{c}}$; glasses masking loss $\mathcal{L}_{\text{m}}$; and glasses segmentation loss $\mathcal{L}_{\text{k}}$. These losses encourage the network to separate identity-dependent deformations from glasses intrinsic deformations, thus helping the networks to generalize on different identities. Please refer supplementary for details.

In addition, we propose a regularization loss $\mathcal{L}_{\text{reg}}$ for training: we use $\mathcal{L}_{\text{KL}}(\cdot)$ the KL-divergence loss between the prior Gaussian distribution and the distribution of the glasses latent space; we also use a l_2-norm for suppressing the delta deformation of face to reduce large displacements of face primitives.

$$\mathcal{L}_{\text{reg}}(\cdot) = \mathcal{L}_{\text{KL}}(\Phi_g) + \mathcal{L}_{\text{L2}}(\Phi_g, \boldsymbol{I}^{i,r}). \quad (13)$$

During training, we set the weights of each loss term as $\lambda_{\text{L1}} = 1, \lambda_{\text{vgg}} = 1, \lambda_{\text{gan}} = 1, \lambda_{\text{c}} = 0.01, \lambda_{\text{m}} = 10, \lambda_{\text{s}} = 10, \lambda_{\text{KL}} = 10^{-4}, \lambda_{\text{L2}} = 10^{-3}$. We train the first stage on a Nvidia Tesla V100 GPU with a batch size of 4 for 300k iterations using Adam optimizer [26] with a learning rate of 10^{-3}, which takes around four days.

Relightable Appearance Training Once the geometry module is trained, we freeze the parameters Φ_g and start training the relightable appearance $\mathcal{A}_f$, $\mathcal{A}_{\delta f}$, and $\mathcal{A}_g$. We denote their parameters as Φ_a, We optimize the parameters Φ_a as follows:

$$\Phi_a' = \arg\min_{\Phi_a} \sum_{N_I} \sum_{N_{G_i}} \sum_{N_C} \mathcal{L}_{\text{group-lit}}(\Phi_a, \boldsymbol{I}^{i,c}), \quad (14)$$

over N_I different subjects; N_C different cameras; and N_{G_i} different group-light frames including with and without wearing glasses on face.

For frames illuminated by group-lights, we take the two nearest fully-lit frames to generate face and glasses geometry using $\mathcal{G}_f, \mathcal{G}_g, \mathcal{G}_{\delta f}, \mathcal{G}_{\delta g}$, and linearly interpolate to get face and glasses geometry for the group-light image.

The objective function for the second stage is mean-square-error photometric loss $\mathcal{L}_{\text{group-lit}}(\cdot) = ||I^i - I^c||_2^2$. The VGG and GAN loss are not used in relightable appearance training since we observe that these loss introduced block-like artifacts in the reconstruction. We use the same optimizer and GPU as in the previous stage. We train the second stage with a batch size of 3 for 200k iterations, which takes around four days.

Components	$l_1(\downarrow)$	PSNR(↑)	SSIM(↑)	LPIPS(↓)
w/o Geo	2.374	32.55	0.8764	0.1712
w/o $\boldsymbol{A}_{\text{shadow}}$	1.870	36.63	0.9227	0.1171
w/o $\boldsymbol{A}_{\text{spec}}$	1.577	37.69	0.9377	0.1087
Full method	**1.558**	**37.98**	**0.9388**	**0.1034**

Table 1. Quantitative ablation of each part of our model.

4. Experiments

In this section, we evaluate each component of our method using the dataset of *Faces with Eyeglasses* and compare extensively with SOTA approaches. We exclude a set of frames and cameras for evaluation.

4.1. Ablation Study

Geometry Guidance. We first show that the proposed geometry-guided losses, including surface normal and segmentation, is essential for achieving crisp and sharp eyeglasses reconstruction. As shown in Fig. 4 and Tab. 1, the model without using geometry guidance is only trained with image-based reconstruction and regularization losses. And it fails to reconstruct the detailed geometry of the eyeglasses, such as the nose pads. In comparison, the model with geometry guidance achieves higher geometric fidelity.
Geometry Interaction. Eyeglasses and faces deform each other at contact points. We show in Fig. 5 that without modeling such deformations, the legs of eyeglasses are rendered incorrectly and penetrate into the head. With the modeling of geometric interactions, our method learns and faithfully represents the deformation of the head as well as the nose.

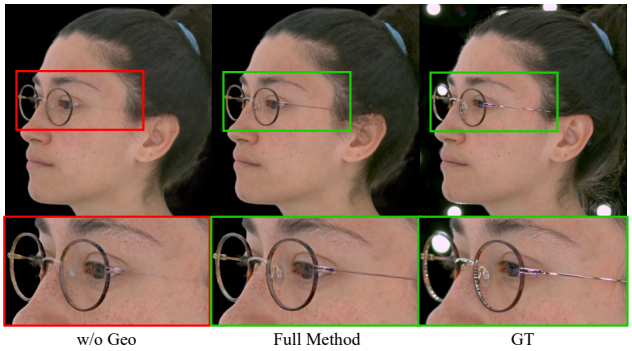

Figure 4. **Ablation study on geometry guidance.** Without geometry guidance lead to blurry results while our full model generates sharp and accurate eyeglasses.

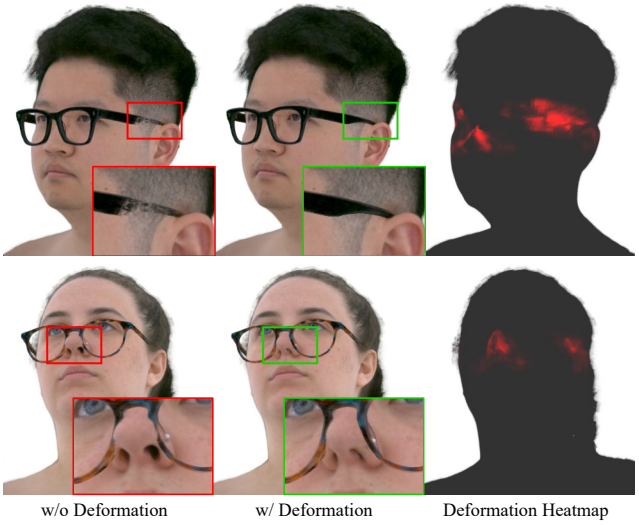

Figure 5. **Effectiveness of deformation.** Face deformation modeling is critical for correctly rendering eyeglasses and face.

Physics-inspired features for neural relighting. Here, we evaluate the effectiveness of the proposed specular and shadow feature on neural relighting. As shown in Fig. 6, the one without using specular features fails to reconstruct specular highlights on the frame. Furthermore, the model without appearance interaction fails to reconstruct correct shadows on the face. We test and evaluate these components on held-out frames and present the quantitative results on Table 1. Adding each component effectively improves the performance on all metrics.

4.2. Comparison

GeLaTO [39]. Previous work [39, 64] enables generative modeling of eyeglasses, but assume that everything except the glasses are static in the scene. In particular, FigNeRF [64] is not applicable to our setup with severe occlusions and head motion. For comparison, we reimplement GeLaTO [39] and train with our datasets. Since GeLaTO

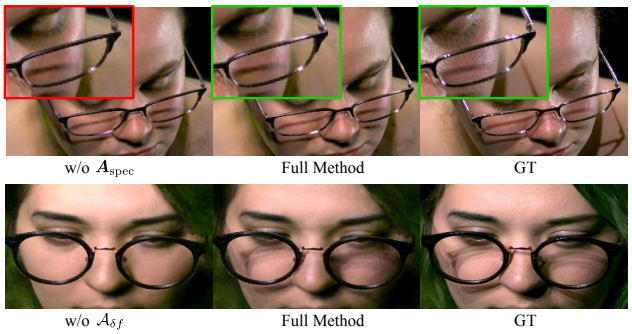

Figure 6. **Ablation study on specular feature and appearance interaction.** Top row: w/o using specular feature and full model. Bottom row: w/o appearance interaction and full model.

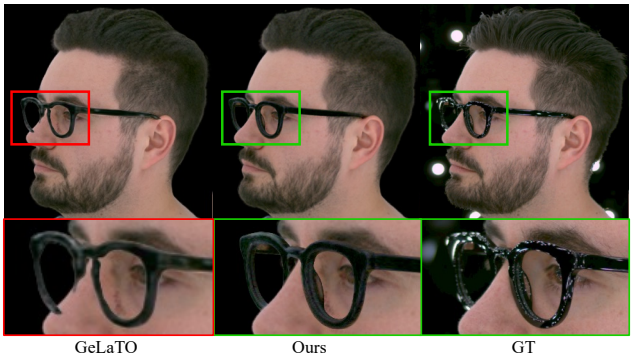

Figure 7. **Comparison with GeLaTO** [39]. Due to the simplified geometry representation, GeLaTO lacks geometry details and suffers from inconsistent occlusions.

Methods	$l_1(\downarrow)$	PSNR(↑)	SSIM(↑)	LPIPS(↓)
GeLaTO [39]	16.561	18.91	0.6479	0.2576
Ours	**9.202**	**21.80**	**0.7690**	**0.1614**

Table 2. Quantitative comparison with GeLaTO.

does not support relighting, we compare only on fully-lit frames. Fig. 7 shows that while GeLaTO lacks geometric details and generates incorrect occlusion boundaries due to the billboard-based geometry, our method achieves high-fidelity results and correctly handles occlusions. Tab. 2 shows that our method also outperforms in all metrics.

GIRAFFE [43] proposed a compositional neural radiance field that supports adding and changing objects in a scene. However, the official implementation only supports objects within the same category. For a fair comparison, we adapt their method to support adding generative objects in multiple categories. Fig. 8 shows that compositional generative modeling in an unsupervised manner still leads to suboptimal fidelity with limited resolution.

VideoEditGAN [66] is a SOTA image-based editing method that allows us to insert glasses on face images. As

Figure 8. **Comparison with GIRAFFE [43] and VideoEditGAN [66].** Compared with our method, other methods fail to render view consistent results.

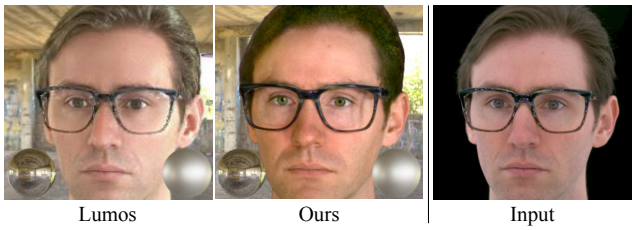

Figure 9. **Comparison with Lumos** [71]. Due to the 3D-aware lighting features, our method yields realistic shadows on the face.

shown in Fig. 8, the image-based approach fails to maintain color and view consistency. Moreover, the approach cannot choose a specific type of glasses. In contrast, our proposed representation enables the accurate reproduction of glasses and faces with consistent rendering in both view and time.

Relighting Comparison with Lumos [71]. All the methods mentioned above do not support relighting of faces and eyeglasses. We compare our relighting results with Lumos [71], a SOTA approach for portrait relighting. Due to the lack of 3D information, Lumos has difficulty rendering non-local light transport effects such as shadows cast by eyeglasses. In contrast, our method generates plausible soft shadows and accurately models photometric interactions between faces and glasses.

4.3. Applications

Generative Eyeglasses. Our model is able to generate new eyeglasses via latent code modification (see supplementary video for more results). Fig. 10 shows that our method supports replacing relightable materials while retaining shapes.

Few-Shot Reconstruction. Our generative glasses model supports differentiable rendering, enabling few-shot reconstruction from a few-view images via inverse rendering. Notably, our non-relightable and relightable appearance

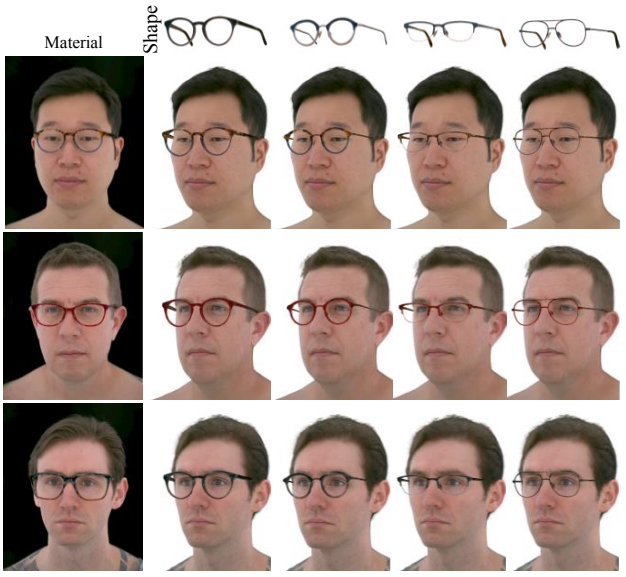

Figure 10. **Material swapping.** Our generative model supports changing materials and shape.

models share the same latent codes. Thus, as shown in Fig. 1, the few-shot reconstruction using only fully-lit illumination can be rendered from novel illuminations.

Lens Insertion. Since our model retains correspondences between primitives, inserting a lens in generated glasses is trivial by selecting control points for the lens contour on a single template. We further incorporate physically-accurate refraction and reflection based on prescription as shown in Fig. 1. Please refer to supplemental for details.

5. Conclusions

We introduced MEGANE, a 3D morphable and relightable model of eyeglasses to create photorealistic compositions of eyeglasses on volumetric head avatars from any view point under novel illuminations. Our experiments show that reproducing geometric and photometric interactions in the real world is now possible by leveraging neural rendering with a hybrid mesh-volumetric generative model. By explicitly controlling the motion of primitives, our approach achieves, for the first time, the learning-based modeling of geometric interactions between glasses and faces. We also examined the effectiveness of physics-inspired lighting features as inputs for neural relighting, and demonstrate that our approach enables relighting with a diverse set of materials that are both transmissive and reflective using a single generative model. Lastly, we show that our generative model allows few-shot fitting to novel glasses, allowing relighting without additional OLAT data.

Future work includes few-shot fitting to in-the-wild images by adopting a test-time finetuning as in [6], or physically accurate fitting of lenses via inverse rendering [32,38].

References

[1] ShahRukh Athar, Zhixin Shu, and Dimitris Samaras. Flame-in-nerf: Neural control of radiance fields for free view face animation. *arXiv preprint arXiv:2108.04913*, 2021. 3

[2] Thabo Beeler, Fabian Hahn, Derek Bradley, Bernd Bickel, Paul Beardsley, Craig Gotsman, Robert W. Sumner, and Markus Gross. High-quality passive facial performance capture using anchor frames. *ACM Trans. Graph.*, 30(4):75:1–75:10, 2011. 2

[3] Sai Bi, Stephen Lombardi, Shunsuke Saito, Tomas Simon, Shih-En Wei, Kevyn Mcphail, Ravi Ramamoorthi, Yaser Sheikh, and Jason Saragih. Deep relightable appearance models for animatable faces. *ACM Transactions on Graphics (TOG)*, 40(4):1–15, 2021. 2, 3, 4, 5

[4] Bernd Bickel, Mario Botsch, Roland Angst, Wojciech Matusik, Miguel Otaduy, Hanspeter Pfister, and Markus Gross. Multi-scale capture of facial geometry and motion. *ACM Trans. Graph.*, 26(3):33:1–33:10, 2007. 2

[5] Derek Bradley, Wolfgang Heidrich, Tiberiu Popa, and Alla Sheffer. High resolution passive facial performance capture. *ACM Trans. Graph.*, 29(4):41:1–41:10, 2010. 2

[6] Chen Cao, Tomas Simon, Jin Kyu Kim, Gabe Schwartz, Michael Zollhoefer, Shun-Suke Saito, Stephen Lombardi, Shih-En Wei, Danielle Belko, Shoou-I Yu, et al. Authentic volumetric avatars from a phone scan. *ACM Transactions on Graphics (TOG)*, 41(4):1–19, 2022. 2, 3, 4, 5, 6, 8

[7] Chen Cao, Hongzhi Wu, Yanlin Weng, Tianjia Shao, and Kun Zhou. Real-time facial animation with image-based dynamic avatars. *ACM Transactions on Graphics*, 35(4), 2016. 3

[8] Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J. Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16123–16133, June 2022. 3

[9] Yu Deng, Jiaolong Yang, Jianfeng Xiang, and Xin Tong. Gram: Generative radiance manifolds for 3d-aware image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10673–10683, June 2022. 3

[10] Unreal Engine. Metahuman creator. https://www.unrealengine.com/en-US/metahuman. 2

[11] John Carl Flugel. The psychology of clothes. *The Sociological Review*, 25(3):301–304, 1933. 1

[12] Yasutaka Furukawa and Jean Ponce. Dense 3D motion capture for human faces. In *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '09, pages 1674–1681. IEEE Computer Society, 2009. 2

[13] Graham Fyffe, Andrew Jones, Oleg Alexander, Ryosuke Ichikari, and Paul Debevec. Driving high-resolution facial scans with video performance capture. *ACM Trans. Graph.*, 34(1):8:1–8:14, 2014. 2

[14] Guy Gafni, Justus Thies, Michael Zollhofer, and Matthias Nießner. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8649–8658, 2021. 2, 3

[15] Stephan J Garbin, Marek Kowalski, Virginia Estellers, Stanislaw Szymanowicz, Shideh Rezaeifar, Jingjing Shen, Matthew Johnson, and Julien Valentin. Voltemorph: Real-time, controllable and generalisable animation of volumetric representations. *arXiv preprint arXiv:2208.00949*, 2022. 3

[16] Pablo Garrido, Levi Valgaerts, Chenglei Wu, and Christian Theobalt. Reconstructing detailed dynamic face geometry from monocular video. *ACM Trans. Graph.*, 32(6):158–1, 2013. 3

[17] Philip-William Grassal, Malte Prinzler, Titus Leistner, Carsten Rother, Matthias Nießner, and Justus Thies. Neural head avatars from monocular rgb videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18653–18664, 2022. 2, 3

[18] Zhenliang He, Wangmeng Zuo, Meina Kan, Shiguang Shan, and Xilin Chen. Attgan: Facial attribute editing by only changing what you want. *IEEE transactions on image processing*, 28(11):5464–5478, 2019. 3

[19] Liwen Hu, Shunsuke Saito, Lingyu Wei, Koki Nagano, Jaewoo Seo, Jens Fursund, Iman Sadeghi, Carrie Sun, Yen-Chun Chen, and Hao Li. Avatar digitization from a single image for real-time rendering. *ACM Transactions on Graphics (ToG)*, 36(6):1–14, 2017. 3

[20] Szu-Hao Huang, Yu-I Yang, and Chih-Hsing Chu. Human-centric design personalization of 3d glasses frame in markerless augmented reality. *Advanced Engineering Informatics*, 26(1):35–45, 2012. 3

[21] Wan-Yu Huang, Chaur-Heh Hsieh, and Jeng-Sheng Yeh. Vision-based virtual eyeglasses fitting system. In *2013 IEEE International Symposium on Consumer Electronics (ISCE)*, pages 45–46. IEEE, 2013. 3

[22] Xiaolei Huang, Song Zhang, Yang Wang, Dimitris N. Metaxas, and Dimitris Samaras. A hierarchical framework for high resolution facial expression tracking. In *Proceedings of the 2004 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, CVPR Workshops '04, page 22, Washington, DC, USA, 2004. IEEE Computer Society. 2

[23] Alexandru Eugen Ichim, Sofien Bouaziz, and Mark Pauly. Dynamic 3d avatar creation from hand-held video input. *ACM Transactions on Graphics (ToG)*, 34(4):1–14, 2015. 3

[24] Alec Jacobson, Ilya Baran, Jovan Popovic, and Olga Sorkine. Bounded biharmonic weights for real-time deformation. *ACM Trans. Graph.*, 30(4):78, 2011. 5

[25] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 2, 3

[26] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[27] Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross Girshick. Pointrend: Image segmentation as rendering. In *Pro-*

ceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 9799–9808, 2020. 5

[28] Guillaume Lample, Neil Zeghidour, Nicolas Usunier, Antoine Bordes, Ludovic Denoyer, and Marc'Aurelio Ranzato. Fader networks: Manipulating images by sliding attributes. *Advances in neural information processing systems*, 30, 2017. 2, 3

[29] Junxuan Li and Hongdong Li. Self-calibrating photometric stereo by neural inverse rendering. In *European Conference on Computer Vision*, pages 166–183. Springer, 2022. 5

[30] Juan Li and Jie Yang. Eyeglasses try-on based on improved poisson equations. In *2011 International Conference on Multimedia Technology*, pages 3058–3061. IEEE, 2011. 3

[31] Wenbo Li, Zhicheng Wang, Binyi Yin, Qixiang Peng, Yuming Du, Tianzi Xiao, Gang Yu, Hongtao Lu, Yichen Wei, and Jian Sun. Rethinking on multi-stage networks for human pose estimation. *arXiv preprint arXiv:1901.00148*, 2019. 5

[32] Zhengqin Li, Yu-Ying Yeh, and Manmohan Chandraker. Through the looking glass: Neural 3d reconstruction of transparent shapes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1262–1271, 2020. 8

[33] Ming Liu, Yukang Ding, Min Xia, Xiao Liu, Errui Ding, Wangmeng Zuo, and Shilei Wen. Stgan: A unified selective transfer network for arbitrary image attribute editing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3673–3682, 2019. 3

[34] Tom Lokovic and Eric Veach. Deep shadow maps. 2000. 5

[35] Stephen Lombardi, Jason Saragih, Tomas Simon, and Yaser Sheikh. Deep appearance models for face rendering. *ACM Transactions on Graphics (ToG)*, 37(4):1–13, 2018. 2, 3

[36] Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. Neural volumes: learning dynamic renderable volumes from images. *ACM Transactions on Graphics (TOG)*, 38(4):1–14, 2019. 2, 3

[37] Stephen Lombardi, Tomas Simon, Gabriel Schwartz, Michael Zollhoefer, Yaser Sheikh, and Jason Saragih. Mixture of volumetric primitives for efficient neural rendering. *ACM Transactions on Graphics (TOG)*, 40(4):1–13, 2021. 2, 3, 4, 5

[38] Jiahui Lyu, Bojian Wu, Dani Lischinski, Daniel Cohen-Or, and Hui Huang. Differentiable refraction-tracing for mesh reconstruction of transparent objects. *ACM Transactions on Graphics (TOG)*, 39(6):1–13, 2020. 8

[39] Ricardo Martin-Brualla, Rohit Pandey, Sofien Bouaziz, Matthew Brown, and Dan B Goldman. Gelato: Generative latent textured objects. In *European Conference on Computer Vision*, pages 242–258. Springer, 2020. 2, 3, 7

[40] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision*, pages 405–421. Springer, 2020. 2, 3

[41] Jacob Munkberg, Jon Hasselgren, Tianchang Shen, Jun Gao, Wenzheng Chen, Alex Evans, Thomas Müller, and Sanja Fidler. Extracting triangular 3d models, materials, and lighting from images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8280–8290, 2022. 2

[42] Koki Nagano, Jaewoo Seo, Jun Xing, Lingyu Wei, Zimo Li, Shunsuke Saito, Aviral Agarwal, Jens Fursund, Hao Li, Richard Roberts, et al. pagan: real-time avatars using dynamic textures. *ACM Trans. Graph.*, 37(6):258–1, 2018. 3

[43] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11453–11464, 2021. 2, 3, 7, 8

[44] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3504–3515, 2020. 2

[45] Arthur Niswar, Ishtiaq Rasool Khan, and Farzam Farbiz. Virtual try-on of eyeglasses using 3d model of the head. In *Proceedings of the 10th International Conference on Virtual Reality Continuum and Its Applications in Industry*, pages 435–438, 2011. 3

[46] Roy Or-El, Xuan Luo, Mengyi Shan, Eli Shechtman, Jeong Joon Park, and Ira Kemelmacher-Shlizerman. Stylesdf: High-resolution 3d-consistent image and geometry generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13503–13513, June 2022. 3

[47] Rohit Pandey, Sergio Orts Escolano, Chloe Legendre, Christian Haene, Sofien Bouaziz, Christoph Rhemann, Paul Debevec, and Sean Fanello. Total relighting: learning to relight portraits for background replacement. *ACM Transactions on Graphics (TOG)*, 40(4):1–21, 2021. 3, 5

[48] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5865–5874, 2021. 2

[49] F. Pighin and J.P. Lewis. Performance-driven facial animation. In *ACM SIGGRAPH Courses*, 2006. 2

[50] Mike Seymour, Chris Evans, and Kim Libreri. Meet mike: Epic avatars. In *ACM SIGGRAPH 2017 VR Village*, SIGGRAPH '17, pages 12:1–12:2, New York, NY, USA, 2017. ACM. 2

[51] Ivan Skorokhodov, Sergey Tulyakov, Yiqun Wang, and Peter Wonka. Epigraf: Rethinking training of 3d gans. *arXiv preprint arXiv:2206.10535*, 2022. 3

[52] Jingxiang Sun, Xuan Wang, Yichun Shi, Lizhen Wang, Jue Wang, and Yebin Liu. Ide-3d: Interactive disentangled editing for high-resolution 3d-aware portrait synthesis. *ACM Transactions on Graphics (TOG)*, 41(6):1–10, 2022. 3

[53] Tiancheng Sun, Jonathan T Barron, Yun-Ta Tsai, Zexiang Xu, Xueming Yu, Graham Fyffe, Christoph Rhemann, Jay Busch, Paul Debevec, and Ravi Ramamoorthi. Single image portrait relighting. *ACM Transactions on Graphics (TOG)*, 38(4):1–12, 2019. 3

[54] Difei Tang, Juyong Zhang, Ketan Tang, Lingfeng Xu, and Lu Fang. Making 3d eyeglasses try-on practical. In *2014 IEEE*

International Conference on Multimedia and Expo Workshops (ICMEW), pages 1–6. IEEE, 2014. 3

[55] Ayush Tewari, Mohamed Elgharib, Gaurav Bharaj, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zollhofer, and Christian Theobalt. Stylerig: Rigging stylegan for 3d control over portrait images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6142–6151, 2020. 3

[56] Ayush Tewari, Ohad Fried, Justus Thies, Vincent Sitzmann, Stephen Lombardi, Kalyan Sunkavalli, Ricardo Martin-Brualla, Tomas Simon, Jason Saragih, Matthias Nießner, et al. State of the art on neural rendering. In *Computer Graphics Forum*, volume 39, pages 701–727. Wiley Online Library, 2020. 2, 3

[57] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2387–2395, 2016. 3

[58] Justus Thies, Michael Zollhöfer, Christian Theobalt, Marc Stamminger, and Matthias Nießner. Headon: Real-time reenactment of human portrait videos. *ACM Transactions on Graphics (TOG)*, 37(4):1–13, 2018. 3

[59] Daoye Wang, Prashanth Chandran, Gaspard Zoss, Derek Bradley, and Paulo Gotardo. Morf: Morphable radiance fields for multiview neural head modeling. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–9, 2022. 3

[60] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *Advances in Neural Information Processing Systems*, 34:27171–27183, 2021. 2, 3, 5

[61] Zhibo Wang, Xin Yu, Ming Lu, Quan Wang, Chen Qian, and Feng Xu. Single image portrait relighting via explicit multiple reflectance channel modeling. *ACM Transactions on Graphics (TOG)*, 39(6):1–13, 2020. 3

[62] Qianyi Wu, Xian Liu, Yuedong Chen, Kejie Li, Chuanxia Zheng, Jianfei Cai, and Jianmin Zheng. Object-compositional neural implicit surfaces. In *European Conference on Computer Vision*, pages 197–213. Springer, 2022. 3

[63] Jianfeng Xiang, Jiaolong Yang, Yu Deng, and Xin Tong. Gram-hd: 3d-consistent image generation at high resolution with generative radiance manifolds. *arXiv preprint arXiv:2206.07255*, 2022. 3

[64] Christopher Xie, Keunhong Park, Ricardo Martin-Brualla, and Matthew Brown. Fig-nerf: Figure-ground neural radiance fields for 3d object category modelling. In *2021 International Conference on 3D Vision (3DV)*, pages 962–971. IEEE, 2021. 2, 3, 7

[65] Yiheng Xie, Towaki Takikawa, Shunsuke Saito, Or Litany, Shiqin Yan, Numair Khan, Federico Tombari, James Tompkin, Vincent Sitzmann, and Srinath Sridhar. Neural fields in visual computing and beyond. In *Computer Graphics Forum*, volume 41, pages 641–676. Wiley Online Library, 2022. 3

[66] Yiran Xu, Badour AlBahar, and Jia-Bin Huang. Temporally consistent semantic video editing. In *European Conference on Computer Vision*, pages 357–374. Springer, 2022. 2, 3, 7, 8

[67] Yinghao Xu, Sida Peng, Ceyuan Yang, Yujun Shen, and Bolei Zhou. 3d-aware image synthesis via learning structural and textural representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18430–18439, June 2022. 3

[68] Bangbang Yang, Yinda Zhang, Yinghao Xu, Yijin Li, Han Zhou, Hujun Bao, Guofeng Zhang, and Zhaopeng Cui. Learning object-compositional neural radiance field for editable scene rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13779–13788, 2021. 3

[69] Xu Yao, Alasdair Newson, Yann Gousseau, and Pierre Hellier. A latent transformer for disentangled face editing in images and videos. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13789–13798, 2021. 2, 3

[70] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Basri Ronen, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. *Advances in Neural Information Processing Systems*, 33:2492–2502, 2020. 2

[71] Yu-Ying Yeh, Koki Nagano, Sameh Khamis, Jan Kautz, Ming-Yu Liu, and Ting-Chun Wang. Learning to relight portrait images via a virtual light stage and synthetic-to-real adaptation. *ACM Transactions on Graphics (TOG)*, 2022. 3, 5, 8

[72] Weidong Yin, Ziwei Liu, and Chen Change Loy. Instance-level facial attributes transfer with geometry-aware flow. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9111–9118, 2019. 3

[73] Miaolong Yuan, Ishtiaq Rasool Khan, Farzam Farbiz, Arthur Niswar, and Zhiyong Huang. A mixed reality system for virtual glasses try-on. In *Proceedings of the 10th international conference on virtual reality continuum and its applications in industry*, pages 363–366, 2011. 3

[74] Kai Zhang, Fujun Luan, Qianqian Wang, Kavita Bala, and Noah Snavely. Physg: Inverse rendering with spherical gaussians for physics-based material editing and relighting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5453–5462, 2021. 2

[75] Li Zhang, Noah Snavely, Brian Curless, and Steven M. Seitz. Spacetime faces: High resolution capture for modeling and animation. *ACM Trans. Graph.*, 23(3):548–558, 2004. 2

[76] Qian Zhang, Yu Guo, Pierre-Yves Laffont, Tobias Martin, and Markus Gross. A virtual try-on system for prescription eyeglasses. *IEEE computer graphics and applications*, 37(4):84–93, 2017. 3

[77] Xiuming Zhang, Pratul P Srinivasan, Boyang Deng, Paul Debevec, William T Freeman, and Jonathan T Barron. Nerfactor: Neural factorization of shape and reflectance under an unknown illumination. *ACM Transactions on Graphics (TOG)*, 40(6):1–18, 2021. 2