UNIVERSITY OF CALIFORNIA SAN DIEGO

Neural Methods for High-Fidelity Reconstruction and Editing

A thesis submitted in partial satisfaction of the
requirements for the degree Master of Science

in

Computer Science

by

Vishal Vinod

Committee in charge:

Professor Manmohan Krishna Chandraker, Chair
Professor Ravi Ramamoorthi
Professor Hao Su

2023

The Thesis of Vishal Vinod is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2023

DEDICATION

To Lord Sri Krishna - the absolute truth, my family, and my teachers.

TABLE OF CONTENTS

LIST OF SUPPLEMENTAL FILES

Video 1. TEGLO - High fidelity 3D reconstruction from single view images

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

Chapter 2, in full, has been submitted for publication of the material as it may appear in a conference, 2024, Vishal Vinod, Tanmay Shah, Dmitry Lagun. The thesis author was the co-primary investigator and author of this paper.

VITA

2020	Bachelor of Engineering, Visvesvaraya Technological University, India
2022–2023	Teaching Assistant, University of California San Diego
2023	Master of Science, University of California San Diego

ABSTRACT OF THE THESIS

Neural Methods for High-Fidelity Reconstruction and Editing

by

Vishal Vinod

Master of Science in Computer Science

University of California San Diego, 2023

Professor Manmohan Krishna Chandraker, Chair

High fidelity reconstruction and editing of objects is a challenging task in the graphics and vision community. Recent work in 3D reconstruction are unable to preserve high-frequency details in addition to enabling tasks such as texture transfer, primarily because they do not disentangle appearance from geometry. Further, reconstruction and editing methods for relighting applications learn a simplified reflectance model and are unable to account for long-range light transport effects such as subsurface scattering. This thesis presents two main directions of research for high fidelity reconstruction and object editing: First, we propose TEGLO (Textured EG3D-GLO) for learning textured 3D representations from single-view image collections. We train a conditional Neural Radiance Field (NeRF) without explicit 3D supervision and creating a

dense correspondence mapping to a 2D canonical coordinate space to equip our method with texture transfer and editing with near perfect reconstruction ($>$74 db PSNR) even at megapixel resolution. Second, we find that recent work in high fidelity relighting explore subsurface scattering with objects where scattering is the primary light transport effect. These methods are unable to model specular highlights which occur when relighting human faces. Toward this, we render a synthetic OLAT dataset of human face images in a virtual light stage with suitable ground truth for reconstruction and relighting. We explore a hybrid physical-neural approach to surface relighting by effectively combining insights from a physically based prior and a neural renderer to improve the fidelity in modeling specular highlights and subsurface scattering effects in relighting human faces.

Chapter 1

Introduction

High fidelity reconstruction and editing of objects is a challenging task with critical applications in virtual reality, content creation and telepresence systems. Recent work in Neural Radiance Fields (NeRFs) [12, 28, 11, 61, 77, 88] explore learning 3D representations from single-view in-the-wild image collections by leveraging the inductive bias across a dataset of class-specific objects. However, these methods face several issues: they fail to preserve high frequency details, have constraints in their rendering resolution and some methods require 3D supervision. Further, previous work do not disentangle appearance from geometry and are hence not suitable for tasks such as texture editing and texture transfer. Toward this, we propose TEGLO (Textured EG3D-GLO) for learning textured 3D representations from single view in-the-wild image collections for a given class of objects. We accomplish this by training a conditional Neural Radiance Field (NeRF) without any explicit 3D supervision. We equip our method with editing capabilities by creating a dense correspondence mapping to a 2D canonical space. We demonstrate that such mapping enables texture transfer and texture editing without requiring meshes with shared topology. Our key insight is that by mapping the input image pixels onto the texture space we can achieve near perfect reconstruction ($\geq$74 dB PSNR at 1024^2 resolution). Our formulation allows for high quality 3D consistent novel view synthesis with high-frequency details at arbitrary resolution (even at megapixel image resolutions).

High fidelity relighting of complex objects such as human faces include modeling

direct lighting effects such as specular highlights and long-range light transport effects such as subsurface scattering. However, prior work model reflectance as diffuse reflectance or as a simple BRDF which are unable to model subsurface scattering effects. Further, Recent work such as OSF [94], an implicit neural rendering model based on NeRFs that learns to approximate cumulative radiance transfer, and [100], a neural implicit rendering model that learns to approximate the radiance transfer gradient, are unable to model specular highlights and are primarily trained and evaluated on objects where subsurface scattering effects are the primary light transport effects. In our experiments with OSF, we augment a pre-trained model with a trainable light-weight neural renderer and observe that the residual specular effects can be learned effectively. We observe that OLAT datasets for human faces are more often than not inaccessible to the general research community primarily due to the privacy and licensing issues involved. While NVPR [97] make available an OLAT light stage dataset for human faces, the dataset does not include camera poses necessary to train a NeRF-based model. Our experiments with a self-supervised physics-based reconstruction method on the NVPR dataset demonstrated the under-constrained nature of the problem and motivates the need for obtaining ground truth for different components of light transport. To address this, we render an OLAT dataset of human faces with ground truth for surface normals, direct lighting, specular map, albedo and subsurface scattering. This motivates our investigation on a hybrid physical-neural rendering based surface relighting method that draws on insights from a physics-based prior for the direct lighting estimate and a neural renderer for the subsurface scattering component.

In Chapter 2, we discuss TEGLO (Textured EG3D-GLO) for textured 3D reconstruction of objects from single-view image collections. Our key insight is that by disentangling texture from geometry by using the 3D surface points (of objects) to learn a dense correspondence mapping via a 2D canonical coordinate space, we can extract a texture for each object. Then, by using the learned correspondences to map the pixels from the input image of the object onto the texture, we enable preserving high-frequency details even at megapixel resolution. As expected, copying the input image pixels onto the texture accurately, allows near perfect reconstruction

while preserving high frequency details with multi-view consistent representations. We show that TEGLO enables several tasks with high fidelity such as texture transfer, texture editing and single view 3D reconstruction.

In Chapter 3, we consider the problem of reconstructing and relighting surfaces modeling complex light transport effects such as soft shadows and subsurface scattering effects. Traditional methods suffer from two drawbacks: First, they consider relighting to be a purely physically-based approach using reconstruction followed by physically-based rendering, or with a purely data-driven approach. We propose a hybrid physical-neural approach to potentially benefit from PBR and neural rendering. Second, current methods model the reflectance as diffuse Lambertian or as a simplified BRDF to account for specular properties. They fail to consider subsurface scattering of skin. In this work, we aim to perform high-quality surface relighting modeling long-range light interactions such as subsurface scattering effects. We also present current progress on 3D-aware photorealistic re-lighting and novel view synthesis. Since a self-supervised formulation by reconstructing the input OLAT images is unconstrained, we render a synthetic dataset of OLAT images of faces with suitable ground truth for supervising our method. We include further details in the synthetic OLAT light stage dataset rendering in Chapter 3.

Chapter 2

Conditional NeRF for High-Fidelity Textured 3D Reconstruction

2.1 Introduction

Reconstructing high-resolution and high-fidelity 3D consistent representations from single-view in-the-wild image collections is critical for applications in virtual reality, 3D content creation and telepresence systems. Recent work in Neural Radiance Fields (NeRFs) [12, 28, 11, 61] aim to address this by leveraging the inductive bias across a dataset of single-view images of class-specific objects for 3D consistent rendering. However, they are unable to preserve high frequency details while reconstructing the input data despite the use of SIREN [70] or positional encoding [50], in part due to the properties of MLPs they use [15]. For arbitrary resolution 3D reconstruction from single-view images, these methods face several challenges. These include image-space approximations that break multi-view consistency constraining the rendering resolution [11], requiring Pivotal Tuning Inversion (PTI) [63] or fine-tuning for reconstruction [28, 11, 71] and the inability to preserve high-frequency details [28, 11, 71, 61]. To address this, we propose TEGLO (Textured EG3D-GLO) that uses a tri-plane representation [11] and Generative Latent Optimization (GLO) [9] based training to enable efficient and high-fidelity 3D reconstruction and novel view synthesis at arbitrary image resolutions from single-view image collections of objects.

Recent works disentangle texture from geometry [15, 90] and enable challenging tasks

Figure 2.1. Teaser - Demonstrating TEGLO for high fidelity 3D reconstruction and multi-view consistent texture representation and texture editing from single-view image collections of objects.

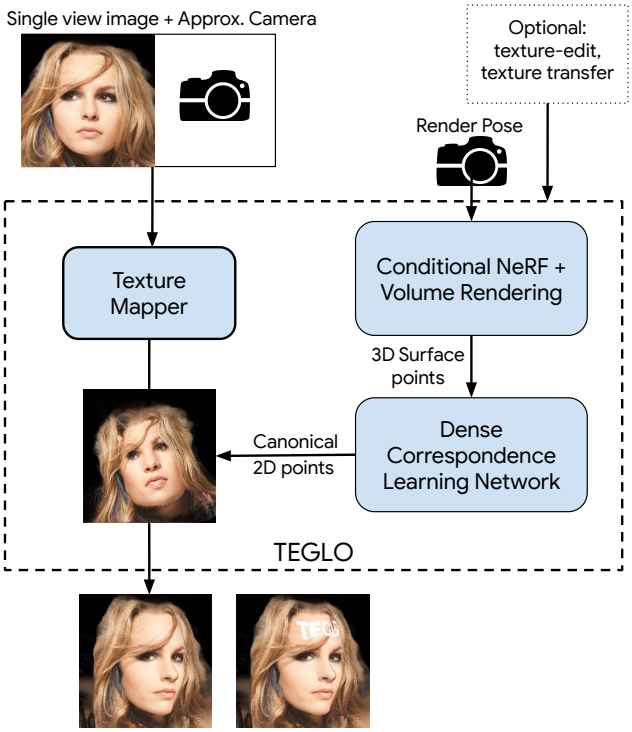


Figure 2.2. Overview - TEGLO enables 3D reconstruction and texture representation from single-view image collections of objects.

such as texture editing and texture transfer. However, they depend on large-scale textured mesh data for high-fidelity 3D reconstruction which is laborious, expensive and time intensive to capture. Further, the use of a capture environment may cause a dataset-shift leading to generalization issues in downstream tasks, and the data use may require custom licensing. All of these factors limit access from the broader research community. This motivates the need for a method to learn textured 3D representations from single-view in-the-wild images of objects. However, the task of disentangling texture and 3D geometry from in-the-wild image collections is a formidable challenge due to the presence of wide variations in poses, partial views, complex details in appearance, geometry, noise in the given image collection. Inspired by surface fields [27], TEGLO leverages the 3D surface points of objects extracted from a NeRF to learn dense correspondences via a canonical coordinate space to enable texture transfer, texture editing and high-fidelity single-view 3D reconstruction.

Our key insight is that by disentangling texture and geometry using the 3D surface points

of objects to learn a dense correspondence mapping via a 2D canonical coordinate space, we can extract a texture for each object. Then, by using the learned correspondences to map the pixels from the input image of the object onto the texture, we enable preserving high-frequency details. As expected, copying the input image pixels onto the texture accurately, allows near perfect reconstruction while preserving high frequency details with multi-view consistent representations. In this work, we present TEGLO, a tri-plane and GLO-based conditional NeRF, and a method to learn dense correspondences to enable challenging tasks such as texture transfer, texture editing and high-fidelity 3D reconstruction even at large megapixel resolutions. We also show that TEGLO enables single-view 3D reconstruction with no constraints on resolution by simply inverting the image into the latent table without any PTI [63] or fine-tuning. We present an overview of TEGLO in Fig.(2.2): TEGLO takes a single-view image and its approximate camera pose to map the pixels onto a texture. Then, to render the object from a different view, we extract the 3D surface points from the trained NeRF and use the dense correspondences to obtain the color for each pixel from the texture. Optionally, TEGLO allows texture edits and texture transfer across objects. In summary, our contributions are:

1. A novel 3D dataset rendering method leveraging the advantages of generative latent optimization and hybrid implicit-explicit representations for textured 3D reconstruction from 2D image collections of objects.
2. A conditional NeRF with a tri-plane representation and GLO auto-decoder based training that enables efficient 3D consistent rendering at arbitrary resolutions.
3. A framework for effectively mapping the pixels from an in-the-wild single-view image onto a texture to enable high-fidelity 3D consistent representations preserving high-frequency details.
4. A method for extracting canonical textures from single-view images enabling tasks such as texture editing and texture transfer for NeRFs.
5. Demonstrating effective mapping of single-view image pixels to a canonical texture space

while preserving 3D consistency and achieving near perfect reconstruction ($\geq$ 74 dB PSNR at 1024^2 resolution).

2.2 Related Work

3D-aware generative models. Learning 3D representations from multi-view images with camera poses have been extensively studied since the explosion of Neural Radiance Fields (NeRFs) [50, 72, 96, 7, 98, 28]. However, these methods require several views and learn a radiance field for a single scene. RegNeRF [53] reduces the need from several views to only a handful, however, the results have several artifacts. Recently, several works learn 3D representations from single-view images [12, 11, 43, 71, 61, 99]. Further, [74, 73, 76, 37] enable multi-view consistent editing, however, they are limited by the rendering resolution. Recent work propose single image 3D consistent novel view synthesis [93, 44, 29, 84], however they are not yet suitable for texture representation. While point cloud based diffusion models [95, 52] enable learning 3D representations, they have limited applicability in textured 3D generation and high fidelity novel view synthesis. In this work, we show that TEGLO learns textured 3D representations from class-specific single-view image collections.

Texture representation. Template based methods [58, 8, 16, 31] deform a template mesh prior for 3D representations and are hence restricted in the topology they can represent. Texture Fields [54] enable predicting textured 3D models given an image and a 3D shape, but are unable to represent high-frequency details. While NeuTex [86] enables texture representation, it does not allow multi-view consistent texture editing at the desired locations due to a contorted UV mapping [90]. NeuMesh [90] learns mesh representations to enable texture transfer and texture editing using textured meshes. However, it performs mesh-guided texture transfer and requires spatial-aware fine-tuning for mesh-guided texture edits. While GET3D [25] learns textured 3D shapes by leveraging tri-plane based geometry and texture generators, it requires 2D silhouette supervision and is limited to synthetic data. AUVNet [15] represents textures from textured meshes by learning an aligned UV mapping and demonstrates texture transfer.

However, it depends on textured mesh data and requires multiple networks to enable single-view 3D reconstruction. In contrast, TEGLO learns textured 3D consistent representations from single-view images by inverting the image into the latent table.

Dense correspondences. Previous work in dense correspondence learning involve supervised [18, 41] or unsupervised [89, 87] learning methods. CoordGAN [51] learns dense correspondences by extracting each image as warped coordinate frames transformed from correspondence maps which is effective for 2D images. However, CoordGAN is unable to learn 3D correspondences. AUVNet [15] establishes dense correspondences across 3D meshes via a canonical UV mapping and asserts that methods that do not utilize color for dense correspondence learning [24, 45] may have sub-par performance in texture representation.

2.3 TEGLO

Given a collection of single-view in-the-wild images of objects and their approximate camera poses, TEGLO aims to learn a textured 3D representation of the data. TEGLO consists of two stages: 3D representation learning and dense correspondence learning. TEGLO Stage-1 consists of a conditional NeRF leveraging a Tri-Plane representation and an auto-decoder training regime based on generative latent optimization (GLO) [9] for 3D reconstruction of the image collection. To train TEGLO Stage-2, we use TEGLO Stage-1 to render a dataset of an object's geometry from five views using the optimized latent code. TEGLO Stage-2 uses the 3D surface points from the rendered dataset to learn dense pixel-level correspondences via a 2D canonical coordinate space. Then, the inference stage uses the learned dense correspondences to map the image pixels from the single-view input image onto a texture extracted from TEGLO-Stage 2. As a result, TEGLO effectively preserves high frequency details at an unprecedented level of accuracy even at large megapixel resolutions. TEGLO disentangles texture and geometry enabling texture transfer (Fig.(2.14)), texture editing (Fig.(2.12)) and single view 3D reconstruction without requiring fine-tuning or PTI (Fig.(2.11)).

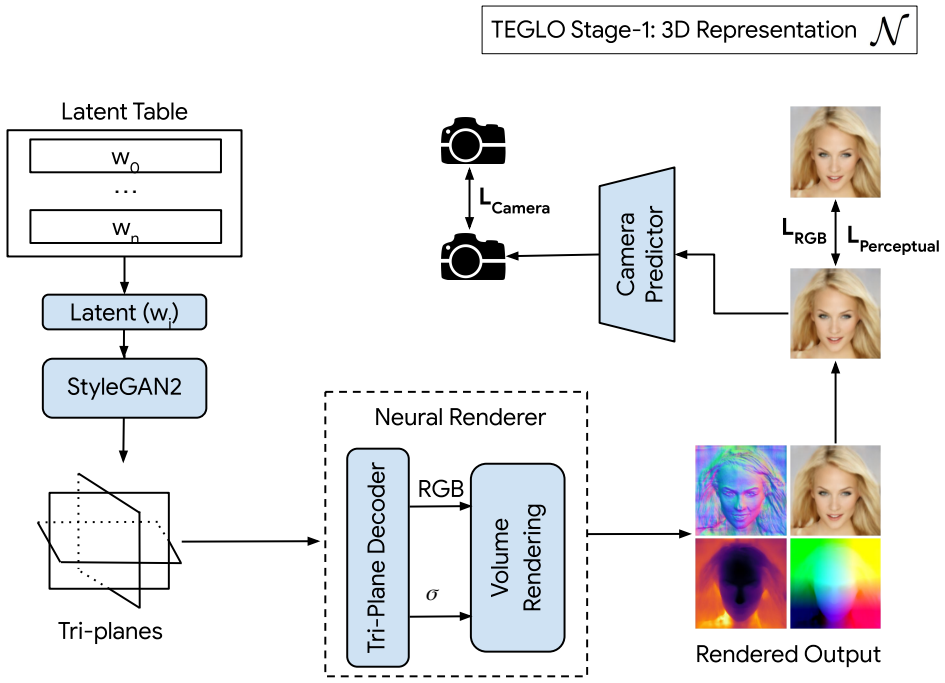


Figure 2.3. TEGLO Stage-1 Architecture - Uses a tri-plane and GLO based conditional NeRF to learn a per-object table of latents to reconstruct the single-view image collection.

2.3.1 Stage-1: Conditional NeRF for 3D Reconstruction

Formulation. We denote the single-view image collection ($\mathscr{I}$) with class specific objects as $\{o_0, o_1, ..., o_n\} \in \mathscr{I}$. To learn 3D representations, TEGLO uses a generative latent optimization (GLO) based auto-decoder framework, where the NeRF is conditioned on an image specific latent vector $\{w_0, w_1, ..., w_n\} \in \mathscr{R}^D$ to effectively reconstruct the image without requiring a discriminator.

Network architecture. The NeRF model $\mathscr{N}$ is represented by TEGLO Stage-1 in Fig.(2.3). The model $\mathscr{N}$ passes the input conditioning latent w_i to a set of CNN-based synthesis layers [36] whose output feature maps are used to construct a k-channel tri-plane. The sampled points on each ray are used to extract the tri-plane features and aggregate the k-channel features. Then the tri-plane decoder MLP outputs the scalar density σ and color which are alpha-composited by volume rendering to obtain the RGB image. Volume rendering along

camera ray $r(t) = O + td$ is:

$$\mathscr{C}_{\text{NeRF}}(r,w) = \int_{b_n}^{b_f} T(t,w)\sigma(r(t),w)\mathbf{c}(r(t),\mathbf{d},w)dt \tag{2.1}$$

$$\text{where } T(t,w) = \exp\left(-\int_{b_n}^{b_f} \sigma(r(s),w)\right)ds$$

Here, the radiance values can be replaced with the depth $d(x)$ or pixel opacity to obtain the surface depth. During inference, the surface depth map and 2D pixel coordinates are used to obtain the 3D surface points via back-projection. The surface normals can be computed as the first derivative of the density σ with respect to the input as follows:

$$\widehat{n}(r,w) = -\int_{b_n}^{b_f} T(t,w)\ \sigma(r(t),w)\ \nabla_{r(t)}(\sigma(r(t),w))dt$$

$$n(r,w) = \frac{\widehat{n}(r,w)}{||\ \widehat{n}(r,w)\ ||_2} \tag{2.2}$$

Thus from an inference step, an RGB image, surface depth map, 3D surface points and the surface normals of the object instance can be obtained. In Fig.(2.4), we show the sample reconstruction results for $\mathscr{N}$ on the CelebA-HQ, AFHQv2 and ShapeNet-Cars datasets. In Fig.(2.5) we show qualitative results for novel view synthesis with $\mathscr{N}$ trained on SRN-Cars and evaluated on a held-out set of views. Since SRN-Cars is a multi-view dataset, we compare the rendered novel views with their corresponding ground-truth views.

Losses. $\mathscr{N}$ is trained by reconstructing the image and simultaneously optimizing a latent (w_i). As noted in [61], this allows the training loss to be enforced on individual pixels enabling training and inference at arbitrary image resolutions. For TEGLO Stage-1 (Fig.(2.3)), three losses are minimized to train $\mathscr{N}$: $\mathscr{L}_{\text{RGB}}$, is an $\mathscr{L}_1$ reconstruction loss between the rendered image and the ground truth image for o_i. The $\mathscr{L}_{\text{Perceptual}}$ loss is a LPIPS (Learned Perceptual Image Patch Similarity) loss between rendered image and the ground truth image. The $\mathscr{L}_{\text{Camera}}$ is the camera prediction $\mathscr{L}_1$ loss between the output of the camera encoder and the ground-truth

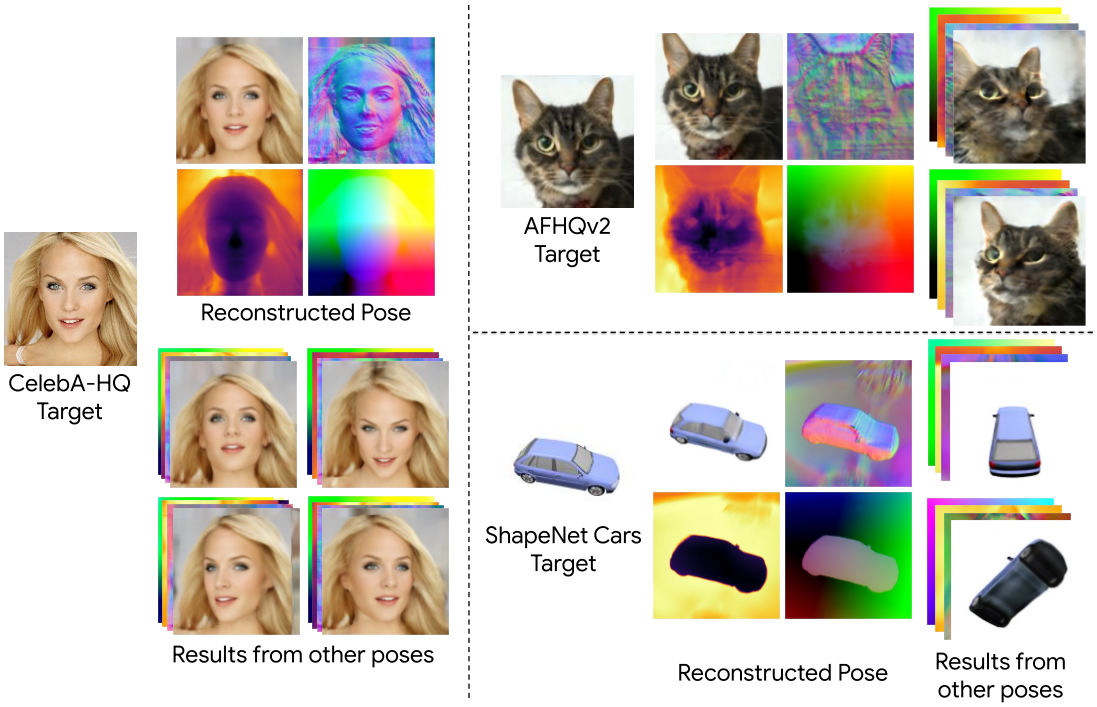


Figure 2.4. Rendering the dataset for TEGLO Stage-2 - Rendering multiple views of images, surface normals, depth maps and 3D surface points from CelebA-HQ, AFHQv2-Cats and ShapeNet-Cars for training TEGLO Stage-2.

camera parameters for the camera pose to learn 3D consistent representations of the object $(o_i \in \mathscr{I})$.

$$\mathscr{L}_{\mathscr{N}} = \mathscr{L}_{\text{RGB}} + \mathscr{L}_{\text{Perceptual}} + \mathscr{L}_{\text{Camera}} \tag{2.3}$$

To train $\mathscr{N}$, we use the single-view image dataset and the approximate pose for each $o_i \in \mathscr{I}$ (Sec.(2.4)). We train the model for 500K steps using the Adam optimizer [38] on 8 NVIDIA V100 (16 GB) taking 36 hours to complete.

Design choices. As noted in Sec.(1), EG3D [11] shows medium resolution (512^2) capacity while using image-space approximations in the super-resolution module which negatively affects the geometric fidelity [71]. While EpiGRAF [71] uses a patch-based discriminator for pure 3D generation, it is still prone to issues in scaling and training with multi-resolution data. Moreover, adversarial training using discriminators leads to training instability. Different from EG3D and EpiGRAF that use an adversarial training paradigm, $\mathscr{N}$ uses a GLO-based auto-

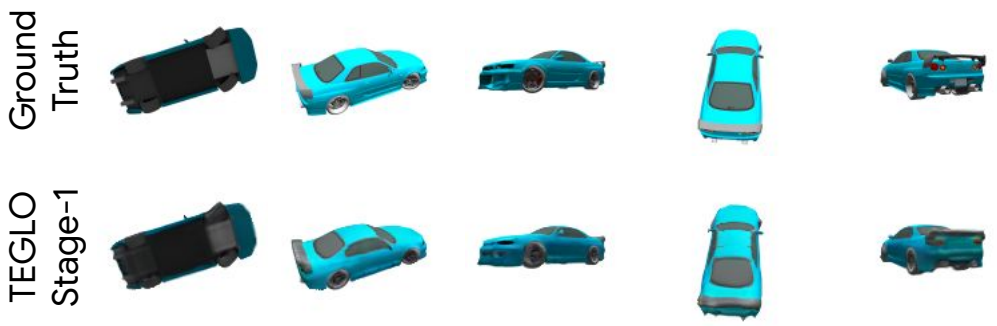


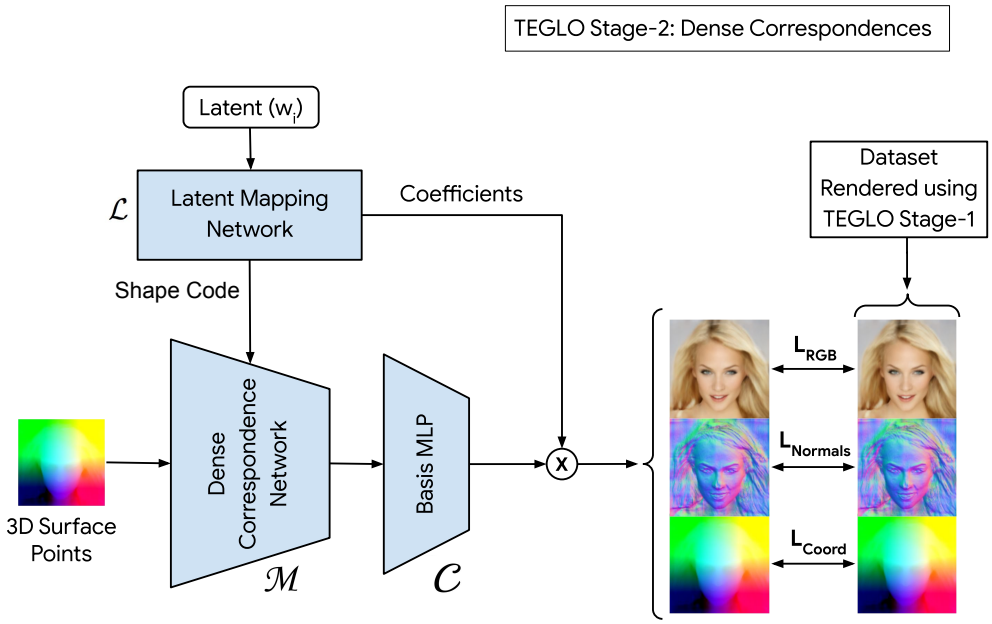
Figure 2.5. Novel View Synthesis - TEGLO Stage-1 results for ShapeNet-Cars data.

decoder training paradigm which jointly optimizes a latent representation and reconstructs the image enabling arbitrary resolution synthesis - even at large megapixel resolutions - without the constraints of a discriminator. Hence, $\mathscr{N}$ enables 3D representations with geometric fidelity while also benefiting from an efficient tri-plane based representation.

EG3D [11] requires camera pose conditioning for the generator and discriminator to establish multi-view consistency. The limitation of a pose-conditioned generator is that it does not completely disentangle the pose from appearance which leads to artifacts such as degenerate solutions (2D billboards), or expressions such as the eye or smile following the camera. Since $\mathscr{N}$ optimizes a latent representation of an object and reconstructs it, we observe that the generator does not require camera pose conditioning and simply using a light-weight camera predictor network and training with a camera prediction loss ($\mathscr{L}_{\text{Camera}}$) is sufficient to learn 3D consistent representations.

2.3.2 Stage-2: Learning Dense Correspondences

Formulation. We render a multi-view dataset ($\mathscr{D}$) using $\mathscr{N}$ trained on single-view image collections for the task of texture representation. We denote each object $e_i \in \mathscr{D}$ comprising of five views: $e_i = \{v_f, v_l, v_r, v_t, v_b\}$ where v denotes the view, and the sub-scripts (j for all v_j) denote frontal, left, right, top and bottom poses respectively (refer Fig.(2.4)). In $\mathscr{D}$, each view $v_j \in e_i$ includes the depth map ($\widehat{d}_j$), RGB image ($\widehat{r}_j$), surface normals ($\widehat{s}_j$), 3D surface points



TEGLO Stage-2

Figure 2.6. TEGLO Stage-2 Architecture - TEGLO Stage-2 learns dense correspondences via a 2D canonical coordinate space mapping.

($\widehat{p}_j$), and the optimized latent, w_i, which is identical for views of e_i as it is independent of camera pose (Fig.(2.4)). For TEGLO Stage 2, we use $\{\{\widehat{r}_j, \widehat{s}_j, \widehat{p}_j\} \in v_j, w_i\} \in e_i\}$.

Learning dense pixel-level correspondences across multiple views of an object is the task of locating the same 3D coordinate point in a canonical coordinate space. Inspired by surface fields [27], we aim to learn dense correspondences using the 3D surface points extracted from $\mathcal{N}$ by back-projecting the depth ($\widehat{d}_j$) and pixel coordinates. Inspired by CoordGAN [51] and AUVNet [15], we propose a dense correspondence learning network in TEGLO Stage-2 trained in an unsupervised manner learning an aligned canonical coordinate space to locate the same 3D surface point across different views (v_j) of the same object (e_i).

Network architecture. TEGLO Stage-2 (Fig.(2.6)) consists of a latent mapping network ($\mathscr{L}$), a dense correspondence network ($\mathscr{M}$) and a basis network ($\mathscr{C}$) - all of which are MLP networks. The 3D surface points ($\widehat{p}_j$) from $v_j \in e_i$) are mapped to a 2D canonical coordinate space conditioned on a shape code mapped from the optimized latent w_i for e_i. We use a Lipschitz

regularization [46] for each MLP layer in the dense correspondence network ($\mathscr{M}$). The latent mapping network ($\mathscr{L}$) is a set of MLP layers that takes the w_i-latent for e_i as input and predicts a shape-code for conditioning $\mathscr{M}$, and coefficients for the deformed basis. Previous work [78, 15] show that if the input is allowed to be represented as a weighted sum of basis images, to obtain a deformed basis before decomposition, then the 2D canonical coordinate space will be aligned. The basis network ($\mathscr{C}$) is similar to [15] and uses the predicted coefficients to decompose the deformed coordinate points. Thus, $\mathscr{M}$ maps the 3D surface points to an aligned 2D canonical coordinate space, enabling the network to learn dense correspondences using $p_j \in \mathscr{S}$ extracted from $\mathscr{N}$. Next, the basis network takes the 2D canonical coordinates as input to predict the deformed basis $\mathscr{B}$. Then, $\mathscr{B}$ is weighted with the predicted coefficients to decompose the basis into the 3D surface points (p_j), surface normals (s_j) and color (r_j).

Losses. TEGLO Stage-2 is trained using three $\mathscr{L}_2$ reconstruction losses: the $\mathscr{L}_{\text{RGB}}$ loss between the rendered RGB image $\widehat{r}_j$ and the predicted RGB image r_j; the $\mathscr{L}_{\text{Normals}}$ loss between the rendered surface normals $\widehat{s}_j$ and the predicted surface normals s_j; $\mathscr{L}_{\text{Coord}}$ loss between the extracted 3D surface points $\widehat{p}_j$ and the predicted 3D surface points p_j.

$$\mathscr{L}_{\text{Stage2}} = \mathscr{L}_{\text{RGB}} + \mathscr{L}_{\text{Normals}} + \mathscr{L}_{\text{Coord}} \tag{2.4}$$

To train TEGLO Stage-2, we use the rendered dataset $\mathscr{D}$ consisting of 1000 objects with five views per object and the optimized latent for each identity. The networks are trained using $\mathscr{L}_{\text{Stage2}}$ loss for 1000 epochs using the Adam [38] optimizer to learn dense correspondences across $e_i \in \mathscr{D}$.

Design choices. We use the optimized w-latent from $\mathscr{N}$ for learning the shape code and coefficients for TEGLO Stage-2 because it represents the 3D geometry and appearance information for object (e_i) independent of camera pose. We observe that using a Lipschitz regularization for every MLP layer in $\mathscr{M}$ suitably regularizes the network to deform the input surface points $\widehat{s}_j$. Interestingly, our experiments show that simply reconstructing the 3D surface

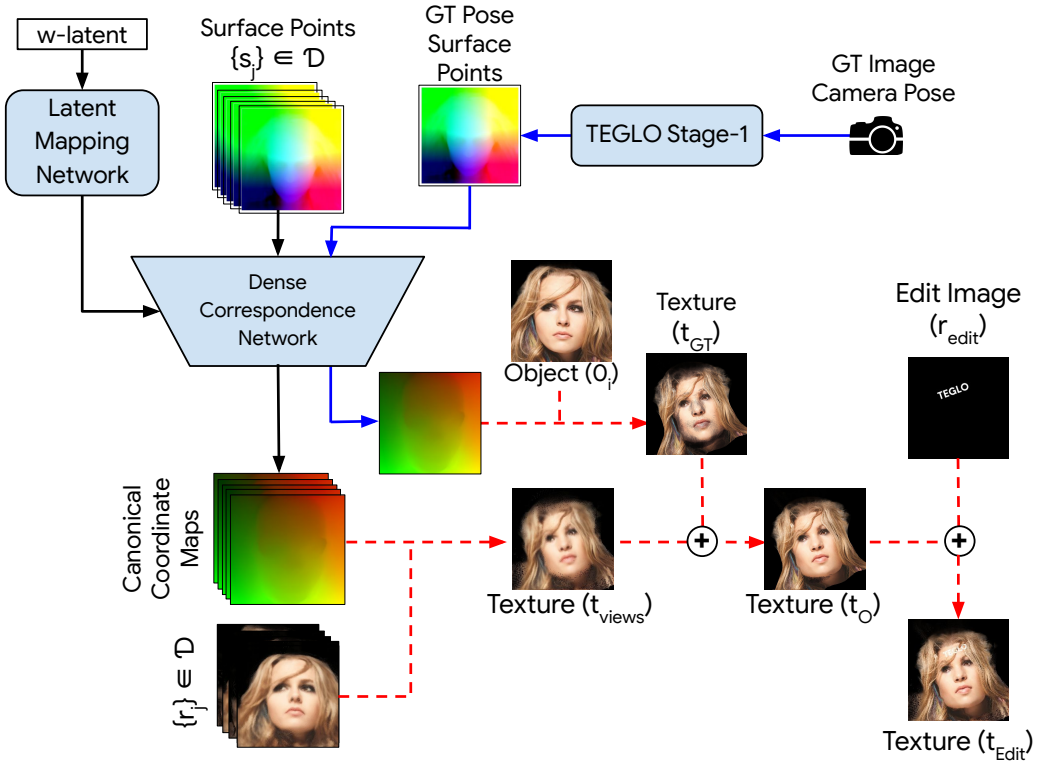


Figure 2.7. Inference - TEGLO texture extraction for texture transfer and editing. Red arrows indicate the use of a K-d tree to store the texture. Blue arrows indicate the use of GT pixels.

points instead of the color, surface points and surface normals also leads to learning reasonable dense pixel-level correspondences. We show qualitative results for TEGLO Stage-2 trained using only $\mathscr{L}_{\text{Coord}}$ loss in Fig.(2.8) as TEGLO-3DP.

2.3.3 Textured 3D Reconstruction

Extracting the texture. We use the learned dense correspondences from TEGLO Stage-2 to extract a texture map for each object $o_i \in \mathscr{I}$. We use the pose of the target image o_i to extract the 3D surface points from $\mathscr{N}$ and use it to map the image pixels to the 2D canonical coordinate space. We denote this as texture t_{GT}. Similarly, we use $\mathscr{M}$ to map the respective RGB values from $\{v_f, v_l, v_r, v_t, v_b\} \in e_i$ using the corresponding 3D surface points (s_j) from five views to the 2D canonical space and denote it as t_{views}. Thus, textures t_{GT} and t_{views} store a mapping the canonical coordinate point and the corresponding RGB values. The procedure is represented in Fig.(2.7) and textures are depicted in Fig.(2.12) and Fig.(2.10). In Fig.(2.7) t_O represents

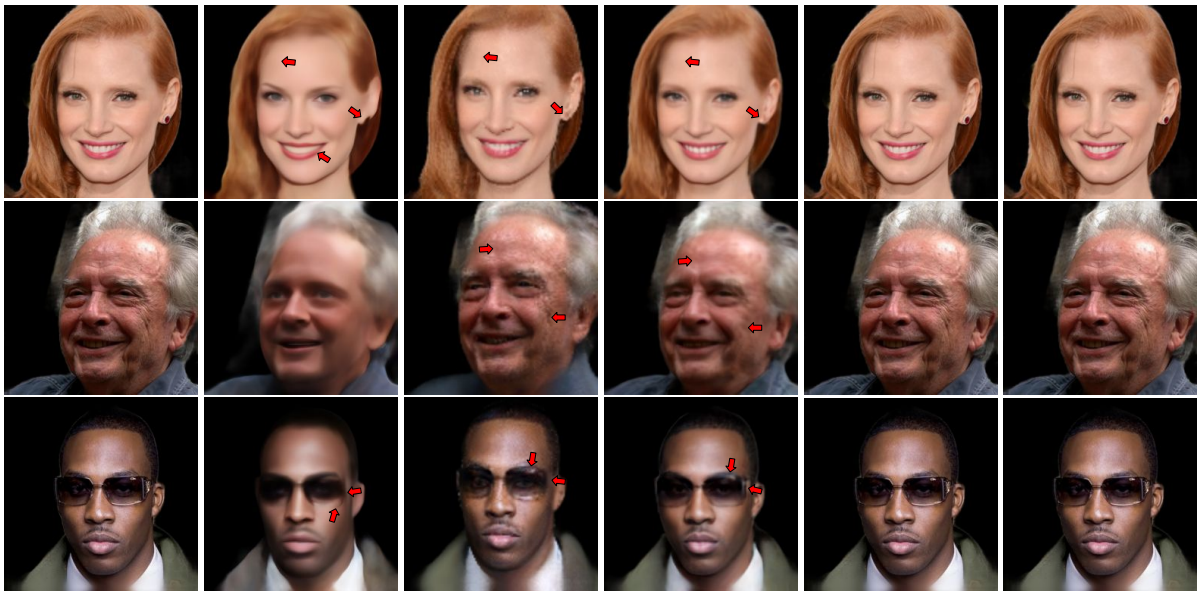


Figure 2.8. Qualitative results - Comparison with relevant 3D-aware generative baseline methods at 256^2 resolution for CelebA-HQ.

the texture obtained by combining t_{GT} and t_{views}. We store this mapping in a K-d tree which enables us to index into the textures using accurate floating point indices to obtain the RGB values. The K-d tree allows querying with canonical coordinates to extract multiple neighbors making TEGLO robust to sparse "holes" in the texture. Refer Fig.(2.9).

Novel view synthesis. For rendering novel views of o_i, we extract the 3D surface points for the pose from $\mathscr{N}$ and obtain the canonical coordinates from $\mathscr{M}$. For each 2D canonical coordinate point c_k, we query the K-d tree for three natural neighbors and obtain indices for the neighbors which are used to obtain the RGB values. Natural Neighbor Interpolation (NNI) [69] enables fast and robust reconstruction of a surface based on a Dirichlet tesselation - unique for every set of query points - to provide an unambiguous interpolation result. We simplify the natural neighbor interpolation (NNI) based only on the distances of the points c_k in the 2D canonical coordinate space to obtain the RGB values from the stored texture. The robust and unambiguous interpolation enables TEGLO to effectively map the ground-truth image pixels from the input dataset $\mathscr{I}$ onto the geometry for novel view synthesis. To extract the Surface Field $\mathscr{S}$, we render e_i from five camera poses which may potentially cause camera pose biases

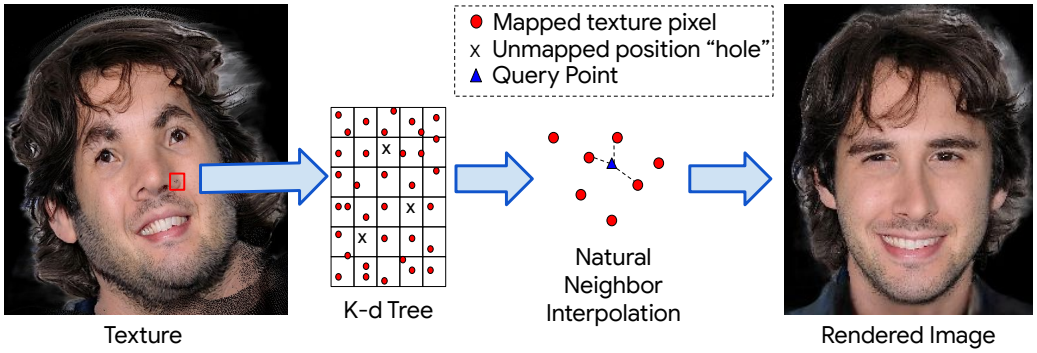


Figure 2.9. Interpolating textures with sparse "holes" - Depicting the KD-Tree and Natural Neighbor Interpolation (NNI) to interpolate any "holes" in the texture for novel view synthesis.

leading to sparse "holes" in the texture. Our formulation uses the K-d tree and NNI to interpolate and index into the textures with sparse "holes" (Refer Fig.(S6)). In Fig.(2.9), each cell in the 5x6 grid represents a discrete pixel in the texture space and the red dot represents a canonical coordinate point. There are three issues that may arise:

1. The canonical coordinate points may not be aligned to the pixel centers and storing them in the discretized texture space may lead to imprecision.
2. There may be multiple canonical coordinates mapped to a discrete integral pixel wherein some coordinates may need to be dropped for an unambiguous texture indexing - leading to loss of information.
3. Some pixels may not be mapped to by any canonical coordinates, creating a "hole" in discretized space. This is represented by "**X**" in the grid in Fig.(2.9).

K-d tree allows extracting multiple neighbors by querying with canonical coordinate points and also enables indexing the texture using floating point values. Hence, using a K-d tree to store the texture helps address (1) and (2). Further, using a K-d tree in conjunction with Natural Neighbor Interpolation (NNI) effectively addresses (3). Natural Neighbor Interpolation (NNI) is formulated as follows:

$$\mathrm{NNI}(x) = \sum_{i=0}^{n} w_i(x) \times f(x_i) \tag{2.5}$$

$$w_i(x) = \frac{\frac{1}{d_i(x)}}{\sum_{j=0}^{n} \frac{1}{d_j(x)}} \tag{2.6}$$

Where x is the query point, $w_i(x)$ is the simplified Laplace weight based on inverse distances to n neighbors corresponding to the polygon potentially encroached by the query point in the Voronoi tessellation plot, and $f(x_i)$ represents the extracted texture pixels. Storing the texture in the K-d tree and using Natural Neighbor Interpolation enables accurate and unambiguous (property of NNI using tessellation plot) floating point indexing into the texture to obtain the RGB color.

Texture editing. Texture with edits are represented as t_{Edit} in Fig.(2.7). We create the edits on a blank image the same size as t_{O} and denote it as r_{edit}. The edit image r_{edit} is considered to be in the canonical space and is directly indexed into the K-d tree to be overlay on t_O. Note that we do not constrain the texture space and it may be visually aligned to a canonical pose as in Fig.(2.12) and Fig.(2.10). The texture with an edit (t_{Edit}) is created by overlaying r_{edit} on t_O. Qualitative results are in Fig.(2.1) and Fig.(2.12).

2.4 Experiments and Results

2.4.1 Datasets

We train TEGLO with single-view image datasets such as FFHQ [36], CelebA-HQ [47, 34] and AFHQv2-Cats [35, 17]. To obtain the approximate camera pose, we follow [61] by first using an off-the-shelf face landmark predictor MediaPipe Face Mesh [3] to extract landmarks appearing at consistent locations. Then, we use a shape-matching least-squares optimization to align the landmarks with 3D canonical landmarks to obtain the approximate pose. We also use a multi-view image dataset - ShapeNet-Cars [14, 13] with results in Fig.(2.1) and Table.(2.4).

Table 2.1. Reconstruction of train images - Quantitative comparison on training data reconstruction at 128^2 resolution.

Method	PSNR (↑)	LPIPS (↓)
π-GAN [12] (CelebA)	23.5	0.226
LoLNeRF [61] (FFHQ)	29.0	0.199
LoLNeRF [61] (CelebA-HQ)	29.1	0.197
ABC [62] (CelebA-HQ)	26.3	-
TEGLO Stage 1 (FFHQ)	29.0	0.294
TEGLO Stage 1 (CelebA-HQ)	28.9	0.317
TEGLO (CelebA-HQ)	**89.5**	**2.3e-7**

Table 2.2. Reconstruction of test images - Quantitative comparison on test data reconstruction at various rendering resolutions.

Method	Res.	PSNR (↑)	LPIPS (↓)
π-GAN [12] (CelebA)	256^2	21.8	0.412
LoLNeRF [61] (FFHQ)	512^2	25.3	0.491
LoLNeRF [61] (CelebA-HQ)	256^2	26.2	0.363
TEGLO Stage 1 (FFHQ)	256^2	27.3	0.334
TEGLO Stage 1 (CelebA-HQ)	256^2	27.5	0.260
TEGLO (FFHQ)	256^2	**84.9**	**2.1e-6**
TEGLO (CelebA-HQ)	256^2	**86.2**	**7.4e-7**
TEGLO (CelebA-HQ)	512^2	82.6	4.4e-6
TEGLO (CelebA-HQ)	1024^2	74.7	6.9e-5

2.4.2 3D Reconstruction

We evaluate TEGLO on the task of reconstructing the input image in the same pose and compare with baseline methods. We report quantitative results for train data reconstruction in Table.(2.1) measuring the PSNR (Peak Signal to Noise Ratio) and LPIPS (Learned Perceptual Image Patch Similarity) metrics for CelebA-HQ and FFHQ. We observe similar results for LoLNeRF and TEGLO Stage-1 at 128^2 resolution. However, as expected, TEGLO attains 89.5 dB PSNR and $7.4e$-7 for LPIPS. We report quantitative results for test data reconstruction from a held-out set at 256^2 resolution for CelebA-HQ and FFHQ data in Table.(2.2) and for AFHQv2-Cats data in Table.(2.3).

We depict qualitative results for CelebA-HQ in Fig.(2.8) where the red arrows indicate missing details. For EG3D-Fit, we invert the image into the EG3D [11] latent space and perform

Table 2.3. Comparing with GLO baselines - Quantitative results for test set reconstruction in PSNR at 256^2 resolution.

Dataset	PSNR (↑)		
	LoLNeRF [61]	TEGLO Stage-1	TEGLO
AFHQv2-Cats [35, 17]	24.94	29.26	**87.38**

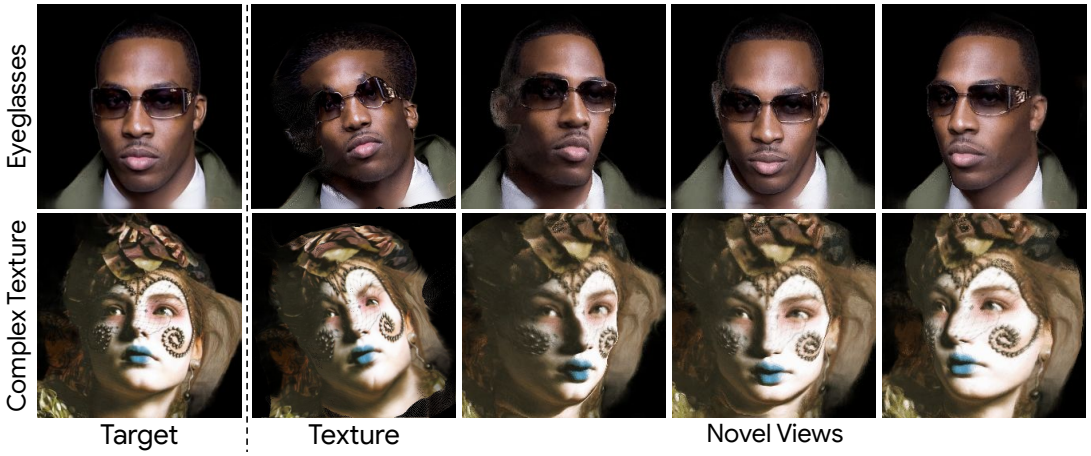


Figure 2.10. Results for complex texture and geometry - Qualitative results for texture representation and novel view synthesis with complex image samples. Compare the results with with Fig.(24) in [10].

Pivotal Tuning Inversion (PTI) [63] for the single-view image. We observe missing details in the results from LoLNeRF [61], EG3D-Fit [11] and TEGLO stage-1 in terms of jewelry, skin wrinkles, eyeglass opacity, eyeglass frame, hair strand etc. As expected, results from TEGLO and TEGLO-3DP (where TEGLO Stage-2 is trained with only surface point supervision) preserve high frequency details missed by baselines methods, demonstrating near perfect reconstruction. In Fig.(2.10), we show qualitative results with the texture (t_O) for complex appearance and geometry such as 3D consistent eyeglasses and make-up. Qualitative results for Cars are depicted in Fig.(2.19) and for AFHQ-Cats in Fig.(2.18).

2.4.3 3D Consistent Novel View Synthesis

To evaluate multi-view consistent synthesis, we report quantitative results for novel view reconstruction on the multi-view SRN-Cars data in Table.(2.4). We observe that TEGLO attains near-perfect reconstruction of test data with 67.5 dB PSNR whereas baselines achieve 30.4 dB

Table 2.4. Novel view reconstruction - Quantitative results for novel view reconstruction on the SRN-Cars dataset [13] at 256^2 resolution to evaluate 3D consistent novel view synthesis. (LoLNeRF result is from the "Concatenation" baseline in ABC [62]).

Dataset	PSNR (↑)			
	LoLNeRF	ABC	TEGLO Stage-1	TEGLO
SRN-Cars [14, 13]	25.80	29.10	30.48	**67.52**

Table 2.5. Comparing with 3D generative baselines - Test data reconstruction with previous state-of-the-art methods.

Method	PSNR (↑)	LPIPS (↓)	SSIM (↑)	ID (↑)	3D Consistency (↑)
EG3D-PTI	26.64	0.323	0.879	0.465	21.20
RealTime-RF [77]	22.29*	0.269	0.665	0.542	-
IDE-3D [73]	26.45	0.273	0.878	0.671	20.69
HFGI3D [88]	29.43	0.172	0.918	0.744	21.69
TEGLO	**84.90**	**2.1e-6**	**0.999**	**0.883**	**33.47**

*From the Supplementary Material of [77].

PSNR. We evaluate the identity consistency across multiple synthesised views using the ID score metric by computing the mean of the MagFace [49] cosine similarity scores from a sampled camera pose. We compare the ID score for TEGLO with other recent 3D GANs and observe that TEGLO outperforms the baselines with a score of 0.883. We also use the 3D consistency metric from [28] to compare the multi-view consistent synthesis of TEGLO with 3D GAN baselines. In brief, we synthesize five novel views near an input camera pose and use IBRNet [82] to predict the input image and then compute the reconstruction PSNR. We report the 3D consistency metric in Table.(2.5) and observe that TEGLO outperforms the 3D GAN methods. [88] notes that "quantitative evaluation of 3D consistency is still an open question" and since 3D consistency in novel view synthesis is better viewed as videos, we urge the reader to refer to the Supplemental Material for this thesis. Overall, we observe that TEGLO attains near-perfect reconstruction of test data attaining $\geq$ 67.5 dB PSNR, $\leq$ 6.9e-5 for LPIPS, $\geq$ 0.999 for SSIM and 33.4 for 3D consistency.

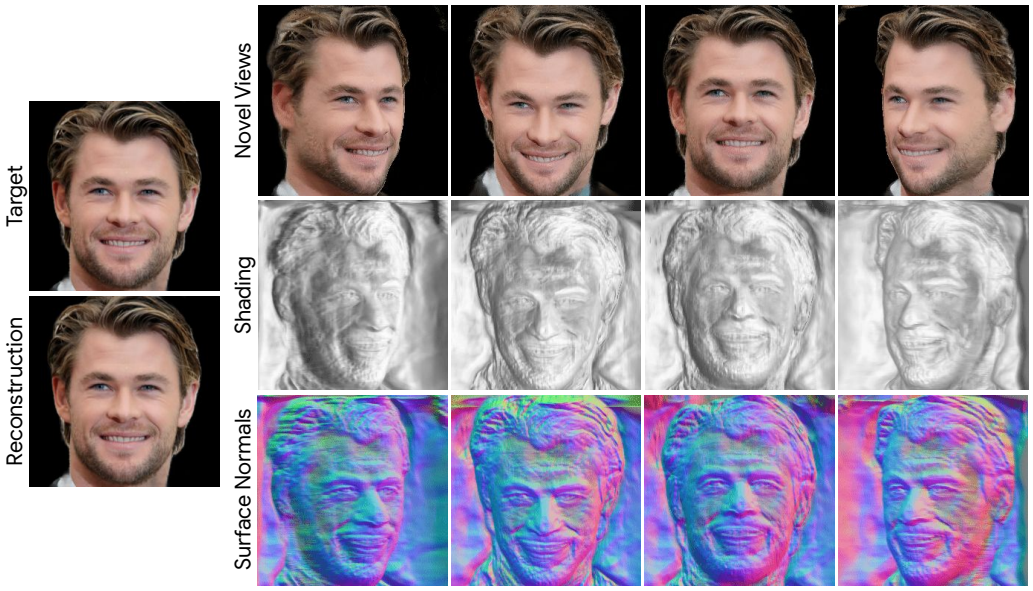


Figure 2.11. Single view 3D reconstruction - 3D reconstruction of test image from CelebA-HQ. Compare with Fig.(25) in [39]).

2.4.4 Single-view 3D Reconstruction

It is the task of representing an in-the-wild or out-of-distribution image using a trained network. Qualitative results for a held-out sample from the CelebA-HQ dataset for pre-trained TEGLO is in Fig.(2.11). Previous work such as AUVNet [15] require additional training of a ResNet-18 [30] for the image encoder and IM-Net [16] for the shape decoder followed by ray marching to obtain the mesh to represent the image while methods such as EG3D [11] require PTI (Pivotal Tuning Inversion [63]) fine-tuning to represent the image. For single-view textured 3D representation in TEGLO, we simply invert the image into the latent with no fine-tuning.

For single view 3D reconstruction and inference, we randomly sample 1K images from the training set and render five views to train TEGLO Stage-2. To evaluate on the test data, we invert the image by optimizing its latent for 200 steps while keeping the network parameters frozen. Then, we render five camera views and back-project to obtain the surface points. We then map the surface points to the canonical coordinate space to register the predicted pixels for those surface points. Similarly, we also map the 3D surface points from the GT camera pose to the canonical space and register GT pixels for those surface points.

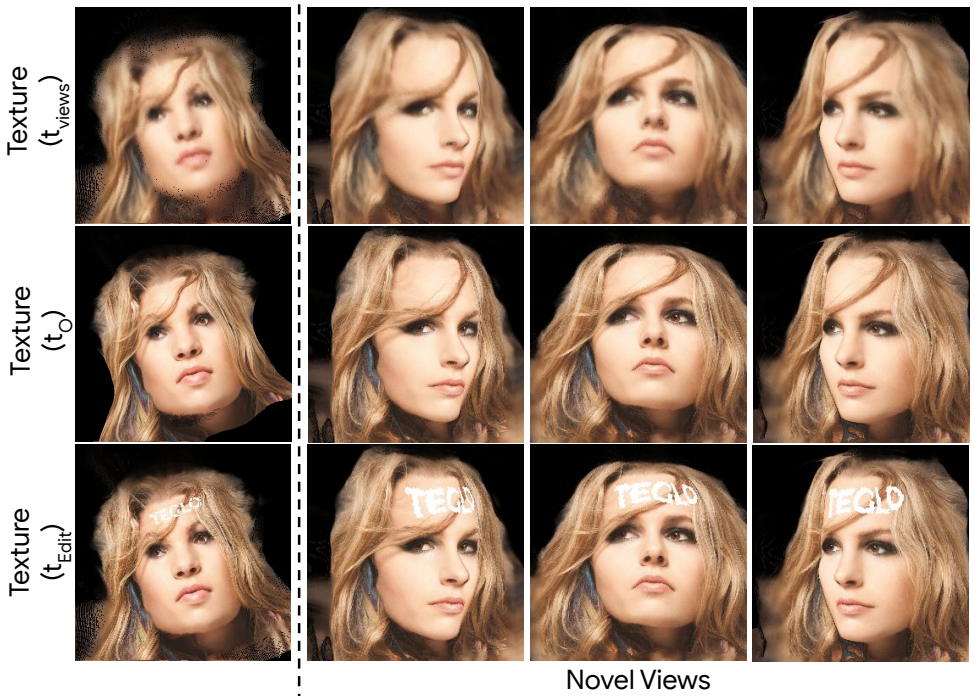


Figure 2.12. Texture editing - Qualitative results for texture edits.

Reconstructing single-view images at arbitrary resolutions while preserving 3D consistency is highly desirable for several applications. However, EG3D [11] is limited by its camera conditioned generator to possess a "baked-in" training resolution. TEGLO does not include any camera conditioning, and as a result, it allows single-view 3D reconstruction and novel view synthesis at arbitrary resolutions without any re-training for different resolutions.

2.4.5 Texture Editing

In Sec.(2.3.3), we describe the procedure to edit textures. Qualitative results with texture editing for CelebA-HQ is in Fig.(2.12) and for AFHQv2-Cats and ShapeNet-Cars in Fig.(2.1). Our edits are class-specific and target image agnostic because edits are performed in the canonical space. Previous work, NeuMesh [90] requires spatial-aware fine-tuning and mesh guided texture editing for precise transfer. However, TEGLO simply maps a texture edit image of the same size as the texture into the K-d tree with an overlay of the pixels (obtaining t_{Edit}) - precisely transferring the edit without requiring any optimization strategies.

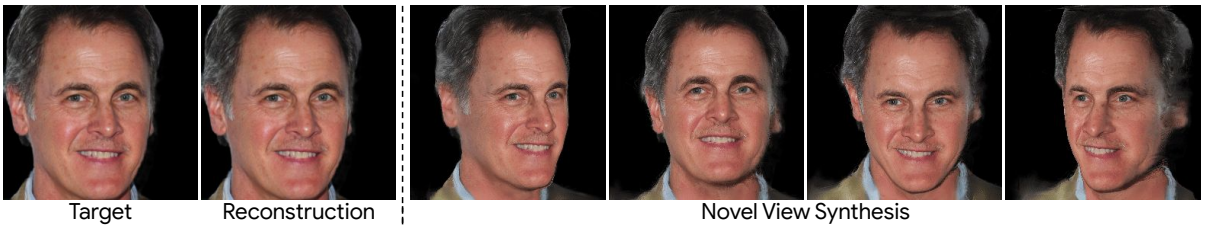


Figure 2.13. Ablation - Results for TEGLO Stage-2 trained with one arbitrary camera pose.

2.4.6 Texture Transfer

As discussed in Sec.(2.3.3), the extracted textures are aligned in a canonical coordinate space allowing texture transfer across geometries. We demonstrate texture transfer in Fig.(2.14(a)). Here, row-1 represents the target image from CelebA-HQ for the geometry learned by TEGLO Stage-1, and column-1 represents the textures (stored in a K-d tree) extracted after TEGLO Stage-2. We observe realistic texture transfer despite arbitrary camera biases in rendering $\mathscr{D}$ which are mitigated by using the K-d tree and NNI. To test if TEGLO is restricted to the range of the five arbitrary views chosen for Stage-2, we show large angle view results for Stage-2 trained with just a single view instead of five in Fig.(2.13) to validate our hypothesis. Fig.(2.14(b)), shows the keypoint correspondences mapped to the canonical coordinate space across different face identities. Since the keypoints from different identities map to the same location in the canonical space, the effectiveness of the correspondences for texture transfer is demonstrated.

2.4.7 Efficient Rendering at High Resolutions

For 3D high-fidelity data rendering from single-view image collections of objects, TEGLO enables arbitrary resolution synthesis. Since the dense correspondences are learned point-wise (using p_j), there is no spatial constraint in querying $\mathscr{M}$ for the canonical coordinate point. Hence, we render images of any size by first dividing the image pixels into 4 tiles, then obtain the surface points from TEGLO Stage-1, map the surface points to the canonical coordinate space using $\mathscr{M}$ and then index into the texture to obtain the RGB color value. Then,

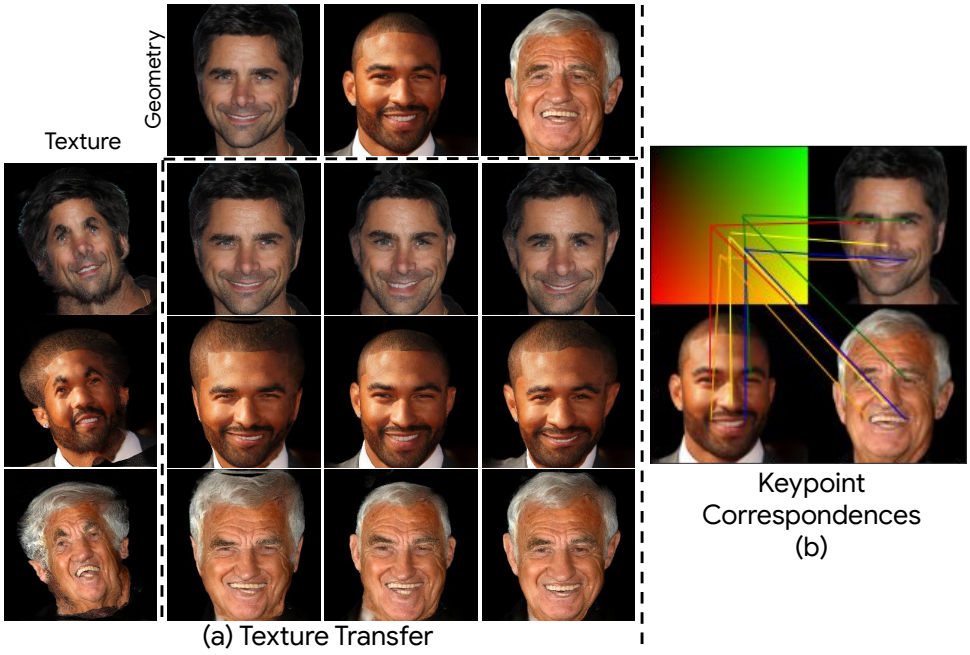


Figure 2.14. Texture transfer (a) Qualitative results for texture transfer with CelebA-HQ. (Top row shows CelebA-HQ image targets). (b) Keypoint correspondences in the canonical space.

after all the tiles are computed, we can combine the divided computations into a single image of high-resolution. The orbit video files in the thesis Supplemental Material and the high-resolution frames at 1024^2 resolution in Fig.(2.17) have been rendered using this approach.

2.4.8 Ablation Experiments

Using $\mathscr{L}_{\text{Camera}}$ loss. EG3D [11] conditions the generator and discriminator with the camera pose to enable 3D consistent novel view synthesis. As noted in the main paper, the pose-conditioned generator does not completely disentangle the camera pose from appearance leading to artifacts such as facial expressions/eyes following the camera. In TEGLO Stage-1, we use the $\mathscr{L}_{\text{Camera}}, \mathscr{L}_{\text{RGB}}$ and $\mathscr{L}_{\text{Perceptual}}$ losses to train $\mathscr{N}$. An ablation experiment without the camera prediction loss led to 2D banner artifacts. This is qualitatively represented in Fig.(2.15) for "Views without $\mathscr{L}_{\text{Camera}}$" with flat and inconsistent geometries for different camera angles. However, the results for training $\mathscr{N}$ using $\mathscr{L}_{\text{Camera}}$ show multi-view consistent representations demonstrating the effectiveness of using the simple camera prediction loss. Furthermore, we

Figure 2.15. Ablation - $\mathscr{L}_{\text{Camera}}$ for TEGLO Stage-1 training.

show that the rendered orbits do not have expressions/eyes following the camera in Supplemental video for this thesis.

TEGLO Stage-2 with $\mathscr{L}_{\text{Coord}}$ loss only. Previous work AUV-Net [15] states that methods [24, 45] that do not use color for learning dense correspondences may learn sub-par texture representations. To verify, we train TEGLO Stage-2 with only $\mathscr{L}_{\text{Coord}}$ reconstruction loss instead of $\mathscr{L}_{\text{Stage2}} = \mathscr{L}_{\text{Coord}} + \mathscr{L}_{\text{Coord}} + \mathscr{L}_{\text{Coord}}$ reconstruction losses. The qualitative results are presented in Fig.(2.8) comparing TEGLO-3DP with TEGLO and other baseline results. Of particular interest is Fig.(2.16) with qualitative results for TEGLO and TEGLO-3DP including the texture image. We note that the reconstruction and novel view synthesis results are nearly identical. However, we also observe TEGLO-3DP including a wayward texture representation near the hair region. While the dense correspondences map the surface points to the appropriate RGB image pixels, there is a scope for null pixel artifacts around the hair region when using NNI. While the 3D reconstruction and novel view synthesis for TEGLO-3DP and TEGLO do not differ, we note the potential for black pixels to be obtained in novel view synthesis leading to lowered qualitative and quantitative results.

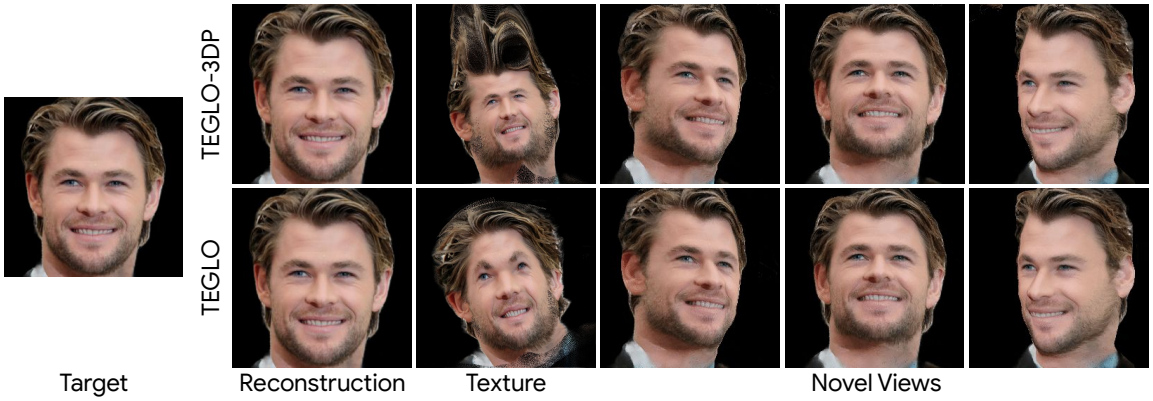


Figure 2.16. Ablation - Qualitative comparison with TEGLO-3DP (TEGLO Stage-2 with 3D surface point reconstruction only)

2.5 Discussion and Limitations

While TEGLO enables near perfect 3D reconstruction of objects from single-view image collections, it requires multi-stage training. We hope that future work can simplify the framework with an elegant end-to-end formulation. A potential next step would be to use StyleGANv2 [36] to generate high quality textures for texture transfer and editing. TEGLO could enable 3D full-body avatars from single views with high frequency details extending methods such as PIFu [66]. Future work could explore representing light stage data across different camera angles in an illumination invariant manner using 3D surface points. One limitation of our method is that the texture does not include ground truth pixels from the obstructed parts of the object. We hope future work can address this limitation. Further, TEGLO is only able to map target image pixels spanning the target image and hence there may be artifacts for camera views with minimal mapped target image pixels. For example, the novel view in row-2, column-3 in Fig.(2.10), the novel view shows a slight twist in the nose geometry partially due to the thin veil on the face which could not be accounted for in Stage-1.

2.6 Conclusion and Broader Impact

In this work, we present TEGLO for high-fidelity canonical texture mapping from single-view images enabling textured 3D representations from class-specific single-view image

Figure 2.17. High resolution rendered views - Qualitative results for novel views in high-resolution with high-frequency details such as freckles, jewelry, make-up, hair, fine details and wrinkles.

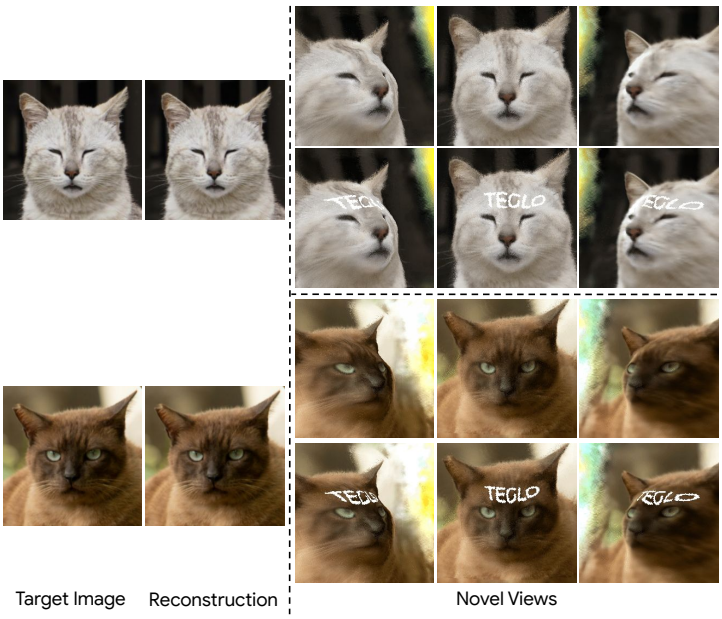


Figure 2.18. Textured synthesis - Target view reconstruction and novel view synthesis for AFHQv2-Cats.

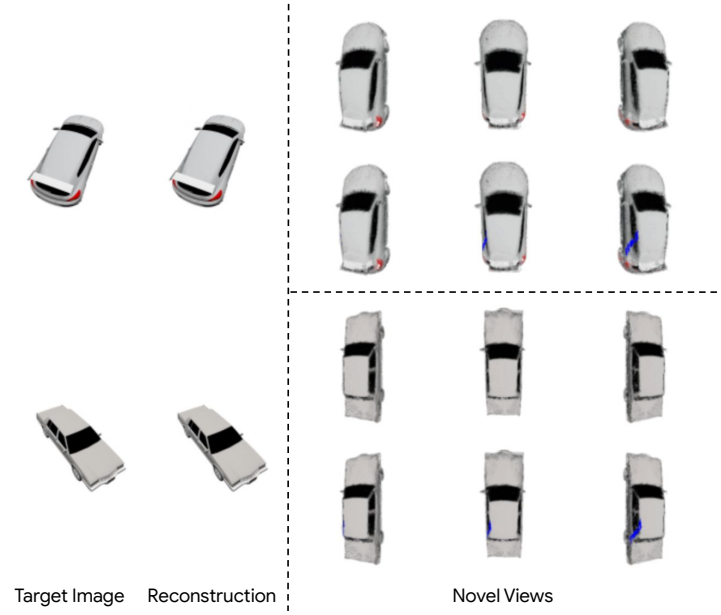


Figure 2.19. Textured synthesis - Target view reconstruction and novel view synthesis for SRN-Cars.

collections. TEGLO consists of a conditional NeRF and a dense correspondence learning network that enable texture editing and texture transfer. We show that by effectively mapping the input image pixels onto the texture, we can achieve near perfect reconstruction ($\geq$ 74 dB PSNR at 1024^2 resolution). TEGLO also allows single-view 3D reconstruction by simply inverting the single-view image into the latent table without requiring any PTI or fine-tuning.

Broader impact. One of the motivating goals for TEGLO stems from the need for photorealistic 3D reconstruction of objects from single-view image collections. As an example, [15] use Tripleganger heads [4] - a dataset containing 515 3D meshes faces at a high cost-per-scan requiring a custom commercial license for use. Similarly, [91] is a dataset of 938 textured meshes of heads made available at no cost. However, the authors allude to the demographic bias in compiling the data, the 68 DSLR camera setup, and the six month effort involved in dataset capture - all of which do not scale and has a high potential for bias and privacy issues. TEGLO enables high-fidelity 3D reconstruction and novel view synthesis from single-view image collections which alleviates these issues and also improves access to high quality data to the broader research community. Hence, TEGLO enables rendering a dataset of diverse objects (improving fairness and mitigating bias) and also reduces the need for large scale data collection (alleviating privacy issues).

Chapter 2, in full, has been submitted for publication of the material as it may appear in a conference, 2024, Vishal Vinod, Tanmay Shah, Dmitry Lagun. The thesis author was the co-primary investigator and author of this paper.

Chapter 3

Hybrid Physical-Neural Approach to Surface Relighting

3.1 Introduction

Relighting objects with complex materials is a challenging task. Learning a relightable implicit neural rendering model for human faces, modeling both direct illumination effects such as specular highlights and indirect illumination effects such as subsurface scattering have limitations based on a radiance transfer formulation. OSFs [94] approximate the cumulative radiance transfer for an object based on the assumption of an unobstructed distant light source. An OSF formulation abstracts the subsurface scattering process as cumulative radiance transfer, eliminating the dual integral required by a BSSRDF formulation, making it feasible for real-time volume rendering. However, OSFs have limitations that make it infeasible to model complex objects such as human faces: the requirement of OLAT (one-light-at-a-time) light stage data with camera poses, assumption of unobstructed distant lighting which is impractical for large objects, and the inability to model specular effects. Similarly, [100] propose an implicit neural rendering model based on approximating the radiance transfer gradient to model subsurface scattering effects, however, the formulation has the same limitations as OSFs: requiring light stage data capture and the inability to model specular highlights. As expected, a simple experiment where a pre-trained OSF is kept frozen and augmented with a light-weight neural renderer shows that the residual specular highlights missed by the OSF can be accounted for. We include the augmented

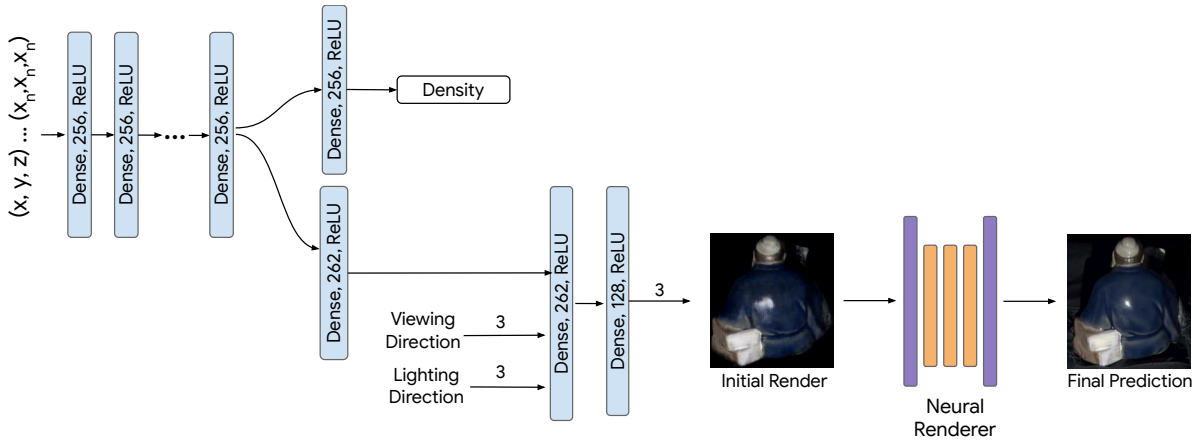


Figure 3.1. OSF + Neural Renderer - Architecture of a simple extension to a trained OSF model with a light-weight neural renderer to model the residual specular highlights unable to be modeled by the OSF.

OSF in Fig.(3.1) and results for the OSF + NR in Fig.(3.2) for reference on the Reading image collection from the DILIGENT-MV [40] dataset where the statue involves significant specular highlight effects.

The process of capturing OLAT (one-light-at-a-time) light stage data with camera poses for human faces is laborious, expensive to set up and is time intensive. Acquiring a large dataset of high resolution face scans is prohibitively expensive and requires a custom licensing for use - all of these factors limit access to the research community. Toward this, we explore a growing trend in computer vision and computer graphics research: the generation of synthetic data for human faces. Furthermore, the Illinois Biometric Information Privacy Act (BIPA) [55] act protects an individual's biometric information, including retina or iris scan, fingerprint, voiceprint, or scan of hand or face geometry, by requiring corporations to make the user aware in writing about the purpose of collecting the data and the intended period of retention [20, 21, 80, 81, 19]. With recent advancements in computer vision, particularly Physically Based Rendering (PBR) and Generative AI, photo-realistic rendering of human faces is possible with ground truth for each components of light transport including the specular maps, subsurface scattering and other indirect illumination effects. In this work, we explore synthetic data rendering for high fidelity face relighting. Using synthetic data for human faces enables privacy preserving training

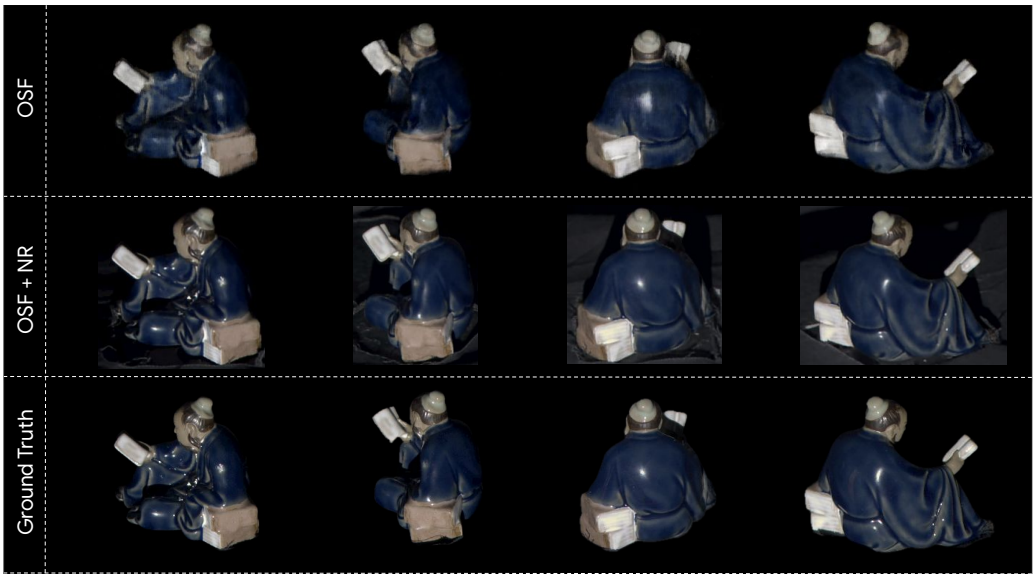
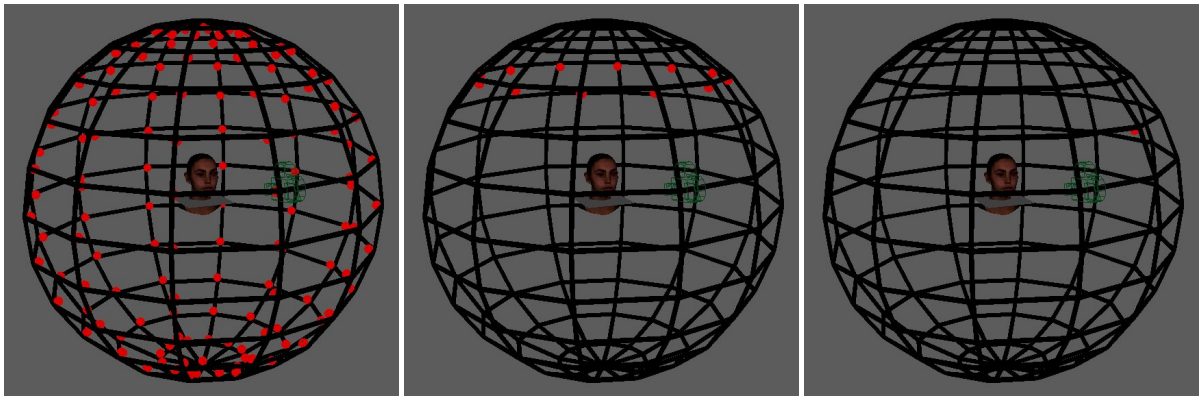


Figure 3.2. Qualitative Results for OSF + NR - Comparing the results from OSF and OSF + NR with the ground truth.

of face recognition and face relighting methods and does not violate BIPA or related policies. Further, high fidelity and photo-realistic generation capabilities in current physics based rendering methods enables generation of large scale data that can be used for photorealistic relighting [92]. Some advantages include: rendering a dataset with multiple objects useful for telepresence applications, developing and using a virtual light stage for data collection (Refer fig.(3.3)) and the ability to obtain ground truth data for physics-based light transport. While there are disadvantages related to scale, obtaining sufficient number of face meshes with displacement maps and the photorealism of the rendered data and the potential domain shift, we seek to attempt this problem by rendering a photo-realistic dataset and train a hybrid physical-neural surface relighting method for high fidelity relighting. If the method demonstrates significant domain shift with respect to real world light stage data (without suitable ground truth) such as [97] or the Imperial Light-Stage Head (ILSH) Dataset [1], we will augment our method with a domain adaptation [75, 79] based training pipeline.

In this work, we consider the problem of reconstructing and relighting human faces from a single input image. Current methods either consider relighting to be a purely physically-based



(a) All lights (b) Top Row (c) OLAT (one-light-at-a-time)

Figure 3.3. Virtual Light Stage - A light stage setup in Arnold Maya with 166 light sources. We use the face mesh of Emily [5]. (a) all 166 light sources turned on; (b) ring lighting sequence. (c) OLAT setup.

approach using reconstruction followed by physically-based rendering [22], or with a purely data-driven approach [57, 48]. We propose a hybrid physical-neural approach to utilize the rich dependencies from a physics-based prior from physically based rendering and from neural rendering. Current methods model the reflectance as diffuse Lambertian or as a simplified BRDF to account for specular properties. However, they fail to consider subsurface scattering of skin or the inter-reflections and volumetric scattering in indoor scenes. This work aims to perform high-quality human face relighting by modeling long-range light interactions such as subsurface scattering. We aim to achieve consistency among the different factors of image formation - geometry, material and lighting. Hence, a neural differentiable renderer that effectively utilizes the learned priors can enable photorealistic relighting with challenging light transport effects such as subsurface scattering and soft shadows.

3.2 Synthetic Dataset Rendering

3.2.1 Related Work

Synthetic 3D faces. Microsoft's "Fake it till you make it" [85] discusses the use of synthetic data for face recognition and other related computer vision tasks. The authors describe a method for rendering 3D face models with high levels of realism and diversity, which can

be used to train machine learning systems. However, there is a clear dataset shift between the synthetic rendering to real world data. The authors argue that synthetic data can match real data and demonstrate the power of domain randomization and careful design choices such as meso-displacement and photo-realistic add-on edits. The authors note that while it requires considerable expertise and investment to develop a synthetic face data generation framework with minimal domain gap, it becomes possible to generate a wide variety of training data with minimal incremental effort. A follow-up work DigiFace-1M [6] presents a large-scale synthetic dataset for face recognition consisting of one million digital face images rendered using a computer graphics pipeline. The authors demonstrate that aggressive data augmentation aka domain randomization can significantly reduce the synthetic-to-real domain gap and show how each attribute affects accuracy. By fine-tuning the network on a smaller number of real face images obtained with user consent (which is a reasonable assumption given that it is necessary for testing and test time adaptation techniques), they achieve accuracy comparable to methods trained on millions of real face images. The authors rightly argue that large-scale web-crawled face recognition datasets including a common benchmark dataset FFHQ [36] raise ethical concerns including privacy and bias issues. An interesting recent method Rodin [83] presents a 3D generative model that uses roll-out diffusion to automatically generate high-fidelity 3D digital avatars represented as NeRFs. Rodin represents a NeRF as multiple 2D feature maps and rolls out these maps into a single 2D feature plane within which 3D-aware diffusion is performed. Rodin is computationally efficient and enables high-fidelity 3D diffusion.Rodin's use of latent conditioning for feature generation for global coherence enables high-fidelity avatars and with semantic editing based from text prompts. Stable Diffusion v5, based on latent diffusion models [64] enables photo-realistic 2D face image generation which was not possible in previous iterations paving way for realistic 2D image generation. While these methods enable rendering realistic human faces with controllable geometry, they do not allow controllable OLAT light stage rendering especially with ground truth for individual components of light transport necessary to train a high-fidelity relighting model.

3D from single views. EG3D [11] discusses a new approach to generating high-quality

multi-view consistent images using unsupervised learning from single-view 2D image collections. The authors introduce a hybrid explicit-implicit network architecture that synthesizes high-resolution, multi-view-consistent images in real-time while producing high-quality 3D geometry. They achieve this by decoupling feature generation and neural rendering, which allows them to leverage state-of-the-art 2D CNN generators for efficiency and expressiveness. However EG3D includes image space approximations that breaks multi-view consistency and is constrained by the tri-plane capacity and is prone to camera-pose biases (such as the eye following the camera) due to the use of camera pose to condition the generator. EpiGRAF [71] aims to address the shortcomings in EG3D by proposing a location-and scale-aware discriminator to work on patches of different sizes and spatial positions. EpiGRAF uses a patch sampling strategy based on an annealed beta distribution to stabilize training and accelerate the convergence and hence enables high resolution synthesis with high fidelity geometry. Further, it includes a novel hypernetwork-modulated discriminator architecture to operate on patches with continuously varying scales - resulting in an efficient, high-resolution, pure 3D generator. While EG3D and EpiGRAF enable 3D from single view, they are not suitable for reconstructing the identities from the input 2D image data collections as they require fine-tuning for each image or Pivotal Tuning Inversion (PTI) [63]. A recent work, LoLNeRF [61] aims to alleviate this issue and enables arbitrary resolution image synthesis by using a Generative Latent Optimization (GLO) [9] based auto-decoder training regime. LoLNeRF enables future work to draw on insights from adversarial approaches to further improve high fidelity 3D reconstruction from single view image collections. Recent work in single view 3D face reconstruction with relighting applications such such as ShadeGAN [56], NeRFFaceLighting [33], LumiGAN [23] and FaceLit [60] enable multi-view consistent relighting - however they approximate a Phong BRDF based lighting model with specular highlights being the primary lighting effect of interest. OSFs [94] and [100] are NeRF-based methods that model subsurface scattering with relighting effects, however they are limited when modeling specular highlights and indirect illumination such as soft shadows. RelightableHands [32] allows relighting hands with visibility information for soft shadows.

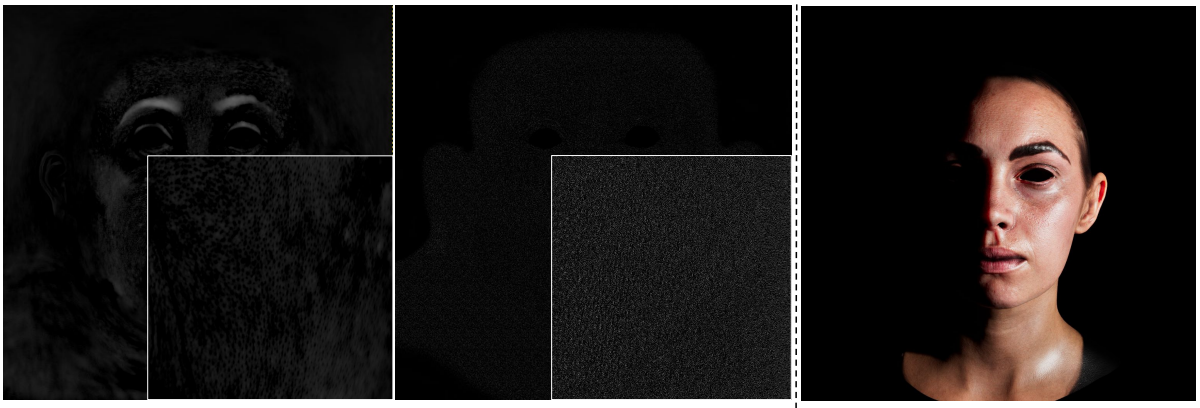


Figure 3.4. Displacement Maps for the Face Mesh - We use a coarse (left) and a micro (middle) displacement map to modify the face mesh in a realistic manner. The rendered RGB (right) uses the *aiStandardSurface* shader from Arnold and includes pore level details in the rendered RGB.

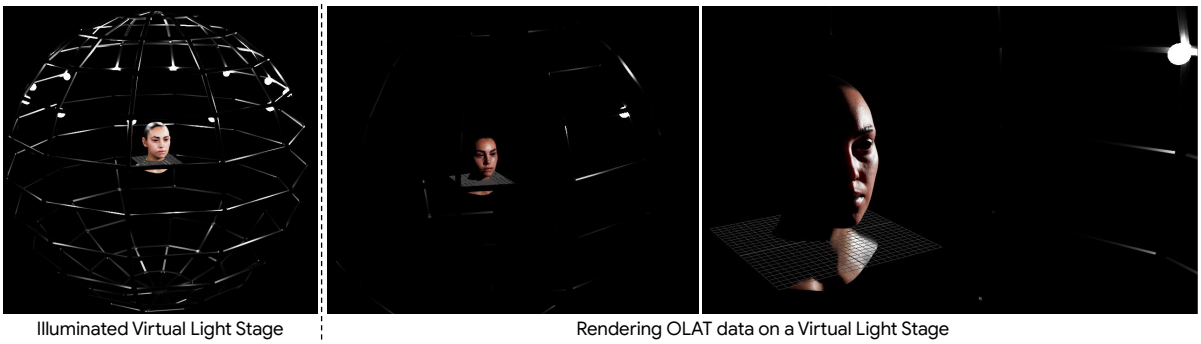


Figure 3.5. Rendering Synthetic Data in the Light Stage - (Left) we show a fully illuminated virtual light stage in Arnold Maya. (Right) we show an OLAT setup and a render from one camera viewing angle. The light stage rig structure and lighting sequence have been adapted from [2].

While the above methods and their variations including allow multi-view consistent generation with ground truth for geometry and albedo, they are not able to provide the necessary lighting components for relighting with a specific focus on subsurface scattering effects. All of these factors motivate the need for a synthetic OLAT light stage dataset with suitable ground truth for the direct and indirect components of light transport.

3.2.2 Dataset Rendering

In this work, our dataset rendering pipeline includes the use of a face mesh and a virtual light stage dataset capture system in Arnold [26] Maya. Arnold is a physics-based production

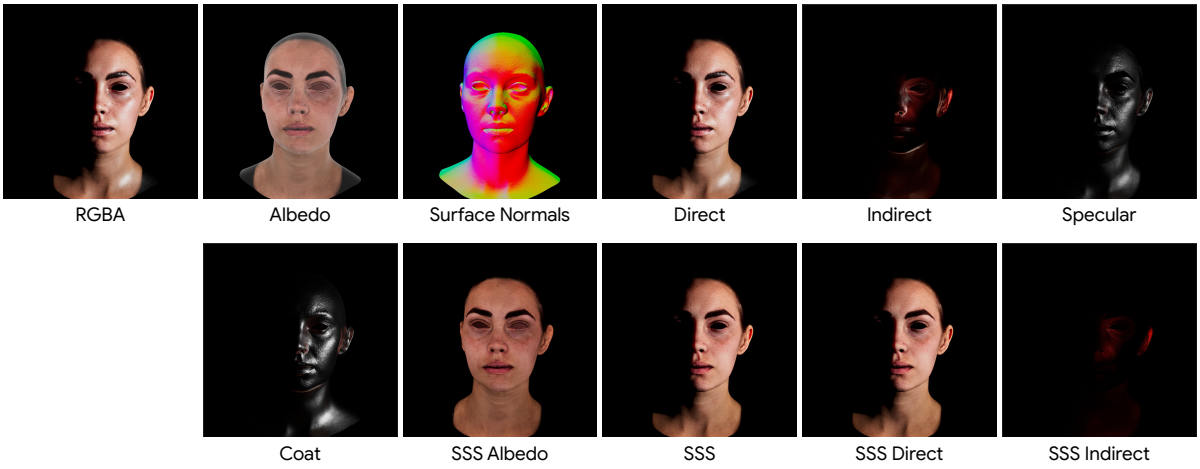


Figure 3.6. Example from the Rendered Dataset - Our rendering consists of a face under 166 different lights in an OLAT setting with multiple camera viewpoints. We aim to scale the virtual light stage to include 150 cameras for large scale experiments with multiple face meshes.

path tracer that simulates physics-based light transport with photorealistic results. We use Arnold to render our training dataset. As depicted in Fig.(3.3), we use the OLAT lighting sequence in the virtual light stage and use Arnold Maya's *aiStandardSurface*[1] to model the light transport based on the displacement maps for the face mesh (depicted in Fig.(3.4)). An interesting note in [100] is that existing datasets for modeling indirect illumination effects such as subsurface scattering include capture of translucent objects at low resolutions which affect the reconstruction fidelity as micro geometry details are not fully captured. Hence, we seek to include a micro displacement map for the face mesh to capture these details to model the subsurface scattering effects.

In Fig.(3.5), we demonstrate the rendering of the dataset in the virtual light stage with a row of lights and a perspective camera to show the full light stage rig illuminated by a ring of lights (left) and in an OLAT sequence on the right. Based on our analysis in Chapter 3 and the limitations of modeling the radiance transfer for relighting human faces, we obtain several AOVs[2] (Arbitrary Output Variables) which allows us to render arbitrary shading components into individual images which can be used as ground truth in our training pipeline. Arnold's AOVs allow us to obtain the direct and indirect lighting contributions which is essential for the

[1]https://help.autodesk.com/view/ARNOL/ENU/?guid=arnold_for_maya_am_Arnold_for_Maya_User_Guide_html

[2]https://help.autodesk.com/view/ARNOL/ENU/?guid=arnold_for_maya_render_settings_aovs_html

high-fidelity reconstruction and relighting framework we propose. Depicted in Fig.(3.6), we obtain the albedo, surface normals, direct lighting contribution, indirect contribution (notice the subsurface scattering and indirect specular contributions), specular, coat, SSS, SSS Direct (direct SSS contribution) and SSS Indirect (scattering within the skin traveling upto a certain mean free path distance). We also obtain the z-depth information that we use in training our proposed inverse rendering framework. For the experiments in this thesis, we work with the Emily face mesh and 166 light sources from multiple camera angles. The next step involves procuring a set of face meshes comprising diverse skin tones and textures to make our dataset more representative for real-world use cases.

3.3 PB-NSR

Given an OLAT dataset of human faces consisting of the RGBA images under varying lighting and camera viewing angles, albedo, surface normals, depth map, direct lighting component and the subsurface scattering component, we aim to reconstruct the geometry, material and lighting and relight the face from a different lighting direction. Our proposed method consists of two stages: In Stage-1 (Fig.(3.7), we perform joint neural reconstruction and relighting with a single decoder multi-decoder framework. The encoder takes in the input image and maps it to a lower dimensionality to be used by the inverse decoder and the relight decoder. The inverse decoder takes the encoder output and provides the albedo, surface normals and depth map as output (in experiments we also work with coat and specular maps). The relight decoder takes the output from the encoder, a new lighting direction and skip connections from the inverse decoder to predict the direct lighting component for the new lighting direction. In Stage-2 (Fig.(3.8)), we aim to use the direct estimate from a physically based renderer as our direct lighting estimate. Note: in this thesis, we use the direct estimate from Stage-1 as a proxy inductive bias with physical cues based on the bidirectional connections used during training. The direct lighting estimate and the geometry and material output from Stage-1 are used as input to the IndirectLightingNet

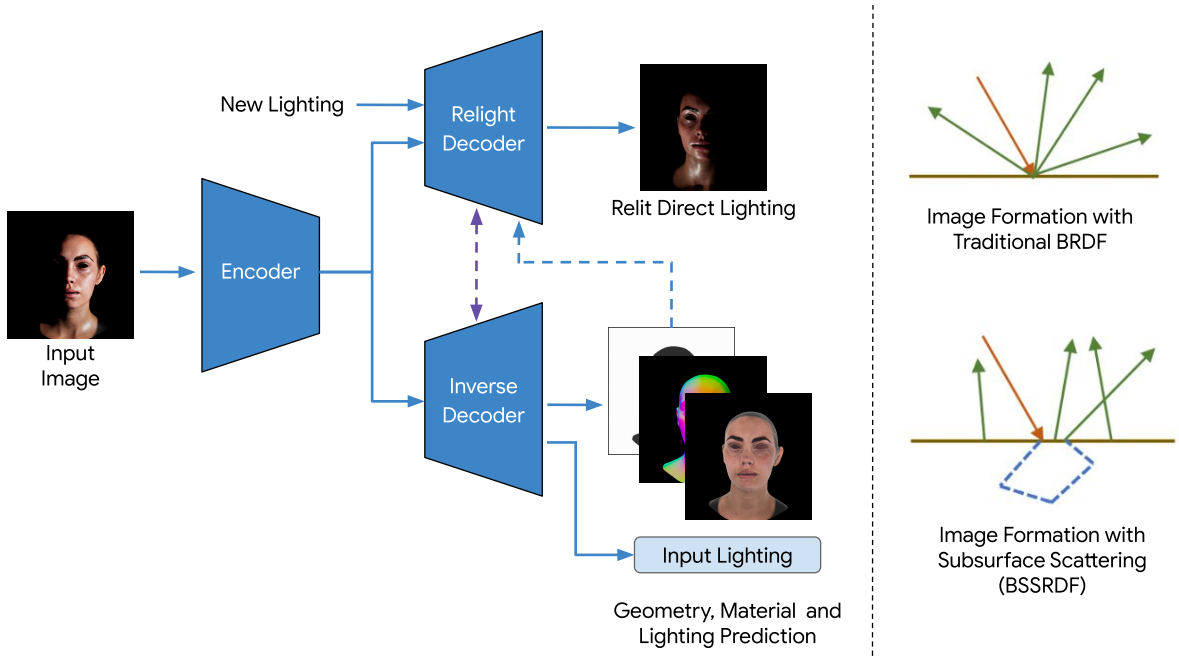


Figure 3.7. Stage-1: Joint Neural Reconstruction and Relighting - Overview of the proposed method with joint reconstruction and relighting. (Right) Comparison between traditional BRDF based light transport with complex subsurface scattering effects modeled by BSSRDF.

to predict the subsurface scattering component. Then, a neural renderer takes the direct lighting component and the indirect lighting as input to predict the final relit output.

3.3.1 Joint Neural Reconstruction and Relighting

Formulation. We denote the OLAT image dataset ($\mathscr{I}$) with several components of shape, material and light transport: RGBA rendered image (r_i), depth map (d_i), surface normal map (s_i), albedo map (a_i), direct lighting estimate ($\mathfrak{d}_i$), the subsurface scattering component (sss_i) and the lighting vector l_i. Hence, we have $\forall i = 0...n$, $\{r_0, d_i, s_i, a_i, \mathfrak{d}_i, \text{sss}_i, l_i\} \in \mathscr{I}$.

Network architecture. We derive inspiration from indoor inverse rendering [42] to develop a single encoder multi-decoder framework with the aim to predict the relit direct estimate that is physically meaningful while preserving input image detail. The encoder takes in the input image r_i based on lighting vector l_i and predicts an intermediate latent representation to be used by the decoders. The first decoder is the inverse decoder that predicts the albedo (a_i), surface normal (s_i), depth map (d_i) and the lighting vector l_i. The relight decoder takes the intermediate

representation from the encoder, a new lighting vector (l_j) and the features of the inverse decoder at each decoder upsampling level as input to predict the direct lighting component for the new lighting vector l_j. Jointly training the inverse and relighting decoders with the bidirectional connection incorporates the inductive bias of appearance variations under varying lighting directions which can help improve the accuracy and the generalization when dealing with complex light transport effects such as subsurface scattering in human faces. We use skip connections between the inverse decoder and the relight decoder to mimic physics-based rendering through the interaction of shape and material parameters with incident illumination. This enables the encoder latent space to model appearance variations for shape and materials under different illumination conditions. Compared to methods that use in-network rendering layers without trainable parameters, our network with bidirectional connections provides inductive biases for shape and material estimation as well as physics based cues for relighting.

In this work, we use the output of Stage-1 as the physically-based direct lighting estimate. For further investigation, we seek to gain further insights from physically-based rendering [59] to compute the Monte Carlo estimate for the direct shading at point $\mathbf{p}$ as:

$$\mathbf{E}_j(\mathbf{p}) = \frac{\text{area}(\mathbf{j})}{N_j} \sum_q \frac{\mathbf{L}_j(\mathbf{p} \rightarrow \mathbf{q}) \max(\cos\theta_{\mathbf{p}} \cos\theta_{\mathbf{q}}, 0)}{\|\mathbf{q} - \mathbf{p}\|_2^2} \tag{3.1}$$

where $\mathbf{p} \rightarrow \mathbf{q}$ is the unit vector from $\mathbf{p}$ to $\mathbf{q}$ on the light source. Toward this, we model the light sources from our virtual light stage as a lamp with a bounding box each predicted by our inverse decoder. To ensure computational tractability, we use the physics-based renderer to predict the direct lighting estimate. We upgrade the direct lighting prediction to include subsurface scattering effects by using a neural renderer to approximate the indirect lighting component. This is a valid next step since our rendered dataset includes ground truth for indirect illumination - specifically the subsurface scattering component. While physically based rendering using OptiX is the proposed approach to include further insights from PBR, we also require a large scale dataset with ground truth for invisible lamp polygons to train the PBR layer. This

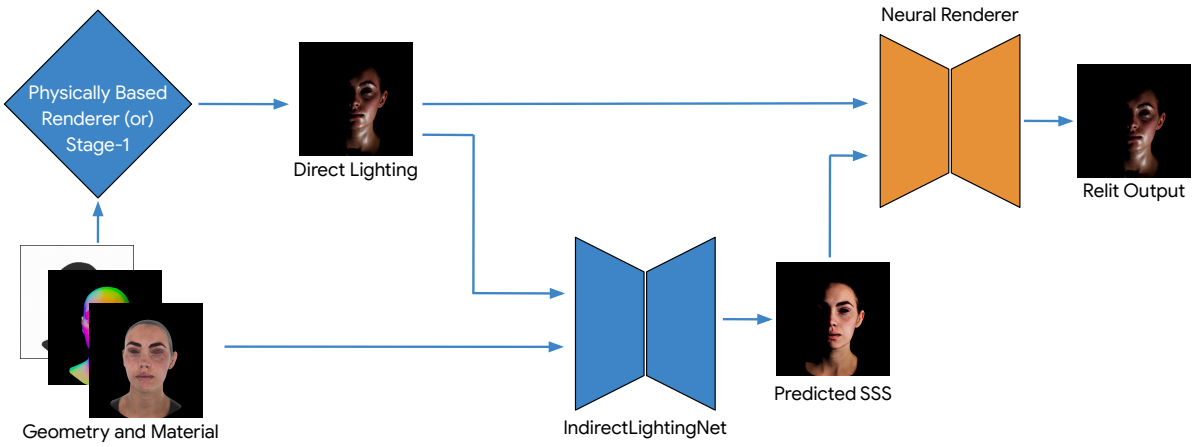


Figure 3.8. Stage-2: Neural Rendering for Indirect Lighting - Overview of the proposed hybrid physical and neural rendering method to consider various components of image formation to relight human faces.

process is in the works and the thesis includes results for the PB-NSR Stage-1 predictions for the direct lighting component.

3.3.2 Neural Rendering for Indirect Lighting

Network architecture. To train PB-NSR Stage-2, we use an IndirectLightingNet based on a standard U-Net [65] architecture to predict the subsurface scattering component ($\widehat{\text{sss}}_i$). The light weights neural renderer is also based on a U-Net but with just 4 blocks and with $c/2$ channels at each layer compared to the original U-Net. The neural renderer takes in the predicted direct lighting estimate ($\widehat{\mathfrak{d}}_i$) and the subsurface scattering component ($\widehat{\text{sss}}_i$) to predict the final lighting estimate. The formulation for the neural renderer have been used by related work such as TotalRelighting [57] and Lumos [92].

Losses. To train Stage-2, we use the direct lighting estimate ($\widehat{\mathfrak{d}}_j$) and the geometry and material (d_i, a_i, s_i) as input to the IndirectLightingNet to predict the subsurface scattering component. From our experiments, we note that predicting the subsurface scattering component (inclusive of both direct and indirect subsurface scattering) and then including a light-weight neural renderer to learn to combine the direct estimate and the subsurface scattering component provided better results than using the IndirectLightingNet to predict the indirect lighting compo-

nent (from Fig.(3.6)) and combining it with the direct estimate. This is primarily due to the minor contributions of the sss_indirect component leading to several images ($\sim$ 35 of 166 lighting directions) with very minor or nearly imperceptible contribution from sss_indirect. We postulate that using the sss_indirect component as the output from IndirectLightingNet will be possible when training with a large scale rendered dataset. As depicted in Fig.(3.8), the light-weight neural renderer takes the direct lighting estimate ($\widehat{\mathfrak{d}}_j$) and the sss prediction (sss_i) as input and predicts the final relit output.

3.3.3 Experiments and Results

We train PB-NSR Stage-1 for 2500 epochs on our dataset of synthetically rendered dataset of Emily's face mesh. We use the Adam [38] optimizer with a learning rate of $1e-4$ and a StepLR schedule reducing the learning rate every 2000 steps. We use the L_1 reconstruction loss with the depth (L_d), the albedo (L_a), the surface normals (L_s) and the direct lighting estimate ($L_{\mathfrak{d}}$) to supervise the training. We also use the LPIPS loss for the direct lighting estimate to improve photo-realism of the prediction (L_{LPIPS}). Our combined loss to train PB-NSR Stage-1 is:

$$L_{\text{Stage-1}} = L_d + L_a + L_s + L_{\mathfrak{d}} + L_{\text{LPIPS}} \tag{3.2}$$

Our Stage-1 training attains an average PSNR of 26.7 dB for the direct lighting estimate.

We train PB-NSR Stage-2 for 3000 epochs on our dataset of synthetically rendered OLAT light stage data. To train the IndirectLightingNet in Stage-2, we use the L_1 reconstruction loss and the LPIPS loss to supervise the subsurface scattering prediction. Similarly, the light-weight neural renderer is trained using the L_1 loss and LPIPS loss for the final relit prediction of PB-NSR Stage-2. The losses used to train PB-NSR Stage-2 are as follows:

$$L_{\text{IndirectLighting}} = L_{\text{sss}} + L_{\text{LPIPS\_sss}} \tag{3.3}$$

$$L_{\text{NR}} = L_{\text{relit}} + L_{\text{LPIPS\_relit}} \tag{3.4}$$

Figure 3.9. Qualitative Results - Qualitative results for train set reconstruction and relighting.

$$L_{\text{Stage-2}} = L_{\text{IndirectLighting}} + L_{\text{NR}} \tag{3.5}$$

The final prediction from PB-NSR Stage-2 attains an average reconstruction PSNR of 32.69 dB. We depict qualitative results for the train set reconstruction with sufficient OLAT lighting wherein the specular highlights are prominent in Fig.(3.9). We observe the direct estimate predictions include some grainy artifacts and hence includes scope for improvement by using a PBR renderer as discussed in the following section. As expected, the grainy artifacts in Stage-1 propagate to the subsurface scattering prediction in Stage-2. We also depict qualitative results for lighting vectors where subsurface scattering is more readily observable in Fig.(3.10). We observe the predictions of PB-NSR to be quite close to the ground truth with minimal graininess. By including further physics based prior from a differentiable PBR layer to Stage-2, we expect our results to improve.

3.4 Discussion and Next Steps

Our initial experiments were primarily related to a NeRF-based reconstruction modeling the radiance transfer and including a neural rendering network to model complex indirect illumination including soft shadows and the residual specular highlights missed by the OSF

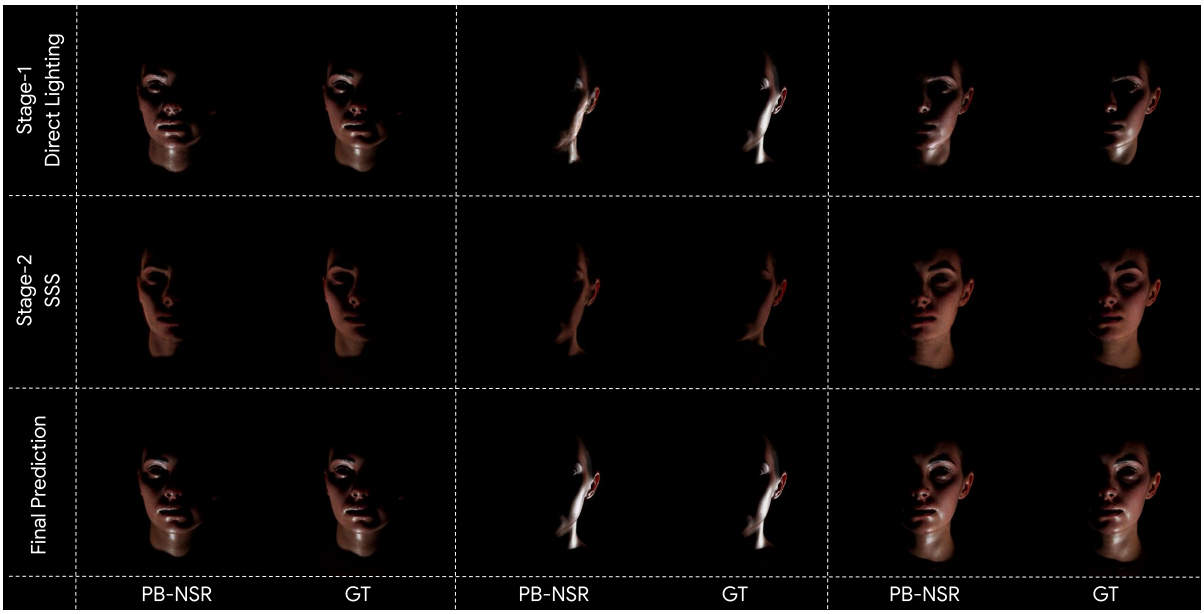


Figure 3.10. Qualitative Results - Qualitative results for direct lighting prediction from Stage-1 (row-1), subsurface scattering prediction from Stage-2 (row-2) and the final relighting prediction (row-3).

and neural radiance transfer gradient formulation. However, we were unable to access OLAT light stage data with camera poses. While some datasets such as ILSH [1] from the ICCV 2023 workshop include light stage data with camera poses useful for novel view synthesis, they lack OLAT data required to model subsurface scattering. Further, the NVPR dataset [97] dataset we accessed included only 4 identities with approval for external use and did not include camera poses. Using multi-view stereo methods such as Colmap [68, 67] fails in cases where the background lacks texture. As a result, we attempted to obtain the surface normals and depth from multi-view data. Our hypothesis was to evaluate the strength of a self-supervised (reconstruction based) method with a physically based differentiable renderer (OptiX) and a neural renderer since only a decomposition that is physically meaningful can be recombined by a physics based renderer to produce appearances that match the input and the relit ground truth image. However, our experiments were unsuccessful primarily due to the lack of a large scale light stage dataset. Initial implementation of the self-supervised hybrid neural physically based rendering without any ground truth except for light stage data diverged early - demonstrating

the need for intermediate supervision to model long range light transport effects especially in a sparse data regime. Another consideration to render our own dataset stemmed from the face that there were no other datasets with densely sampled camera views for objects demonstrating subsurface scattering effects. While [100] claim to release a dataset of 8 scenes with objects where subsurface scattering is the primary light transport effect, the dataset is not relevant to our use case of high fidelity face relighting.

Our next steps include rendering a large scale OLAT dataset of human faces with micro displacement details. This will allow us to train PB-NSR Stage-2 with the OptiX physically-based renderer to predict the direct lighting component. We postulate that further insights from PBR in addition to the physically based prior learned in Stage-1 will improve our results. Further, this will enable us to use only the sss_indirect component for Stage-2 instead of using sss along with the light-weight neural renderer. We also seek to evaluate the effectiveness of our method using the large scale synthetic dataset recently made available by [92].

3.5 Conclusion

In this work, we aim to perform high fidelity relighting of human faces with a focus on subsurface scattering effects. Initially, we explored OSF, a NeRF-based method to approximate the cumulative radiance transfer after all the light transport effects have occurred. However, we soon realized that getting access to OLAT light stage dataset of human faces with camera poses is a challenge. We worked with the NVPR dataset to obtain the shape and geometry ground truth suitable for our reconstruction based formulation. However, the size of the dataset and the lack of ground truth for several important aspects of light transport prompted us to explore realistic synthetic dataset rendering. Hence, we setup a virtual light stage in Arnold for Maya and render an OLAT dataset using Emily's face mesh with all necessary AOVs required for training our hybrid physical-neural rendering formulation. We then propose a two stage training regime that benefits from a physically based prior and from using a neural renderer to estimate

the direct and indirect components of light transport. We show that our proposed method can suitably approximate the subsurface scattering components and enables high fidelity relighting with complex light transport effects.

Bibliography

[1] Imperial light-stage head (ilsh) dataset. https://sites.google.com/view/vschh/home#h.s92120bbpjva. Online; Accessed: 2023-07-07.

[2] Interactive lookdev lighting rig. https://www.artstation.com/artwork/bK94Va. Online; Accessed: 2023-09-14.

[3] Mediapipe face mesh. https://google.github.io/mediapipe/solutions/face_mesh.html. Online; Accessed: 2022-06-20.

[4] Triplegangers heads. In *https://triplegangers.com/*, 2022.

[5] O Alexander, M Rogers, W Lambeth, J Chiang, W Ma, C Wang, and P Debevec. The digital emily project: Achieving a photoreal digital actor. *IEEE Computer Graphics and Applications*, 30, 2009.

[6] Gwangbin Bae, Martin de La Gorce, Tadas Baltrušaitis, Charlie Hewitt, Dong Chen, Julien Valentin, Roberto Cipolla, and Jingjing Shen. Digiface-1m: 1 million digital face images for face recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3526–3535, 2023.

[7] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5855–5864, 2021.

[8] Anand Bhattad, Aysegul Dundar, Guilin Liu, Andrew Tao, and Bryan Catanzaro. View generalization for single image textured 3d models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6081–6090, 2021.

[9] Piotr Bojanowski, Armand Joulin, David Lopez-Paz, and Arthur Szlam. Optimizing the latent space of generative networks. *arXiv preprint arXiv:1707.05776*, 2017.

[10] Chen Cao, Tomas Simon, Jin Kyu Kim, Gabe Schwartz, Michael Zollhoefer, Shun-Suke Saito, Stephen Lombardi, Shih-En Wei, Danielle Belko, Shoou-I Yu, Yaser Sheikh, and Jason Saragih. Authentic volumetric avatars from a phone scan. *ACM Trans. Graph.*, 41(4), jul 2022.

[11] Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. Efficient Geometry-aware 3D Generative Adversarial Networks. In *CVPR*, 2022.

[12] Eric R Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5799–5809, 2021.

[13] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv: 1512.03012*, 2015.

[14] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. ShapeNet: An Information-Rich 3D Model Repository. Technical Report arXiv:1512.03012 [cs.GR], Stanford University — Princeton University — Toyota Technological Institute at Chicago, 2015.

[15] Zhiqin Chen, Kangxue Yin, and Sanja Fidler. Auv-net: Learning aligned uv maps for texture transfer and synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1465–1474, 2022.

[16] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5939–5948, 2019.

[17] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8188–8197, 2020.

[18] Christopher B Choy, JunYoung Gwak, Silvio Savarese, and Manmohan Chandraker. Universal correspondence network. *Advances in neural information processing systems*, 29, 2016.

[19] ClassAction.org. Snapchat violated ill. users' privacy by collecting biometric information without consent, class action alleges., 2020.

[20] The New York Times Company. The best law you've never heard of., 2021.

[21] The New York Times Company. How illinois is winning in the fight against big tech., 2022.

[22] Paul Debevec. The light stages and their applications to photoreal digital actors. *SIGGRAPH Asia*, 2(4):1–6, 2012.

[23] Boyang Deng, Yifan Wang, and Gordon Wetzstein. Lumigan: Unconditional generation of relightable 3d human faces. *arXiv preprint arXiv:2304.13153*, 2023.

[24] Yu Deng, Jiaolong Yang, and Xin Tong. Deformed implicit field: Modeling 3d shapes with learned dense correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10286–10296, 2021.

[25] Jun Gao, Tianchang Shen, Zian Wang, Wenzheng Chen, Kangxue Yin, Daiqing Li, Or Litany, Zan Gojcic, and Sanja Fidler. Get3d: A generative model of high quality 3d textured shapes learned from images. In *Advances In Neural Information Processing Systems*, 2022.

[26] Iliyan Georgiev, Thiago Ize, Mike Farnsworth, Ramón Montoya-Vozmediano, Alan King, Brecht Van Lommel, Angel Jimenez, Oscar Anson, Shinji Ogaki, Eric Johnston, Adrien Herubel, Declan Russell, Frédéric Servant, and Marcos Fajardo. Arnold: A brute-force production path tracer. *ACM Trans. Graph.*, 37(3), aug 2018.

[27] Lily Goli, Daniel Rebain, Sara Sabour, Animesh Garg, and Andrea Tagliasacchi. nerf2nerf: Pairwise registration of neural radiance fields. *arXiv preprint arXiv:2211.01600*, 2022.

[28] Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. Stylenerf: A style-based 3d-aware generator for high-resolution image synthesis. *arXiv preprint arXiv:2110.08985*, 2021.

[29] Jiatao Gu, Alex Trevithick, Kai-En Lin, Josh Susskind, Christian Theobalt, Lingjie Liu, and Ravi Ramamoorthi. Nerfdiff: Single-image view synthesis with nerf-guided distillation from 3d-aware diffusion. *arXiv preprint arXiv:2302.10109*, 2023.

[30] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[31] Paul Henderson, Vagia Tsiminaki, and Christoph H Lampert. Leveraging 2d data to learn textured 3d mesh generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7498–7507, 2020.

[32] Shun Iwase, Shunsuke Saito, Tomas Simon, Stephen Lombardi, Timur Bagautdinov, Rohan Joshi, Fabian Prada, Takaaki Shiratori, Yaser Sheikh, and Jason Saragih. Relightablehands: Efficient neural relighting of articulated hand models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16663–16673, 2023.

[33] Kaiwen Jiang, Shu-Yu Chen, Hongbo Fu, and Lin Gao. Nerffacelighting: Implicit and disentangled face lighting representation leveraging generative prior in neural radiance fields. *ACM Transactions on Graphics*, 42(3):1–18, 2023.

[34] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[35] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, 34:852–863, 2021.

[36] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

[37] Hyunsu Kim, Gayoung Lee, Yunjey Choi, Jin-Hwa Kim, and Jun-Yan Zhu. 3d-aware blending with generative nerfs. *arXiv preprint arXiv:2302.06608*, 2023.

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

[39] Jiaman Li, Zhengfei Kuang, Yajie Zhao, Mingming He, Karl Bladin, and Hao Li. Dynamic facial asset and rig generation from a single scan. *ACM Trans. Graph.*, 39(6):215–1, 2020.

[40] Min Li, Zhenglong Zhou, Zhe Wu, Boxin Shi, Changyu Diao, and Ping Tan. Multi-view photometric stereo: A robust solution and benchmark dataset for spatially varying isotropic materials. *IEEE Transactions on Image Processing*, 29:4159–4173, 2020.

[41] Xueting Li, Sifei Liu, Shalini De Mello, Xiaolong Wang, Jan Kautz, and Ming-Hsuan Yang. Joint-task self-supervised learning for temporal correspondence. *Advances in Neural Information Processing Systems*, 32, 2019.

[42] Zhengqin Li, Mohammad Shafiei, Ravi Ramamoorthi, Kalyan Sunkavalli, and Manmohan Chandraker. Inverse rendering for complex indoor scenes: Shape, spatially-varying lighting and svbrdf from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2475–2484, 2020.

[43] Connor Z Lin, David B Lindell, Eric R Chan, and Gordon Wetzstein. 3d gan inversion for controllable portrait image animation. *arXiv preprint arXiv:2203.13441*, 2022.

[44] Kai-En Lin, Yen-Chen Lin, Wei-Sheng Lai, Tsung-Yi Lin, Yi-Chang Shih, and Ravi Ramamoorthi. Vision transformer for nerf-based view synthesis from a single input image. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 806–815, 2023.

[45] Feng Liu and Xiaoming Liu. Learning implicit functions for topology-varying dense 3d shape correspondence. *Advances in Neural Information Processing Systems*, 33:4823–4834, 2020.

[46] Hsueh-Ti Derek Liu, Francis Williams, Alec Jacobson, Sanja Fidler, and Or Litany. Learning smooth neural functions via lipschitz regularization. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–13, 2022.

[47] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[48] R. MallikarjunB., Ayush Tewari, Abdallah Dib, T. Weyrich, B. Bickel, H. Seidel, H. Pfister, W. Matusik, Louis Chevallier, Mohamed A. Elgharib, and C. Theobalt. Photoapp: Photorealistic appearance editing of head portraits. *Acm Trans. Graph.*, 2021.

[49] Qiang Meng, Shichao Zhao, Zhida Huang, and Feng Zhou. Magface: A universal representation for face recognition and quality assessment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14225–14234, 2021.

[50] 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. *Communications of the ACM*, 65(1):99–106, 2021.

[51] Jiteng Mu, Shalini De Mello, Zhiding Yu, Nuno Vasconcelos, Xiaolong Wang, Jan Kautz, and Sifei Liu. Coordgan: Self-supervised dense correspondences emerge from gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10011–10020, 2022.

[52] Alex Nichol, Heewoo Jun, Prafulla Dhariwal, Pamela Mishkin, and Mark Chen. Point-e: A system for generating 3d point clouds from complex prompts. *arXiv preprint arXiv:2212.08751*, 2022.

[53] Michael Niemeyer, Jonathan T Barron, Ben Mildenhall, Mehdi SM Sajjadi, Andreas Geiger, and Noha Radwan. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5480–5490, 2022.

[54] Michael Oechsle, Lars Mescheder, Michael Niemeyer, Thilo Strauss, and Andreas Geiger. Texture fields: Learning texture representations in function space. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4531–4540, 2019.

[55] State of Illinois. Illinois civil liabilities (740 ilcs 14/) biometric information privacy act., 2008.

[56] Xingang Pan, Xudong Xu, Chen Change Loy, Christian Theobalt, and Bo Dai. A shading-guided generative implicit model for shape-accurate 3d-aware image synthesis. *Advances in Neural Information Processing Systems*, 34:20002–20013, 2021.

[57] 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.

[58] Dario Pavllo, Jonas Kohler, Thomas Hofmann, and Aurelien Lucchi. Learning generative models of textured 3d meshes from real-world images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13879–13889, 2021.

[59] Matt Pharr, Wenzel Jakob, and Greg Humphreys. *Physically based rendering: From theory to implementation*. MIT Press, 2023.

[60] Anurag Ranjan, Kwang Moo Yi, Jen-Hao Rick Chang, and Oncel Tuzel. Facelit: Neural 3d relightable faces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8619–8628, 2023.

[61] Daniel Rebain, Mark Matthews, Kwang Moo Yi, Dmitry Lagun, and Andrea Tagliasacchi. Lolnerf: Learn from one look. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1558–1567, 2022.

[62] Daniel Rebain, Mark J Matthews, Kwang Moo Yi, Gopal Sharma, Dmitry Lagun, and Andrea Tagliasacchi. Attention beats concatenation for conditioning neural fields. *arXiv preprint arXiv:2209.10684*, 2022.

[63] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *ACM Transactions on Graphics (TOG)*, 42(1):1–13, 2022.

[64] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[65] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[66] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2304–2314, 2019.

[67] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[68] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016.

[69] Robin Sibson. A brief description of natural neighbour interpolation. *Interpreting multivariate data*, pages 21–36, 1981.

[70] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. *Advances in Neural Information Processing Systems*, 33:7462–7473, 2020.

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

[72] Pratul P Srinivasan, Boyang Deng, Xiuming Zhang, Matthew Tancik, Ben Mildenhall, and Jonathan T Barron. Nerv: Neural reflectance and visibility fields for relighting and view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7495–7504, 2021.

[73] 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. *arXiv preprint arXiv:2205.15517*, 2022.

[74] Jingxiang Sun, Xuan Wang, Yong Zhang, Xiaoyu Li, Qi Zhang, Yebin Liu, and Jue Wang. Fenerf: Face editing in neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7672–7682, 2022.

[75] Daichi Tajima, Yoshihiro Kanamori, and Yuki Endo. Relighting humans in the wild: Monocular full-body human relighting with domain adaptation. In *Computer Graphics Forum*, volume 40, pages 205–216. Wiley Online Library, 2021.

[76] Junshu Tang, Bo Zhang, Binxin Yang, Ting Zhang, Dong Chen, Lizhuang Ma, and Fang Wen. Explicitly controllable 3d-aware portrait generation. *arXiv preprint arXiv:2209.05434*, 2022.

[77] Alex Trevithick, Matthew Chan, Michael Stengel, Eric R Chan, Chao Liu, Zhiding Yu, Sameh Khamis, Manmohan Chandraker, Ravi Ramamoorthi, and Koki Nagano. Real-time radiance fields for single-image portrait view synthesis. *arXiv preprint arXiv:2305.02310*, 2023.

[78] Matthew Turk and Alex Pentland. Eigenfaces for recognition. *Journal of cognitive neuroscience*, 3(1):71–86, 1991.

[79] Giuseppe Vecchio, Simone Palazzo, and Concetto Spampinato. Surfacenet: Adversarial svbrdf estimation from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12840–12848, 2021.

[80] LLC. VOX MEDIA. Judge approves $650 million facebook privacy settlement over facial recognition feature., 2021.

[81] LLC. VOX MEDIA. Google to pay $100 million to illinois residents for photos' face grouping feature., 2022.

[82] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P Srinivasan, Howard Zhou, Jonathan T Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2021.

[83] Tengfei Wang, Bo Zhang, Ting Zhang, Shuyang Gu, Jianmin Bao, T. Baltrušaitis, Jingjing Shen, Dong Chen, Fang Wen, Qifeng Chen, and B. Guo. Rodin: A generative model for sculpting 3d digital avatars using diffusion. *Computer Vision and Pattern Recognition*, 2022.

[84] Daniel Watson, William Chan, Ricardo Martin-Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel view synthesis with diffusion models. *arXiv preprint arXiv:2210.04628*, 2022.

[85] Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Sebastian Dziadzio, Thomas J Cashman, and Jamie Shotton. Fake it till you make it: face analysis in the wild using synthetic data alone. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3681–3691, 2021.

[86] Fanbo Xiang, Zexiang Xu, Milos Hasan, Yannick Hold-Geoffroy, Kalyan Sunkavalli, and Hao Su. Neutex: Neural texture mapping for volumetric neural rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7119–7128, 2021.

[87] Taihong Xiao, Sifei Liu, Shalini De Mello, Zhiding Yu, Jan Kautz, and Ming-Hsuan Yang. Learning contrastive representation for semantic correspondence. *International Journal of Computer Vision*, 130(5):1293–1309, 2022.

[88] Jiaxin Xie, Hao Ouyang, Jingtan Piao, Chenyang Lei, and Qifeng Chen. High-fidelity 3d gan inversion by pseudo-multi-view optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 321–331, 2023.

[89] Jiarui Xu and Xiaolong Wang. Rethinking self-supervised correspondence learning: A video frame-level similarity perspective. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10075–10085, 2021.

[90] Bangbang Yang, Chong Bao, Junyi Zeng, Hujun Bao, Yinda Zhang, Zhaopeng Cui, and Guofeng Zhang. Neumesh: Learning disentangled neural mesh-based implicit field for geometry and texture editing. *arXiv preprint arXiv:2207.11911*, 2022.

[91] Haotian Yang, Hao Zhu, Yanru Wang, Mingkai Huang, Qiu Shen, Ruigang Yang, and Xun Cao. Facescape: a large-scale high quality 3d face dataset and detailed riggable 3d face prediction. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 601–610, 2020.

[92] 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)*, 41(6):1–21, 2022.

[93] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4578–4587, 2021.

[94] Hong-Xing Yu, Michelle Guo, Alireza Fathi, Yen-Yu Chang, Eric Ryan Chan, Ruohan Gao, Thomas Funkhouser, and Jiajun Wu. Learning object-centric neural scattering functions for free-viewpoint relighting and scene composition. *arXiv preprint arXiv:2303.06138*, 2023.

[95] Xiaohui Zeng, Arash Vahdat, Francis Williams, Zan Gojcic, Or Litany, Sanja Fidler, and Karsten Kreis. Lion: Latent point diffusion models for 3d shape generation. *arXiv preprint arXiv:2210.06978*, 2022.

[96] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. Nerf++: Analyzing and improving neural radiance fields. *arXiv preprint arXiv:2010.07492*, 2020.

[97] Longwen Zhang, Qixuan Zhang, Minye Wu, Jingyi Yu, and Lan Xu. Neural video portrait relighting in real-time via consistency modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 802–812, 2021.

[98] 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.

[99] Xiaoming Zhao, Fangchang Ma, David Güera, Zhile Ren, Alexander G Schwing, and Alex Colburn. Generative multiplane images: Making a 2d gan 3d-aware. *arXiv preprint arXiv:2207.10642*, 2022.

[100] Shizhan Zhu, Shunsuke Saito, Aljaz Bozic, Carlos Aliaga, Trevor Darrell, and Christop Lassner. Neural relighting with subsurface scattering by learning the radiance transfer gradient. *arXiv preprint arXiv:2306.09322*, 2023.