High-Quality Facial Geometry from Sparse Heterogeneous Cameras under Active Illumination

Lewis Bridgeman
lewis.bridgeman@lumirithmic.com
Lumirithmic Ltd
London, UK

Gilles Rainer
g.rainer@imperial.ac.uk
Imperial College London
London, UK

Abhijeet Ghosh
ghosh@imperial.ac.uk
Imperial College London,
Lumirithmic Ltd
London, UK

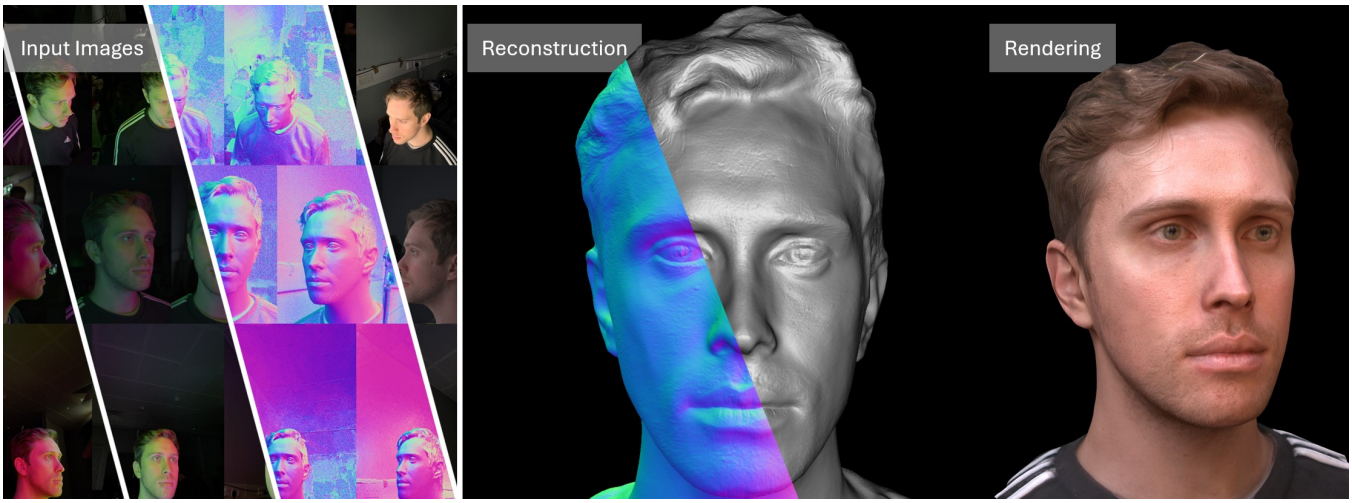

Figure 1: Given multi-view images and their reflectance decomposition from active illumination (*left*), our method reconstructs high-quality surface geometry (*middle*), which can be used for highly realistic rendering (*right*).

Abstract

High-resolution facial geometry is essential for realistic digital avatars. Traditional reconstruction methods, such as multi-view stereo, often struggle with materials like skin, which exhibit complex light reflection, absorption, and scattering properties. Neural reconstruction methods have shown greater robustness to these view-dependent effects. However, positional-encoding-based implementations are typically slow, while faster hash-encoded methods may falter under sparse camera views. We present a geometry reconstruction method tailored for an active-illumination facial capture setup featuring sparse cameras with varying characteristics. Our technique builds upon hash-encoded neural surface reconstruction, which we enhance with additional active-illumination-based supervision and loss functions, allowing us to maintain high reconstruction speed and geometrical fidelity even with reduced camera coverage. We validate our approach through qualitative evaluations across diverse subjects, and quantitative evaluation using a synthetic dataset rendered with a virtual reproduction of our capture setup. Our results demonstrate that our method significantly outperforms previous neural reconstruction techniques on datasets with sparse camera configurations.

CVMP '24, November 18–19, 2024, London, United Kingdom
2024. ACM ISBN 979-8-4007-1281-4/24/11
https://doi.org/10.1145/3697294.3697296

CCS Concepts

• **Computing methodologies** → **Reconstruction**; **Shape representations**; *Computer graphics.*

Keywords

Facial, Geometry, Reconstruction, Neural, Active, Illumination, Mesh, BRDF, Avatar, Surface, Mesh, Detail

ACM Reference Format:
Lewis Bridgeman, Gilles Rainer, and Abhijeet Ghosh. 2024. High-Quality Facial Geometry from Sparse Heterogeneous Cameras under Active Illumination. In *Proceedings of the 21st ACM SIGGRAPH European Conference on Visual Media Production (CVMP '24).* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3697294.3697296

1 Introduction

High-quality digital assets of human faces, coupled with the ability to scan people, are highly valuable across numerous domains such as visual effects, medicine, cosmetics, gaming, and more. They are also central to virtual human interactions, such as in the *Meta-Verse.* The standard for simulating and rendering digital faces is exceptionally high, because humans are adept at observing and analyzing faces. Even slight inaccuracies or implausibilities may be noticed, often unconsciously, which can elicit negative reactions. This phenomenon is known as the *Uncanny Valley* [Mori 2012]: the

more realistic and human-like the facial simulation, the higher its fidelity needs to be in order to achieve plausibility and realism.

In this paper, we focus on reconstructing accurate facial geometry for a specific scanning setup developed by Lattas et al. [Lattas et al. 2022]. Various facial scanning setups have been introduced in the past (e.g. the famous *Light Stage* [Debevec et al. 2023]); the configuration used in this paper consists of eight iPads (with controllable screens and selfie cameras) and five DSLR cameras. The iPad screens, arranged on the surface of a hemisphere, are used to sequentially display two colored binary patterns, and both the iPad and DSLR cameras capture one image of the subject under each lighting condition. This active lighting capture is used to decompose the radiance observations into reflectance maps (albedos and normals) using the analytic method in [Lattas et al. 2022]. Our method takes the captured images, plus the derived appearance maps, from these 13 cameras as input; these are used for additional supervision of the geometry reconstruction, which is one of our contributions.

The traditional geometry reconstruction pipeline for uncalibrated cameras consists of *Structure from Motion* (SfM) [Ullman 1979] and *Multi-view Stereo* (MVS) [Hernandez Esteban et al. 2008]. In the first pass, each camera's intrinsics and pose are estimated, while in MVS, the calibrated cameras are used to infer the object's surface. One underlying assumption of this method is that the surfaces are Lambertian (diffuse), such that every point on the surface will produce the same observation in all cameras. While generally robust, MVS can struggle with shiny surfaces (reconstructing bumps where highlights are observed) or texture-less surfaces.

More recently, neural methods have become the new standard for high-quality geometry reconstruction. Pioneered by first applications of per-scene representational learning such as NeRF [Mildenhall et al. 2020], small neural networks regressing density in 3D space have shown exceptional results for multi-view geometry estimation. However their training can be prohibitively long, taking up to days where MVS takes minutes. This led to the introduction of acceleration structures such as a Hash Grid of neural features [Müller et al. 2022]. These accelerated neural reconstructions are able to achieve speeds comparable to MVS, but for challenging datasets such as ours with sparse views and wide baselines, reconstruction quality tends to deteriorate.

Our method is based on a hash-encoding accelerated version of the original *NeuS* [Wang et al. 2021], but we introduce several modifications to the training objectives that make the method more robust and accurate, especially for face captures from a sparse set of cameras. Specifically, we incorporate:

- a monocular depth loss,
- active-illumination-based supervision,
- a masking loss for face and background disambiguation.

To further enhance the recovered facial geometry, we incorporate additional processing stages for clean mesh re-topology and the transfer of fine detail from the measured normals to the geometry itself. We evaluate these improvements against baseline neural and traditional reconstruction methods across a variety of real-world captures, as well as a synthetic dataset that we render in a virtual replica of the acquisition setup, in order to also obtain quantitative geometric errors.

2 Related Work

2.1 Face Scanners

Face scanning has been a prolific area of research in Computer Graphics for a long time, initially motivated by VFX applications. The *Light Stage* [Debevec et al. 2023] and its numerous variations have established themselves as the standard setups for scanning of human faces, combining polarization imaging and illumination [Ghosh et al. 2011; Ma et al. 2007]. Other setups have removed the need for full spherical lighting coverage, the requirement of active illumination [Riviere et al. 2020], and the need for polarized cameras [Gotardo et al. 2018], but still employ high-quality hardware on a calibrated, fixed setup. Recently, Lattas et al. [Lattas et al. 2022] proposed a low-cost scanning setup consisting of uncalibrated cameras and screens for active lighting. The two-shot capture (illustrated in Fig.4) produces reflectance maps for each camera, which can then be reprojected onto the reconstructed geometry.

The traditional geometry reconstruction pipeline involves obtaining the camera poses, either through explicit calibration using known target objects [Zhang 2004], or via Structure-from-Motion [Ullman 1979]. The Multi-View Stereo [Hernandez Esteban et al. 2008] pipeline then uses the posed cameras to triangulate points in the scene, which are subsequently merged into a triangle mesh. While this approach is not specific to human faces, photogrammetry or variants [Beeler et al. 2010] are still the most commonly used approach across face scanning setups, and generally performed as pre-processing before any further tasks such as registration [Li et al. 2017] or deep learning [Cao et al. 2022]. The reconstructed geometry can be further enhanced to the pore level by embossing normal maps into the surface [Nehab et al. 2005].

2.2 Neural Geometry Reconstruction

Neural Radiance Fields (NeRF [Mildenhall et al. 2020]) introduced a radically new approach to scene and geometry modelling, as they encode matter as density in space learnt by a small neural network, as opposed to the traditional mesh of triangles. Although initially designed for view synthesis, a surface can again be extracted from the learnt density via classic techniques such as marching cubes [Lorensen and Cline 1987]. Slightly modifying the model, NeuS [Wang et al. 2021] also regresses matter in space using a small Multi-Layer Perceptron, except here the network learns the value of a signed distance function (*SDF*), from which volume density can be derived. Once training is finished, a surface can naturally be extracted (and meshed) as the zero-level set of this neural SDF.

Acceleration. Although the results were incredible, initial neural scene representations suffered from prohibitively long training times, up to several days [Barron et al. 2021]. Extensive research has gone into acceleration structures for neural representations, the most effective being the Hash Grid [Müller et al. 2022]. Initially developed for NeRFs, Hash Grids can also be employed for neural surface reconstruction, effectively decreasing the convergence time from hours to minutes. NeuS2 [Wang et al. 2023] presents an implementation of such which is entirely implemented in CUDA for rapid training, and extend the process to dynamic scenes. Neuralangelo [Li et al. 2023b] uses numerical gradients to overcome locality in the computation of analytical gradients of a hash-encoding.

Regularizers & Priors. Another initial drawback of neural scene representations was their reliance on well-posed data. Various modifications to the training supervision and objectives have been proposed to deal with datasets imperfections, most commonly sparse views [Long et al. 2022; Niemeyer et al. 2022; Yu et al. 2021] or lighting changes between views [Martin-Brualla et al. 2021]. Other efforts attempted to better constrain the results by using additional information in the training, e.g. supervising with depth information [Deng et al. 2022], reflectance and normal information [Brument et al. 2023], or even independent monocular geometric cues that are extracted from each input view [Yu et al. 2022].

Gaussian Splatting. Most recently, Gaussian Splatting [Kerbl et al. 2023] was proposed as a new volumetric representation for view synthesis. Although there are many similarities with neural implicit representations, and techniques have been proposed to extract mesh geometry from splats [Guédon and Lepetit 2024; Huang et al. 2024; Turkulainen et al. 2024], we believe neural implicit representations to be better regularizers in our setting and to lead to more accurate and robust surface estimation.

Meshing. Despite Marching Cubes being the most prominent method for meshing isosurfaces, it fails to recover sharp features and can suffer from artefacts. Neural Marching Cubes [Chen and Zhang 2021] boosts the performance of the classic method using machine learning to better preserve finer details. Subsequent work extends Dual Contouring with machine learning [Chen et al. 2022], reducing the vertex count required to achieve similar mesh fidelity. DMTet [Shen et al. 2021] represents geometry as a SDF defined on a deformable tetrahedral grid, allowing the surface to be recovered through marching tetrahedra.

2.3 Neural Representations for Faces

Implicit neural representations also find applications in facial modelling, their main advantage being the high quality view synthesis. Beyond static rendering, encoding information in a neural network or a neural feature grid opens doors for learning dynamic faces across time [Gafni et al. 2021; Kirschstein et al. 2023; Lombardi et al. 2021; Park et al. 2021] as well as across lighting conditions [Lombardi et al. 2018; Rainer et al. 2023; Rao et al. 2022], or even across a database of faces [Chan et al. 2021].

However, in these methods the scene representation remains neural, extracting a mesh would lose desirable properties of the volumetric representation which can much more smoothly interpolate than a triangle mesh where artifacts are extremely visible. Burkov et al. [Burkov et al. 2022] leverage the NeuS representation across a database of portraits, to estimate head geometry given a single image, but the results suffer in this case from being overly smooth. Giebenhain et al. [Giebenhain et al. 2023, 2024] use an ensemble of small neural networks to learn a parametric head model across a database of subjects and expressions – although these output a mesh, they also inherit the smoothness from the low-dimensional parametric encoding.

We propose to similarly use a neural implicit geometry representation to reconstruct our scans from multi-view input. While these representations achieve impressive results in various synthesis tasks, they are tailored to these applications and can't easily be edited by artists or integrated in common rendering engines. The standard, default representation for computer graphics assets remain meshes and textures, so our goal is to output a high fidelity mesh for every scan: the neural representation serves as a robust model for the optimization, where we introduce novel additional objectives and losses, but the final output is a high-resolution mesh.

3 Method

In this section, we present our method for generating high-quality geometry in a setting where common methods struggle due to the difficult capture conditions: low-light and high exposure imaging; heterogeneous cameras with varying camera and color parameters; use of iPad selfie cameras, which are low resolution and prepossessed with proprietary software; a relatively sparse arrangement of cameras.

By incorporating additional priors and utilizing active illumination data, our reconstruction becomes robust to these sub-optimal capture settings. Our approach is based upon SDF-based volume-rendering, using multi-resolution hash grid encoding. We extend on this foundation by integrating images captured under active-illumination, and introducing additional losses to guide the optimization. These new losses aim to maintain the quality of the geometry under challenging capture conditions; in our case, minimal coverage from cameras that range in image quality. The key components of our method are:

- **Active-illumination Supervision:** Employing multiple images captured under different illumination conditions, along with estimated surface normals, to provide a stronger supervisory signal.
- **Monocular Depth Loss:** Using depth images estimated from each camera view as a prior, aiding the reconstruction of concave surfaces with reduced visibility.
- **Mask Loss:** Incorporating a monocular estimate of object segmentation as a shape-from-silhouette term.

Volume Rendering of an SDF:. Our method builds upon the neural SDF representation introduced in NeuS [Wang et al. 2021]. Here, a surface $\mathcal{S}$ is represented by the implicit function $\mathcal{S} = \{x \in \mathbb{R}^3 | f(x) = 0\}$, where the zero-level set denotes the underlying surface. NeuS proposes a formula for converting the value of the signed-distance function into an opacity value:

$$\alpha_i = \max\left(\frac{\Phi_s(f(\mathbf{p}(x_i))) - \Phi_s(f(\mathbf{p}(x_{i+1})))}{\Phi_s(f(\mathbf{p}(x_i)))}, 0\right) \tag{1}$$

where Φ_s is the sigmoid function, and α_i is the resulting opacity at point x_i. This conversion allows the SDF to be utilized in the NeRF volume-rendering function [Mildenhall et al. 2020] for optimization of the underlying surface.

Positional Encoding and Hash Encoding: The original formulation of NeuS utilizes positional encoding, where the xyz positional coordinates are mapped into a higher-dimensional space $f : \mathbb{R}^3 \rightarrow \mathbb{R}^{3\times 2L}$. This encoding is provided as input to the MLP that learns the SDF, enabling the network to capture high-frequency details more effectively than using xyz coordinates alone. In contrast, hash encoding, introduced in [Müller et al. 2022], employs multi-resolution grids where each cell maps to an entry in a hash

table, storing learnable features. Given a position xyz, the corresponding features across all grid resolutions are concatenated to form the feature vector input to the MLP. Hash encoding allows for a smaller network compared to positional encoding, and is significantly faster to optimise; learning the SDF in [Li et al. 2023b; Wang et al. 2023] takes minutes, whereas the original NeuS requires hours. However, hash encoding lacks the implicit regularization provided by the MLP, and training all resolutions simultaneously can lead to overfitting, which may introduce noise or artefacts in the geometry. To address this, existing hash-encoded SDF methods [Li et al. 2023b; Wang et al. 2023] use a progressive training approach, where higher-resolution grids are added gradually, mitigating these issues for typical neural rendering datasets. Despite this, when camera coverage is reduced or image quality diminishes, hash encoding still tends to degrade in quality compared to positional encoding results, even with the progressive hash-encoding approach. Our extension to hash-encoded SDF methods is designed to preserve geometry quality even in datasets with sparse or varying camera quality.

3.1 Active-Illumination Supervision

For our experiments we use data captured by the specific facial scanning setup developed by Lattas et al. [Lattas et al. 2022]. However, this method is equally applicable to any active illumination capture setup capable of photographing a subject under multiple lighting conditions. In our setup, we capture subjects in a static pose using two complementary binary patterns. Following the approach described in [Lattas et al. 2022], we derive facial reflectance maps from these captures. By adjusting the output size of the color MLP, we can train the network not only to learn RGB values under a single illumination condition, but also to handle multiple active lighting conditions or analytically derived maps. Through empirical experimentation, we determine that concatenating the mixed albedo, the two binary gradients, and the derived specular surface normals yields the best result on our datasets. This selection of training images is shown in Fig. 1.

3.2 Monocular Depth Loss

In datasets with sparse camera coverage, the network may struggle to converge to the true surface in areas lacking sufficient multi-view observations. This issue is exacerbated for concave objects, where self-occlusion further diminishes camera visbility. To help the network converge to the true surface with fewer observations, we incorporate the prior from a monocular depth estimation network.

We first output the estimated depth values from our SDF using a modified version of the discrete ray-accumulation formula originally applied to color in [Wang et al. 2021]:

$$\hat{D} = \sum_{i=1}^{n} T_i \alpha_i d_i \tag{2}$$

where T_i is the accumulated transmittance $T_i = \prod_{j=1}^{i-1}(1-\alpha_j)$, and d_i represents the discrete depth values along the ray. Depth estimates D for each input view are generated using [Birkl et al. 2023]. We then compute a scale-invariant depth loss, as used in training monocular depth estimation networks in [Eigen et al. 2014].

$$L_{depth}(D,\hat{D}) = \frac{1}{n}\sum_{p} d_p^2 - \frac{\lambda}{n^2}\left(\sum_{p} d_p\right)^2 \tag{3}$$

where $d_p = \log D_p - \log \hat{D}_p$ for each pixel p. In this formula, setting $\lambda = 0$ is equivalent to $L2$, while $\lambda = 1$ yields a fully scale-invariant function. We set λ to 1, to avoid assumptions about the absolute scale of the monocular depth estimates, focusing instead their inter-pixel relative depths.

3.3 Mask Loss

In our multi-view facial datasets, we care only about reconstructing the geometry within the capture volume, *i.e.* within the region covered by active illumination. To this end, we model the capture region with a volume-rendered SDF, and model the outside region with a hash-encoded NeRF++ [Zhang et al. 2020]. However, with sparse camera supervision, the optimization may have trouble automatically distinguishing between foreground and background, potentially leading to artefacts in the geometry. To improve the delineation of the foreground region, we introduce a mask loss that utilizes the prior from a pre-trained matting network, in order to guide the SDF during the early stages of training. For each input image, we generate alpha mattes S using the method described in [Ke et al. 2022]. During training, we estimate current foreground segmentation $\hat{S}$ at the current step using the discrete ray-accumulation formula, accumulating opacity values:

$$\hat{S} = \sum_{i=1}^{n} T_i \alpha_i \alpha_i \tag{4}$$

We can compute the masking loss using the mean squared error:

$$L_{mask}(S,\hat{S}) = \frac{1}{n}\sum_{p}^{N}(S-\hat{S})^2 \tag{5}$$

We apply the masking loss for the first 25% of training steps, gradually reducing its influence to zero. This approach allows the network to refine the foreground segmentation details directly from the images.

3.4 Re-topology

The usual approach to meshing an SDF is the marching cubes algorithm, which uses a lookup table to determine triangle connectivity given a voxelised representation of the distance field. However, marching cubes suffers from several issues that affect the quality of the final geometry: "staircase" artefacts, topological ambiguities, and over-smoothing of sharp geometry features. Some of these artefacts can be observed in the left panel on Fig. 2. We apply post-processing to our marching cubes output to generate a mesh with clean topology but equivalent detail. First, we re-topologize our geometry using the Discrete Voronoi Diagrams method presented in [Valette et al. 2008], as shown in the centre panel of Fig. 2. We generate a coarse mesh at this stage and follow up with several subdivision steps. Finally, we optimize the positions of the new vertices such that they lie on the zero-level set of our learned SDF

using the loss function:

$$L_{SDF} = \frac{1}{\hat{V}} \sum_{i=0}^{\hat{V}} f(\hat{v}_i)^2 \tag{6}$$

where $\hat{v}_i$ are randomly sampled surface points on the mesh for each iteration. We also use a normal term to maintain surface detail:

$$L_{normals} = 1 - \frac{1}{\hat{V}} \sum_{i=0}^{\hat{V}} \frac{\Delta\hat{v}_i \cdot \Delta f(\hat{v}_i)}{\|\Delta\hat{v}_i\| \|\Delta f(\hat{v}_i)\|} \tag{7}$$

where $\Delta\hat{v}_i$ is the normal corresponding to mesh point $\hat{v}_i$, and $\Delta f(.)$ is the gradient of the SDF, which is conveniently already included in the NeuS formulation as part of the Eikonal term. This optimization takes less than a minute, and the results can be seen in the right panel of Fig. 2. This approach results in a much cleaner mesh topology, making it more suitable for normal embossing (section 4.6), while maintaining the original level of detail represented in the SDF.

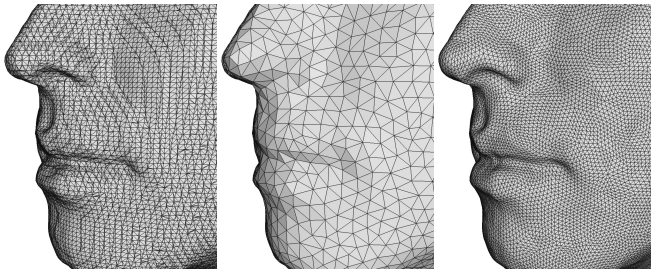

Figure 2: *Left:* The result of meshing the learned SDF using traditional marching cubes. *Centre:* The SDF field re-topologized with Voronoi-clustering [Valette et al. 2008]. *Right:* The result of the re-topologized and subdivided mesh optimized to fit the zero-level set of the SDF.

4 Results and Evaluation

4.1 Implementation Details

We implement our method in Pytorch using the the NerfAcc library [Li et al. 2023a]. For our Signed Distance Function (SDF) we utilize a hash grid with 14 levels, and a smaller grid with 10 levels for the outer NeRF. The SDF network is an MLP with one hidden layer of size 64, while the color network is an MLP with two hidden layers each of size 64. The SDF network is initialized to a sphere of radius 0.3 meters, and we train the network with input images resized to 800×1200 pixels. All experiments are conducted on an Nvidia GeForce 3090 GPU over 10,000 iterations.

4.2 Data Capture & Input

Our datasets are captured with the setup described in [Lattas et al. 2022], comprising 13 cameras: 5 Canon EOS M50s, and 8 iPad Pro selfie cameras (5th generation). The Canon cameras are arranged in a semi-circular configuration in front of the subject, while the iPads are arranged in two rows above and below. Subjects are captured under two color multiplexed binary gradient patterns which are displayed on the iPad screens. Each pattern is captured with a 330ms exposure time. An optical flow step is used to correct for any motion between the two binary images.

Following the method of [Lattas et al. 2022], the binary gradient patterns are decomposed into a set of reflectance maps: diffuse and specular albedos, and normals. We use structure-from-motion to find the camera parameters, using Agisoft Metashape [Agisoft LLC 2024]. We use a combination of reflectance maps in training our SDF representation (section 3.1), the specular normals for embossing the geometry (section 4.6), and the full set of BRDF maps are used for rendering (Fig. 1).

4.3 Quantitative Evaluation on Synthetic Data

Data Generation. Obtaining ground truth geometry for a quantitative comparison of our reconstruction to previous methods is a time-consuming process that would require another measurement technique as reference, such as Structured Light or Laser scanning, or using a reference object with known geometry. We instead opt to use a synthetic model to generate input data to our method; the reconstruction can then be compared to the model that was used to generate the data. We choose the Digital Emily [Alexander et al. 2009; The Wikihuman Project 2015] model (artist-cleaned mesh and appearance maps) and build a virtual replica of the scanning setup of Lattas et al. [Lattas et al. 2022] in Blender [Blender Foundation 2018], as shown in Fig. 4. The two lighting conditions are depicted on the left, simulating the iPad lighting with area light sources, and rendering those produces the binary-illumination, multi-view images shown on the right. From these, we compute the additional reflectance maps to input to our method: If our reconstruction is perfect, we should obtain the same mesh as the original Digital Emily model (reconstructions shown in Fig. 3).

We compare our method to the Multi-view Stereo (MVS) result of Agisoft Metashape [Agisoft LLC 2024], the original NeuS method [Wang et al. 2021], and the accelerated NeuS2 [Wang et al. 2023]. Each of these methods was trained or ran using the the mixed albedo images, while ours uses a larger selection of BRDF maps as described. We use two primary metrics to evaluate the performance of the reconstruction methods: the average perpendicular distance between the ground truth geometry and recovered geometry; and the angular difference in surface normals between the ground truth and reconstructed meshes. These results of our comparison are summarized for all methods in Table 1.

Our method demonstrates superior performance in both the mesh distance and angular difference metrics across all tested methods. Achieving both a smaller average perpendicular distance and a smaller difference in normals indicates that our method generates the most complete and most accurate surface, while also replicating the finer surface details. Notably, we are able to achieve a similar runtime to NeuS2 without sacrificing the quality of the results. In addition to the quantitative metrics, the reconstructed meshes from each method are displayed in Fig. 3. These results demonstrate that our method produces the most complete reconstruction while maintaining high-frequency details.

4.4 Qualitative Comparisons

A qualitative comparison of geometry reconstruction for all methods can be seen on a selection of captured datasets in Fig. 5. As can

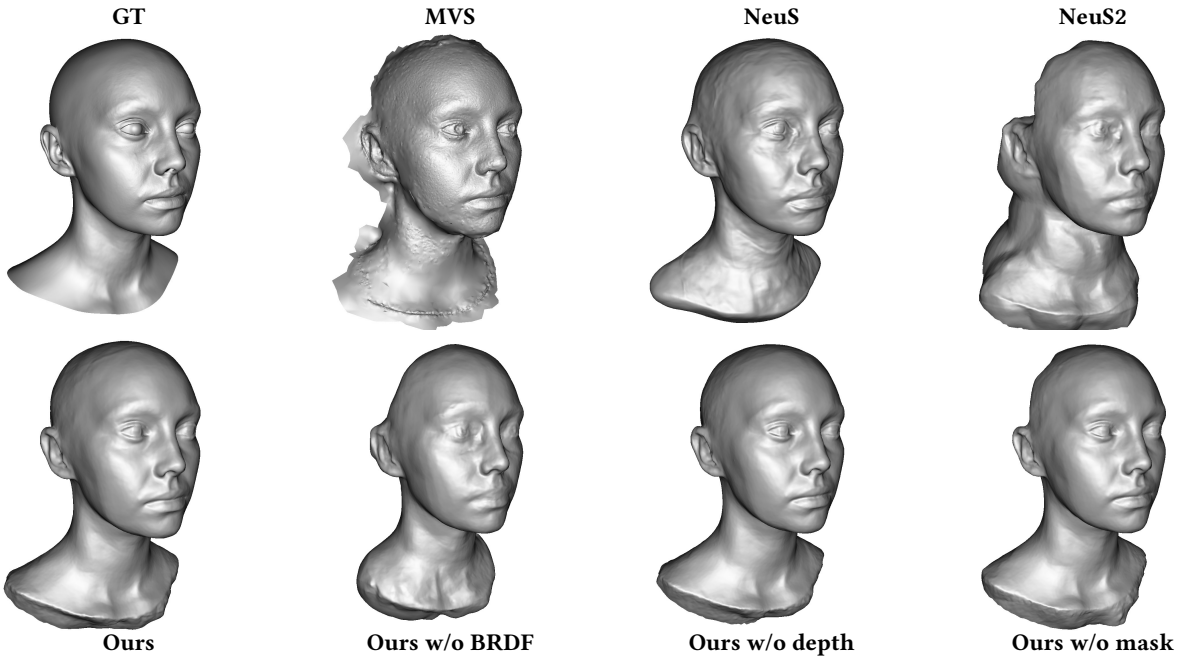

Figure 3: Qualitative visualization of the reconstructions of various methods and ablation study reported in Tab. 1, on the Digital Emily dataset [The Wikihuman Project 2015].

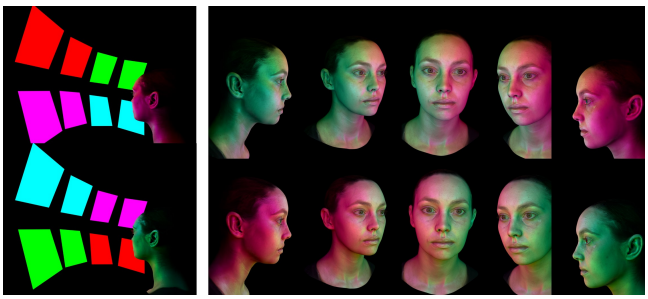

Figure 4: *Left*: Digital replica of the capture setup of Lattas et al. [Lattas et al. 2022], using Digital Emily [The Wikihuman Project 2015]. *Right*: Renderings given to our reconstruction method to compute quantitative metrics against the ground truth geometry.

be seen, MVS demonstrates its ability to capture high-frequency detail, however suffers from increased noise which can obscure the finer details; it also often fails to recover the full shape, resulting in incomplete reconstructions. NeuS generally provides a more complete reconstruction, however it often fails to distinguish between the foreground from background with the reduced number of cameras. NeuS2 achieves a complete reconstruction, but struggles to resolve the true surface in some circumstances, resulting in over-smoothed face details. In comparison, our method provides the highest quality results overall. It manages to resolve a high level of detail without introducing noise or any of the noticeable artefacts that present themselves in the results of previous methods.

4.5 Ablation Study

We ablate the key contributions of our method to evaluate their necessity and impact on the overall performance. We perform a quantitative and qualitative evaluation of the ablated methods on the Digital Emily dataset, and these results are presented in Tab. 1 and Fig. 3 respectively. Here we demonstrate the results of removing the active-illumination supervision (w/o BRDF), the monocular depth loss (w/o depth), and the mask loss (w/o mask). The quantitative results demonstrate that the active-illumination supervision and mask loss make a clear improvement to the accuracy of the recovered surface, and the normal detail. The influence of the monocular depth loss is less significant on the Digital Emily dataset. However we observe that while all of our additional losses are necessary for achieving consistently accurate results across various datasets, not every loss significantly affects the results for every individual dataset. We also present additional qualitative ablation examples in Fig 6. These examples have been chosen to highlight datasets where each loss demonstrably impacts quality of the result. As shown, the mask loss aids in foreground-background disambiguation, the depth loss helps resolve the surface in concave regions, and the active-illumination loss enhances surface detail.

4.6 Further Geometry Processing

The reconstructed face geometry can further be augmented to pore-level detail by embossing the given normal maps, as initially proposed by Nehab et al. [Nehab et al. 2005]. We show the result of such a post-processing step in Fig. 7. The normals of the reconstructed meshes are visualized before and after embossing, compared to the given camera space normals on the left. As shown

Table 1: Reconstruction error (↓) with regards to ground truth on the Digital Emily 2 [The Wikihuman Project 2015] model.

Method	MVS	NeuS	NeuS2	**Ours**	Ours w/o BRDF	Ours w/o depth	Ours w/o mask
Mean Dist. (mm)	20.42 ± 47.56	3.63 ± 13.12	8.42 ± 19.52	**2.03** ± 8.51	4.16 ± 14.19	2.18 ± 8.29	2.91 ± 11.01
Mean Angle (°)	13.38 ± 17.23	14.42 ± 18.19	18.06 ± 19.73	**12.09** ± 16.79	16.87 ± 18.69	12.16 ± 17.08	12.87 ± 17.18
Median Dist. (mm)	0.41	0.36	0.91	**0.30**	0.59	0.31	0.36
Median Angle (°)	6.07	6.22	8.93	4.64	8.77	**4.54**	5.02
Runtime	**3m**	10h	5m	10m	-	-	-

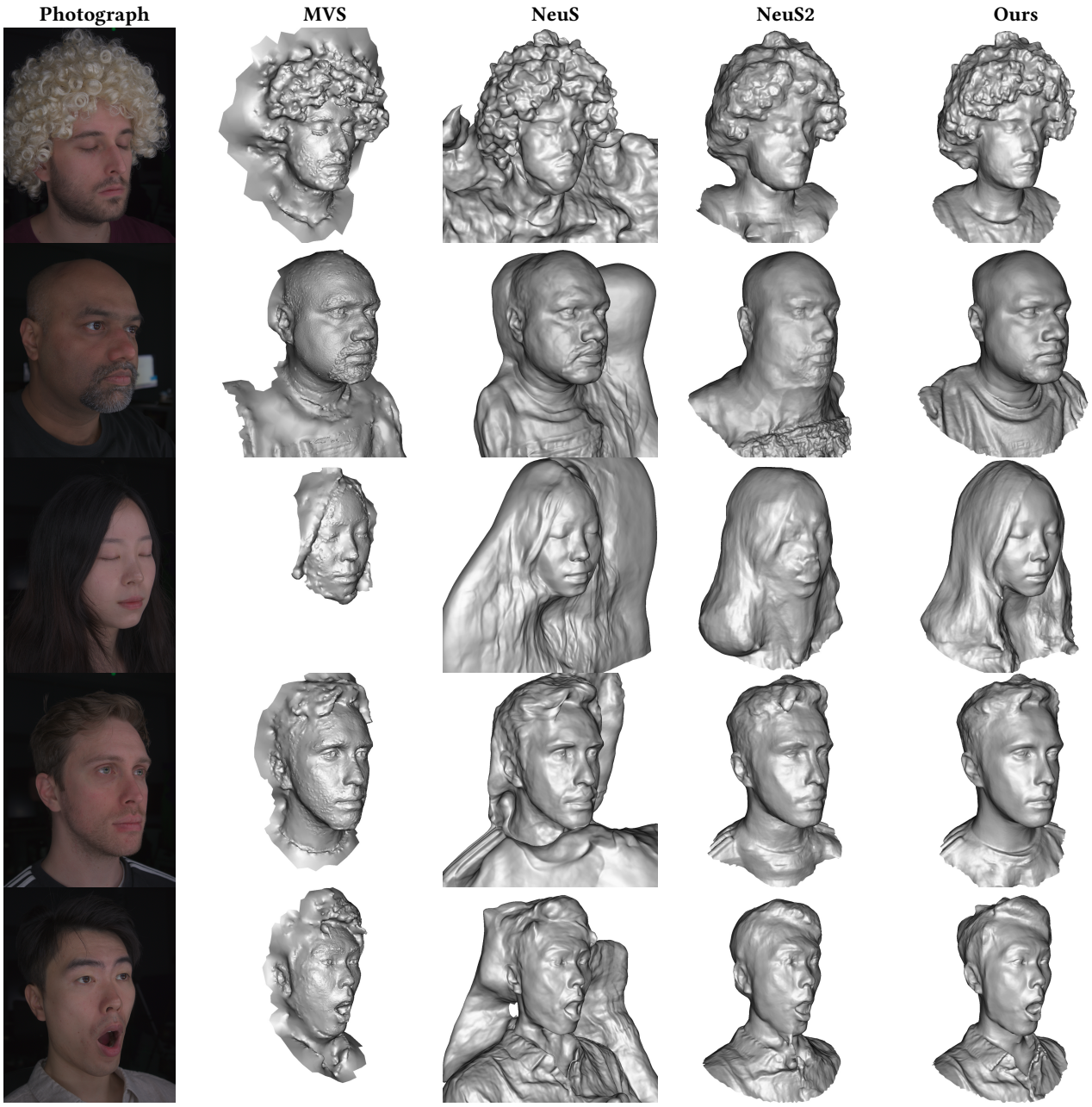

Figure 5: Visualization of geometries reconstructed by different methods, for subjects of different genders, skin types, and hairstyles. Across the board, our method outperforms competitors in terms of both details and robustness.

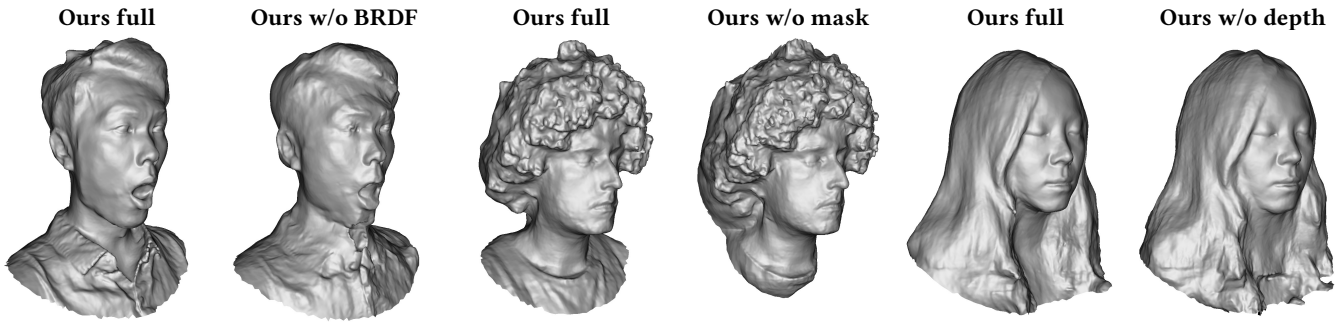

Figure 6: Visualization of geometries reconstructed with our full method and ablated versions of our method.

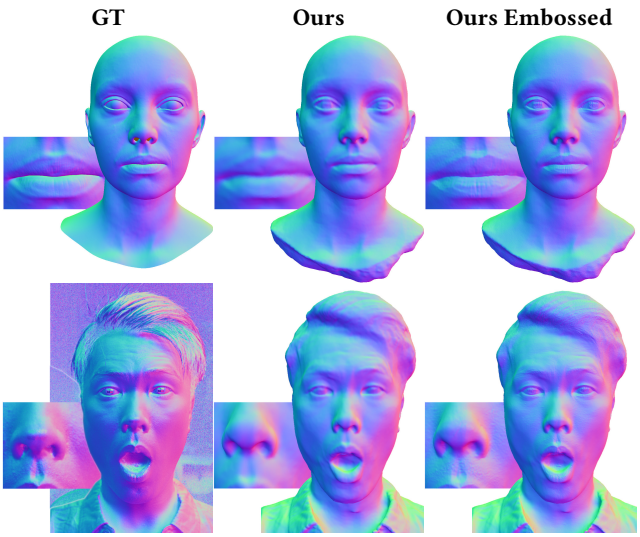

Figure 7: *Left*: Given camera-space normal map. *Middle*: Reprojection of our output geometry orientation into the same view. *Right*: Final mesh normals after embossing of the given camera-space normals. Synthetic data (*top*), real (*bottom*).

in the insets, the initial reconstruction, although already high in fidelity, appears smooth compared to the pixel-detail level of the camera-space normal maps. After embossing, the insets reveal that much of the detail is accurately transferred to the surface geometry.

4.7 Object Scanning.

While our implementation for the mask loss (see Sec. 3.3) currently uses a face matting network [Ke et al. 2022], this can easily be replaced by an object detection network or a foreground/background segmentation network, in order to apply the proposed pipeline to arbitrary objects (as demonstrated in Fig. 8). All other components of the proposed method apply to any object or scene that can be expressed as a surface with given, pre-estimated reflectance components.

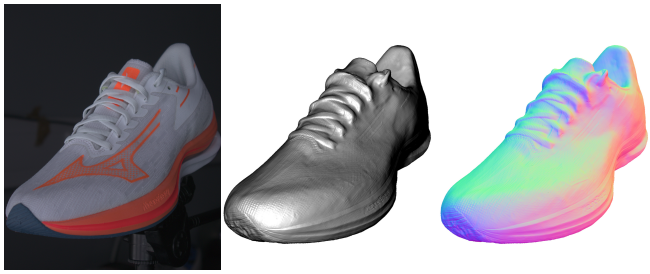

Figure 8: The proposed reconstruction method is not restricted to faces, it can also handle objects with intricate geometries such as the shoelaces here. *Left to right*: Photograph, reconstructed mesh and normals.

4.8 Limitations and Failure Cases

Our additional losses are able to assist the network in converging to the correct surface in the case of reduced supervision from a sparser set of cameras. However, the reliance on additional supervisory signals, such as monocular depth or matting estimates, can lead to failure when these are erroneous. The monocular depth and alpha matting networks [Birkl et al. 2023; Ke et al. 2022] struggle particularly in low-light images, so our active illumination capture must be sufficiently bright.

5 Conclusions and Future Work

In this paper, we introduced a novel method for reconstructing high-quality 3D face geometry in a challenging capture setting. By leveraging the additional supervision provided by active-illumination capture, and the priors of pre-trained monocular depth estimation and alpha matting networks, we are able to improve on the quality of state-of-the-art neural surface reconstruction methods while greatly reducing the number of observations. Furthermore, we maintain a competitive runtime versus recent accelerated neural reconstruction methods. We show that our method is able to consistently produce complete reconstructions with accurate normal detail on a wide range of human subjects, and the method naturally extends to objects. The method seamlessly inserts itself into a wider face geometry and appearance capture pipeline, allowing the generation of hyper-realistic renderings.

Future work could explore using the given normals to directly supervise the gradient of the level-set of the neural SDF during training, rather than simply using it as a rendering supervision signal; however, achieving this requires managing the low-frequency bias that presents itself in the measured normals from the active-illumination pipeline.

Acknowledgments

This work was partly supported by EPSRC grant EP/X011364/1 GNOMON.

References

Agisoft LLC 2024. *Metashape Professional (Version 2.1.2) (Software).* Agisoft LLC. https://www.agisoft.com/.

Oleg Alexander, Mike Rogers, William Lambeth, Matt Chiang, and Paul Debevec. 2009. The Digital Emily Project: Photoreal Facial Modeling and Animation. In *ACM SIGGRAPH 2009 Courses* (New Orleans, Louisiana) *(SIGGRAPH '09)*. Association for Computing Machinery, New York, NY, USA, Article 12, 15 pages. https://doi.org/10.1145/1667239.1667251

Jonathan T. Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P. Srinivasan. 2021. Mip-NeRF: A Multiscale Representation for Anti-Aliasing Neural Radiance Fields. *ICCV* (2021).

Thabo Beeler, Bernd Bickel, Paul Beardsley, Bob Sumner, and Markus Gross. 2010. High-quality single-shot capture of facial geometry. *ACM Trans. Graph.* 29, 4, Article 40 (jul 2010), 9 pages. https://doi.org/10.1145/1778765.1778777

Reiner Birkl, Diana Wofk, and Matthias Müller. 2023. MiDaS v3.1 – A Model Zoo for Robust Monocular Relative Depth Estimation. *arXiv preprint arXiv:2307.14460* (2023).

Blender Foundation 2018. *Blender - a 3D modelling and rendering package.* Blender Foundation, Stichting Blender Foundation, Amsterdam. http://www.blender.org

Baptiste Brument, Robin Bruneau, Yvain Quéau, Jean Mélou, François Lauze, Jean-Denis Durou, and Lilian Calvet. 2023. RNb-Neus: Reflectance and normal Based reconstruction with NeuS. arXiv:2312.01215

Egor Burkov, Ruslan Rakhimov, Aleksandr Safin, Evgeny Burnaev, and Victor S. Lempitsky. 2022. Multi-NeuS: 3D Head Portraits From Single Image With Neural Implicit Functions. *IEEE Access* 11 (2022), 95681–95691. https://api.semanticscholar.org/CorpusID:252185485

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. 2022. Authentic volumetric avatars from a phone scan. *ACM Trans. Graph.* 41, 4, Article 163 (jul 2022), 19 pages. https://doi.org/10.1145/3528223.3530143

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. 2021. Efficient Geometry-aware 3D Generative Adversarial Networks. In *arXiv*.

Zhiqin Chen, Andrea Tagliasacchi, Thomas Funkhouser, and Hao Zhang. 2022. Neural dual contouring. *ACM Transactions on Graphics (TOG)* 41, 4 (2022), 1–13.

Zhiqin Chen and Hao Zhang. 2021. Neural marching cubes. *ACM Transactions on Graphics (TOG)* 40, 6 (2021), 1–15.

Paul Debevec, Tim Hawkins, Chris Tchou, Westley Sarokin, and Mark Sagar. 2023. *Acquiring the Reflectance Field of a Human Face* (1 ed.). Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3596711.3596762

Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. 2022. Depth-supervised NeRF: Fewer Views and Faster Training for Free. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

David Eigen, Christian Puhrsch, and Rob Fergus. 2014. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems* 27 (2014).

Guy Gafni, Justus Thies, Michael Zollhöfer, and Matthias Nießner. 2021. Dynamic Neural Radiance Fields for Monocular 4D Facial Avatar Reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 8649–8658.

Abhijeet Ghosh, Graham Fyffe, Borom Tunwattanapong, Jay Busch, Xueming Yu, and Paul Debevec. 2011. Multiview face capture using polarized spherical gradient illumination. *ACM Trans. Graph.* 30, 6 (dec 2011), 1–10. https://doi.org/10.1145/2070781.2024163

Simon Giebenhain, Tobias Kirschstein, Markos Georgopoulos, Martin Rünz, Lourdes Agapito, and Matthias Nießner. 2023. Learning Neural Parametric Head Models. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*.

Simon Giebenhain, Tobias Kirschstein, Markos Georgopoulos, Martin Rünz, Lourdes Agapito, and Matthias Nießner. 2024. MonoNPHM: Dynamic Head Reconstruction from Monocular Videos. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*.

Paulo Gotardo, Jérémy Riviere, Derek Bradley, Abhijeet Ghosh, and Thabo Beeler. 2018. Practical dynamic facial appearance modeling and acquisition. *ACM Trans. Graph.* 37, 6, Article 232 (dec 2018), 13 pages. https://doi.org/10.1145/3272127.3275073

Antoine Guédon and Vincent Lepetit. 2024. SuGaR: Surface-Aligned Gaussian Splatting for Efficient 3D Mesh Reconstruction and High-Quality Mesh Rendering. *CVPR* (2024).

Carlos Hernandez Esteban, George Vogiatzis, and Roberto Cipolla. 2008. Multiview Photometric Stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30, 3 (2008), 548–554. https://doi.org/10.1109/TPAMI.2007.70820

Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2024. 2D Gaussian Splatting for Geometrically Accurate Radiance Fields. In *SIGGRAPH 2024 Conference Papers*. Association for Computing Machinery. https://doi.org/10.1145/3641519.3657428

Zhanghan Ke, Jiayu Sun, Kaican Li, Qiong Yan, and Rynson WH Lau. 2022. Modnet: Real-time trimap-free portrait matting via objective decomposition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 1140–1147.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Transactions on Graphics* 42, 4 (July 2023). https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/

Tobias Kirschstein, Shenhan Qian, Simon Giebenhain, Tim Walter, and Matthias Nießner. 2023. NeRSemble: Multi-View Radiance Field Reconstruction of Human Heads. *ACM Trans. Graph.* 42, 4, Article 161 (jul 2023), 14 pages. https://doi.org/10.1145/3592455

Alexandros Lattas, Yiming Lin, Jayanth Kannan, Ekin Ozturk, Luca Filipi, Giuseppe Claudio Guarnera, Gaurav Chawla, and Abhijeet Ghosh. 2022. Practical and Scalable Desktop-Based High-Quality Facial Capture. In *Computer Vision – ECCV 2022*, Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (Eds.). Springer Nature Switzerland, Cham, 522–537.

Ruilong Li, Hang Gao, Matthew Tancik, and Angjoo Kanazawa. 2023a. Nerfacc: Efficient sampling accelerates nerfs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 18537–18546.

Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. 2017. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)* 36, 6 (2017), 194:1–194:17. https://doi.org/10.1145/3130800.3130813

Zhaoshuo Li, Thomas Müller, Alex Evans, Russell H Taylor, Mathias Unberath, Ming-Yu Liu, and Chen-Hsuan Lin. 2023b. Neuralangelo: High-Fidelity Neural Surface Reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Stephen Lombardi, Jason Saragih, Tomas Simon, and Yaser Sheikh. 2018. Deep Appearance Models for Face Rendering. *ACM Trans. Graph.* 37, 4, Article 68 (July 2018), 13 pages.

Stephen Lombardi, Tomas Simon, Gabriel Schwartz, Michael Zollhoefer, Yaser Sheikh, and Jason Saragih. 2021. Mixture of Volumetric Primitives for Efficient Neural Rendering. *ACM Trans. Graph.* 40, 4, Article 59 (jul 2021), 13 pages. https://doi.org/10.1145/3450626.3459863

Xiaoxiao Long, Cheng Lin, Peng Wang, Taku Komura, and Wenping Wang. 2022. Sparseneus: Fast generalizable neural surface reconstruction from sparse views. In *European Conference on Computer Vision*. Springer, 210–227.

William E. Lorensen and Harvey E. Cline. 1987. Marching cubes: A high resolution 3D surface construction algorithm. In *Proceedings of the 14th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '87)*. Association for Computing Machinery, New York, NY, USA, 163–169. https://doi.org/10.1145/37401.37422

Wan-Chun Ma, Tim Hawkins, Pieter Peers, Charles-Felix Chabert, Malte Weiss, and Paul Debevec. 2007. Rapid acquisition of specular and diffuse normal maps from polarized spherical gradient illumination. In *Proceedings of the 18th Eurographics Conference on Rendering Techniques* (Grenoble, France) *(EGSR'07)*. Eurographics Association, Goslar, DEU, 183–194.

Ricardo Martin-Brualla, Noha Radwan, Mehdi S. M. Sajjadi, Jonathan T. Barron, Alexey Dosovitskiy, and Daniel Duckworth. 2021. NeRF in the Wild: Neural Radiance Fields for Unconstrained Photo Collections. In *CVPR*.

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *ECCV*.

Masahiro Mori. 2012. The Uncanny Valley. *IEEE Robotics & Automation Magazine* 19, 2 (2012), 98–100.

Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. 2022. Instant Neural Graphics Primitives with a Multiresolution Hash Encoding. *ACM Trans. Graph.* 41, 4, Article 102 (July 2022), 15 pages. https://doi.org/10.1145/3528223.3530127

Diego Nehab, Szymon Rusinkiewicz, James Davis, and Ravi Ramamoorthi. 2005. Efficiently combining positions and normals for precise 3D geometry. *ACM Trans. Graph.* 24, 3 (jul 2005), 536–543. https://doi.org/10.1145/1073204.1073226

Michael Niemeyer, Jonathan T. Barron, Ben Mildenhall, Mehdi S. M. Sajjadi, Andreas Geiger, and Noha Radwan. 2022. RegNeRF: Regularizing Neural Radiance Fields

for View Synthesis from Sparse Inputs. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*.

Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M. Seitz. 2021. HyperNeRF: A Higher-Dimensional Representation for Topologically Varying Neural Radiance Fields. *ACM Trans. Graph.* 40, 6, Article 238 (dec 2021).

G. Rainer, L. Bridgeman, and A. Ghosh. 2023. Neural Shading Fields for Efficient Facial Inverse Rendering. *Computer Graphics Forum* 42, 7 (2023), e14943. https://doi.org/10.1111/cgf.14943 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.14943

Pramod Rao, Mallikarjun B R, Gereon Fox, Tim Weyrich, Bernd Bickel, Hans-Peter Seidel, Hanspeter Pfister, Wojciech Matusik, Ayush Tewari, Christian Theobalt, and Mohamed Elgharib. 2022. VoRF: Volumetric Relightable Faces. In *Proceedings of the British Machine Vision Conference (BMVC)*. BMVA Press. selected for oral presentation.

Jérémy Riviere, Paulo Gotardo, Derek Bradley, Abhijeet Ghosh, and Thabo Beeler. 2020. Single-shot high-quality facial geometry and skin appearance capture. *ACM Trans. Graph.* 39, 4, Article 81 (aug 2020), 12 pages. https://doi.org/10.1145/3386569.3392464

Tianchang Shen, Jun Gao, Kangxue Yin, Ming-Yu Liu, and Sanja Fidler. 2021. Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. *Advances in Neural Information Processing Systems* 34 (2021), 6087–6101.

The Wikihuman Project. 2015. *Digital Emily 2*. https://vgl.ict.usc.edu/Data/DigitalEmily2/

Matias Turkulainen, Xuqian Ren, Iaroslav Melekhov, Otto Seiskari, Esa Rahtu, and Juho Kannala. 2024. DN-Splatter: Depth and Normal Priors for Gaussian Splatting and Meshing. arXiv:2403.17822 [cs.CV]

S. Ullman. 1979. The Interpretation of Structure from Motion. *Proceedings of the Royal Society of London Series B* 203, 1153 (Jan. 1979), 405–426. https://doi.org/10.1098/rspb.1979.0006

Sébastien Valette, Jean Marc Chassery, and Rémy Prost. 2008. Generic remeshing of 3D triangular meshes with metric-dependent discrete Voronoi diagrams. *IEEE Transactions on Visualization and Computer Graphics* 14, 2 (2008), 369–381.

Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. 2021. NeuS: Learning Neural Implicit Surfaces by Volume Rendering for Multi-view Reconstruction. *NeurIPS* (2021).

Yiming Wang, Qin Han, Marc Habermann, Kostas Daniilidis, Christian Theobalt, and Lingjie Liu. 2023. NeuS2: Fast Learning of Neural Implicit Surfaces for Multi-view Reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. 2021. pixelNeRF: Neural Radiance Fields from One or Few Images. In *CVPR*.

Zehao Yu, Songyou Peng, Michael Niemeyer, Torsten Sattler, and Andreas Geiger. 2022. MonoSDF: Exploring Monocular Geometric Cues for Neural Implicit Surface Reconstruction. *Advances in Neural Information Processing Systems (NeurIPS)* (2022).

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

Zhengyou Zhang. 2004. Camera calibration with one-dimensional objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 26, 7 (2004), 892–899. https://doi.org/10.1109/TPAMI.2004.21