Analyzing the Input Sensitivity in Neural and Gaussian Splatting Methods for 3D Modelling of Human Head

Masterarbeit im Fach Informatik

vorgelegt von

Mohit Choithwani

geb. am 27.01.1997 in Mumbai, India

angefertigt am

Department Informatik
Lehrstuhl Graphische Datenverarbeitung
Friedrich-Alexander-Universität Erlangen-Nürnberg

Betreuer: M.Sc. Maximilian Weiherer, Prof. Dr. Bernhard Egger

Betreuender Hochschullehrer: Prof. Dr. Marc Stamminger

Beginn der Arbeit: 01.12.2023

Abgabe der Arbeit: 03.06.2024

Abstract

Reconstructing high-quality 3D models from multi-view images remains an ill-posed problem in the field of computer vision. Over recent years, Neural Radiance Field (NeRF)[46] has emerged as a leading approach for addressing various computer vision tasks, including image-based 3D reconstruction. NeRF has demonstrated superior results in both novel view synthesis and 3D reconstruction of scenes/objects. Image-based neural rendering methods for surface reconstruction give considerable advantages over traditional surface reconstruction methods, producing highly detailed 3D models. However, these methods typically rely on a high-quality and dense collection of input images for effective 3D reconstruction. In particular, the sparsity of views in input images significantly constrains the high-fidelity reconstruction capabilities of neural rendering methods. Obtaining such high-quality and dense data is not always feasible in real-world settings, thereby limiting their practical applicability. Meanwhile, Gaussian splatting [36] methods offer a significant advantage in terms of rendering speed because they efficiently use GPUs, which ensures rapid performance in novel view synthesis. However, methods based on NeRF and Gaussian splatting often rely on assumptions about input parameters, such as the need for precise camera poses and specific lighting conditions. These requirements can make them less practical for real-world applications. In this work, we systematically examine the influence of these parameters on the quality of 3D reconstruction models. By evaluating these methods under difficult conditions, we aim to determine robust 3D surface reconstruction methods that can deliver fair results even with imperfect input conditions. We believe our findings will encourage more researchers to work further on NeRF-based and Gaussian splatting-based methods in terms of robustness which will help in future innovations in 3D reconstruction, thereby making the methods more versatile and efficient in diverse real-world applications.

Acknowledgement

Firstly, I would like to express my sincere gratitude to my thesis advisor, Maximilian Weiherer, for his unwavering support throughout my master's thesis, especially during challenging times. I am also sincerely grateful to Prof. Dr. Bernhard Egger for providing me with the opportunity to write this thesis and for suggesting the initial idea. Additionally, I extend my heartfelt gratitude to all the members of the Chair of Visual Computing for their valuable feedback during interim presentations. I appreciate the access to the LGDV lab and the quick assistance from the admins with any installation issues. Lastly, I extend my sincere gratitude to the Erlangen National High-Performance Computing Center (NHR) at FAU for providing valuable HPC resources.

Contents

Chapter 1

Introduction

There have been remarkable advancements in the field of 3D reconstruction technology within the last few years. The concept of constructing 3D models from images or video has become a cornerstone technology in various fields, such as augmented reality, virtual reality, biometric identification, medical imaging, animation, and the film industry [81]. These advancements in 3D modelling hold the potential to transform several industries, such as personalized shopping with a single image for sunglasses or jewellery where a user could 'wear' it on their face through face reconstruction. By simply taking a picture of their face, users can get a preview of how exactly those products will look on them, making online shopping personalized and convenient.

Briefly, 3D reconstruction is the process of creating a virtual 3D representation of an object from videos or images by extracting the 3D information, such as structure and geometric details, from the 2D input data. Specifically, images-based 3D reconstruction of the human head focuses on accurately reconstructing the 3D shape and features of a face from a single or multiple facial images. This process can be formulated as follows: With a set of 2D images, $\mathbf{I} = \{\mathbf{I}_m \mid m = 1, \ldots, n\}$, the goal is to reconstruct the surface of the object $\hat{\mathbf{X}}$ that closely matches the actual 3D structure $\mathbf{X}$. Figure 1.1 illustrates the common workflow of the reconstruction process using multiple images. This process involves capturing several images from different viewpoints using a camera. These images, along with supplementary inputs, are then processed by the 3D reconstruction method, which ultimately produces the final 3D model. Traditional 3D reconstruction approaches, including statistical model fitting and photogrammetry, have been widely used for accurate reconstructions. Statistical model fitting adjusts the parameters that define the shape, texture and pose of the 3D model such that the difference between the input images and the rendered projections of the 3D model should be minimal. At the same time, photogrammetry reconstructs 3D geometry by analyzing the differences in perspective and relative positions of features across multiple 2D images. Although these methods generate high-quality results, they often require extensive preprocessing or controlled environmental settings. Despite their effectiveness, these methods are computationally intensive and fail to capture fine details of the face, such as wrinkles.

In the last four years, Neural Radiance Fields [46] have posed a ground-breaking paradigm shift for the novel view synthesis and 3D reconstruction tasks, leading to significant advance-

ments in computer graphics and computer vision. NeRF was proposed by Mildenhall et al. in 2020, which quickly became a popular research direction. NeRF utilizes an MLP to encode the 3D scene by mapping the spatial location and viewing direction to color and density values. Unlike conventional methods that discretize 3D space, NeRF employs implicit representations to describe scenes or objects. Building on these advances in neural implicit representations, neural surface reconstruction methods have shown promising results in producing high-quality metric 3D reconstructions. Methods such as NeuS [64], VolSDF [75], and UniSurf [50] leverage neural rendering to jointly optimize implicit geometry and the radiance field by minimizing the difference between rendered views and ground-truth views. Apart from neural approaches, recently, Gaussian Splatting [36] has become very popular as it yields realistic rendering while being significantly faster to train than NeRFs, establishing it as a compelling alternative to neural approaches. One of the primary advantages of neural approaches over traditional 3D reconstruction methods is their capability for end-to-end learning, allowing the entire reconstruction pipeline to be optimized jointly, thereby eliminating the need for multiple stages required in traditional methods, such as key point detection and feature mapping.

NeRF methods have two principal drawbacks despite their ability to generate convincing geometry and photorealistic novel views. Firstly, they depend heavily on a large number of high-quality images with enough overlap between the input images. Moreover, they often rely on strong assumptions about their input. For instance, many methods [64, 50] assume that the camera poses are accurate and that the input images are clean and not blurred. Moreover, it becomes more challenging when only a sparse set of input images is available, and existing neural rendering approaches typically yield incomplete or distorted results. In many practical scenarios, where data is often corrupted, obtaining such a comprehensive set of input images with accurate camera calibration is neither feasible nor readily available. Secondly, many NeRF-based methods are computationally expensive, utilizing high-end GPU computation power and extensive training time. For example, training a classic NeRF [46] model on a scene with RTX 3080 GPU can take several hours, making real-time application impractical. These limitations restrict the applicability of neural rendering methods for real-world scenarios where the input data may be noisy, incomplete, or distorted [15].

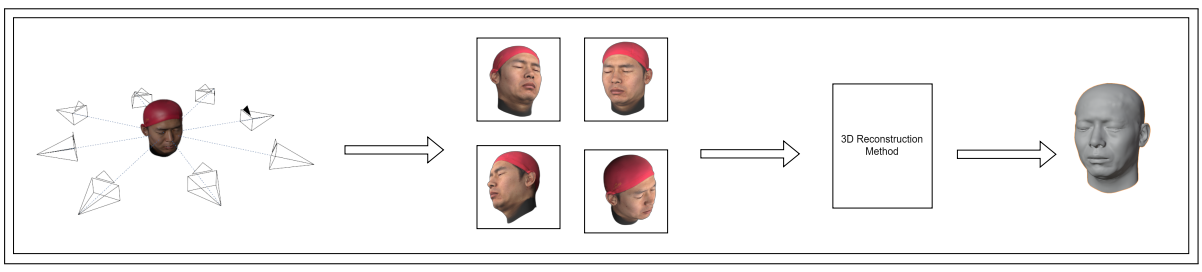


Figure 1.1: Multi-view Image Based 3D Reconstruction.

1.1 Motivation

Our aim is to investigate the potential of cutting-edge surface reconstruction methods, primarily focusing on neural approaches and with a secondary emphasis on Gaussian splatting techniques in handling input variability. Specifically, it investigates how variations in the quantity and quality of input images affect the quality of reconstructed 3D face models. The fundamental idea for this research comes from the aforementioned questions: To what extent do poor quality input images, including sparse or noisy data, affect the quality of reconstructed 3D models and analyze the robustness challenges of current SOTA methods in surface reconstruction, which is crucial for enhancing the practicality and reliability of these SOTA methods in real-world settings where accurate input data cannot be guaranteed.

In real-life situations, it is common to capture sub-optimal images due to various factors such as poor lighting, low resolution, and the complexity of the scene. Therefore, understanding the influence of these factors on the reconstruction process is essential. This thesis seeks to determine the minimum quality and quantity of input data necessary to achieve accurate 3D face reconstructions using surface reconstruction methods based on neural radiance and Gaussian splatting. By systematically analyzing these parameters, our goal is to pinpoint the threshold at which reconstruction quality significantly degrades for selected SOTA methods and determine the methods that allow for high-fidelity reconstructions despite suboptimal inputs.

This research aims to contribute to the field of 3D modelling using images by providing a sufficient understanding of the abilities and limitations of SOTA 3D reconstruction methods based on neural radiance and Gaussian splatting techniques under real-world constraints. The findings from this work will serve as a guide for future research and applications, leading to the development of more robust solutions that can reliably handle imperfect input data for reconstruction.

In summary, the motivation behind this thesis is to analyze the robustness and effectiveness of SOTA surface reconstruction methods for 3D modelling of human heads under various circumstances, including sparse and imperfect inputs. This analysis involves evaluating the reconstruction quality of these SOTA methods to determine the most effective methods for 3D reconstruction in suboptimal circumstances.

1.2 Contributions

In this work, we thoroughly examine how state-of-the-art surface reconstruction methods, including neural radiance and Gaussian-splatting techniques, perform well in 3D human head modelling in challenging conditions. This work attempts to fill the gap of insufficient research on how these novel methods' performance changes with varying input quality and quantity. Thus, the main focus is on measuring how a diverse set of non-optimal inputs, such as sparse input images (e.g., only three images), perturbed poses, and poor lighting conditions, affect the performance of novel 3D reconstruction methods.

To summaryize, the key contributions of our work include:

- **Performance Evaluation with Sparse and Dense Images:** We systematically evaluate the performance of SOTA methods when provided when presented with a varying number of input images, ranging from sparse (as few as three images) to dense (as many as 100 images). Our findings reveal that all the methods that are evaluated demonstrate a decrease in reconstruction accuracy when trained and rendered with sparse input images compared to a dense number of images. Additionally, our results indicate that for most methods, there is no significant improvement in the quality of the reconstructed human head when using more than 15 images.
- **Robustness to Inaccurate Camera Poses:** We assess the robustness of SOTA methods against inaccuracies in camera poses by reconstructing 3D models of the head using imprecise poses. Our findings indicate that the imperfect camera poses negatively impact the performance of all tested methods, resulting in reconstructions that are less precise and lack fine details.
- **Performance Under Sub-optimal Lighting Conditions:** We investigate how low lighting conditions of the surroundings in which the images are captured affect the reconstruction performance of SOTA methods. Our results show a significant drop in performance when lighting is sub-optimal, leading to poorer reconstruction quality compared to well-lit conditions. This is primarily because low lighting makes it harder to capture accurate surface details and textures.
- **Determining of Robust Methods:** We performed a thorough assessment of the SOTA methods for 3D reconstruction by testing them in both optimal and substandard conditions. Through a comparative analysis of the resulting 3D reconstruction, we identified methods that demonstrate robust performance even in sub-optimal conditions, highlighting their reliability and effectiveness in practical applications.

1.3 Thesis Structure

This thesis is divided into the following chapters. Chapter 2 provides an overview of related work and recent advancements in 3D surface reconstruction. Chapter 3 delves into the background of Neural Radiance Fields. Chapter 4 describes the methodology utilized in the experiments. Chapter 5 presents the quantitative and qualitative results of all experiments. Chapter 6 discusses the limitations and proposes future research directions. And Chapter 7 concludes the thesis.

Chapter 2

Related Work

Reconstructing realistic human face models from multi-view RGB images is an ill-posed problem in computer vision and computer graphics [30]. Unlike reconstructing a rigid object, reconstructing a facial model of a human is considerably more challenging due to the high variability in poses, expressions, and textures, which makes it difficult to capture the 3D geometry accurately. Additionally, facial features involve complex structures and fine-level details that demand advanced techniques for precise reconstruction. Various methods are used for modelling 3D head and facial geometry, such as:

2.1 3D Morphable Model-based Reconstruction

To simplify the inherent complexity of the 3D facial geometry reconstruction, in 1992, Blanz and Vetter proposed a statistical model known as the 3D Morphable Model (3DMM) [7]. 3DMM represents the human face by decomposing the face attributes into low-dimensional representations. This helps to capture the variations in facial shape and texture, allowing for an accurate reconstruction of facial geometry from 2D images. Paysan et al. work [51] expanded the statistical model by offering much better accuracy in shape and texture, leading to the Basel Face Model (BFM), the first publicly available Morphable Model. This improvement was achieved through the use of a superior scanning device and the reduction of correspondence artifacts, enabled by an advanced registration algorithm. The use of 3DMM provides a detailed representation of the human face, allowing for more precise guidelines in face reconstruction and simplifying the process of face morphing. Further, the generative abilities of 3DMM [37, 24] enable the generation of geometrically consistent faces from within the model's framework [17]. Notably, ControlFace [57], trained on SynthFace, has achieved state-of-the-art results on the NoW benchmark without the need for 3D supervision. However, its focus is on shape, excluding expressive variations and biases within the dataset that may influence the performance of 3D reconstruction.

2.2 Modern Deep Learning-based Methods

Deep learning-based methods have shown remarkable performance across various computer vision problems, including 3D reconstruction. CNN-based architectures are extensively used for modelling 3D shapes, capable of reconstructing 3D facial geometry of humans from sparse multi-view images [27, 11, 69] and even from a single image [12, 65]. Combining 3D-CNN with techniques such as stereo matching enables the model to infer depth from the images, thereby helping in high-fidelity model reconstruction. However, the process of training a CNN-based model is computationally expensive and time-intensive [33]. To address this issue, a method proposed by Guo et al.[27] which operates on cascaded CNN models, with an initial coarse-layer CNN and a subsequent fine-layer CNN, to reconstruct detailed 3D facial surfaces from a single image. Despite this advancement, these methods require large datasets for training and often face challenges in generalizing to unseen environments.

The generative adversarial network (GAN) [25] proposed by Goodfellow et al. in 2014 has received significant attention for generating realistic images, beneficial for 3D reconstruction tasks. A GAN consists of a generator that creates realistic images and a discriminator that distinguishes these generated images from the real ones. Building on this framework, GAN-FIT [24] leverages GANs and DCNNs to create high-fidelity 3D models from limited input images by training a powerful GAN-based facial texture generator and employing nonlinear optimization to find optimal latent parameters of the 3DMM. However, GAN requires a large volume of training data for optimal performance. The training process itself is challenging, requiring significant computational resources, and they often experience stability issues due to their sensitivity to parameter choices. Additionally, 3DGANs suffer from mode collapse, which limits the diversity of the outputs [1]. To address stability issues and mode collapse, Arjovsky et al. introduced the Wasserstein GAN (WGAN) [2], which utilizes a special loss function that improves optimization stability. Along with this, Moschoglou et al. implemented an autoencoder-based GAN, 3DFaceGAN [47], to model 3D faces, combining reconstruction and adversarial loss with training the generator and discriminator, while retaining high-frequency details. Despite these advancements, the challenging training process and computational demands of GANs limit their applicability to real-time 3D face reconstruction.

2.3 Learning-based 3D Reconstruction

Learning-based methods use advanced machine learning techniques to directly infer complex mappings from the 2D input images to the resulting 3D representation, thus eliminating the traditional MVS pipeline.

Based on the output representation they use, these methods are broadly categorized as follows (see Fig. 2.1):

2.3.1 Voxel

Voxel models utilize small cubes (voxel grids) to represent the object's shape and colour. In the past, voxel-based methods, like 3D-R3N2 [13], employed 3D convolutional encoders and

decoders to predict 32^2 voxel grids from multi-view images. However, voxel-based methods were limited by high memory and computation demands, restricting them to small voxel grids (up to 128^3) and causing difficulties in representing large scenes and complex objects.

2.3.2 Point Cloud Representation

Point clouds represent the object's surface in 3D space using a collection of points, requiring less memory and resulting in smoother shapes. Qi et al. proposed PointNet [53], a deep learning framework that leverages point clouds while maintaining permutation invariance and robustness to input perturbations and corruption. Fan et al. [18] suggested using point clouds for 3D reconstruction from a single image, addressing ground truth ambiguity with a conditional shape sampler that predicts multiple plausible 3D point clouds. However, point cloud-based methods need non-trivial post-processing steps to create final 3D meshes, making them less commonly used.

2.3.3 Mesh Representation

A mesh-based method describes the surface of the 3D object using polygons (triangles and quadrilaterals). These meshes only capture the surface geometry and not the object's volume. Methods such as Pixel2Mesh [63] achieve accurate geometry reconstruction by progressively deforming the vertices of an ellipsoid mesh, utilizing perceptual features extracted from the input image. While mesh-based methods generate ready-to-use mesh models and are effective, they are computationally expensive and require significant processing time.

2.3.4 Neural Representation

Neural representations utilize neural networks to implicitly encode the 3D geometry without spatial discretizing. Building on this concept, techniques like Neural Radiance Fields [46] model complex scenes by learning radiance fields directly from images.

2.4 3D Reconstruction using NeRF-Based Methods and Key Challenges

In recent years, NVS and 3D reconstruction have seen significant progress with the rise of 3DGS [36] and NeRF [46]. Several survey papers [16, 71, 21, 54, 19] have been published to highlight the rapid advances in NeRF and 3DGS, offering a comprehensive overview of their evolution and multidimensional understanding of these technology.

The introduction of NeRF [46] has revolutionized 3D reconstruction and view synthesis. By utilizing an MLP to encode a 3D scene, the MLP maps 3D spatial location to color and volume density. Using implicit 3D representation and classical volume rendering techniques enables the synthesis of novel views of the scene from any direction. The advanced view synthesis capabilities reduce visual distortions, outperforming traditional methods such as COLMAP [58]. NeRF's impressive results in novel view synthesis have led to numerous

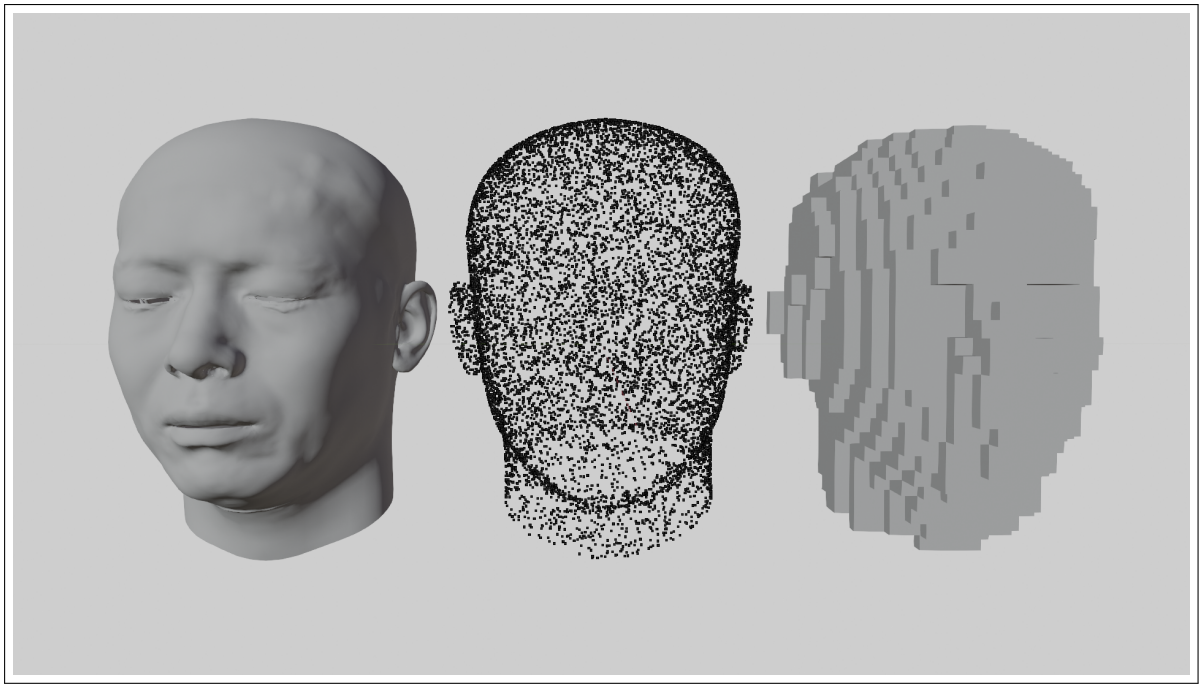


Figure 2.1: 3D face model representations: Mesh-based (left), point cloud (center), and voxel-based (right).

improvements that focus on improving NeRF in terms of image quality [4], robustness [39, 49], training speed [48] and rendering speed [28, 10]. However, these enhancements are often optimized for view synthesis rather than accurate geometry reconstruction. NeRF and its variants [80, 49, 22] can capture the physical characteristics of the scene, such as shape, material and texture, and generate photorealistic images. However, they face challenges in accurately defining isosurfaces of the volume density to represent the underlying 3D geometry of the scene. This has led to the development of various NeRF variants that focus on methods specifically designed to extract precise 3D meshes.

Previous studies [29, 34] in deep neural networks have shown that neural network models are sensitive to input data. Various benchmarks, such as ImageNet-C [29] and ImageNet-V2 [56], have been developed to assess the robustness of image classifiers against simulated corruptions and natural distribution shifts. Adversarial training is a standard method to enhance models' robustness against corruption. For instance, Xie et al. [70] demonstrate that adversarial perturbations can augment data and improve image classification accuracy. Data augmentation techniques like Mixup [79] and CutMix [78] significantly boost generalization performance; however, they are not suitable for NeRF because they disrupt the necessary spatial consistency and view-dependent effects, which are required for accurate 3D reconstruction.

Neural rendering techniques typically assume that input data is extensive and high-quality, including precise camera positions and consistent lighting across perspectives. When these specified conditions are not satisfied, the resulting 3D representation suffers from severe per-

formance degradation. [61] in their work has demonstrated that SOTA methods for novel view synthesis experience a significant performance decline in the presence of corruptions such as Gaussian noise and fog. Interestingly, not all types of corruption affect performance equally; for example, JPEG compression results in a 15% performance drop, whereas fog causes a 50% drop in quality. Ensuring the robustness of 3D reconstruction models, particularly those leveraging NeRF, is critical for their deployment in real-world applications and, as it ensures the models' reliability under diverse and challenging conditions.

2.4.1 Methods Optimized for Sparse Views

MixNeRF [59] aims to synthesize novel views from sparse input images by predicting the joint distribution of RGB colors across the ray samples using a mixture of distributions. It introduces an additional element of estimating light depth associated with the geometric height of the 3D scene in the model structure, helping to better understand the geometry of the scene. PixelNeRF [76] can accurately predict the 3D representation of multiple objects from one or several images without requiring additional supervision or optimization. It also showcases the robustness in handling larger viewpoint changes and more complex scenes. SparseNeRF [62] addresses the challenge of synthesizing new views from limited perspectives by extracting depth priors from pre-trained depth models. This improves depth accuracy and maintains spatial continuity, enhancing the performance of new view synthesis. SparseNeuS [40] uses an SDF for surface representation and extracts priors from image features through geometry encoding volumes. This method enables high-quality 3D reconstruction from sparse images, with just two or three, by applying these priors across various scenes. Deep 3D Capture [5] learns geometry and material properties from sparse images using a deep multi-view stereo network. It estimates and aligns per-view depth maps to construct high-quality 3D geometry but requires specialized illumination rigs, limiting its practical applicability. Many approaches require supplementary supervision to work in sparse settings. For example, DietNeRF [32] is able to train with fewer input images but requires precise camera positioning and semantic supervision.

2.4.2 Methods Optimized for Noisy Poses and Sub-Optimal Lighting

Several advanced methods have been developed to address the challenges of inconsistent and low-lighting images. NeRF in the Wild [43] trains NeRF on low-quality tourist images characterized by inconsistent lighting, enabling it to adapt and produce realistic renderings. Meanwhile, NeRF in the Dark [45], and HDR-NeRF [31] focus on generating high dynamic range (HDR) images from noisy, low-light conditions, enhancing the quality of images captured in suboptimal lighting environments. To increase the robustness of NeRF against inaccurate poses, Aug-NeRF [9] introduces a triple-level augmentation training pipeline designed to handle noisy inputs. FreeNeRF enhances NeRF rendering performance with fewer samples through frequency regularization, using frequency range regularization and near/camera density field penalties to achieve results comparable to more complex methods without additional pretraining or supervision. Mip-NeRF improves standard NeRF volume rendering

by employing cone tracing and integrated positional encoding, effectively addressing aliasing and blurring to enhance rendering quality. Deblur-NeRF [42] is a model specifically tailored to address the challenges presented by blurred images. Meanwhile, Nan [52] is designed to effectively mitigate the impact of burst noise and ensure accurate and reliable output. All the above methods assume the availability of precise camera poses. To overcome this issue, BARF [39] adopts a more flexible methodology by relaxing the need for exact camera positions. However, it requires a dense set of training images for high-quality 3D reconstruction.

2.4.3 Methods Optimized for Surface Reconstruction

Various NeRF-based approaches are tailored to reconstruct the 3D surface of the object. Neural RGB-D [3] adapt NeRF by incorporating a truncated signed distance function to leverage depth information for capturing detailed surfaces, allowing high-quality 3D reconstructions. Wang et al. proposed NeuS[64], which represents an object's surface as the zero-level set of an SDF and introduces a new volume rendering method to train this SDF representation. By eliminating first-order bias which is inherent in traditional volume rendering methods, NeuS achieves accurate surface reconstruction without mask supervision. VolSDF [75] defines the volume density function as the cumulative form of a Laplace distribution applied to an SDF representation, which provides a beneficial inductive bias for the geometry. Also, it allows efficient unsupervised disentanglement of shape and appearance in volume rendering for accurate geometry reconstruction. Building upon VolSDF's [75] hybrid volume-surface neural representation, BakedSDF [74] extends its application to unbounded real-world scenes. However, the training and rendering time for many methods is significantly large. To tackle this issue, InstantNGP [48] introduces a hybrid 3D grid structure with multi-resolution hash encoding combined with an MLP to accelerate the training and rendering of NeRFs. This advancement makes neural rendering more practical and brings it closer to real-world applications. Rakotosaona at el. [55] proposes a novel method that enables easy 3D surface reconstruction from any NeRF-based approach. After training the radiance field, the volumetric 3D representation is transformed into a Signed Surface Approximation, enabling easy extraction of the 3D mesh and its appearance. Neuralangelo [38] uses numerical gradients and multi-resolution to reconstruct detailed scenes. It effectively recovers dense 3D surface structures from multi-view images with high fidelity, even without auxiliary inputs like depth maps, surpassing previous methods. NeuS2 [67] achieves two orders of magnitude acceleration without sacrificing reconstruction quality by utilizing multi-resolution hash encodings and lightweight second-order derivative calculations with CUDA parallelism. Plenoxels [20] propose a multi-resolution representation using a view-dependent sparse voxel grid that stores opacity and spherical harmonic coefficients. This approach optimizes calibrated images via gradient methods and regularization without neural networks.

2.5 Methods Based on Gaussian Splatting

Gaussian Splatting [36] introduces a new paradigm for neural rendering by encoding scenes as collections of 3D Gaussian splats, which are splatted onto the image plane for rendering.

Gaussian splats are a collection of spatially distributed particles; instead of representing a 3D scene with polygons or distance fields, This method represents a 3D scene through millions of particles, each characterized by several attributes such as each particle, represented as a 3D Gaussian, has a specific position, rotation, and non-uniform scale in 3D space. Additionally, each particle possesses an opacity value and a color defined by third-order Spherical Harmonics coefficients, allowing the perceived color to vary based on the view direction. By representing scenes using 3D points, covariances, color, and opacity, Gaussian Splatting offers a significant advantage in rendering speed. The efficient utilization of GPUs ensures rapid performance, making this method perfect for lengthy computer vision tasks [19]. Based on rasterization techniques, 3D Gaussian splatting (3DGS) learns a point cloud of splats from input images, which allows for novel perspective views by placing the camera at any desired position.

However, faster processing speed comes at the cost of increased storage requirements. Due to the large number of splats it generates, 3DGS requires ten times more memory than NeRF. Another key challenge in Gaussian Splatting is extracting an object's 3D geometry from millions of unorganized 3D Gaussian splats. SuGAR [26] addresses this by regularizing 3D Gaussians and using Poisson reconstruction to extract high-quality meshes from the splats, providing a robust solution to this complex problem. Darmon et al. [15] further advance this field by addressing camera pose distortion and motion blur, modelling motion blur as a Gaussian distribution over camera poses. This approach integrates seamlessly with the probabilistic formulation of 3DGS, maintaining training efficiency and rendering speed. Framework proposed by [72] incorporates structure priors through techniques like visual hull and floater elimination and utilizes a diffusion model-based Gaussian repair process to enhance rendering quality, achieving photorealistic outputs and reconstructing high-quality 3D objects from sparse views.

Despite the advancements in NeRF-based and Gaussian splatting-based methods, their robustness under imperfect input conditions remains an open challenge. Both approaches typically assume near-perfect input parameters, which is rarely true in real-world settings. The need for high-fidelity data and particular environmental factors highlights an essential area for further research in developing robust 3D reconstruction methods that can perform well under suboptimal circumstances. Our work aims to address this gap by systematically evaluating the impact of various input parameters on the quality of 3D reconstruction by utilizing SOTA surface reconstruction methods. Our research focuses on identifying robust 3D surface reconstruction methods that maintain high performance even with less-than-ideal input data.

Chapter 3

Background

This chapter focuses on Neural Radiation Fields [45], which dominate image-based 3D reconstruction. While Gaussian splatting [36] offers speed advantages, NeRF's detailed and accurate outputs highlight its primary role over other methods.

3.1 NeRF

Neural Radiance Fields [46] stand out as the leading architecture in the realm of neural fields. Initially proposed to tackle the challenge of view synthesis, aiming to generate novel photo-realistic views of a 3D object or scene from a set of images captured from various viewpoints. The tasks of view synthesis and 3D reconstruction are closely related, as both involve creating 3D representations from 2D images.

3.1.1 NeRF Architecture

NeRF represents a 3D scene or object as a continuous volumetric function, which accepts a single continuous 5D coordinate as input, comprising a 3D position $\mathbf{x} = (x, y, z)$ and a viewing direction $\mathbf{d} = (\theta, \phi)$. This input is fed into an MLP, which outputs the corresponding radiance $\mathbf{c} = (r, g, b)$ and volume density σ of the scene.

Formally, NeRF can be parameterized as an MLP:

$$f : (\mathbf{x}, \mathbf{d}) \rightarrow (\mathbf{c}, \sigma)$$

Training NeRF involves challenges due to the unknown scene density and color. A differentiable approach is utilized to map the output back to 2D images, which are then compared to ground-truth images. This comparison formulates a rendering loss, which is used to optimize the network.

3.1.2 Volumetric Rendering

The mapping of the neural field output into 2D images is achieved through volume rendering. This involves casting rays from each pixel of the original images and sampling points along

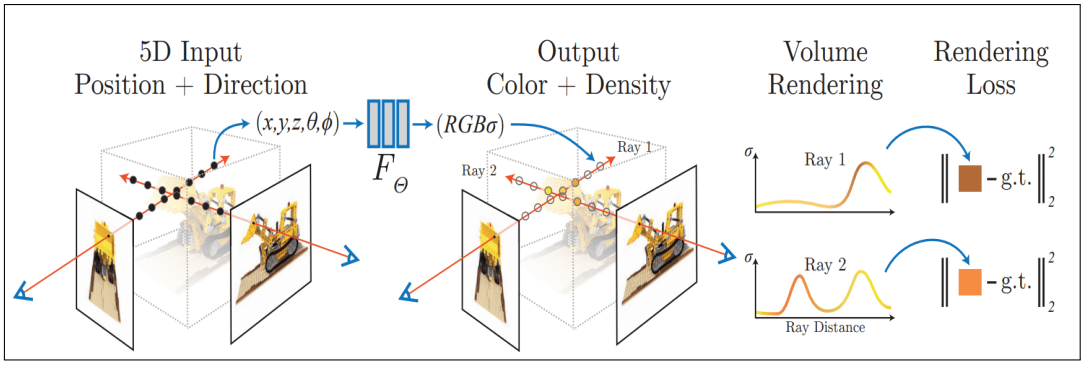


Figure 3.1: NeRF Training and Rendering Process [46].

these rays. Each sampled point is defined by its spatial location, color, and volume density, which are used as inputs to the neural field.

A ray can be defined by its origin $\mathbf{o}$, direction $\mathbf{d}$, and samples at timesteps t:

$$\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$$

The color on the ray $C(r)$ can be expressed by integrating the emitted radiance and the accumulated transmittance along the ray path from the near point tn to the far point tf:

$$C(r) = \int_{t_n}^{t_f} T(t)\sigma(r(t))c(r(t), \mathbf{d})\, dt$$

Where:

- $T(t)$: Cumulative transmittance from the ray starting point t_n to t, representing the probability that the ray travels without being scattered. Transmittance $T(t)$ is mathematically expressed as:

$$T(t) = \exp\left(-\int_{t_n}^{t} \sigma(r(s))\, ds\right)$$

- $\sigma(r(t))$: Volume density at point $r(t)$.
- $c(r(t), \mathbf{d})$: Color at point $r(t)$ in direction $\mathbf{d}$.

3.1.3 Sampling

A stratified sampling technique is employed to uniformly sample the points along the ray, as MLP can only query a finite number of points. This sampling technique improves the approximation accuracy without requiring an excessively large number of samples. First, it divides the region $[t_n, t_f]$ which represents the near and far bounds of the ray within the scene, is divided into N uniform sub-intervals. Then, within each sub-interval, a point is sampled from a uniform distribution. The formula for sampling a point within each sub-interval is expressed as:

$$t_i \sim U\left[t_n + \frac{i-1}{N}(t_f - t_n),\, t_n + \frac{i}{N}(t_f - t_n)\right]$$

Here, $U[a, b]$ denotes a uniform distribution between a and b.

Once the sampling points are determined, the integral can be approximated as a discrete sum. The color $\hat{C}(r)$ along the ray is approximated by adding the contributions from sampled points and can be formulated as:

$$\hat{C}(r) = \sum_{i=1}^{N} T_i \alpha_i c_i$$

- The cumulative transmittance T_i is given by:

$$T_i = \exp\left(-\sum_{j=1}^{i-1} \sigma_j \delta_j\right)$$

- The weighting factor α_i indicates the contribution of the i-th sample to the final color. It is calculated as:

$$\alpha_i = 1 - \exp(-\sigma_i \delta_i)$$

Where δ_i is the distance between nearby sampling points.

3.1.4 Position Encoding

NeRF uses positional encoding to map the 3D coordinates and viewing directions to a higher-dimensional space, which allows to capture high-frequency details of the scene. The position encoding function, denoted as $\gamma(v)$, applies a series of sine and cosine transformations to the input. Here is the mathematical representation:

$$\gamma(v) = (\sin(2^0 \pi v), \cos(2^0 \pi v), \sin(2^1 \pi v), \cos(2^1 \pi v), \ldots, \sin(2^{L-1} \pi v), \cos(2^{L-1} \pi v))$$

Where

- L is the number of frequency bands. Higher the value enables the network to capture finer details.

3.1.5 Multi-Stage Sampling and Loss Function

To improve efficiency, a multi-stage voxel sampling technique, ranging from coarse to fine, is employed. This method ensures a more efficient distribution of samples and enhances rendering quality by concentrating computational efforts on significant areas.

Coarse Sampling: Initially, a set of sparse sampling points is uniformly distributed across the scene. These points are processed by the coarse network to estimate the initial

rendering outcomes. Each sampling point is assigned a weight based on its significance. These weights are normalized to form a PDF, guiding the distribution of fine sampling points.

Fine Sampling: Using the PDF obtained from coarse sampling, additional fine sampling points are selected using the Inverse Transform Sampling method. These points are primarily concentrated in regions likely to contain significant content. The fine network then evaluates both coarse and fine sampled points to obtain detailed color and volume density information. The final rendered color $\hat{C}_f(r)$ of the ray is calculated using all $N_c + N_f$ sample points:

$$\hat{C}_f(r) = \sum_{i=1}^{N_c+N_f} (\hat{w}_i \cdot c_i)$$

where $\hat{w}_i$ is the normalized weight, and c_i is the color of the sampled point.

Loss Function: The loss function used to optimize the NeRF network incorporates both the coarse and fine rendering results. The goal is to minimize the total squared error between the predicted colors and the true pixel colors. The loss function L is given as:

$$L = \sum_{r \in R} \left[\|\hat{C}_c(r) - C(r)\|_2^2 + \|\hat{C}_f(r) - C(r)\|_2^2 \right]$$

where:

- R represents the collection of rays in each batch.
- $C(r)$ describes the true RGB color of ray r.
- $\hat{C}_c(r)$ and $\hat{C}_f(r)$ are the coarse and fine volume-predicted colors of ray r, respectively.

The coarse network's output is optimized to ensure that its weight distribution helps in effectively distributing samples for the fine network. This two-stage process ensures that computational resources are focused on the most significant parts of the scene, enhancing the overall rendering quality.

In summary, within the NeRF structure, the methodology for creating novel perspectives can be seen as:

- NeRF generates novel views by leveraging a continous 3D representation learned from multi-view 2D images. For each pixel in the target image, a light ray is cast through the pixel using camera parameters and scene geometry. Along each ray, spatial points are systematically sampled.
- For each sampled point, an MLP network predicts the color and density value based on the spatial coordinates and viewing direction of that point.
- The color and density values of all sampled points are integrated using volume rendering techniques to synthesize the final pixel color, ensuring high-fiedelity image reconstruction.
- This whole process is optimized by a loss function that minimizes the error between the rendered images and ground-truth images.

Chapter 4

Methodology

4.1 Overview

The concept of 3D reconstruction is based on transforming 2D visual information into 3D spatial representations. It is often quite challenging to obtain high-fidelity and accurate 3D reconstructions depending on various factors, such as the object's surface properties and texture, the quality and quantity of the input images used for the reconstruction process, and the algorithm employed for the reconstruction. Capturing images for 3D reconstruction in real-world settings can be particularly challenging due to a multitude of factors. These include variations in lighting conditions, precise camera calibration, the inherent complexity of the scenes being captured, as well as other environmental and technical aspects. These variations of input have a significant impact on the accuracy of reconstructed 3D models, often leading to distortions and artifacts in the resultant 3D model. When it comes to reconstructing 3D facial models, these challenges intensify as the goal shifts towards generating realistic and high-fidelity 3D facial models. Considering these challenges, the overall approach of our work has been strategically planned to provide an in-depth analysis of the following question: How do variations in input data influence the outcomes of the resulting 3D face model when using the SOTA learning-based surface reconstruction techniques?

This section is structured as follows: Sec 4.2 discusses the input assessment criteria, including the impact of image quantity, camera calibration parameters, and lighting conditions. Sec 4.3 talks about the selection of state-of-the-art methods. Sec 4.4 addresses the dataset used and Blender setup for rendering the images which are used for training the models. Sec 4.5 discusses the overview of the data used in all the experiments. Sec 4.6 describes the sampling algorithm used in our approach. Sec 4.7 focuses on the evaluation metric utilized to assess the performance of the 3D reconstruction.

4.2 Input Assessment Criteria

To thoroughly understand the influence of input factors on the quality of 3D reconstruction, we define three key criteria for evaluation, which are as follows:

- **Impact of image quantity:** By utilizing SOTA methods, we perform a comprehensive set of experiments covering a wide range of image densities, from sparsely populated to densely packed settings. This analysis aims to determine the optimal minimum number of images required for reconstructing a high-quality 3D facial model. Understanding the correlation between the number of input images and the performance of the reconstructed 3D model is essential. This relationship directly influences the practicality and efficiency of the reconstruction process in real-world scenarios. We know that more images generally lead to better reconstruction quality. However, acquiring more images is only sometimes practical. Therefore, exploring the trade-offs between image quantity and reconstruction quality is crucial. Through our analysis, we aim to understand better how minimal image inputs can lead to high-quality reconstructions. This understanding can be leveraged to achieve high-quality reconstructions without significantly increasing the costs associated with data acquisition. Furthermore, this analysis considers the influence of different input image densities on the performance of various learning-based reconstruction methods, thereby providing a comprehensive comparison of SOTA methods on the 3D face reconstruction process from a sparse to a dense setting. Additionally, this study seeks to determine the most effective learning-based method for achieving optimal results with the sparse set of input images.

- **Influence of camera calibration:** Firstly, we address camera calibration, specifically reference object-based calibration. This process involves observing a calibration object/subject whose geometry in 3D space is known with high precision. To accurately assess the influence of inaccuracies in camera calibration on the performance of 3D reconstruction, we incorporate controlled perturbations into the known camera poses. Camera poses consist of intrinsic parameters (focal length, principal point, and lens distortion) and extrinsic parameters (position and orientation). By introducing Gaussian noise in the extrinsic parameters, specifically the translation vector and rotation matrix, while keeping the intrinsic parameters constant, we simulate deviations from the actual poses. These perturbations mimic real-world errors, where rotational measurements might be imprecise due to factors such as sensor errors, mechanical slippage, or calibration inaccuracies. We utilize these noisy poses along with the input images to train the 3D reconstruction models. This analysis carefully examines the influence of camera pose errors on the fidelity of 3D reconstructions. Furthermore, by systematically varying the number of input images used in the reconstruction process, we seek to better understand how the introduction of noisy camera poses impacts the quality and accuracy of 3D reconstructions across a spectrum of image densities. This analysis includes scenarios ranging from sparse image sets, with only a few images, to dense configurations, with a large number of images. We aim to understand the interplay between image quantity, and camera pose inaccuracies and how these factors collectively influence the robustness and performance of learning-based surface reconstruction models under diverse real-world conditions.

- **Influence of lighting conditions:** To examine how different aspects of lighting—such

as intensity, direction, and uniformity—impact the 3D face reconstruction quality. The goal is to comprehensively understand how sub-optimal lighting variation affects the accuracy and robustness of learning-based surface reconstruction models. Lighting is a pivotal factor in image capture for 3D modelling, significantly influencing the visibility of features and the overall quality of the input data. We carefully designed the experiment setup to simulate low lighting conditions, focusing on scenarios where lighting may pose limitations. This meticulous setup ensures that our findings accurately reflect real-world conditions. We use images generated in a controlled environment with non-uniform lighting for this analysis, employing a directional point light source. The detailed explanation for the data used in this experiment is further elaborated in Sec 4.5, which is used for the reconstruction process. By incorporating a systematic variation in the number of input images for the reconstruction process, this setup allows us to address the challenges posed by lighting conditions rigorously and to thoroughly assess model performance across a wide range of input images. For instance, we investigate the models' capabilities in reconstructing a 3D face using image sets ranging from as few as three images to as many as a hundred, all captured under identical lighting conditions.

4.3 Selection of the SOTA Methods

Determining an optimal SOTA method for 3D reconstruction involves careful consideration of several critical criteria to ensure that the chosen approach yields high-quality 3D reconstructions and aligns with real-world application constraints. We outline four essential criteria for evaluating and selecting effective 3D learning-based surface reconstruction methods.

- **Sparse Input Data:** The first criterion requires the method's capability to generate reasonable 3D reconstruction using only seven RGB images. This requirement is significant because it assesses the method's efficiency and robustness in inferring 3D geometry from minimal information.
- **No Auxiliary Supervision:** The second criterion highlights the importance of a method's ability to operate independently without the need for additional information for supervision, such as depth maps or segmentation masks. This criterion is particularly relevant for simplifying the data acquisition process and reducing the cost associated with 3D reconstruction technologies, as it eliminates the need for special hardware that collects the additional information. However, it is noteworthy that while our primary objective was to minimize reliance on auxiliary supervision, we made an exception to include the GaussianObject [72] method in our evaluation. Although this particular method necessitates the use of masks for 3D reconstruction, it demonstrates the ability to effectively work with limited input images.
- **Computational Efficiency:** The third criterion emphasizes the computational constraints of the methods, particularly regarding training and execution time. We limited our selection to models with a training time of less than 24 hours on a reasonably powerful GPU. This restriction is crucial for implementing these methods in real-world

situations. By enforcing this criterion, we ensure that the chosen methods are not only academically interesting but also viable for real-time or near-real-time applications.

- **Exclusion of Pre-trained Models:** The fourth criterion focuses exclusively on methods that are not pre-trained on any dataset. This ensures that the performance assessment is based purely on the inherent capabilities of the methods themselves, without the influence of prior knowledge gained from large-scale datasets. By excluding pre-trained models, we ensure a uniform starting point, facilitating an unbiased evaluation of each method under the same initial conditions, enabling to focus on their intrinsic performance. However, we include Dust3r [66] methods that utilize a CroCo [68] pretrained model. This inclusion is due to its unique approach, which takes only images as input and outputs the 3D model.

Based on these criteria, the following methods are particularly promising for generating 3D objects from a set of images.

- **NeuS: Learning Neural Implicit Surfaces by Volume Rendering for Multi-view Reconstruction.** NeuS [64] employs a novel volume rendering technique to train a neural signed distance representation that ensures unbiased surface reconstruction in the first-order approximation of the SDF. The method achieves high-fidelity surface reconstructions, specifically with objects that have complex structures. When provided with calibrated multi-view images of an object, NeuS implicitly represent the surface of the object using two funcions, $f(\mathbf{x}) : \mathbb{R}^3 \to \mathbb{R}$ which maps the 3D position $x \in \mathbb{R}^3$ to its signed distance value to the object, and $c : \mathbb{R}^3 \times \mathbb{S}^2 \to \mathbb{R}^3$ which encodes the color and viewing direction associated with a point. Thus, the surface can be represented as

$$S = \{x \in \mathbb{R}^3 \mid f(x) = 0\}.$$

. Then, with the help of novel PDF $\phi_s(f(x))$, called S-density field, SDF network is trained only with 2D images via volumetric rendering. In our experiments, we trained the NeuS model using images and their corresponding camera poses. Each dataset was trained for 100k iterations without using a mask, taking approximately 9 hours on a single GPU. This demonstrates the significant computational intensity and time required for high-quality surface reconstructions.

- **Instant Neural Graphics Primitives with a Multiresolution Hash Encoding.** By implementing a novel encoding technique using a hierarchy of hash tables to map neural network inputs to a higher-dimensional space, Instant-NGP [48] demonstrates efficient training and high-quality approximations. Instant-NGP presumes that the object to be reconstructed is confined within multi-resolution voxel grids. These voxel grids at each resolution are then mapped to a hash table containing a fixed-size array of learnable feature vectors. To obtain a hash encoding for a 3D position $\mathbf{x} \in \mathbb{R}^3$, it calculates $h^i(\mathbf{x}) \in \mathbb{R}^d$ at each level (where d is the dimension of a feature vector, $i = 1, \ldots, L$)) by interpolating the feature vectors allotted to the surrounding voxel grids at that level. The hash encodings at all L levels are then combined to form the multi-resolution hash

encoding $h(\mathbf{x}) = \{h^i(\mathbf{x})\}_{i=1}^{L} \in \mathbb{R}^{L \times d}$. Another key aspect is its effective utilization of GPU parallelism, which avoids control flow divergence and pointer-chasing, thereby significantly accelerating the training process. Nonetheless, Instant-NGP struggles to achieve high accuracy in reconstructing intricate geometries. The dependence on hash tables can lead to hash collisions, potentially introducing errors in the generated graphics. Our research utilized the Instant-NGP model, trained on images and corresponding camera poses. Each dataset was subjected to 50,000 epochs of training, and the entire process took around 8 minutes on a single GPU. This illustrates the model's rapid and efficient training process.

- **NeuS2: Fast Learning of Neural Implicit Surfaces for Multi-view Reconstruction.** The method's key feature is its utilization of multi-resolution hash encodings of learnable features to represent the neural network-encoded SDF. NeuS2 [67] maps each 3D position $\mathbf{x}$ to multi-resolution hash encodings $h_\Omega(x)$ with learnable hash table entries Ω, combining the hash encodings with the position $\mathbf{x}$ as input to the SDF network $f_\Theta(x, h_\Omega(x))$ for efficient rendering and training. It implements a strategy of progressive learning to optimize the multi-resolution hash encoding from coarse to fine, stabilizing and expediting training processes. This method incorporates a novel lightweight approach for the calculation of second-order derivatives that are tailored for ReLU-based MLPs to utilize CUDA parallelism, enhancing speed and efficiency. This method excels in significantly accelerating training speed compared to previous methods, such as NeuS. It drastically reduces the computational workload, enabling the reconstruction of static scenes in minutes and dynamic scene training at a rate of 20 seconds per frame. Despite the expedited training, NeuS2 maintains reconstruction quality with accurate geometry and visual fidelity. This method surpasses the state-of-the-art in both training speed and reconstruction accuracy. In the context of our experiments, we trained the NeuS2 model on each dataset of images and their corresponding camera poses. The training involved 100K iterations per dataset, and the training duration was approximately 20 minutes on a single GPU.

- **BakedSDF: Meshing Neural SDFs for Real-Time View Synthesis.** BakedSDF [74] introduces a novel approach to reconstructing unbounded real-world scenes with high precision using a hybrid neural volume-surface representation. The method integrates the advantages of both volumetric and surface-based techniques, ensuring smooth level sets and allowing high-resolution mesh extraction through the utilization of Marching Cubes [41]. This method optimizes the geometry and appearance of an object using a surface-based representation and NeRF-like volume rendering [46] with a learned function that maps a 3D position $\mathbf{x}$ and outgoing ray direction $\mathbf{d}$ to a volumetric density τ and color $\mathbf{c}$. To render the color of a single pixel in a target camera view, first, it computes the ray corresponding to that pixel $\mathbf{r} = \mathbf{o} + t\mathbf{d}$, and then evaluates the NeRF at a series of points $\{t_i\}$ along the ray. The resultant outputs $\tau_i, \mathbf{c}_i$ at each point are

combined together into a single output color value C:

$$C = \sum_i \exp\left(-\sum_{j<i} \tau_j \delta_j\right) \left(1 - \exp(-\tau_i \delta_i)\right) \mathbf{c}_i, \quad \delta_i = t_i - t_{i-1}.$$

This above definition of C is a quadrature-based approximation of the volume rendering equation [44]. After optimizing the neural volumetric representation, a triangle mesh is created from the recovered MLP-parameterized SDF by querying it on a regular 3D grid and then running Marching Cubes [41]. Finally, to enhance the view-dependent appearance, spherical Gaussians embedded within each vertex of the mesh are used, improving the realism and fidelity of rendered images. However, this method struggles to capture very fine or thin structures, such as dense foliage and may have difficulties with semi-transparent materials like glass and fog due to the limitations inherent in mesh-based reconstructions. In our study, the BakedSDF model (SDFstudio implementation) was trained using images and their corresponding camera poses. Each dataset underwent 100K training iterations, with the training and inference steps taking around 12 hours on a single GPU.

- **Volume Rendering of Neural Implicit Surfaces** Unlike other traditional methods of modelling geometry as a volume density function, VolSDF [75] uses an SDF representation for modelling volume density as a function of the geometry. The volume density, $\sigma(x)$, is defined as :

  $$\sigma(x) = \alpha \Psi_\beta(-d_\Omega(x))$$

  In the above equation, α and β are learnable parameters, Ψ_β is the Cumulative Distribution Function (CDF) of the Laplace distribution, and $d_\Omega(x)$ represents the signed distance from point x to the surface. This innovative approach of using an SDF for modelling volume density significantly improves the geometry representation and reconstruction in neural volume rendering. VolSDF successfully achieves disentanglement of shape and appearance, allowing for independent manipulation of these components. This means that the geometric structure (shape) and visual characteristics (appearance) can be switched between different scenes. This flexibility is not possible with traditional NeRF-based models, thus enhancing the usability VolSDF usability. This method enhances the capability of generating high-fidelity 3D models from sparse input images by generating a more accurate and reliable geometry reconstruction. However, this method has limitations in representing non-watertight manifolds and surfaces with boundaries, such as zero-thickness surfaces. Our experiments utilized the VolSDF model, which was trained using images and their camera poses. Each dataset was subjected to 100k iterations, with the entire process taking approximately 9 hours on a single GPU. This underscores the significant computational effort and time commitment needed for precise surface reconstructions.

- **Neus-facto:** SDF Studio [77] is a comprehensive framework that integrates multiple neural implicit surface reconstruction methods. This unified framework benefits from

the strengths of multiple neural implicit methods. One of the key methods within SDF Studio is Neus-facto. This approach draws inspiration from the Nerfacto method developed in nerfstudio [60]. In Neus-facto, the proposal network from mip-NeRF360 [4] is utilized to sample points along the ray. The concept of employing a proposal network for point sampling, originally from mip-NeRF360, is integrated with NeuS [64]. By adopting this sampling strategy, the process is significantly accelerated, and the number of samples required for each ray is drastically reduced. This optimization, which involves efficient sampling along the ray using the proposal network, enhances the efficiency of neural volume rendering while preserving high-fidelity multiview 3D reconstruction capabilities. The Neus-facto model was trained in our experiments using image datasets and their corresponding camera poses. For each dataset, the model was trained for 100K iterations, taking roughly 45 minutes on a single GPU. This highlights the model's rapid training capabilities.

- **SuGaR: Surface-Aligned Gaussian Splatting for Efficient 3D Mesh Reconstruction and High-Quality Mesh Rendering.** SuGaR [26] has pioneered a technique that allows for the rapid and accurate extraction of scene geometry from 3D Gaussian Splatting [36] representations. The approach introduces a regularization term, which ensures that the Gaussian distributions align closely with the surface geometry of the scene. The alignment of these Gaussian distributions to the scene surface is then exploited to extract a scene geometry from the Gaussians using Poisson reconstruction [35]. The Poisson reconstruction is fast, scalable, and preserves fine details within the reconstructed scene. To further enhance the quality of the reconstruction, an optional refinement strategy is employed. This strategy involves binding the Gaussians to the surface of the mesh and simultaneously optimizing both the Gaussians and the mesh through Gaussian splatting rendering. This technique refines the mesh by fine-tuning the positions and parameters of the Gaussians to achieve a more precise and detailed representation. SuGAR simplifies the creation of high-quality and readily deployable meshes. This method offers significantly faster processing, while it may involve a trade-off in rendering quality when compared to slower and more computationally intensive methods such as NeuS [64].

- **GaussianObject: Just Taking Four Images to Get A High-Quality 3D Object with Gaussian Splatting.** GaussianObject can reconstruct and render high-quality complex 3D objects from highly sparse views using only four input images. The major focus of this technique is Structure-Prior optimization. Techniques such as visual hull and floater elimination are introduced to explicitly incorporate structure priors into the initial optimization process. This helps to establish multi-view consistency and create a coarse 3D Gaussian representation. Then, it utilizes a new Gaussian repair model based on diffusion models to fill in incomplete object information. It uses self-generating strategies to provide enough image pairs for constructing the Gaussian repair model. To achieve photo-realistic rendering from any viewpoint, the framework employs 3D Gaussian Splatting that enhances structure prior to embedding and fast

rendering capabilities. Each 3D Gaussian that represents a scene is parameterized as $G = \{\mu_i, q_i, s_i, \sigma_i, sh_i\}$, where μ denotes the center location, q the rotation quaternion, s the scaling vector, σ the opacity, and sh the spherical harmonic (SH) coefficients.

$$C(u) = \sum_{i \in N} SH(sh_i, v_i)\alpha_i \prod_{j=1}^{i-1}(1 - \alpha_j), \alpha_i = \sigma_i G_i(p)$$

This equation describes the color $C(u)$ of a pixel u based on the view direction v_i.

This method is restrained by its dependence on accurate camera poses for creating a visual hull. When dealing with extreme viewpoints, this technique may produce artifacts that negatively impact the visual quality and impede the reconstruction performance.

- **DUSt3R: Geometric 3D Vision Made Easy.** DUSt3R [66] is the first complete 3D reconstruction pipeline, unifying monocular and binocular 3D reconstruction from uncalibrated and unposed images. It's a novel approach for Dense Unconstrained Stereo 3D Reconstruction from uncalibrated and unposed cameras. Its innovative capability lies in its ability to operate without requiring any prior knowledge of camera calibration or scene parameters. Instead, it employs a fully data-driven strategy using a generic transformer architecture. A key aspect of DUSt3R lies in its global alignment strategy for mapping the points in multi-view 3D reconstruction. Unlike traditional methods that rely heavily on camera calibrations and minimize reprojection errors, DUSt3R optimizes camera pose and alignment of the object's geometry directly in 3D space. This global alignment represents a significant advancement, ensuring that the reconstructed 3D model is coherent and accurate across all views. By aligning all point maps simultaneously, DUSt3R eliminates the need for iterative reprojection error minimization, thereby streamlining the reconstruction process and enhancing robustness. However, this global optimization strategy has some limitations. DUSt3R requires extensive datasets for effective training, which can be resource-intensive. Moreover, the global optimization strategy was not designed to efficiently handle extensive image collections, which could potentially hinder its scalability. Additionally, the method removes the necessity for explicit camera parameters, resulting in reduced accuracy when compared to traditional methods that utilize these parameters for sub-pixel triangulation.

4.4 Dataset and Rendering Setup

4.4.1 Rendering Synthetic Face Images using the FaceScape Dataset

For the generation of training images, which are used to reconstruct a 3D model, we employed the FaceScape [73] dataset, a large-scale 3D face dataset renowned for its remarkably detailed 3D representations of human faces. The FaceScape dataset provides 18,760 textured 3D faces captured from 938 subjects, each showcasing 20 distinct expressions. The 3D models demonstrate facial geometry at the pore level and have been uniformly standardized topologically to ensure consistent geometry throughout the dataset.

The collection of high-resolution, textured 3D facial models from the FaceScape dataset serves dual purposes in our study:

- **Synthetic Data Generation:** We render synthetic face images under a customizable environment that allows precise control over camera settings and lighting conditions in Blender [14] software. By utilizing FaceScape's high-resolution 3D face models with detailed textures, we generate synthetic face data. We selected the FaceScape dataset because of its high-quality 3D face models, characterized by detailed textures and complex geometric structures. These models are rendered to produce synthetic images that closely resemble real human faces, which is crucial for reconstructing accurate 3D models. The fine level of detail in the FaceScape models ensures that the synthetic data used for training the 3D reconstruction model can mimic the intricate appearance of real faces, thereby enhancing the model's effectiveness in real-world applications.

- **FaceScape Benchmarking:** The FaceScape models act as the ground truth in our experiments, enabling us to evaluate the accuracy of reconstruction methods under various conditions. By utilizing these models as a standard, we are able to precisely quantify the accuracy of the resultant 3D reconstruction of the face. This assessment is done using the Chamfer distance metric [8], which serves as the primary measure used for evaluating the completeness of the reconstructed facial geometries. Further details about the evaluation process are covered in Section 4.7

4.4.2 Blender Setup

We utilize Blender 3.5 [14] to generate the synthetic face data. We configured a virtual camera setup within the Blender viewport, positioning multiple cameras strategically around the centrally placed 3D face model from the FaceScape dataset [73]. Each 3D model is selected randomly and consistently used for generating synthetic data. The cameras were placed on a sphere with a radius of 2m, as illustrated in Figure 4.1. This equidistant placement of cameras ensures that the 3D face model is uniformly and fully covered, resulting in multi-view consistent rendered images—all the cameras point at the 3D face model from the FaceScape dataset. Our approach for generating multi-view images is designed to provide a comprehensive view of the subject's face captured from multiple cameras, with a particular emphasis on capturing the frontal side.

To effectively capture a diverse range of viewpoints from different perspectives, we carefully adjust two critical parameters of the camera to control its position: yaw and pitch. Each parameter is carefully calibrated to maximize the coverage.

- **Yaw:** Yaw refers to the horizontal rotation of an object around its vertical axis. For every camera, the yaw angle varies from -60° to 60°, and the yaw angle is incremented by 10° for each step. This results in a series of camera positions at -60°, -50°, -40°, ..., 50°, 60°. This broad range allows the capture of the model's side profiles from extreme side views to nearly frontal views.

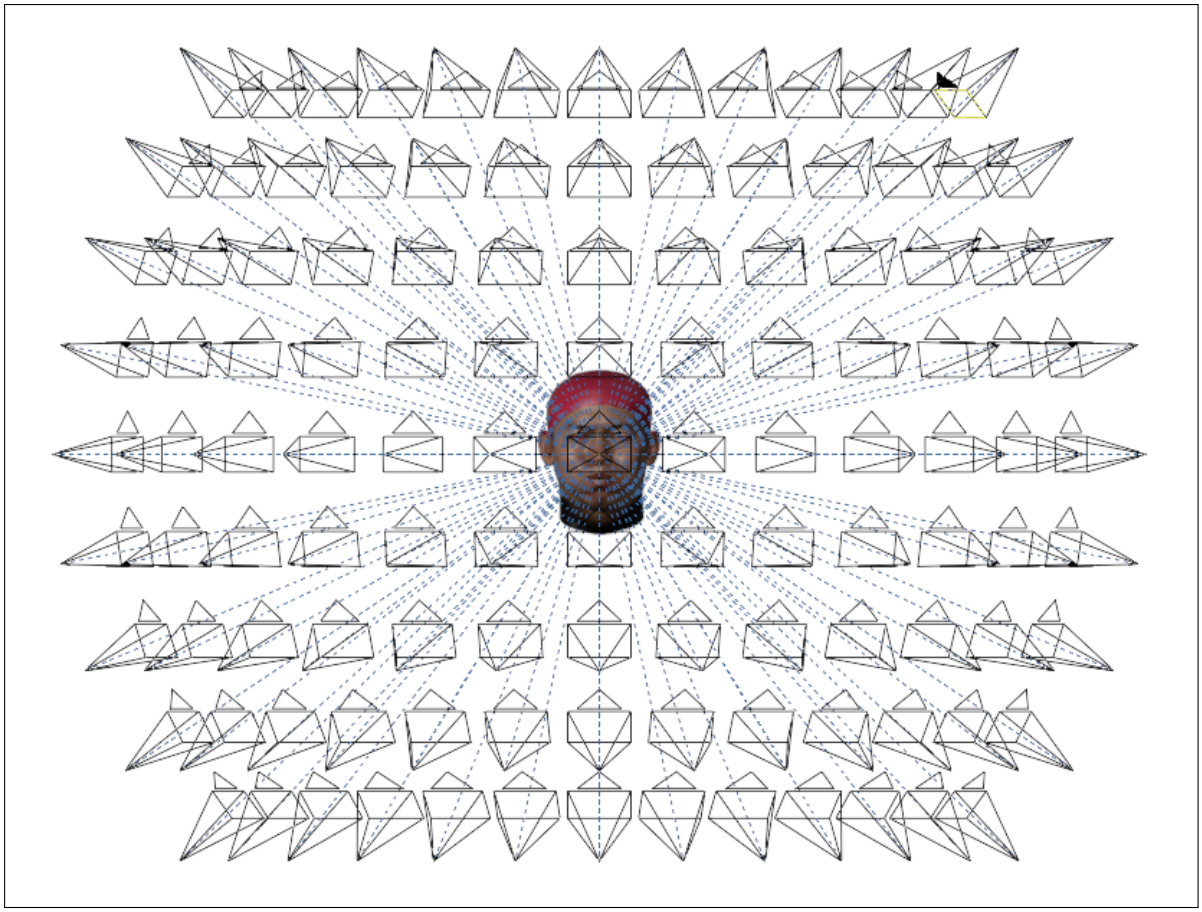


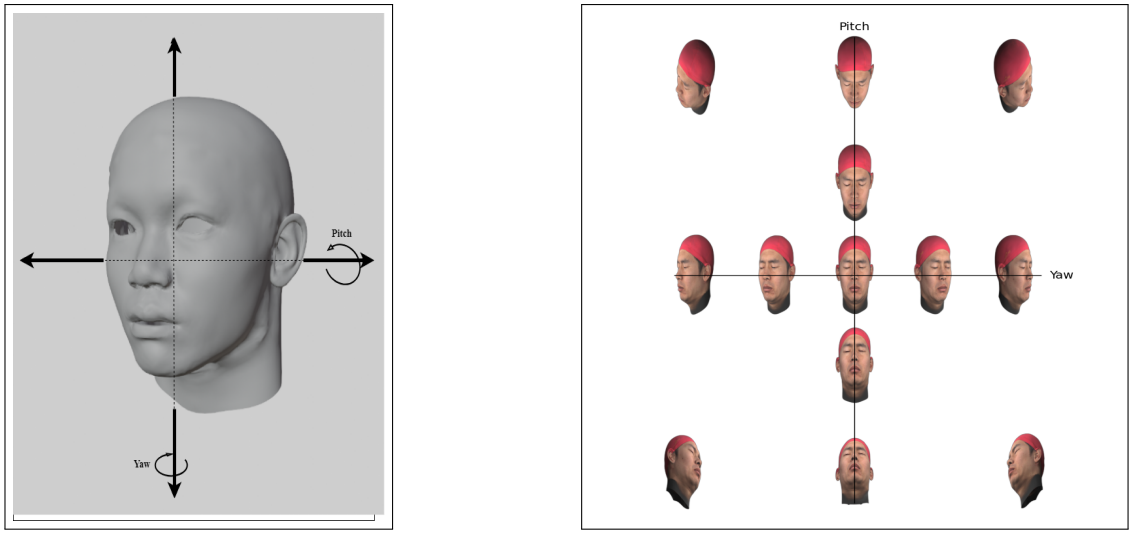
Figure 4.1: Setup with 117 cameras capturing 3D face model from multiple angles, 2m apart.

- **Pitch:** Pitch refers to the vertical rotation of an object around its horizontal axis. Similar to the yaw setup, the pitch angle varies from -40° to 40°, with increments of 10° for every camera. This setup encompasses camera placements at -40°, -30°, -20°, ..., 30°, 40°. This configuration enables the capture of the upper and lower parts of the face.

Figure 4.2 depicts a visual representation of the rotational adjustments of yaw and pitch on a 3D face model. Yaw movement involves the rotation of the model around its vertical axis, and pitch entails vertical rotation around the model's horizontal axis.

The systematic variation in yaw and pitch angles, resulting in a total of 117 distinct camera positions, encapsulates the essence of this design. These positions are precisely arranged to comprehensively capture the wide range of viewpoints of the 3D face model. The collective viewing of the 117 rendered images from their respective camera positions forms a panoptic rig. All camera's resolution is set to a standard HD format (1920×1080) to ensure high-quality render images that are precisely calibrated. To achieve a high level of precision, we define both the intrinsic and extrinsic parameters of each camera accurately.

- **Intrinsic Parameters:** These camera's parameters, which include the focal length, optical centre, and lens distortion, are precisely defined within Blender to ensure that



(a) **(b)**

Figure 4.2: (a) Illustration of Yaw and Pitch Rotation. We rotate around horizontal and vertical axes to get pitch and yaw variation. (b) Variation in head orientation with respect to yaw and pitch angles.

each rendered image accurately reflects realistic camera settings.

- **Extrinsic Parameters:** The spatial relationships and viewing angles for each camera in our setup are carefully recorded, providing valuable data on the spatial relationships and viewing angles relative to the 3D face model. The general form of the extrinsic matrix **E** is:

$$\mathbf{E} = \begin{bmatrix} \mathbf{R} & \mathbf{t} \\ 0 & 1 \end{bmatrix}$$

where:

  - **R** is the 3x3 rotation matrix.
  - **t** is the 3x1 translation vector.
  - The last row $\begin{bmatrix} 0 & 0 & 0 & 1 \end{bmatrix}$ ensures homogeneous coordinates.

For effective modelling of 3D geometry of face from 2D images, it is crucial to accurately determine the poses of the cameras used. This involves defining both the intrinsic and extrinsic parameters, which are essential for establishing the precise positioning of the cameras.

We use a Python script to ensure consistency and automate the rendering process in Blender. This script efficiently configures 117 cameras at predetermined yaw and pitch angles. It imports a 3D face model from the FaceScape dataset, positioning it centrally in the Blender scene to ensure optimal alignment with the configured camera setup. The script initiates the rendering process for each camera angle, systematically producing a total of 117 images.

Alongside image rendering, the script also generates a JSON file containing vital camera parameters, including both intrinsic parameters (like focal length and sensor size) and extrinsic parameters (specifically the position and orientation of each camera relative to the 3D model).

4.5 Overview of Data used in Experiments

In this section, we comprehensively describe the synthetic image data employed in three experiments. The data is used to evaluate the 3D reconstruction models under various configurations, from sparse to dense image settings. In the context of sparse 3D reconstruction, we utilize sets of 3, 5, and 7 images to generate the 3D model. Conversely, in the dense 3D reconstruction setting, we employ sets of 10, 15, 30, 50, and 100 images to create a more detailed and comprehensive model. The selection of images for each configuration is executed using a Sampling Algorithm, which is elaborated upon in Section 4.6. To ensure the robustness of our evaluation, we have generated data across three distinct configurations to capture a wide range of viewpoints. This strategy is critical because, even within a specific configuration, the inclusion of poorly chosen poses could potentially compromise the accuracy of the 3D reconstruction evaluation. In order to ensure the reliability of the 3D reconstruction methods, our methodology is specifically tailored to optimize the diversity of viewpoints captured in the images while minimizing any redundant data within the image set. By doing so, we can effectively enhance the overall quality and precision of our assessment of the 3D reconstruction models.

- **Experiment 1: Optimal Lighting and Accurate Poses** In this experiment, synthetic images were rendered under optimal lighting conditions utilizing FaceScape [73] models in Blender[14]. The images were captured from all the 117 cameras positioned at distinct locations. Figure 4.3 depicts a series of rendered images captured from various camera angles. We ensured the acquisition of precise camera poses by utilizing predefined camera positions. Subsequently, we curated a set of images comprising various quantities of images: 3, 5, 7, 10, 15, 30, 50, and 100. These image sets, along with their accurate poses, were then employed to reconstruct the 3D face model using chosen SOTA methods. Optimal lighting conditions were carefully configured to ensure uniform illumination throughout all the areas. This effectively reduces shadows in the rendered images, improving the accuracy of the 3D reconstruction. The background of each image was removed as part of the preprocessing step. This step is crucial for isolating the subject from the background and eliminating any nonessential information that could adversely affect the reconstruction process. The background removal was carried out using a tool called Rembg [23] tool based on U^2-Net to ensure consistency across all images.

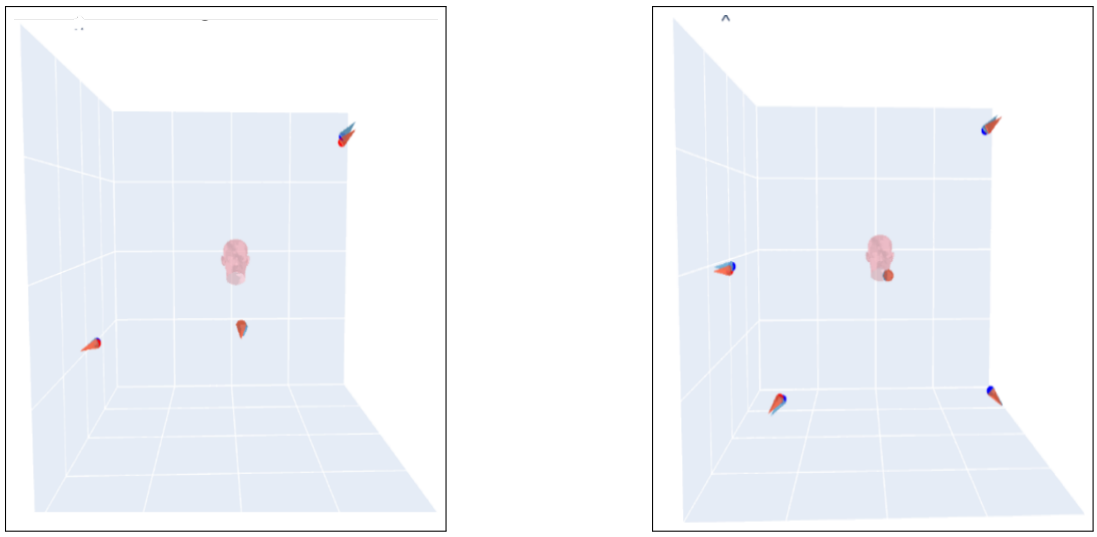
- **Experiment 2: Pose Perturbations** For this experiment, we generated distorted poses by introducing controlled perturbations to evaluate the robustness of the 3D reconstruction methods in the presence of the noisy camera pose. The same set of images from Experiment 1 was reused; however, to introduce some variability in the

Figure 4.3: Collection of images rendered from diverse camera positions.

poses, Gaussian noise with a mean of 0 and a standard deviation of 0.005 was added to the camera poses. The following practices were used to introduce noise into the camera poses:

- **Translation Vector Noise:** We apply Gaussian noise with a mean of 0 and a standard deviation of 0.005 directly to the translation matrix elements for each camera's pose. This process simulates potential errors in estimating the position of the camera in the scene.
- **Rotation Matrix Noise:** Unlike the translation matrix, the perturbation of the rotation matrix involves a more complex process due to maintaining a valid rotation after adding the noise. Initially, we convert the rotation matrix into Euler angles, which accurately represent 3D rotational movements along the principal axes. Then, we apply Gaussian noise with a mean of 0 and a standard deviation of 0.005 to these Euler angles. After noise addition, the perturbed Euler angles convert into a rotation matrix. After forming the new perturbed rotation matrix, the next step is to verify that it represents a valid rotation. This verification process involves confirming the newly formed rotation matrix's orthogonality and ensuring that the determinant of the matrix is exactly 1. This ensures that the rotation does not have any scaling or skewing effects that distort the 3D reconstruction process.

This addition of noise is designed to replicate realistic variations in pose accuracy, reflecting the potential inaccuracies encountered in real-world scenarios. Figure 4.4 depicts the deviation between the original camera poses and perturbated poses for three and five sampled camera positions. The original camera poses are represented in blue, while the noisy camera poses are shown in red. This visual representation in the Figure effectively demonstrates the magnitude of the perturbations and their spatial distribution. These perturbed camera poses, along with the corresponding images, were then input into learning-based 3D surface reconstruction methods. This setup allowed us to examine how minor inaccuracies in camera positioning affect the performance and reliability of the reconstruction techniques. The evaluation aimed to understand how sensitive the models are to changes in camera position and to determine to what extent these inaccuracies in poses downgrade the quality of the resulting 3D face models.



(a) (b)

Figure 4.4: Visualization of the deviation between original (blue) and perturbed (red) camera poses for three (a) and five (b) camera positions.

- **Experiment 3: Suboptimal Lighting Conditions** To simulate real-world lighting challenges, we utilized the FaceScape [73] model to render images under suboptimal lighting conditions using Blender [14] software. The 3D model was illuminated by a low-intensity point light placed at the extreme left, creating an artificial effect with selectively illuminated facial features. The rendered images are specifically tailored to evaluate the performance of 3D reconstruction models under challenging lighting conditions. A few rendered synthetic images demonstrating these lighting effects are illustrated in Figure 4.5. These images effectively highlight the variability and complexity introduced by the suboptimal lighting. Similar to Experiment 1, we removed the background of the rendered images using the Rembg tool as a preprocessing step. This step was crucial to isolate the facial features from the background. The preprocessed images, combined with accurate camera poses, were then input into selected SOTA reconstruction methods to evaluate their performance under these challenging lighting conditions.

4.6 Sampling Algorithm

We implemented a customized sampling technique based on the Farthest Point Sampling (FPS) algorithm, specifically designed to sample images (with diverse viewpoints and less overlapping in sparse settings) from a pool of 117 synthetic images rendered using the FaceScape [73] model in Blender [14] software. The collective viewing of the 117 rendered images from their specific camera positions forms a panoptic rig. Our sampling technique started at the centre region of the panoptic rig, serving as the initial reference point for subsequent selections. The central region consists of images with yaw angles ranging from -20° to 20° and

Figure 4.5: Rendered images taken from different camera perspectives under suboptimal lighting conditions.

pitch angles ranging from -10° to 10°. This strategic sampling technique ensured that the chosen images captured a diverse range of perspectives, a crucial aspect of the 3D reconstruction process. In this process, a critical element was the use of a predefined minimum spatial threshold requirement that guarantees the capture of the distinct spatial characteristics of each sampled image. A sampled image was considered valid if it was located at least a minimum spatial threshold distance away from all previously selected images. Images failing to meet this criterion were excluded from the selection pool. The sampling algorithm was designed with adaptability in mind. When no valid images can be found after multiple attempts, the algorithm dynamically relaxes the minimum spatial threshold requirement. This is done by reducing the distance criterion by half, which increases the chances of finding valid images over time. This adaptive approach allows for efficient sampling while maintaining a good balance between image diversity and spatial coverage, which are important for modelling 3D geometry.

The pseudocode in Algorithm 4.1 outlines the detailed steps of our customized sampling algorithm and illustrates how the selection process is carried out:

4.7 Evaluation Metric

This section discusses the evaluation metric utilized to measure the completeness of reconstructed 3D models generated by SOTA methods across all the experiments. For this purpose, we employed the **Chamfer Distance** (CD) initially proposed by Barrow et al. [8] for quantitative evaluation. In the context of 3D reconstruction, CD is chosen as the loss function for optimizing the network. Additionally, it is also widely used to evaluate the performance of reconstruction methods by measuring the similarity between the ground truth mesh and the reconstructed mesh.

The CD is computed as the average pair-wise distance between two point sets, specifically, the ground truth point set $\boldsymbol{S_1}$ and the reconstructed mesh vertex set $\boldsymbol{S_2}$. The use of CD offers a significant advantage due to its ability to compute efficiently and its versatile applicability to point sets of varying sizes. Various studies have demonstrated the effectiveness of CD in comparing the quality of 3D reconstructions, notably in the 3D MoMa project **??**.

```
Require: Num_images ← {3, 5, 7, 10, 15, 30, 50, 100}
Require: MinSpatialThreshold ← 6
for i ← 1 to len(Num_images) do
    Set SampledImages ← empty list
    FirstSample ← randomly_sample_image from the center region
    Add FirstSample to SampledImages
    for each value in i − 1 do
        NextSample ← randomly_sample_image from the panoptic rig
        # Check the constraint
        if distance(NextSample) > MinSpatialThreshold then
            # Valid sample
            Add NextSample to SampledImages
        else
            goto sample
        end if
        # If after a few iterations a valid sample is not found, then relax the distance constraint
        MinSpatialThreshold ← MinSpatialThreshold / 2
        goto sample
    end for
end for
return SampledImages
```

Algorithm 4.1: Sample_Images

Mathematically, the CD is defined as:

$$d_{\mathbf{CD}}(\boldsymbol{S_1}, \boldsymbol{S_2}) = d_{\mathbf{CD1}}(\boldsymbol{S_1}, \boldsymbol{S_2}) + d_{\mathbf{CD2}}(\boldsymbol{S_1}, \boldsymbol{S_2}) = \frac{1}{|\boldsymbol{S_1}|} \sum_{x \in \boldsymbol{S_1}} \min_{y \in \boldsymbol{S_2}} \|x - y\|_2^2 + \frac{1}{|\boldsymbol{S_2}|} \sum_{y \in \boldsymbol{S_2}} \min_{x \in \boldsymbol{S_1}} \|x - y\|_2^2 \tag{4.1}$$

$\boldsymbol{S_1} \subset \mathbb{R}^3$ is the ground truth point set and $\boldsymbol{S_2} \subset \mathbb{R}^3$ is the reconstructed mesh vertex set. The Chamfer Distance is composed of two parts: $d_{\mathbf{CD1}}(\boldsymbol{S_1}, \boldsymbol{S_2})$ and $d_{\mathbf{CD2}}(\boldsymbol{S_1}, \boldsymbol{S_2})$. The term $d_{\mathbf{CD1}}(\boldsymbol{S_1}, \boldsymbol{S_2})$ computes the average distance from a point in $\boldsymbol{S_1}$ to its nearest vertex in $\boldsymbol{S_2}$, while $d_{\mathbf{CD2}}(\boldsymbol{S_1}, \boldsymbol{S_2})$ evaluates the average distance from a vertex in the mesh to its nearest point in $\boldsymbol{S_1}$.

Chapter 5

Results

This chapter presents the results of three experiments conducted to evaluate the performance of the SOTA methods for human head reconstruction using synthetic data. We have calculated the Chamfer Distance [8] between the ground truth mesh and the reconstructed 3D model for evaluation. Since the Chamfer Distance is the measure of the discrepancy, the lower value implies a higher similarity between the ground truth and the reconstructed mesh, reflecting a better reconstruction performance. We have provided both quantitative results and qualitative results. The quantitative results include a thorough analysis of all selected SOTA methods and a comparison of their performance of surface reconstruction across all the experiments. The qualitative results focus on the top four performing methods (with lower Chamfer Distance) in each experiment, providing visual representations that illustrate the effectiveness of the top four methods in reconstructing the detailed 3D head models.

5.1 Experiment 1: Impact of image quantity

In this experiment, we have investigated the effect of varying the number of input images (from sparsely populated to densely populated) on the performance of nine SOTA 3D reconstruction methods. We utilized eight sets of images, each containing a different number of images: 3, 5, 7, 10, 15, 30, 50, and 100. These images are selected from a total pool of 117 images via our sampling algorithm 4.6. The rendering of these synthetic images is covered in the data section ¸4.5. We utilized three different configurations for each of the eight sets of images and then trained all selected SOTA methods using these input sets, covering both sparse and dense settings. For instance, consider an image set containing three images. We reconstructed three models for this set, each with a different configuration. This process was repeated for all sets: 3, 5, 7, 10, 15, 30, 50, and 100 images. As a result, we constructed 24 models for a single method (8 sets of images $\times$ 3 configurations). This comprehensive approach allows us to thoroughly evaluate each method's performance across various configurations and dataset sizes.

5.1.1 Quantitative Results

Figure 5.1 illustrates the quantitative comparison of all the selected SOTA models, and Figure 5.2 represents the performance of the top four performing methods based on their reconstruction quality, which is evaluated by calculating the Chamfer Distance metric between the ground truth and the reconstructed 3D representation. The x-axis represents the number of images, while the y-axis shows the Chamfer Distance in millimetres. Our results demonstrate that for each method, the quality of 3D reconstruction improves with an increasing number of images, achieving the best results (i.e., the lowest Chamfer Distance value) when each method is trained with the complete set of 100 images. However, there is a noticeable degradation in the reconstruction quality as the number of images decreases, and only a few methods—such as NeuS2, NeuS, VolSDF, BakedSDF, and Neus-facto, demonstrate the capability to construct a reasonable 3D reconstruction under sparse image settings.

By testing these SOTA methods with multiple sets, mainly varying in image quantity, our findings further suggest that for most of the methods, there is no significant improvement in the reconstruction quality beyond 15 images. Specifically, the Chamfer Distance shows minimal reduction when the number of images is increased from 15 to 30 and remains almost unchanged for 50 and 100 images. This plateau in performance suggests that a much smaller set of images, about 15, could be sufficient to achieve near-optimal reconstruction quality for various methods.

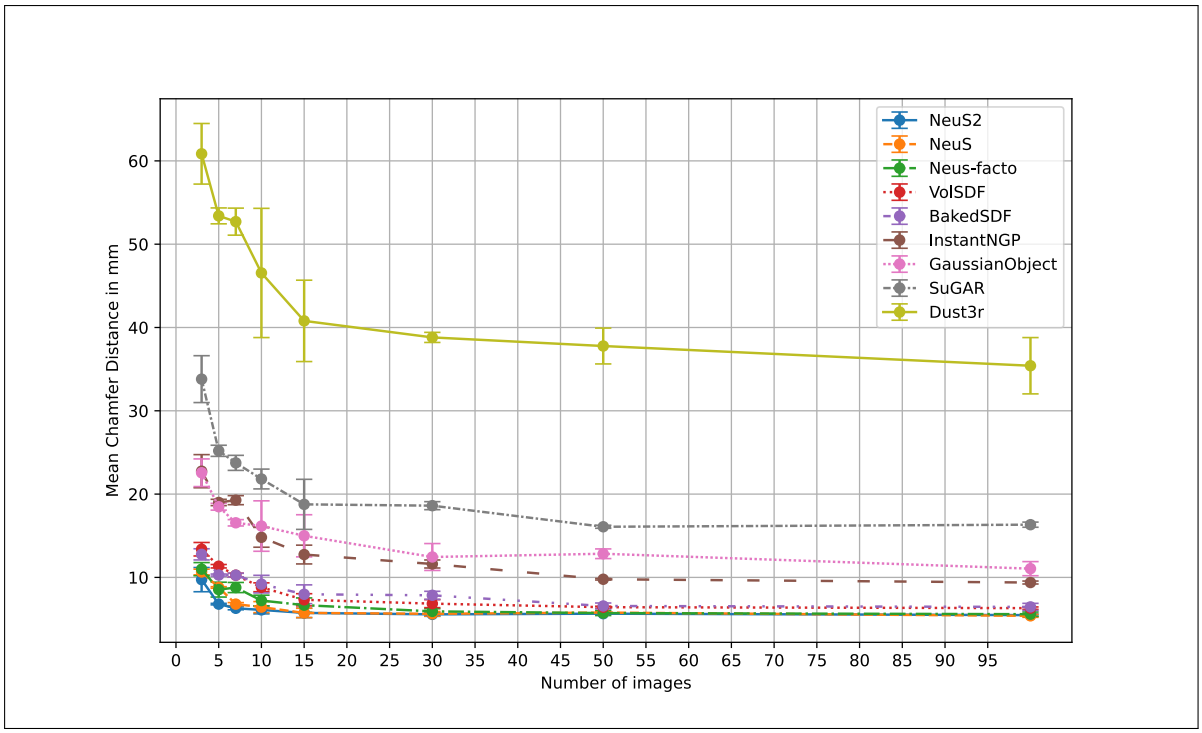


Figure 5.1: Performance analysis of nine SOTA 3D reconstruction methods with varying input images.

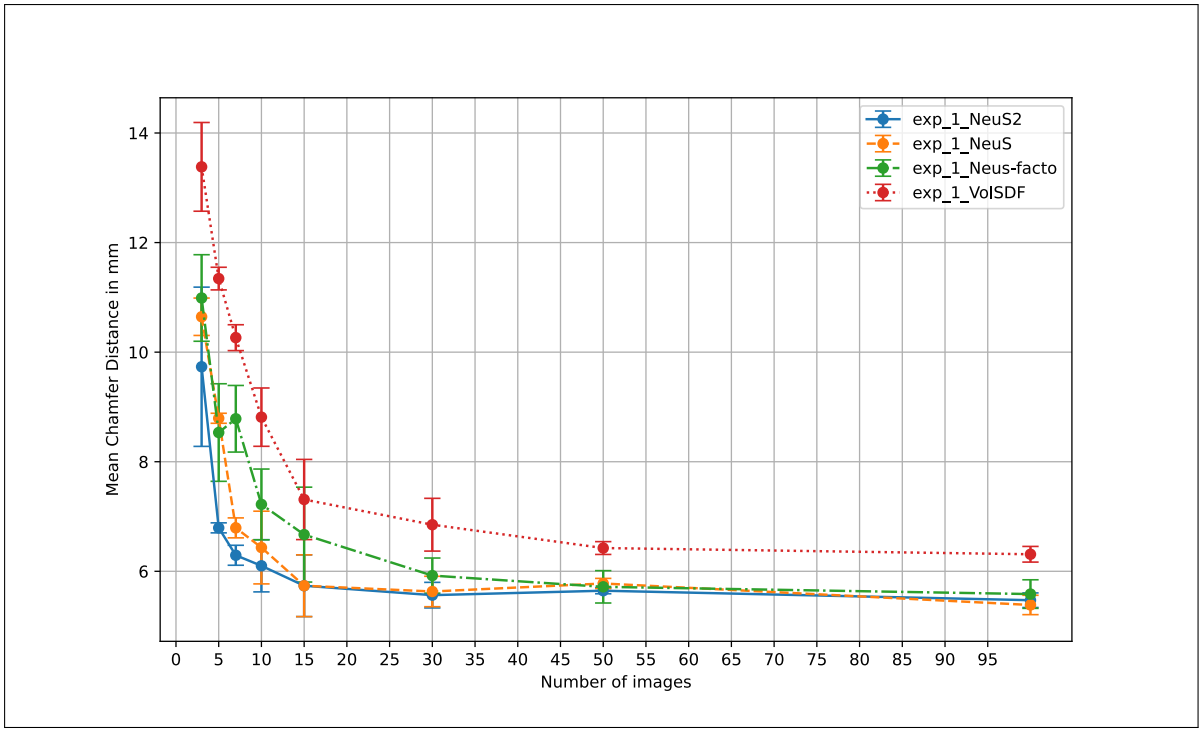


Figure 5.2: Performance analysis of top four performing methods with varying input images.

5.1.2 Qualitative Results

We presented the top four performing methods—NeuS, NeuS2, Neus-facto, and VolSDF. Each figure includes nine 3D models. The leftmost model represents the ground truth mesh, serving as the benchmark for accuracy. The top row of each figure illustrates the reconstruction results obtained from 3, 5, 7, and 10 images from left to right. In the bottom row, the reconstruction results using 15, 30, 50, and 100 images are displayed in the left-to-right sequence.

In the provided figures, Figure 5.3 illustrates the resultant 3D output from the NeuS2 method, while Figure 5.4 presents visual representations of 3D models generated using the NeuS method. Additionally, Figure 5.5 shows the results of the reconstruction using the NeuS-facto method, and Figure 5.6 shows the results achieved by the VolSDF method. These qualitative results visually validate the quantitative findings, demonstrating that although an increase in the number of input images leads to better reconstruction, the degree of improvement reduces significantly beyond a certain threshold (which is 15 in most methods). This plateau indicates that additional images contribute progressively less to enhancing reconstruction quality beyond this threshold. From the results, it is evident that both the NeuS2 and NeuS methods perform equally well overall. However, in sparse settings, specifically with 3, 5, 7, and 10 images, NeuS2 shows a slight performance advantage over NeuS.

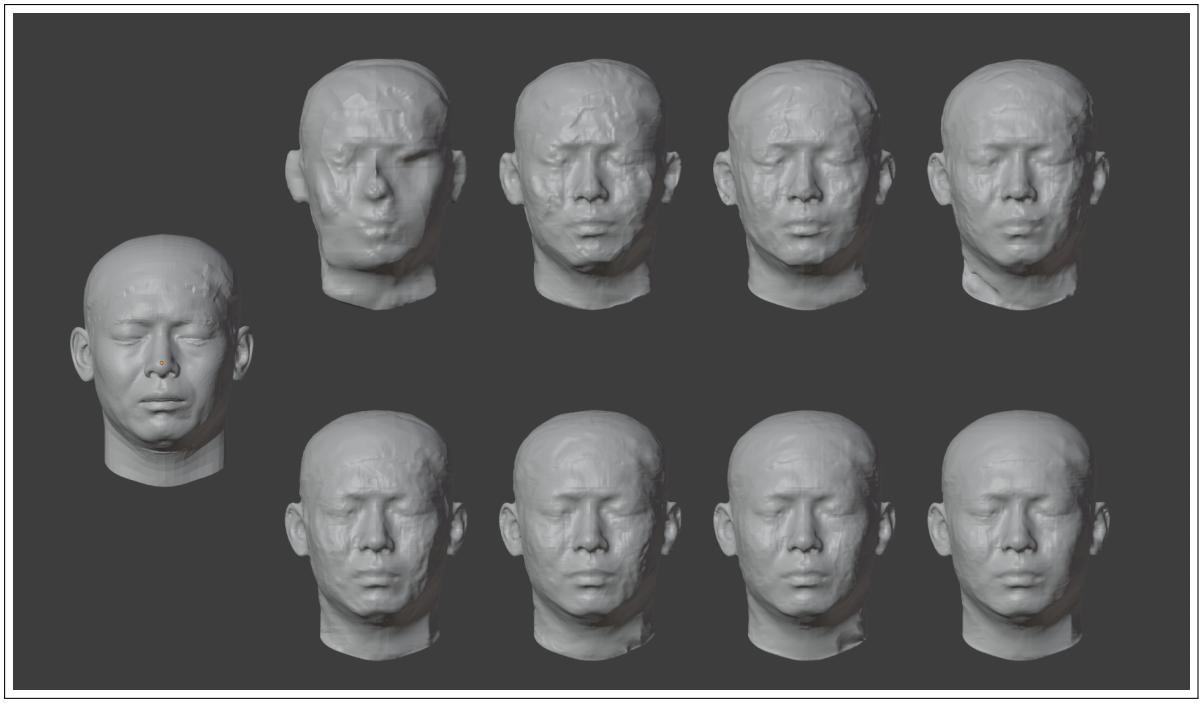


Figure 5.3: Qualitative results: 3D meshes constructed with the NeuS2 method. The ground truth mesh is in the top left. The top row features reconstructions using 3, 5, 7, and 10 images, and the bottom row displays reconstruction using 15, 30, 50, and 100 images.

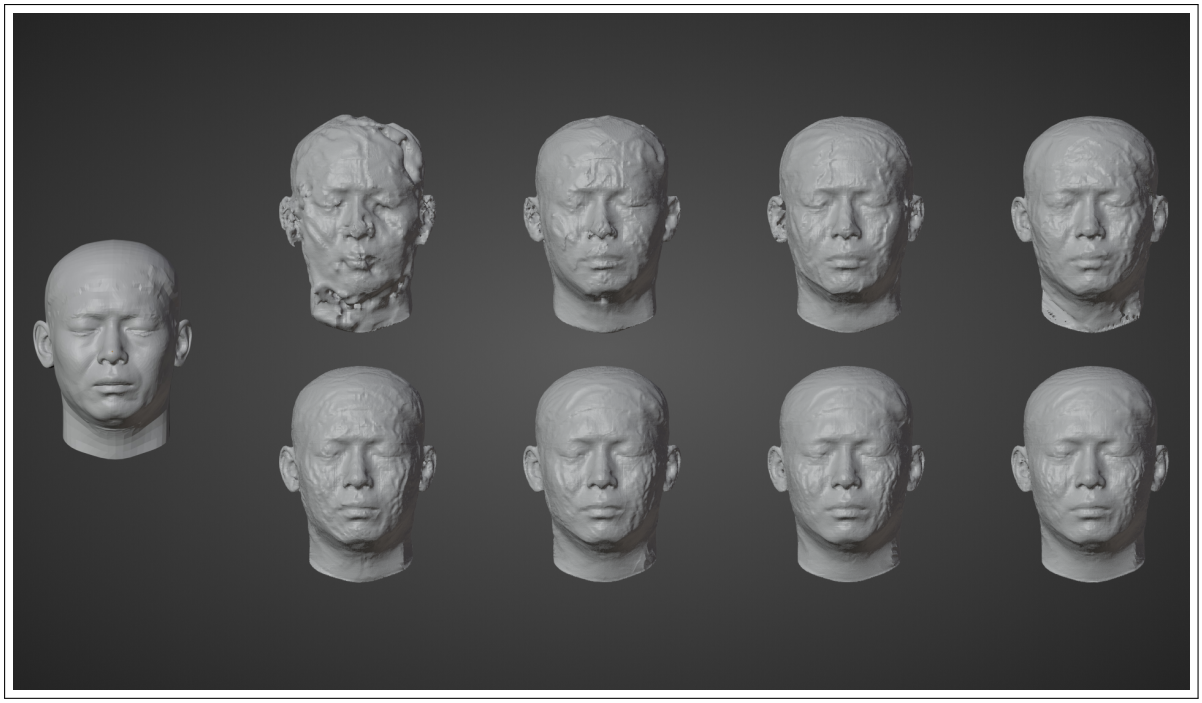


Figure 5.4: Qualitative results: 3D meshes constructed with the NeuS method. The ground truth mesh is in the top left. The top row features reconstructions using 3, 5, 7, and 10 images, and the bottom row displays reconstruction using 15, 30, 50, and 100 images.

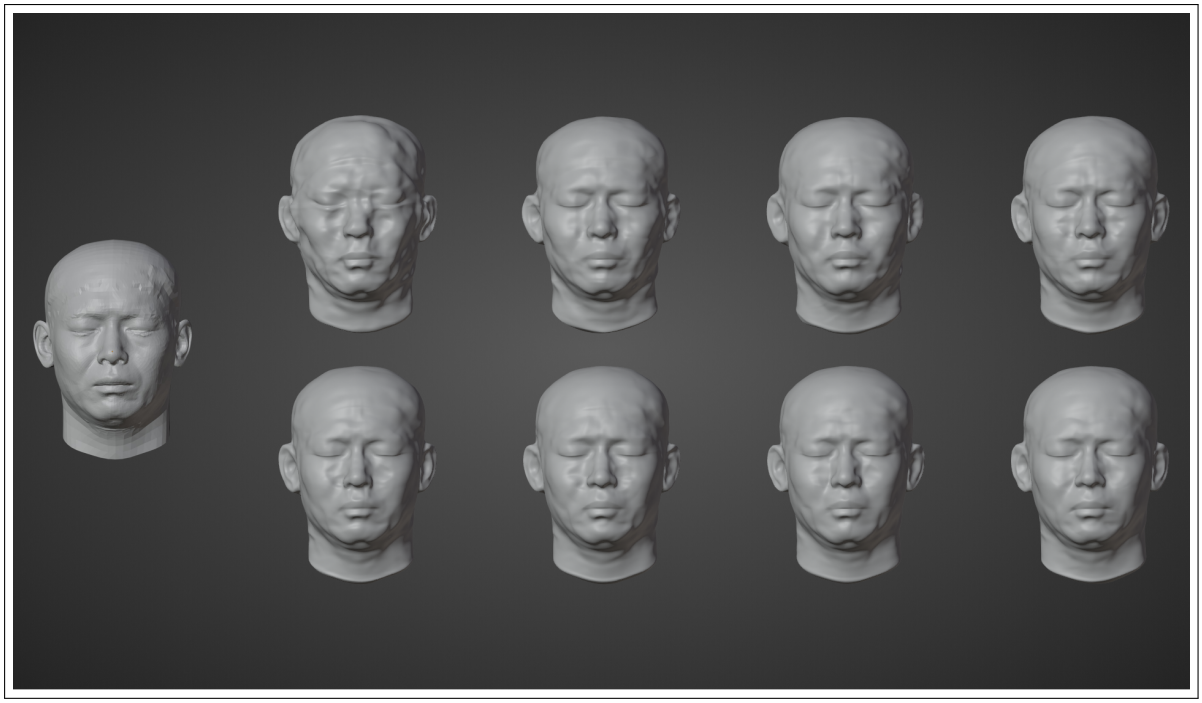


Figure 5.5: Qualitative results: 3D meshes constructed with the Neus-facto method. The ground truth mesh is in the top left. The top row features reconstructions using 3, 5, 7, and 10 images, and the bottom row displays reconstruction using 15, 30, 50, and 100 images.

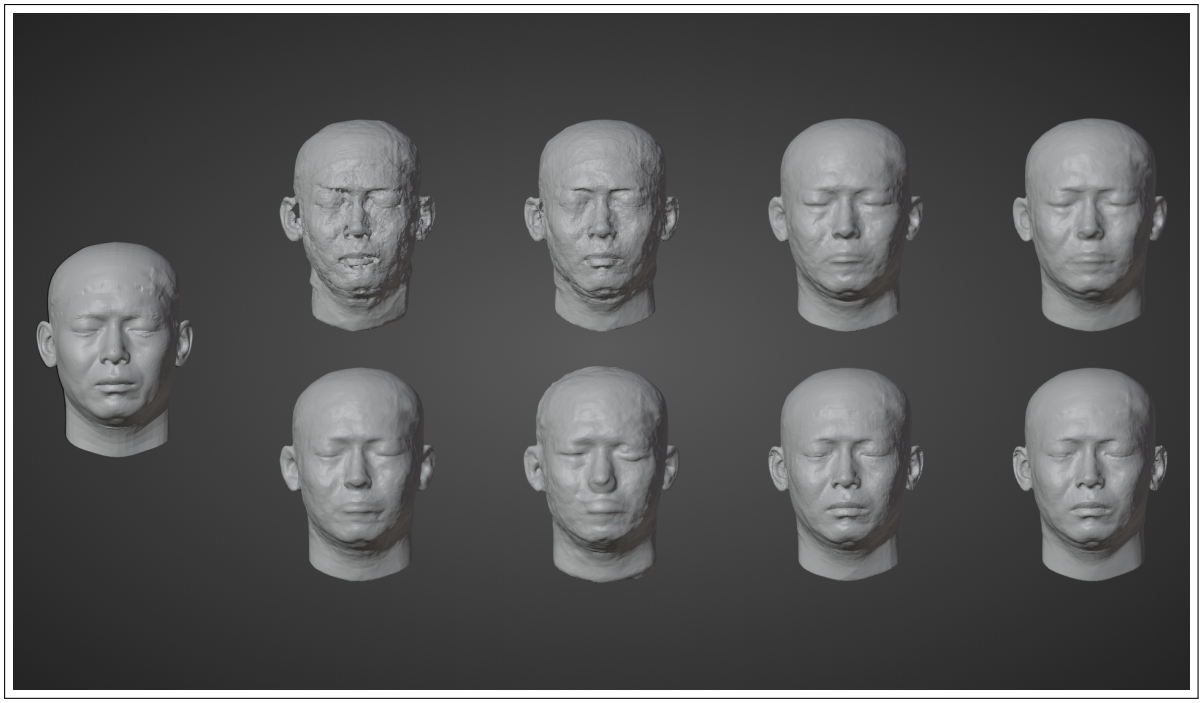


Figure 5.6: Qualitative results: 3D meshes constructed with the VolSDF method. The ground truth mesh is in the top left. The top row features reconstructions using 3, 5, 7, and 10 images, and the bottom row displays reconstruction using 15, 30, 50, and 100 images.

5.2 Experiment 2: Impact of Pose Perturbations

In this experiment, we estimated the influence of inaccuracies in camera calibration on the performance of 3D reconstruction of the human head. We utilized the same dataset of synthetic images as in Experiment 1. Each image set consists of noisy camera poses, which are generated by adding Gaussian noise with a 0 mean and 0.005 standard deviations into the actual camera poses to introduce controlled perturbations. These perturbed poses, along with the corresponding images, were utilized for 3D reconstruction to assess the robustness of all the selected SOTA methods against noisy poses. The experimental procedure is the same as that of Experiment 1, by utilizing eight distinct sets of images, each containing a different number of images: 3, 5, 7, 10, 15, 30, 50, and 100, across three different configurations. As a result, we reconstructed 24 human head models for each method (8 sets of images $\times$ 3 configurations). By systematically varying the number of input images used in the reconstruction process, we understand how the presence of noisy camera poses impacts the accuracy of 3D reconstructions across a spectrum of image densities. This comprehensive approach allows us to thoroughly evaluate the impact of inaccurate poses on the reconstruction quality. Dust3r [66] method is not included in this experiment because it solely relies on images for reconstructing an object's geometry.

5.2.1 Quantitative Results

Figure 5.7 illustrates the performance of all the selected SOTA models, and Figure 5.8 represents the top four performing methods in terms of their reconstruction quality when subjected to perturbated poses. The x-axis represents the number of images, while the y-axis indicates the Chamfer Distance in millimetres. Our results show that introducing Gaussian noise with 0 mean and 0.005 standard deviations significantly degrades 3D reconstruction performance when each method is trained with inaccurate poses. However, with the increase in the number of images, the influence of noise decreases, as evident in the results. Despite this noise impact reduction, it is not eradicated completely as the fine details are missing in the resultant 3D reconstruction even with a complete set of 100 images, which can be seen in the qualitative results. The reconstructed model generated with various image densities results in a high Chamfer Distance compared to models reconstructed under optimal conditions. This analysis highlights the persistent challenge of maintaining high reconstruction quality in the presence of pose inaccuracies despite increasing the number of images used.

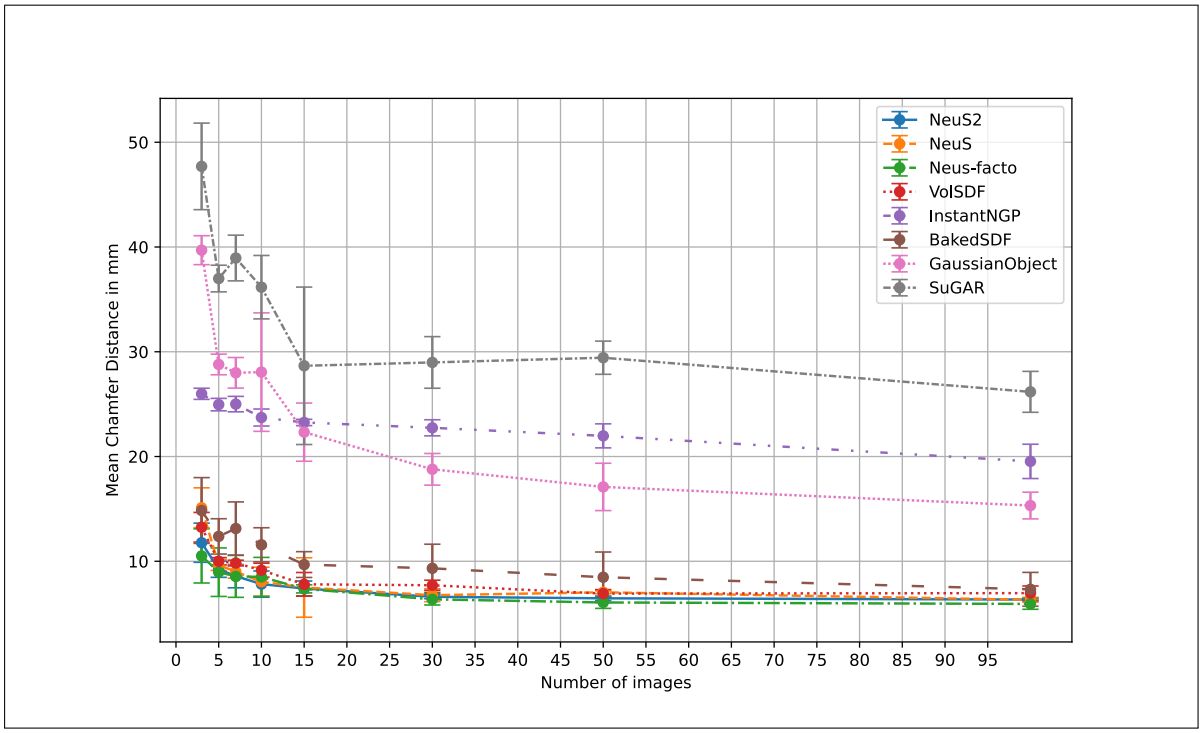


Figure 5.7: Performance evaluation of eight SOTA 3D reconstruction methods with varying input images under pose perturbations.

5.2.2 Qualitative Results

Figures **??** illustrates the quantitative results of the pose perturbation experiment for the top four performing models reconstructed using sets of 3, 5, 7, 10, 15, 30, 50, and 100 images. Among all the SOTA methods, NeuS-facto achieved the highest performance, consistently

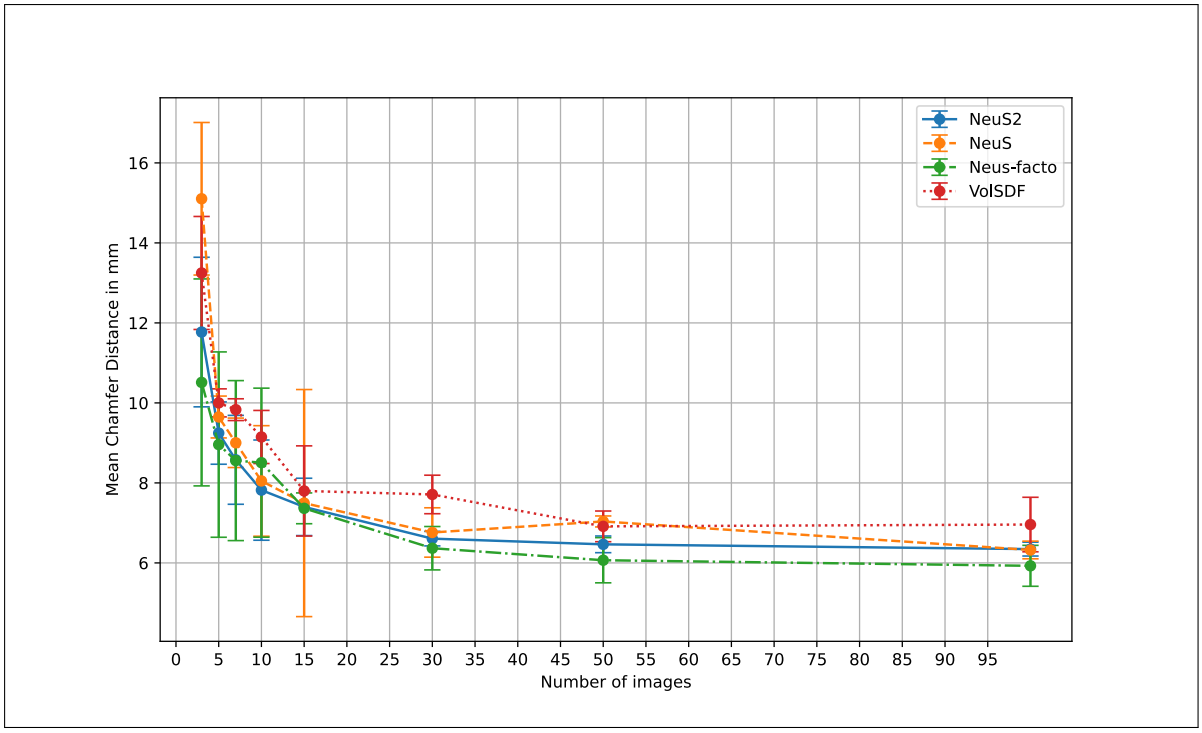


Figure 5.8: Performance evaluation of top four performing methods with varying input images under pose perturbation

giving better results across various image sets, followed by NeuS2, which also has similar robust performance across varying conditions. NeuS method achieved the next best results, while VolSDF was slightly behind, still showing good performance. These rankings were determined based on the Chamfer Distance metric, which evaluates the accuracy of the reconstructed 3D models. The results focus on the effectiveness of the top four methods in handling pose perturbations, showing their superior robustness and precision in 3D reconstruction tasks compared to the other models. Additionally, the performance trends from this experiment highlight the sensitivity of all the methods towards the noisy camera poses and their impact on the 3D reconstruction quality.

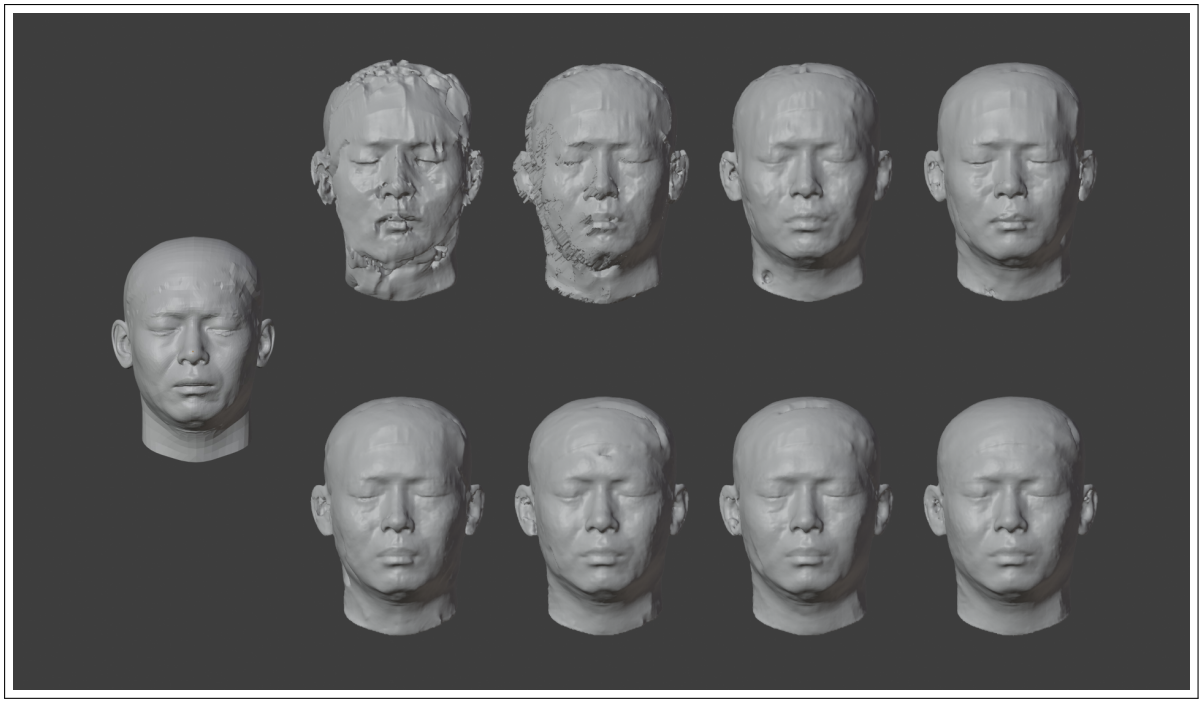


Figure 5.9: Qualitative results: 3D meshes constructed with the NeuS2 method under pose perturbation. The ground truth mesh is in the top left. The top row features reconstructions using 3, 5, 7, and 10 images, and the bottom row displays reconstruction using 15, 30, 50, and 100 images.

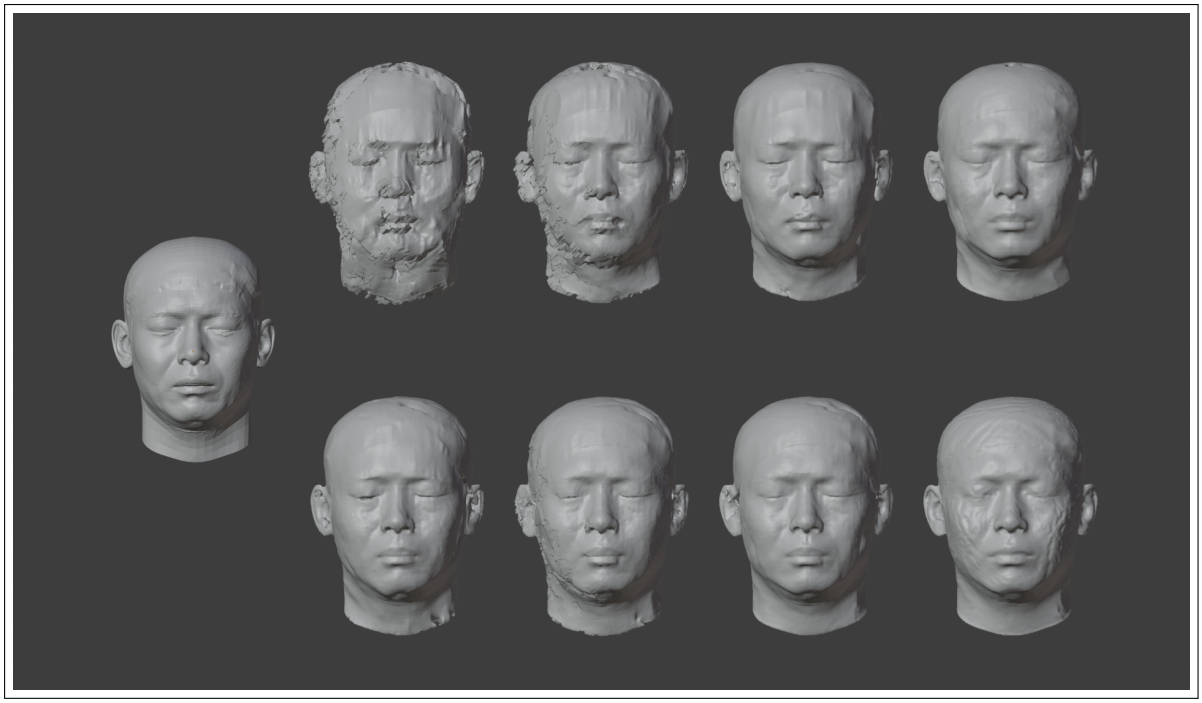


Figure 5.10: Qualitative results: 3D meshes constructed with the NeuS method under pose perturbation. The ground truth mesh is in the top left. The top row features reconstructions using 3, 5, 7, and 10 images, and the bottom row displays reconstruction using 15, 30, 50, and 100 images.

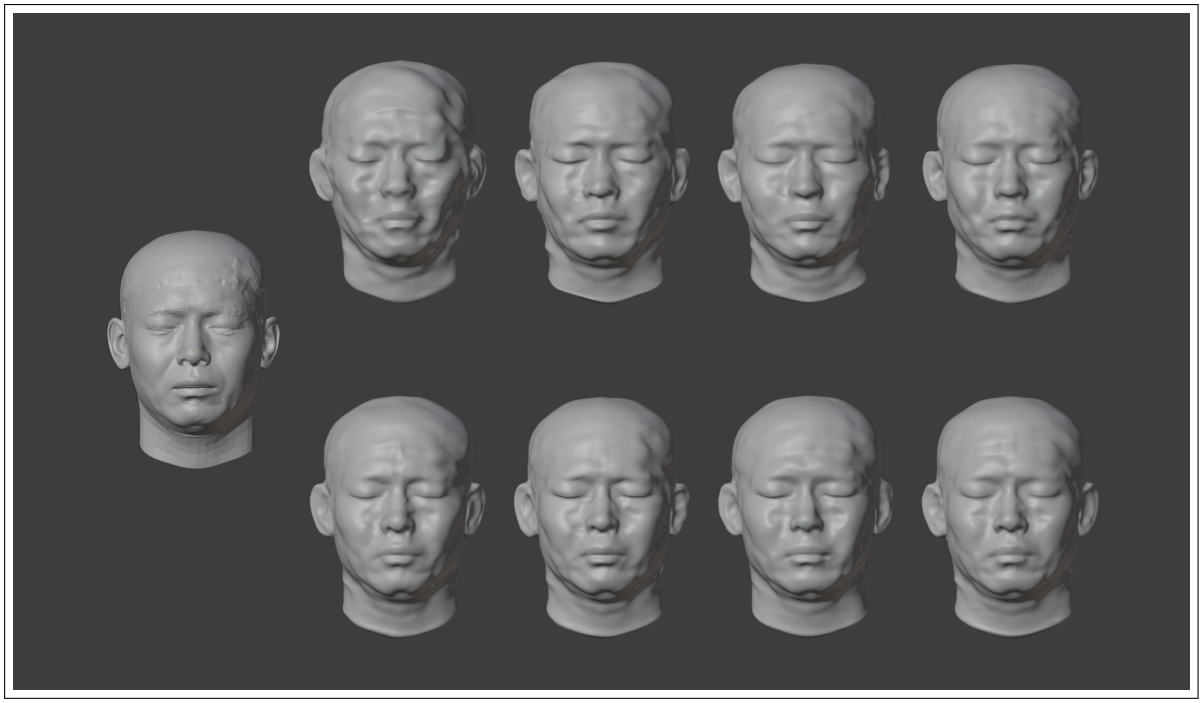


Figure 5.11: Qualitative results: 3D meshes constructed with the Neus-facto method under pose perturbation. The ground truth mesh is in the top left. The top row features reconstructions using 3, 5, 7, and 10 images, and the bottom row displays reconstruction using 15, 30, 50, and 100 images.

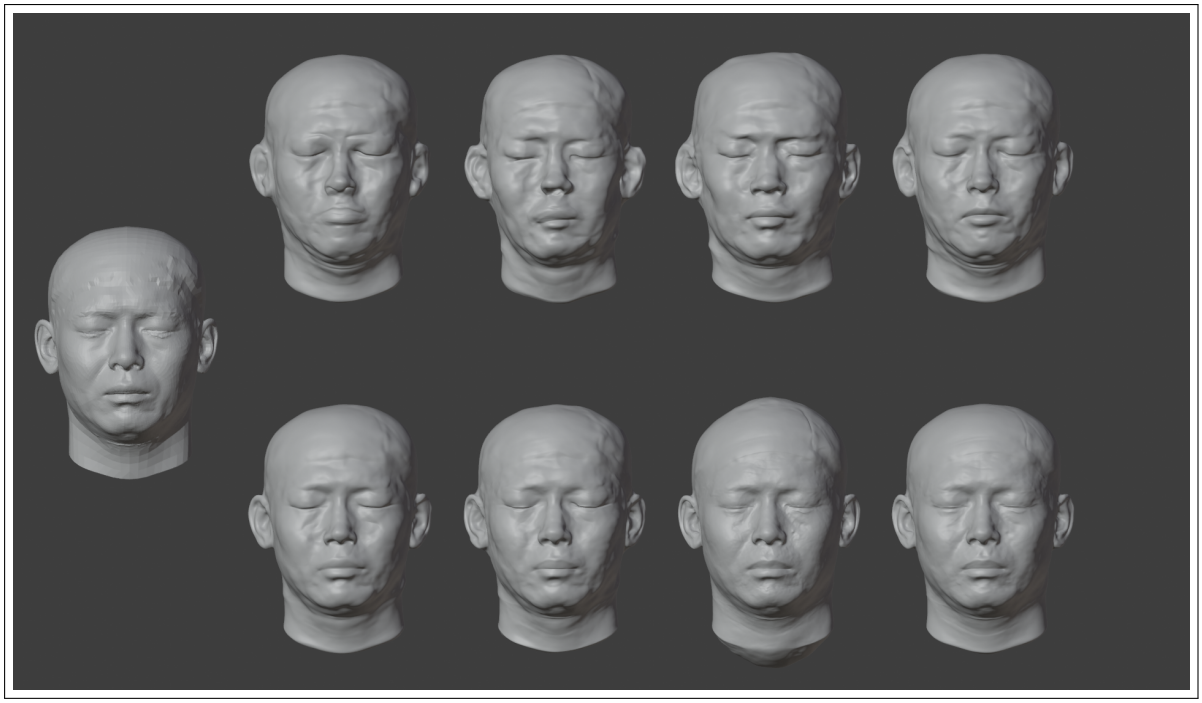


Figure 5.12: Qualitative results: 3D meshes constructed with the VolSDF method under pose perturbation. The ground truth mesh is in the top left. The top row features reconstructions using 3, 5, 7, and 10 images, and the bottom row displays reconstruction using 15, 30, 50, and 100 images.

5.3 Experiment 3

In this experiment, we evaluated the reconstruction quality of all the SOTA methods under challenging lighting conditions to examine their robustness against sub-optimal lighting conditions. We used low-intensity synthetic images generated in a controlled lighting setup for the reconstruction process. The detailed information about the data is explained in [data]. These low-lighting synthetic images, along with their respective camera poses, were employed as input for all the selected surface reconstruction methods to model the 3D human head. We then evaluated the quality of the reconstructed model using Chamfer Distance, focusing on their completeness.

5.3.1 Quantitative Results

Quantitative results (Performance analysis of all nine SOTA methods and top four best methods are illustrated in Figure 5.13 and Figure 5.14) show a significant degradation in reconstruction accuracy across all methods due to the sub-optimal lighting. This degradation is primarily attributed to the difficulty in capturing accurate features from 2D images, leading to compromised 3D representations that lack fine details. Additionally, the performance trends indicate that increasing the number of images generally improves reconstruction quality. However, even with the large sets of images, the 3D methods could not reconstruct fine

details of the face areas because those areas were hidden due to low lighting. These results highlight the impact of inadequate lighting on 3D reconstruction quality with various number of input images. NeuS2 and NeuS, both methods managed to perform relatively better under sparse settings (particularly with 5, 7, 10 and 15 images), while Neus-facto and BakedSDF performed better in dense settings (with 50 and 100 images).

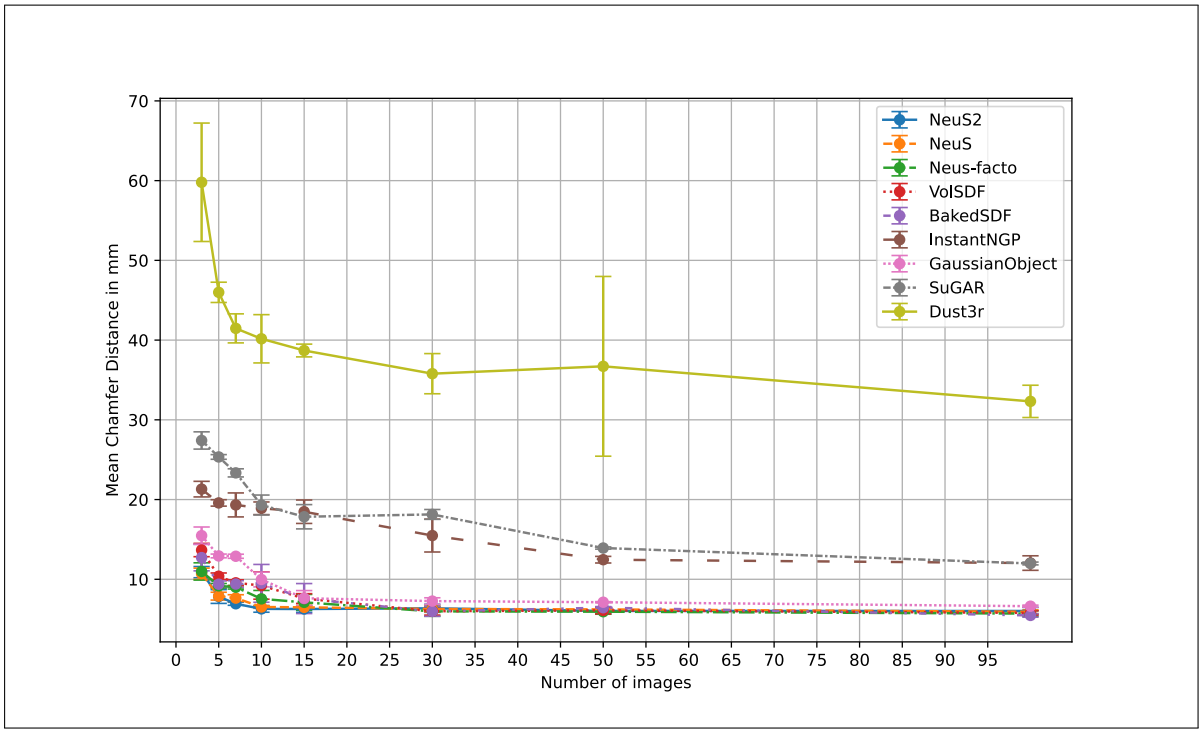


Figure 5.13: Performance analysis of nine SOTA 3D reconstruction methods with varying input images under sub-optimal lighting conditions

5.3.2 Qualitative Results

Qualitatively, NeuS2, NeuS, BakedSDF, and Neus-facto are the top-performing methods compared to other SOTA methods, yet all models showed varying degrees of inconsistencies and loss of detail (see Figures 5.15, 5.16, 5.17, and 5.18 respectively). This shows that the reconstructed models under insufficient lighting conditions demonstrate visible discrepancies, with fine details, particularly in poorly lit areas, such as the inside structure of ears, often inaccurately constructed. This qualitative analysis highlights the challenges of SOTA methods in maintaining high reconstruction fidelity under adverse lighting conditions.

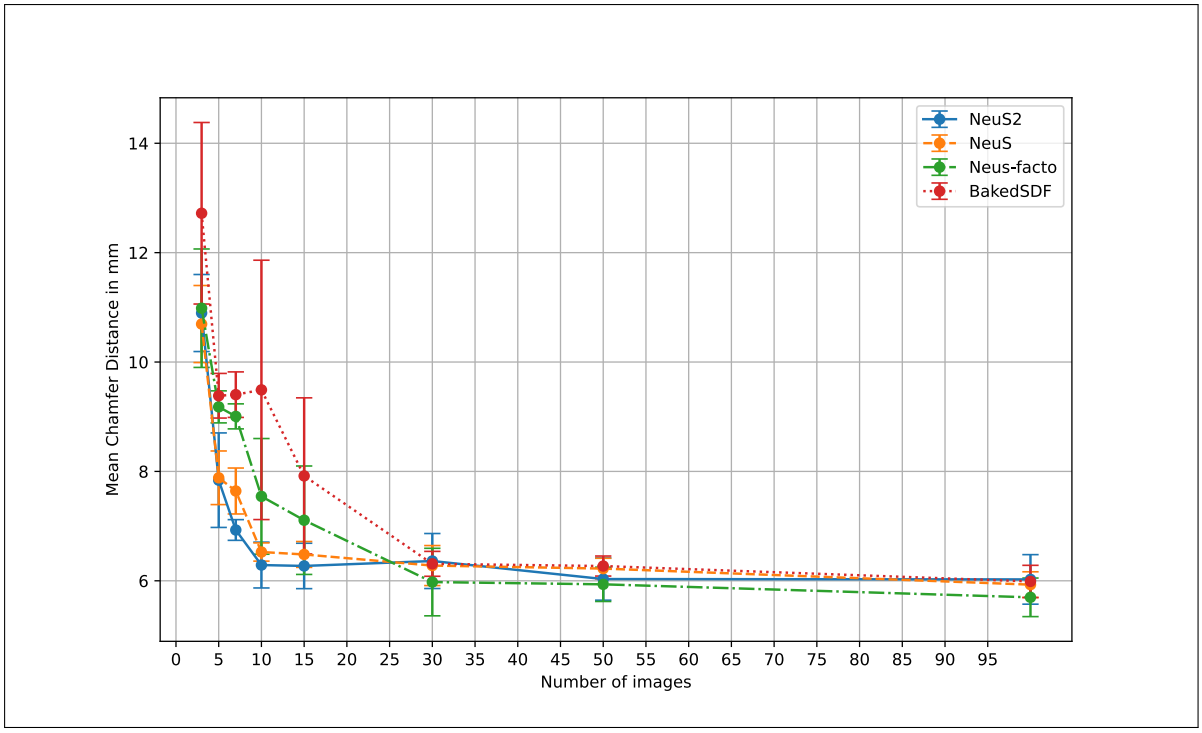


Figure 5.14: Performance analysis of top four performing methods with varying input images under sub-optimal lighting conditions

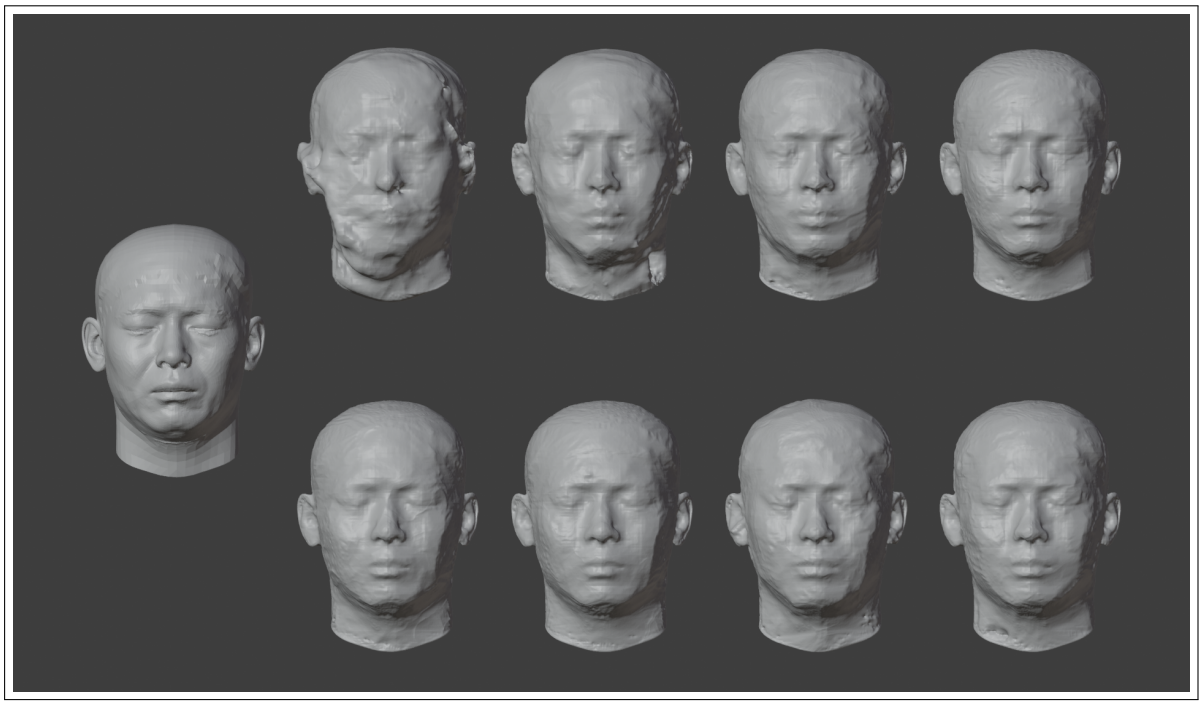


Figure 5.15: Qualitative results: 3D meshes constructed with the NeuS2 method under low-lighting conditions. The ground truth mesh is in the top left. The top row features reconstructions using 3, 5, 7, and 10 images, and the bottom row displays reconstruction using 15, 30, 50, and 100 images.

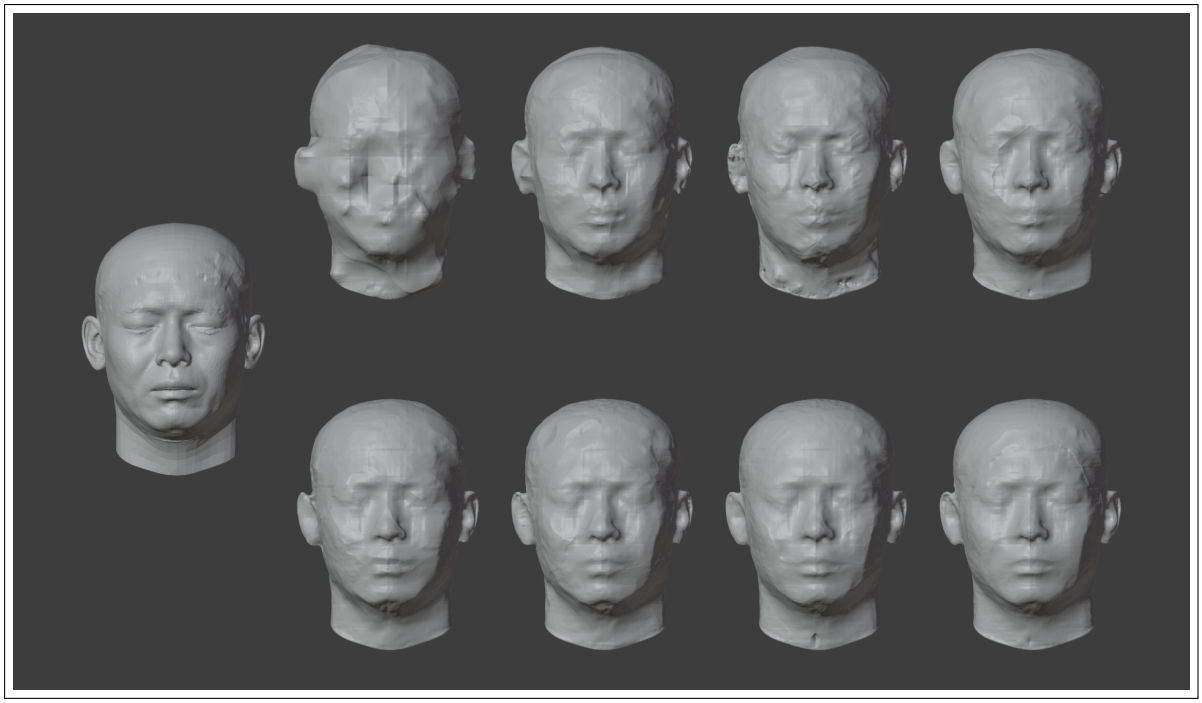


Figure 5.16: Qualitative results: 3D meshes constructed with the NeuS method under low-lighting conditions. The ground truth mesh is in the top left. The top row features reconstructions using 3, 5, 7, and 10 images, and the bottom row displays reconstruction using 15, 30, 50, and 100 images.

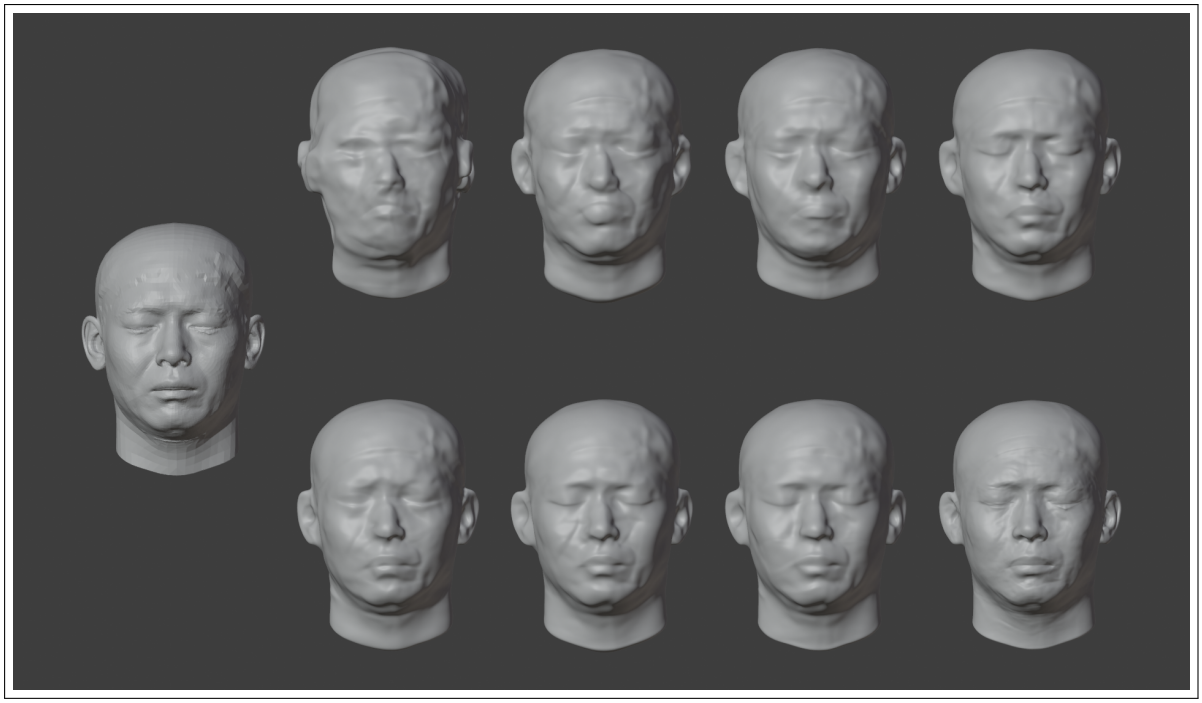


Figure 5.17: Qualitative results: 3D meshes constructed with the Neus-facto method under low-lighting conditions. The ground truth mesh is in the top left. The top row features reconstructions using 3, 5, 7, and 10 images, and the bottom row displays reconstruction using 15, 30, 50, and 100 images.

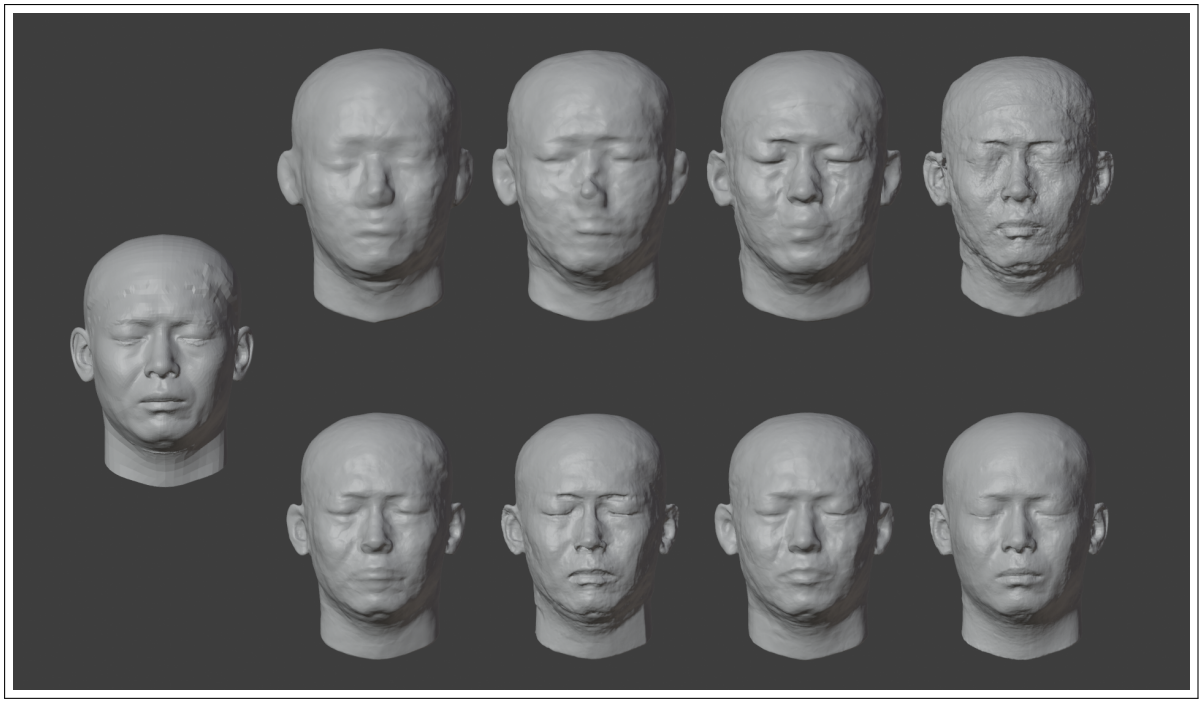


Figure 5.18: Qualitative results: 3D meshes constructed with the BakedSDF method under low-lighting conditions. The ground truth mesh is in the top left. The top row features reconstructions using 3, 5, 7, and 10 images, and the bottom row displays reconstruction using 15, 30, 50, and 100 images.

In conclusion, NeuS2, NeuS, and NeuS-facto methods have consistently shown superior performance across various challenging conditions. In optimal settings, we identified 15 images as the threshold for achieving reliable reconstruction quality. Our research also revealed that different types of inaccuracies affect the reconstruction models in diverse ways. More specifically, our results have shown that low lighting conditions lead to better reconstruction outcomes compared to the effects of noisy poses. Additionally, our results highlight the comparative performance of all nine SOTA methods, with neural-based methods outperforming Gaussian splatting-based methods (SuGAR [26] and GaussianObject [72]) in surface reconstruction tasks. Despite Dust3r's [66] innovative approach for reconstructing 3D models without camera poses and fast rendering, it ranked lower in accuracy compared to both neural and Gaussian splatting-based reconstruction methods.

Chapter 6

Limitations and Future Work

6.1 Limitations

The FaceScape dataset [73] presents a significant limitation despite its high-quality textures and finely detailed geometry. It does not accurately represent the actual human head in terms of scale. The scale mismatch directly affects the validity of our quantitative results when applied to real human heads. While the trends in the accuracy of the reconstructed models with different sets of input images remain consistent, the Chamfer Distance varies by a few millimetres due to the FaceScape models being larger than real human heads. Thus, when evaluating the Chamfer distance between the actual human head sizes and its 3D reconstruction, the distance tends to reduce due to this scale discrepancy. Hence, the 3D models from the FaceScape dataset may not generalize well to real-world applications, potentially reducing the direct impact of our results.

Another limitation of our research is that we only use synthetic data rendered through the FaceScape [73] model and Blender [14] software, as explained in the 4.4. While these generated images are nearly photorealistic, they lack essential elements found in real-world settings, such as realistic facial features, occluders such as glasses, or realistic backgrounds. Moreover, the 3D models from the FaceScape dataset are presented with a cap covering the head, which is not representative of typical human conditions as the hair is covered; this could hinder the generalizability of the neural rendering methods when applied to real-world settings.

The scope of our research is restricted by focusing only on front-facing images of the head, which proves beneficial for capturing detailed facial features, such as the contours of the eyes, nose, and mouth that are critical for many face analysis tasks. However, our method fails to adequately represent the back of the head and the area behind the ears. These regions are essential for creating a complete and accurate 3D model of the human head. This constraint limits the comprehensive analysis and reconstruction capabilities of the 3D models, potentially impacting applications that require a full 360-degree view of the head.

6.2 Future Work

The field of neural rendering is evolving rapidly, and further exploration of new models promises to provide valuable insights into the model's robustness and the potential applications of this new technology. Our hypothesis that synthetic data would adequately represent real-world data may be partially accurate, as the synthetic dataset does not include essential facial features, such as beards or hair. This limitation can affect the performance of neural rendering models in reconstructing the head with fine details. Therefore, future work should explore the evaluation of surface reconstruction models using real-world data, which includes occluders such as glasses and facial jewellery, to enhance the realism of 3D reconstructions. Another promising direction for future work is determining the optimal number of cameras and their spatial positioning to generate the best view that is required for accurate reconstruction. This involves determining the configuration with the minimum number of cameras to achieve precise 3D head reconstruction. Reducing the number of cameras necessary for data acquisition can significantly lower the overall costs associated with the process, making it more accessible and practical for broader applications.

Chapter 7

Conclusion

In this thesis, we have examined nine SOTA methods for 3D modelling of human head, analyzing their performance under various challenging conditions. We have primarily focused on neural rendering surface reconstruction methods, including VolSDF [75], BakedSDF [74], NeuS-Facto [77], NeuS [64], NeuS2 [67], and InstantNGP [48], along with two Gaussian splatting methods, SuGAR [26] and GaussianObject [72], and a multi-view stereo approach, Dust3r [66]. We quantified the performance of all the mentioned methods based on their ability to reconstruct 3D models of the human head with limited and substandard inputs. The reason behind this is to measure the influence of input parameters on the 3D reconstruction process, which helps to understand the robustness and effectiveness of the SOTA methods for 3D reconstruction when working with poor input data. In order to make this assessment, we performed a series of experiments by training the models with synthetic face images, which were rendered using the FaceScape [73] 3D model. We quantitatively measured the performance using the Chamfer Distance between the ground truth and the reconstructed 3D models.

Our results demonstrated a noticeable decline in the performance of 3D reconstructions under challenging conditions, such as sparse input images, inaccurate poses, and poor lighting conditions. These findings highlight the sensitivity of the surface reconstruction methods to the quality and quantity of the input data.

Particularly, our results showed that methods such as NeuS2, NeuS, and Neus-facto are more robust and consistently showed superior performance across various challenging conditions. Their performances highlight that all three methods (NeuS, NeuS2 and Neus-facto) are capable of producing reasonable reconstructions even with noisy input and limited input data, provided the views sufficiently cover the subject to extract maximum 3D information. Interestingly, we found that increasing the number of input images beyond 15 did not significantly enhance the reconstruction performance for the top-performing methods, indicating a threshold where additional images could not be beneficial and emphasizing the importance of selecting images wisely.

Further, we also evaluated the impact of perturbed camera poses on the quality of the 3D reconstruction of the human head model. Our findings indicate that noisy poses significantly degrade the quality of 3D reconstructions across all selected SOTA methods. Gaussian noise with a mean of 0 and a standard deviation of 0.005 in the camera poses was sufficient to

prevent the models from capturing the fine details of the faces; this points to the necessity for 3D reconstruction methods that rely less on precise camera calibrations and implicitly estimate the accurate camera poses, such as NopeNeRF [6] and Dust3r [66]. Our experiments covered a range of sparse to dense images and their corresponding perturbed poses and found that noisy poses with sparse inputs heavily impacted the performance of 3D reconstruction. However, a dense set of images with perturbated poses mitigated this impact, though the fine details of the face were still not captured. Moreover, we assessed the performance of the selected SOTA reconstruction methods in poorly lit conditions. We found a noticeable decline in the quality of 3D reconstruction, particularly in the regions of the face hidden by shadows or insufficient lighting. The darker regions of the face were not reconstructed well. Our work also demonstrated that different inaccuracies influence the reconstruction models in distinct ways. Specifically, our results indicate that low lighting conditions result in better reconstruction outcomes compared to the adverse effects caused by noisy poses.

Further, our results reveal that Gaussian splatting methods tend to be less effective at capturing fine surface details of the face compared to other NeRF methods. Gaussian splatting methods need more precision to reconstruct 3D geometry, leading to less accurate reconstructions. Despite not relying on precise camera poses, the dust3r [66] method often struggles to generate smooth surfaces, resulting in a rough and uneven reconstructed surface.

In conclusion, this thesis provides a detailed analysis of the capabilities and limitations of the current SOTA surface reconstruction methods based on neural radiance and Gaussian splatting. Among all the methods used in our experiments, NeuS2, NeuS, and Neus-facto methods demonstrated reasonable reconstruction results with sparse inputs under suboptimal conditions, achieving a Chamfer Distance between 6mm and 8mm for 15 views. Their impressive performance showcased their potential utility in practical applications where such challenges are unavoidable.

List of Figures

Bibliography

[1] Alankrita Aggarwal, Mamta Mittal, and Gopi Battineni. "Generative adversarial network: An overview of theory and applications". In: *International Journal of Information Management Data Insights* 1.1 (2021), p. 100004.

[2] Martin Arjovsky, Soumith Chintala, and Léon Bottou. "Wasserstein generative adversarial networks". In: *International conference on machine learning*. PMLR. 2017, pp. 214–223.

[3] Dejan Azinović et al. "Neural rgb-d surface reconstruction". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 6290–6301.

[4] Jonathan T. Barron et al. "Mip-NeRF 360: Unbounded Anti-Aliased Neural Radiance Fields". In: *CVPR* (2022).

[5] Sai Bi et al. *Deep 3D Capture: Geometry and Reflectance from Sparse Multi-View Images*. 2020. arXiv: `2003.12642 [cs.CV]`.

[6] Wenjing Bian et al. "Nope-nerf: Optimising neural radiance field with no pose prior". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 4160–4169.

[7] Volker Blanz and Thomas Vetter. "A Morphable Model For The Synthesis Of 3D Faces". In: *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*. 1st ed. New York, NY, USA: Association for Computing Machinery, 2023. ISBN: 9798400708978. URL: `https://doi.org/10.1145/3596711.3596730`.

[8] L. G. Brown. "Parametric correspondence and chamfer matching: Two new techniques for image matching". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 10.1 (1992), pp. 11–21.

[9] Tianlong Chen et al. "Aug-nerf: Training stronger neural radiance fields with triple-level physically-grounded augmentations". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 15191–15202.

[10] Zhiqin Chen et al. "Mobilenerf: Exploiting the polygon rasterization pipeline for efficient neural field rendering on mobile architectures". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 16569–16578.

[11] Nikolai Chinaev, Alexander Chigorin, and Ivan Laptev. "Mobileface: 3D face reconstruction with efficient cnn regression". In: *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*. 2018, pp. 0–0.

[12] Jongmoo Choi et al. "3D Face Reconstruction Using a Single or Multiple Views". In: *2010 20th International Conference on Pattern Recognition* (2010), pp. 3959–3962. URL: https://api.semanticscholar.org/CorpusID:3345824.

[13] Christopher B Choy et al. "3d-r2n2: A unified approach for single and multi-view 3d object reconstruction". In: *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*. Springer. 2016, pp. 628–644.

[14] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation. Stichting Blender Foundation, Amsterdam, 2018. URL: http://www.blender.org.

[15] François Darmon et al. "Robust Gaussian Splatting". In: *arXiv preprint arXiv:2404.04211* (2024).

[16] Frank Dellaert and Lin Yen-Chen. *Neural Volume Rendering: NeRF And Beyond*. 2021. arXiv: 2101.05204 [cs.CV].

[17] Bernhard Egger et al. "3d morphable face models—past, present, and future". In: *ACM Transactions on Graphics (ToG)* 39.5 (2020), pp. 1–38.

[18] Haoqiang Fan, Hao Su, and Leonidas J Guibas. "A point set generation network for 3d object reconstruction from a single image". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 605–613.

[19] Ben Fei et al. "3D Gaussian Splatting as New Era: A Survey". In: *IEEE Transactions on Visualization and Computer Graphics* (2024).

[20] Sara Fridovich-Keil et al. "Plenoxels: Radiance fields without neural networks". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 5501–5510.

[21] Kyle Gao et al. *NeRF: Neural Radiance Field in 3D Vision, A Comprehensive Review*. 2023. arXiv: 2210.00379 [cs.CV].

[22] Stephan J. Garbin et al. *FastNeRF: High-Fidelity Neural Rendering at 200FPS*. 2021. arXiv: 2103.10380 [cs.CV].

[23] D. Gatis. *Rembg*. 2022. URL: %5Curl%7Bhttps://github.com/danielgatis/rembg%7D.

[24] Baris Gecer et al. "Ganfit: Generative adversarial network fitting for high fidelity 3d face reconstruction". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 1155–1164.

[25] Ian Goodfellow et al. "Generative adversarial networks". In: *Communications of the ACM* 63.11 (2020), pp. 139–144.

[26] Antoine Guédon and Vincent Lepetit. "SuGaR: Surface-Aligned Gaussian Splatting for Efficient 3D Mesh Reconstruction and High-Quality Mesh Rendering". In: *arXiv preprint arXiv:2311.12775* (2023).

[27] Yudong Guo et al. "Cnn-based real-time dense face reconstruction with inverse-rendered photo-realistic face images". In: *IEEE transactions on pattern analysis and machine intelligence* 41.6 (2018), pp. 1294–1307.

[28] Peter Hedman et al. "Baking neural radiance fields for real-time view synthesis". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 5875–5884.

[29] Dan Hendrycks and Thomas Dietterich. "Benchmarking neural network robustness to common corruptions and perturbations". In: *arXiv preprint arXiv:1903.12261* (2019).

[30] Yang Hong et al. *HeadNeRF: A Real-time NeRF-based Parametric Head Model*. 2022. arXiv: `2112.05637 [cs.CV]`.

[31] Xin Huang et al. "Hdr-nerf: High dynamic range neural radiance fields". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 18398–18408.

[32] Ajay Jain, Matthew Tancik, and Pieter Abbeel. *Putting NeRF on a Diet: Semantically Consistent Few-Shot View Synthesis*. 2021. arXiv: `2104.00677 [cs.CV]`.

[33] Yan Jiao and Pin-han Ho. "Design of Binocular Stereo Vision System Via CNN-based Stereo Matching Algorithm". In: *2021 International Conference on Networking and Network Applications (NaNA)* (2021), pp. 426–431. URL: `https://api.semanticscholar.org/CorpusID:245014201`.

[34] Oğuzhan Fatih Kar et al. "3d common corruptions and data augmentation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 18963–18974.

[35] Michael Kazhdan, Matthew Bolitho, and Hugues Hoppe. "Poisson surface reconstruction". In: *Proceedings of the Fourth Eurographics Symposium on Geometry Processing*. SGP '06. , Cagliari, Sardinia, Italy, Eurographics Association, 2006, pp. 61–70. ISBN: 3905673363.

[36] Bernhard Kerbl et al. "3D Gaussian Splatting for Real-Time Radiance Field Rendering". In: *ACM Transactions on Graphics* 42.4 (July 2023). URL: `https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/`.

[37] Biwen Lei et al. "A hierarchical representation network for accurate and detailed face reconstruction from in-the-wild images". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 394–403.

[38] Zhaoshuo Li et al. "Neuralangelo: High-fidelity neural surface reconstruction". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 8456–8465.

[39] Chen-Hsuan Lin et al. *BARF: Bundle-Adjusting Neural Radiance Fields*. 2021. arXiv: 2104.06405 [cs.CV].

[40] Xiaoxiao Long et al. "Sparseneus: Fast generalizable neural surface reconstruction from sparse views". In: *European Conference on Computer Vision*. Springer. 2022, pp. 210–227.

[41] William E. Lorensen and Harvey E. Cline. "Marching cubes: A high resolution 3D surface construction algorithm". In: *SIGGRAPH Comput. Graph.* 21.4 (Aug. 1987), pp. 163–169. ISSN: 0097-8930. DOI: 10.1145/37402.37422. URL: https://doi.org/10.1145/37402.37422.

[42] Li Ma et al. "Deblur-nerf: Neural radiance fields from blurry images". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 12861–12870.

[43] Ricardo Martin-Brualla et al. "Nerf in the wild: Neural radiance fields for unconstrained photo collections". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 7210–7219.

[44] N. Max. "Optical models for direct volume rendering". In: *IEEE Transactions on Visualization and Computer Graphics* 1.2 (1995), pp. 99–108. DOI: 10.1109/2945.468400.

[45] Ben Mildenhall et al. "Nerf in the dark: High dynamic range view synthesis from noisy raw images". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 16190–16199.

[46] Ben Mildenhall et al. "NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis". In: *ECCV*. 2020.

[47] Stylianos Moschoglou et al. "3dfacegan: Adversarial nets for 3d face representation, generation, and translation". In: *International Journal of Computer Vision* 128 (2020), pp. 2534–2551.

[48] Thomas Müller et al. "Instant Neural Graphics Primitives with a Multiresolution Hash Encoding". In: *ACM Trans. Graph.* 41.4 (July 2022), 102:1–102:15. DOI: 10.1145/3528223.3530127. URL: https://doi.org/10.1145/3528223.3530127.

[49] Michael Niemeyer et al. *RegNeRF: Regularizing Neural Radiance Fields for View Synthesis from Sparse Inputs*. 2021. arXiv: 2112.00724 [cs.CV].

[50] Michael Oechsle, Songyou Peng, and Andreas Geiger. "Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 5589–5599.

[51] Pascal Paysan et al. "A 3D Face Model for Pose and Illumination Invariant Face Recognition". In: *2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance*. 2009, pp. 296–301. DOI: 10.1109/AVSS.2009.58.

[52] Naama Pearl, Tali Treibitz, and Simon Korman. "Nan: Noise-aware nerfs for burst-denoising". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 12672–12681.

[53] Charles R Qi et al. "Pointnet: Deep learning on point sets for 3d classification and segmentation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 652–660.

[54] AKM Shahariar Azad Rabby and Chengcui Zhang. *BeyondPixels: A Comprehensive Review of the Evolution of Neural Radiance Fields*. 2024. arXiv: 2306.03000 [cs.CV].

[55] Marie-Julie Rakotosaona et al. "Nerfmeshing: Distilling neural radiance fields into geometrically-accurate 3d meshes". In: *arXiv preprint arXiv:2303.09431* (2023).

[56] Benjamin Recht et al. "Do imagenet classifiers generalize to imagenet?" In: *International conference on machine learning*. PMLR. 2019, pp. 5389–5400.

[57] Will Rowan et al. *Fake It Without Making It: Conditioned Face Generation for Accurate 3D Face Reconstruction*. 2023. arXiv: 2307.13639 [cs.CV].

[58] Johannes Lutz Schönberger et al. "Pixelwise View Selection for Unstructured Multi-View Stereo". In: *European Conference on Computer Vision (ECCV)*. 2016.

[59] Seunghyeon Seo et al. "Mixnerf: Modeling a ray with mixture density for novel view synthesis from sparse inputs". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 20659–20668.

[60] Matthew Tancik et al. "Nerfstudio: A Modular Framework for Neural Radiance Field Development". In: *ACM SIGGRAPH 2023 Conference Proceedings*. SIGGRAPH '23. 2023.

[61] Chen Wang et al. "Benchmarking robustness in neural radiance fields". In: *arXiv preprint arXiv:2301.04075* (2023).

[62] Guangcong Wang et al. "Sparsenerf: Distilling depth ranking for few-shot novel view synthesis". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 9065–9076.

[63] Nanyang Wang et al. "Pixel2mesh: Generating 3d mesh models from single rgb images". In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 52–67.

[64] Peng Wang et al. "NeuS: Learning Neural Implicit Surfaces by Volume Rendering for Multi-view Reconstruction". In: *NeurIPS* (2021).

[65] Shan Wang, Xukun Shen, and Kun Yu. "Real-Time 3d Face Reconstruction From Single Image Using End-To-End Cnn Regression". In: *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE. 2021, pp. 3293–3297.

[66] Shuzhe Wang et al. *DUSt3R: Geometric 3D Vision Made Easy*. 2023. arXiv: 2312.14132 [cs.CV].

[67] Yiming Wang et al. "NeuS2: Fast Learning of Neural Implicit Surfaces for Multi-view Reconstruction". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2023.

[68] Philippe Weinzaepfel et al. "Croco: Self-supervised pre-training for 3d vision tasks by cross-view completion". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 3502–3516.

[69] Fanzi Wu et al. "MVF-Net: Multi-View 3D Face Morphable Model Regression". In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), pp. 959–968. URL: https://api.semanticscholar.org/CorpusID:104291924.

[70] Cihang Xie et al. "Adversarial examples improve image recognition". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 819–828.

[71] Yiheng Xie et al. *Neural Fields in Visual Computing and Beyond*. 2022. arXiv: 2111.11426 [cs.CV].

[72] Chen Yang et al. *GaussianObject: Just Taking Four Images to Get A High-Quality 3D Object with Gaussian Splatting*. 2024. arXiv: 2402.10259 [cs.CV].

[73] Haotian Yang et al. *FaceScape: a Large-scale High Quality 3D Face Dataset and Detailed Riggable 3D Face Prediction*. 2020. arXiv: 2003.13989 [cs.CV].

[74] Lior Yariv et al. "BakedSDF: Meshing Neural SDFs for Real-Time View Synthesis". In: *arXiv preprint arXiv:2302.14859* (2023).

[75] Lior Yariv et al. "Volume rendering of neural implicit surfaces". In: *Thirty-Fifth Conference on Neural Information Processing Systems*. 2021.

[76] Alex Yu et al. "pixelnerf: Neural radiance fields from one or few images". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 4578–4587.

[77] Zehao Yu et al. *SDFStudio: A Unified Framework for Surface Reconstruction*. 2022. URL: https://github.com/autonomousvision/sdfstudio.

[78] Sangdoo Yun et al. "Cutmix: Regularization strategy to train strong classifiers with localizable features". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 6023–6032.

[79] Hongyi Zhang et al. "mixup: Beyond empirical risk minimization". In: *arXiv preprint arXiv:1710.09412* (2017).

[80] Kai Zhang et al. *NeRF++: Analyzing and Improving Neural Radiance Fields*. 2020. arXiv: 2010.07492 [cs.CV].

[81] Michael Zollhöfer et al. "State of the art on monocular 3D face reconstruction, tracking, and applications". In: *Computer graphics forum*. Vol. 37. 2. Wiley Online Library. 2018, pp. 523–550.

Erklärung

Ich versichere, dass ich die Arbeit ohne fremde Hilfe und ohne Benutzung anderer als der angegebenen Quellen angefertigt habe und dass die Arbeit in gleicher oder ähnlicher Form noch keiner anderen Prüfungsbehörde vorgelegen hat und von dieser als Teil einer Prüfungsleistung angenommen wurde. Alle Ausführungen, die wörtlich oder sinngemäß übernommen wurden, sind als solche gekennzeichnet.

Ich bin damit einverstanden, dass die Arbeit veröffentlicht wird und dass in wissenschaftlichen Veröffentlichungen auf sie Bezug genommen wird.

Der Universität Erlangen-Nürnberg, vertreten durch den Lehrstuhl für Graphische Datenverarbeitung, wird ein (nicht ausschließliches) Nutzungsrecht an dieser Arbeit sowie an den im Zusammenhang mit ihr erstellten Programmen eingeräumt.

Erlangen, den 03.06.2024

(Mohit Choithwani)