3D SHAPE FIDELITY MEASUREMENT, RECOVERY, AND CREATION

by

Tianyu Luan

May 20th, 2024

A dissertation proposal submitted to the
Faculty of the Department of Computer Science and Engineering of
the University at Buffalo, State University of New York
in partial fulfilment of the requirements for the
degree of

Doctor of Philosophy

Department of Computer Science and Engineering

Table of Contents

Abstract

Augmented Reality (AR) and Virtual Reality (VR) technologies show great potential in fields like education, remote work, retail, real estate, entertainment, *etc*.. High fidelity in AR and VR is important for user immersion, presence, and emotional engagement. Studies show that realistic scenes increase immersion and emotional arousal, while a strong sense of presence boosts emotional intensity. However, current methods for creating high-fidelity 3D worlds are limited, with existing metrics like chamfer distance not accurately measuring realism. Our research tackles these issues with a five-part plan: (1) Create a user study dataset for initial human fidelity annotations, (2) Design a reality-reference 3D shape metric to measure fidelity differences and adjust it with user study data, (3) Turn this metric into a loss function to guide high-fidelity 3D shape reconstruction, (4) Develop a non-reference 3D shape metric for measuring fidelity without real-world references, and (5) Use this non-reference metric to guide 3D shape generation for higher-fidelity results. Preliminary results include a user study dataset with 1,008 videos of 3D objects with different distortions, scored by 868 participants providing 24,304 scores. We developed a 3D shape metric aligned with human perception and used it to enhance the fidelity of human hand mesh reconstructions. This research aims to set a standard for assessing fidelity based on human perception, improving high-fidelity 3D reconstruction and generation.

1 Backgrounds

Augmented Reality (AR) and Virtual Reality (VR) technologies have shown their vast potential across various domains. In education, VR enables immersive learning environments that enhance students' comprehension and memory through virtual field trips and scientific experiments. For remote work, VR is used to replicate office settings, allowing team members to engage in face-to-face meetings and collaborative efforts in virtual spaces, thus boosting communication efficiency and team bonds. In retail, AR allows consumers to try on clothing or visualize how furniture will look in their homes before purchasing, using smartphones or specialized glasses. Moreover, VR is employed in the real estate sector to offer virtual tours of properties, enabling potential buyers to remotely explore and experience the layout of homes. Finally, AR and VR are popular in the entertainment and gaming industries, providing unparalleled gaming experiences and interactive opportunities.

In AR & VR applications, fidelity is a hard-to-evaluate yet crucial factor. Previous works show evidence that high fidelity enhances user immersion, presence, and co-presence, and hence brings emotional impact on the user such as emotional arousal, enhancement, and emotional interaction.

Fidelity, immersion, and emotional arousal. In psychology, immersion is defined as the degree to which an individual feels absorbed by or engrossed in a particular experience [74]. For immersions, experiments in [72] found that looking at the unrealistic scene significantly lowers the immersion feeling score of the subjects compared with looking at the realistic scene. Subsequently, the sense of immersion also impacts people's emotional arousal. Very recent research [33] analyze previous works including [6, 25, 57, 81, 83] and found that higher immersions of nature exposure would significantly decrease the arousal of fatigue in the subjects.

Fidelity, presence, and emotional enhancement. Presence is defined as an experience of being in one place or environment, even when one is physically situated in another [74]

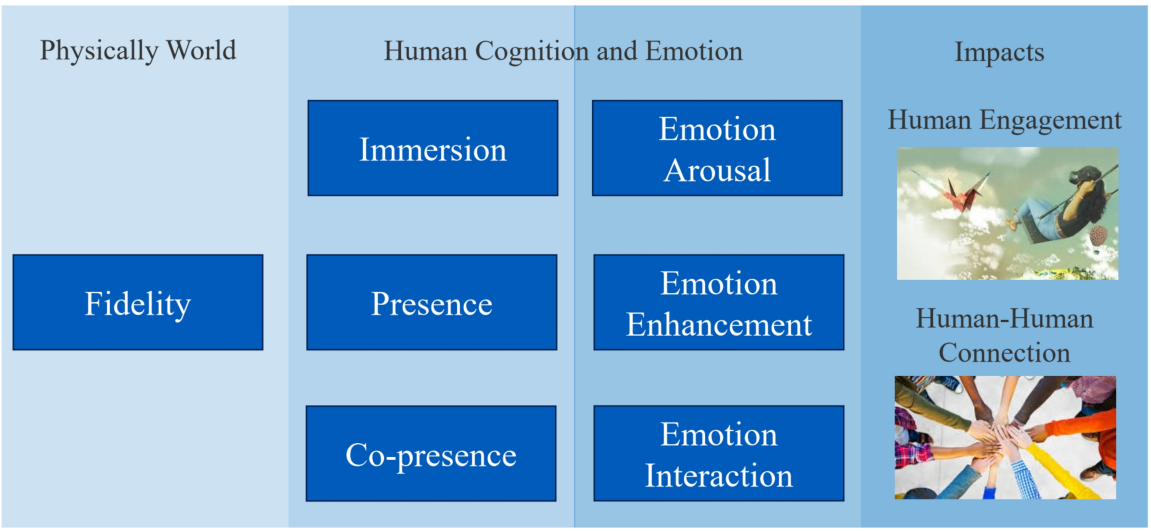

Figure 1: The motivation of acquiring high-fidelity in 3D virtual world.

(sometimes also called *situated immersion*). [54] proposed a framework named *Servotte-Ghuysen framework* to analyze the sense of presence. In the Servotte-Ghuysen framework, fidelity serves as a crucial system factor for users to have a sense of presence, and a high-fidelity environment can enhance users' sense of presence. Research on VR-related psychology also found that the sense of presence enhances the intensity of people's emotional feelings. For instance, [52] shows that the intensity of the subjects' happiness feeling shows a positive correlation with the sense of presence in a relaxing environment, and a negative correlation with the sense of presence in an anxious environment, while the intensity of the subjects' anxious and sadness gives the opposite results.

Fidelity, co-presence and emotional interaction. Co-presence exists when people sense that they are able to perceive others and that others are able to actively perceive them [32] found the sense of co-presence increases when the level of realism increases. The relation between co-presence and emotional interaction is not evidently clear yet. In psychology, people would empirically use human interaction behavior to measure the sense of co-presence, such as in [47]. Thus, it is highly possible that the co-presence would strongly affects the emotional interaction among people.

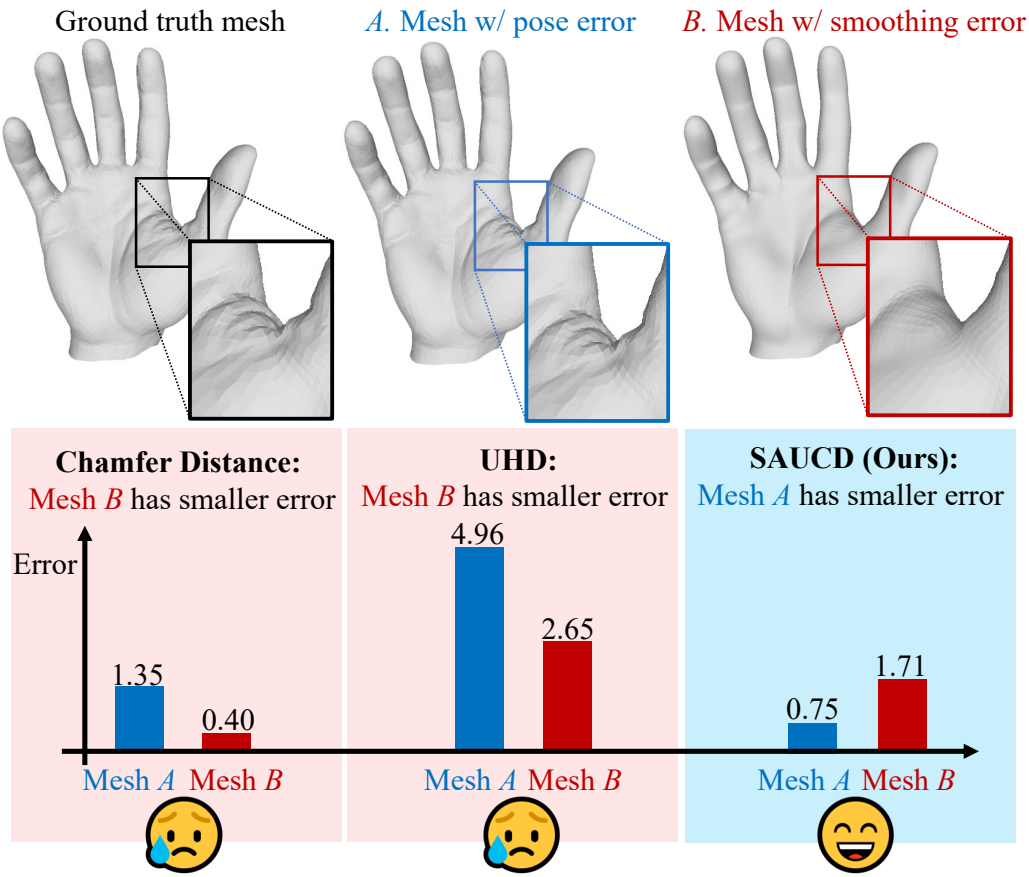

Figure 2: An example of how previous spatial domain 3D shape metrics (Chamfer Distance [8] and UHD [77]) deviate from human evaluation. We create Mesh A by adding a small pose error to the ground truth mesh, and by applying a large smoothing kernel to ground truth, we create Mesh B. Contrary to human perception, previous spatial domain metrics evaluate Mesh B better than Mesh A. This indicates that while they are sensitive to general shape differences, they tend to overlook high-frequency details. Note that different metrics use different units of measurement.

Although fidelity is so important for people's engagement in the VR-world and human-human connection, existing research on high-fidelity 3D worlds remains at a trivial exploratory stage. High-fidelity mesh reconstruction is a widely studied direction, and significant efforts have been made to enhance the details of 3D meshes. But we cannot answer if these efforts truly enhanced fidelity. These works lack rigorous standards to demonstrate whether they have indeed enhanced fidelity. Their evaluations rely on traditional metrics based on L2 distance, such as chamfer distance, and visualization. These evaluation standards have clear flaws when assessing realism. For instance, Figure 1 shows an example

concerning chamfer distance, illustrating the misalignment between this metric and human perception of fidelity. Specifically, when we remove the wrinkles from the ground truth mesh (resulting in Mesh B), the errors detected by previous metrics are not as significant as when we slightly change the pose of the hand (Mesh A). However, humans tend to sense a significant difference between ground truth and Mesh B, but barely recognize the difference between ground truth and Mesh A. Other previous works evaluated the results through visualization. These measurements may vary for different people, failing to capture a statistical understanding of the realism perceived by the population, that is, how realistic their results appear to the entire population. Moreover, visualizing a few results does not reflect the method's performance, especially since most of these methods do not claim to be randomly selected. These issues are not only present in the 3D reconstruction field. As generative AI evolves, the demand for generating high-fidelity 3D worlds is increasing, and more attempts are being made to produce high-fidelity works, yet no one has answered what fidelity is, how to measure fidelity, and what methods can truly reconstruct and generate fidelity.

We are trying to address these challenges. Initially, we discovered that human perception of fidelity follows predictable patterns. Various papers have studied human sensations of realism and immersion, revealing statistical consistencies in these perceptions. [41] found that the fidelity scores a group of subjects have on a certain object or behavior have an obvious consistency. For most objects, the standard derivation of scoring distribution would be less than 1 on a scoring scale of 1 to 5. This consistency among subjects enables us to model human's sense of fidelity. Specifically, it is necessary to collect data on people's reactions to various 3D objects and scenes first. Based on this data, we then plan to establish a standard for assessing fidelity—a standard that reflects the statistical expectation of evaluations from the entire dataset, not just individual responses. With this established standard, we would further utilize it as a loss function to pursue high-fidelity 3D reconstruction and generation. The detailed plan can be found in Sec. 4.

2 Related Works

2.1 High-fidelity 3D Reconstruction & Generation

Current high-fidelity 3D reconstruction & Generation techniques utilize three main types of shape representations: 3D meshes, voxels, and point clouds.

3D mesh-based methods. Previous works on mesh-based 3D reconstruction and generation have focused on the human body [76, 89, 27, 88, 37], human hands [68, 61, 38], human face [26, 75, 63, 11], and generation objects [59, 10, 73, 87, 86, 71]. These methods have made significant progress in 3D reconstruction & generation, but they have not yet answered the question of what is high-fidelity, how to measure fidelity, and how to really reconstruct/generate high-fidelity shapes.

Voxel-based methods. Voxel-based methods offer approaches to represent the 3D shape in voxels such as in [40, 84, 4, 90, 35, 58]. They are tasked with mapping three-dimensional spatial coordinates to corresponding color and volume density outputs, which are then integrated into final images through a process known as volume rendering. Despite these strengths, voxel-based methods encounter difficulties in accurately defining isosurfaces within the volume density, which are essential for representing the 3D geometry of the scene. The prevalent method involves applying heuristic thresholds to these density values to determine the isosurfaces. However, this approach tends to produce surfaces that are noisy and imprecise due to insufficiently constrained level sets, leading to representations that fail to accurately depict the structural details of the shape [67, 82]. Thus, in our work, we mainly focus on 3D mesh-based representation.

Point-cloud-based methods. Point-cloud-based reconstruction & generation also has a lot of existing works such as [95, 79, 48, 18, 23]. Many of them would have high-fidelity results. However, considering the sparse and unconnected nature of point cloud representation. high-fidelity point cloud results often require much finer-grained representation (probably

millions of points) and computational resources. By contrast, using 3D mesh representation is more computationally friendly when achieving the same level of fidelity.

2.2 Image Quality Assessment

Image quality assessment methods are close to our research in 3D shape evaluation. It could be a good reference and inspiration for our research on design 3D shape metrics.

Full-reference image quality assessment (FR-IQA). FR-IQA evaluates image quality by comparing a reference image to a distorted version, focusing on their dissimilarities. A traditional and widely used metric in this domain is the peak signal-to-noise ratio (PSNR). PSNR, known for its simplicity, calculates pixel-wise fidelity, providing a direct measure of image degradation. However, it falls short in accounting for the complexities of the human visual system (HVS), which does not strictly interpret image quality based on pixel accuracy. This discrepancy led Wang et al. to develop the Structural Similarity (SSIM) index, which assesses the similarity in local patches, offering a more perceptually-aligned measure of image quality [69]. This innovation has spurred extensive subsequent research, introducing more sophisticated, hand-crafted features that more closely mimic human perception. Notable developments include advanced metrics discussed in studies by [80, 12, 93, 30, 91]. These metrics enhance the traditional approach by integrating various perceptual aspects into the assessment of image quality.

No-reference image quality assessment (NR-IQA). No-reference image quality assessment (NR-IQA) presents unique challenges due to the absence of reference images to guide the evaluation. Within NR-IQA, there are two distinct subtasks: technical quality assessment [20] and aesthetic quality assessment [45]. Technical quality assessment is primarily concerned with the technical attributes of an image, such as sharpness, brightness, and noise. This approach is typically used to gauge the fidelity of an image relative to the original scene and assesses the accuracy of image acquisition, transmission, and reproduction processes. It focuses on objective measures that are quantifiable and replicable. In contrast, aesthetic

quality assessment deals with the subjective perceptions of viewers regarding an image's visual appeal. This subtask considers aesthetic elements like composition, lighting, color harmony, and overall artistic impression. The evaluation of aesthetic quality is inherently more subjective than technical quality assessment, heavily depending on individual preferences and cultural influences. Both subtasks of NR-IQA, despite their differing focuses, involve assessments that are influenced by factors such as lighting, color accuracy, and sharpness. Traditional NR-IQA methods typically rely on natural scene statistics (NSS) [92, 39, 44, 42]. However, recent advancements by [7] have introduced new features based on pseudo-reference images, which have shown significant improvements over methods based solely on NSS. These developments underscore the evolving nature of image quality assessment in the face of emerging technologies and methodologies.

2.3 Mesh Quality Assessment

Mesh quality assessment methods are also close to our research in 3D shape evaluation. However, existing approaches cannot satisfy our requirements to measure the fidelity of a given 3D shape. Previous works include the following categories.

Metrics in 3D mesh reconstruction. Chamfer Distance [8] is a popular metric used in 3D mesh reconstruction tasks such as those in [34, 70, 96, 85, 55, 51, 94, 28]. Other spatial domain metrics, such as 3D Intersection over Union (IoU) in [22, 13, 46, 17, 60, 53]. F-score in [66, 16, 5, 62], and Unidirectional Hausdorff distance (UHD) in [77] are commonly focused on the geometry accuracy of mesh shapes. These metrics can provide accurate geometry measurements, but they are not designed to align with human evaluation. Deep-learning-based methods such as Single Shape Fréchet Inception Distance [78] are also used in 3D reconstruction. While these metrics have the capacity to adapt from human evaluation, they are more like black boxes, with performances subject to dataset size and annotation bias. Moreover, most previous works miss out on user study validation to verify if their metrics align with human evaluation.

3D shape generation metrics. Multiple metrics have been used in 3D shape generation, such as Minimal Matching Distance (MMD) [3], Jensen-Shannon Divergence (JSD) [29], Total Mutual Difference (TMD) [77], Fréchet Pointcloud Distance (FPD) [56], *etc.*. These metrics are designed to measure the differences between the generated distributions, while our task is to build a metric to compare the shape of two meshes.

3D mesh compression and watermarking metrics. Previous works [64, 9, 31, 14] focused on evaluating mesh errors in mesh compression and watermarking. Since compression and watermarking pursue mesh errors that cannot be detected by humans, they mainly focus on barely noticed errors. However, our task is to build a metric that can handle generally occurring errors that happen in 3D reconstruction tasks and applications.

3 Problem Formulation

3.1 Fidelity

Fidelity is a personalized concept. The fidelity of the same object can be different in different people's eyes. Thus, the mathematical definition of fidelity must be based on the statistical result of a group of people's opinions. Apart from the shape, the context of the shape can also influence the human's cognition of fidelity. Typically, shape context includes color, lighting, materials, and backgrounds. Thus, in our research, we defined the fidelity of a 3D shape as:

$$F(x; c) = E_{s \in S} f(x; s, c), \tag{1}$$

where $f(x; s, c)$ is a scalar that represents the fidelity of shape x under context c in subject s's opinion. The larger the better. S is a collection of subjects. Thus, $F(x; c)$ is the fidelity of shape x under context c in subject set S. Theoretically, S could be any subject collection. In our research, we aim to find out the object fidelity in the opinion of the whole population, so we define S as the whole population, and we sample a subset $\tilde{S}$ from S to do our user study.

Here, we make the assumption that $\tilde{S}$ is uniformly sampled from the whole population S as many dataset annotations and user study tasks did.

To build a fidelity metric, we need to design a function to fit $F(x; c)$ in Eq. (1) as

$$\hat{F}(x; c) = \arg\min_{\hat{F}'} \Sigma_x (\hat{F}'(x; c) - F(x; c))^2, \tag{2}$$

where $\hat{F}(\cdot)$ is the fitting function of $F(\cdot)$. In practice, $\hat{F}(\cdot)$ could be designed as an analytic-based function or a neural network.

3.2 Reality-referenced Fidelity

Fidelity should be a non-reference concept, meaning $F(x; c)$ in Eq. (1) does not require a comparison with another shape. However, the measurement of fidelity could be challenging. Considering that the real objects in most cases have very high fidelity, using the real objects to serve as a reference for fidelity would make the evaluation of fidelity easier, while still giving us reasonable results. Thus, in certain applications such as reconstruction, we try to use reality-referenced fidelity as a substitute for fidelity. The reality-referenced fidelity is defined as:

$$F_r(x, x_0; c) = E_{s \in S} f_r(x, x_0; s, c), \tag{3}$$

where x_0 is the reference shape, $F_r(\cdot)$ and $f_r(\cdot)$ are the reality-referenced fidelity of subject set S and subject s, respectively. Different from $F(\cdot)$ and $f(\cdot)$, $F_r(\cdot)$ and $f_r(\cdot)$ are the lower the better. The rest of the variables are defined the same as in Eq. (1).

To build a Reality-referenced fidelity metric, we need to design a function to fit $F_r(x, x_0; c)$ in Eq. (3) as

$$\hat{F}_r(x, x_0; c) = \arg\min_{\hat{F}_r'} \Sigma_x (\hat{F}_r'(x, x_0; c) - F_r(x, x_0; c))^2, \tag{4}$$

where $\hat{F}_r(\cdot)$ is the fitting function of $F_r(\cdot)$. In practice, $\hat{F}_r(\cdot)$ could be designed as an

analytic-based function or a neural network.

3.3 High-fidelity Reconstruction

In high-fidelity reconstruction, we aim to reconstruct a 3D shape that has lower reality-referenced fidelity to the real ground-truth shape. Specifically, high-fidelity reconstruction can be defined as:

$$\hat{x_r} = \arg\min_{x} \hat{F}_r(x, x_0; c_0), x = R(I; c_0), \tag{5}$$

where $\hat{x_r}$ is the high-fidelity reconstructed shape, $R(\cdot)$ is a reconstruction function, which is typically and neural network. x_0 is the ground-truth shape. I is the input, which could be an image, a coarse shape, or other input data structures. Other variables are defined the same as in Eq. (3).

3.4 High-fidelity Generation

In high-fidelity generation, we aim to generate a 3D shape that has high fidelity. Specifically, high-fidelity reconstruction can be defined as:

$$\hat{x_g} = \arg\max_{x} \hat{F}(x; c_0), x = G(C; c_0), \tag{6}$$

where $\hat{x_g}$ is the high-fidelity generated shape, $G(\cdot)$ is a generative function, which is typically a generative neural network. C is the condition of $G(\cdot)$, which could be an image, a coarse shape, other input data structures, or a constant(unconditional generation). Other variables are defined the same as in Eq. (1).

4 Research Plan

As illustrated in Fig. 3, our research plan has 5 parts, covering dataset, metric, and model design: **(1)** Create a user study dataset to provide initial human annotations of fidelity. **(2)**

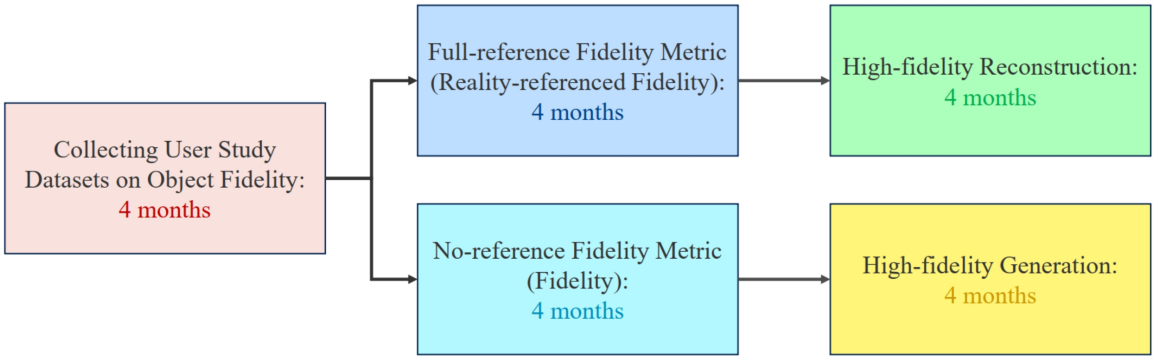

Figure 3: Research Plan. Please see Sec. 4 for further details.

We plan to design a reality-reference 3D shape metric to measure the fidelity difference between reconstructed 3D shape and real-world 3D shape and adjust it using the annotations in the user study dataset in (1). **3)** Transform this metric into a loss function and use it to guide 3D shape high-fidelity reconstruction. We will conduct this comprehensive plan in 2 years. Below, we provide an overview of the core concepts of each phase. **(4)** We plan to design a non-reference 3D shape metric to measure the fidelity without the reference of real-world shapes. **(5)** Use the non-reference metric to guide 3D shape generation, enabling the generative model to create higher-fidelity 3D shapes. We plan to spend 4 months on each of the parts. The detailed plan will be illustrated in the following sections.

4.1 User Study Dataset: 4 months

To develop a 3D model that aligns with human perception, it's essential to first gather the annotations of human preferences regarding 3D objects. Our approach involves creating a dataset of diverse 3D shapes, each subjected to various distortions that are encountered in real-world generative models. We will invite multiple subjects to evaluate each object and distortion, assigning scores that reflect their preferences. Through the aggregation and analysis of these scores, we can establish a set of human preference annotations for the 3D shapes in our dataset.

The scoring of our dataset would be 2-fold. For the first part of the annotation, we would

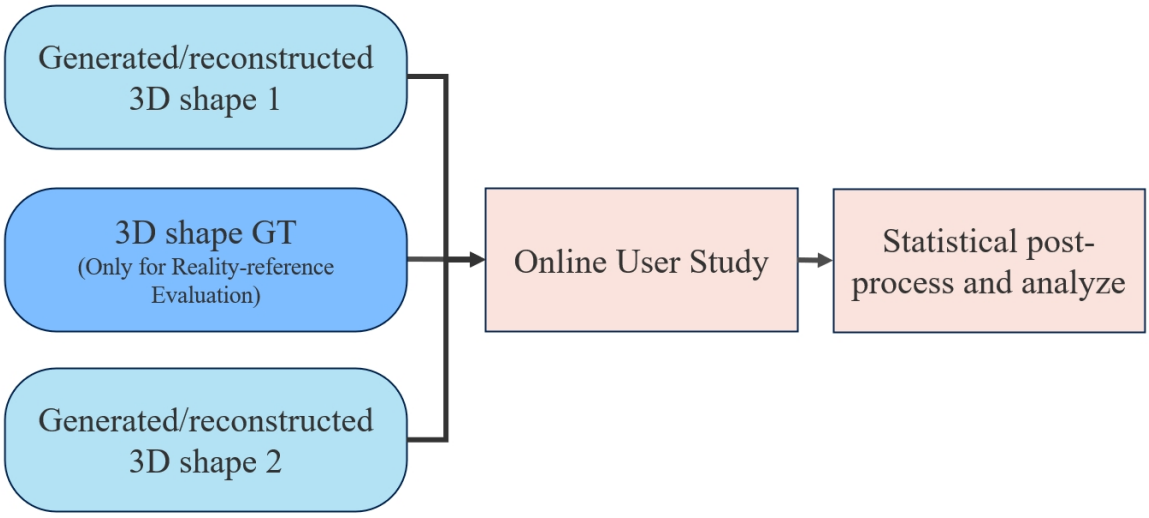

Figure 4: User study process design. Please see Sec. 4.1 for further details.

ask the subjects to do a reality-referenced evaluation using a tournament-style approach. Specifically, as shown in Fig. 4, we would present to the subjects 2 shapes that originated from the same 3D real object. For each time of the evaluation, we would give the subject a ground-truth shape as a reference. The subject will evaluate the differences between each given shape and the groundtruth shape, and give an opinion of which one is better. After several rounds of comparison, we can determine a ranking of which subject performs better, and then convert this ranking into scores. If we acquire each subject's fidelity opinion by directly asking them to give a score, the scales of scoring among different subjects are inconsistent. Subjects with a wider scoring range and more dispersed distribution will have a higher scoring weight. A tournament-style evaluation strategy can effectively avoid the bias of unequal scoring weights among different subjects. By using this method, we can obtain human annotations of object fidelity that are referenced to reality.

For the second part of the annotation, we continue to use a tournament strategy similar to the first part. Unlike the first part, here we do not provide the subjects with a ground truth. Instead, we need to compare which of the two shapes, both derived from the same real-world 3D shape, has higher fidelity without a reality reference. Please note that the

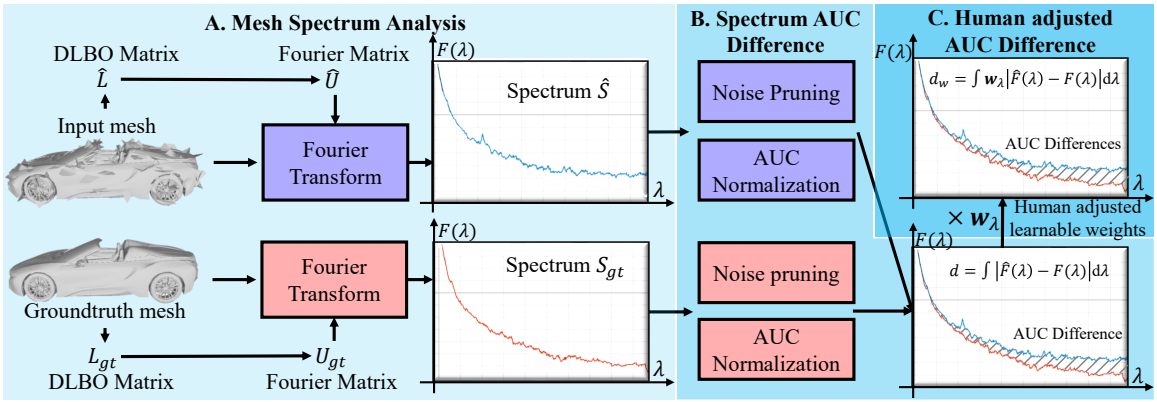

Figure 5: Reality-referenced 3D shape fidelity metric design.

scores obtained from this comparison are still non-referenced evaluations. The comparison here is merely a step of the tournament; it qualitatively assesses which one is better rather than quantitatively measuring the distance of fidelity like a full-referenced evaluation would do. Thus, we can obtain non-referenced human annotations of fidelity through this part of the annotation. Additionally, to remove outliers from the annotations, we will employ the Interquartile Range (IQR) method [15] in both parts of the annotation, a robust statistical technique designed to identify and exclude anomalous data points.

As mentioned in Section Sec. 3, the impact of context on shape evaluation is also significant. Thus, we plan to design various types of contextual changes to analyze the similarities and differences in human scoring results under different contexts. Our context variations will include common factors that affect human perception, such as changes in lighting, surface color of the shape, and texture variations, *etc.*. After analyzing the differences and similarities brought by each context, our shape fidelity score will be the average of these contexted scorings.

A preliminary dataset has been constructed to test our method, with further details on its design and analysis available in Sec. 5.1. In the future, we plan to enrich our dataset with a broader range of objects and distortion types generated from recent 3D shape generation approaches.

4.2 Reality-referenced 3D Shape Metric: 4 months

To develop a network that is aligned with human perception for 3D models, we recognize the necessity of annotations based on human preferences. However, the extensive process of annotating a diverse array of objects and their distortions is resource-intensive, presenting a challenge in scaling our dataset. This limitation raises concerns about potential overfitting if the network is trained directly on a restricted dataset. To mitigate this, we propose to design some analytic-based features captured from the observation of human perception.

Specifically, we plan to leverage frequency analysis of 3D shapes, which is refined by human evaluations from our user study dataset. The core idea is to treat each 3D shape as a signal and perform a frequency analysis on the signal. This method integrates the energy of the spectrum, with the consideration of the varying human sensitivity to different frequencies. By optimizing a vector that represents human sensitivity weights, we can construct a metric based on the user study dataset annotations. This frequency metric design is less susceptible to the limitations posed by the dataset's size and diversity, which minimizes the risk of overfitting and enables a more robust evaluation of 3D shapes that align with human perception.

An initial version of our metric has been developed, and preliminary comparisons with established metrics are detailed in Tab. 2. In the future, we plan to refine this metric further, leveraging an expanded dataset from a broader user study to enhance its accuracy.

4.3 High-fidelity 3D Hand Reconstruction: 4 months

With the development of a reality-referenced 3D metric, our next step is to utilize this metric as guidance to do high-fidelity 3D reconstruction. This network is designed to operate in the frequency domain, utilizing 3D mesh as the representation for shapes. Given that the frequency information of 3D mesh is essentially one-dimensional, we leverage signal processing techniques for the analysis and generation of 3D mesh shapes. This idea allows

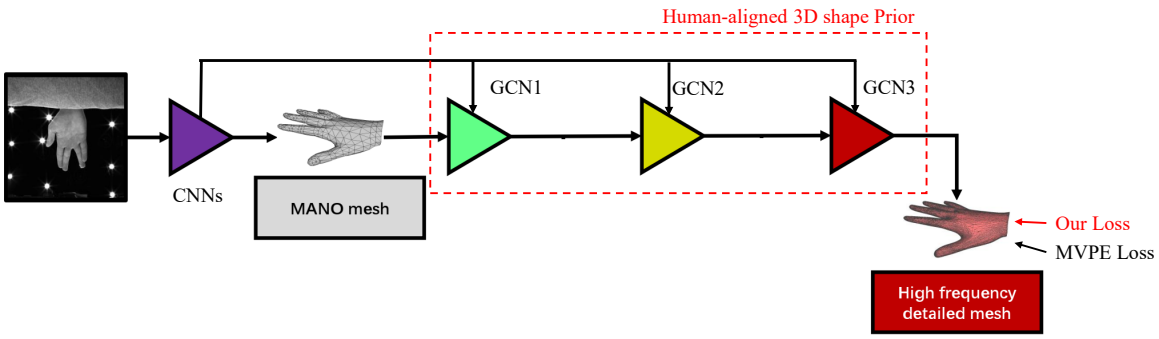

Figure 6: Simple 3D prior network on human hand reconstruction task.

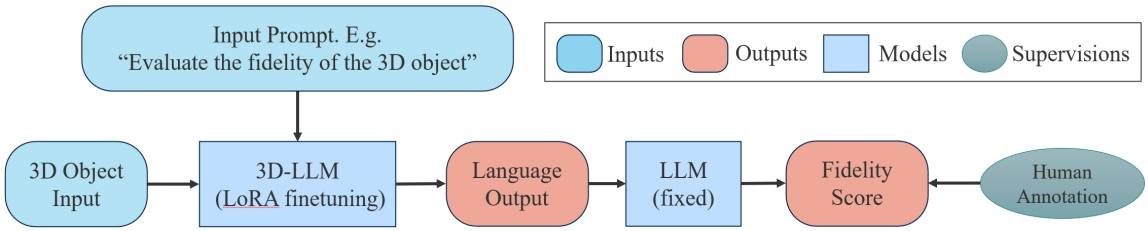

Figure 7: Non-referenced 3D shape fidelity metric design.

us to adapt the extensive range of existing models in one-dimensional signal processing tasks, such as those used for audio and medical signals. Furthermore, the metric previously developed will be converted into a loss function, ensuring that the reconstruction of the high-fidelity aligns more closely with the human perception of fidelity in the 3D world.

In Fig. 6 we show an initial example of our human-aligned 3D prior network designed for human hand inputs. The results can be found in Fig. 10. In future work, we will deploy our network for more generation 3D objects.

4.4 Non-referenced 3D Shape Fidelity Metric: 4 months

Considering the need for a deeper understanding of 3D shape for non-referenced metrics, and given that our training data is quite limited, we plan to leverage a pre-trained large language model to design our non-referenced 3D shape fidelity metric. As shown in Fig. 7, we intend to use a 3D-LLM model (such as [19]) as the metric function. Its input will be the

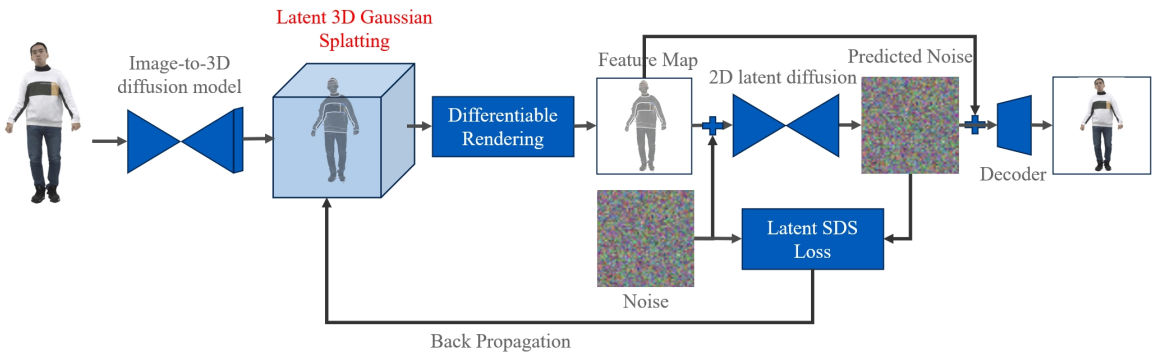

Figure 8: High-fidelity 3D human generation. Please see Sec. 4.5 for further details.

evaluated 3D shape and a single-sentence prompt. We will set the prompt as a fixed input, such as ”Evaluate the fidelity of the 3D shape”. Then, we will input the language output into another LLM to obtain a fidelity score. During training, we will fix the latter LLM and fine-tune the former 3D-LLM using LoRA [21]. With this design, we can fine-tune an existing LLM with minimal data to achieve the required metric.

4.5 High-fidelity 3D Human Generation: 4 months

We plan to use a voxel-based approach for high-fidelity human body generation. As shown in Fig. 8, our approach involves incorporating latent features into voxels and rendering these features to obtain feature maps from different angles. In the post-processing stage, we use techniques similar to latent diffusion to recover the feature maps and employ a decoder to generate high-fidelity rendered images. By calculating the SDS loss [50] during the diffusion process, we can optimize our voxel representation. Furthermore, with our specially designed non-referenced metric, we can transform the generated results into 3D mesh and further enhance the fidelity of the generated outcomes.

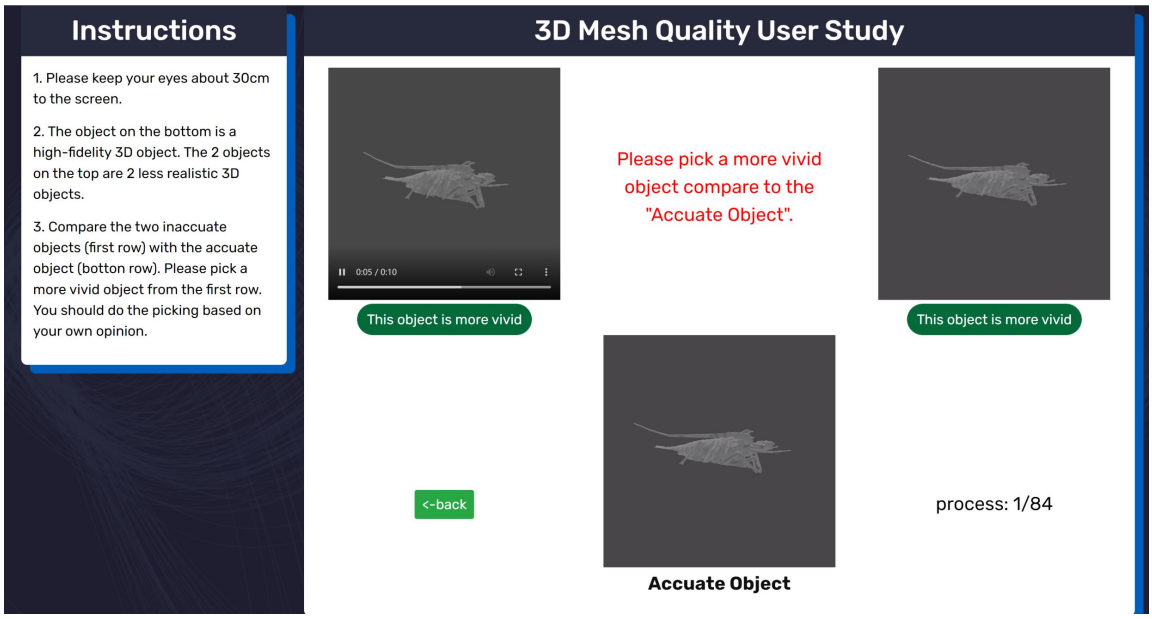

Figure 9: The preliminary system of our online user study [36].

5 Preliminary Results

5.1 User Study Dataset

We try to build a simple user study dataset as illustrated in Sec. 4.1. We choose 12 objects as ground truth 3D triangle mesh from public object/scene/human mesh datasets such as [43, 24, 65] and commercial datasets such as [1, 2]. These objects span various categories, including humans, animals, buildings, and plants. For each, we generate 7 types of common distortions in 3D reconstruction, each with 4 levels of severity, resulting in 28 variants per object. Each variant is rendered into three videos using different mesh materials, totaling 1,008 videos. We adopt a pairwise comparison method for scoring, similar to the approach in [49], ensuring a fair distribution of scores among the variants. Overall, 868 participants (536 males, 316 females, and 16 others) provided 24,304 scores across all objects and materials. In Fig. 9 we show our annotation website design and in Tab. 1, we show our dataset error analysis.

Dataset	Raw	w/ IQR removal
number of valid scores	24304	23775
Scoring range	$[0, 6]$	$[0, 6]$
95% confidence interval	0.318	0.303
Relative 95% confidence interval	5.33%	**5.04%**

Table 1: Dataset statistics and error analysis.

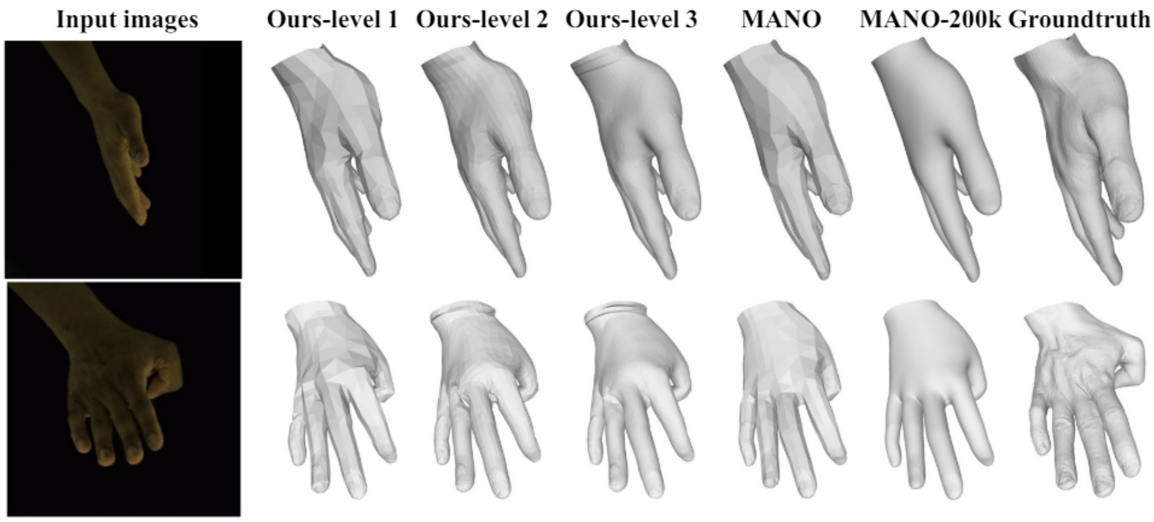

Figure 10: The preliminary results of high-fidelity human hand reconstruction [38].

5.2 Analytic-based 3D Shape Metric

We designed a human-aligned 3D shape metric as illustrated in Sec. 4.2. We optimized and evaluated our metric on the previously built dataset. The results are shown in Tab. 2.

5.3 High-fidelity Human Mesh Reconstruction

We show some preliminary results in Fig. 10 for the network designed in Fig. 6. We observe the detail enhancement on the human hand surface, which has higher fidelity in human perception.

MetricsObject No.	1	2	3	4	5	6	7	8	9	10	11	12	Overall
Chamfer Distance [8]	0.54	0.15	-0.10	0.57	-0.06	-0.12	-0.20	0.07	0.04	0.30	-0.20	0.17	0.097
Point-to-Surface	0.45	0.19	-0.04	0.66	-0.08	-0.25	-0.32	-0.20	0.01	0.13	-0.21	-0.12	0.017
Normal Difference	0.46	0.11	0.06	0.28	0.11	0.21	0.29	0.47	0.27	0.39	0.11	0.27	0.253
IoU [17]	0.60	0.63	0.01	0.51	0.30	0.02	-0.07	0.20	0.14	0.47	-0.09	-0.01	0.225
F-score [66]	0.58	0.09	0.05	0.33	0.03	0.06	0.16	0.34	0.27	0.25	0.01	0.34	0.208
SSFID [78]	0.71	**0.74**	-0.04	**0.74**	**0.39**	0.24	0.13	0.32	0.25	0.64	0.25	-0.02	0.363
UHD [77]	0.29	0.22	0.11	0.15	-0.04	0.18	0.41	0.55	0.13	0.18	0.25	0.33	0.231
(Ours)	0.73	0.21	**0.60**	0.63	0.31	0.51	**0.83**	0.65	**0.77**	**0.80**	**0.69**	0.08	0.567
Adjusted (Ours)	**0.79**	0.19	0.56	0.64	0.36	**0.54**	0.79	**0.76**	0.75	0.77	0.67	**0.36**	**0.598**

a. Pearson's linear correlation coefficient.

MetricsObject No.	1	2	3	4	5	6	7	8	9	10	11	12	Overall
Chamfer Distance [8]	0.33	0.14	-0.09	0.43	-0.08	-0.06	-0.15	0.17	-0.04	0.24	-0.16	0.22	0.079
Point-to-Surface	0.42	0.39	0.14	0.59	0.11	0.05	-0.10	0.20	0.18	0.40	-0.11	0.18	0.205
Normal Difference	0.44	0.22	0.33	0.42	0.19	0.29	0.33	0.56	0.33	0.32	0.21	0.34	0.331
IoU [17]	0.57	0.61	0.28	0.50	0.36	0.21	0.12	0.31	0.262	0.56	0.03	0.30	0.342
F-score [66]	0.47	0.25	0.20	0.52	0.21	0.11	0.07	0.36	0.30	0.42	-0.01	0.35	0.27
SSFID [78]	0.63	**0.81**	0.28	**0.70**	0.33	0.23	0.10	0.33	0.32	0.65	0.16	0.34	0.407
UHD [77]	0.38	0.20	0.11	0.32	0.13	0.35	0.41	0.60	0.06	0.27	0.37	0.35	0.296
(Ours)	0.79	0.25	**0.57**	0.59	0.36	0.56	**0.83**	0.79	0.69	**0.69**	**0.83**	0.24	0.598
Adjusted (Ours)	**0.83**	0.21	0.55	0.59	**0.38**	**0.60**	0.82	**0.80**	**0.69**	0.68	0.75	**0.42**	**0.611**

b. Spearman's rank order correlation coefficient.

MetricsObject No.	1	2	3	4	5	6	7	8	9	10	11	12	Overall
Chamfer Distance [8]	0.25	0.14	-0.08	0.31	-0.04	-0.02	-0.09	0.15	0.013	0.19	-0.07	0.22	0.080
Point-to-Surface	0.33	0.30	0.07	0.45	0.10	0.08	-0.03	0.17	0.13	0.30	-0.01	0.16	0.171
Normal Difference	0.34	0.16	0.17	0.31	0.18	0.22	0.26	0.44	0.25	0.23	0.16	0.27	0.250
IoU [17]	0.42	0.44	0.24	0.37	0.28	0.22	0.14	0.26	0.20	0.41	0.10	0.23	0.275
F-score [66]	0.37	0.17	0.14	0.42	0.15	0.11	0.09	0.28	0.23	0.34	0.01	**0.30**	0.216
SSFID [78]	0.48	**0.62**	0.24	**0.51**	0.25	0.24	0.12	0.29	0.26	**0.48**	0.17	0.23	0.322
UHD [77]	0.27	0.13	0.07	0.22	0.09	0.26	0.29	0.42	0.048	0.19	0.28	0.24	0.209
(Ours)	0.60	0.16	**0.42**	0.41	0.27	0.45	**0.65**	0.57	**0.55**	0.47	**0.60**	0.19	0.445
Adjusted (Ours)	**0.64**	0.14	0.40	0.41	**0.29**	**0.48**	0.63	**0.59**	0.55	0.45	0.57	0.29	**0.453**

c. Kendall's rank order correlation coefficient.

Table 2: Correlations between different metrics and human annotation [36].

6 Limitations

Our methodology's current limitations and potential improvements span several key areas that pave the way for future enhancements. Initially, our reliance on mesh-based 3D shape representations limits the variety and complexity of shapes we can model. Expanding to include point clouds and voxels would capture more intricate geometries and improve the versatility of our models. Additionally, our reality-referenced metric, based on analytic methods, may not effectively capture the subtle human perceptions of fidelity. Implementing few-shot learning techniques could dynamically adapt to individual variations in fidelity

perception, providing a more nuanced assessment tool. Currently, the context within our dataset is predefined, which might not reflect the full diversity of real-world scenarios. A thorough analysis of context at both the dataset and metric levels in future studies would enhance the generalizability and applicability of our method across a broader range of environments. Lastly, our method currently lacks a mechanism for integrating direct human feedback as it relies solely on a pre-annotated dataset for training. Incorporating online reinforcement learning would allow us to utilize real-time feedback to continuously fine-tune both the metric and the reconstruction results, likely improving system responsiveness and increasing user satisfaction by aligning outputs more closely with human evaluative standards. These improvements are crucial for advancing the capabilities and accuracy of 3D reconstruction technologies.

Reference

[1] Gobotree - photos, cut-outs, 3d people. https://www.gobotree.com/.

[2] Sketchfab - the best 3d viewer on the web. https://sketchfab.com/.

[3] Panos Achlioptas, Olga Diamanti, Ioannis Mitliagkas, and Leonidas Guibas. Learning representations and generative models for 3d point clouds. pages 40–49, 2018.

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

[5] Jan Bechtold, Maxim Tatarchenko, Volker Fischer, and Thomas Brox. Fostering generalization in single-view 3d reconstruction by learning a hierarchy of local and global shape priors. In *CVPR*, pages 15880–15889, 2021.

[6] Ernest Bielinis, Jenni Simkin, Pasi Puttonen, and Liisa Tyrväinen. Effect of viewing video representation of the urban environment and forest environment on mood and level of procrastination. *International Journal of Environmental Research and Public Health*, 17(14):5109, 2020.

[7] Yochai Blau, Roey Mechrez, Radu Timofte, Tomer Michaeli, and Lihi Zelnik-Manor. The 2018 pirm challenge on perceptual image super-resolution. In *Proceedings of the European Conference on Computer Vision Workshops*, pages 334–355, Sep 2018.

[8] Gunilla Borgefors. Distance transformations in arbitrary dimensions. *Computer vision, graphics, and image processing*.

[9] Abdullah Bulbul, Tolga Capin, Guillaume Lavoué, and Marius Preda. Assessing visual quality of 3-d polygonal models. *IEEE Signal Processing Magazine*.

[10] Anh-Quan Cao and Raoul de Charette. Scenerf: Self-supervised monocular 3d scene reconstruction with radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9387–9398, 2023.

[11] Zenghao Chai, Tianke Zhang, Tianyu He, Xu Tan, Tadas Baltrusaitis, HsiangTao Wu, Runnan Li, Sheng Zhao, Chun Yuan, and Jiang Bian. Hiface: High-fidelity 3d face reconstruction by learning static and dynamic details. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9087–9098, 2023.

[12] Damon M Chandler. Most apparent distortion: Full-reference image quality assessment and the role of strategy. *Journal of Electronic Imaging*, 19, 2010.

[13] Zhiqin Chen, Vladimir G Kim, Matthew Fisher, Noam Aigerman, Hao Zhang, and Siddhartha Chaudhuri. Decor-gan: 3d shape detailization by conditional refinement. In *CVPR*, pages 15740–15749, 2021.

[14] Massimiliano Corsini, Mohamed-Chaker Larabi, Guillaume Lavoué, Oldřich Petřík, Libor Váša, and Kai Wang. Perceptual metrics for static and dynamic triangle meshes. In *Comput. Graph. Forum*.

[15] Frederik Michel Dekking, Cornelis Kraaikamp, Hendrik Paul Lopuhaä, and Ludolf Erwin Meester. *A Modern Introduction to Probability and Statistics: Understanding why and how*, volume 488. Springer, 2005.

[16] Kyle Genova, Forrester Cole, Avneesh Sud, Aaron Sarna, and Thomas Funkhouser. Local deep implicit functions for 3d shape. In *CVPR*, pages 4857–4866, 2020.

[17] Paul Henderson and Vittorio Ferrari. Learning to generate and reconstruct 3d meshes with only 2d supervision. *arXiv preprint arXiv:1807.09259*, 2018.

[18] Fangzhou Hong, Liang Pan, Zhongang Cai, and Ziwei Liu. Garment4d: Garment reconstruction from point cloud sequences. In *NeurIPS*, 2021.

[19] Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3d-llm: Injecting the 3d world into large language models. *Advances in Neural Information Processing Systems*, 36:20482–20494, 2023.

[20] Vlad Hosu, Hanhe Lin, Tamas Sziranyi, and Dietmar Saupe. Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Transactions on Image Processing*, 29:4041–4056, 2020.

[21] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[22] Tao Hu, Liwei Wang, Xiaogang Xu, Shu Liu, and Jiaya Jia. Self-supervised 3d mesh reconstruction from single images. In *CVPR*, pages 6002–6011, 2021.

[23] P. Jenke, M. Wand, M. Bokeloh, A. Schilling, and W. Strasser. Bayesian point cloud reconstruction. *Computer Graphics Forum*, 2006.

[24] Rasmus Jensen, Anders Dahl, George Vogiatzis, Engin Tola, and Henrik Aanæs. Large scale multi-view stereopsis evaluation. In *CVPR*, pages 406–413, 2014.

[25] Hyun In Jo, Kounseok Lee, and Jin Yong Jeon. Effect of noise sensitivity on psychophysiological response through monoscopic 360 video and stereoscopic sound environment experience: a randomized control trial. *Scientific reports*, 12(1):4535, 2022.

[26] Yueying Kao, Bowen Pan, Miao Xu, Jiangjing Lyu, Xiangyu Zhu, Yuanzhang Chang, Xiaobo Li, and Zhen Lei. Towards 3d face reconstruction in perspective projection: Estimating 6dof face pose from monocular image. IEEE, 2023.

[27] Hyomin Kim, Hyeonseo Nam, Jungeon Kim, Jaesik Park, and Seungyong Lee. Laplacianfusion: Detailed 3d clothed-human body reconstruction. *ACM Transactions on Graphics (TOG)*, 41(6):1–14, 2022.

[28] Audrius Kulikajevas, Rytis Maskeliunas, Robertas Damasevicius, and Tomas Krilavicius. Auto-refining 3d mesh reconstruction algorithm from limited angle depth data. *IEEE Access*.

[29] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*.

[30] Valero Laparra, Johannes Ballé, Alexander Berardino, and Eero P Simoncelli. Perceptual image quality assessment using a normalized laplacian pyramid. *Electronic Imaging*, 28:1–6, 2016.

[31] Guillaume Lavoué. A local roughness measure for 3d meshes and its application to visual masking. *ACM Transactions on Applied perception*.

[32] Karl Lenz. Behavior in public places. notes on the social organization of gatherings. In *Goffman-Handbuch: Leben–Werk–Wirkung*, pages 291–297. Springer, 2022.

[33] Hongyi Li, Yujun Ding, Bing Zhao, Yuhang Xu, and Wei Wei. Effects of immersion in a simulated natural environment on stress reduction and emotional arousal: A systematic review and meta-analysis. *Frontiers in Psychology*, 13:1058177, 2023.

[34] Peizhen Lin, Hongliang Zhong, Lei Wang, and Jun Cheng. 3d mesh reconstruction of indoor scenes from a single image in-the-wild. In *International Conference on Graphics and Image Processing*.

[35] Ange Lou, Benjamin Planche, Zhongpai Gao, Yamin Li, Tianyu Luan, Hao Ding, Terrence Chen, Jack Noble, and Ziyan Wu. Darenerf: Direction-aware representation for dynamic scenes. *arXiv e-prints*, pages arXiv–2403, 2024.

[36] Tianyu Luan, Zhong Li, Lele Chen, Xuan Gong, Lichang Chen, Yi Xu, and Junsong Yuan. Spectrum auc difference (saucd): Human-aligned 3d shape evaluation. *arXiv preprint arXiv:2403.01619*, 2024.

[37] Tianyu Luan, Yali Wang, Junhao Zhang, Zhe Wang, Zhipeng Zhou, and Yu Qiao. Pc-hmr: Pose calibration for 3d human mesh recovery from 2d images/videos. In *AAAI*, pages 2269–2276, 2021.

[38] Tianyu Luan, Yuanhao Zhai, Jingjing Meng, Zhong Li, Zhang Chen, Yi Xu, and Junsong Yuan. High fidelity 3d hand shape reconstruction via scalable graph frequency decomposition. In *CVPR*, pages 16795–16804, 2023.

[39] Chao Ma, Chia-Yen Yang, Xiaokang Yang, and Ming-Hsuan Yang. Learning a no-reference quality metric for single-image super-resolution. *Computer Vision and Image Understanding*, 158:1–16, May 2017.

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

[41] Steven C Mills and Tillman J Ragan. A tool for analyzing implementation fidelity of an integrated learning system. *Educational Technology research and development*, 48(4):21–41, 2000.

[42] Anish Mittal, Anish K Moorthy, and Alan C Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12):4695–4708, Dec 2012.

[43] Gyeongsik Moon, Takaaki Shiratori, and Kyoung Mu Lee. Deephandmesh: A weakly-supervised deep encoder-decoder framework for high-fidelity hand mesh modeling. In *ECCV*, 2020.

[44] Anish K Moorthy and Alan C Bovik. Blind image quality assessment: From natural scene statistics to perceptual quality. *IEEE Transactions on Image Processing*, 20(12):3350–3364, Dec 2011.

[45] Naila Murray, Luca Marchesotti, and Florent Perronnin. Ava: A large-scale database for aesthetic visual analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2408–2415, 2012.

[46] Yinyu Nie, Xiaoguang Han, Shihui Guo, Yujian Zheng, Jian Chang, and Jian Jun Zhang. Total3dunderstanding: Joint layout, object pose and mesh reconstruction for indoor scenes from a single image. In *CVPR*, pages 55–64, 2020.

[47] Kristine Nowak. Defining and differentiating copresence, social presence and presence as transportation. In *presence 2001 conference, Philadelphia, PA*, volume 2, pages 686–710. Citeseer, 2001.

[48] Songyou Peng, Chiyu Jiang, Yiyi Liao, Michael Niemeyer, Marc Pollefeys, and Andreas Geiger. Shape as points: A differentiable poisson solver. In *NeurIPS*, 2021.

[49] Nikolay Ponomarenko, Vladimir Lukin, Alexander Zelensky, Karen Egiazarian, Marco Carli, and Federica Battisti. Tid2008-a database for evaluation of full-reference visual quality assessment metrics. *Advances of Modern Radioelectronics*, 10(4):30–45, 2009.

[50] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022.

[51] Marie-Julie Rakotosaona, Paul Guerrero, Noam Aigerman, Niloy J Mitra, and Maks Ovsjanikov. Learning delaunay surface elements for mesh reconstruction. In *CVPR*, pages 22–31, 2021.

[52] Giuseppe Riva, Fabrizia Mantovani, Claret Samantha Capideville, Alessandra Preziosa, Francesca Morganti, Daniela Villani, Andrea Gaggioli, Cristina Botella, and Mariano Alcañiz. Affective interactions using virtual reality: the link between presence and emotions. *Cyberpsychology & behavior*, 10(1):45–56, 2007.

[53] Hari Santhanam, Nehal Doiphode, and Jianbo Shi. Automated line labelling: Dataset for contour detection and 3d reconstruction. pages 3136–3145, 2023.

[54] Jean-Christophe Servotte, Manon Goosse, Suzanne Hetzell Campbell, Nadia Dardenne, Bruno Pilote, Ivan L Simoneau, Michèle Guillaume, Isabelle Bragard, and Alexandre Ghuysen. Virtual reality experience: Immersion, sense of presence, and cybersickness. *Clinical Simulation in Nursing*, 38:35–43, 2020.

[55] Rakesh Shrestha, Zhiwen Fan, Qingkun Su, Zuozhuo Dai, Siyu Zhu, and Ping Tan. Meshmvs: Multi-view stereo guided mesh reconstruction. In *3DV*, pages 1290–1300. IEEE, 2021.

[56] Dong Wook Shu, Sung Woo Park, and Junseok Kwon. 3d point cloud generative adversarial network based on tree structured graph convolutions. In *ICCV*, pages 3859–3868, 2019.

[57] Brid Sona, Erik Dietl, and Anna Steidle. Recovery in sensory-enriched break environments: integrating vision, sound and scent into simulated indoor and outdoor environments. *Ergonomics*, 62(4):521–536, 2019.

[58] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5459–5469, 2022.

[59] Jiaxiang Tang, Xiaokang Chen, Jingbo Wang, and Gang Zeng. Point scene understanding via disentangled instance mesh reconstruction. pages 684–701, 2022.

[60] Jiaxiang Tang, Xiaokang Chen, Jingbo Wang, and Gang Zeng. Point scene understanding via disentangled instance mesh reconstruction. In *ECCV*, pages 684–701, 2022.

[61] Xiao Tang, Tianyu Wang, and Chi-Wing Fu. Towards accurate alignment in real-time 3d hand-mesh reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11698–11707, 2021.

[62] Maxim Tatarchenko, Stephan R Richter, René Ranftl, Zhuwen Li, Vladlen Koltun, and Thomas Brox. What do single-view 3d reconstruction networks learn? In *CVPR*, pages 3405–3414, 2019.

[63] Hitika Tiwari, Vinod K. Kurmi, K.S. Venkatesh, and Yong-Sheng Chen. Occlusion resistant network for 3d face reconstruction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 813–822, 2022.

[64] Kai Wang, Guillaume Lavoué, Florence Denis, Atilla Baskurt, and Xiyan He. A benchmark for 3d mesh watermarking. In *Shape Modeling International Conference*, pages 231–235. IEEE, 2010.

[65] Lizhen Wang, Zhiyuan Chen, Tao Yu, Chenguang Ma, Liang Li, and Yebin Liu. Faceverse: a fine-grained and detail-controllable 3d face morphable model from a hybrid dataset. In *CVPR*, pages 20333–20342, 2022.

[66] Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2mesh: Generating 3d mesh models from single rgb images. In *ECCV*, pages 52–67, 2018.

[67] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021.

[68] Shuaibing Wang, Shunli Wang, Dingkang Yang, Mingcheng Li, Ziyun Qian, Liuzhen Su, and Lihua Zhang. Handgcat: Occlusion-robust 3d hand mesh reconstruction from monocular images. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pages 2495–2500. IEEE, 2023.

[69] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13:600–612, 2004.

[70] Xingkui Wei, Zhengqing Chen, Yanwei Fu, Zhaopeng Cui, and Yinda Zhang. Deep hybrid self-prior for full 3d mesh generation. In *ICCV*, pages 5805–5814, 2021.

[71] Yi Wei, Linqing Zhao, Wenzhao Zheng, Zheng Zhu, Jiwen Lu, and Jie Zhou. Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving. In *International Conference on Computer Vision (ICCV)*, 2023.

[72] Robert B Welch, Theodore T Blackmon, Andrew Liu, Barbara A Mellers, and Lawrence W Stark. The effects of pictorial realism, delay of visual feedback, and observer interactivity on the subjective sense of presence. *Presence: Teleoperators & Virtual Environments*, 5(3):263–273, 1996.

[73] Mingyun Wen and Kyungeun Cho. Object-aware 3d scene reconstruction from single 2d images of indoor scenes. *Mathematics*, 11(2):403, 2023.

[74] Bob G Witmer and Michael J Singer. Measuring presence in virtual environments: A presence questionnaire. *Presence*, 7(3):225–240, 1998.

[75] Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Matthew Johnson, Jingjing Shen, Nikola Milosavljević, Daniel Wilde, Stephan Garbin, Toby Sharp, Ivan Stojiljković, et al. 3d face reconstruction with dense landmarks. In *European Conference on Computer Vision*, pages 160–177. Springer, 2022.

[76] Chenyan Wu, Yandong Li, Xianfeng Tang, and James Wang. Mug: Multi-human graph network for 3d mesh reconstruction from 2d pose. *arXiv preprint arXiv:2205.12583*, 2022.

[77] Rundi Wu, Xuelin Chen, Yixin Zhuang, and Baoquan Chen. Multimodal shape completion via conditional generative adversarial networks. In *ECCV*, pages 281–296, 2020.

[78] Rundi Wu and Changxi Zheng. Learning to generate 3d shapes from a single example. *arXiv preprint arXiv:2208.02946*, 2022.

[79] Xianzu Wu, Xianfeng Wu, Tianyu Luan, Yajing Bai, Zhongyuan Lai, and Junsong Yuan. Fsc: Few-point shape completion. *arXiv preprint arXiv:2403.07359*, 2024.

[80] Weisi Xue, Lei Zhang, Xuanqin Mou, and Alan C Bovik. Gradient magnitude similarity deviation: A highly efficient perceptual image quality index. *IEEE Transactions on Image Processing*, 23:684–695, 2014.

[81] Wenfei Yao, Xiaofeng Zhang, and Qi Gong. The effect of exposure to the natural environment on stress reduction: A meta-analysis. *Urban Forestry & Urban Greening*, 57:126932, 2021.

[82] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. *Advances in Neural Information Processing Systems*, 34:4805–4815, 2021.

[83] Nicola L Yeo, Mathew P White, Ian Alcock, Ruth Garside, Sarah G Dean, Alexander J Smalley, and Birgitta Gatersleben. What is the best way of delivering virtual nature for improving mood? an experimental comparison of high definition tv, 360 video, and computer generated virtual reality. *Journal of environmental psychology*, 72:101500, 2020.

[84] Alex Yu, Sara Fridovich-Keil, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. *arXiv preprint arXiv:2112.05131*, 2021.

[85] Rongfei Zeng, Mai Su, and Xingwei Wang. Cd˜2: Fine-grained 3d mesh reconstruction with twice chamfer distance. *arXiv preprint arXiv:2206.00447*, 2022.

[86] Yuanhao Zhai, Mingzhen Huang, Tianyu Luan, Lu Dong, Ifeoma Nwogu, Siwei Lyu, David Doermann, and Junsong Yuan. Language-guided human motion synthesis with atomic actions. In *ACM MM*, pages 5262–5271, 2023.

[87] Yuanhao Zhai, Tianyu Luan, David Doermann, and Junsong Yuan. Towards generic image manipulation detection with weakly-supervised self-consistency learning. In *ICCV*, pages 22390–22400, 2023.

[88] Hongwen Zhang, Yating Tian, Yuxiang Zhang, Mengcheng Li, Liang An, Zhenan Sun, and Yebin Liu. Pymaf-x: Towards well-aligned full-body model regression from monocular images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[89] Junhao Zhang, Yali Wang, Zhipeng Zhou, Tianyu Luan, Zhe Wang, and Yu Qiao. Learning dynamical human-joint affinity for 3d pose estimation in videos. *IEEE Transactions on Image Processing*, 30:7914–7925, 2021.

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

[91] Lin Zhang, Yuming Shen, and Hongyu Li. Vsi: A visual saliency-induced index for perceptual image quality assessment. *IEEE Transactions on Image Processing*, 23:4270–4281, 2014.

[92] Lin Zhang, Lei Zhang, and Alan C Bovik. A feature-enriched completely blind image quality evaluator. *IEEE Transactions on Image Processing*, 24(8):2579–2591, Aug 2015.

[93] Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. Fsim: A feature similarity index for image quality assessment. *IEEE Transactions on Image Processing*, 20:2378–2386, 2011.

[94] Zhihao Zhang, Xinyang Ren, and Xianqiang Yang. Parametric chamfer alignment based on mesh deformation. *Measurement and Control*.

[95] Chuhang Zou and Derek Hoiem. Silhouette guided point cloud reconstruction beyond occlusion. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 41–50, 2020.

[96] Xinxin Zuo, Sen Wang, Minglun Gong, and Li Cheng. Unsupervised 3d human mesh recovery from noisy point clouds. *arXiv preprint arXiv:2107.07539*, 2021.