Understanding Perceptual Mesh Quality in Virtual Reality and Desktop Settings

Dalia Ahmed M. ALfarasani

School of Computer Science and Informatics

Cardiff University

A thesis submitted in partial fulfilment of the requirement for the degree of

Doctor of Philosophy

February 2024

Abstract

This thesis focuses on 3D mesh quality, essential for immersive VR applications. It examines subjective methodologies for Quality of Experience (QoE) assessments and then develops objective quality metrics incorporating QoE influencing factors. Existing studies consider 3D mesh quality on the desktop. The perceptual quality in a Virtual Reality (VR) setting can be different, this inspired us to measure mesh quality in a VR setting, which has been the subject of limited studies in this area. We consider how different 3D distortion types affect perceptual quality of 3D when viewed in a VR setup. In our experiment findings, in the VR setting, perception appears more sensitive to particular distortions than others, compared with the desktop. This can provide helpful guidance for downstream applications. Furthermore, we evaluate state-of-the-art perceptually inspired mesh difference metrics for predicting objective quality scores captured in VR and compare them with the desktop. The experimental results show that subjective scores in the VR setting are more consistent than those on desktop setting.

As we focus on a better understanding of perceptual mesh quality, we further consider the problem of mesh saliency, which measures the perceptual importance of different regions on a mesh. However, existing mesh saliency models are largely built with hard-coded formulae or utilise indirect measures, which cannot capture true human perception. In this thesis, to generate ground truth mesh saliency, we use subjective studies that collect eye-tracking data from participants and develop a method for mapping the eye-tracking data of individual views consistently onto a mesh. We further evaluate existing methods of measuring saliency and propose a new machine learning-based method that better predicts

subjective saliency values. The predicted saliency is also demonstrated to help with mesh quality prediction as salient regions tend to be more important perceptually, leading to a novel effective mesh quality measure.

Contents

Abstract **i**

Contents **iii**

List of Tables **ix**

List of Figures **xi**

List of Abbreviations **xv**

Acknowledgements **xvii**

1 Introduction **1**

1.1 Aims 4

1.2 Background and Motivation 5

1.3 Research Questions and Contributions 10

1.4 Thesis Structure 11

1.5 Publications Related to This Thesis 12

1.6 Summary 12

2 Background and Related Work **14**

2.1 Introduction 14

2.2 2D Image Quality Assessment 18

2.2.1 Full-Reference Image Quality Assessment Methods 19

2.2.2 Reduced-Reference Image Quality Assessment Methods 24

2.2.3 No-Reference (Blind-Reference) Image Quality Assessment Methods 25

2.3 3D Mesh Visual Quality Assessment 26

2.4 Subjective 3D Mesh Quality Databases 31

2.5 Subjective Methodologies 32

2.5.1 Single Stimulus Methodologies 32

2.5.2 Double Stimulus Methodologies 34

2.6 Limitations of Perceptual Metrics 36

2.6.1 Perceptual Metrics Based on Images 36

2.6.2 Perceptual Metrics Based on 3D Geometry 37

2.7 Eye Movement 39

2.7.1 Eye Movement Metrics 41

2.8 3D Mesh Saliency 44

2.9 Virtual Reality 45

2.9.1 Virtual Reality Applications 48

2.10 Summary 51

3 Subjective Study of 3D Mesh Quality Scores in Virtual Reality 52

3.1 Introduction 52

3.2 Subjective Experiment 55

3.2.1 Evaluation Methodology 57

3.2.2 Stimuli Generation 58

3.2.3 Display 58

3.2.4 Participants and Training 59

3.2.5 Procedure 61

3.2.6 Duration 61

3.2.7 Experiment Design 62

3.2.8 Ethical Approval 63

3.3 Data Analysis 64

3.3.1 Screening Participants and Computing Mean Opinion Scores (MOS) 64

3.3.2 Observer Agreement Analysis and Correlation Analysis 66

3.4 Results 67

3.4.1 Distortion by Shape 67

3.4.2 Distortion by Type and Location 72

3.5 Summary 76

4 Learning to Predict 3D Mesh Saliency 77

4.1 Introduction 78

4.2 Related Work 79

4.3 3D Mesh Saliency Applications 82

4.3.1 3D View Selection 83

4.3.2 3D Mesh Simplification 85

4.4 Voronoi Tessellation and Delaunay Triangulation 86

4.5 Integration of Structural Similarity Index Model (SSIM) in a Subjective Experiment Utilising Eye-tracking 87

4.5.1 Design of User Experiments 94

4.6 Experimental Setup 97

4.6.1 Experimental Procedure 98

4.6.2 Ethical Approval 100

4.7 Obtaining Ground Truth and Evaluation of Existing Methods 101

4.8 Learning New Methods of Measuring 3D Mesh Saliency 104

4.8.1 Least Squares Regression 106

4.8.2 Support Vector Regression 106

4.8.3 Machine Learning based on Neural Networks 107

4.9 Results and Discussions 109

4.9.1 Eye Tracking and Mesh Saliency Ground Truth Results 109

4.9.2 Evaluation Results of Existing and Our Learning Methods 111

4.10 Summary 114

5 Objective Quality Assessment Measures for Mesh Quality 115

5.1 Introduction 116

5.2 Mesh Difference Metrics 117

5.3 Comparison of Objective Quality Measures for VR and Desktop 123

5.4 Data Analysis 124

5.5 Comparison of MOS Scores for VR and Desktop 124

5.6 Saliency-Weighted Objective Quality Metrics 127

5.7 Summary 129

6 Conclusion & Future Work 130

6.1 Summary 130

6.2 Novel Contribution 131

6.2.1 Subjective Study of 3D Mesh Quality Scores in Virtual Reality . . 131

6.2.2 Learning to Predict 3D Mesh Saliency 133

6.2.3 Objective Quality Assessment Measures for Mesh Quality 133

6.3 Future Work 134

Bibliography 138

Appendices **168**

Appendix A Subjective Study of 3D Mesh Quality Scores in Virtual Reality **169**

A.1 Distortion by Shape 169

A.2 Distortion by Type, Location and Levels 169

Appendix B Learning to Predict 3D Mesh Saliency **181**

B.1 Dataset 181

B.2 Related Saliency Models and Geometric Features 182

B.2.1 Geometric Characteristics for Mesh Saliency 183

B.2.2 Geometric Feature Extraction 183

B.3 Implementation Details for 3D Mesh Saliency 186

List of Tables

3.1 The geometric information of the reference meshes in the LIRIS/EPFL general-purpose database 57

3.2 Pearson and Spearman correlation analysis comparing VR and desktop MOS scores for different stimuli (the distortion type followed by distortion location. 69

3.3 Pearson and Spearman coefficient correlations between MOS scores from VR and desktop settings, grouped based on distortion types and locations. 73

4.1 Average SSIM value and Mean Square Error (MSE) for each existing method and our learning-based method for evaluating the quality of predicted saliency maps against the ground truth derived from eye tracking. The only test set is used to ensure a fair comparison. For SSIM, larger is better, and for MSE, smaller is better. 112

5.1 Pearson (PLCC) & Spearman (SROCC) correlations value (%) between Mean Opinion Scores in VR and desktop settings, and values from the objective mesh quality metrics for the General-purpose dataset. 123

5.2 Pearson (PLCC) & Spearman (SROCC) correlations value (%) between Mean Opinion Scores in VR setting and MSDM2 along with different saliency weighting: Lee [115], Song [193], and our least squares regression model, denoted as LSR (see Chapter 4). We also compare with saliency-weighted MSDM2 based on ground truth saliency for the Armadillo model (as it is the only shared model used in our saliency subjective study (see Chapter 4), denoted as MSDM2-GT. 128

A.1 Comparison of MOS scores for both VR and desktop settings organised based on shape (Armadillo, Venus, Dyno and Rocker-Arm) and distortion type/location. 170

A.2 Comparison of MOS scores for both VR and desktop settings organised based on distortion Type, Location and Levels (Noise Uniform, Noise Rough, Noise Intermediate, Noise Smooth, Taubin Uniform, Taubin Rough and Taubin Intermediate) 174

List of Figures

1.1 An example of a 3D reference mesh of an armadillo, along with different types of distortion that affect visual quality. 3

1.2 Visual media life cycle. In each step, quality degradation may occur, and a quality estimator (Q in the figure) is used to decide on appropriate enhancement. 7

1.3 Scope and contribution of this thesis. Each box represents a different element: (Q) presents the research questions, (CQ) presents the contributions and green boxes correspond to the chapters in the thesis. 9

2.1 A diagram summarising background and related work in quality assessment. 17

2.2 Example of the single stimuli method [153] 33

2.3 Example of the double stimulus method [153] 35

2.4 Distorted versions of the Horse model, all associated with the equal maximum Root Mean Square Error. (a) Original model. Results after (b) watermarking (MSDM2=0.14), (c) Laplacian smoothing (MSDM2=0.40), (d) watermarking (MSDM2=0.51), (e) simplification (MSDM2=0.62) and (f) Gaussian noise addition (MSDM2=0.84) [108]. 38

2.5 Corneal reflection and bright pupil as seen in the infrared camera image. . 42

3.1 Examples of 3D meshes belonging to the LIRIS/EPFL General-Purpose database. The top row shows the 4 reference meshes. The second row presents 4 distorted 3D meshes: e) 3D mesh Armadillo affected with noise on rough regions, (f) 3D mesh Venus affected uniformly with noise, (g) 3D mesh Dinosaur uniformly smoothed, (h) 3D mesh Rocker-Arm affected with noise on smooth regions. 53

3.2 Examples of 3D meshes belonging to the LIRIS/EPFL General-Purpose database. The top row shows the four reference meshes (from left to right: Armadillo, Venus, Dyno and Rocker-Arm) and the rest shows that we have 21 models with different types/levels of distortion for each shape (only selected ones are shown), so the total number of shapes is 88. 56

3.3 Example of the Home page showing our experimental environment of the subjective test. 59

3.4 Example of the main page where the user can interact with the models in the experiment environment. The left side contains the reference model, and the distorted model is on the right side (B). 60

3.5 Illustration for Interquartile Range (IQR) 64

3.6 The reference mesh, three distorted Armadillo model meshes, and the enlarged views of some representative distorted regions on the meshes as marked in the rectangles. 68

3.7 The reference mesh, three distorted Dyno model meshes, and the enlarged views of some representative distorted regions on the meshes as marked in the rectangles. 69

3.8 Comparison of MOS for both VR and desktop settings in the Pairwise Comparison (PC) experiment for all the stimuli shapes. 71

3.9 An example of the distortion types. (a) Original Venus model and illustration of the different types of regions; (b) high-level noise applied on rough regions; (c) medium-level noise applied on smooth regions. 74

3.10 Comparison of MOS scores between the VR and desktop settings. Each figure shows a type of distortion (Noise, Smoothing and Taubin) applied to certain locations (Uniform, Rough, Intermediate and Smooth). The x-axis corresponds to VR MOS scores or Desktop MOS scores, and the y-axis shows the (normalised) MOS scores averaged over all subjects for the distorted shapes. The blue and green plots correspond to the VR and desktop settings, respectively. 75

4.1 This illustration shows the Voronoi and Delaunay area in a 3D mesh around a given vertex. 86

4.2 This illustration shows a diagram of the structural similarity (SSIM) measurement system [223] 88

4.3 Example of 3D shape is rendered from 20 different views. 92

4.4 Example of an object from two views and its corresponding vertex maps. 94

4.5 Example views of one mesh (Armadillo) white background. 8 out of 20 views are shown here. 95

4.6 Example views of one mesh (Armadillo) with the black background. 20 views are shown here. 96

4.7 Examples of 4 views in each shape and the eye fixation of a participant. . 98

4.8 Example of how we run the experiment using eye tracker after each shape, we use the grey background to give a participant break. 99

4.9 Example of a stigmatised failure to capture eye tracking. 101

4.10 Examples of ground truth salient maps derived from eye tracking data; red and yellow are salient areas, while green and blue are non-salient areas. Source of the data present in the Appendix B.1 108

4.11 Examples of comparison of saliency maps before and after normalising views. From left to right Armadillo (before), Armadillo (after), Kitten (before) and Kitten (after) 109

4.12 Examples of 2D fixation maps and the results of fusing them to form consistent saliency maps on 3D models; red and yellow are salient areas, while green and blue are non-salient areas. 110

4.13 Examples of ground truth salient maps derived from eye tracking data; red and yellow are salient areas, while green and blue are non-salient areas. From left to right Falling, Bulldog and Gargoyle shapes. 110

4.14 Example of Ant's saliency results: (a) ground truth, (b,c,d) our learning-based methods, (e,f) existing methods and (g-j) geometry feature-based baseline methods. 113

A1 Comparison of MOS scores averaged over all the shapes for both VR and desktop settings for each type of location with changing level of distortion strength. 180

B1 Illustrate data flow for eye tracking stimuli and remapping scripts. 186

B2 Illustrate data flow for existing method and evaluating scripts.. 187

B3 Illustrate data flow for existing method and evaluating scripts.. 188

B4 Illustrate data flow for existing method and evaluating scripts. 189

List of Abbreviations

ACR	Absolute Category Rating
ACR-HR	Absolute Category Rating with Hidden Reference
AR	Augmented Reality
CNN	Convolutional Neural Network
DAME	Dihedral Angle Mesh Error
DCR	Degradation Category Rating
DMOS	Differential Mean Opinion Score
DNN	Deep Neural Network
DS	Double Stimulus
DSCS	Double Stimulus Comparison Scale
DSCQS	Double-Stimulus Continuous Quality-Scale
FMPD	Fast Mesh Perceptual Distance
FNN	Feed-forward Neural Network
FR	Full-Reference
GL	General Laplacian
HCI	Human-Computer Interaction
HD	Hausdorff Distances
HKS	Heat Kernel Signature
HMD	Head-Mounted Display
HVS	Human Visual System
IVQ	Image Visual Quality

LSR	Least Squares Regression
MOS	Mean Opinion Score
MSDM	Mesh Structural Distortion Measure
MSDM2	Multi-scale Mesh Structural Distortion Measure
MVQ	Mesh Visual Quality
NR	No-Reference
PC	Pairwise Comparison
PSNR	Peak Signal Noise Ratio
QoE	Quality of Experience
RMS	Root Mean Square
RR	Reduced-Reference
SAMVIQ	Subjective Assessment Methodology for Video Quality
SHOT	Signature and Histogram of Orientation
SIFT	Scale-Invariant Feature Transform
SS	Single Stimulus
SSCQE	Single Stimulus Continuous Quality Evaluation
SSIM	Structural Similarity Index Model
SVR	Support Vector Regression
TPDM	Tensor-based Perceptual Distance Measure
VQA	Visual Quality Assessment
VR	Virtual Reality
XR	Extended Reality
3DWP	Direct Reconstruction of 3D Worker Pose
6DoF	Six Degrees of Freedom

Acknowledgements

I sincerely thank my supervisors, Prof. Yukun Lai and Prof. Paul Rosin, for their continuous support and guidance. They accorded me throughout my PhD study immense knowledge, patience, and motivation. They were not only encouraging but also made the enormous task of writing a dissertation lighter. I could not have asked for better supervisors, mentors, and advisors for my PhD study. Also, I would like to thank the School Research Ethics Committee for giving the permission to do my two experiments involving human participants at Cardiff University.

Finally, I would like to express my gratitude to my family and friends for their support and motivation throughout my PhD study and during the difficult moments when I had to work extra hard to complete this research project.

Chapter 1

Introduction

Multimedia technologies have become an important area that benefits significantly from recent advances of computer vision. In recent times, most people have spent a large part of their life interacting with multimedia technology, such as surfing the internet or using streaming services. We often depend on multimedia systems when working on various tasks, such as using security cameras to regulate our house or smartwatches to track our sleeping. As more companies compete to deliver multimedia services to users, to meet the expectations of consumers, multimedia systems must be created with the goal of maximising user satisfaction with services. This measure of quality is known as Quality of Experience (QoE). Quality of Experience is *"the degree of delight or annoyance of the user of an application or service"* [114].

Multimedia systems and services are designed around the consumption of visual media (e.g., images, 3D meshes and videos). Some examples of such multimedia systems include virtual and augmented reality, video surveillance, and mobile services. Visual media consumption is important for various applications, systems, and services. According to the Cisco Visual Networking Index [90], the visual quality of more than 70% of internet traffic worldwide is expected to increase. Also, according to statistics [141, 59], we download around 3.9 trillion images daily [141] and upload at least 1.8 billion images daily to various online platforms [59].

3D technology, such as 3D TVs and 3D gaming devices, provides a new opportunity for improving user experience in 3D surroundings [2]. Owing to improved availability

and quality, 3D models are becoming more popular as a new primary technology in the digital age [47]. 3D mesh data are often used in digital entertainment, scientific visualisation, and the preservation of cultural artefacts. The increasing visualisation capabilities of viewing devices and widely available online content lead to increased computation and storage demands. Network-based applications are particularly affected because they often need to apply a certain level of compression for 3D models to increase the transmission speed [78]. 3D meshes and other 3D representations are required for a wide range of applications, such as medical applications [176] and software for surgery [61]. It is common for 3D mesh models to have a high number of vertices and faces to be displayed or streamed in real time [177]. A model with more vertices and faces can often appear more realistic or of higher quality because of the improved level of detail.

There is a well-documented trade-off between visual quality and processing time. For some tasks such as simplification, in order to reduce the processing time. It needs to reduce the complexity of the meshes [136]. The quality of the visual representation of the data is impacted, which may result in distortions. As a result, the quality of the 3D mesh must be assessed. Many computer vision applications take into account the importance of visual quality. Indeed, the quality of the data may significantly impact the performance of a computer vision program. As a result, we need data grading on how users evaluate quality. Assessments of 3D meshes determine how much the original model has been deformed. As 3D forms become more prevalent in many application domains, this issue becomes more critical.

Metrics for assessing the 3D mesh visual quality (MVQ) have been carefully used to assess the perceptual effects of distortion to forecast distorted 3D data visual quality compared to the original data [60]. This has led to the development of several criteria for estimating the negative effect of visual artefacts. For example, some metrics are based on Laplacian coordinates, forms of curvature calculation, geometric features, and standard geometric distances [78].

Subjective evaluation and objective measurements may be used to evaluate perceptual

quality. Objective metrics consist of automated prediction techniques for visual quality degradation (i.e., the annoyance level of visual artefacts). Alternatively, during user studies for subjective evaluation, a group of volunteers evaluate the visual quality of test data during subjective investigations. As humans are the eventual judges for visual quality, these subjective studies are probably the most reliable method for generating ground-truth datasets that may be used to analyse human psychological preferences and behaviour (when grading multimedia information) and to assess and modify objective quality criteria.

This thesis aims to focus on 3D mesh visual quality assessment (see Figure 1.1 for an example with different types/levels of distortion). It also examines how the way the user interacts with technology tools such as Virtual Reality (VR) and regular desktop settings affect perceptual 3D mesh quality when the shapes are distorted with different distortion types and levels. Although the focus is on 3D shapes, some works used in the image quality assessment area that can help build metrics in the field of 3D meshes will also be discussed.

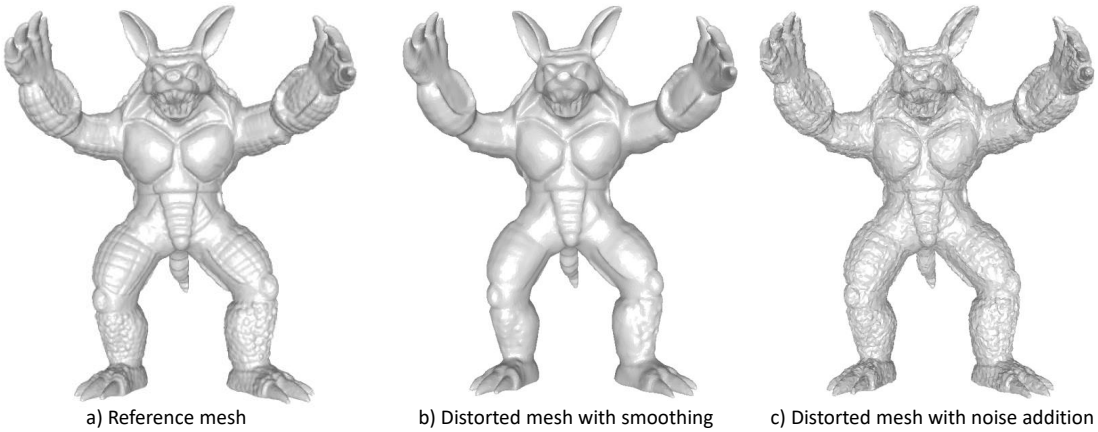


Figure 1.1: An example of a 3D reference mesh of an armadillo, along with different types of distortion that affect visual quality.

The remainder of this chapter discusses the aims and motivations behind this research

in Sections 1.1 and 1.2, and summarises the novel contributions in Section 1.3. In addition, Section 1.4 provides an outline of what is covered in each chapter. We conclude by presenting a list of publications resulting from this thesis research in Section 1.5.

1.1 Aims

This thesis focuses on visual quality assessment of 3D meshes as a commonly used representation for 3D shapes. To address the limited studies on 3D mesh quality assessment in VR and how that compares with the normal display, the thesis in particular investigates how different types and levels of distortions affect the perceptual quality of 3D meshes comparing VR and desktop settings. To achieve this, we collected subjective user scores on distorted shapes in the VR setting compared to desktop studies to measure the quality between these settings. This provides insights regarding how human perceptual quality differs with different settings. We also build a learning-based model to predict scores using the data. We then compare and evaluate different objective quality assessment methods, which are methods focused on measuring the dissimilarity between two meshes (reference and distorted mesh) in normal display desktop and VR settings. We further hypothesise that salient regions have more impact on visual perceptual quality. To study this, we exploit 3D mesh saliency, using objective and objective measures, incorporating user fixations. This involves collecting eye-tracking data and fusing information from different views to form consistent 3D saliency maps on meshes. The thesis then explores ways to predict 3D mesh saliency. The existing studies used hard-coded formulas which are unreliable with human perceptions. In our experiment, there is an improvement in existing models by incorporating machine learning to help combine geometric features to better predict mesh saliency.

1.2 Background and Motivation

In recent years, the use of 3D meshes has increased in the general public and industry, owing to the emergence and improvements in 3D imaging (3D scanners, 360-degree cameras, MRI, etc.), 3D modelling tools became available, including affordable virtual reality and mixed reality (MR) head-mounted displays (HMD) (e.g., Oculus Rift, Meta/Oculus Quest and HTC Vive). These technologies are expanding the size and complexity of 3D data. Indeed, a 3D scene may contain thousands or even millions of geometric primitives and an array of appearance attributes to reproduce realistic material attributes.

Improvements and developments in the field of extended reality (XR), which includes (VR) and augmented reality (AR), are generally considered future goals of computer science. VR systems with six degrees of freedom (6DoF) provide a variety of possibilities for immersive, realistic interaction. One significant obstacle that prevents the quality of display and interaction in XR in large and complex 3D scenes is latency difficulties that arise while streaming the 3D scene to the client device for networked applications.

The number of VR and AR applications requiring 3D data saved from online servers exacerbates these issues. 3D content needs to be simplified and compressed to be compatible with HMDs, as well as mitigating latency issues caused by transmission. These losses result in visual degradation that may reduce the perceived quality of a 3D scene and, thus, negatively affect user QoE. Thus, to find the right compromise between visual quality and data size/LoD (Level of Detail), it is essential to define measures to assess the impact of these distortions accurately. It is necessary to use quality assessment methodologies for this purpose. Metrics and datasets available for public quality evaluation of 3D graphics are few, and mainly for models containing colour or texture attributes. Many previous studies focus on image or video quality assessment. However, this is beginning to change with the increasing popularity of new VR technologies, which provide an immersive experience for users.

Visual quality assessment (VQA) is a growing requirement for digital images, 3D models, and video technologies in entertainment, communications, security, monitoring, and medical imaging. The demand for 3D visual quality has pushed visual media quality assessment to the forefront. Acquisition, processing, compression, transmission, display, printing, and reproduction systems are some aspects that may impact or impair the quality of visual media. Also, other factors can also introduce visible artefacts, such as noise, simplification and watermarking. The objective of the visual quality evaluation is to measure the quality of visual media such as still images, image sequences (video) and 3D models using quality metrics. These measurement tools need to be appropriate to the applications being evaluated.

Indeed, there are generally two typical types of methods used in VQA: subjective and objective metrics [110]. Subjective VQAs have long been used to evaluate visual quality. Whether or not a visual reference is present, human participants are asked to assess the perceived visual quality of the shown media using a quality scale. Responses from multiple human subjects for a specific stimulus are often summarised using the subjective mean opinion scores (MOSs) [153]. MOSs can effectively measure how the quality of visual stimuli is perceived. However, these are expensive and time-consuming to obtain, and as such difficult to integrate into real-world systems to enable real-time visual quality monitoring and control.

Owing to these issues, it has become necessary to develop reliable objective quality metrics that can automatically assess the quality of visual media as seen by humans. Such automatically predicted MOS scores can also help improve future work by giving designers and testers objective metrics during the design and testing stages and reducing the need for testing with costly human subjects. The objective assessment process, however, must be well correlated with the subjective assessment process. Similar to subjective visual quality assessment, objective visual quality assessment can also be broken down into three types: full-reference (FR), reduced-reference (RR), and no-reference (NR) or blind [238]. This thesis focuses on the FR type as undistorted shapes are often available in many

applications, and the subjective and objective quality assessment studies are explored in Chapters 3, 4 and 5.

Life Cycle of Visual Media Processing

The visual media life cycle is essential for processing all multimedia data (images, 3D shapes and videos). During the different stages of producing and utilising visual media, various artefacts may be introduced. So we summarise each stage of the visual media life cycle, including acquisition, processing, storage and transmission, and display, as shown in Figure 1.2.

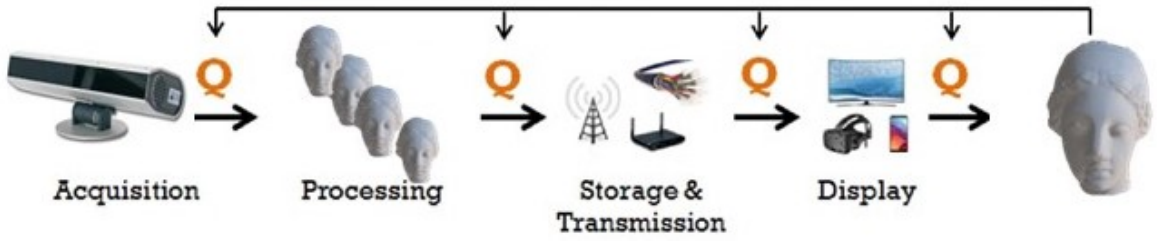


Figure 1.2: Visual media life cycle. In each step, quality degradation may occur, and a quality estimator (Q in the figure) is used to decide on appropriate enhancement.

Acquisition: The acquisition or capture phase of the visual media production is problematic in various ways, such as if there is camera movement, or if an object moves during visual capture. Indeed, blur is one kind of artefact that occurs through movement [85]. Sensor technology limitations in the camera are further limitations during image acquisition. Captured images may also contain noise, especially in low light settings, and distortions due to imperfection of optical components in the camera. In the case of an improper sampling rate, an A/D converter on a camera may produce aliasing artefacts during capture.

Processing: The most common processing techniques are compression (encoding and decoding), enhancement, and multiple image fusion to create a panoramic image or construct a 3D shape, all of which may involve degradation of the captured data. For instance, image compression using block-based codecs often results in block artefacts, whereas

image compression using wavelet-based codecs typically results in ringing or aliasing artefacts [125, 123, 135]. Meanwhile, automated image stitching to generate panoramic images may contain parallax errors [24], and depth triangulation in the construction of 3D shapes may contribute to geometry noise [95].

Storage and transmission: Artefacts created at this stage result from the transformation technique employed to modify visual media such as images and 3D shapes. Information loss frequently happens when storing or transferring (e.g., streaming) visual media due to network or device limitations. Packet loss and visual delay are examples of artefacts that occur at this step [23, 164].

Display: Artefacts that result from this stage occur because of the requirement to transfer visual media such as an image or video to a particular device configuration. This procedure consists of resampling, trans-coding, and tone-mapping [103, 75, 29]. Also, artefacts can be caused by the display technology itself. For instance, delayed temporal response of liquid crystal display (LCD) and the hold-type LCD rendering approach can produce motion blur artefacts [157].

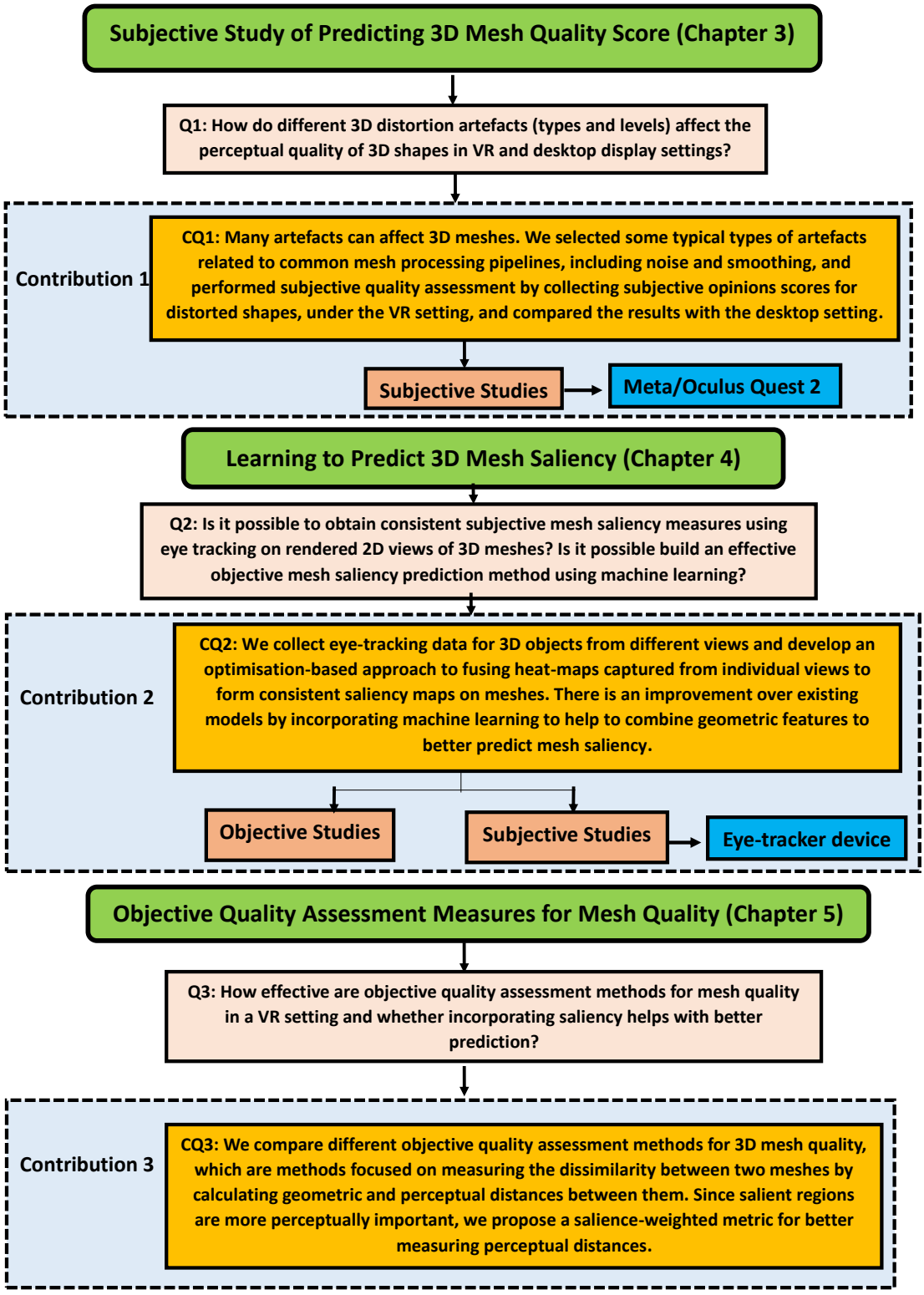


Figure 1.3: Scope and contribution of this thesis. Each box represents a different element: (Q) presents the research questions, (CQ) presents the contributions and green boxes correspond to the chapters in the thesis.

1.3 Research Questions and Contributions

This thesis investigates the following research questions. As shown in Figure 1.3, there is a list of research questions related to contributions and chapters.

As discussed before, the thesis addresses subjective and objective quality assessment for 3D meshes, comparing the VR and desktop settings, and studies related problems of mesh saliency, which can be an important aspect to understand perceived quality.

- **Q1.** How do different 3D distortion artefacts (types and levels) affect the perceptual quality of 3D shapes in VR and desktop display settings?
- **Q2.** Is it possible to obtain consistent subjective mesh saliency measures using eye tracking on rendered 2D views of 3D meshes? Is it possible build an effective objective mesh saliency prediction method using machine learning?
- **Q3.** How effective are objective quality assessment methods for mesh quality in a VR setting and whether incorporating saliency helps with better prediction?

To answer these questions, the following contributions are outlined:

- (CQ1) Many artefacts can affect 3D meshes. We selected some typical types of artefacts related to common mesh processing pipelines, including noise and smoothing, and performed subjective quality assessment by collecting subjective opinions scores for distorted shapes, under the VR setting, and compared the results with the desktop setting (Chapter 3).
- (CQ2) We collect eye-tracking data for 3D objects from different views and develop an optimisation-based approach to fusing heat-maps captured from individual views to form consistent saliency maps on meshes. There is an improvement over existing models by incorporating machine learning to help to combine geometric features to better predict mesh saliency (Chapter 4).

- (CQ3) We compare different objective quality assessment methods for 3D mesh quality, which are methods focused on measuring the dissimilarity between two meshes by calculating geometric and perceptual distances between them. Since salient regions are more perceptually important, we propose a salience-weighted metric for better measuring perceptual distances (Chapter 5).

1.4 Thesis Structure

Chapter 2: **Background and Related Work** provides an introduction to the computer vision literature on VQA with regards to image and 3D mesh metrics, and examines what affects the perceived quality of multimedia. It also discusses how various artefacts affect VQA in normal display and VR settings.

Chapter 3: **Subjective Study of 3D Mesh Quality Scores in Virtual Reality** describes how a subjective VQA experiment was conducted, which aims to understand the impact of several factors that created 3D mesh distortions, including noise and lack of details. The experiment shows how these types/levels of distortions have an effect on the perceived quality of 3D meshes and enables improved user experience in VR settings.

Chapter 4: **Learning to Predict 3D Mesh Saliency** as saliency is related to the perceptual quality of the 3D mesh, we present new quantitative methodologies for predicting 3D mesh saliency, measuring the perceptual importance of different regions on a mesh. This chapter reports on the collection of eye-tracking data for 3D objects from different views and develops an optimisation-based approach to fusing heat-maps captured from individual views to form consistent saliency maps on meshes. We further develop a learning-based approach that regresses local surface characteristics to predict mesh saliency on a new shape.

Chapter 5: **Objective Quality Assessment Measures for Mesh Quality** presents a comparison of different quality assessment methods for mesh quality. These methods

measure how different two meshes are from each other by determining their geometric and/or perceptually related distances. We then compare the results of these methods with subjective scores for both VR and desktop settings. We further demonstrate incorporating mesh saliency helps with better objective mesh quality assessment.

Chapter 6: **Conclusion and Future Work** summarises the outcomes of this thesis and discusses directions for future research, in terms of including technological improvement, human perspective, and 3D complexity, as well as an objective approach to study perceived visual quality in 3D mesh assessment.

1.5 Publications Related to This Thesis

Listed below are the refereed publications resulting from my PhD dissertation research:

1) ALfarasani, Dalia A., Thomas Sweetman, Yu-Kun Lai, and Paul L. Rosin. “Learning to predict 3D Mesh saliency.” In 2022 26th International Conference on Pattern Recognition (ICPR), pp. 4023-4029. IEEE, 2022.

2) Alfarasani, Dalia A., Lai, Yu-Kun and Rosin, Paul L. 2023. ”Subjective Study of 3D Mesh Quality Scores in Virtual Reality.” In 2023 7th International Conference on Virtual and Augmented Reality Simulations (ICVARS), pp. 7-13. ACM, 2023

1.6 Summary

In this chapter, the background and motivation for the present work were discussed. The hypothesis and the main research questions were presented, and there was an overview provided regarding the structure and research contributions. Before moving to the main technical discussion of the thesis, the next chapter will provide a more detailed background and exposition of related work in visual quality assessment, including subjective

and objective image quality assessment and 3D mesh quality as well as mesh saliency, which sets the thesis against the context of existing work in this area.

Chapter 2

Background and Related Work

Overview

In this chapter, we illustrate the background and related work in computer vision on visual quality assessment (VQA) and explore the artefacts that might affect media quality. The 3D mesh quality assessment literature review is limited compared with VQA in 2D imaging, and the methodological background is necessarily generalised to incorporate a wider pool of literature. We present various methods used in 2D and 3D VQA. Moreover, we present 3D mesh saliency to predict the importance of local regions of the shape. Also, we look at display settings, such as VR, and how the quality of the visual stimuli affects the viewer's perceptions in VR setting. Finally, We looked at various applications that use 2D desktop and VR settings.

2.1 Introduction

In this review of computer vision literature, 3D mesh visual quality (MVQ) metrics are classified into three categories based on the availability of reference objects [8]. Full-reference (FR) is a category when the reference is entirely available. No-reference (NR) is used when no information about the reference is available. Reduced-reference (RR) is applied when only a portion of the reference is available, such as a subset of features extracted from the reference. Having briefly presented the three types of MVQ metrics, we focus now on the full-reference (FR) MVQ metrics, which are most related to the work

in this thesis.

The most common existing methods work on full reference (FR) metrics, such as mesh structural distortion measure (MSDM) [111], Multiscale Mesh Structural Distortion Measure (MSDM2) [108], Dihedral Angle Mesh Error (DAME) [212], Fast Mesh Perceptual Distance (FPDM) [216], and Tensor-based Perceptual Distance Measure (TPDM) [209, 56]. These metrics are generally computed in two steps: first, by calculating the vertex-based or edge-based quality values from the reference and the distorted mesh and; second, by aggregating the local quality values into a single score based on spatial pooling to reflect the overall quality of the distorted mesh.

3D mesh quality research is more in the developing stage compared to image visual quality. A number of studies on MVQ measures [111] have appeared to be motivated by studies on IVQ metrics [223]. Li et al. [117] developed a spatial pooling technique for IVQ measurements through machine learning techniques. They retrieved statistical descriptors from the local quality map to describe the overall quality of the image and showed that quality is sensitive to local quality map distribution. Such works show that there is some similarity between MVQ and IVQ measurements, given that both are intended to correspond with visual perception. Research in the VQA literature indicates that image quality evaluation and mesh quality assessment have broadly comparable perceptual considerations [111, 223].

Research on VQA in image quality evaluation has undertaken several types of research on IVQ measures using machine learning techniques [112, 70, 150, 151, 86, 69, 231] while the majority of studies on mesh quality assessment have explored MVQ metrics [111, 108, 212, 216, 209, 56]. By doing so, such methods have explicitly created specific visual perception models for mesh quality. Using machine learning approaches, Lavoue et al. [109] have developed the multi-attribute computational model (MACM). However, the MACM measure does not account for the perceptual properties of the human visual system, and its capacity to generalise across databases has not been assessed in the computer vision literature [109]. We anticipate that machine learning approaches

will play an increased role in MVQ measures in the short- and long-term, especially when a sizeable subjective database for mesh quality evaluation becomes more accessible in the future.

Mesh saliency detection [127] as a way to predict quality assessment has also garnered interest in visual perception as a research topic connected to mesh quality. Most existing methods of measuring salience use heat-maps that focus on highlighting salient areas. Several computational saliency algorithms [115, 192, 120, 155, 204] have been developed to discover the perceptually significant mesh parts where human visual attention is concentrated. Since the human visual system is the recipient of both mesh visual quality and mesh saliency, there is potential to enhance the performance of MVQ metrics by including mesh saliency in the metric. Kim et al. [102] performed user research with an eye-tracking experiment and assessed the association between the mesh saliency map produced by the approach [115] and the fixation map collected from the eye-tracking experiment. Chen et al. [37] constructed a benchmark using pseudo ground truth mesh saliency maps based on focus points (Schelling points) and used a regression model to predict the saliency map of 3D meshes with the benchmark. Tasse et al. [205] developed criteria for evaluating the performance of computational 3D saliency approaches [192, 186, 205] using the benchmark [37]. There is currently a lack of quantitative research on the accuracy and dependability of current mesh saliency detection techniques. In the published work [115, 192, 120, 155, 204], the efficacy of the mesh saliency detection approach was mostly supported by application-guided assessment [115, 192] or subjective visual analysis [120, 155, 204]. A number of studies show that mesh saliency can improve the results of graphics applications, such as mesh simplification and viewpoint selection [115, 192].

In the following figure, we present a diagram which summarises the quality assessment techniques that can be used in 2D and 3D. As the 3D mesh literature review is limited, we start with showing 2D IQA, which is closely related. There are three types of methods used both in 2D and 3D, namely full reference, reduced-reference and no-

reference. All of these methods have some metrics that can be used to predict the quality. As we focus on full-reference cases, we present in Figure 2.1 chapter structure which shows how we start with 2D IQA as this has extensive metrics that can be used in 3D MQA, and extended to use in eye movement and virtual reality.

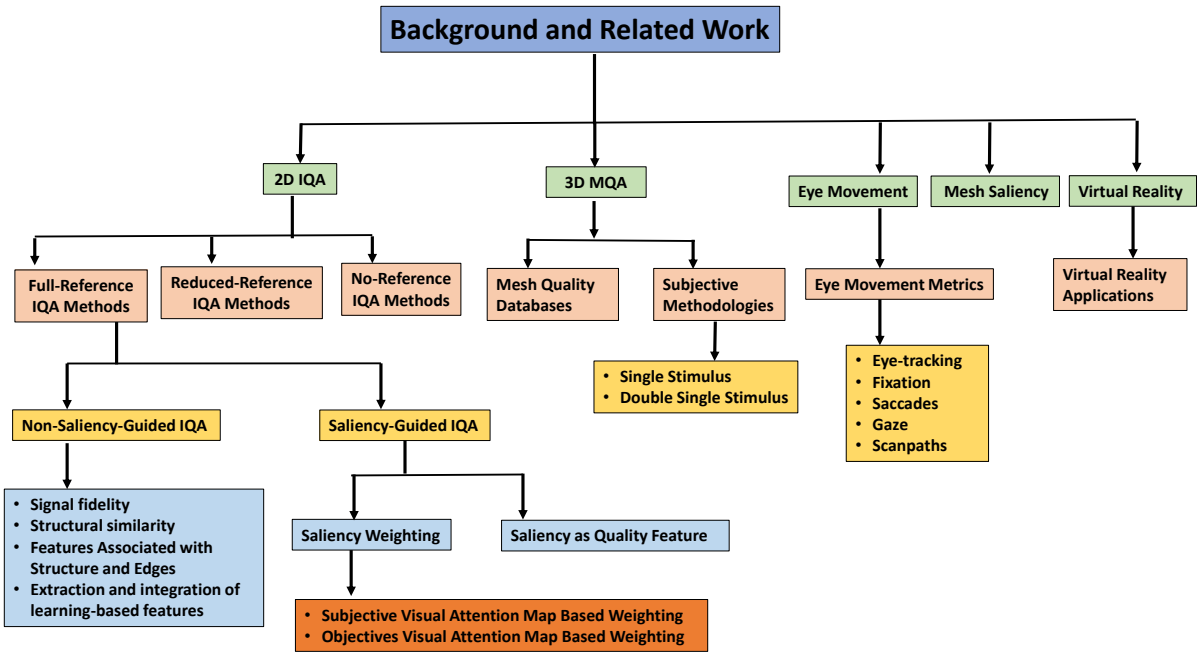


Figure 2.1: A diagram summarising background and related work in quality assessment.

2.2 2D Image Quality Assessment

Since 3D mesh quality assessment is quite limited, most methods were adapted from 2D image quality assessment (IQA). More than a hundred FR measures for 2D images may be found in [160]. Most blind quality measurements use degradation-based approaches, focusing on the most common degradation types (blocking, ringing, blur) [222, 182]. Accordingly, RR metrics provide important alternative methods since only a portion of the original image features are expected to be available. The same methods described in [41, 251, 181] are used to evaluate the visual quality of stereoscopic images. Several spatial pooling algorithms emphasising low-quality image areas have been developed for IVQ metrics [221]. Wang et al. [220] suggested various spatial pooling techniques for the IVQ measures and showed that, among these spatial pooling strategies, the local information content weighted pooling approach delivers the most effective performance for this metric. Wang et al. [221] subsequently provided a sophisticated statistical technique to assess the local information content utilised to weight the quality map. Moorthy et al. [146] introduced the percentile weighting approach for spatial pooling, which assigns more significant weights to low-quality image areas.

Several articles in image quality evaluation have previously studied the introduction of visual attention or computational image saliency into IVQ measures [145, 124, 62, 122, 244]. These are based on the assumption that distortions occurring in more salient regions of an image have a more significant impact on image quality, either the visual fixation map or the image saliency map has been incorporated into the IVQ metric to improve the performance of the metric. The experimental results have confirmed this assumption. Fewer studies have investigated the link between mesh saliency and meshed visual quality, not to mention the incorporation of the mesh saliency map into MVQ measures, compared to those on image quality evaluation. In [246], the image saliency map was incorporated into IVQ measurements based on statistical analysis. Based on the results of their study, they concluded that computational saliency models improve IVQ

metrics when incorporating image saliency maps. Computational saliency metric and IVQ metrics must be combined to determine the exact amount of performance gain [246].

In this section, we will present objective IQA metrics that have been used in computer vision research. These measures were developed for use in general-purpose IQA, and it is presumed that they can handle a wide variety of distortions. IQA measurements can be classified as either full reference (FR), reduced reference (RR) and no-reference (NR), depending on whether or not a distortion-free reference image is readily available [219, 221, 146, 232].

2.2.1 Full-Reference Image Quality Assessment Methods

In FR IQA, the original reference image is used to predict the quality of the distorted image. Most FR IQA measures follow a similar framework for the most straightforward IQA tasks, namely feature extraction from both images (reference and distorted) followed by distance calculation. Usually, feature extraction can be collected from the spatial domain. We review IQA metrics in the following sections using the underlying features as a lead since feature extraction is key to FR IQA measures. We cover two forms of IQA measurement in this subsection: methods based on uniformly evaluating IQA over the whole images, and saliency-guided IQA where more emphasis is put on salient regions.

2.2.1.1 Non-Saliency IQA Methods

- **Signal fidelity**: Traditional signal fidelity measurements such as mean square error (MSE) and peak signal-to-noise ratio (PSNR), are often disputed since they do not consider image signal properties or the human visual system (HVS).
- **Structural similarity**: The structural similarity index measure (SSIM) was first hypothesised by Wang et al. [223]. Since the HVS is sensitive to image structure, some methods extract image structures and calculate structural similarity as quality.

The brightness, structure and contrast characteristics are taken from the image. Following this, the SSIM index is computed using the reference image and the image that has been distorted. More details are discussed in Chapter 4.

- **Features associated with structure and edges**: Structure and edge information provide useful clues for measuring image similarities. Zhan et al. [243] analysed image quality by combining the distribution of several types of structural distortion with the degree of structural differences. To estimate the image quality of a distorted image compared to a reference image, the position of structural information is calculated, and the category distortion index is determined by the mutual correlation between the gradients of the two images. Zhang et al. [247] suggested using an approach known as the non-shift edge-based ratio (NSER). To get an accurate assessment of the level of quality, the authors used the variance in the total number of edge points in a non-shift edge map. Capodiferro et al. [30] developed an approach to combine a structure loss measure with a definite indication of impairment types. This would result in a more accurate assessment of structural loss. Two different measures were used by di Claudio et al. [52] to quantify the influence of detail losses and misleading details on perception. Ding et al. [53] used anisotropy and local directionality to assess major structural information change. Images were judged based on superpixel luminance similarity, chrominance superpixel similarity, and pixel gradient comparisons [203].

- **Extraction and integration of learning-based features**: Some IQA methods are based on the extraction and integration of learning-based features. They use IQA measures based on machine learning approaches for discovering and integrating elements linked to image quality. Using machine learning in feature integration has the benefit of allowing the model to learn potentially complicated relationships between images and quality measures by exploiting the knowledge from the training data, resulting in better performance. Singular value decomposition (SVD) based features were used by Narwaria et al. [150], and support vector regression (SVR) was used as a feature fusion technique [185]. The distance between the singular val-

ues of the reference image blocks and the distorted image blocks. The final quality is obtained by calculating a global rating for each block. Liu et al. [126] present a new parallel boosting measure built on the strength of previous FR measures. For this purpose, the authors used the SVR to combine the quality features extracted by the most advanced state-of-the-art FR measurements.

A nonnegative matrix factorisation (NMF) parts-based model was used by Wang et al. [215] to estimate image distortions. The extreme learning machine (ELM) output is the final quality score. Support vector classification and k-nearest-neighbour regression were used by Peng et al. [162] to develop a two-stage framework. The authors suggest a probabilistic strategy for distortion-specific features by integrating the SSIM, VSNR (Visual Signal-to-Noise Ratio) [32], and VIF (Vision Information Fusion) [182] measurements using the k-nearest-neighbour regression method.

Another well-known learning-related technique for determining which features to employ is sparse representations. According to Yuan et al. [241], the quality map was created by comparing the sparse representations of reference and distorted patches to each other. Using the kernel ridge regression (KRR) [213], the local quality was incorporated into the overall quality score. Ahar et al. [9] developed a sparse coding method. In their approach, an initial Fourier basis ranked the sparse coefficient amplitudes and then tested their correspondence to the reference images.

Deep learning has shown success on various visual problems. Also, some methods have used deep learning to predict the quality [158, 252]. A deep neural network (DNN) and deep similarity (DeepSim) were proposed by Gao et al. [68]. The DNN in [68] computed the final quality by combining the local similarities of DNN features between two images. Initially, a local linear model (LLM) was used to detect the degradation between the reference and distorted images. Then they proposed a distortion-specific compensation technique to deal with the offset induced by different image distortions. In Wang et al. [215], the score offset was calculated using a convolutional neural network (CNN). An end-to-end approach to feature learning and regression using a neural network was developed by Bosse et al. [22] in their

study of FR and NR IQA. This approach may be used for FR or NR MQA with few modifications and it learns the local quality and weights simultaneously.

2.2.1.2 Saliency-Guided IQA

Image quality and visual attention are two areas of study that are closely linked. This is partly because, for viewers, image quality is significantly connected with artefacts in the salient regions. Human attention to visual information may be shown as a weighted map to emphasise salient regions and encourage image quality assessment measures. The technique of selecting and focusing on a specific aspect of stimuli is referred to as human visual attention and refers intrinsically to a behavioural and cognitive process. Two categories may be used to define saliency-guided IQA methods: subjective visual attention map-based weighting and objective saliency map-based weighting.

- **Subjective visual attention map based weighting:** In recent years, an eye-tracking data-guided pooling technique has been used to enhance the performance of IQA measurements. Larson et al. [214] performed eye-tracking tests and studied changes in visual attention in terms of various types, levels, and viewing techniques with regard to image distortion. Liu et al. [124] used eye-tracking data as the weighting information in IQA measures. They examined the efficacy of visual attention data obtained during free-viewing and quality-rating activities and found that the free-viewing task promoted IQA measures more effectively. Liu et al. [124] also explored the effects of image content on the promotion effect of visual attention data on IQA measures. They indicated that visuals with minor inter-observer attention differences benefited greatly from saliency pooling.

  Min et al. [142] compiled eye-tracking data for seven well-known IQA datasets and used the human fixation data for the quality map pooling step. Wang et al. [217] provided two unique pooling algorithms to include the saliency map in IQA measures, and they showed that distortion types significantly impacted IQA metric

performance gain. A suitable pooling mechanism must be chosen for the particular IQA measure. Rai et al. [168] performed an eye-tracking experiment using an HMD and developed a saliency-guided pooling technique to produce IQA measurements for VR. Zhang et al. [247] improved IQA measurement using eye-tracking data from distorted images. They indicated that eye-tracking data of both the reference and distorted images might improve the performance of IQA measurements.

- **Objective visual attention map based weighting:** Owing to the expense and limited availability of eye-tracking sensors, the objective saliency model is also used for large-scale IQA applications. Ma et al. [132] presented a pooling technique for applying saliency data to MSSSIM and vision information fusion (VIF). They split the image into overlapping blocks and computed the local mean value of each block as a weighted coefficient. Zhang et al. [249] weighed 12 IQA indicators using 20 saliency models. The statistical findings demonstrated that saliency-guided weighting improved IQA metrics and that the performance improvement was strongly dependent on distortion types and saliency models. Wen et al. [225] developed an FR IQA measure with saliency weighting, and a Fourier transform-based objective saliency model was employed for weighting.

  Zhang et al. [248] suggest an effective method of measuring saliency dispersion for identifying stimuli, using an adaptive strategy for incorporating saliency into IQA indicators. Mittal et al. [143] introduced an objective salient area recognition technique for JPEG-distorted images by combining low-level characteristics such as contrast, brightness, and quality index. This saliency model promotes the implementation of high-quality map pooling algorithms. Harel et al. [81] computed the saliency map using the graph-based visual saliency (GBVS) [82], and included the Itti et al.'s models [92] to weight the quality measure and improve performance.

  By including saliency-based pooling, Nasrinpour et al. [152] enhanced the tone-mapped image quality index (TMQI) [234]. The enhanced technique separated the image into small patches and determined the weighting factor of each patch using the attention by information maximisation (AIM) [25] saliency model. Similar to

[152], Kundu et al. [104] enhanced the traditional TMQI [234] by saliency-based pooling based on the model of Itti et al. [94]

2.2.1.3 Using Saliency as Quality Feature

It is possible to employ visual saliency as a quality feature, as an alternative to a weighted map. Various elements, such as image compression, various image transformations, degradation, sound, high-level face information, and mental health, may also affect visual attention and image saliency [124, 248, 142]. As an NR IQA feature, saliency information is employed [88]. The saliency-guided natural scene statistics (NSS) feature is shown to be an effective descriptor in assessing image quality. By considering saliency change, Zhang et al. [244] devised an FR IQA measure called visual saliency-induced VSI. The small diamond search pattern (SDSP) [247] saliency model produces the saliency map employed as a weighting map in the final pooling process.

2.2.2 Reduced-Reference Image Quality Assessment Methods

Reduced-Reference Image Quality Assessment is a type of method used to evaluate the quality of an image based on a limited amount of reference information. The goal of reduced-reference image quality assessment is to estimate the perceptual quality of an image without having access to the original, pristine reference image. In this context, Wu et al. [230] proposed visual information fidelity, which refers to the preservation of important visual features, such as edges, textures, and structures, in the distorted image compared to the reference. The method employs a reduced-reference approach, meaning that it uses a subset of features from the reference image to evaluate the quality of the distorted image. This reduces the computational complexity and the reliance on full reference images. The chosen features are typically extracted using computer vision techniques and can include statistical measures, local image descriptors, or higher-level visual attributes.

The quality assessment process involves comparing the selected features between the reference and distorted images. Quality can be measured by a variety of metrics, including mean squared error and structural similarity [171]. These metrics quantify the differences between the reference and distorted images providing an estimate of the perceived image quality. It provides a valuable tool for evaluating image quality since it preserves essential visual information without a full reference image. As a compromise between full-reference and no-reference approaches, it can be applied to a variety of image processing applications to assess quality accurately.

2.2.3 No-Reference (Blind-Reference) Image Quality Assessment Methods

Blind Image Quality Assessment (BIQA) is a field of study focused on developing algorithms and techniques to assess the quality of images without relying on reference images or human subjective judgments. The goal of BIQA is to automatically evaluate the visual quality of images based on their inherent characteristics and perceptual attributes. Li [118] proposed a BIQA method that analyses various image features such as sharpness, contrast, colour accuracy, noise, and distortion to determine the perceived quality. These algorithms aim to replicate human perception by modelling the visual system and understanding how different factors affect image quality. By considering both low-level features (e.g., pixel-level information) and high-level features (e.g., semantic content), BIQA algorithms can provide comprehensive evaluations. To achieve this, some BIQA approaches [143, 182] employ various techniques such as traditional machine learning, statistical modelling, and deep learning. They learn from large datasets of images that are annotated with quality scores to establish relationships between image features and subjective quality judgments. The trained models can then predict the perceived quality of new images.

BIQA has numerous applications, including image compression, image enhancement,

image restoration, and image retrieval. By automatically assessing image quality, these algorithms can assist in optimising image processing pipelines, enhancing user experience in multimedia applications, and aiding in image-based decision-making tasks. BIQA is a field dedicated to developing algorithms that can automatically evaluate the visual quality of images without relying on human judgments or reference images.

2.3 3D Mesh Visual Quality Assessment

Several perceptually motivated metrics have been developed for 3D meshes inspired by image quality metrics. Abouelaziz et al. [3] produce 2D projections of 3D models from various viewpoints, and then use patches from the generated images to input into a Convolutional Neural Network (CNN) model to generate feature vectors for the reference and test meshes. The quality score is calculated using the Kullback-Leibler (KL) divergence between feature vectors. Karni and Gotsman [98] were the first to try to incorporate some perceptual insights to improve the accuracy of geometric distortion measurements. They proposed combining the Root Mean Square (RMS) distance between corresponding vertices with the RMS distance of their Laplacian coordinates (which reflect the degree of smoothness of the surface) to improve the accuracy of geometric distortion measurements. Chetouani et al. [40] suggested employing a Support Vector Regression (SVR) model to combine various commonly used full-reference quality measures to increase the correlation between prediction and human observations.

Perceptual metrics based on global roughness variation are proposed by Corsini et al. [43] to quantify the quality of a watermarked mesh. They defined roughness as the variance of the difference between a 3D model and its smoothed counterpart and the variance of the dihedral angles between neighbouring faces assessed at multiple resolutions. Karni and Gotsman [99] use the Geometric Laplacian (GL) to assess compression techniques, which depends on each vertex's smoothness. For the evaluation of 3D models, Pan et al. [157] offer a metric based on the geometric and texture resolutions. Their research

shows that for textured surfaces, image texture is generally more important than model geometry in perceptual contribution. Lavoué et al. [111] developed MSDM, inspired by human perception to evaluate the quality of watermarking algorithms. Bian et al. [20] created a geometry-based perceptual metric that employs strain energy, which measures the energy that causes mesh deformation between the reference mesh and the distortion version. This measure has been used to assess watermarking, compression, and filtering operations. The global roughness difference calculated by Corsini et al. [44] and Wang et al. [216] is straightforward because they compute multiple global roughness values for each model. Some other works involve bottom-up quality measurements, incorporating perceptually motivated techniques, such as visual masking [209, 212, 216]. A recent study [45] examined these works and compared their performance against the correlation with mean opinion ratings obtained from subjective rating experiments. MSDM2 [108], FMPD [216], and DAME [212] were highly predictive visual quality metrics identified in this investigation.

According to the existing study, Lavoue et al.'s MSDM2 [109], Wang et al.'s FMPD [216], and Vasa and Rus's DAME [212] are strong predictors of visual quality. Aside from these works on global visual fidelity assessment (suitable for supra-threshold distortions), several recent relevant works have been introduced: for example, in 2016, Nader et al. [149] introduced a bottom-up visibility threshold predictor for 3D meshes (assuming a flat-shaded rendering). Also, Guo et al. [76] investigated the local visibility of geometric artefacts and demonstrated that curvature might be a good predictor of local distortions. Finally, Lavoue et al. [112] presented comprehensive research that looked at the usage of image metrics computed on rendered images for measuring the visual quality of 3D models (without texture). It demonstrates that some of them (in particular, Multi-Scale SSIM (MSSSIM)) may provide great performance. Yildiz et al. [238] conduct a crowdsourcing project to get data on mesh quality from human observers. They define the distance between two meshes as the weighted Euclidean distance of their feature vectors, constructed using the histogram statistics of several mesh descriptors such as curvature and roughness. Feng et al. [64] suggest a spatial pooling method. They construct the

distortion distribution from a reference and a test mesh and extract statistical characteristics (standard deviation, mean, max, min, and three quartiles) as features. To compute the local distortion distribution, Torkhani et al. [209] utilise their proposed tensor-based perceptual distance measure (TPDM). They used machine learning to train an SVR (Support Vector Regression) model to determine the link between the distortion distribution and quality scores by pairing feature vectors with Mean Opinion Scores (MOS) for the related meshes.

Similarly, various NR measures have been created for MVQ evaluation. Abouelaziz et al. [6] suggested a blind technique based on mean curvature features and the general regression neural network (GRNN) for feature learning and quality prediction. Nouri et al. [154] employed visual saliency and support vector regression (SVR). The authors developed an NR technique known as the 3D blind mesh quality assessment index (BMQI). The works [8, 4] also used SVR to build their models. In [5], hand-crafted perceptual characteristics (dihedral angles and mesh shape) taken from the 3D mesh and displayed as 2D patches of a predetermined size are used to feed into a CNN. The CNN is fed with rendered images from 3D objects in [7], and the view is altered by rotating the 3D mesh by 60 degrees along the X and Y axes. In [4], a patch-selection technique based on mesh saliency was established to provide greater weight to interesting areas.

The literature studies for mesh quality assessment like Lavoué et al.'s MSDM2 [109], Wang et al.'s FMPD [216], and Váša and Rus's DAME [212] are strong predictors of visual quality. Váša and Rus [212] studied dihedral angle discrepancy, whereas Lavoué et al. [107] suggested metrics based on local variances in curvature statistics. Local changes of attribute values at the vertex or edge level are included in these metrics, which are subsequently aggregated into a global score. On the other hand, Corsini et al. [45] calculated global roughness values per model before working out global roughness differences as measures. Torkhani et al. [209] incorporated perceptually motivated methods such as visual masking, which are similar to bottom-up image quality measurements.

There are existing research works that utilised 3D models to estimate quality of 3D

distorted meshes like [42, 44, 107, 111, 163, 189, 212, 28, 228, 183, 174, 207] which conducted subjective assessments using 3D static or dynamic models, but none of them used a virtual reality setting. The authors conducted a subjective assessment survey to determine how downsampling or introducing coordinate noise to a 3D point cloud impacts perceived quality [13, 95]. They also presented a subjective analysis of 3D point cloud denoising algorithms using the Double Stimulus-Impairment-Scale (DSIS) approach and the correlation with objective measures which show a high correlation.

The most recent work by Bulbul et al. [27] presented an excellent review and comparison of different environments, approaches and materials. For example, Nehme et al. [153] used a virtual reality experiment to see how the explicit reference affects the quality evaluation of coloured 3D models. They conducted a psycho-visual study to compare the performance of two methods: ACR-HR (with hidden references) and DSIS (with explicit references). They used two sets of observers, and two tests were given to each group in a different order. The experiment utilised the HTC Vive Pro virtual reality headset in fixed position mode in an immersive virtual world. Their focus is to analyse the subjective quality assessment methods for coloured meshes. In contrast, our work measures subjective quality assessment for meshes with geometry only and compares the results with desktop settings.

The 3D metrics used for simplification [100, 89] are local error measures from vertex to vertex to generate a single distance value between two meshes. Existing global 3D metrics, on the other hand, are designed to measure specific artefacts produced by watermarking [71, 14], or compression [99] algorithms; these artefacts are mostly uniform noise, so these metrics are not appropriate for evaluating smoothing, simplification, or other non-uniform processing against a mesh. Such metrics can play a crucial role in computer graphics by replacing typical geometric distances for measuring and driving 3D mesh processing systems and algorithms. However, for 3D objects, objective quality assessment research is still in its early stages; only a few metrics have been proposed, and they come with many limitations (for example, objects to compare must have the same

connectivity or sampling density).

When it comes to assessing mesh visual quality, there are several limitations associated with the metrics used for evaluation. Here are some common limitations of mesh visual quality assessment metrics:

Subjectivity: Assessing visual quality is inherently subjective, as it relies on human perception and preferences. Different individuals may have varying opinions on what constitutes good or bad visual quality. Metrics that attempt to quantify visual quality may not fully capture the subjective aspects of human perception.

Lack of Ground Truth: Unlike other domains such as image or video processing, there is often no universally accepted ground truth for mesh visual quality. While there are objective measures to assess geometric fidelity (e.g., distance-based metrics), evaluating overall visual quality is challenging due to the lack of a definitive reference standard.

Simplified Criteria: Many existing metrics for mesh visual quality assessment focus on specific aspects such as geometric distortion, surface smoothness, or curvature preservation. While these criteria are essential, they may not encompass the full range of factors relevant to overall visual quality, including shading, lighting, texture mapping, or material appearance.

Computational Complexity: Some advanced metrics that aim to capture perceptual quality require computationally expensive operations such as ray-tracing or global illumination simulations. Applying these metrics to large-scale or real-time applications may be impractical due to their computational demands.

Lack of Consensus: The research community lacks a metric for a single comprehensive metric for mesh visual quality assessment. Various metrics exist, each with its own strengths, weaknesses, and underlying assumptions. This lack of consensus makes it challenging to compare and interpret results across different studies.

Addressing these limitations is an ongoing research effort. Researchers continue to ex-

plore new metrics, combine multiple criteria, and develop subjective evaluation methodologies to better assess the visual quality of meshes. However, it remains a complex and challenging problem due to the subjective nature of visual perception and the multidimensional aspects of mesh representation. As the focus here is 3D mesh quality assessment, I will cover this in greater detail in Chapters 3, 4 and 5.

2.4 Subjective 3D Mesh Quality Databases

The work proposed by Watson et al. [224] is the first study to assess 3D static mesh quality measured visual fidelity by simplifying meshes. A previous study Corsini et al. [44] examined watermarking 3D meshes to accustom existing experimental protocols to subjective evaluation of 3D mesh quality. Recently, several publicly available databases of static meshes with associated mean opinion scores (MOS) have been released. *The LIRIS/EPFL General-Purpose-Database* we explain more in Section 3.2 as it is used as a basis for our study.

The LIRIS Masking Database [107] was created to study the spatial visual masking effect. This database includes 22 impaired models derived from 4 reference meshes. Models were carefully selected to offer a broad range of roughness, and the noise was added in either rough or smooth regions.

The IEETA Simplification Database [187] includes 30 simplified models (obtained by using different vertex reduction algorithms) derived from 5 reference meshes.

The UWB Compression Database [212] contains 63 impaired, geometrically compressed meshes derived from 5 reference models.

The above databases have been used to evaluate and compare the most recent objective perceptual quality metrics for 3D static meshes [45], such as MSDM2 [108], DAME [212], FMPD [216] and TPDM [209].

2.5 Subjective Methodologies

This section reviews subjective methods for evaluating perceptual visual quality. The next subsections detail the different subjective quality assessment methodologies that can be used in different media types (including 3D meshes).

2.5.1 Single Stimulus Methodologies

The single stimulus (SS) approach entails showing the experiment participants a series of images, one at a time, and asking them to score their visual quality, as shown in Figure 2.2. The rating scale varies across experiments, and a training period is normally conducted before the trial begins. Because of its simplicity and the minimal number of stages, this subjective technique is a popular choice. The single stimulus approach, for example, has been employed in many studies, including those done by Sheikh et al. [183] and Cheng et al. [39]. The following are some of the most often-used approaches for (SS) subjective quality assessment:

Absolute Category Rating (ACR): ACR [87, 172, 227] is a single subjective quality experiment in which test stimuli are shown one at a time and subjects are asked to score the visual quality of the images on a discrete scale rating from 1 to 5: 1. Bad, 2. Poor, 3. Fair, 4. Good, and 5. Excellent. The advantages of such a method are its simplicity in design and the computation of subjective ratings. Still, it generally necessitates a lengthy training session to familiarise the participants with the grading scale. Furthermore, it has been discovered that subjective judgment is occasionally impacted by participants' thoughts about the stimulus's contents. The ACR-HR approach described below is used to reduce the effect of image content on subjective evaluations.

Absolute Category Rating with Hidden Reference (ACR-HR): ACR-HR, as introduced above, is a variation of the ACR where the original image is "hidden" among the distorted stimuli without informing the subjects of such occurrences. This experimental design

eliminates variation attributable to the participants' opinions on the content and the computation of the Differential Mean Opinion Scores (DMOS) rather than the mean opinion score (MOS), resulting in an exact assessment of the stimuli's quality. This approach is frequently employed in state-of-the-art methods due to its trade-off between simplicity and accuracy. This technique, for example, was used in the Laboratory for Image and Video Engineering (LIVE) 6 large-scale and publicly accessible subjective quality evaluation research and in Cheng et al. [39] to evaluate the perceived quality of learning-based image compression algorithms.

Single Stimulus Continuous Quality Evaluation (SSCQE): SSCQE is a single stimulus subjective quality experiment similar to the ACR but uses a continuous rather than discrete rating scale. An electronic recording device attached to the computer should be used, as indicated in BT.500-14 [26]. The continuous quality scale is similar to the continuous grading scale of objective quality indicators, allowing for a more accurate comparison using such a quality evaluation technique.

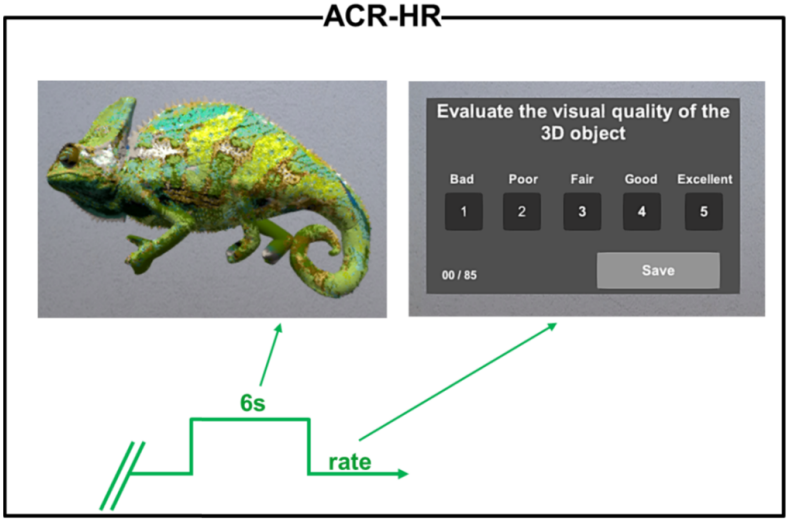


Figure 2.2: Example of the single stimuli method [153]

2.5.2 Double Stimulus Methodologies

The double stimulus (DS) technique is a sort of subjective quality experiment in which respondents are shown two stimuli opposite to each other and asked to rate the difference in quality between the two images as shown in Figure 2.3 [212, 224, 173, 13, 188, 201]. The grading scale varies from experiment to experiment and will be discussed in the next paragraph. In general, double-stimulus techniques take longer than single-stimulus methodologies, but they are more accurate for particular sorts of artefacts, such as variations in the colours of the stimuli. As a result, the approach has lately been employed in a number of subjective quality investigations. Ascenso et al. [17] and Testolina et al. [207] use DS trying to evaluate the quality of learning-based image coding solutions. The following are some of the most often-used approaches for DS subjective quality assessment:

Degradation Category Rating (DCR): It involves displaying to participants two stimuli side by side and asking them to score the extent of impairment of one versus the other, which serves as a reference, using a discrete grading scale. The scale 1 worse quality and 5 best quality) 1. Very annoying 2. Annoying 3. Slightly annoying 4. Perceptible but not annoying 5. Imperceptible. The main disadvantage of this strategy is that it yields fewer ratings in the same amount of time as the ACR since the subjects are asked to view two stimuli rather than one, making it a slower alternative. However, because the scoring is based on the impairment between images rather than the overall quality, it benefits from not being impacted by the participant's evaluation of the content. Furthermore, compared to ACR, DSIS makes it easier to identify colour degradation between two images. The grading subject is always given with the same reference stimulus in the same place. DSIS is frequently utilised in the realm of subjective image quality evaluation, as demonstrated by Ascenso et al. [17].

Double-Stimulus Continuous Quality-Scale (DSCQS): is a subjective quality evaluation approach comparable to DSIS in which participants are asked to score the overall quality of both given stimuli on a continuous quality rating scale. The reference stimu-

lus is exhibited at a random position that is unknown to the participant in this manner. This approach is the slowest of those mentioned above since the subjects are required to assess the quality of two stimuli at each phase. This approach is efficient for analysing learning-based compression algorithms [16].

Double Stimulus Comparison Scale (DSCS): In DSCS subjective quality experiments, participants are asked to rate the visual quality of the first stimulus at each stage using the second as a reference. The grades on the discrete grading scale are as follows: -3. Much worse -2. Worse -1. Slightly worse 0. The same 1. Slightly better 2. Better 3. Much better. For example, individuals score all images of the same content including the reference and test stimuli in a randomised sequence. As a result, this is the experiment that requires the most effort and hence the most extended duration. While this approach is the most accurate in comparing the quality of different compression algorithms, it has the disadvantage that the bitrates of the compared stimuli should be as close as possible to ensure a fair comparison. Other scaling possibilities in this strategy might include three (Better, Same, Worse) or perhaps simply two (Better, Worse). Lastly, this technique has proposed partial comparisons of stimuli to shorten the testing time.

This subjective study is the most used method for a range of multimedia types. In 3D point cloud models with varying noise distortion levels and geometry resolution in an AR environment, a subsequent study [12] proposed using the DSIS technique and Mean-Opinion-Scores. The study finds that the geometric complexity of the model influences the assessment score and that objective and subjective measures have little connection.

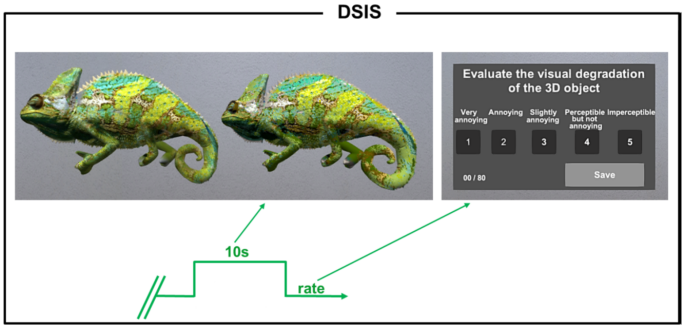


Figure 2.3: Example of the double stimulus method [153]

2.6 Limitations of Perceptual Metrics

The computer graphics community has often exploited knowledge about the Human Visual System (HVS) for several purposes. It is possible to reduce the computational time in a sophisticated rendering system by accurately rendering only the visually essential parts of a 3D scene. Another example is the algorithms that attempt to simplify the geometry while preserving the overall appearance of the 3D model using perceptual criteria. These works are reviewed in the following paragraphs as they are categorised as image-based and model-based (or geometry-based). The image-based metrics render 3D shapes to individual views and are therefore view-dependent, whereas geometry-based perceptual metrics work directly on the 3D models, so are view-independent. Because of this advantage, our study uses geometry-based perceptual metrics as the basis.

2.6.1 Perceptual Metrics Based on Images

Black-box approaches are not concerned with how visual systems work but with defining a function capable of predicting how a human observer will perceive specific visual artefacts, given the visual stimulus as input. An example is the work of Marziliano et al. [135], which aims to detect and quantify JPEG artefacts such as ringing and blur. The advantage of this approach is that it can be used when it is difficult to determine a way to integrate different visual stimuli. The use of mechanistic and black-box perceptual metrics in computer graphics is applied in many applications, such as perceptually-driven rendering and evaluating specific processes like compression or watermarking and mesh simplification.

To evaluate the perceptual impact of mesh simplification from a perceptually-based perspective, Lindstrom and Turk [121] proposed using a simplified version of the Sarnoff model to render the shape simplified from various viewpoints. Williams et al. [228] developed a view-dependent simplification algorithm using a simple Contrast Sensitivity

Function (CSF) model that considers texture and lighting effects. Qu and Meyer [166] suggested that the visual masking effect of 2D texture maps leads to the simplification and remeshing of textured models. Using image-based perceptual metrics, Ferwerda et al. [66] established a masking model, extending Daly's Visible Difference Predictor (VDP) operator [48], that illustrates how surface texture may hide specific visual errors, particularly regarding polygonal tessellation. Perceptual assessment has recently been elevated to a higher level of investigation involving visual systems; for example, Ramanarayanan et al. [169] investigated how changes in the lighting environment impact the perception of geometry, material, and illumination in a scene.

2.6.2 Perceptual Metrics Based on 3D Geometry

Computer graphics applications have a significant challenge when using image-based metrics: the perceived deterioration of still images may not be suitable to measure the perceived degradation of a 3D model as mentioned by Rogowitz and Rushmeier [173] who did subjective experiments that supported this conclusion. According to their findings, when evaluating the quality of a simplified 3D model, the observer's subjective opinions differ depending on whether animation or a collection of static frames from the same animation is utilised. It is challenging to include changes in perceived differences in the perceptual metric when the object moves. The work of Yee et al. [236] was an early effort to combine image movement, visual attention, and saliency. To speed up global illumination rendering, Myszkowski et al. [148] suggested an enhancement to the VDP for the quality assessment of computer-generated animations.

Perceptual metrics based on 3D geometry are used in different applications. Mesh simplification algorithms are one technique for reducing the number of vertices in a model while reducing the impact on visual quality. According to Kim et al. [100], human vision is sensitive to curvature changes, which suggests a Discrete Differential Error Measure (DDEM). The perceptual assessment process of Williams et al. [228] is similarly based on models' geometry due to the view dependence of the simplification technique. Some

geometric distance-based methods are widely used, although such simple geometric measurements like HD, PSNR and RMS error are purely geometry based, and often do not correlate well with subjective perception by the human visual system. More details of these methods on 5.2.

For example, all deformed models in Figure 2.4 have equal RMS, compared with the original model, but their visual characteristics range from excellent (top row) to poor (bottom row). As a result, some objective quality metrics have been developed, the goal of which is to generate a score that predicts the subjective visual quality (or visual impact of the distortion) of a distorted 3D model in comparison to a reference; these objective scores should be statistically consistent with those of human observers.

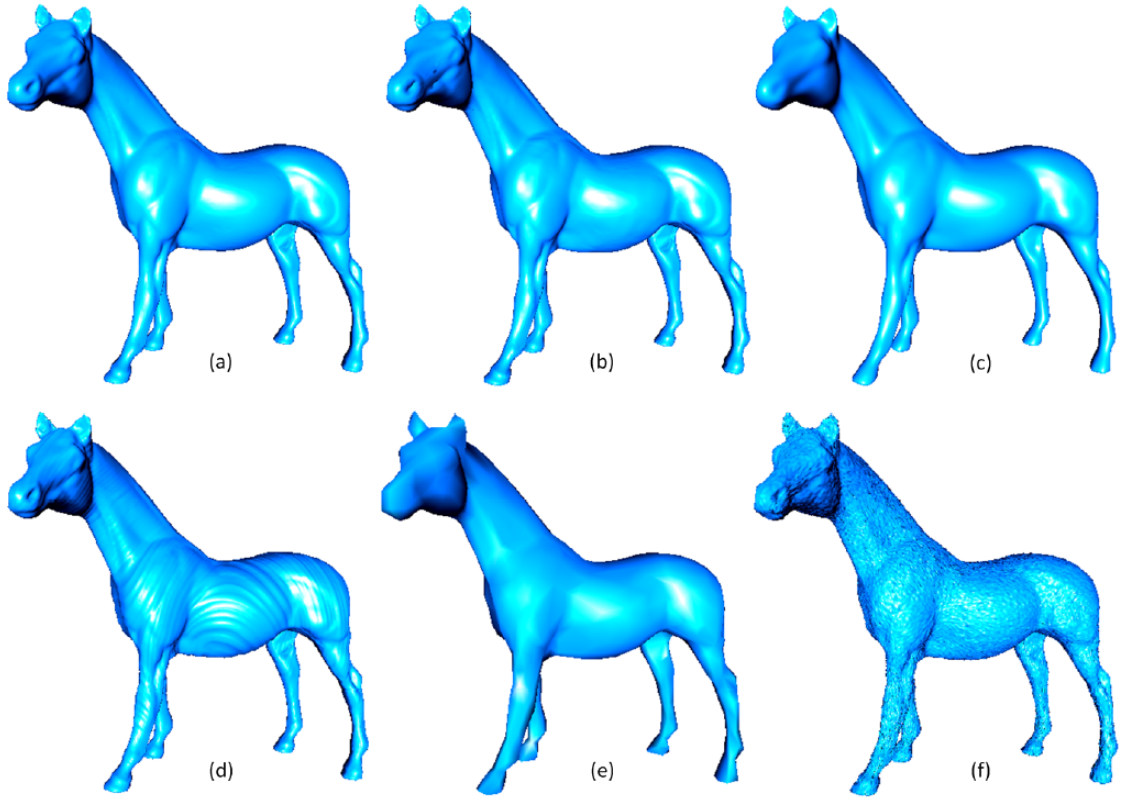


Figure 2.4: Distorted versions of the Horse model, all associated with the equal maximum Root Mean Square Error. (a) Original model. Results after (b) watermarking (MSDM2=0.14), (c) Laplacian smoothing (MSDM2=0.40), (d) watermarking (MSDM2=0.51), (e) simplification (MSDM2=0.62) and (f) Gaussian noise addition (MSDM2=0.84) [108].

2.7 Eye Movement

As we focus on mesh quality assessment, and we use an eye tracker to collect data for perceptual saliency, we summarise the most recent work on using an eye tracker in MQA. A user study experiment was conducted by Mantiuk et al. [134] with the use of an eye tracker, which tracked the human eye fixations on 3D meshes. They computed the correlation between human eye fixations and saliency maps in MQA which shows a good correlation between human eye perception and saliency. Wang et al. [215] conducted an eye-tracking experiment on 3D-printed objects and produced a 3D mesh dataset with fixation maps. They gathered gaze data from human observers using a monocular pupil eye tracker and extracted fixations from the pupil positions. This was the first study of visual attention on three-dimensional objects. Lavoue et al. [110] conducted an eye-tracking investigation on the rendered 3D shapes and developed fixation density maps for 3D meshes by mapping human eye fixations onto the 3D shapes. On the basis of acquired human eye fixations, they investigated the effect of shape, camera position, material, and luminance on visual attention.

There are many tools to predict 3D mesh quality. These techniques are useful for this study because they can be modified and adjusted to offer an understanding of human-computer interaction in different tasks (e.g., medical, business and VR contexts, etc.). The goal of this section is to examine a few key techniques regarding eye-tracking for IQA. Eye-tracking is becoming more frequently used in computer vision, human-computer interaction, and related fields. For individuals interested in employing eye-tracking to leverage eye movement data as an input mechanism to drive system interaction, this section provides a background on the fundamentals of eye-movement-tracking technology. We will examine some of the problems that need to be overcome to utilise the approach to analyse eye movement metrics in complex interactive systems and the future potential for eye movement tracking research.

Eye movement data can show mental attention in terms of the context of a visual dis-

play. Measuring various characteristics of eye movements, such as fixations (when eyes are still "encode" information), can also indicate periods when mental processing is taking place. In practice, for psychology and human-computer interaction (HCI) researchers to get useful information from eye-movement recordings, they have to define "areas of interest" over certain parts of a display or interface they are testing and then look at the eye movements that happen in those "areas of interest". So, specific interface elements can be evaluated objectively for their visibility, usefulness and placement, and the results can be used to improve the design of the interface [73]. For example, in a task where people are asked to locate an object, if they look at it for longer than expected before choosing it, this may not signify any important mental processing and may suggest the task needs to be redesigned.

Psychological research on eye tracking often provides indications and assessments about how people solve image problems, and how they think and imagine [19, 97, 240, 242]. Since eye movements indicate patterns of perceptual and cognitive processing, eye-movement analysis has great potential as a tool for usability research in human-computer interaction (HCI) and related fields, such as human factors and cognitive ergonomics. In HCI, the use of eye-movement analysis is in its early stages. Indeed, there is much work to be done in using eye movements to examine the attributes of websites that correlate with usability [46, 1].

In applied human factors research, eye trackers have been used to study and improve doctor performance in medical procedures [80], to evaluate cockpit controls to prevent errors by pilots [83], and to measure and improve situational awareness in air-traffic-control training [83, 113, 139]. Eye-tracking technology is also becoming increasingly popular in business research, to find out, for example, what types of advertisements get the most attention [129] and if internet users observe banner ads on websites [11]. In the following sections, we discuss various eye-tracking metrics and how to understand them in greater depth.

2.7.1 Eye Movement Metrics

In this section, we discuss how eye-tracking research uses fixations and saccades as its main measurements. These basic measurements provide a foundation for several derived metrics, including "gaze" and "scanpath" measurements.

Eye-tracking is the technique that eye movements are analysed to infer where an individual is directing at any particular time, and also to determine how the participant's eyes are moving from one point to another with reference to a visual image [165]. According to [167], the history of eye-tracking dates back to the 19th century, when initial techniques of tracking eye fixation and gaze emerged.

Significant improvements to the eye-tracking techniques were made in the 1970s, with corresponding advances in eye-tracking technology, and relations with a psychological theory that related eye-tracking data to cognitive processes [167]. Such pioneering work required enormous effort to measure eye movement (using a movie camera and cockpit-mounted mirrors) and assess eye movement data. The researchers made some important conclusions that remain useful today despite the challenges. For instance, the researchers suggest that fixation frequency can be used to assess a display's importance. In contrast, fixation duration can be applied to understanding the difficulty of extracting and interpreting information. The pattern of fixation transitions between displays can provide valuable information about the efficiency of the structure of specific display elements.

The technical problems and issues that are known to hold back research and eye-tracking techniques are increasingly being overcome. For instance, modern eye-tracking systems are relatively easy to use [167]. Commercially available eye trackers appropriate for usability research are based mainly on eye video images. In particular, [74] noted that most commercial eye trackers used today focus on measuring point-of-regard through a method known as "corneal-reflection/pupil-centre", as shown in Figure 2.5. These eye-tracking systems rely on a standard desktop computer with an infrared camera mounted next to or beneath a display monitor, using image processing software that seeks to locate

and identify eye features necessary for tracking [165]. Despite advances in eye-tracking techniques, there are still some limitations since modern eye-tracking systems have still proved unreliable in 10-20% of participants [74, 167].

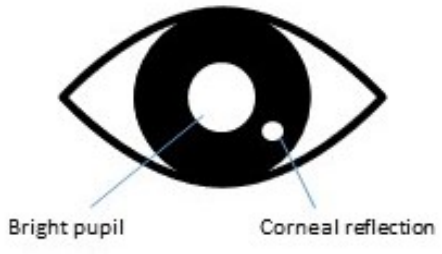


Figure 2.5: Corneal reflection and bright pupil as seen in the infrared camera image.

Despite some limitations, eye-tracking techniques remain useful in HCI research and researchers have argued for the primacy of such techniques in HCI research, promoting the "based on the eye-mind" hypothesis. This is the hypothesis that eye movement measurements can offer a dynamic trace that shows the focus of attention of an individual with respect to an observed visual display [165]. In practice, the use of eye-tracking data to make sense of human attention requires that the HCI researcher defines and establishes "areas of interest" from other areas of an interface or display and evaluates the movement of the eye with reference to such sites [165, 73]. Goldberg et al. [73] argue that through the techniques and methodologies of eye tracking, the researcher objectively analyses the meaningfulness, placement, and visibility of particular interface elements, which can then be used to enhance the interface design. Also, researchers have highlighted several key metrics important in eye-tracking research, in connection with fixation, saccades, scanpath, and gaze measurement [165].

Fixation refers to the moments when the eyes are relatively stationary in between movement, presumable when encoding or "taking in" information [133]. Normally, fixations last for an average period of 218 milliseconds, as noted in Poole et al. [165]. For instance, in tasks that involve encoding, such as browsing a website, a higher fixation frequency on a specific point is interpreted as an indication that the user has a greater interest in that target [165]. It can also imply that the target is quite complex such that

the user finds it difficult to encode the information [167]. However, by contrast, Jacob et al. [167] argue that such interpretations may warrant the opposite conclusion in a search task. For instance, clusters of fixations or a higher proportion of single fixations may indicate greater uncertainty in identifying a specific target point. Researchers also argue that fixation duration is associated with the processing time used to encode an image being fixated [133]. Importantly, it is widely believed that external representations linked to long fixations do not offer as meaningful information as those linked to short fixations [73].

Saccades, the quick eye movements between fixations, are also involved in key eye-tracking metrics. Although in [165] it is argued that since encoding does not occur during saccades and therefore these cannot offer information about the saliency or complexity of an image, a number of aspects of this metric are quite informative. For instance, regressive saccades, also known as back-tracking eye movements, have been identified as an important measure of processing encoding difficulty [170].

Gaze is described as a fixation-derived metric [165]. It is also known as the fixation cycle and dwell fixation cluster [83, 139]. According to Hauland et al. [83], gaze is often the total number of fixation durations found in a particular area defined by the researcher. It is widely applied in contrasting attention distributed between areas [139]. It has also been found useful in determining anticipation in context awareness, especially in situations where longer gazes appear on the target (image or shape) before a possible event [83, 139, 165].

Scanpaths refer to the overall saccade fixation and saccade sequence. In [73] it is argued that when using scanpaths in a search task, a straight line is the best scanpath with a relatively short fixation duration at the target. This metric often relies on quantitative analysis and uses many derived measures, including transition matrix, saccade fixation ratio, scanpath direction, scanpath regularity, spatial density and scanpath length [1, 73].

2.8 3D Mesh Saliency

The concept of saliency in computer vision has been well-documented and investigated by researchers. According to [94], saliency refers to the subjective perceptual quality that makes particular objects or images distinctive from others, making it possible for the eye to concentrate on those objects or images. Significant progress in understanding 2D image saliency has been made through several attempts to establish low-level features in images, including orientation, colour, and intensity, while employing these features in various spatial scales and considering the surrounding area. Itti et al. explain that it is possible to predict a viewer's eye focus by relying on low-level features [94]. However, semantic information would also grab the viewer's attention, such as when an image contains a text or person not necessarily highlighted by the low-level features. Significantly, some methods have been developed to detect saliency in a 3D model; however, evaluating the effectiveness of existing methods has often proved difficult. Although mesh saliency was exploited to guide mesh simplification, which showed its effectiveness for preserving interesting areas [94], early work often produced inconsistent results with human tracking data, which are largely subjective. This has prompted the development of new methods for measuring saliency, and new methods have emerged to improve earlier methods.

The first method to measure mesh saliency is that developed by Lee et al. [115], which is based on the idea that areas of an object that have different geometric characteristics as compared to adjacent areas would be salient because the eye tends to focus on phenomena or objects that are outstanding or are out of context in relation to the whole shape. For instance, a large spike appearing in the middle of a flat surface would attract significant attention from the viewer, while an equally flat section in an area full of large spikes would also attract attention. Lee et al. [115] explain that a Gaussian weighted average of the curvatures of vertices appearing in a radius could be used. The difference of Gaussians would then be aggregated together using a nonlinear normalisation to come up with the complete computed saliency.

Another method used to measure saliency is known as a multi-scale computational model, which is broadly similar to the method developed by Lee et al. [115]. However, it relies on spectral processing to detect saliency instead of the Gaussian weighted average used in Lee et al. [115]. Song et al. also developed a multi-scale computational model [192], which uses a set of meshes simplified to various degrees for multiscale analysis. It calculates the scale saliency map related to each scale by computing the spectral mesh saliency for every scale. Song et al. note that the scale saliency maps can be used to produce a final map [192]. It is possible to develop a more accurate model by learning a new method that combines several methods. Judd et al. [96] evaluated existing models based on eye-tracking data. A review of their methods showed that existing models were inconsistent with the human eye tracking data, which suggested they could add primitive methods to account for the inconsistencies.

2.9 Virtual Reality

As previously mentioned, in this thesis, we consider the perceptual quality of 3D meshes also in the VR setting, which can have a different impact on the subjective perception of quality, compared with the ordinary desktop setting. This will provide background for the subjective and objective quality studies in Chapters 3 and 5.

In recent years, VR and its applications have advanced rapidly with the advent of popular consumer Head Mounted Displays (HMDs), such as Meta/Oculus Rift/Quest, HTC Vive, and PlayStation VR. In 2022, VR application, gaming and video revenues have grown to more than five times the value of 2017 [211]. Owing to recent developments in headset technology, such as Meta Quest Pro, with faster and more reliable 5G wireless networks, it is anticipated that headset installations will also grow significantly in the coming years. There is a wide range of applications for VR in the consumer world, such as gaming, viewing 360-degree videos, and immersive education. Platforms such as YouTube, Facebook, and Netflix now facilitate viewing 360-degree images and videos

and provide various online tools, encouraging greater consumer engagement in VR.

VR imaging often uses a 360-degree camera equipped with multiple lenses that capture all 360 degrees of a scene. For instance, the Samsung Gear 360 VR Camera is a portable VR device for consumers with 180° dual lenses and a maximum image resolution of 5472×2736. The new Insta360 Titan is a professional 360-degree camera with eight 200-degree fisheye lenses capable of capturing 11K 2D and 3D (stereo) images. Since several lenses concurrently record images, they must be "stitched together" to create a complete spherical image. Typically, the spherical image is recorded in an equirectangular projection format for 2D 360° content, and in an over-under equirectangular format for 3D (stereo) 360° content. In addition to 360° videos, 3D shapes are also widely used for VR to model 3D objects and scenes. Once modelled, they can be rendered from arbitrary viewpoints to generate images for VR viewing.

VR is arguably a more immersive setting than normal 2D viewing. It contrasts with situations where individuals observe images and videos on flat-panel desktops and mobile devices. In VR, an image on the left may be positioned over the image on the right. Since the image may fill the whole viewing area, viewers are free to see it from any angle. Typically, just a small portion of the image is visible when the subject gazes an image from an oblique. Consequently, the information a user perceives depends on the spatial distribution of image content, the object being focused on, and the spatial distribution of visual attention. Accordingly, unrestricted viewing of high-resolution, immersive VR requires a substantial data volume, which causes difficulties in storing, transferring, and processing the visuals, which all affect viewing quality.

In view of the factors examined, it is crucial to study and predict the perceptual quality of VR images. Both subjective and objective quality methods are useful to comprehend and evaluate the visual quality of immersive VR environments. Subjective VR image quality assessment (VR-IQA) is a method in which human volunteers evaluate the quality of VR images. Based on gathered opinion ratings, it is possible to develop and evaluate prediction models. Most distortion types such as image compression artefacts, Gaussian

noise, and Gaussian blur, fail to capture the unique distortions of panoramic VR (2D and 3D) images. To further develop this area, recently, Chen et al. [35] have compiled a more comprehensive database containing regular image distortions and VR-specific stitching distortions. Also, they incorporate eye-tracking data collected during the subjective research.

When it comes to the quality of virtual reality (VR), there are several limitations that can impact the overall experience. The Field of View (FoV) in VR headsets is typically narrower compared to natural human vision. Limited FoV can create a sense of tunnel vision and restrict the immersive experience, affecting the perceived quality. High latency and motion blur can lead to a mismatch between head movements and the visual response in VR. This can cause discomfort, and disorientation, and reduce the perceived quality by introducing motion artefacts [196]. Due to bandwidth limitations and storage constraints, VR content may be compressed, resulting in artefacts such as blurring, or colour distortions. These compression artefacts can degrade the visual quality and impact the overall immersion. Also, VR headset lenses can introduce various types of distortion, such as chromatic aberration or geometric distortion. These optical issues can affect the clarity and visual fidelity of the VR experience. Some VR applications require substantial computational power to render complex and realistic environments in real-time. The limitations of current hardware can restrict the graphics quality, texture details, and overall visual quality in VR. The quality of VR content, including textures, lighting, and models, can significantly impact the overall visual experience [138]. Poorly optimised or low-quality content can lead to reduced realism and immersion. These limitations require continuous advancements in VR hardware, software, and content development. Improvements in display technology, resolution, FoV, and reduction of latency and artefacts are areas of active research and development to enhance the quality of virtual reality experiences. We will describe further issues regarding the subjective analysis of VR, towards examining the results of distortions in the VR setting in Chapter 3.

2.9.1 Virtual Reality Applications

We now review common virtual reality applications, most of which are based on omnidirectional images. One of the VR developments has been to include users in VR spaces. In particular, there have been efforts to generate individualised 3D models of human beings to improve image processing in a virtual environment [50]. One of the VR applications uses saliency maps for omnidirectional images. It demonstrated that saliency maps for omnidirectional images (ODIs) can be observed through HMDs without an eye-tracking device. In [50] viewport centre trajectories (VCTs) were collected, and a way to make saliency maps from the data collected was suggested. Images are also used to compare the saliency maps. Then, because ODIs tend to be biased toward the equator, the author suggests a post-processing method called fused saliency maps (FSM) to make current saliency models fit the needs of ODIs.

Another application, discussed in [31], compares VR viewing and viewing on a 2D desktop. The evaluation gives an in-depth critique of how visuals are fundamental for immersive, real-time rendered VR. The analysis of the results of both studies shows that slowing down navigation reduces the effect of depth cues on visual salience and increases the effect based only on 2D image features. Even though scores vary depending on content, it is clear that saliency prediction methods based on boundary connectivity and surroundedness are effective in most settings.

Another application of 3D imagining is for human body model generation, which involves using 3D body scanners to produce a virtual 3D model of an original person [15]. This technique has been considered effective in capturing data with high precision and resolution, although it is associated with some challenges. For instance, the process can be time-consuming, especially when post-processing the raw data. Also, it is usually difficult to reconstruct the human body and to ascertain its position in controlling the movement of images and objects in a virtual environment [91]. The limitation of such techniques is normally evident for instance in computer game contexts, where a gamer is

asked to respond in real-time to stimuli in a virtual space [91]. The movement of gestures and body parts is recognised as generally immersive for a gamer and, with the right tools, can lead to a satisfactory immersive experience. Nevertheless, improvements made in this area are evident in the increased application of 3D cameras to virtual environments.

One special technique in this domain is the gesture recognition interface (GRI), which is a type of perceptual computing user interface that allows computers to capture and interpret human gestures as commands. This technique has emerged in recent years as a particularly important development in the gaming industry [91]. The 3D cameras have been applied in virtual reality environments in two distinct ways with respect to the GRI. One of the applications involves using fingers and the hand to prompt specific commands, such as moving images on the screen, including driving and shooting. The second application of 3D cameras involves complete body immersion into the virtual environment [91].

Based on the technology of 3D cameras for GRI, one study [91] has developed a technique for detecting saliency in a virtual environment using 3D cameras. The technique uses both the actual data generated by the camera and the physical constraints of the VR environment to ensure the user is completely immersed in the virtual environment. Drawing from the work of Yao et al. [235] regarding how to use anthropomorphic measurements to detect 3D images in a virtual environment, [91] explains that the technique uses an image processing algorithm that can detect the main parts of the body of a user. Such image processing, using anthropomorphic measurements of different body parts, involves a connection between human body parts and the corresponding VR points. Although experimental studies are required to explore the effectiveness of this technique, it is considered appropriate for enhancing the immersive experiences of users in a VR environment.

Some studies using VR can be conducted on gaze and head orientation [180]. For example, collected experimental data relating to the gaze and head orientation of individuals. The study allowed observers to examine omnidirectional stereo panoramas in VR

environments, both in seated and standing positions. They also obtained data from users looking at the same images in a desktop scenario, experiencing monoscopic panoramas through mouse-based interaction.

The researchers made essential conclusions that inspire this work. Firstly, they concluded that gaze data and saliency in a virtual environment with traditional displays, indicating that existing saliency predictors apply to VR after making a few simple modifications. This assertion is consistent with the views of Monroy et al. [144], which noted that saliency prediction techniques initially used to detect traditional 2D images are not directly applicable to omnidirectional images (ODIs) because of the heavy distortions that exist in VR due to projection, and observed biases as a significant feature. De Abreu et al. [50] also showed that most saliency techniques for conventional 2D images could only enhance the performance when detecting saliency for omnidirectional images if modified by eliminating the centre bias common in most cases.

Secondly, Sitzmann et al. [180] note that gaze and head interaction are connected to VR viewing conditions. They demonstrate that head orientation measured using inertial sensors may be adequate in predicting saliency with significant accuracy without using expensive eye tracking.

Thirdly, the researchers showed that time-dependent viewing behaviour could be accurately measured shortly after the user is exposed to a new image. However, owing to great inter-user variance, such data cannot be captured for more extended periods.

Fourthly, they also note that the fewer salient regions in a VR image, the faster user attention is focused on a particular region and the more attention is concentrated.

Lastly, the researchers note that there were two view modes among the users – reorientation and attention, which are distinguishable through gaze or head movement in real-time and thus provide important insights into interactive applications [180]. Significantly, Startsev et al. [198] demonstrate that regular image saliency models can be applied in VR environments. They show that traditional image saliency predictors could perform

saliency prediction for panoramic 360° scenes. However, panoramic 360° scenes are typically represented using equirectangular images, leading to some problems that create image distortion [198]. Monroy et al. [144] show that omnidirectional saliency prediction may be achieved by separating an omnidirectional image into patches before adding a saliency refinement architecture that considers spherical coordinates to a convolutional neural network (CNN). Although this is a new study requiring further validation, the conclusions could have a wide range of influences on VR tasks as it extends the traditional image detection approaches identified in [93].

2.10 Summary

In this chapter, we explored the fundamental background of visual quality evaluation. To understand the associated problems comprehensively, we looked at the current research on visual quality evaluation in 2D images, 3D meshes, and omnidirectional images in a VR setting. We also explain several commonly used 2D image and 3D model saliency prediction techniques. These approaches are related to the work for VR setting in Chapters 3 and 5, as well as 3D mesh saliency reported in Chapter 4. Methods for predicting media quality were also examined, such as eye movement, fixation and gaze. We emphasised the importance of examining the current literature that employs the collective prediction method. We can now proceed to the next chapter and explore the subjective study of 3D mesh quality scores in virtual reality.

Chapter 3

Subjective Study of 3D Mesh Quality Scores in Virtual Reality

Overview

As previously mentioned, although 3D mesh quality has been extensively researched, there is limited work that measures the 3D mesh quality in a VR setting. This is an important topic for downstream applications such as the Metaverse because massive amounts of data are necessary to support AR and VR experiences. The chapter is organised as follows. After introducing the problem in Section 3.1, we show the details of the subjective experiment in Section 3.2. Sections 3.3 and 3.4 give the details of our findings. Finally, Section 3.5 contains the conclusion and future work.

3.1 Introduction

Recent advances in 3D mesh modelling, representation, and rendering have progressed to the point that they are now extensively employed in many applications, such as networked 3D gaming, 3D virtual reality, augmented reality and immersive worlds, and 3D visualisation. With VR, users can experience high-quality, photo-realistic images and immersive virtual environments in real-time using the latest advancements in computer graphics hardware and software [65]. Increasing the visual quality of a mesh by using a large number of vertices and faces provides a more detailed representation. However,

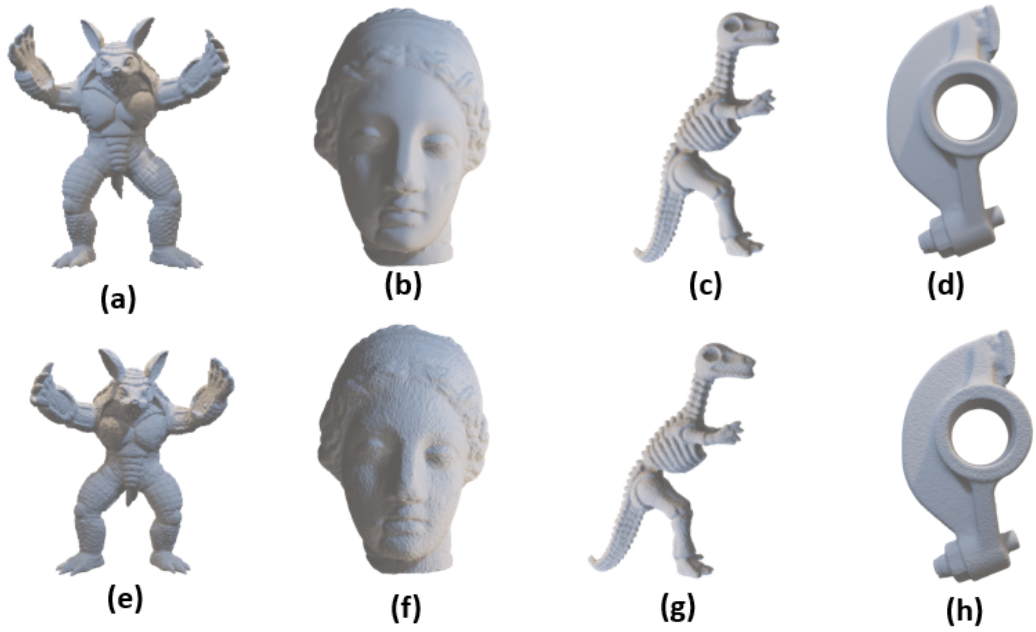


Figure 3.1: Examples of 3D meshes belonging to the LIRIS/EPFL General-Purpose database. The top row shows the 4 reference meshes. The second row presents 4 distorted 3D meshes: e) 3D mesh Armadillo affected with noise on rough regions, (f) 3D mesh Venus affected uniformly with noise, (g) 3D mesh Dinosaur uniformly smoothed, (h) 3D mesh Rocker-Arm affected with noise on smooth regions.

the added complexity leads to increased requirements for data storage, processing power (CPU and GPU), and network bandwidth, especially for real-time applications and when data needs to be transmitted over the network. As a result, a trade-off between graphical model visual quality and processing time frequently arises, necessitating determining the quality of 3D graphic resources.

Several geometric modifications may be applied to 3D mesh models like compression, simplification, and watermarking. These processing procedures may influence the appearance and visual quality of the 3D models (see Figure 3.1) and, as a result, the quality of the user experience (QoE). Thus, subjective quality evaluation tests are essential for evaluating visual quality as perceived by human observers. Subjective methods involve a group of human participants being asked to rate the quality of a collection of 3D meshes that have been subjected to different types and levels of distortion. The output of the subjective method is a set of mean opinion scores (MOS), which enables predictive models to be developed and evaluated, taking subjective scores as ground truth. Some subjective

methods can be used in 2D image and 3D graphical areas, for example, single stimulus, double stimulus, subjective assessment methodology for video quality, and pairwise comparison [39, 183, 17, 207].

Nevertheless, choosing the appropriate subjective technique is not easy since we must verify that such methods produce accurate and reliable findings. In our case, we used a pairwise comparison of 3D meshes, where participants were asked to rate a collection of different levels of 3D mesh distortion in terms of visual quality, compared with the reference undistorted shape presented to the user along with the distorted shape. Pairwise comparisons are simpler and more intuitive for users, ensuring that users can concentrate on judging the quality of the distorted shape compared to the given reference shape.

We propose to measure how different distortions (noise, smoothing, etc.) of 3D shapes affect the perceptual quality of 3D objects in a virtual reality environment by collecting subjective scores for distorted shapes. We compare the MOS between virtual reality (VR) and traditional desktop display settings. Moreover, we analyse different 3D mesh distortion types with different 3D shapes to determine which distortion type/shape shows significant results. To compare VR and normal desktop settings more easily, we used an existing database evaluated on the traditional desktop display. As VR is becoming a popular way of consuming and visualising 3D content with high resolutions, we build an application to carry out VR experiments using a Meta/Oculus Quest 2 headset.

Previous subjective tests in the field of computer graphics were conducted to evaluate the visual quality of static and animated 3D models [44, 76, 112]. As shown in most papers, there is no agreement on the appropriate approach to assessing the quality of 3D models [153]. As we focus on the human visual system (HVS) strongly linked with perceptual quality measures, we concentrate on perceived 3D mesh perceptual quality measures using a VR headset, not purely geometric measurements that ignore human perception. Bulbul et al. [27], Lavoué and Mantiuk [112] and Muzahid et al. [147], are mostly working in the perceptual area and provide reviews of more broad 3D visual quality evaluation techniques.

3.2 Subjective Experiment

In our experiment, we use a public database to evaluate the 3D mesh quality level of distortion using a virtual reality headset (HMD). Figure 3.2 presents the LRIS/EPEL general-purpose database, which contains four reference models (Armadillo, Venus, Dyno and Rocker Arm) and 84 distorted meshes (21 distorted meshes for each reference mesh) [111]. This dataset used two different types of degradation, noise and Taubin smoothing [206], to simulate typical degradation of mesh quality due to e.g., compression and watermarking [111]. These distortions have different levels of strength and four types of locations on meshes: uniformly on the whole mesh, smooth areas, rough areas, and intermediate areas, where different areas are identified based on local curvature variations. Note that Taubin smoothing is not applied to smooth areas as the effect is hard to notice.

For noise addition, subjective quality scores are provided for each distorted mesh in the form of MOS, ranging from 0 (worst quality) to 10 (best quality). Lavoué et al. [111] created noise by altering the coordinates of the mesh's vertices with a randomly calculated offset between 0 and the specified maximum deviation. Smoothing was accomplished by applying Taubin [206] smoothing filter to the mesh's vertices.

These distortions were applied at three distinct intensities (visually selected): high, medium, and low (these levels correlate to the number of smoothing iterations and the maximum deviation value for noise addition). Finally, these distortions were applied in different locations on the meshes: evenly (across the whole object), only to smooth areas, rough areas, and intermediate areas. Each model generated 21 degraded versions: three noise strengths in four types of locations and three Taubin smoothing strengths in three types of locations (i.e., excluding smooth areas).

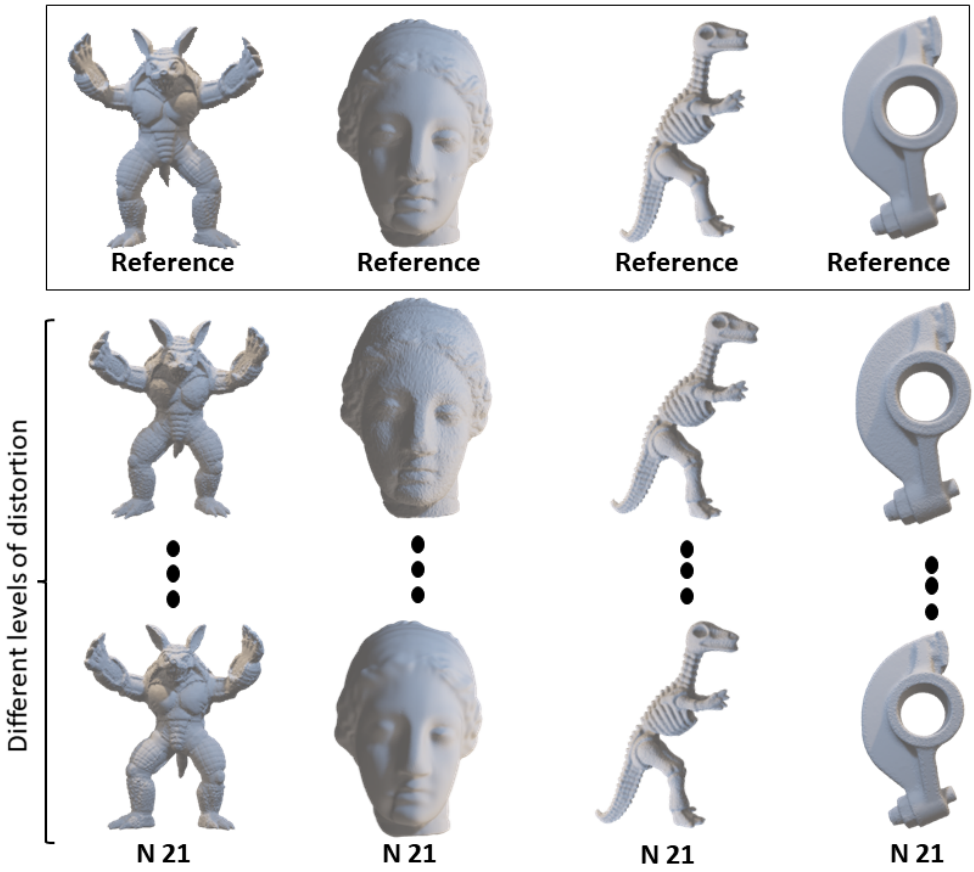


Figure 3.2: Examples of 3D meshes belonging to the LIRIS/EPFL General-Purpose database. The top row shows the four reference meshes (from left to right: Armadillo, Venus, Dyno and Rocker-Arm) and the rest shows that we have 21 models with different types/levels of distortion for each shape (only selected ones are shown), so the total number of shapes is 88.

In this work, we compare VR and desktop settings. In the desktop setting experiment, Lavoué et al. [111] used a 1280 $\times$ 720 resolution monitor, and each model was displayed in a 600 $\times$ 600-pixel window. All models have a resolution of between 50,000 and 100,000 triangles, as illustrated in Table 3.1, so the details of the model can be viewed well. They used rotation, interaction and zoom operations to allow the participant to interact with the model (e.g. mouse clicks) in their experiment. Also, the paper [111] showed participants the models (both reference and distorted versions) in a desktop display setting. The participants were allowed to browse through shapes so that they could memorise the worst/best quality shapes.

Our study uses a VR setting that does not allow participants to see all the models simultaneously. However, we show participants a trailer with a different dataset to make

sure the participant has an idea of how the experiment will be. These models were obtained from different sources which used different scanners. For example, the Armadillo model is a manifold/simplified version of the original model that was created from scanning data by the Stanford Computer Graphics Laboratory. The Dinosaur model is courtesy of Cyberware Inc. The Venus and Rocker Arm models are courtesy of the AIM@SHAPE project. Our subjective study was conducted using a pairwise comparison (PC) method in a virtual reality setting.

Table 3.1: The geometric information of the reference meshes in the LIRIS/EPFL general-purpose database

Mesh	Number of Vertices	Number of Faces	Number of Edges
Armadillo	40002	80000	120000
Venus	49666	99328	148992
Dinosaur	42146	84288	126432
Rocker-Arm	40177	80354	120531

3.2.1 Evaluation Methodology

In our experiment, we followed the same methodology used in the desktop setting as Lavoué et al. [111]. We used a pairwise comparison approach to evaluate the quality of distorted meshes with respect to the reference (undistorted) 3D mesh. More precisely, throughout the experiment, each participant was provided with a distorted version to compare against the reference version. Then, each participant was asked to measure the quality of the distorted version compared with the reference using a slider bar (0 worst quality, 10 best quality). Pairwise comparison is simpler to perform and requires less mental effort from participants.

This experiment shows a new approach to using a VR setting to compare how different platforms (VR versus a traditional desktop display) affect perceived mesh quality, which has not been done before. We will provide a new comparison approach between VR and desktop settings to simulate real environments and identify similarities and dissimilarities between human perception in these settings.

3.2.2 Stimuli Generation

In the many previous subjective research experiments incorporating 3D material, the 3D models were displayed to the viewers using a variety of approaches, including still images, free interactivity, and animations. Still images are insufficient to assess the visual quality of 3D models, as demonstrated by Rogowitz et al [173]. As a result, the object must move for the observer to notice the dynamic impacts of shading on the shape. It is also critical for the observer to perceive the entire object rather than focusing on a particular point of view. Participants could notice small areas not visible in 2D views due to the model's free rotation. To generate the mean opinion score (MOS), we used the same stimuli generated by [173].

3.2.3 Display

In our experiment, the display technology consists of a Meta/Oculus Quest 2 HMD with Qualcomm Snapdragon XR2 Platform, a single Fast-Switch LCD 1832 $\times$ 1920 pixels per eye with refresh rate 72Hz and tracking inside and outside 6 DOF (degrees of freedom).

The experiment was built as an application in Unity3D and rendered with a resolution of 1832 $\times$ 1920, as shown in Figure 3.3. The experiment is based on the pairwise comparison method; the participants rate the quality between two models where one is the (undistorted) reference, and the other is a distorted version. Participants were allowed to explore the 3D mesh object by using touch controllers. In the experiment, the participant's head's position and rotation are used in the Unity3D application to provide a first-person perspective to explore the object [197].

Since depth perception could also play a significant role in the selection, we presented the objects equidistantly at 20cm distance from the participants' eyes, and they could freely move closer and further from the objects to explore them. In Figure 3.4 shows how the experiment looks like in the VR environment using first person.

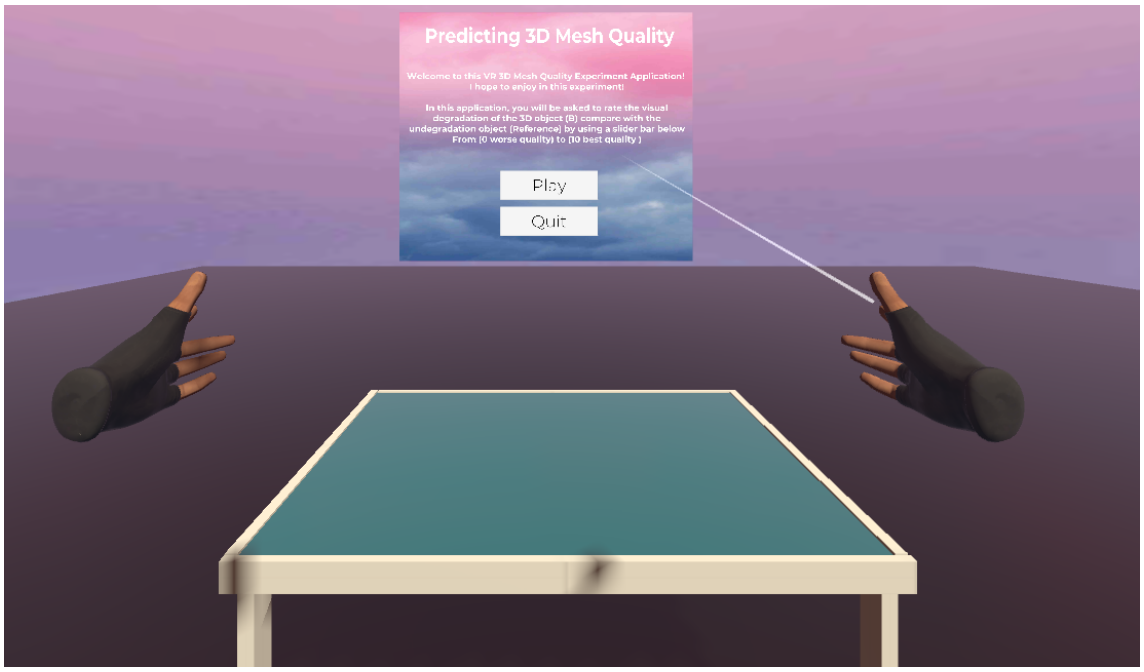


Figure 3.3: Example of the Home page showing our experimental environment of the subjective test.

3.2.4 Participants and Training

Participants in our experiment participated voluntarily without being awarded monetary or other rewards. A consent form was given to the participants before the experiment began. They were also informed that they could leave the experiment at any time and that they were not required to complete it. The names and VR data of the participants were kept anonymous. The VR device we used to collect their MOS data was used during their participation. Before we started the experiment, we began with a trial session, as recommended by the ITU-R BT.500 [179], such that participants become acquainted with the virtual environment and task, to ensure they fully comprehend the experiment's task. This stimulus outcome was not recorded. The main reason for this trailer is to enable participants who are not familiar with VR devices to learn how to rotate, scale and transform 3D objects.

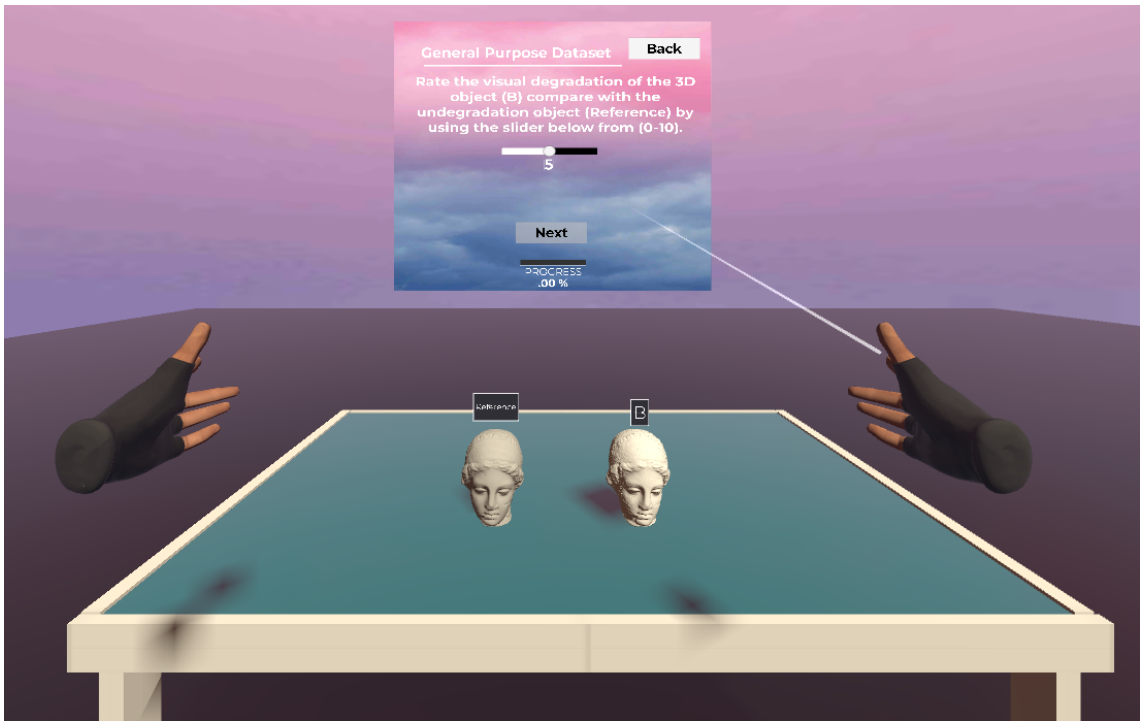


Figure 3.4: Example of the main page where the user can interact with the models in the experiment environment. The left side contains the reference model, and the distorted model is on the right side (B).

The slider bar provides subjective quality rankings for distorted mesh in the form of MOS, rating from 0 (worst quality) to 10 (highest quality) as illustrated in Figure 3.4. At the end of the stage, we introduced the experiment application in which the participant was shown two objects, one undistorted (Reference) compared with the distorted shape and asked to judge their quality based on the undistorted mesh quality. This step was intended to allow the observer to become acquainted with the experiment, to focus adequately, and to ensure that observers fully comprehend the experiment's task. Figure 3.4 shows the experiment environment.

The experiment was conducted at Cardiff University and involved students aged between 20 and 40, with twenty females and thirty males. All participants reported normal or corrected-to-normal vision.

3.2.5 Procedure

The experiment used a Meta/Oculus Quest 2 headset and asked participants to rate the quality of the 3D distorted models. The experiment showed all (84) 3D scanned meshes from the LIRIS General-purpose database and distorted versions. In the experiment, we did not specify the duration of time to complete it as some of the participants were not familiar with the VR setting so we did as the participants enjoyed and finished it with no concern for the duration. However, we mentioned in the trailer that the experiment should be no more than 30 minutes to avoid sickness symptoms. The participants took an average of 22-27 minutes to view the models. The experiment was carried out through a computer application that presents the 3D objects on the VR headset in random order, with paired 3D objects appearing side by side. At each comparison, participants compare the reference model with the distorted model by using a VR touch controller to scroll the slider bar score as illustrated in Figure 3.4. This way allows us to collect the MOS scores and analyse correlations of individual distortion types by using Pearson and Spearman coefficient correlation.

3.2.6 Duration

The overall length of the experiment affects the efficiency of the experimental method, especially in virtual reality where most of the subjects have not used the VR headset before and tend to exhibit symptoms of cybersickness both during and after the virtual environment experience [106]. To avoid these issues, we chose to display the reference and the test stimulus simultaneously side by side in the same scene. In this way, the number of presentations is halved. To avoid fatigue, boredom and cybersickness, we allow the participant to move around the lab or sit in a chair to reduce any motion sickness. Each subject's session took place on a single day in order to prevent any learning effect between stimuli. The stimuli were displayed in a random order (i.e reference models, distortion types and distortion levels) to each participant. Each stimulus was presented once; the

participant was not able to replay the scene.

3.2.7 Experiment Design

The design choices used in an experiment are of great importance because they can bias the results significantly, especially for computer-generated stimuli, where almost every element can be controlled. Effective parameters were controlled in several quality assessment studies so we chose these elements (lighting, background, stimuli order and duration). We describe the elements below.

Lighting is the most significant element in controlling the environment. The position of the 3D object and the type of light sources are critical factors that have a significant impact on the viewing conditions. According to Rogowitz et al. [173], models lit from the front provide different subjective scores than those lit from above. Light is fixed and shines from a left-above direction. As we used the Unity game engine there are different types of lighting such as directional lighting, spotlight, point light and area light. In our experiment, we chose a spotlight which is useful for creating focused lighting effects and highlighting specific objects to give a clearer perception of 3D shapes, according to the human visual system [175].

Background colour can influence perceived quality by altering the visibility of the model's borders. While some user studies [224, 175] utilise a uniform black environment, Corsini et al. [44] use a nonuniform background that transitions from blue to white to avoid overestimation of contours. However, in our case, as our models have no colour we used dark green to distinguish between the scene background and the table colour. Also, it let the participiant see the shadow of the model on the table to know how far/close the object is to the table, as shown in Figure 3.4.

Stimuli order considers whether the stimuli (reference and distorted meshes) should be displayed to the user simultaneously as the same protocol in the desktop setting (e.g., side by side) or in sequence in comparison-based tests (e.g., first the reference, then the

tested models). When displayed in order, users are allowed to return to the reference model, as in Rogowitz and Rushmeier's experiment [173] to provide a more comprehensive comparison. Furthermore, the stimuli's arrangement and placement should be chosen so that external factors such as observer movement and ambient light have the least impact. Taking these into account, we showed stimuli side-by-side in our experiments to make comparison easier, especially as meshes can be interactively rotated to allow the complete shapes to be seen.

Duration of which the tested models are shown to the subjects may also affect the evaluation results. The average duration time is around 20 to 25 minutes per participant. We chose a duration such that users were given enough time to examine the shapes, while avoiding taking an excessive amount of time during the user evaluation which may lead to the degradation of user response data due to fatigue.

3.2.8 Ethical Approval

The Human Ethics Committee of the Cardiff University SREC reference: COMSC/ Ethics/ 2021/ 089 has authorised this study. COVID-19 regulations (University guidelines, Welsh and UK government policies) require us to examine how to conduct research securely using a head-mounted display, especially regarding participant safety. Additionally, hand sanitiser, anti-bacterial wipes, hand sanitiser, and disposable masks were provided along with the COVID-19 screening form. As a result of these extra procedures, the user study took longer to complete, but they were a reasonable solution.

3.3 Data Analysis

The following sections analyse and discuss the results of the experiment described above.

3.3.1 Screening Participants and Computing Mean Opinion Scores (MOS)

We follow the ITU-R BT.500-13 recommendation [179], where we show a trailer with a different dataset to the participants to make sure they understand how the experiment works. Once we finished the experiment, we could collect the MOS, but before performing any data analysis, we tested the participants' performance to ensure the collected data was meaningful.

To compute the Interquartile Range (IQR) [179] of our data, we first need to identify outliers. We identify the first quartile (Q1), the median, and the third quartile (Q3). So we calculate IQR = Q3 – Q1. Calculate upper value = Q3 + $1.5\times$ IQR. Finally, calculate lower value = Q1 – $1.5\times$ IQR as shown in Figure 3.5. One outlying participant was found in both settings and was rejected from the dataset.

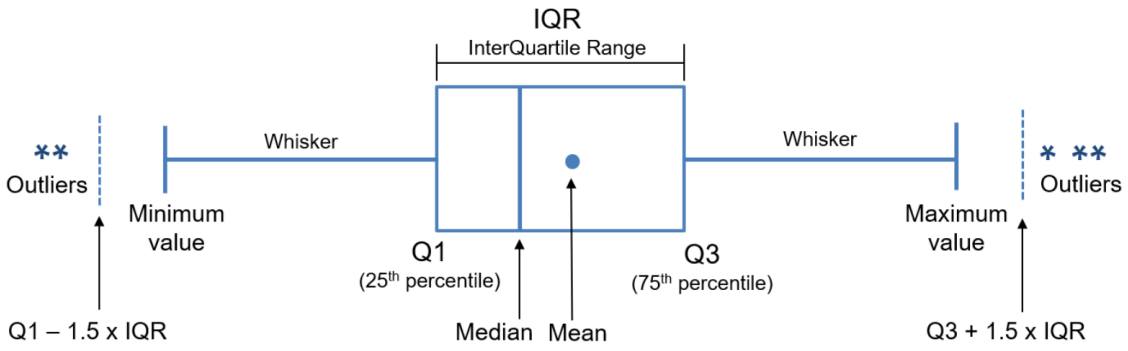


Figure 3.5: Illustration for Interquartile Range (IQR)

To analyse user ratings, a common method is to compute the mean opinion score (MOS) for each stimulus.

$$MOS_e = \frac{1}{10 \times N} \sum_{i=1}^{N} s_{ie}, \tag{3.1}$$

where s_{ie} refers to the score assigned by participant i to the stimulus e, and N denotes the number of (valid) subjects. We further divided the scores by 10 to normalize them in the range of $[0, 1]$.

We follow most of the existing work [10, 35] and set the scores such that 0 means the worst quality, and 10 is the best quality. So we expect the MOS to decrease as the distortion level increases. In Armadillo, we notice a strong consistency between the VR setting and traditional desktop display setting, as the participants in both settings showed almost the same behaviour for each type of distortion. However, for the rest of the models, we may see some disparities in the rating scores of the two settings. In fact, in some cases, desktop participants' scores are not consistent with the stimuli, i.e., the quality does not always drop when the level of distortion increases (e.g. for the Venus model), but VR viewers give scores more consistent with the distortion levels.

Furthermore, we found that VR observers were able to detect distortions that desktop observers missed. These initial findings suggest some discrepancies in the human perception w.r.t. different display techniques. In the next section, we will examine whether these differences are statistically significant and seek to explain their origins.

3.3.2 Observer Agreement Analysis and Correlation Analysis

We calculated the standard deviation σ values of the MOS scores from both the VR and desktop settings. In the VR setting $\sigma_{VR} = 0.0220$ whereas in the desktop setting $\sigma_D = 0.0278$. Since the standard deviation in the VR setting σ_{VR} is lower than that of the desktop setting σ_D, it shows that the user ratings are more consistent for the VR setting.

In order to further analyse the similarity and dissimilarity between subjective mesh quality, we look at the correlations between subject evaluations. To begin with, we examine the correlation between the VR setting and traditional desktop display setting for distortions of each 3D object of the dataset.

We also check the correlation for each type of distortion. In this chapter, Pearson and Spearman statistics are commonly used in statistical analysis to assess relationships between variables. The following two measures are used to measure the correlation between virtual reality and desktop display settings. The Pearson linear correlation coefficient (PLCC or r_p) measures the prediction accuracy of MOS, while the Spearman rank-order correlation coefficient (SROCC or r_s) measures the prediction monotonicity [221]. Both values of PLCC and SROCC range from -1 to 1, where 1 indicates a total positive correlation, -1 indicates a total negative correlation, and 0 indicates no correlation.

Suppose in our case we have two (VR & desktop) settings $x = \{x_1, x_2, \ldots, x_n\}$ and $y = \{y_1, y_2, \ldots, y_n\}$, both containing n stimuli. The Pearson linear correlation coefficient r_p between settings x and y is calculated as follows.

$$r_p = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{3.2}$$

where

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \text{ and } \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{3.3}$$

MOS scores x and y are sorted in the same order, in either ascending or descending order.

Let X_i be the rank of x_i in x, while Y_i be the rank of y_i in y. We generate two new sequences $X = \{X_1, X_2, \ldots, X_n\}$ and $Y = \{Y_1, Y_2, \ldots, Y_n\}$. Let $d_i = X_i - Y_i$, the Spearman rank-order correlation coefficient r_s between settings x and y is calculated as follows.

$$r_s = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}. \tag{3.4}$$

3.4 Results

In this section, we divide our results into two parts, distortion by shapes (Armadillo, Venus, Rocker Arm and Dyno) as illustrated in Figure 3.8 and distortion by types/locations as illustrated in Figure 3.10 (Noise Uniform, Noise Rough, Noise Intermediate, Noise Smooth, Taubin Uniform , Taubin Rough and Taubin Intermediate).

3.4.1 Distortion by Shape

We now analyse MOS scores between VR and desktop settings on the basis of individual test shapes. Since the two modes of display have their own characteristics, as illustrated in Figure 3.8, the perceived quality is more consistent on some shapes than others. The x-axis corresponds to different distortion types, locations and strengths, and the y-axis shows the (normalised) MOS scores.

We can see that Armadillo has the highest correlations in both Pearson linear coefficient correlation (PLCC) and Spearman coefficient correlation (SROCC) ($r_p = 0.754$, $r_s = 0.685$) compared with the rest of the models. Armadillo contains some details, which means it is easy for participants to notice the different quality between the reference model and the distorted version as shown in Figure 3.6.

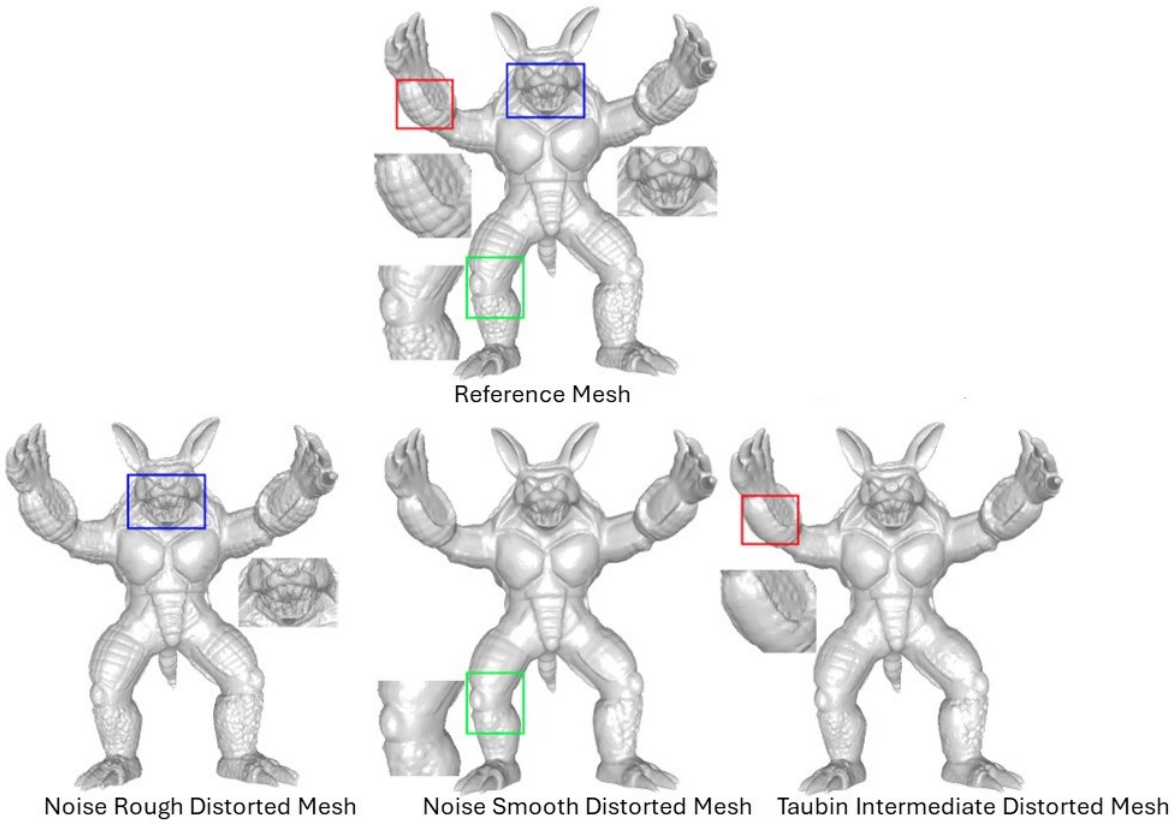


Figure 3.6: The reference mesh, three distorted Armadillo model meshes, and the enlarged views of some representative distorted regions on the meshes as marked in the rectangles.

In comparison, the Venus model has a lower linear relationship and a higher nonlinear relationship ($r_p = 0.698$, $r_s = 0.701$) because a participant in the VR setting is easy to zoom and rotate the model compared with the desktop, which helps the participant to notice small areas that might not appear well on the traditional desktop.

The third model is Rocker Arm, which shows a fair relationship between VR setting and traditional desktop setting ($r_p = 0.623$, $r_s = 0.513$). In both settings, participants often did not notice if there was a distortion in the shape because the shape did not have much details they could detect. The last model is Dyno which has the worst result compared with the rest of the models. The linear coefficient correlation is better than the nonlinear coefficient correlation ($r_p = 0.624$, $r_s = 0.499$). The reason behind the worst correlation is that the Dyno model has less detail area which makes it hard to detect the quality even if the reference is available as shown in Figure 3.7. Note the Figures 3.6 and

3.7 show according enlarged parts in reference model for comparison.

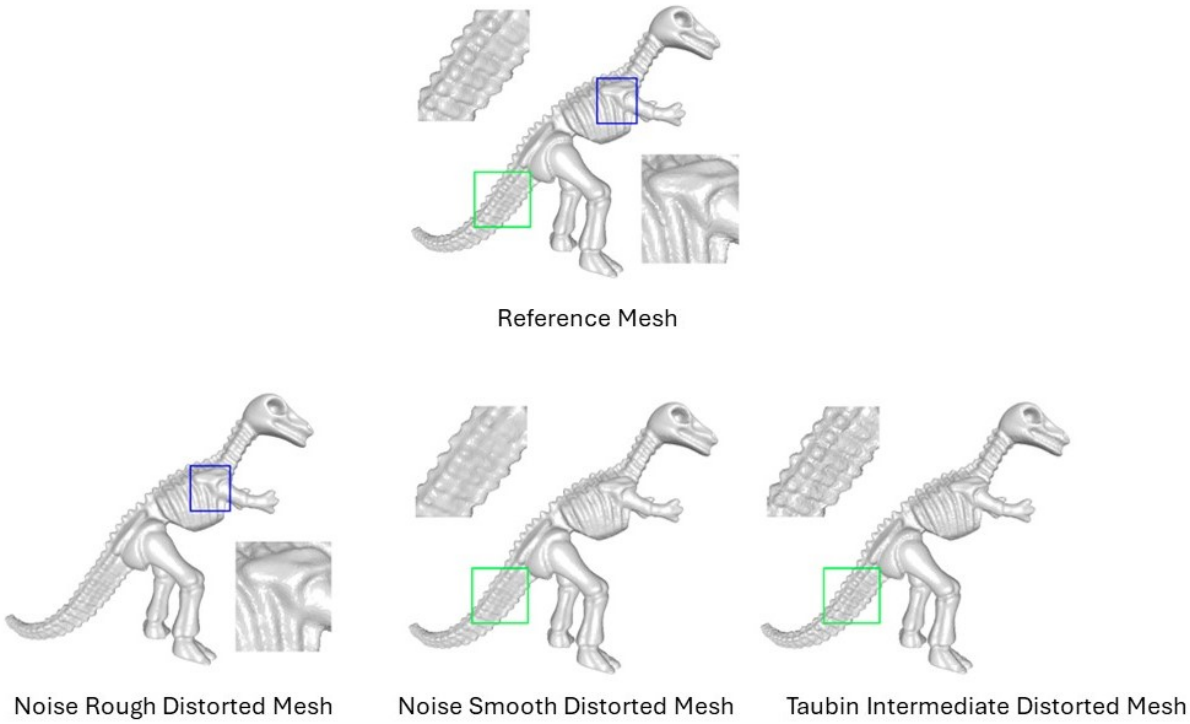
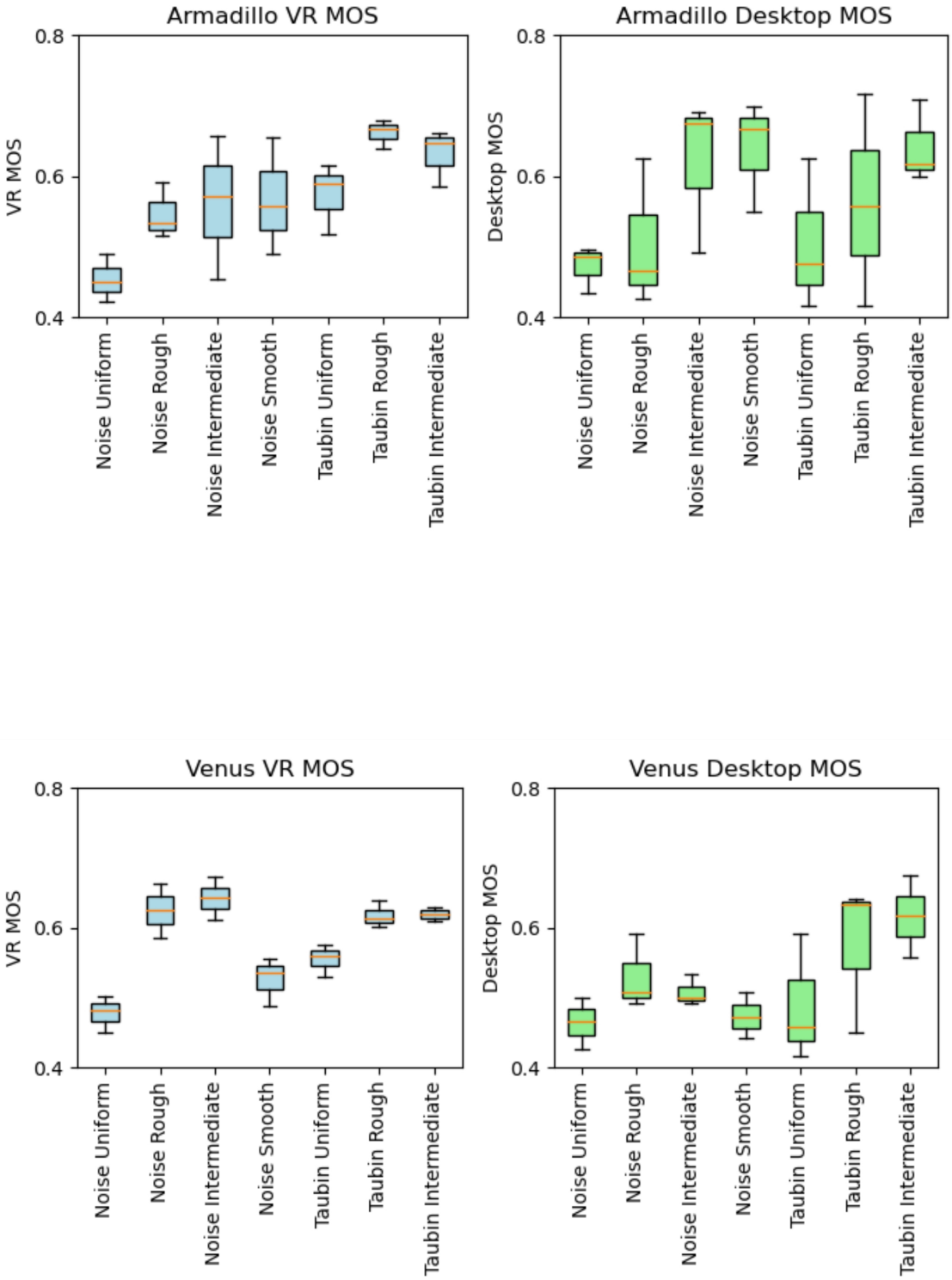


Figure 3.7: The reference mesh, three distorted Dyno model meshes, and the enlarged views of some representative distorted regions on the meshes as marked in the rectangles.

After calculating PLCC and SROCC for individual distortion types, we further compute PLCC and SROCC for all the distortions for each shape, using the same formulae, as indicated in the 'All' row in Table 3.2.

Table 3.2: Pearson and Spearman correlation analysis comparing VR and desktop MOS scores for different stimuli (the distortion type followed by distortion location.

Distortion Type and Location	Armadillo		Venus		Dyno		Rocker Arm	
	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC
Noise Uniform	0.850	1	0.934	1	1	1	0.549	0.500
Noise Rough	0.993	1	0.882	1	0.657	0.500	0.093	0
Noise Intermediate	0.925	1	0.828	0.500	-0.348	0.500	0.729	0.500
Smooth	0.667	0.500	0.981	1	0.667	0.866	0.995	1
Taubin Uniform	0.564	0.500	0.686	0.500	0.975	1	0.183	0.500
Taubin Rough	0.873	1	0.269	0.500	0.500	0.500	-0.660	-0.500
Taubin Intermediate	0.538	0.500	0.995	1	0.829	1	-0.868	-1
All	0.754	0.685	0.698	0.701	0.624	0.499	0.623	0.513


Armadillo VR MOS
VR MOS
0.8
0.6
0.4
Noise Uniform
Noise Rough
Noise Intermediate
Noise Smooth
Taubin Uniform
Taubin Rough
Taubin Intermediate
Armadillo Desktop MOS
Desktop MOS
0.8
0.6
0.4
Noise Uniform
Noise Rough
Noise Intermediate
Noise Smooth
Taubin Uniform
Taubin Rough
Taubin Intermediate
Venus VR MOS
VR MOS
0.8
0.6
0.4
Noise Uniform
Noise Rough
Noise Intermediate
Noise Smooth
Taubin Uniform
Taubin Rough
Taubin Intermediate
Venus Desktop MOS
Desktop MOS
0.8
0.6
0.4
Noise Uniform
Noise Rough
Noise Intermediate
Noise Smooth
Taubin Uniform
Taubin Rough
Taubin Intermediate

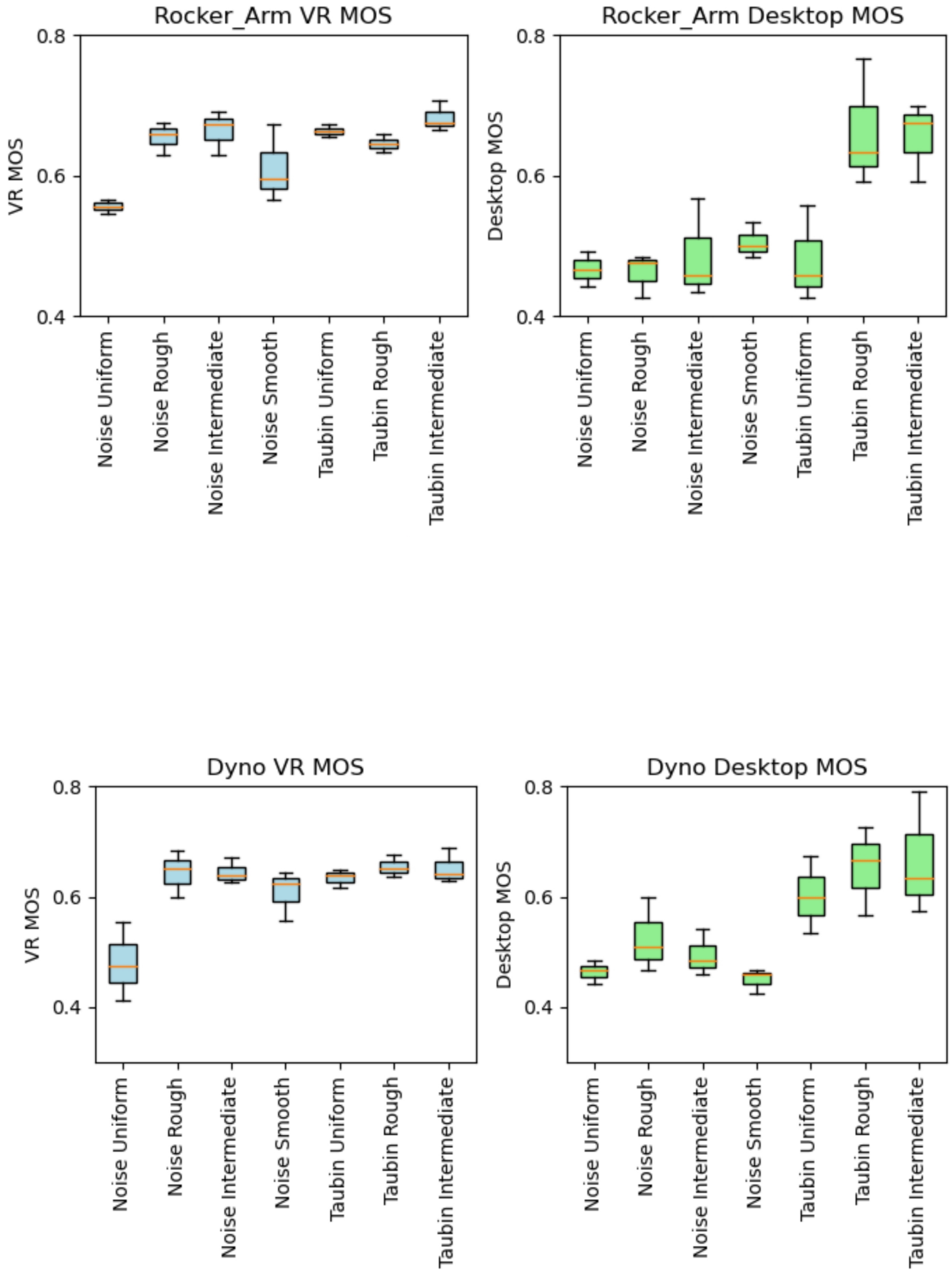


Figure 3.8: Comparison of MOS for both VR and desktop settings in the Pairwise Comparison (PC) experiment for all the stimuli shapes.

3.4.2 Distortion by Type and Location

As we explained in the previous section, this dataset has two main distortion types: adding noise and Taubin (smoothing), but each type has different levels of strength and different locations. These distortions have different levels of strength and four types of locations on meshes: uniformly on the whole mesh, smooth areas, rough areas, and intermediate areas, where different areas are identified based on local curvature variations.

We now group the results based on type and location, and for each type, we show different distorted shapes and levels of distortion strength (see Figure 3.10). The x-axis corresponds to each distorted shape (a combination of 4 shapes and three levels of distortion strength as shown in Appendix A.1. As we only have three values of distortion level strength (Low, Medium and High), we calculated them using Pearson and Spearman correlations. The PLCC and SROCC correlations between VR and desktop settings for different distortion types and locations are summarised in Table 3.3 (For results in the 'All' row, we computed the same formulae but using the whole dataset) with a detailed breakdown given in Table 3.2. More details are presented in Appendix A.2.

In the VR setting, users perceive 3D models differently compared to the desktop setting due to the higher resolution and enhanced interactivity that VR offers. Consequently, Mean Opinion Score (MOS) evaluations appear more sensitive to the location of distortions on 3D models. A clear example of this is the Noise Uniform distortion. This distortion, which uniformly adds noise across all shapes, was the most easily detected by participants. It produced consistent results in both VR and desktop settings, as indicated by high correlation coefficients ($r_p = 0.732$, $r_s = 0.753$). For the Noise Rough and Noise Intermediate distortions, noise is selectively added to the rough and intermediate regions of the mesh. These areas typically already contain detailed information, making the added noise less perceptible. As depicted in Figure 3.10, the x-axis groups three samples, each representing a different level of distortion strength (low, medium, and high) across four distinct shapes. Although adding more noise generally results in lower Mean Opinion

Scores (MOS) in both VR and desktop settings, the difference between these two environments is not substantial. Importantly, the decrease in MOS scores due to noise in rough or intermediate areas is significantly less than when noise is added uniformly across the entire mesh. In the case of adding noise to the smooth regions, the MOS scores have larger drops with increasing strength of noise, close to the Noise Uniform case. This shows that with better observation/interaction, subjective scores are more sensitive to where distortion, especially noise, is applied. In contrast, the results of the desktop settings show little difference between locations. For Taubin distortions we noticed that the VR setting is more consistent than the desktop setting between different subjects.

Table 3.3: Pearson and Spearman coefficient correlations between MOS scores from VR and desktop settings, grouped based on distortion types and locations.

Distortion Type/Location	Pearson Correlation (PLCC)	Spearman Correlation (SROCC)
Noise Uniform	0.732	0.753
Noise Rough	0.441	0.448
Noise Intermediate	0.273	0.266
Noise Smooth	0.335	0.387
Taubin Uniform	0.526	0.413
Taubin Rough	0.307	0.123
Taubin Intermediate	-0.078	-0.004
All	**0.929**	**0.929**

A visual example is shown in Figure 3.9 where (b) is the shape with high-level noise applied on the rough regions, whereas (c) is with medium-level noise applied on the smooth regions.We have compared the same mesh using the same distortion type but different locations to determine if the location affected the shape visualisation and MOS score. It is obvious that the distortion in (c) is more visible than in (b), which is correctly reflected in the MOS scores in the VR setting but not so in the desktop setting, where the strength of distortion rather than the location has more impact on the perceptual quality. Because of such differences, the correlations in these locations are significantly lower than in the uniform case.

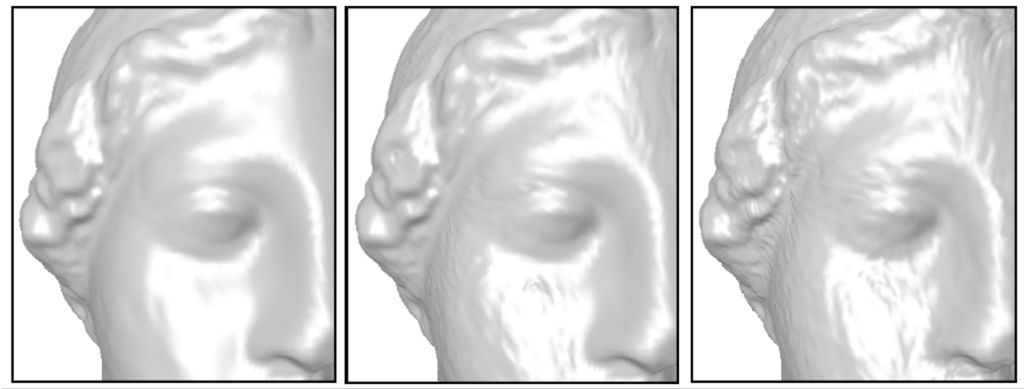


Figure 3.9: An example of the distortion types. (a) Original Venus model and illustration of the different types of regions; (b) high-level noise applied on rough regions; (c) medium-level noise applied on smooth regions.

We now compare smoothing (Taubin) and adding noise. As shown in Figure 3.10, Noise Uniform is highly correlated between VR and desktop settings. Also, this type of distortion shows a lower score in MOS on both settings which starts from 0.4 to 0.55. However, in Noise Rough and Intermediate distortion, VR participants give a high MOS score distortion on these types of distortion compared to desktop participants. MOS scores where smoothing is applied tend to be higher than with noise added, especially when distortions are applied uniformly. In contrast, in the desktop setting, these types of distortions have lower MOS scores. Similarly, different strength levels also have less effect on MOS scores than the desktop setting. These also lead to lower correlations between VR and desktop settings in smoothing. In the case of Taubin distortion, we notice that VR and desktop give high MOS distortion. This means this type of distortion is more visible in both settings. Nevertheless, if we consider all samples (shapes, distortions, locations and strength levels), the MOS scores remain highly correlated between VR and desktop settings ($r_p = 0.929$, $r_s = 0.929$).

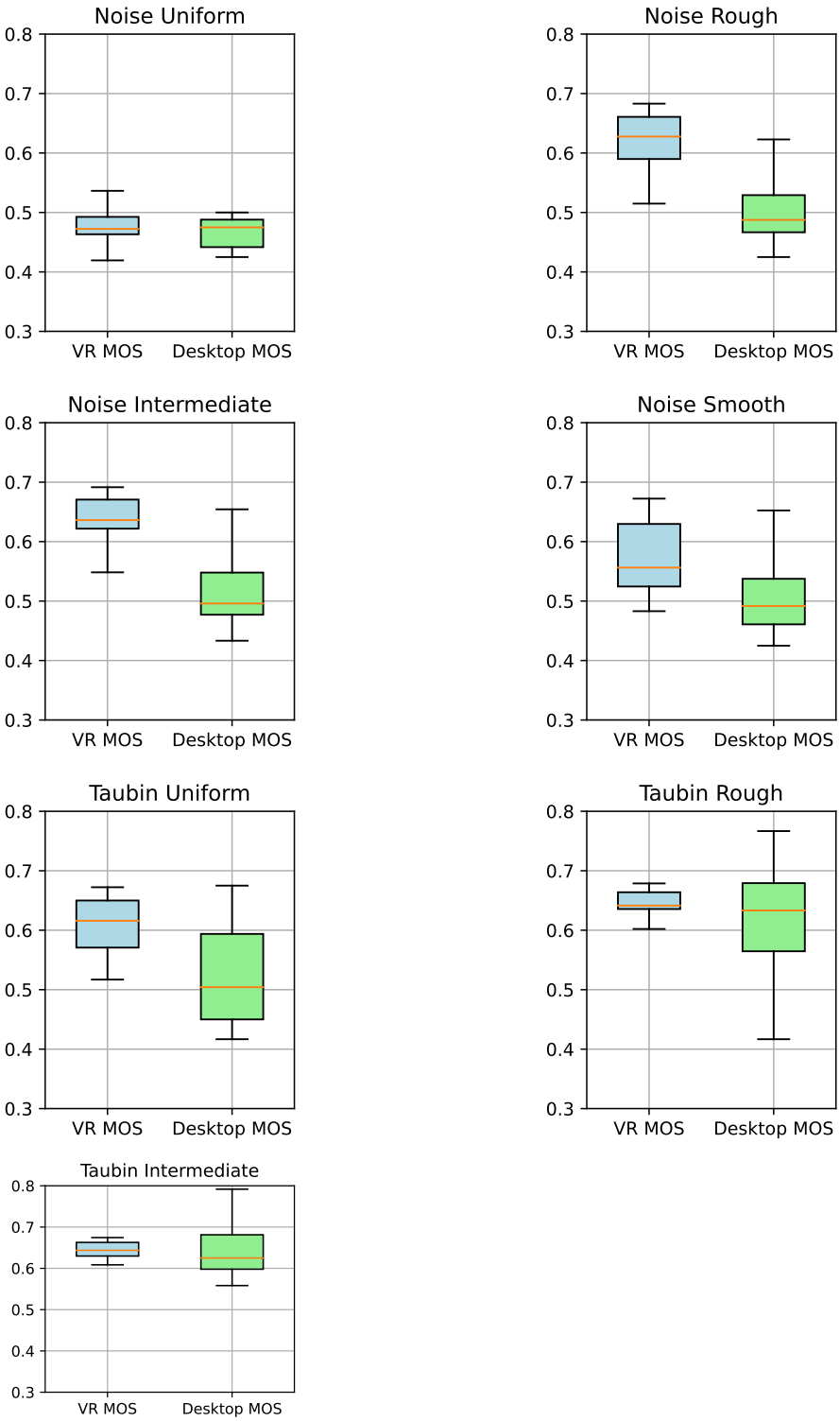


Figure 3.10: Comparison of MOS scores between the VR and desktop settings. Each figure shows a type of distortion (Noise, Smoothing and Taubin) applied to certain locations (Uniform, Rough, Intermediate and Smooth). The x-axis corresponds to VR MOS scores or Desktop MOS scores, and the y-axis shows the (normalised) MOS scores averaged over all subjects for the distorted shapes. The blue and green plots correspond to the VR and desktop settings, respectively.

3.5 Summary

This chapter proposed a subjective study using linear and nonlinear correlation coefficients to compare MQA in virtual reality and traditional desktop display settings. Our analysis indicates the actual perceived mesh quality varies according to shapes and is sensitive to different types/locations of distortion. We notice that overall MOS score distributions are highly correlated between the VR and desktop settings. However, in the VR settings, noise is more noticeable than a loss of details, when compared with the desktop setting. In particular, noise added to the entire shapes or smooth regions tends to be more noticeable than in other regions, and the differences are much more significant in the VR setting. The findings can provide useful guidance when processing 3D shapes for VR applications. In future work, we aim to construct a larger-scale database of perceptual quality under different combinations of distortions.

Chapter 4

Learning to Predict 3D Mesh Saliency

Overview

As previously mentioned, saliency is related to the perceptual quality of 3D shapes. This chapter introduces novel methods for measuring salience on a 3D model, measured by running experiments with an eye tracker, which produces a saliency measurement that is accurate to human visual judgment. After giving a brief introduction of the problem in Section 4.1, relevant literature is discussed in the context of this chapter in Section 4.2, which extends the literature reviewed in Chapter 2. Section 4.3 details some existing applications that benefit from 3D mesh saliency. Section 4.4 and Section 4.5 show Voronoi tessellation and SSIM method used in our 3D mesh saliency and our method to build subjective quality perception for mesh saliency via eye-tracking .The user study experiment conducted for measuring 3D mesh saliency is described in Section 4.6. With the data collected during the user study, we obtain ground truth mesh saliency by fusing eye fixations from individual views and different participants. The ground truth mesh saliency is further used to evaluate existing mesh saliency methods 4.7. Learning new methods of measuring 3D mesh saliency is presented in Section 4.8, and the results are shown in Section 4.9, which achieves better results than existing methods. Finally, some conclusion thoughts and suggestions for future work are given in Section 4.10. In the next chapter, we will demonstrate that incorporating saliency weighting into objective quality metrics lead to improved performance.

4.1 Introduction

Mesh saliency can play an important role in computer graphics in determining the outcomes of many tasks, such as feature detection [191], shape recognition [200], mesh segmentation [36], mesh watermarking [137], 3D printing [218], etc. Mesh saliency measures the perceptual importance of local regions on a mesh, which is subjective. It can be considered from a generic perspective where some 3D surface regions are considered more important than others [115, 192], but also task-dependent, e.g. in relation to touching [105]. Mesh saliency is also related to other metrics that measure "uniqueness" or "distinctiveness", e.g. surface distinction [184] and region distinctness [116]. However, distinctiveness measures focus on regions that set a shape apart from others. The ground truth for mesh saliency is typically determined by human perception, and subjective judgement is therefore involved in assessing the performance of such approaches.

Many computational models for image visual saliency have been proposed and implemented. In 1985, Itti et al. [93] proposed an early model which stated that image locations with saliency would have some distinction from their surrounding environment. Some researchers in this field, such as [84, 140, 210] have described in their works various other models of saliency. In 2002, Stove and Straßer [199] used saliency information acquired from an individual's eye movements to simplify images, generating a non-photorealistic, painterly rendering. However, these works focus on eye tracking for image saliency rather than the saliency of 3D shapes, which we investigate in this chapter.

Although there are many ways to detect 3D mesh saliency, few techniques have been developed to evaluate their effectiveness. Many papers utilise heatmaps for showing salient model parts or utilise saliency-led mesh simplification to demonstrate the methodology while preserving attention-grabbing parts. Even though such approaches show how they operate at a high level, they make it hard to compare the effectiveness of various methods as they only provide a subjective evaluation or may not measure the saliency directly. The challenges in existing mesh saliency models are largely built with hard-

coded formulae, which cannot capture true human perception. Some existing techniques utilise indirect measures to capture user perception (e.g., mouse clicks), which can be unreliable. Our contribution is to investigate a methodology to produce ground truth mesh saliency through the fusion of eye-tracking data for different views of rendered 3D shapes. Based on this, we further develop machine learning methods for predicting saliency on 3D meshes. In this chapter:

- We investigate using eye-tracking data of rendered views of 3D shapes to obtain ground truth saliency on meshes.
- As each view is only able to cover part of the mesh, and different views may contain shape parts with significantly different levels of saliency, an optimisation approach is developed to fuse saliency derived from individual views to take into account their relative saliency levels while ensuring consistent saliency values in the shared regions.
- We further build machine learning models to predict mesh saliency based on local geometric features and existing 3D saliency prediction models. Our experiments show that a learning-based approach performs better than existing saliency methods on unseen shapes.

4.2 Related Work

Several algorithms for computing the saliency of 3D models have been developed recently. Lee et al. [115] were the first to introduce the concept of mesh saliency, a computational measure of regional importance on a mesh. Their approach is based on differences in Gaussian, a geometric measure aiming to approximate human perceptual importance. Kim et al. [101] conducted a user study comparing earlier mesh-saliency approaches to human eye movements using a 2D method. To measure the association between mesh saliency and fixation positions for 3D rendered images, they implemented the standard-

ised chance-adjusted saliency and demonstrated that the existing computational models of mesh saliency could significantly better predict human eye movements than a purely random model or a curvature model. Their goal is that the computational model of mesh saliency has a better correlation with human eye fixations than a random model regardless of viewing direction for the first few seconds after stimulus onset.

Although the importance of regions on 3D shapes can be considered general, it can also be task-specific. For example, the work [105] considers the problem of tactile mesh saliency, where saliency is defined in the context of grasping, pressing and touching. This chapter focuses on general visual saliency (i.e., without a specific task).

Many algorithms are based on the 'centre-surround' method of Lee et al. [115] that uses the absolute difference between the Gaussian-weighted average of the mean curvature at scales σ, and 2σ, with the Gaussian filtering limited to neighbourhoods of size 2σ. Several scales with different σ values are jointly used to capture saliency at different scales. Yang et al. [233] proposed a method for quantitatively calculating visual attention based on eye-tracking data for 3D scene maps by obtaining the participants' gaze behaviour differences to establish a quantitative relationship between eye movement indexes and visual saliency.

Liu et al.'s [128] use of virtual agents to simulate how humans interact with objects helps to understand shapes and to identify their salient parts in relation to their functions. Moreover, Chen et al. [37] investigated human perception and considered 3D mesh Schelling points, which are feature points people choose in a coordination task. They found that Schelling point sets are usually highly symmetric, and local curvature properties are the most useful method for identifying Schelling points. They propose using sophisticated deep learning approaches to discover mesh Schelling points automatically, without the need for participant observations. The authors use mesh convolution and pooling to extract meaningful characteristics from mesh objects and then predict the 3D heat map of Schelling points end-to-end [34].

The multi-scale computational model was developed by Song et al. [192] and uses a set of meshes which are simplified to various degrees. It also calculates the scale saliency map related to each and every scale through the computation of the spectral mesh saliency for every scale. The scale saliency maps can then be put together to produce a final saliency map.

Most research is dedicated to detecting saliency on 2D images and 3D meshes; little work has been conducted on 3D point clouds for accurate saliency detection. Guo et al. [77] introduce a new saliency detection approach for point clouds by using principal component analysis (PCA) in a sigma-set feature space, a method that is introduced to transform covariance descriptors to Euclidean space. In this method, they construct local shape descriptors based on covariance matrices for saliency detection, considering that covariance matrices can naturally model nonlinear correlations of different low-level compact and rotational-invariant features. By transforming these covariance matrices to vector descriptors in Euclidean vector space by applying the sigma-point technique, which keeps the inherent statistics of regions of 3D point clouds. PCA is employed in the descriptor space for identifying saliency patterns in a point cloud based on their informative descriptors. This method shows its advantages of being structure-sensitive, capturing geometry information and being computationally efficient.

Mesh saliency has many interesting applications. Leifman et al. [116] presented an algorithm for identifying regions of interest on 3D surfaces. Their method studies 3D regions of distinctiveness from local and global perspectives and demonstrates that saliency derived from their method is effective for viewpoint selection. Howlett et al. [89] demonstrated the value of saliency for guiding 3D simplification, where saliency was captured using an eye-tracker for recording the two-dimensional image area in which an individual has looked at a three-dimensional model.

4.3 3D Mesh Saliency Applications

This section reviews some applications that could benefit from 3D mesh saliency. Rather than analysing each application in depth and comparing it to possibly related approaches, this section aims to emphasise the relevance and utility of 3D mesh saliency for various applications to stimulate further work in this domain.

Before discussing the 3D view selection and 3D mesh simplification applications in detail, we consider the importance of these applications on 3D mesh quality and saliency. 3D mesh view selection is important in mesh quality assessment because it allows for a more detailed analysis of the mesh structure and geometry. When viewing a mesh from different angles, it is easier to identify various issues that may be present, such as mesh distortion, irregularity, or discontinuity. For example, when analysing the mesh quality of a complex surface or object, a mesh view selection can help to identify regions with more important details and therefore require higher mesh density. Similarly, by examining the mesh from different angles, it is possible to identify regions with poor connectivity or mesh distortion caused by poor element shape. Furthermore, 3D mesh view selection can be used to check for mesh quality at different levels of detail, from a global view of the entire mesh to a more detailed examination of specific regions or elements.

As a common approach to reduce computational and storage costs, 3D mesh simplification is one of the most important techniques because it reduces vertices, edges, and faces while preserving the mesh's overall shape and important features. This process can be used to improve the efficiency of algorithms that require mesh processing, such as simulation or rendering. Therefore, it is important for mesh quality assessment methods to evaluate the perceptual quality for simplified meshes.

Simplification can also be used as a preprocessing step for mesh analysis algorithms that are sensitive to the number of vertices in a mesh. For example, computing the curvature of a mesh requires the computation of local neighborhoods around each vertex,

which can be computationally expensive for large meshes. A simplified mesh reduces the number of vertices, allowing curvature computation to be performed more efficiently.

As saliency helps with mesh quality, it can be used to guide the mesh in these applications by identifying the important features and regions that should be preserved. Saliency measures can help identify regions of high curvature, sharp edges, or features that are important for simulation or rendering. By preserving these salient features, the 3D view selections allow for a detailed analysis of the mesh structure and geometry and the simplified mesh can retain the essential information while reducing the complexity of the mesh. There are several techniques for salience-based mesh simplification, including the use of salience measures based on curvature, distance, or feature detection. For example, curvature-based salience measures can identify regions of the mesh with high or low curvature, and guide the simplification process to preserve the high-curvature regions while simplifying the low-curvature regions.

4.3.1 3D View Selection

Although it is often problem-dependent, 3D view selection aims to find the most informative views for 3D shapes. The best view is often selected by computing the geometric complexity based on view descriptors. A descriptor calculates the geometric complexity from the visible surface of the 3D object. The viewpoint which maximises the geometric complexity is considered to have the most information on that view of the object. Lee et al. [115] suggested that the best view contains the most important features of a 3D object. The view selection algorithms assess the quality of the view according to the descriptors. To develop the best view selection algorithms, mesh saliency can be utilised that detect the salient regions of the 3D mesh model. Researchers in [57] developed a best view selection method and introduced seven descriptors, including view area, mesh saliency, the ratio of the visible area, silhouette length, surface area entropy, silhouette entropy, and curvature entropy for the performance evaluation. These descriptors are also considered to be used to solve computer vision research problems in the future.

Researchers in [193] introduced the best view selection technique based on a deep CNN model named Classification-for-Saliency CNN (CFS-CNN). First, they rendered the mesh into 24 2D views. These views are selected between -30° to +30° elevation. Then a weakly supervised learner is designed using the pre-trained VGG-19 network [33]. The VGG-19 network introduced two new layers: the View Saliency (VS) layer and the Saliency-based Pooling (SP) layer. The VS layer learns the saliency features from different layers of VGG-19 and feeds the features into the SP layer to further use these features in the classification. After training the CFS-CNN, the CAM (Class Activity Map) method [190] is utilised to calculate the saliency map for each pixel. After that, they generate the view-based saliency maps from a 2D to a 3D saliency map. The proposed method learns the vertex-level annotation from scene saliency to select the best view in the 3D scenes. The score S_i of the i-th view Z_i is defined as

$$S_i = \frac{L_i \sum_a M_a(Z_i)}{\sum_j (L_j \sum_a M_a(Z_j))} \tag{4.1}$$

Here, $M_a(Z_i)$ represents the 3D saliency of vertex a, and L_i is the saliency of the view Z_i. The best view should be prominent when compared with other views.

Song et al. [194] introduced an unsupervised multi-view CNN (UMVCNN) framework to select the best view for 3D objects. The UMVCNN model is based on the VGG-19 network [33] and extracts features from the Softmax Layer associated with the FC9 layer. Then the saliency map is generated by back-propagating the vector with entries from the Softmax Layer to input views. After that, a 3D saliency map M_i is obtained by transforming the saliency from 2D to 3D. Then the salient viewpoint is selected where the sum of saliency maps of the 3D object is maximised. Giorgi et al. [72] presented a salient view selection technique based on semantics. They divide the shape into features such as blobs and tubes using the Plumber method [159]. For most shapes, the salient views correspond to the tubular parts. Here, the semantic-based technique assigns higher weights to the tubular features than blobs. A scoring function is designed to evaluate the different viewpoints. The scoring function considers visibility, relevance, and feature type

for salient view selection.

4.3.2 3D Mesh Simplification

3D Mesh simplification decreases the number of faces, vertices, and edges of 3D mesh models. Mesh simplification is performed after the acquisition of geometric data. The faces and vertices are merged appropriately, simplifying the complex mesh model.

However, informative features such as sharp area, volume, and boundary should be preserved. Pellizzoni et al. [161] proposed a mesh simplification algorithm named iterative edge contraction (IEC). Their method is based on [84] which utilised the quadratic error metric to find the optimal position of the vertex. Unlike the classic method [84], the discrete curvature is computed using the Gauss-Bonnet method where the discrete Gaussian curvature at a vertex is defined as 2π minus the total angles formed by edges emanating from the vertex, which more aggressively simplifies rounded and flat regions.

In [119], the simplification process works to preserve the topological structure as much as possible. The original 3D mesh model is approximated using fewer faces and points. Asgharian et al. [18] presented a technique for simplifying a complex 3D mesh. They reduce the vertices by mesh re-sampling utilising the Nyquist theorem that specifies the most relevant samples necessary for simplification. The Nyquist method selects the maximum and minimum curvature in different directions in the original 3D mesh for adaptive sampling. The sampling approach obtains the optimal number of samples in the simplified 3D mesh. The presented simplification model observed that visual quality could be preserved without significantly losing precision. To simplify the triangular meshes, a novel Laplacian-based technique [131] is introduced, which can also preserve the different geometric features. The Laplacian-based method produced fast and efficient results. To classify the vertices, the Laplacian descriptor is combined with the K-means clustering method. The introduced method showed promising results in terms of accuracy and preserving geometric features.

4.4 Voronoi Tessellation and Delaunay Triangulation

In geometry, Voronoi tessellation and Delaunay triangulation are used for partitioning and analysing points as shown in Figure 4.1. Algorithms like these are usually used in 2D, but they can also be applied to 3D meshes [67]. These are used in our pipeline to construct ground truth mesh saliency maps from eye fixation data.

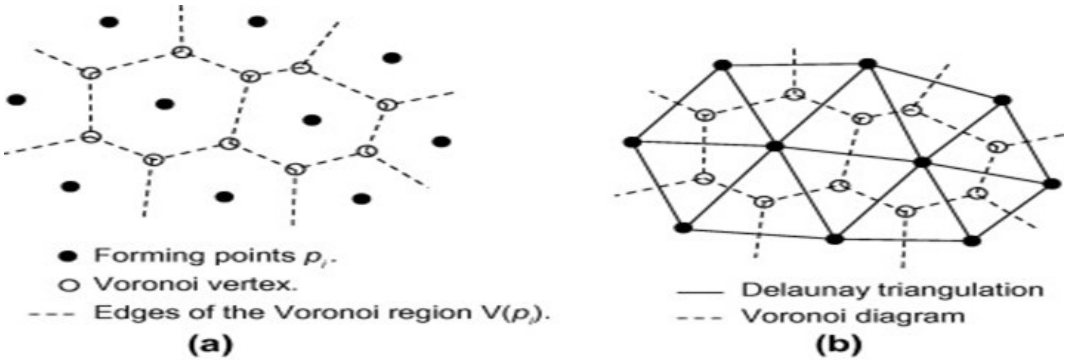


Figure 4.1: This illustration shows the Voronoi and Delaunay area in a 3D mesh around a given vertex.

Voronoi diagrams divide planes into regions based on their proximity. Each region contains points that are closest to a particular input point. By doing this, Voronoi cells or Voronoi polygons can be formed, covering the entire plane. By using Voronoi tessellation, 3D space can be divided into polyhedral cells. All points in space closer to an input point are considered to be part of the Voronoi cell for that input point. They are typically bound by planar surfaces called facets. Voronoi cells are formed by connecting pairs of neighbouring input points with perpendicular bisectors [55]. In the case of a 3D point cloud, the goal is to create polyhedral cells that cover 3D space. A point cloud consists of cells that represent specific points. Each Voronoi cell contains all the points in space that are closest to a particular point. Those lines that connect neighbouring points in the cloud are the perpendicular bisectors that define their boundaries. There are several algorithms available for computing Voronoi tessellation for 3D point clouds, such as Fortune's algorithm and incremental algorithm. Based on proximity relationships between points, these algorithms iteratively add points and construct Voronoi cells. As a result, the 3D space is divided into polyhedral cells, where each cell represents a point in the cloud [202].

The Delaunay triangulation, on the other hand, is a triangulation of input points. A triangle formed by input points will not have any points inside the circumcircle. Delaunay triangulations maximise the minimum angle of all triangles in 2D by connecting input points to form triangles. The Delaunay triangulation can be extended to 3D meshes by connecting the input points to form tetrahedra, which are the 3D equivalents of triangles. By connecting four non-coplanar points in 3D Delaunay triangulation, no point lies inside the circumspect of any tetrahedron created by the input points [38]. As for the Delaunay triangulation of a 3D point cloud, the goal is to create a triangulated mesh. In the point cloud, triangles are formed by connecting three non-collinear points. There should be no point in a Delaunay triangulation that lies inside a tetrahedron formed by the points. A Bowyer-Watson algorithm or an incremental algorithm can be used to compute the Delaunay triangulation of a 3D point cloud. In these algorithms, points are added and tetrahedra are formed based on Delaunay criteria. In this case, the triangles are non-overlapping and have the minimum angle that covers the point cloud. Delaunay triangulation and Voronoi tessellation can both provide valuable information about spatial relationships between points in 3D point clouds. The Voronoi tessellation illustrates how space is divided around each point. Delaunay triangulations provide a connectivity structure that is useful for generating meshes and reconstructing surfaces.

4.5 Integration of Structural Similarity Index Model (SSIM) in a Subjective Experiment Utilising Eye-tracking

The quality estimation model proposed by Wang et al. [223] determines the structural similarity between two images and is used as the framework of quality measurement, so a distorted image can be compared with the original one to figure out how visually precise the indistinct image is to the novel image. They proposed a metric for structural similarity measure (SSIM). The SSIM method joins three mechanisms to create a similarity measure between two images through contrast, luminance and structure.

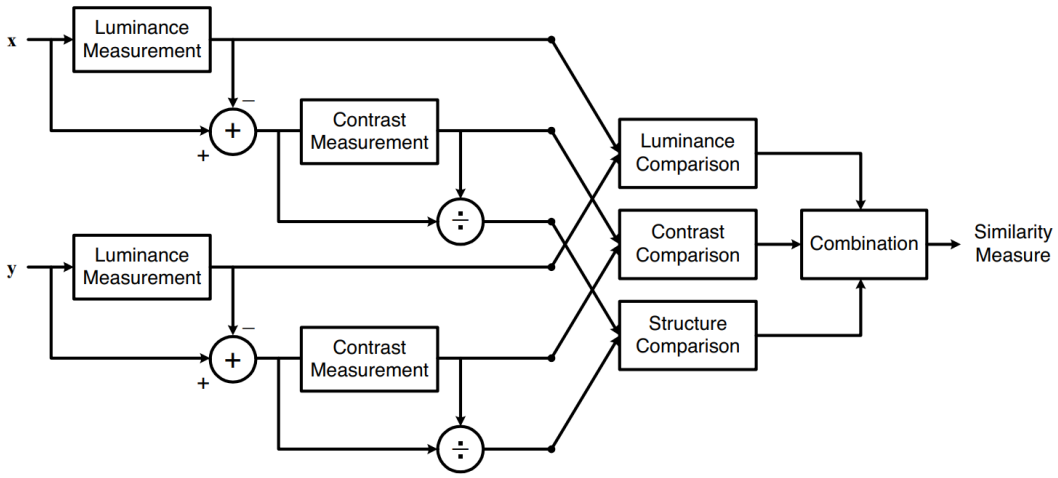


Figure 4.2: This illustration shows a diagram of the structural similarity (SSIM) measurement system [223]

Comparing two images A and B, the SSIM method takes a spatial patch around each pixel to get local information around the pixel. Then a Gaussian function scales the pixel intensity to give more influence to points closer to the centre of the patch. The mean patch intensity is then calculated to produce μ_x and μ_y where x and y are spatial patches of pixels in the same spatial position in images A and B. We would like to measure the similarity between two saliency maps. In Figure 4.2, the system separates the task of similarity measurement into three comparisons: luminance, contrast and structure. First, the luminance of each signal is compared. Assuming discrete signals, this is estimated as the mean intensity. The luminance comparison function $l(x, y)$ is then a function of μ_x and μ_y. Second, the mean intensity from the signal. In discrete form, the resulting signal corresponds to the projection of the vector onto the hyperplane. Using the standard deviation as an estimate of the signal contrast, the contrast comparison $c(x, y)$ is then the comparison of σ_x and σ_y. Third, the signal is normalised (divided by its own standard deviation), so that the two signals being compared have unit standard deviation. The structure comparison $s(x, y)$ is conducted on these normalised signals $(x - \mu_x)/\sigma_x$ and $(y - \mu_y)/\sigma_y$. Finally, the three components are combined to yield an overall similarity measure:

$$S(x, y) = f\left(l(x, y), c(x, y), s(x, y)\right). \tag{4.2}$$

It is important to note that the three components are relatively independent. The image

structure does not change when luminance or contrast is changed.

It is necessary to define the three functions $l(x, y), c(x, y), s(x, y)$, as well as the combination function, in order to define the similarity measure in (4.2). The similarity measure should also meet the following criteria: Symmetry: $S(x, y) = S(y, x)$. Boundedness: $S(x, y) \leq 1$. Unique maximum: $S(x, y) = 1$ if and only if $x = y$ (in discrete representations, $x_i = y_i$ for all $i = 1, 2, \ldots, N$, where N is the number of pixels).

This can be used to construct a luminance comparison function below

$$l(x, y) = \frac{(2\mu_x\mu_y + C_1)}{(\mu_x^2 + \mu_y^2 + C_1)} \tag{4.3}$$

where the constant C_1 is included to stabilise the function when μ_x + μ_y approaches zero. The standard deviations of the spatial patches are used as a contrast metric; with this calculation, a contrast comparison function can be rendered close to the luminance comparison function.

$$c(x, y) = \frac{(\sigma_x\sigma_y + C_2)}{(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{4.4}$$

where the C_2 constant is another constant for stabilisation. In order to calculate the final part structure.

$$s(x, y) = \frac{(\sigma_{xy} + C_3)}{(\sigma_x\sigma_y + C_3)} \tag{4.5}$$

We combine the three comparisons of (4.3), (4.4) and (4.5) and name the resulting similarity measure the SSIM index between signals x and y.

$$SSIM(x, y) = [l(x, y)]^\alpha \cdot [c(x, y)]^\beta \cdot [s(x, y)]^\gamma \tag{4.6}$$

where $\alpha > 0$, $\beta > 0$, and $\gamma > 0$ are parameters used to adjust the relative importance of the three components. It is easy to verify that this definition satisfies the three conditions given above. In order to simplify the expression, we set $\alpha = \beta = \gamma \geq 1$.

Finally, once we obtain the similarity measured SSIM index between signals x and y as in equation (4.6), we now calculate the SSIM index between the neighbourhoods of

two pixels, taking $C_3 = \frac{C_2}{2}$ produces the following function:

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + C_1) + (2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{4.7}$$

As SSIM is a method used for 2D images, it focuses on luminance, contrast and structure. In this work, we translate it to 3D meshes. The luminance equivalent consists of a vertex property on the mesh, such as colour, normal, or vertex position instead of pixel intensity. As we use SSIM to measure the similarity of mesh saliency, the mesh saliency value at each vertex is used in place of luminance. The contrast would be variability in the vertices' properties, e.g. the variability of saliency values. In 3D meshes, unlike 2D images where the neighbourhood is directly defined, the topological structure is determined by the adjacency or connectivity of vertices and edges. To address this, the only change made in this work is how to extend the SSIM function to 3D meshes and how to analyse the spatial patches (as neighbourhoods on meshes). Instead of inputting 2D images, a value list is sent for each vertex representing the saliency at that vertex is used for each input mesh.

The area covered by a pixel in an image is uniform for all pixels in that image, but this is not necessarily true for 3D models. Each vertex is connected to faces of different sizes, and each vertex has a different degree of influence on how the model looks. To account for the varied vertex effect, the saliency value of each vertex will be multiplied by the Voronoi area of the vertex (see Figure 4.1), which for a triangular mesh is the sum of 1/3 of the area of each face of the vertex. Changing SSIM to operate on 3D heat-maps requires adapting the neighbourhoods of the standard SSIM. SSIM takes a window around each pixel when working on 2D images, but this does not work directly for meshes due to their irregular connectivity. We, therefore, replace such windows with neighbourhoods on meshes within a certain distance to the vertex of concern. For this purpose, a smaller neighbourhood than that used in eye-tracking mapping is more meaningful. We set the SSIM neighbourhood distance threshold as $0.02 \times d_{\max}$ where $d_{\max}$ is the farthest distance between pairs of vertices on the mesh. Increasing the size of the neighbourhood in

SSIM calculations to 3% or higher of $d_{\max}$ would generally make the method less sensitive and could potentially reduce the method's ability to detect small important structural differences.

We explained in Section 2.7 eye tracking and the metrics that are used in different applications. In our experiment, we used eye tracking (fixations). Eye tracking data is widely used for the analysis of the behaviour of visualisation research, human-computer interaction, scene perception and visual quality. Hanhela et al. [79] developed a model to extort 3D gaze data in sequence from a camera eye-tracking data. Per eye gaze data has been analysed by using the stereoscopic of the human visual system by the conversion of data into stereoscopic volume-of-interest through a 3D heat-map of the eye tracking experiment. By finding a connection between observers and increasing the tracking precision in this model, it is possible to optimise the participation of observers concerned in 3D gaze-tracking experimentation which helps to achieve a high precision level in 3D gaze tracking. In our experiment, the eye tracking experimental design includes several steps, the first showing the 3D shape to the participant. The second step is mapping the eye-tracking data to produce a heat-map describing the 3D saliency of the model. The implementation detail is given in Appendix B.3 Figure B1.

Unlike subjective image saliency measure which is straightforward to present to the participants, measuring 3D mesh saliency increases the question of how we present models to participants. We could show a rotating animated model to the participant; however, this may introduce a bias as people lose focus towards the end of the image view. This project takes 20 2D images of a model from various positions around the model to show all parts of a model without introducing a bias (see Figure 4.3). The images are taken from the centre of different faces of an icosahedron scaled to surround the model to ensure even coverage. Icosahedrons have 20 equal triangular faces, 30 edges, and 12 vertices. When it is subdivided (each triangular face is split into smaller triangles), it looks more like a sphere. Due to the fact that it uses a relatively small number of vertices and faces, it provides a good approximation of a sphere. Note eye tracking provides an intuitive way

for collecting user interest given visual stimuli. However, the participant cannot see the entire shape at once. To address this, we place each shape at the centre of origin, scale it to fit in the unit sphere, and render 20 evenly distributed views (using face centres of an icosahedron as the camera location with the camera direction pointing to the origin) to provide sufficient coverage of the shape shown Figure 4.3.

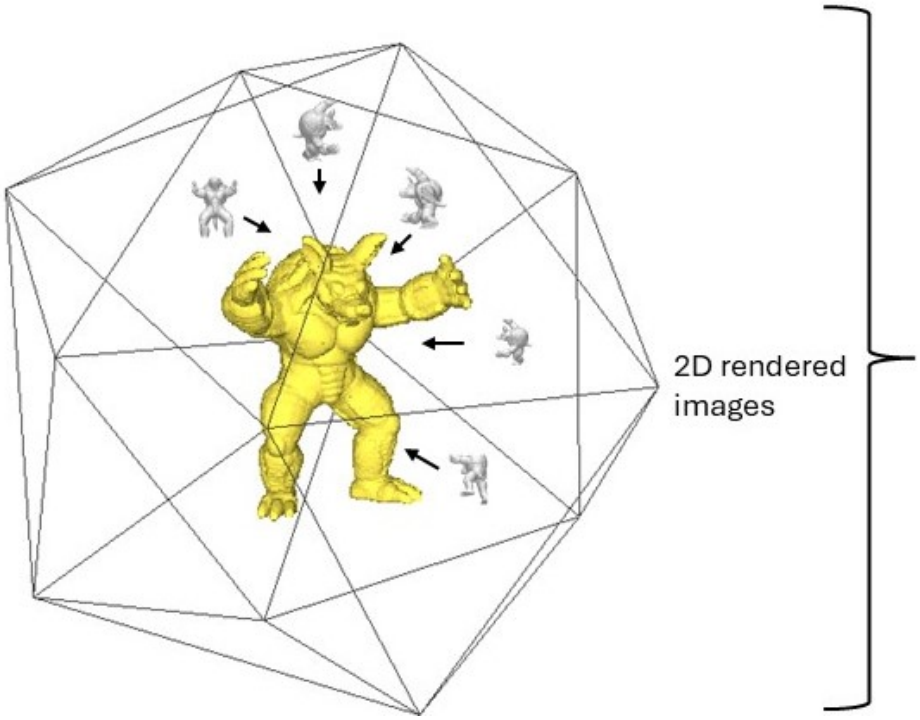


Figure 4.3: Example of 3D shape is rendered from 20 different views.

Several lights are placed around each point of the camera to ensure equal lighting. Specular lighting strength was set to 0 as bright spots on a model could draw people's attention. This would give areas of relatively low geometric interest high saliency. To facilitate mapping between the mesh and 2D rendered images, we also generate another set of images. A 3D scatterplot marks the location of each vertex on the mesh. Each vertex marker is colour coded by the vertex index using RGB encoding. Each marker takes up exactly one pixel of the image. When given a fixation location in pixels the remapping code can identify which vertex is at that location (or closest to it). The mesh and the background of the image are then coloured black. This ensures that the only non-zero RGB values are the vertices, otherwise, the remapping code could take the grey areas of the mesh as a colour-coded vertex and either have array-bound errors or just incorrect vertex selection. The benefit of having the mesh in the scene blacked out is that vertices behind the mesh that would not be visible to a viewer do not show up on the vertex map as they are blocked by the mesh as shown in Figure 4.4. This stops the situation where a fixation may lie between two vertices and the remapping decides the fixation is on a vertex on the opposite side of the mesh, where the participant could not see at all. After these images have been generated, they will be used to obtain saliency data for each image in an eye-tracking experiment.

Note that although 2D-rendered images are used to collect eye-tracking data, the saliency values are associated with mesh vertices, and thus are ensured to be view-independent. However, the number of views can affect the quality of mesh saliency maps. In cases where fewer than 20 images are used, the mesh model's coverage may be limited. Therefore, it may be difficult to determine salient regions accurately if we do not have a comprehensive understanding of the model's appearance from different angles. Due to the limited viewpoints, the saliency map may be more sensitive to the views used in data collection. However, the specific impact will depend on the saliency algorithm and the quality of the data used. It is possible to estimate saliency using around 20 images and provide reasonably good coverage of viewpoints. It is more likely that a saliency map will capture a range of salient features with various viewpoints. Using more images can result

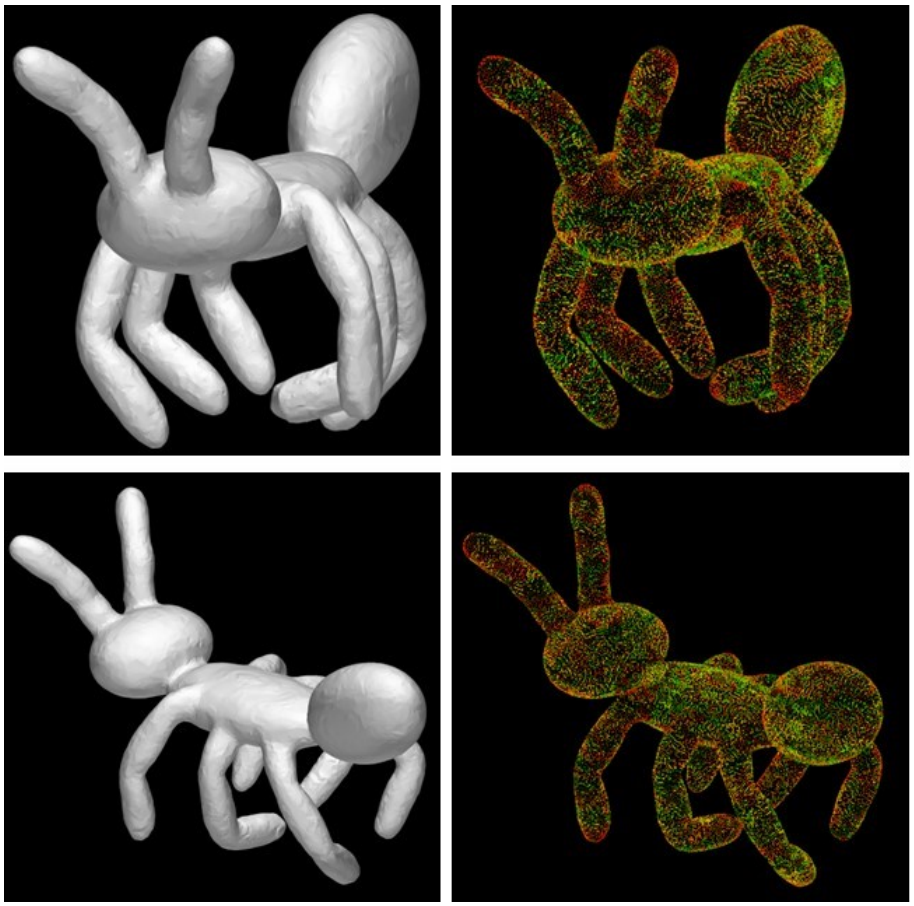


Figure 4.4: Example of an object from two views and its corresponding vertex maps.

in more accurate and saliency maps. However, the actual effect depends on factors such as the model's complexity and its distribution. A saliency algorithm, dataset, and model characteristics can all influence the optimal number of images for saliency estimation. To determine which method is most appropriate for a particular application, it will need to experiment with different numbers of images and evaluate the resulting saliency maps.

4.5.1 Design of User Experiments

The design of our experiment is shown in Figure B1 in the Appendix. When producing these view images, we must ensure they express clear clues for 3D shapes but avoid introducing artefacts that may distract user attention. We have done a demo to check the suitable background colour that makes the participants focus on the shape rather than the background colour. We found the black background colour is more suitable to focus on

compared with the white background Figure 4.5.

Our preliminary user evaluation shows that the black background is less distracting than the white background (see Figure 4.6), so participants will concentrate on the actual shapes rather than their attention wandering around in the background. We also set the light source to be in the same location as the camera and pointing towards the object's centre, ensuring the captured view is well-lit (but not over-exposed).

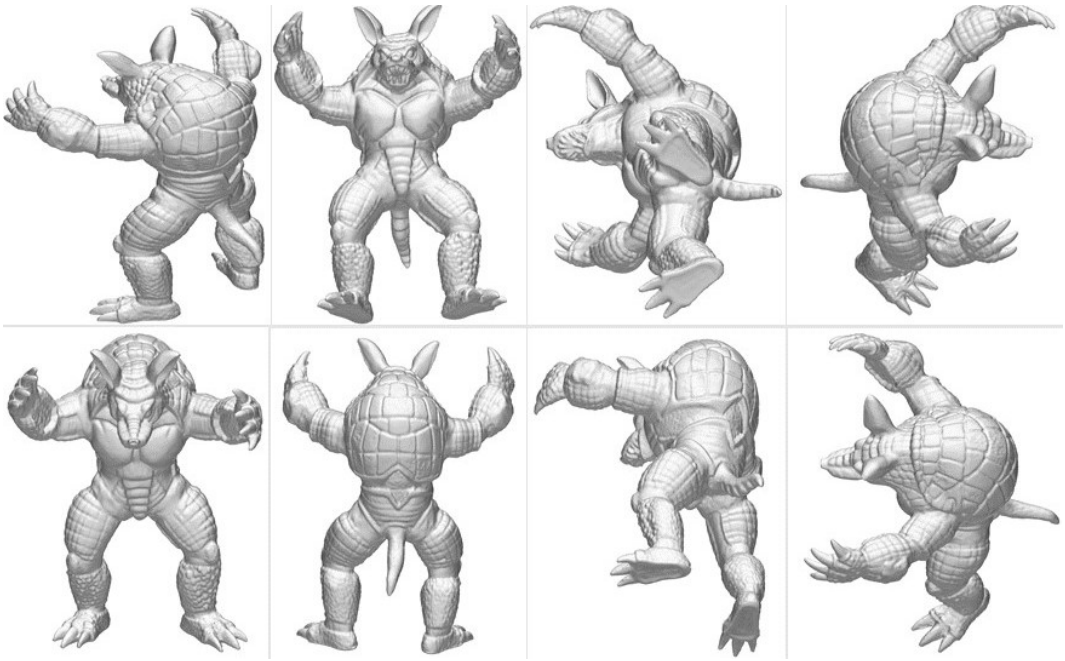


Figure 4.5: Example views of one mesh (Armadillo) white background. 8 out of 20 views are shown here.

Figure 4.6: Example views of one mesh (Armadillo) with the black background. 20 views are shown here.

4.6 Experimental Setup

To make the task manageable, in particular, to ensure that participants can concentrate while doing the study, we select 20 shapes as shown in the references in Appendix B.1 in our study. This leads to 400 rendered images, 20 views per shape at 1920×1080 pixel resolution. As described before, each participant is shown four views of each shape, leading to a total of 80 images. The eye-tracking data of all users is then fused to produce ground truth mesh saliency on 20 shapes. When applying our dataset for machine learning, We have used 5-fold cross-validation, a method of cross-validation that randomly partitions the original sample into 5 equal subsamples. Among the 5 subsamples, one subsample is retained as validation data, and the remaining 4 subsamples are used as training data. During cross-validation, each subsample is used as validation data exactly once, as the process is repeated five times. The average of the five results can then be used to produce a single estimate. Note that although 20 shapes are not many, each shape contains thousands of vertices, and thus it is sufficient for training machine learning models when applied at the vertex level.

As the subjective result is hard to measure, before we started the experiment, we showed the participants a trailer of the experiment to ensure that the participants fully understood the task before starting. Some other studies allow the user to rotate and zoom in/out of the model, but our study is different because we use an eye tracker camera to calculate the time users need for each view. Also, we let the participants ask questions during the trailer to feel more confident about getting accurate results. The participants are then shown images of these rendered views, and their eye-tracking data is captured. However, adjacent views naturally have large overlaps, which not only happens naturally but is also useful, as this allows saliency captured from different views to be reliably fused. However, this leads to a potential problem of memorising: when a user is presented with a similar view shortly before, he/she may not actively try to explore interesting features of the shape.

To avoid this, as we present in Figure 4.7 we carefully group the 20 views into five groups, each with four views, such that these four views are as widespread as possible, and each user is only asked to look at one group of views for each shape. The user is then asked to watch these views (4 views per shape) in a random order to avoid bias, with each rendered view shown for 5 seconds [54]. If the participant has less than 5 seconds, it may not be enough to capture all the primary points of interest, although it may be adequate for very simple stimuli like asking what the object in the scene is. However, when viewing a scene for longer than 5 seconds, the viewer may also look at less important regions. Each rendered image is followed by 2 seconds grey background to break any fixation from the previous image and to provide a pause to allow the subjects' eyes to relax and focus (see Figure 4.8).

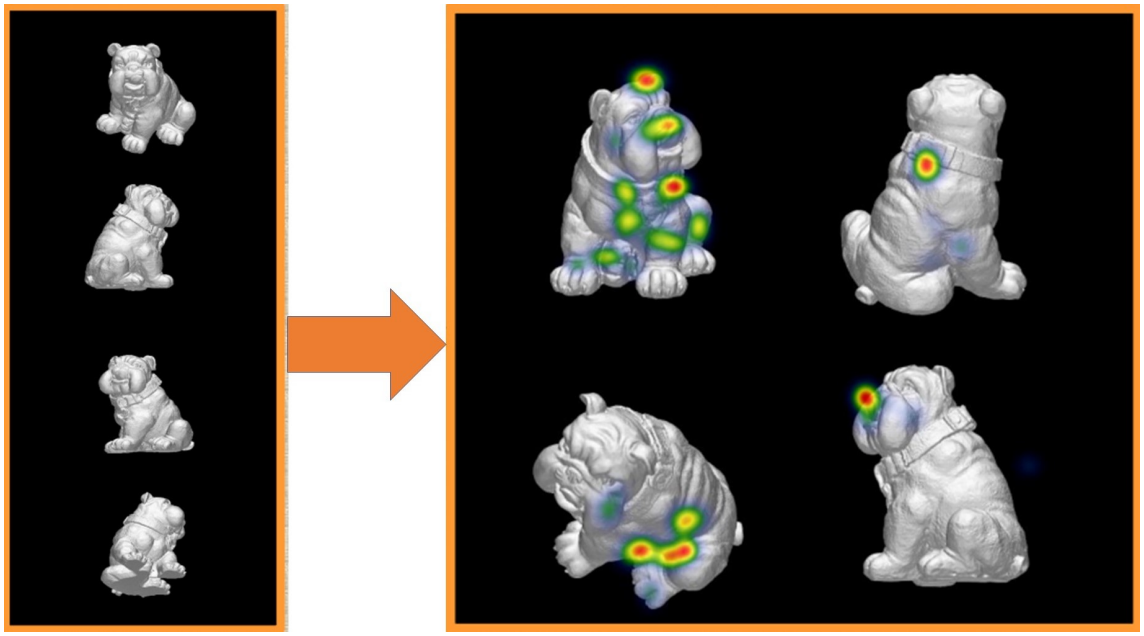


Figure 4.7: Examples of 4 views in each shape and the eye fixation of a participant.

4.6.1 Experimental Procedure

The experiments were administered within the School of Computer Science and Informatics, Cardiff University. Participants in our experiment participated voluntarily without being awarded monetary or other rewards. A consent form was given to the participants

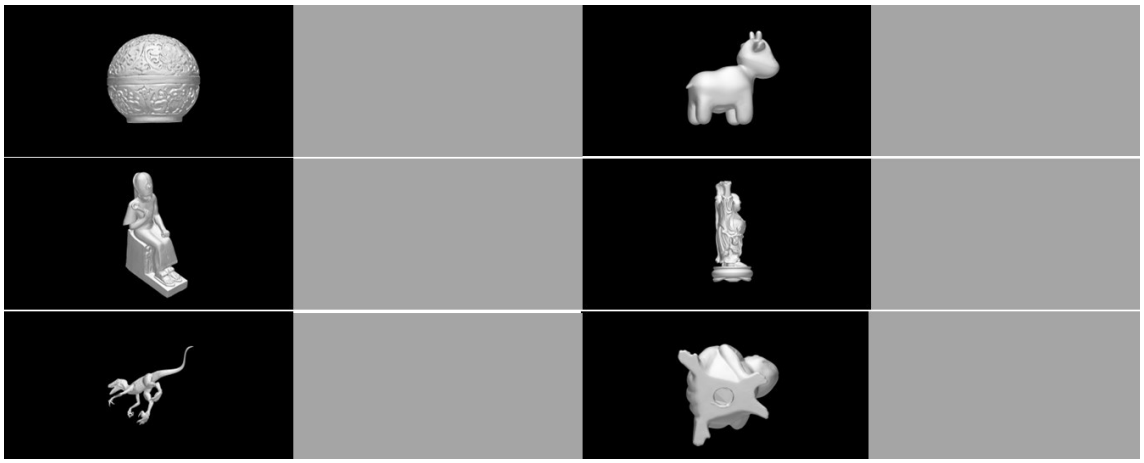


Figure 4.8: Example of how we run the experiment using eye tracker after each shape, we use the grey background to give a participant break.

before the experiment began. They were also informed that they could leave the experiment at any time and that they were not required to complete it. The names and gaze data of the participants were kept anonymous. The eye tracker we used to collect their gaze data was used during their participation. The experiment was carried out in a computer vision lab with an occasional reflective surface and constant close light. The viewing distance was maintained at around 60 cm.

The eye movements of the participants were measured employing a non-invasive SensoMotoric Instrument (SMI) Red-m eye-tracker operating at a rate of 250 Hz. Eye tracking manufacturers often report accuracy at $<0.5°$ [63], but the SMI device we use reports accuracy at $0.4°$. In order to evaluate the accuracy of existing methods using the data from the eye-tacking experiment, it is necessary to compare the saliency map from the existing work with the saliency map derived from the eye-tracking (ground truth) experiment which we normalise from 0 to 1. By the means of SMI's BeGaze™Analysis Software, gaze data was extracted from the raw eye-tracking data obtained throughout the experiment.

For every 3D mesh, this data contains the number of fixation points, and for each fixation point, its coordinates and duration. Fixation was strictly outlined by SMI's computer software using the distribution and duration-based formula established, with a minimum

period of 100 ms [226]. The mean duration of fixations μ_i for a subject i is:

$$\mu_i = \frac{1}{n}\sum_{j=1}^{n} x_j \tag{4.8}$$

where n is the total variety of fixations recorded over the 80 stimuli utilised in our study and x_j is the period of the fixation j.

Participants: Forty female and twenty male members from the School of Computer Science at Cardiff University, with ages in the range of 20 to 39, volunteered to participate.

Design and procedure: Participants were informed that their task in the experiment was to look at the region on the model that they thought was of most interest, and the participant was not allowed to move their head during the experiment so that if there was an issue we can pause the experiment and then recommence. The experiment took only ten minutes per participant.

Failure of eye tracking: During the experiment, we selected the normal vision participant, but we noticed some failures in capturing eyes during the experiment. We have found 5 participants who have stigmatised or tired eyes if they are focusing they lose consternation. We removed their data from our work to avoid any misleading results, as shown in Figure 4.9.

4.6.2 Ethical Approval

The Human Ethics Committee of the Cardiff University SREC reference: COMSC/Ethics/2019/212 has authorised this study. All information that is collected about the participants during the experiment of this research is kept strictly confidential. The information is stored securely through Cardiff One Drive, for a period of five years. We may share the data we collect with researchers at other institutions, but any information that leaves Cardiff University will have participant personal details removed. In any sort of output we might publish, we will not include information that will make it possible for other people to

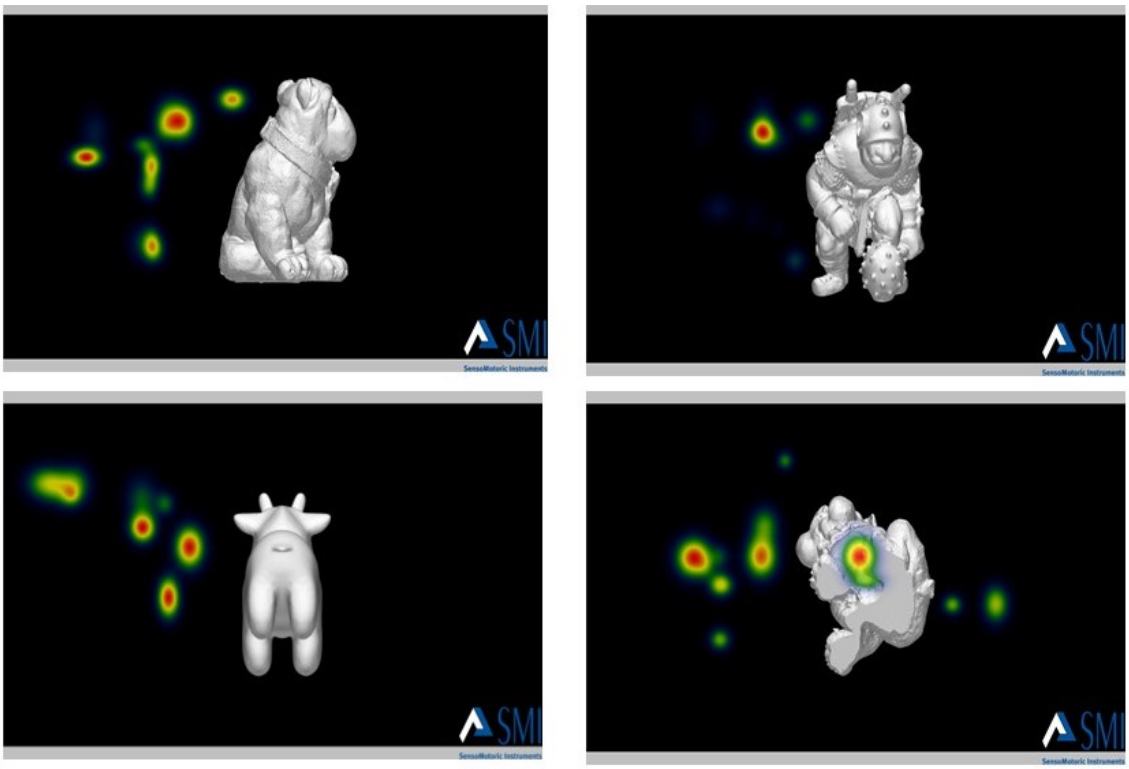


Figure 4.9: Example of a stigmatised failure to capture eye tracking.

know the participant's name or identify them in any way.

4.7 Obtaining Ground Truth and Evaluation of Existing Methods

In the following, we discuss how we work out the ground truth saliency map on a mesh M from subjective eye-tracking data (see Figure 4.10).

Let R_i ($i = 1, 2, \ldots, 20$) be the 20 rendered views of M. For each view R_i, the eye tracking data of all the users is collected and represented as a sequence of eye fixation points $(x_j^{(i)}, y_j^{(i)}, t_j^{(i)})$, where j is the fixation index, $(x_j^{(i)}, y_j^{(i)})$ are the coordinates of the fixation point in the image domain, and $t_j^{(i)}$ is the duration of the fixation. The duration of fixations in our experiment is a minimum of 100 and the maximum is 200 milliseconds However, since we are only focusing on one object, this period of time should avoid bias [226]. As mentioned in other studies [178], the duration should be between 100 and 300

ms depending on the task.

As fixation points tend to be sparse, following the common practice in image saliency research that applies Gaussian blurring to the fixation map to estimate the saliency map. To determine the most effective sigma value for Gaussian blurring in our saliency map generation method, we conducted experiments with and compare the performance of different sigma values such as 1.0, 1.5, 2.0 and 2.5 etc. With lower sigma, obtained saliency maps do not provide a good coverage for salient regions due to the discrete fixation points. On the other hand, higher sigma values such as 2.5 over-smooth the maps, leading to saliency regions covering not so significant part of the mesh. For our specific saliency map generation method, a sigma value of 1.5 is most effective. Compared to other tested sigma values, it provides a balanced trade-off between coverage and detail preservation. We map discrete fixation maps to meshes to obtain per-vertex saliency values as follows. We first map 2D fixation point $(x_j^{(i)}, y_j^{(i)})$ to the corresponding fixation vertex $v_j^{(i)}$ on the 3D mesh M. It iterates over each fixation in the experiment. Each fixation takes the vertex map corresponding to the 2D image fixation, takes the fixation x and y position in pixels finds the nearest coloured pixel in the vertex map and decodes the RGB value into a vertex index. Let $d_{\max}$ be the distance between the two farthest apart vertices on the mesh. Each vertex v in the neighbourhood $\mathcal{N}_j^{(i)}$ on the mesh M receives a saliency contribution from the fixation vertex $v_j^{(i)}$ according to the following formula:

$$s(v, v_j^{(i)}) = \exp\{-d(v, v_j^{(i)})/\bar{d}\} \cdot t_j^{(i)} \qquad (\forall v \in \mathcal{N}_j^{(i)}). \tag{4.9}$$

In practice, $\bar{d}$ is set to 0.05 times $d_{\max}$, and $\mathcal{N}_j^{(i)}$ is defined as those vertices v with distance to the fixation vertex $d(v, v_j^{(i)}) \leq \bar{d}$. This ensures each fixation point influences a reasonably sized neighbourhood, with the influence dropping where the distance from the fixation point increases. The distance measure $d(\cdot, \cdot)$ is ideally geodesic distances, although, in practice, Euclidean distance gives a decent approximation and is used in our experiments due to the relatively small neighbourhood size and shapes not having highly folded structures. Then, the contributions of all fixation points from the same view are

summed up to work out the saliency value for each vertex w.r.t. the given view $s_v^{(i)}$:

$$s_v^{(i)} = \sum_j s(v, v_j^{(i)}). \tag{4.10}$$

However, the saliency values for different views are not directly comparable. For example, if one view contains highly regions, e.g., faces, some potentially important but less significant regions, e.g. hands, may receive low saliency, whereas if the hands are seen without faces at the same time, they may be seen as highly salient in that particular view. Therefore, the relative importance of each vertex needs to be normalised when fusing inputs from different views. Let the rendered view $\mathcal{V}^{(i)}$ be the vertices that are visible from view R_i. We further introduce a weight w_i for the i-th view and use the commonly seen regions as anchors for normalisation, formulated as the following optimisation problem:

$$\min_{w_1,w_2,\ldots,w_{20}} \sum_{i_1,i_2\in\{1,2,\ldots,20\},i_1\neq i2} \sum_{v\in\mathcal{V}^{(i_1)}\bigcap\mathcal{V}^{(i_2)}} \left(w_{i_1} s_v^{(i_1)} - w_{i_2} s_v^{(i_2)}\right)^2, \tag{4.11}$$

where i_1 and i_2 iterate over all adjacent views (with at least one shared vertex). This ensures shared vertices across multiple views have saliency values as consistent as possible. To avoid getting trivial solutions with $w_1 = w_2 = \cdots = w_{20} = 0$, we additionally introduce a constraint:

$$\sum_i w_i = 1. \tag{4.12}$$

The above least-squares optimisation problem can be easily solved by solving a (small) linear system with the weights of individual views as unknowns. The final saliency value for s_v is obtained by averaging over values, linearly scaled to $[0, 1]$:

$$s_v = \frac{\sum_{i=1,v\in\mathcal{V}^{(i)}}^{20} w_i \cdot s_v^{(i)} - s_{\min}}{s_{\max} - s_{\min}}, \tag{4.13}$$

where $s_{\min}$ and $s_{\max}$ are the minimum and maximum values of s_v (before linear scaling).

Our collected ground truth saliency maps can be used to evaluate the effectiveness

of existing mesh saliency (MeshSIFT [49], SHOT [208], Gaussian curvature and Off-centre bias [21]) and existing 3D saliency models, namely Lee et al. [115, 184] and Song et al. [192] in a quantitative way. To evaluate the existing methods of measuring saliency, methods are required for comparing two saliency maps on the same mesh, i.e. the ground truth generated by the eye tracking experiment and the saliency map output by a saliency prediction method. The prediction method's performance is better if it has a closer distribution to the ground truth.

A basic measure for the similarity between these maps is Mean Square Error (MSE), which is 0 if they are a perfect match, and a high value if they are dissimilar. This measure is simple, but it only works well when the absolute salient values of two saliency maps are close. In practice, however, it is the relative importance which is more important. For instance, if one region is more important than another, it is hard to know how much the salience value of the first region should be larger than that of the second. To address this, we utilise the SSIM method (see Section 4.5), a measure widely used in image analysis and is known to be better correlated to perceptual similarity and less sensitive to absolute value differences. We extend the standard SSIM defined in the image domain to 3D mesh heat maps.

4.8 Learning New Methods of Measuring 3D Mesh Saliency

Once we obtained the ground truth now we used machine learning to learn a salience model based on a combination of geometric features (i.e. MeshSIFT [49], SHOT [208], Gaussian curvature and Off-centre bias [21]) and existing 3D saliency models, namely Lee et al. [115, 184] and Song et al. [192]. Existing methods for mesh saliency are largely based on handcrafted rules. In this work, we investigate using learning-based approaches to predict mesh saliency. To make this task feasible,we take features at each vertex as input and predict saliency values so that they are as close as possible to the ground truth saliency described in the previous Section 4.7.

Here, Off-center bias measures the Euclidean distance of a vertex position from the object's centre such that the further away from the centre, the more salient it is. This is intuitive as protrusions tend to have higher saliency values. We used a non-linear suppression operator similar to the one proposed by Itti et al. [94]. Let $\mathbf{f}_v = (f_{v,1}, f_{v,2}, \ldots, f_{v,N})$ be the feature vector containing both geometry-related and existing saliency estimation results for vertex v, where N is the total number of feature values for a vertex. Machine learning models are built using all the vertices of the meshes in the training set, and then we retain test set mesh vertices for testing purposes. For this purpose,

There are three separate built linear and nonlinear combination models that we used in this study to get more accurate in measuring 3D mesh saliency, i.e. Least Square Regression (LSR), Feed-forward Neural Network (FNN) and Support Vector Regression (SVR). The advantages of using these methods are discussed as follows. There are many reasons why least squares regression (LSR) is a simple algorithm to implement. It is also computationally efficient, which is very useful when dealing with large 3D meshes. Also, the coefficients of the least squares regression model can be interpreted directly, which could help us understand how different 3D mesh features contribute to its saliency. Moreover, least squares regression can provide a good fit to the data when the relationship between features and saliency is linear or approximately linear. Support Vector Regression (SVR) uses a subset of the training points to calculate the decision function, making it relatively robust to outliers. A nonlinear relationship between a feature and a target value can be modelled by SVR by using different kernel functions. It provides flexibility and potentially higher predictive accuracy for complex, nonlinear data relationships. Lastly, Feed-forward Neural Networks (FNNs) have only one direction of input to output. The difference between them and recurrent neural networks is that they do not have cycles or loops. Feedforward neural networks are easily parallelisable because each layer has independent nodes. It can be helpful when using hardware accelerators like GPUs.

4.8.1 Least Squares Regression

The first method is the Least Square Regression (LSR), a straightforward method for the data set that minimises each method's (MSE) against the ground truth by linearly scaling each method by weight for its data set. Then these weights can be applied to each of the test set models' saliency maps to produce a saliency map weighted by the learned model. Figure B4 in the Appendix show the flow of how LSR predicts the weight. SSIM can then evaluate the method's accuracy for the test set against ground truth. The least squares regression aims to work out the optimal per-feature weight ω_k and bias b such that the model best predicts the saliency values in the training set, i.e.

$$\min \sum_{v} \left(\sum_{k=1}^{N} \omega_k f_{v,k} + b - s_v \right)^2 . \tag{4.14}$$

4.8.2 Support Vector Regression

Support Vector Regression (SVR) is a type of Support Vector Machine (SVM) that is used for regression problems. For 3D mesh saliency, SVR could be used to learn a mapping between local geometric features and saliency values as shown in Appendix Figure B4. To calculate the local geometric features of each vertex in the mesh, there are several properties that are used, such as mean curvature, Gaussian curvature, and shape index. It is possible to represent each vertex's features as a vector. We have used the same process as least squares regression. SVR, the standard model, takes per-vertex features as input and predicts the saliency value for the vertex [237].

$$s = f(x) = \sum_{n} (\alpha_n - \alpha_n*)(x_n * x) + b, \tag{4.15}$$

where s represents the predicted saliency of a vertex. $f(x)$ represents the function learned by the SVR model. x represents the feature vector for a vertex, which includes its geometric properties. n is a sum over all the training vertices, α_n and α_n* are the Lagrange

multipliers. They are non-zero only for the support vectors. $(x_n * x)$ is the dot product of the nth training example x_n and the new feature vector x. b is the bias term, which adjusts the average predicted saliency. Using this equation, we calculate a weighted sum of the dot products between the new feature vector x and each of the training vertices x_n. Using the difference between the Lagrange multipliers, we can determine how much each training vertex influences the prediction.

4.8.3 Machine Learning based on Neural Networks

The third method is to use the Feed-forward Neural Network (FNN) of MATLAB in the toolbox of the neural network. The FNN is different from the least squares regression in a way that replaces the linear model with the neural network. This data will be trained by the FNN until it can no longer be improved. For the remaining of the selected models, the network will then simulate the output against the ground truth and SSIM will assess the similarity between the network and ground truth as shown in Appendix Figure B3.

In this network, three layers are used where the input layer contains N nodes corresponding to the input features, the hidden layer contains ten nodes, and the output layer contains one node corresponding to the predicted mesh saliency value [156]. The output is in only one direction, forward. The network has no phases or loops. The main purpose of using a feed-forward network is because this task has only one single layer perception forward from input nodes to hidden nodes.

Our dataset only contains a small number of 3D shapes, so learning at the shape level is not practical. Instead, our neural network is trained to predict mesh saliency at the vertex level, which is feasible. As saliency prediction based on vertex geometric features is relatively straightforward, we found that a shallow feedforward network with 3 layers achieves better performance than alternative architectures.

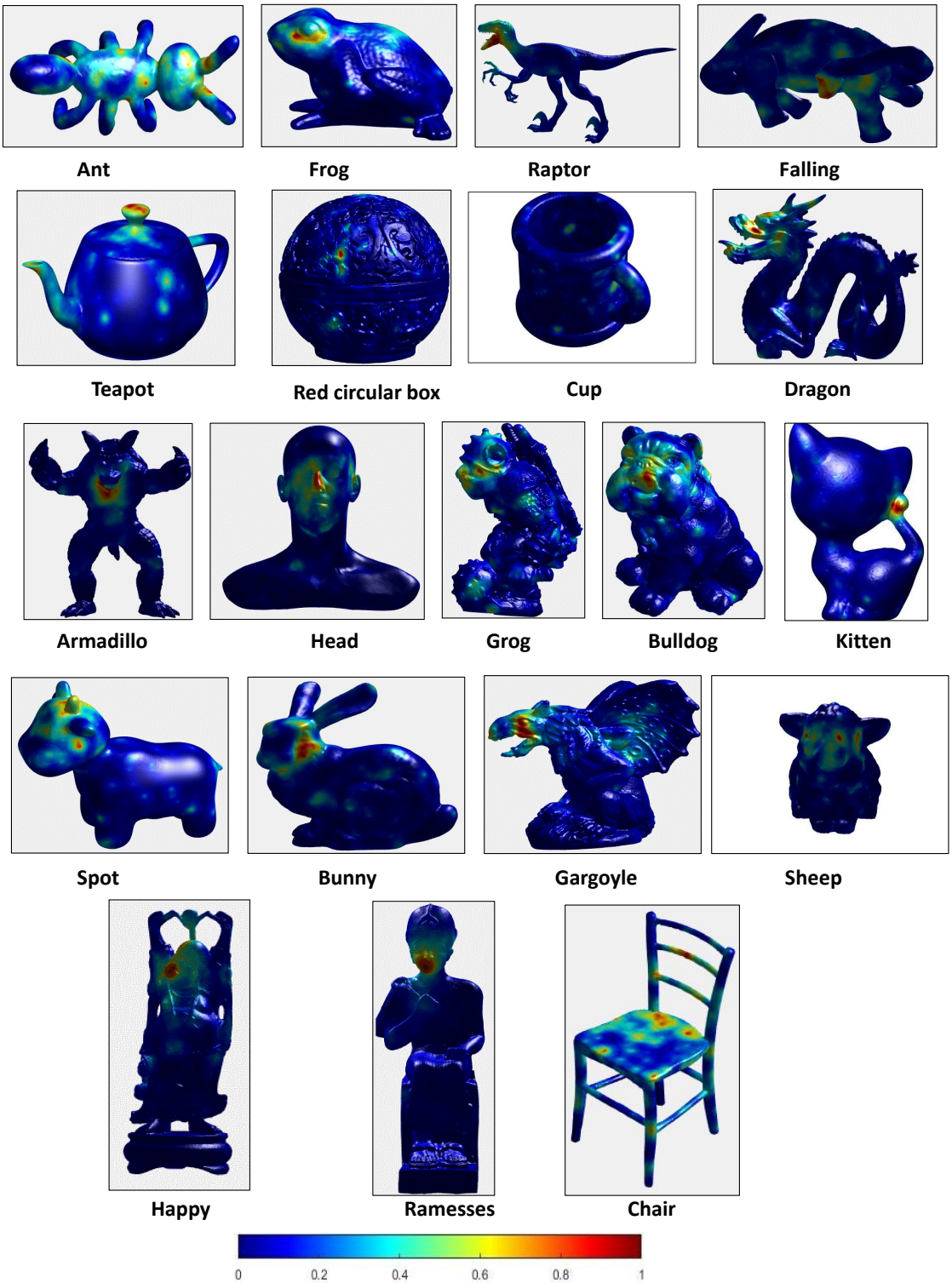


Figure 4.10: Examples of ground truth salient maps derived from eye tracking data; red and yellow are salient areas, while green and blue are non-salient areas. Source of the data present in the Appendix B.1

4.9 Results and Discussions

Figure 4.11 illustrates the saliency maps before and after normalisation. It is evident that normalising views was practical. Areas that garnered attention from multiple viewpoints, especially the faces shown in Figure 4.10, were accentuated due to participants focusing on them. However, participants did not focus on faces in some shapes like Armadillo and Kitten, directing their attention elsewhere. Compared to predominant areas like faces, these areas seem less emphasised which is a valuable result for future work.

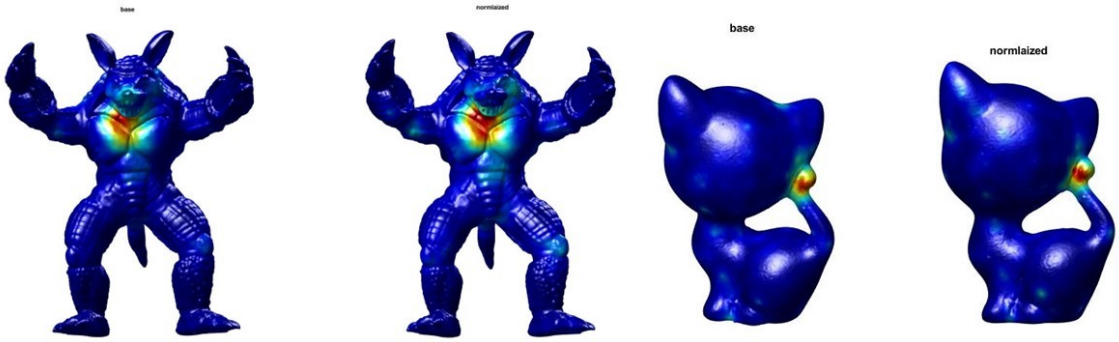


Figure 4.11: Examples of comparison of saliency maps before and after normalising views. From left to right Armadillo (before), Armadillo (after), Kitten (before) and Kitten (after)

4.9.1 Eye Tracking and Mesh Saliency Ground Truth Results

We now show some examples demonstrating the effectiveness of our fusion strategy for saliency from individual views. As shown in Figure 4.12, the initial eye-tracking data is captured on individual views, which are then fused using our method to produce a consistent saliency map on each mesh (with a normalised saliency value assigned to each vertex of the mesh). As shown, the fusion works well, with salient areas that get a lot of attention from multiple different views, like both faces in the examples in Figure 4.12. Regions which receive little attention from the participants correspond to those boring/less distinctive areas of the model, and they have low saliency values.

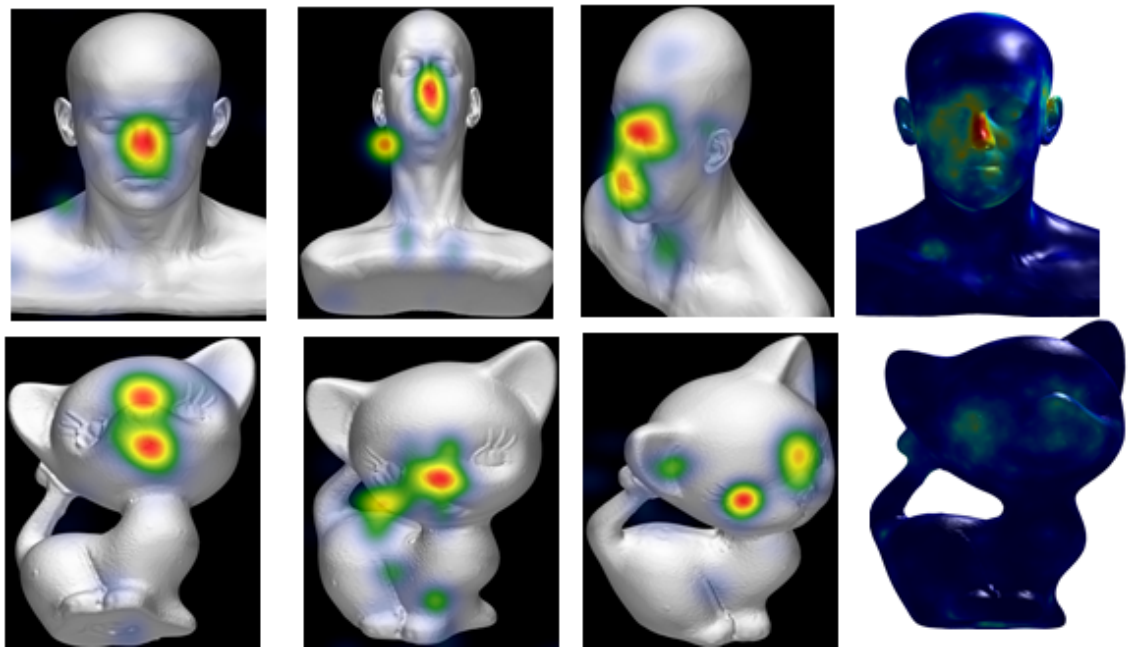


Figure 4.12: Examples of 2D fixation maps and the results of fusing them to form consistent saliency maps on 3D models; red and yellow are salient areas, while green and blue are non-salient areas.

More examples of ground truth saliency maps from eye tracking are shown in Figure 4.13. Similar trends are observed, although, for objects not including faces (e.g. the Falling shape), salient regions tend to be more flexible, and some regions on the body are also relatively salient (although less so than faces); see, e.g. the Gargoyle shape. Participants paid the most attention to models with visible facial features.

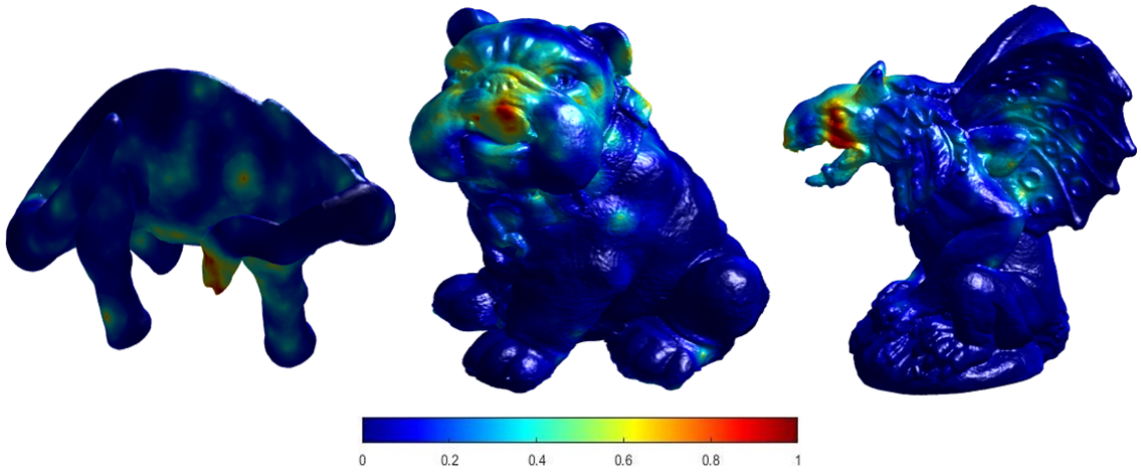


Figure 4.13: Examples of ground truth salient maps derived from eye tracking data; red and yellow are salient areas, while green and blue are non-salient areas. From left to right Falling, Bulldog and Gargoyle shapes.

As shown in Figure 4.13, the saliency map of the Falling mesh is far from the rest of the models, while other models show the head has a very high salience. Furthermore, this finding is to be predicted; very little else is known to be notable on models with faces. This might be because the saliency values around the face greatly outweigh any other values, as each participant will look at the face at some point. This means there is a minor salience quality when the saliency map is normalised 0 to 1 area that is not in the facial region.

One way of circumventing this would be to use a non-linear scaling to stop super salient areas from blocking out the rest of the dataset. The second possibility is that, obviously, people first look at the most salient area of an image. The most salient area in the case of these models might be the head. However, as participants only have 5 seconds to examine each image, they may not have time to look at other slightly less salient areas on the mesh. For example, the body or other parts. For this problem, it could be possible to reduce the exposure time of each object and assign a higher saliency weight to fixations at the beginning of a viewing, so that the first objects viewed would be given greater importance than the last point viewed.

4.9.2 Evaluation Results of Existing and Our Learning Methods

We now apply our evaluation methodology to existing mesh saliency methods and our learning-based methods. To ensure a fair comparison, in particular between existing methods and learning-based methods, we only report the average performance on the test set. For existing methods, we test representative methods Lee et al. [115], and Song et al. [192], and baseline methods Gaussian curvature and off-centre-bias. Quantitative evaluation on our eye tracking-based test set is reported in Table 4.1. As can be seen, Song et al.'s method achieves better performance than other existing and baseline methods, according to both SSIM and MSE. Other methods tend to perform similarly, with Lee et al.'s results better than the three baseline methods in both metrics.

The variants of learning-based methods perform better than existing methods, according to both SSIM and MSE metrics. We found that least squares regression outperformed more complicated methods, including feed-forward neural networks and SVR. This is probably because of the relatively limited data, and the simpler linear model avoids overfitting and generalizes better to unseen data. Our learning-based method achieves (0.906 SSIM and 0.004 MSE), which are significantly better than state-of-the-art methods (0.751 SSIM and 0.010 MSE) for Song et al [192].

In Figure 4.14, we show the visual comparison of different results of Ant shape, along with SSIM values. Moreover, we explain before how we generated the existing methods and geometric features to produce the result. This figure shows the output of these methods together. As can be seen, ground truth (captured using an eye tracker) is generally plausible, and learning-based methods, particularly those based on least-squares regression, predict saliency maps more similar to the ground truth. SSIM is a method used for 2D images. As described before, it is translated to the 3D meshes. Note, SSIM does not consider colour differences, only luminance, contrast and structure changes, it may not accurately reflect perceptual similarity, particularly for complex 3D objects. This means there could be cases where SSIM scores are acceptable, but visual quality is not expected.

Table 4.1: Average SSIM value and Mean Square Error (MSE) for each existing method and our learning-based method for evaluating the quality of predicted saliency maps against the ground truth derived from eye tracking. The only test set is used to ensure a fair comparison. For SSIM, larger is better, and for MSE, smaller is better.

	Models	SSIM	MSE
Existing models	Off-center bias	0.620	0.020
	Lee et al.	0.629	0.016
	Song et al.	0.751	0.010
	Gaussian curvature	0.616	0.022
	Mesh SIFT	0.720	0.012
	SHOT	0.620	0.016
Learnt models	Least squares regression (LSR)	0.906	0.004
	Feed-forward neural network (FNN)	0.895	0.006
	Support vector regression (SVR)	0.861	0.009

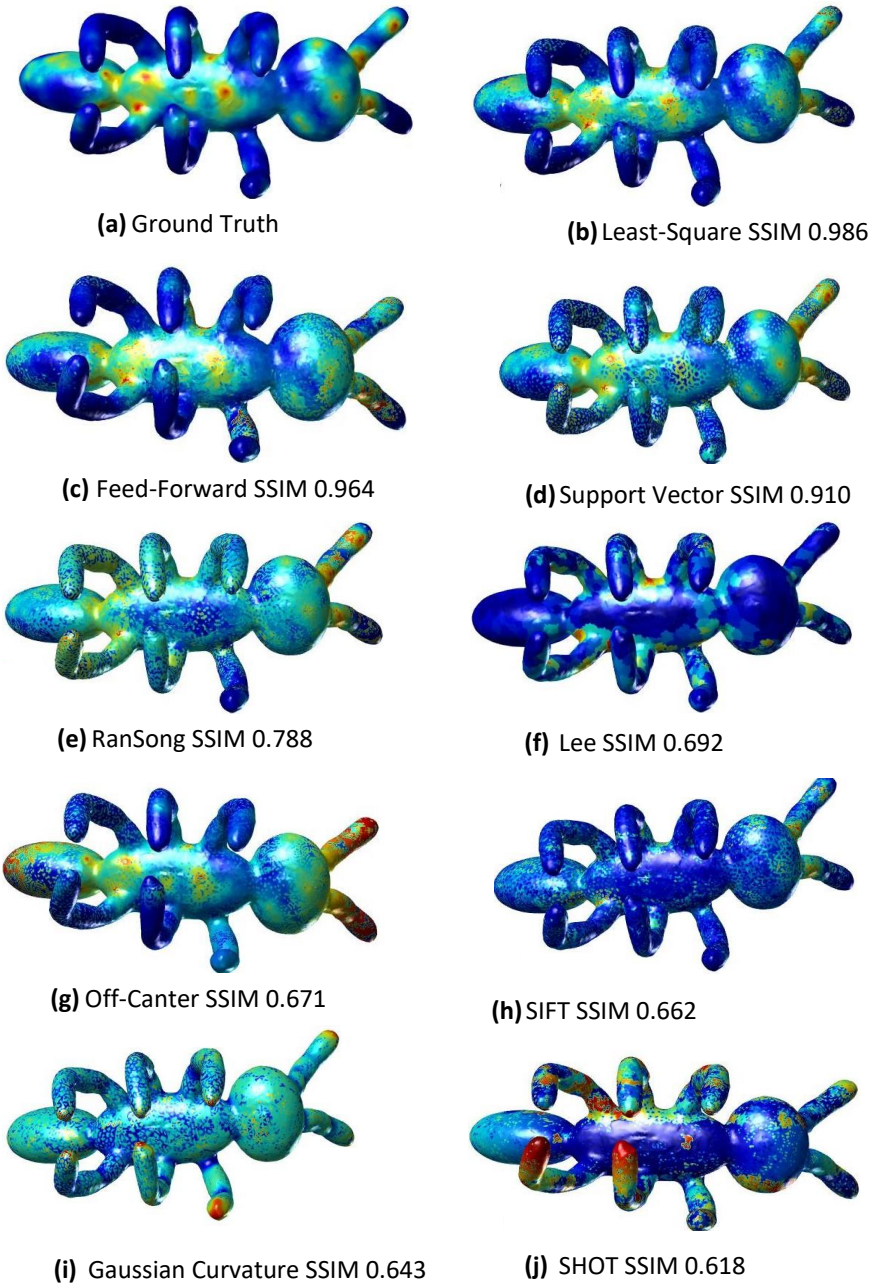


Figure 4.14: Example of Ant's saliency results: (a) ground truth, (b,c,d) our learning-based methods, (e,f) existing methods and (g-j) geometry feature-based baseline methods.

4.10 Summary

Estimating saliency on meshes is a fundamental tool that benefits many downstream applications. Existing methods largely focus on developing dedicated formulas to achieve this, but it is difficult to fully capture perceptual importance using these methods. This chapter investigates a methodology to produce ground truth saliency maps on meshes using eye-tracking data. In particular, we fuse saliency maps from individual views to produce a single consistent saliency map for a given mesh. We further develop learning-based methods that take existing saliency prediction results and geometric features at each vertex as input to predict the local saliency value. Qualitative and quantitative results show that our learning-based methods, particularly the model based on least squares regression, outperform state-of-the-art methods. In future work, we would like to build a larger dataset and evaluate the effectiveness of more machine learning methods, including methods based on deep neural networks.

Chapter 5

Objective Quality Assessment Measures for Mesh Quality

Overview

Chapter 3 presented our methodology on 3D mesh quality scores using a subjective study while wearing a VR device. This chapter focuses on objective quality measures used in a traditional normal display desktop setting and compares them with our VR subjective study. The objective study in this chapter compares different quality assessment methods, which are methods focused on measuring the dissimilarity between two meshes by calculating the geometry distance between them. We evaluate the state-of-the-art mesh perceptual difference metrics for predicting the objective quality scores captured in the VR setting and compare these with the desktop setting. Since salient regions are more important than others and can have a more significant impact on perceptual quality, we further propose a new mesh perceptual difference measure incorporating mesh saliency.

After a brief introduction, in Section 5.1 we provide an overview of perceptual objective quality measures on 3D meshes and summarise state-of-the-art objective quality methods in Section 5.2. In Section 5.3, we present the details of comparing the subjective and objective quality assessment studies. Section 5.4 and Section 5.5 present our data and results of the proposed objective quality metric on both settings. Finally, Section 5.7 presents the conclusion of our work and future work.

5.1 Introduction

Many applications including Virtual Reality benefit from 3D object processing, transmission and visualisation. When dealing with 3D models (often represented by polygonal meshes), several processing methods, such as filtering, denoising, simplification, watermarking and compression, are routinely used. These algorithms have different objectives, but the way they modify the mesh's visual appearance is an important consideration. A watermarking scheme seeks maximum robustness while keeping the geometric modification as unnoticeable as possible; similarly, a compression or simplification algorithm seeks minimum streaming size or triangle number while keeping the visual difference with the original mesh as slight as possible. Denoising or filtering algorithms are used to enhance the model's quality while keeping its original form.

As we discussed earlier, these algorithms generally use geometric distances (e.g., Euclidean vertex-to-vertex distances, Hausdorff distances) that do not capture visual quality or the perception of the difference between two 3D models. It may be possible to capture the perception of visuals by conducting subjective experiments in which human observers provide their opinions about the processed models. The problem with such subjective evaluations is that they are both time-consuming and expensive and cannot be incorporated into automatic systems. Several perceptual-based objective quality metrics have been proposed in the computer graphics community to better reflect differences of 3D objects in visual perception.

Subjective research and objective measurements can be used to assess perceptual quality. A subjective experiment based on human judgment of a set of distorted objects is used to analyse and compare with geometric measures. Objective metrics are methods that are supposed to predict/or measure visual quality loss. Most present processing techniques (simplification, watermarking, and compression) are driven and/or assessed by basic metrics such as Hausdorff distance (HD) and root mean square error (RMS), but these are not consistent with human vision, which is a significant issue. In this regard, we first compare

the VR and desktop settings using perceptual-based metrics such as mesh structural distortion measure (MSDM) and multi-scale mesh structural distortion measure (MSDM2) [108], and compare their prediction quality for VR and desktop settings. We further propose a new metric that incorporates mesh saliency.

5.2 Mesh Difference Metrics

Several objective quality criteria for 3D models have been established, inspired by the large quantity of past work on image and video quality evaluation. These are typically full-reference comparing the distorted model to its original/reference (as we discussed in Chapter 3) and employ the traditional technique used in image quality assessment: vertex-level local feature differences are determined, which are then pooled across the whole 3D model to generate a global quality score.

In the following sections, we provide a brief explanation of the metrics that were examined. Two traditional geometric measurements are HD and RMSE. Karni and Gotsman [99] and Sorkine et al. [195] proposed two new RMSE combinations and the GL, which we have also incorporated. Finally, we have considered four of the more recent model-based perceptual measures: Lavoué et al. mesh structural distortion measure [111], Corsini and Drelie Gelasca et al. [43] roughness-based metrics. We do not use perceptual image-based metrics in our studies since they are less reliable for predicting the perceived visual impairment on 3D models, as mentioned at the start of Section 2.6.2.

Hausdorff Distance (HD) in 3D space, the distance between a point p and an object A is represented by the calculation $e(p, A)$:

$$e(p, A) = \min_{v_i^A \in A} d\left(p, v_i^A\right) \tag{5.1}$$

This equation demonstrates the relationship between the ith on vertex object A and

the d Euclidean distance. The following calculation is the asymmetric Hausdorff Distance between two objects, A and B.

$$H_a(A,B) = \max_{v_i^A \in A} e\left(v_i^A, B\right). \tag{5.2}$$

After that, the symmetric Hausdorff Distance is defined as follows:

$$H_d(A,B) = \max\left\{H_a(A,B), H_a(B,A)\right\}. \tag{5.3}$$

Root Mean Square Error (RMS) is based on the assumption that correspondence vertices between the two objects are available. Because of this, it can only be used to compare two meshes with the same connectivity. In formula

$$RMS(A,B) = \left(\frac{1}{n}\sum_{i=1}^{n}\left\|v_i^A - v_i^B\right\|^2\right)^{1/2} \tag{5.4}$$

Where n is the number of vertices in each mesh and v_i^B is the vertex of B matches v_i^A of A.

Geometric Laplacian Measures (GL1) and (GL2): Karni and Gotsman [99] invented the GL. It is based on a measure of the vertices' smoothness. In particular, given a vertex v.

$$GL(v) = v - \frac{\sum_{i\in n(v)} l_i^{-1} v_i}{\sum_{i\in n(v)} l_i^{-1}} \tag{5.5}$$

where l_i is the Euclidean distance from v to v_i and $n(v)$ is the set of indices for v's neighbours. After a Laplacian smoothing step, $GL(v)$ shows the difference and its new position. Taking into account Eq. 5.5, Karni and Gotsman [99] has created a visual metric GL_1 between two objects A and B defined as

$$GL(A,B) = \alpha\, RMS(A,B) + (1-\alpha)\left(\frac{1}{n}\sum_{i=1}^{n}\left\|GL\left(v_i^A\right) - GL\left(v_i^B\right)\right\|^2\right)^{1/2} \tag{5.6}$$

Following previous work, two settings are considered: GL_1 where $\alpha = 0.5$, and GL_2

where $\alpha = 0.15$. The latter puts more emphasis on geometric Laplacian differences.

Mesh Structural Distortion Measure (MSDM): Lavoué et al.'s[111] MSDM measure is based on the idea of structural similarity for evaluating the quality of 2D images by Wang et al. [223]. It is calculated in a bottom-up approach, where local distortion measure is first calculated. The local neighbourhood around vertex v is defined as vertices within the sphere centred at v with radius r (set to 0.5 % of the bounding box length of the mesh). It is likely that some edges intersect with the sphere, so to improve accuracy, especially with poor tessellation, edge points are also added, which are defined as intersection points of edges with the sphere. Curvatures at the edge points are obtained by a simple linear interpolation of curvatures at the endpoints of edges.

$LMSDM$ defines how to measure the distance between two local mesh windows a and b:

$$LMSDM(a,b) = \left(0.4 \times L(a,b)^3 + 0.4 \times C(a,b)^3 + 0.2 \times S(a,b)^3\right)^{\frac{1}{3}}. \tag{5.7}$$

where

$$\begin{aligned} L(a,b) &= \frac{\|\mu_a - \mu_b\|}{\max(\mu_a, \mu_b)} \\ C(a,b) &= \frac{\|\sigma_a - \sigma_b\|}{\max(\sigma_a, \sigma_b)} \\ S(a,b) &= \frac{\|\sigma_a\sigma_b - \sigma_{ab}\|}{\sigma_a\sigma_b} \end{aligned} \tag{5.8}$$

which are defined using the mean μ and standard deviation σ of curvatures within local regions a and b respectively. These distance functions have a solid intuitive relationship to psychovisual ideas and several previous efforts on 3D perception. A normalised curvature distance is denoted by L. The curvature distance, which Kim et al. [100] also discuss, is inherently tied to normal directions, which drive rendering and hence the visual appearance of the 3D object. The standard deviations, indicating the roughness of the surfaces, are used to calculate C. Several researchers emphasised the importance of roughness dis-

tance for perceptual measurements [99, 43]. Finally, S, like Howlett et al. [89] and Lee et al. [115], seeks to identify changes in key features by evaluating the covariance between the local windows.

The psychovisual research community [58] acknowledges that there are three main significant groups of regions in an image or a 3D object edge, textured or smooth. These categories are related to the notion of masking; textured (or rough) parts have a high degree of masking, but geometric changes on edges or smooth regions are considerably more evident. The suggested metric captures these behaviours: The C coefficient will indicate a geometric change on a smooth zone that changes the roughness degree, but the structural S coefficient will highlight a change on an edge region.

A Minkowski sum of the n_w local window distances between two meshes A and B determines their global measure:

$$MSDM(A,B) = \left(\frac{1}{n_w} \sum_{j=1}^{n_w} LMSDM(a_j, b_j)^3 \right)^{1/3} \in [0,1) \tag{5.9}$$

n_w is the total number of local mesh windows, and b_j is B's local window that corresponds to A's window of a_j. Practically, this measure is asymmetric and considers one local window per vertex of the original mesh. When the measured objects are visually significantly different from others, their value goes toward 1 (the theoretical limit), whereas it is equal to 0 for identical ones.

Multi-scale Mesh Structural Distortion Measure Method (MSDM2). Similar to MSDM, this technique is primarily influenced by Wang et al. [223] 2D image SSIM metric, which argues that the human visual system is well-suited to retrieving structural information. As a result, this metric is based on structural differences (as measured by curvature statistics) derived from corresponding local regions from the meshes being compared. Compared with MSDM, MSDM2 is calculated in a multiscale manner. As Zhu et al. [253] have demonstrated, the perceptibility of distortion on a 3D object is dependent on its level of detail and viewing conditions (e.g. display resolution and viewing

distance). As a result, a single-scale technique may be applicable only in certain circumstances. As a result, it develops a multi-scale distortion measure, similar to what Lee et al. [115] proposed for their saliency model, to capture distortion at all perceptually significant scales, improving efficiency and robustness. Different scales are implemented through using different radii of local neighbourhoods.

The visual distortion measure is computed as follows. Given a distorted mesh M_d and the associated reference (i.e. original) mesh M_r, scale-dependent curvatures are computed on vertices from both meshes, using fast projection and barycentric interpolation, and each vertex of the distorted mesh M_d is matched with its corresponding 3D point and curvature value from the reference mesh M_r. A local distortion measure is calculated for each vertex of M_d as the difference of Gaussian-weighted statistics obtained across a local spherical region of radius r. The global multiscale distortion score is then calculated using Minkowski pooling to combine the local values. The above steps are repeated at multiple scales, resulting in several distortion maps, and the final distortion map is created by combining all of the local distortion maps.

The MSDM2 method starts with a fast asymmetric match between the distorted object M_d and the original object M_r, then computes Gaussian-weighted curvature statistics at multiple scales over local windows for each vertex to produce a local distortion map pooled into a single global multiscale distortion score (GMD) [111]. The final metric is obtained by averaging forward ($M_d \longrightarrow M_r$) and backward ($M_r \longrightarrow M_d$) global distortion scores. In practice, three scales are used, where $r = 2\varepsilon, 3\varepsilon, 4\varepsilon$, where ε is 0.5% of the model's bounding box's maximum length. The key differences between the planned MSDM2 and its predecessor MSDM [111] are as follows: The curvature scale size (improving robustness). MSDM requires an implicit vertex-vertex correlation, but MSDM2 does a rapid projection and curvature interpolation. It suggests that there is no connection limitation, which increases the quality of the matching. Curvature statistics have been mostly improved; moreover, they have been standardised using Gaussian weighting methods, and their combination has been modified. The connection with human judgement has

improved significantly.

Direct Reconstruction of 3D Worker Pose (3DWPM1 and 3DWPM2): Taking into account that visual artefacts caused by watermarking can be measured by the amount of roughness introduced on the surface, Corsini et al. [43] proposed two perceptual metrics for the quality evaluation of watermarking algorithms. The watermarking visual impairment is evaluated by considering the increment of roughness between the original model and the watermarked model in the following way:

$$3DWPM(A,B) = \log\left(\frac{\rho(B) - \rho(A)}{\rho(A)} + k\right) - \log(k) \tag{5.10}$$

There are two total roughness values: $\rho(A)$ represents the roughness of the original model and $\rho(B)$ represents the roughness of the watermarked model. In order to avoid numerical instability, the constant k is used. There are two ways to measure the roughness of a model.

The first roughness measure 3DWPM1 is a variant of the method by Wu et al. [229]. Dihedral angles, the angle between two adjacent normals, are used in this metric to measure per-face roughness. The dihedral angle [43] is related to surface roughness. As a result, the dihedral angles between adjacent faces of a smooth surface are close to zero because the face normals vary slowly over the surface. An evaluation of roughness takes into account the scale of the roughness by converting the average roughness per face into an average roughness per vertex and considering rings of varying sizes (1-ring, 2-ring, etc.). The roughness of the 3D object is the sum of the roughness of all vertices. The second method 3DWPM2 is by [43], artefacts are more easily perceived on smooth surfaces. A smoothing algorithm is then applied to the surface and the roughness is measured as the variance between the smoothed version and the original.

5.3 Comparison of Objective Quality Measures for VR and Desktop

To understand the relationship between the subjective and objective quality measures, we apply existing geometric and perceptually inspired geometry-based 3D metrics for triangular meshes, including HD, RMS, as well as GL_1 [99] and GL_2 [195] which combine RMS with Laplacian coordinates, roughness-based $3DWPM_1$ and $3DWPM_2$ [44], as well as MSDM and MSDM2 [108]. As mesh difference measures tend to be at significantly different scales, comparing their absolute values is not meaningful. Therefore, we use Pearson Linear Correlation Coefficient (PLCC) and Spearman Rank Order Correlation Coefficient (SROCC) to evaluate these methods. For MSDM and MSDM2, we use author provided code in their MePP/MePP2 projects. MeshLab/Metro is used to calculated HD, and we re-implemented other metrics.

Table 5.1: Pearson (PLCC) & Spearman (SROCC) correlations value (%) between Mean Opinion Scores in VR and desktop settings, and values from the objective mesh quality metrics for the General-purpose dataset.

Metric	Armadillo		Venus		Dyno		Rocker Arm		All	
	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC
Hausdorff	30.2	69.5	0.8	1.6	22.6	30.9	5.5	18.1	1.3	13.8
VR-Hausdorff	30.5	32.5	18.7	6.1	26.7	2.0	31.6	12.5	13.2	1.9
RMS	32.2	62.7	77.3	90.1	0.0	0.3	3.0	7.3	7.9	26.8
VR-RMS	21.7	11.2	76.7	68.2	22.4	34.1	3.3	14.7	25.1	21.8
GL_1	43.7	70.2	80.2	92.0	3.2	15.5	8.4	14.2	12.6	33.1
VR-GL_1	37.1	25.5	79.1	67.8	36.7	45.6	14.6	19.7	34.4	26.2
GL_2	55.5	77.8	77.6	91.0	12.5	30.6	17.1	29.0	18.0	39.3
VR-GL_2	52.9	37.0	78.5	63.9	53.0	54.3	28.3	30.5	43.3	30.7
$3DWPM_1$	35.7	65.8	46.6	71.6	35.7	62.7	53.2	87.5	38.3	69.3
VR-3DWPM$_1$	70.0	65.1	54.5	31.6	66.9	43.2	76.8	56.4	60.8	46.6
$3DWPM_2$	43.1	74.1	16.4	34.8	19.9	52.4	29.9	37.8	24.6	49.0
VR-3DWPM$_2$	50.4	64.6	28.5	22.9	44.9	46.2	53.8	54.0	40.9	45.5
MSDM	70.0	84.8	72.3	87.6	56.8	73.0	75.0	89.8	56.4	73.9
VR-MSDM	74.7	70.4	86.6	87.0	76.5	60.3	70.7	44.5	77.1	65.6
MSDM2	72.8	81.6	76.5	89.3	73.5	85.9	76.1	89.6	66.2	80.4
VR-MSDM2	93.5	91.7	84.5	82.1	75.6	58.6	76.9	51.5	76.8	68.5

5.4 Data Analysis

We completed the experiment and can now collect the MOS, but before we do any data analysis, we tested the participants' performance in order to ensure the data collected is meaningful. We follow the ITU-R BT.500-13 recommendation [179] and show the participants a trailer with a different dataset to ensure they understand how the experiment works. To calculate the Interquartile Range (IQR) [179] of our data, we must first identify outliers. The first quartile (Q1), median, and third quartile (Q3) are identified. As a result, we compute IQR = Q3 - Q1. We then calculate the maximum value as Q3 + ($+1.5\times$ IQR). Finally, compute the lower value (Q1 - ($+1.5\times$ IQR). One outlier was discovered in both settings and was removed from the data set.

To analyse user ratings, a common method is to compute the mean opinion score (MOS) for each stimulus.

$$MOS_e = \frac{1}{10 \times N} \sum_{i=1}^{N} s_{ie}, \tag{5.11}$$

where s_{ie} refers to the score assigned by participant i to the stimulus e, and N denotes the number of (valid) subjects. We further divided the scores by 10 to normalise them in the range of $[0, 1]$. We follow most of the existing work [10, 35] and set the scores such that 0 means the worst quality, and 1 is the best quality. So we expect the MOS to decrease as the distortion level increases.

5.5 Comparison of MOS Scores for VR and Desktop

As shown in Table 5.1, we compare results between Mean Opinion Scores (MOS) in VR and desktop, with mesh differences predicted by objective measures. To make results easier to read, we group them in pairs, showing the comparative performance with desktop and VR MOS scores. Since these methods predict mesh difference (i.e., smaller means better) whereas MOS scores are designed such that larger values mean better quality, we

report correlations with the negative sign dropped (equivalent to the correlations between MOS scores and negated mesh difference values). The "All" performance is obtained by analysing predictions of all shapes together as shown in Table 5.1. PLCC for "All" is defined as follows

$$All = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{\sum(x - \bar{x})^2 \sum(y - \bar{y})^2}} \tag{5.12}$$

where $\bar{x}, \bar{y}$ are the sample means for (x) and and (y). So, if the value is close to +1, it indicates a strong positive correlation; if the value is close to -1, it shows a strong negative correlation. SROCC is similarly calculated on all the shapes together.

The overall trend of the performance of different methods is consistent between desktop and VR: Purely geometric metrics such as Hausdorff and RMS perform poorly in both cases. GL_1 and GL_2 metrics work slightly better, but it remains a simple geometric metric, so there is a big gap between the performance of GL_1/GL_2 and other metrics which are more perceptually inspired, including 3DWPM$_1$, 3DWPM$_2$, MSDM and MSDM2. In both desktop and VR settings, 3DWPM$_1$ shows better prediction capability than 3DWPM$_2$, in both PLCC and SROCC for most shapes and overall. In comparison, MSDM/MSDM2 perform consistently better than other analysed metrics. The gap between MSDM and MSDM2 is smaller for the VR setting compared with the desktop setting. In the VR setting, MSDM2 with multiscale analysis achieves lower PLCC than MSDM (by 0.3%) but higher SROCC (by 2.9%), whereas in the desktop setting, MSDM2 is better than MSDM by a large margin (9.8% for PLCC and 6.5% for SROCC overall). Nevertheless, MSDM2 remains the best-performing metric overall for VR quality prediction. We summerise the result in the table Table 5.1 :

- There is consistency in the performance trends of the different methods between desktop and VR settings.
- Pure geometric metrics like Hausdorff and RMS deliver poor results in both environments.
- GL_1 and GL_2 metrics, though a bit better, are still simple geometric metrics and

thus underperform compared to more perceptually-inspired metrics.

- Metrics such as $3DWPM_1$, $3DWPM_2$, MSDM, and MSDM2, which are inspired by perceptual cues, perform significantly better.
- In both VR and desktop settings, $3DWPM_1$ predicts better than $3DWPM_2$ in terms of both PLCC and SROCC for most shapes overall.
- MSDM and MSDM2 consistently outperform all other metrics that were analysed.
- The performance gap between MSDM and MSDM2 is smaller in the VR setting compared to the desktop setting.
- In the VR setting, MSDM2 achieves a slightly lower PLCC but higher SROCC compared to MSDM.
- In the desktop setting, MSDM2 surpasses MSDM by a significant margin, 9.8% for PLCC and 6.5% for SROCC.
- Despite variations in different settings and metrics, MSDM2 proves to be the best-performing metric for predicting VR quality overall.

Another interesting observation is that depending on the objects, comparing VR and desktop correlations, PLCC correlations are higher for VR compared with desktop, and SROCC correlations are higher for desktop than VR. As a large number of objective quality metrics are evaluated, this behaviour is likely due to the distribution of subjective scores. The majority of these metrics show a linear relationship with the amount of distortion, which suggests that under VR settings, participants are more capable of detecting geometric distortions. This leads to a more linear relationship between Mean Opinion Scores (MOS) and the level of distortions. In contrast, despite a consistent ranking in desktop settings, the relationship between MOS scores and geometric distortions tends to be less linear, indicating a more complex or non-linear interaction between these variables.

5.6 Saliency-Weighted Objective Quality Metrics

Inspired by the image-based IQA metrics, salient regions are more important for human perception, and as a result, it can be hypothesised that these regions should have a more significant importance when measuring visual quality (difference). Since MSDM2 performs best in the previous study, we incorporate saliency measures as weights to combine local MSDM (LMSDM) to form the global MSDM values. We define the saliency weight MSDM2 below:

$$SalMSDM2(A,B) = \left(\frac{1}{\sum_j \hat{s}_j} \sum_{j=1}^{n_w} \hat{s}_j LMSDM(a_j, b_j)^3 \right)^{1/3} \tag{5.13}$$

where n_w is the number of neighbourhoods, in practice, these can be the same as the number of vertices. $\hat{s}_j$ is the adjusted saliency value at j-th vertex, calculated as

$$\hat{s}_j = \gamma s_j + (1-\gamma), \tag{5.14}$$

where s_j is the saliency value for the j-th vertex, and γ is a weight to control the contribution of salient vertices. $\gamma \in [0,1]$. Setting $\gamma = 0$ reduces to the traditional non-salient MSDM2. Setting $\gamma = 1$ means non-salient regions (with $s_j = 0$) will have no contributions. In practice, we find that $\gamma = 1$ works well and is set as our default (but this may depend on particular saliency maps/measures).

Table 5.2 shows the results comparing MSDM2 and saliency weight MSDM2 in predicting mesh quality on our VR perceptual dataset. Existing mesh saliency measures Lee [115], Song [193], and our learning-based least squares regression model LSR (see Chapter 4 for details). Overall, our MSDM2-LSR improves PLCC by 1.1% and SROCC by 1.0% compared with MSDM2 in terms of overall performance. Given the already decent performance of MSDM2, this demonstrates the effectiveness of saliency-weighting for objective mesh quality prediction. Our saliency-weighted MSDM2 also achieves comparable or better PLCC and SROCC correlations for individual shapes, compared with

MSDM2, showing its robust performance.

Table 5.2: Pearson (PLCC) & Spearman (SROCC) correlations value (%) between Mean Opinion Scores in VR setting and MSDM2 along with different saliency weighting: Lee [115], Song [193], and our least squares regression model, denoted as LSR (see Chapter 4). We also compare with saliency-weighted MSDM2 based on ground truth saliency for the Armadillo model (as it is the only shared model used in our saliency subjective study (see Chapter 4), denoted as MSDM2-GT.

Metric	Armadillo		Venus		Dyno		Rocker Arm		All	
	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC
MSDM2	93.5	91.7	84.5	82.1	75.6	58.6	76.9	51.5	76.8	68.5
MSDM2-Lee	95.5	95.1	83.6	80.4	74.9	58.8	75.5	50.9	76.0	67.9
MSDM2-Song	93.5	91.7	84.5	82.1	75.6	58.6	76.9	51.5	75.9	69.6
MSDM2-LSR	93.5	92.0	84.6	82.7	75.8	58.6	77.3	52.4	77.9	69.5
MSDM2-GT	94.4	94.8	-	-	-	-	-	-	-	-

In comparison, when adding saliency weighting using other saliency methods, the overall performance is not improved, except for MSDM2-Song which achieves 69.6% SROCC, compared with 68.5% obtained by MSDM2. Nevertheless, for individual shapes, these can still be effectively. One of the examples is Armadillo, where MSDM2-Lee performs particularly well, achieving 95.5% PLCC and 95.1% SROCC, compared to 93.5% PLCC and 91.7% SROCC for MSDM2. This indicates that individual saliency methods may be advantageous for specific inputs, and there may be scope for improving saliency measures for specific applications such as mesh quality prediction.

Out of the 4 shapes used in the LIRIS/EPFL General-Purpose dataset, the Armadillo shape is also included in our eye-tracking-based subjective mesh saliency study (see Chapter 4 for details), so we also report the performance when ground truth (GT) saliency is used in SalMSDM2. As can be seen, ground truth saliency is effective, achieving 94.4% PLCC and 94.8% SROCC, compared to 93.5% PLCC and 91.7% SROCC for MSDM2, better than MSDM2-Song and MSDM2-LSR. It is interesting to note that, MSDM2-Lee in this case actually outperforms MSDM2-GT. This shows that an effective predicted saliency map can be equally as good as ground truth saliency maps, for helping with mesh quality prediction. Nevertheless, it can still be difficult to ensure consistent quality for all kinds of input shapes. It is worth noting that although the Armadillo model is in-

cluded in our subjective saliency dataset, but it is only used for testing so our least-squares regression model does not have the unfair advantage.

5.7 Summary

This chapter evaluates different geometric and perceptually inspired metrics for measuring the dissimilarity of shapes. We compare their performance with subjective MOS scores obtained from VR and desktop settings. The overall trend remains consistent, and perceptually inspired metrics work consistently better than geometric metrics. In particular, MSDM/MSDM2 perform better than alternative metrics for predicting perceptual mesh quality under the VR setting. Comparing VR with desktop, it was found that VR MOS scores tend to preserve better linearity in terms of geometric distortions. To further improve the mesh quality prediction, we proposed a saliency-weighted MSDM2 and demonstrated that using our least-squares regression-based saliency maps, the method improves both PLCC and SROCC correlations consistently. We present our work that the visual saliency can be used to improve Mesh Quality Assessment in Chapters 4 and 5. Such techniques can be useful for real-time VR applications where geometry can be selectively simplified during streaming.

Chapter 6

Conclusion & Future Work

Overview

This thesis aims to measure the 3D mesh visual quality assessment. This chapter presents a summary of the thesis's work in Section 6.1. In Section 6.2 we discussed the novel contributions. A summary of the most important findings from each contribution chapter is provided. Finally, Section 6.3 discusses the work that carry out in the future.

6.1 Summary

In response to the limited research on the quality of 3D meshes, especially within VR environment, our study concentrates on visual quality assessment in this area. Given the increasing prevalence of 3D meshes as vital graphical elements for constructing immersive VR experiences, we explore the effects of diverse 3D distortion types on the perceptual quality of 3D shapes in VR setups. Our primary aim is to comprehend how different types and intensity levels of 3D mesh distortions impact VR perceived quality and user experience a critical consideration for real-time applications. We employ both Pearson and Spearman Correlation Coefficients to analyze the correlations between VR and desktop settings under varying types and levels of distortions. Our experimental findings reveal a positive linear relationship between the two settings.

In addition, we introduce a method of predicting 3D mesh saliency using neural net-

works. A mesh saliency algorithm was devised to identify crucial regions of the 3D mesh. By leveraging eye-tracking experiment data, machine learning optimises both linear and nonlinear models to measure geometric features. This approach shows that mesh saliency significantly enhances mesh quality prediction.

6.2 Novel Contribution

Mesh Quality Assessment (MQA) is relatively new compared to Image Quality Assessment (IQA). This thesis focuses on 3D mesh quality assessment as there is limited work on 3D mesh quality in a VR setting. Most existing studies measure the quality of 3D mesh in a desktop environment, which inspired our work to measure the quality in a virtual reality environment. However, in the VR setting, this is a crucial area for downstream applications. This is because massive amounts of data are necessary to support augmented reality and virtual reality. Also. The existing studies used hard-coded formula methods to measure 3D mesh saliency, which are difficult to measure. We introduce new methods of measuring 3D mesh saliency. We ran an experiment using an eye tracker, which produced a salient measurement accurate to human visual judgement. The thesis included the following contributions that can be discussed in each contribution.

6.2.1 Subjective Study of 3D Mesh Quality Scores in Virtual Reality

Perceiving distortion in VR and desktop settings can differ due to the unique characteristics and user experiences offered by each setting. A number of factors can affect perceived distortion, including the quality of the VR headset, the capabilities of the graphics processor, the complexity of the virtual scene, and the user's own perception. Chapter 3 presented a subjective study comparing the quality of 3D mesh distorted with reference shape (undistorted) in VR and traditional desktop display settings. In addition, we observed how various forms of distortions impact the visual quality of the 3D mesh by

gathering subjective scores for deformed shapes utilising linear and nonlinear correlation coefficients. The study reveals that the perceived quality of 3D meshes varies and is sensitive to the type and location of distortion. The distributions of total MOS scores are substantially associated between VR and desktop settings. However, noise is more visible in the VR setting than detail loss, as compared to the desktop setting. Mainly, noise applied to complete meshes or smooth parts tends to be more evident than noise added to other regions, and the variations are far more pronounced in VR. The results may guide the processing of 3D meshes for virtual reality applications.

As we focused above on 3D mesh distortion, there are other factors that can be affect the quality of 3D mesh in VR and desktop settings. Artefacts in VR distort the accurate representation of the virtual world. Users of VR headsets typically see virtual content through lenses. There is a possibility that these lenses will introduce optical distortions such as chromatic aberration or barrel distortion, which can adversely affect the quality of the image. In some cases, users may notice colour fringing or image warping at the edges of their view. The Field of View (FoV) of VR headsets is limited as compared to human vision. In the case of users experiencing a ”tunnel vision” effect, their peripheral vision may appear less detailed or distorted as a result of a restricted field of vision. Further, VR experiences often involve head movements, and any delay or lag can result in motion-related distortions. There may be a delay between the movement of the user's head and the corresponding change in the virtual world if the system fails to update the view quickly enough. Motion sickness or disconnection may happen as a result of perceiving the quality incorrectly.

When it comes to desktop settings, distortion usually refers to visual artefacts or anomalies that adversely affect the quality of content displayed. Streaming or file compression can result in compression artifacts when applied to reduce file sizes. As a result of these artefacts, visual distortions appear, such as blockiness, blurring, or pixelation. As a result, mesh clarity and perceived quality are reduced.

6.2.2 Learning to Predict 3D Mesh Saliency

This chapter is about learning to predict 3D mesh saliency with a focus on general visual saliency (i.e.,without specific task). Chapter 4 estimating saliency on a mesh is an essential technique with numerous downstream applications. Existing approaches rely mostly on building specialised formulae, which is hard to measure, and capturing perceptual significance using these methods is challenging. This thesis examines a technique for producing ground truth saliency maps on meshes using eye-tracking data. In particular, we combine saliency maps from several viewpoints to build a single saliency map that is defined for a particular mesh (so not view-dependent). We continue to build learning-based algorithms that use previous saliency prediction findings and geometric features at each vertex to predict the local saliency value. According to qualitative and quantitative findings, the learning-based approaches outperform state-of-the-art methods, especially the model based on least squares regression.Our work can be useful for employing deep learning approaches and compare deep learning results with traditional machine learning to make improvements. Also, it helps to understand how the deep learning accuracy and speed are different with the traditional machine learning. As 3D mesh data is relatively small, more data will improve the measurements of saliency for 3D shapes.

6.2.3 Objective Quality Assessment Measures for Mesh Quality

Chapter 5 evaluates different methods that measure how different two meshes are from each other by determining their geometry and/or perceptual distance. We compare the state-of-the-art mesh perceptual difference metrics for predicting the objective quality scores captured in the VR setting with the desktop setting. We found that our methods, VR-MSDM and VR-MSDM2, are better at measuring 3D mesh quality than the state-of-the-art. We further incorporated mesh saliency into the mesh difference measure, which is shown to improve the perceptual quality prediction for the VR setting.

6.3 Future Work

This thesis has contributed to understanding 3D mesh quality both subjectively and objectively; nonetheless, techniques for measuring 3D mesh quality are still in earlier stages, especially in the VR area, and further study is required. Each chapter's contribution proposes some possible future research areas (see Chapters 3, 4 and 5). Here, further thoughts are provided.

- **Objects quality in Virtual Reality (VR):** Further research will focus on how visibility information can be effectively incorporated into objective quality measures. Objects in VR scenes may change their position and orientation based on designed behaviour or in response to user interaction. The visibility of objects in VR scenes is therefore affected, along with other factors such as occlusion and lighting, which can influence the perceptual quality of 3D objects. Also, larger subjective databases of 3D quality perception could enable deep-learning models to be effectively built for 3D mesh quality measure especially in VR setting, similar to the success of Learned Perceptual Image Patch Similarity (LPIPS) for image quality measures [245].

- **Saliency-based metrics:** The saliency of visual content, such as images, 3D meshes and videos, plays a crucial role in quality assessment and evaluation. Saliency refers to the areas of an image or 3D mesh that catch the viewer's attention. In quality assessment, it is essential to understand visual saliency, since it facilitates the determination of the most visually significant or important regions of an image or 3D mesh. A visual saliency assessment is used in quality assessment in order to evaluate the overall aesthetic appeal, composition, and attention-grabbing aspects of visual content. Analysing the salient regions allows the identification of potential areas that may affect the perceived quality of the content. Visual quality can be assessed using a variety of approaches based on saliency. The purpose of saliency metrics is to identify the salient regions in 3D mesh and assign weights or importance values

to these regions. It is possible for the algorithm to take into account the visual significance of different regions by incorporating these weights into the overall quality assessment. For practical applications, such saliency prediction methods need to be sufficiently fast, especially for real-time adaptive streaming applications. Current mesh saliency methods are generally time-consuming, especially on complicated 3D shapes which need such techniques for saliency-guided simplification. In the future, with the increasing availability of labelled data, graph convolutional neural networks can potentially be built on 3D meshes to efficiently predict saliency maps in an end-to-end manner.

- **Deep learning-based methods:** In order to predict visual saliency, deep learning techniques can be used, such as convolutional neural networks (CNNs). By using these models, quality assessment algorithms are able to detect salient features and regions.In recent years, much progress has been made in understanding the representations learned by CNNs to recognise scenes and objects (e.g. [250]). However, there needs to be more knowledge of how deep saliency models learn. How does saliency computation arise in deep saliency architectures, and how do the patterns learn at various network levels diverge from those learned for object recognition? To investigate this, one may compare the results of two different deep networks trained on an identical set of stimuli and labelled with object category labels and saliency annotations, respectively (e.g. clicks).

- **Region of Interest (ROI) detection:** The purpose of this technique is to identify the areas in 3D mesh that are likely to be visually appealing and informative. It is possible to determine the quality of 3D mesh by the presence of salient objects or areas. Some of the possibilities can be used for real-time implications. The ability to detect ROI in real-time is crucial for applications such as live video streaming, augmented reality (AR), and virtual reality (VR). Research in the future should focus on developing efficient ROI detection algorithms that can be run in real-time on resource-constrained devices. It may be necessary to explore techniques such as model compression, optimisation, and hardware acceleration in order to achieve

real-time performance. Also, subjective evaluation, although objective measures play an important role in quality assessment, subjective evaluation is equally important in capturing human perceptions and preferences. In the future, user studies can be conducted to collect subjective ratings and preferences for various salient regions. Based on these data, subjective quality assessment models can be developed that align with human perceptions and preferences.

- **Eye-tracking data:** Human visual attention can be studied through eye-tracking experiments. Using eye-tracking data, researchers can determine which areas attract the most attention and evaluate the quality of visual content based on those areas. Some of the future work personalised saliency models. Most saliency models are designed based on the average behaviour of the human gaze. Despite this, individuals may have unique gaze patterns which are influenced by factors such as culture, experience, or cognitive differences. It is possible that future research will focus on developing personalised saliency models that can adapt to individual users and provide more accurate predictions of their visual attention. Also, in real-world environments, the majority of eye-tracking studies are conducted in controlled laboratory settings, which may not be representative of real-life conditions. To understand how gaze behaviour and perception of visual quality are affected by eye tracking and saliency visual quality in naturalistic environments, such as outdoor scenes or interactive settings, future research may examine these factors.

- **Datasets and benchmarks on quality assessment:** Regarding future research on this thesis, a substantial amount of effort is required to build appropriate benchmarks that allow the creation and assessment of 3D mesh quality and correspondence strategies for various scenarios. Specifically, more research needs to be done on how different materials affect the shape and how that might fit with or against the assumptions of the current 3D mesh quality assessment and correspondence methods. Generally, the 3D dataset is relatively small, especially 3D mesh saliency datasets are still orders of magnitude smaller than their corresponding datasets in other fields of computer vision, such as image analysis, where it is common to have

datasets with over 1 million images [51].

There are still more substantial datasets available in the image domain. However, 3D mesh datasets are usually much smaller. For instance, Yohanandan et al. [239] amassed a dataset of 300 million images called the JFT-300M and demonstrated that object identification models performed more effectively when trained using this dataset. Because of this, future studies in saliency will benefit significantly from collecting additional data. A more considerable amount of data opens up new dimensions to evaluate models. Also, in future research, we want to construct a more extensive dataset and examine the efficacy of other machine-learning approaches, particularly those based on deep neural networks. In the future, when richer datasets are available, our method will demonstrate its robust capacity for learning. One of our future efforts is to construct a subjective database with complete coverage of several elements, including various forms of distortion and models with varying geometries. This study might assist authors seeking to perform a subjective quality experiment to evaluate the predicted 3D mesh quality of distortion models in choosing the appropriate experimental approach for their particular challenge.

Bibliography

[1] Antti Aaltonen, Aulikki Hyrskykari, and Kari-Jouko Räihä. “101 spots, or how do users read menus?” In: *Proceedings of the SIGCHI conference on Human factors in computing systems*. 1998, pp. 132–139.

[2] Zeineb Abderrahim and Mohamed Salim Bouhlel. “Interactive multiresolution visualization of 3D Mesh”. In: *International Journal of Computer Applications* 67.14 (2013), pp. 33–39.

[3] Ilyass Abouelaziz, Aladine Chetouani, Mohammed El Hassouni, and Hocine Cherifi. “Reduced reference mesh visual quality assessment based on convolutional neural network”. In: *2018 14th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*. IEEE. 2018, pp. 617–620.

[4] Ilyass Abouelaziz, Aladine Chetouani, Mohammed El Hassouni, Longin Jan Latecki, and Hocine Cherifi. “Convolutional neural network for blind mesh visual quality assessment using 3D visual saliency”. In: *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE. 2018, pp. 3533–3537.

[5] Ilyass Abouelaziz, Mohammed El Hassouni, and Hocine Cherifi. “A convolutional neural network framework for blind mesh visual quality assessment”. In: *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE. 2017, pp. 755–759.

[6] Ilyass Abouelaziz, Mohammed El Hassouni, and Hocine Cherifi. “A curvature based method for blind mesh visual quality assessment using a general regres-

sion neural network". In: *2016 12th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*. IEEE. 2016, pp. 793–797.

[7] Ilyass Abouelaziz, Mohammed El Hassouni, and Hocine Cherifi. "Blind 3D mesh visual quality assessment using support vector regression". In: *Multimedia Tools and Applications* 77.18 (2018), pp. 24365–24386.

[8] Ilyass Abouelaziz, Mohammed El Hassouni, and Hocine Cherifi. "No-reference 3d mesh quality assessment based on dihedral angles model and support vector regression". In: *International Conference on Image and Signal Processing*. Springer. 2016, pp. 369–377.

[9] Ayyoub Ahar, Adriaan Barri, and Peter Schelkens. "From sparse coding significance to perceptual quality: A new approach for image quality assessment". In: *IEEE Transactions on Image Processing* 27.2 (2017), pp. 879–893.

[10] Pinar Akyazi and Touradj Ebrahimi. "A new objective metric to predict image quality using deep neural networks". In: *Applications of Digital Image Processing XLI*. Vol. 10752. SPIE. 2018, pp. 554–567.

[11] W Albert. "Do web users actually look at ads? A case study of banner ads and eye-tracking technology". In: *Proceedings of the 11th Annual Conference of the Usability Professionals' Association*. 2002.

[12] Evangelos Alexiou, Evgeniy Upenik, and Touradj Ebrahimi. "Towards subjective quality assessment of point cloud imaging in augmented reality". In: *2017 IEEE 19th International Workshop on Multimedia Signal Processing (MMSP)*. IEEE. 2017, pp. 1–6.

[13] Evangelos Alexiou, Irene Viola, Tomás M Borges, Tiago A Fonseca, Ricardo L De Queiroz, and Touradj Ebrahimi. "A comprehensive study of the rate-distortion performance in MPEG point cloud compression". In: *APSIPA Transactions on Signal and Information Processing* 8 (2019).

[14] Patrice Rondao Alface, Mathieu De Craene, and Benoit B Macq. “Three-dimensional image quality measurement for the benchmarking of 3D watermarking schemes”. In: *Security, Steganography, and Watermarking of Multimedia Contents VII*. Vol. 5681. SPIE. 2005, pp. 230–240.

[15] Brett Allen, Brian Curless, and Zoran Popović. “Articulated body deformation from range scan data”. In: *ACM Transactions on Graphics (TOG)* 21.3 (2002), pp. 612–619.

[16] Th Alpert and JP Evain. “Subjective quality evaluation: the SSCQE and DSCQE methodologies”. In: *EBU technical review* (1997), pp. 12–20.

[17] Joao Ascenso, Pinar Akyazi, Fernando Pereira, and Touradj Ebrahimi. “Learning-based image coding: early solutions reviewing and subjective quality evaluation”. In: *Optics, Photonics and Digital Technologies for Imaging Applications VI*. Vol. 11353. International Society for Optics and Photonics. 2020, 113530S.

[18] Lida Asgharian and Hossein Ebrahimnezhad. “How many sample points are sufficient for 3D model surface representation and accurate mesh simplification?” In: *Multimedia Tools and Applications* 79.39 (2020), pp. 29595–29620.

[19] Linden J Ball, Erica J Lucas, Jeremy NV Miles, and Alastair G Gale. “Inspection times and the selection task: What do eye-movements reveal about relevance effects?” In: *The Quarterly Journal of Experimental Psychology Section A* 56.6 (2003), pp. 1053–1077.

[20] Zhe Bian, Shi-Min Hu, and Ralph Martin. “Comparing small visual differences between conforming meshes”. In: *International Conference on Geometric Modeling and Processing*. Springer. 2008, pp. 62–78.

[21] Ali Borji and James Tanner. “Reconciling saliency and object center-bias hypotheses in explaining free-viewing fixations”. In: *IEEE Transactions on Neural Networks and Learning Systems* 27.6 (2015), pp. 1214–1226.

[22] Sebastian Bosse, Dominique Maniry, Klaus-Robert Müller, Thomas Wiegand, and Wojciech Samek. “Deep neural networks for no-reference and full-reference image quality assessment”. In: *IEEE Transactions on image processing* 27.1 (2017), pp. 206–219.

[23] Fadi Boulos, Benoît Parrein, Patrick Le Callet, and David Hands. “Perceptual effects of packet loss on H. 264/AVC encoded videos”. In: *Fourth International Workshop on Video Processing and Quality Metrics for Consumer Electronics VPQM-09*. 2009.

[24] Matthew Brown and David G Lowe. “Automatic panoramic image stitching using invariant features”. In: *International journal of computer vision* 74.1 (2007), pp. 59–73.

[25] Neil Bruce and John Tsotsos. “Attention based on information maximization”. In: *Journal of Vision* 7.9 (2007), pp. 950–950.

[26] I BT. “Methodologies for the subjective assessment of the quality of television images, document Recommendation ITU-R BT. 500-14 (10/2019)”. In: *ITU, Geneva, Switzerland* (2020).

[27] Abdullah Bulbul, Tolga Capin, Guillaume Lavoué, and Marius Preda. “Assessing visual quality of 3-D polygonal models”. In: *IEEE Signal Processing Magazine* 28.6 (2011), pp. 80–90.

[28] Martin Čadík, Robert Herzog, Rafał Mantiuk, Radosław Mantiuk, Karol Myszkowski, and Hans-Peter Seidel. “Learning to predict localized distortions in rendered images”. In: *Computer Graphics Forum*. Vol. 32. 7. Wiley Online Library. 2013, pp. 401–410.

[29] Martin Čadík, Michael Wimmer, Laszlo Neumann, and Alessandro Artusi. “Evaluation of HDR tone mapping methods using essential perceptual attributes”. In: *Computers & Graphics* 32.3 (2008), pp. 330–349.

[30] Licia Capodiferro, Giovanni Jacovitti, and Elio D Di Claudio. “Two-dimensional approach to full-reference image quality assessment based on positional structural information”. In: *IEEE transactions on image processing* 21.2 (2011), pp. 505–516.

[31] Ufuk Celikcan, Mehmet Bahadir Askin, Dilara Albayrak, and Tolga K Capin. “Deep into visual saliency for immersive VR environments rendered in real-time”. In: *Computers & Graphics* 88 (2020), pp. 70–82.

[32] Damon M Chandler and Sheila S Hemami. “VSNR: A wavelet-based visual signal-to-noise ratio for natural images”. In: *IEEE transactions on image processing* 16.9 (2007), pp. 2284–2298.

[33] Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. “Return of the devil in the details: Delving deep into convolutional nets”. In: *arXiv preprint arXiv:1405.3531* (2014).

[34] Geng Chen, Hang Dai, Tao Zhou, Jianbing Shen, and Ling Shao. “Automatic Schelling Points Detection from Meshes”. In: *IEEE Transactions on Visualization and Computer Graphics* (2022).

[35] Meixu Chen, Yize Jin, Todd Goodall, Xiangxu Yu, and Alan Conrad Bovik. “Study of 3D virtual reality picture quality”. In: *IEEE Journal of Selected Topics in Signal Processing* 14.1 (2019), pp. 89–102.

[36] Xiaobai Chen, Aleksey Golovinskiy, and Thomas Funkhouser. “A benchmark for 3D mesh segmentation”. In: *ACM transactions on graphics (tog)* 28.3 (2009), pp. 1–12.

[37] Xiaobai Chen, Abulhair Saparov, Bill Pang, and Thomas Funkhouser. “Schelling points on 3D surface meshes”. In: *ACM Transactions on Graphics (TOG)* 31.4 (2012), pp. 1–12.

[38] Siu-Wing Cheng, Tamal K Dey, and Jonathan Shewchuk. *Delaunay mesh generation*. CRC Press, 2012.

[39] Zhengxue Cheng, Pinar Akyazi, Heming Sun, Jiro Katto, and Touradj Ebrahimi. “Perceptual quality study on deep learning based image compression”. In: *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE. 2019, pp. 719–723.

[40] Aladine Chetouani. “A 3D mesh quality metric based on features fusion”. In: *Electronic Imaging* 2017.20 (2017), pp. 4–8.

[41] Aladine Chetouani. “Full reference image quality metric for stereo images based on cyclopean image computation and neural fusion”. In: *2014 IEEE Visual Communications and Image Processing Conference*. IEEE. 2014, pp. 109–112.

[42] Ioan Cleju and Dietmar Saupe. “Evaluation of supra-threshold perceptual metrics for 3D models”. In: *Proceedings of the 3rd symposium on Applied perception in graphics and visualization*. 2006, pp. 41–44.

[43] Massimiliano Corsini, E Drelie Gelasca, and Touradj Ebrahimi. “A multi-scale roughness metric for 3D watermarking quality assessment”. In: *Workshop on Image Analysis for Multimedia Interactive Services 2005*. CONF. 2005.

[44] Massimiliano Corsini, Elisa Drelie Gelasca, Touradj Ebrahimi, and Mauro Barni. “Watermarked 3-D mesh quality assessment”. In: *IEEE Transactions on Multimedia* 9.2 (2007), pp. 247–256.

[45] Massimiliano Corsini, Mohamed-Chaker Larabi, Guillaume Lavoué, Oldřich Petřík, Libor Váša, and Kai Wang. “Perceptual metrics for static and dynamic triangle meshes”. In: *Computer Graphics Forum*. Vol. 32. 1. Wiley Online Library. 2013, pp. 101–125.

[46] Laura Cowen, Linden Js Ball, and Judy Delin. “An eye movement analysis of web page usability”. In: *People and computers XVI-memorable yet invisible*. Springer, 2002, pp. 317–335.

[47] Leonard Daly and Don Brutzman. “X3D: Extensible 3D graphics standard [standards in a nutshell]”. In: *IEEE Signal Processing Magazine* 24.6 (2007), pp. 130–135.

[48] Scott J Daly. “Visible differences predictor: an algorithm for the assessment of image fidelity”. In: *Human Vision, Visual Processing, and Digital Display III*. Vol. 1666. SPIE. 1992, pp. 2–15.

[49] Tal Darom and Yosi Keller. “Scale-invariant features for 3-D mesh models”. In: *IEEE Transactions on Image Processing* 21.5 (2012), pp. 2758–2769.

[50] Ana De Abreu, Cagri Ozcinar, and Aljosa Smolic. “Look around you: Saliency maps for omnidirectional images in VR applications”. In: *2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE. 2017, pp. 1–6.

[51] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. “Imagenet: A large-scale hierarchical image database”. In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255.

[52] Elio D Di Claudio and Giovanni Jacovitti. “A detail-based method for linear full reference image quality prediction”. In: *IEEE Transactions on Image Processing* 27.1 (2017), pp. 179–193.

[53] Li Ding, Hua Huang, and Yu Zang. “Image quality assessment using directional anisotropy structure measurement”. In: *IEEE Transactions on Image Processing* 26.4 (2017), pp. 1799–1809.

[54] Soussan Djamasbi, Marisa Siegel, and Tom Tullis. “Generation Y, web design, and eye tracking”. In: *International journal of human-computer studies* 68.5 (2010), pp. 307–323.

[55] Adam Dobrin. “A review of properties and variations of Voronoi diagrams”. In: *Whitman College* 10.1.453 (2005), p. 9156.

[56] Lu Dong, Yuming Fang, Weisi Lin, and Hock Soon Seah. “Perceptual quality assessment for 3D triangle mesh based on curvature”. In: *IEEE Transactions on Multimedia* 17.12 (2015), pp. 2174–2184.

[57] Helin Dutagaci, Chun Pan Cheung, and Afzal Godil. “A benchmark for best view selection of 3D objects”. In: *Proceedings of the ACM workshop on 3D object retrieval*. 2010, pp. 45–50.

[58] Michael P Eckert and Andrew P Bradley. “Perceptual quality metrics applied to still image compression”. In: *Signal processing* 70.3 (1998), pp. 177–200.

[59] Jim Edwards. “Planet selfie: We’re now posting a staggering 1.8 billion photos every day”. In: *Business Insider* (2014).

[60] MAM El-Bendary, M El-Tokhy, and HB Kazemian. “Efficient image transmission over low-power IEEE802. 15.1 network over correlated fading channels”. In: *2012 6th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT)*. IEEE. 2012, pp. 563–567.

[61] MAM El-Bendary, M El-Tokhy, F Shawki, and FE Abd-El-Samie. “Studying the throughput efficiency of JPEG image transmission over mobile IEEE 802.15. 1 network using EDR packets”. In: *2012 6th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT)*. IEEE. 2012, pp. 573–577.

[62] Mylène CQ Farias and Welington YL Akamine. “On performance of image quality metrics enhanced with visual attention computational models”. In: *Electronics letters* 48.11 (2012), pp. 631–633.

[63] Anna Maria Feit, Shane Williams, Arturo Toledo, Ann Paradiso, Harish Kulkarni, Shaun Kane, and Meredith Ringel Morris. “Toward everyday gaze input: Accuracy and precision of eye tracking and implications for design”. In: *Proceedings of the 2017 Chi conference on human factors in computing systems*. 2017, pp. 1118–1130.

[64] Xiang Feng, Wanggen Wan, Richard Yi Da Xu, Haoyu Chen, Pengfei Li, and J Alfredo Sánchez. “A perceptual quality metric for 3D triangle meshes based on spatial pooling”. In: *Frontiers of Computer Science* 12.4 (2018), pp. 798–812.

[65] Belen Jiménez Fernández-Palacios, Daniele Morabito, and Fabio Remondino. “Access to complex reality-based 3D models using virtual reality solutions”. In: *Journal of cultural heritage* 23 (2017), pp. 40–48.

[66] James A Ferwerda, Peter Shirley, Sumanta N Pattanaik, and Donald P Greenberg. “A model of visual masking for computer graphics”. In: *Proceedings of the 24th annual conference on Computer graphics and interactive techniques*. 1997, pp. 143–152.

[67] Steven Fortune. “Voronoi diagrams and Delaunay triangulations”. In: *Computing in Euclidean geometry* (1995), pp. 225–265.

[68] Fei Gao, Yi Wang, Panpeng Li, Min Tan, Jun Yu, and Yani Zhu. “Deepsim: Deep similarity for image quality assessment”. In: *Neurocomputing* 257 (2017), pp. 104–114.

[69] Paolo Gastaldo and Judith A Redi. “Machine learning solutions for objective visual quality assessment”. In: *6th international workshop on video processing and quality metrics for consumer electronics, VPQM*. Vol. 12. 2012.

[70] Paolo Gastaldo, Rodolfo Zunino, and Judith Redi. “Supporting visual quality assessment with machine learning”. In: *EURASIP Journal on Image and Video Processing* 2013.1 (2013), pp. 1–15.

[71] Elisa Drelie Gelasca, Touradj Ebrahimi, Massimiliano Corsini, and Mauro Barni. “Objective evaluation of the perceptual quality of 3D watermarking”. In: *IEEE International Conference on Image Processing 2005*. Vol. 1. IEEE. 2005, pp. I–241.

[72] Daniela Giorgi, Michela Mortara, and Michela Spagnuolo. “3D shape retrieval based on best view selection”. In: *Proceedings of the ACM workshop on 3D object retrieval*. 2010, pp. 9–14.

[73] Joseph H Goldberg and Xerxes P Kotval. “Computer interface evaluation using eye movements: methods and constructs”. In: *International journal of industrial ergonomics* 24.6 (1999), pp. 631–645.

[74] Joseph H Goldberg, Mark J Stimson, Marion Lewenstein, Neil Scott, and Anna M Wichansky. “Eye tracking in web search tasks: design implications”. In: *Proceedings of the 2002 symposium on Eye tracking research & applications*. 2002, pp. 51–58.

[75] Lutz Goldmann, Francesca De Simone, Frederic Dufaux, Touradj Ebrahimi, Rudolf Tanner, and Mauro Lattuada. “Impact of video transcoding artifacts on the subjective quality”. In: *2010 Second International Workshop on Quality of Multimedia Experience (QoMEX)*. IEEE. 2010, pp. 52–57.

[76] Jinjiang Guo, Vincent Vidal, Irene Cheng, Anup Basu, Atilla Baskurt, and Guillaume Lavoue. “Subjective and objective visual quality assessment of textured 3D meshes”. In: *ACM Transactions on Applied Perception (TAP)* 14.2 (2016), pp. 1–20.

[77] Yu Guo, Fei Wang, and Jingmin Xin. “Point-wise saliency detection on 3D point clouds via covariance descriptors”. In: *The Visual Computer* 34.10 (2018), pp. 1325–1338.

[78] Meenakshi Gupta and Atul Garg. “A Comparative Analysis of Content Delivery Network and Other Techniques for Web Content Delivery”. In: *International Journal of Service Science, Management, Engineering, and Technology (IJSSMET)* 6.4 (2015), pp. 43–58.

[79] Marianne Hanhela, Atanas Boev, Atanas Gotchev, and Miska Hannuteela. “Fusion of eye-tracking data from multiple observers for increased 3D gaze tracking precision”. In: *2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*. IEEE. 2012, pp. 420–424.

[80] Eamonn KS Hanson. “Focus of attention and pilot error”. In: *Proceedings of the 2004 symposium on Eye tracking research & applications*. 2004, pp. 60–60.

[81] J Harel, C Koch, and P Perona. “Graph-based visual saliency”. In: *Proceedings of the 20th Annual Conference on Neural*.

[82] Jonathan Harel, Christof Koch, and Pietro Perona. “Graph-based visual saliency”. In: *Advances in neural information processing systems* 19 (2006).

[83] Gunnar Hauland. “Measuring team situation awareness by means of eye movement data”. In: *Proceedings of HCI International 2003*. Vol. 3. 2019, pp. 230–234.

[84] P.S. Heckbert and M. Garland. “Optimal triangulation and quadric-based surface simplification”. In: *Computational Geometry* 14.1-3 (1999), pp. 49–65.

[85] Sheila S Hemami and Amy R Reibman. “No-reference image and video quality estimation: Applications and human-motivated design”. In: *Signal processing: Image communication* 25.7 (2010), pp. 469–481.

[86] Andrew Hines, Paul Kendrick, Adriaan Barri, Manish Narwaria, and Judith A Redi. “Robustness and prediction accuracy of machine learning for objective visual quality assessment”. In: *2014 22nd European Signal Processing Conference (EUSIPCO)*. IEEE. 2014, pp. 2130–2134.

[87] Tobias Hoßfeld, Matthias Hirth, Judith Redi, Filippo Mazza, Pavel Korshunov, Babak Naderi, Michael Seufert, Bruno Gardlo, Sebastian Egger, and Christian Keimel. “Best Practices and Recommendations for Crowdsourced QoE-Lessons learned from the Qualinet Task Force” Crowdsourcing””. In: (2014).

[88] Weilong Hou and Xinbo Gao. “Be natural: A saliency-guided deep framework for image quality”. In: *2014 IEEE International Conference on Multimedia and Expo (ICME)*. Chengdu: IEEE, July 2014.

[89] S. Howlett, J. Hamill, and C. O’Sullivan. “An experimental approach to predicting saliency for simplified polygonal models”. In: *In Proceedings of the 1st Symposium on Applied Perception in Graphics and Visualization*. Vol. 57-64. ACM, 2004.

[90] Cisco Visual Networking Index. “Forecast and methodology, 2016–2021”. In: *White paper, Cisco public* 6 (2017).

[91] Dan Ionescu, Bogdan Ionescu, Shahidul Islam, Cristian Gadea, and Eric McQuiggan. “Using depth measuring cameras for a new human computer interaction in augmented virtual reality environments”. In: *2010 IEEE International Conference on Virtual Environments, Human-Computer Interfaces and Measurement Systems*. IEEE. 2010, pp. 114–119.

[92] L Itti, C Koch, and E Niebur. “A model of saliency-based visual attention for rapid scene analysis”. In: *IEEE Trans. Pattern Anal. Mach. Intell.* 20.11 (1998), pp. 1254–1259.

[93] L. Itti, C. Koch, and E. Niebur. “A model of saliency-based visual attention for rapid scene analysis”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20 (1998), p. 11.

[94] Laurent Itti, Christof Koch, and Ernst Niebur. “A model of saliency-based visual attention for rapid scene analysis”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20.11 (1998), pp. 1254–1259.

[95] Alireza Javaheri, Catarina Brites, Fernando Pereira, and João Ascenso. “Subjective and objective quality evaluation of 3D point cloud denoising algorithms”. In: *2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. IEEE. 2017, pp. 1–6.

[96] T. Judd. “Learning to predict where humans look”. In: *IEEE International Conference on Computer Vision (ICCV)*. 2009.

[97] Marcel Adam Just and Patricia A Carpenter. “Eye fixations and cognitive processes”. In: *Cognitive psychology* 8.4 (1976), pp. 441–480.

[98] Zachi Karni and Craig Gotsman. “Compression of soft-body animation sequences”. In: *Computers & Graphics* 28.1 (2004), pp. 25–34.

[99] Zachi Karni and Craig Gotsman. “Spectral compression of mesh geometry”. In: *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*. 2000, pp. 279–286.

[100] Sun-Jeong Kim, Soo-Kyun Kim, and Chang-Hun Kim. “Discrete differential error metric for surface simplification”. In: *10th Pacific Conference on Computer Graphics and Applications, 2002. Proceedings.* IEEE. 2002, pp. 276–283.

[101] Vladimir G. Kim, Siddhartha Chaudhuri, Leonidas Guibas, and Thomas Funkhouser. “Shape2pose: Human-centric shape analysis”. In: *ACM Transactions on Graphics (TOG)* 33.4 (2014), pp. 1–12.

[102] Youngmin Kim, Amitabh Varshney, David W Jacobs, and François Guimbretiere. “Mesh saliency and human eye fixations”. In: *ACM Transactions on Applied Perception (TAP)* 7.2 (2010), pp. 1–13.

[103] Matthias Kirchner and Rainer Bohme. “Hiding traces of resampling in digital images”. In: *IEEE Transactions on Information Forensics and Security* 3.4 (2008), pp. 582–592.

[104] Debarati Kundu and Brian L Evans. “Visual attention guided quality assessment of Tone-Mapped images using scene statistics”. In: *2016 IEEE International Conference on Image Processing (ICIP)*. Phoenix, AZ, USA: IEEE, Sept. 2016.

[105] Manfred Lau, Kapil Dev, Weiqi Shi, Julie Dorsey, and Holly Rushmeier. “Tactile mesh saliency”. In: *ACM Transactions on Graphics* 35.4 (2016), p. 52.

[106] Joseph J LaViola Jr. “A discussion of cybersickness in virtual environments”. In: *ACM Sigchi Bulletin* 32.1 (2000), pp. 47–56.

[107] Guillaume Lavoué. “A local roughness measure for 3D meshes and its application to visual masking”. In: *ACM Transactions on Applied perception (TAP)* 5.4 (2009), pp. 1–23.

[108] Guillaume Lavoué. “A multiscale metric for 3D mesh visual quality assessment”. In: *Computer Graphics Forum*. Vol. 30. 5. Wiley Online Library. 2011, pp. 1427–1437.

[109] Guillaume Lavoué, Irene Cheng, and Anup Basu. “Perceptual quality metrics for 3D meshes: towards an optimal multi-attribute computational model”. In: *2013*

IEEE International Conference on Systems, Man, and Cybernetics. IEEE. 2013, pp. 3271–3276.

[110] Guillaume Lavoué, Frédéric Cordier, Hyewon Seo, and Mohamed-Chaker Larabi. "Visual attention for rendered 3D shapes". In: *Computer Graphics Forum*. Vol. 37. 2. Wiley Online Library. 2018, pp. 191–203.

[111] Guillaume Lavoué, Elisa Drelie Gelasca, Florent Dupont, Atilla Baskurt, and Touradj Ebrahimi. "Perceptually driven 3D distance metrics with application to watermarking". In: *Applications of Digital Image Processing XXIX*. Vol. 6312. International Society for Optics and Photonics. 2006, p. 63120L.

[112] Guillaume Lavoué and Rafał Mantiuk. "Quality assessment in computer graphics". In: *Visual signal quality assessment*. Springer, 2015, pp. 243–286.

[113] Benjamin Law, M Stella Atkins, Arthur E Kirkpatrick, and Alan J Lomax. "Eye gaze patterns differentiate novice and experts in a virtual laparoscopic surgery training environment". In: *Proceedings of the 2004 symposium on Eye tracking research & applications*. 2004, pp. 41–48.

[114] Patrick Le Callet, Sebastian Möller, Andrew Perkis, et al. "Qualinet white paper on definitions of quality of experience". In: *European network on quality of experience in multimedia systems and services (COST Action IC 1003)* 3.2012 (2012).

[115] C. H. Lee, A. Varshney, and D. W. Jacobs. "Mesh saliency". In: *ACM Trans. Graph. (Proc. SIGGRAPH)* 24.3 (2005), pp. 659 –666.

[116] G. Leifman, E. Shtrom, and A. Tal. "Surface regions of interest for viewpoint selection". In: *IEEE Trans. Pattern Anal. Mach. Intell* 38.12 (2016), 2544–2556.

[117] Qiaohong Li, Yu-Ming Fang, and Jing-Tao Xu. "A novel spatial pooling strategy for image quality assessment". In: *Journal of Computer Science and Technology* 31.2 (2016), pp. 225–234.

[118] Xin Li. "Blind image quality assessment". In: *Proceedings. International Conference on Image Processing*. Vol. 1. IEEE. 2002, pp. I–I.

[119] Yaqian Liang, Fazhi He, and Xiantao Zeng. “3D mesh simplification with feature preservation based on Whale Optimization Algorithm and Differential Evolution”. In: *Integrated Computer-Aided Engineering* Preprint (2020), pp. 1–19.

[120] Max Limper, Arjan Kuijper, and Dieter W Fellner. “Mesh Saliency Analysis via Local Curvature Entropy.” In: *Eurographics (Short Papers)*. 2016, pp. 13–16.

[121] Peter Lindstrom and Greg Turk. “Image-driven simplification”. In: *ACM Transactions on Graphics (ToG)* 19.3 (2000), pp. 204–241.

[122] Hantao Liu, Ulrich Engelke, Junle Wang, Patrick Le Callet, and Ingrid Heynderickx. “How does image content affect the added value of visual attention in objective image quality assessment?” In: *IEEE Signal Processing Letters* 20.4 (2013), pp. 355–358.

[123] Hantao Liu and Ingrid Heynderickx. “A perceptually relevant no-reference blockiness metric based on local image characteristics”. In: *EURASIP Journal on Advances in Signal Processing* 2009 (2009), pp. 1–14.

[124] Hantao Liu and Ingrid Heynderickx. “Visual attention in objective image quality assessment: Based on eye-tracking data”. In: *IEEE transactions on Circuits and Systems for Video Technology* 21.7 (2011), pp. 971–982.

[125] Hantao Liu, Nick Klomp, and Ingrid Heynderickx. “A no-reference metric for perceived ringing artifacts in images”. In: *IEEE Transactions on Circuits and Systems for Video Technology* 20.4 (2009), pp. 529–539.

[126] Tsung-Jung Liu and Kuan-Hsien Liu. “No-reference image quality assessment by wide-perceptual-domain scorer ensemble method”. In: *IEEE Transactions on Image Processing* 27.3 (2017), pp. 1138–1151.

[127] Xianyong Liu, Ligang Liu, Weijie Song, Yanping Liu, and Lizhuang Ma. “Shape context based mesh saliency detection and its applications: A survey”. In: *Computers & Graphics* 57 (2016), pp. 12–30.

[128] Zhenbao Liu, Xiao Wang, and Shuhui Bu. “Human-centered saliency detection”. In: *IEEE Transactions on Neural Networks and Learning Systems* 27.6 (2015), pp. 1150–1162.

[129] Gerald L Lohse. “Consumer eye movement patterns on yellow pages advertising”. In: *Journal of Advertising* 26.1 (1997), pp. 61–73.

[130] David G. Lowe. “Distinctive image features from scale-invariant keypoints”. In: *International Journal of Computer Vision* 60.2 (2004), pp. 91–110.

[131] Wei Lyu, Wei Wu, Lin Zhang, Zhaohui Wu, and Zhong Zhou. “Laplacian-based 3D mesh simplification with feature preservation”. In: *International Journal of Modeling, Simulation, and Scientific Computing* 10.02 (2019), p. 1950002.

[132] Qi Ma and Liming Zhang. “Image quality assessment with visual attention”. In: *2008 19th International Conference on Pattern Recognition*. Tampa, FL, USA: IEEE, Dec. 2008.

[133] Päivi Majaranta and Andreas Bulling. “Eye tracking and eye-based human–computer interaction”. In: *Advances in physiological computing*. Springer, 2014, pp. 39–65.

[134] Rafał Mantiuk, Kil Joong Kim, Allan G Rempel, and Wolfgang Heidrich. “HDR-VDP-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions”. In: *ACM Transactions on graphics (TOG)* 30.4 (2011), pp. 1–14.

[135] Pina Marziliano, Frederic Dufaux, Stefan Winkler, and Touradj Ebrahimi. “Perceptual blur and ringing metrics: application to JPEG2000”. In: *Signal processing: Image communication* 19.2 (2004), pp. 163–172.

[136] Atef Masmoudi, Med Salim Bouhlel, and William Puech. “Image encryption using chaotic standard map and engle continued fractions map”. In: *2012 6th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT)*. IEEE. 2012, pp. 474–480.

[137] Nassima Medimegh, Samir Belaid, Mohamed Atri, and Naoufel Werghi. “3D mesh watermarking using salient points”. In: *Multimedia Tools and Applications* 77.24 (2018), pp. 32287–32309.

[138] Arian Mehrfard, Javad Fotouhi, Giacomo Taylor, Tess Forster, Nassir Navab, and Bernhard Fuerst. “A comparative analysis of virtual reality head-mounted display systems”. In: *arXiv preprint arXiv:1912.02913* (2019).

[139] Claudia Mello-Thoms, Calvin F Nodine, and Harold L Kundel. “What attracts the eye to the location of missed and reported breast cancers?” In: *Proceedings of the 2002 symposium on Eye tracking research & applications*. 2002, pp. 111–117.

[140] R. Milanese, H. Wechsler, S. Gill, J. Bostl, and T. Pun. “Integration of bottom-up and top-down cues for visual attentionusing non-linear relaxation”. In: *IEEE Conference on Computer Vision and Pattern Recognition*. 1994, pp. 781–785.

[141] Melissa Miles and Edward Welch. *Photography and Its Publics*. Bloomsbury Visual Arts, 2020.

[142] Xiongkuo Min, Guangtao Zhai, Zhongpai Gao, and Ke Gu. “Visual attention data for image quality assessment databases”. In: *2014 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE. 2014, pp. 894–897.

[143] Anish Mittal, Anush K Moorthy, Alan C Bovik, and Lawrence K Cormack. “Automatic prediction of saliency on JPEG distorted images”. In: *2011 Third International Workshop on Quality of Multimedia Experience*. IEEE. 2011, pp. 195–200.

[144] Rafael Monroy, Sebastian Lutz, Tejo Chalasani, and Aljosa Smolic. “Salnet360: Saliency maps for omni-directional images with cnn”. In: *Signal Processing: Image Communication* 69 (2018), pp. 26–34.

[145] Anush Krishna Moorthy and Alan Conrad Bovik. “Visual importance pooling for image quality assessment”. In: *IEEE journal of selected topics in signal processing* 3.2 (2009), pp. 193–201.

[146] Anush Krishna Moorthy and Alan Conrad Bovik. “Visual quality assessment algorithms: what does the future hold?” In: *Multimedia Tools and Applications* 51.2 (2011), pp. 675–696.

[147] AA M Muzahid, Wanggen Wan, and Xiang Feng. “Perceptual quality evaluation of 3d triangle mesh: a technical review”. In: *2018 International Conference on Audio, Language and Image Processing (ICALIP)*. IEEE. 2018, pp. 266–272.

[148] Karol Myszkowski. “Perception-based global illumination, rendering, and animation techniques”. In: *Proceedings of the 18th spring conference on Computer graphics*. 2002, pp. 13–24.

[149] Georges Nader. “Evaluating the visibility threshold for a local geometric distortion on a 3D mesh and its applications”. PhD thesis. Université de Lyon, 2016.

[150] Manish Narwaria and Weisi Lin. “Objective image quality assessment based on support vector regression”. In: *IEEE Transactions on Neural Networks* 21.3 (2010), pp. 515–519.

[151] Manish Narwaria and Weisi Lin. “SVD-based quality metric for image and video using machine learning”. In: *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 42.2 (2011), pp. 347–364.

[152] Hamid Reza Nasrinpour and Neil DB Bruce. “Saliency weighted quality assessment of tone-mapped images”. In: *2015 IEEE International Conference on Image Processing (ICIP)*. IEEE. 2015, pp. 4947–4951.

[153] Yana Nehmé, Jean-Philippe Farrugia, Florent Dupont, Patrick Le Callet, and Guillaume Lavoué. “Comparison of subjective methods for quality assessment of 3D graphics in virtual reality”. In: *ACM Transactions on Applied Perception (TAP)* 18.1 (2020), pp. 1–23.

[154] Anass Nouri, Christophe Charrier, and Olivier Lézoray. “3d blind mesh quality assessment index”. In: *Electronic Imaging* 2017.20 (2017), pp. 9–26.

[155] Anass Nouri, Christophe Charrier, and Olivier Lézoray. “Multi-scale mesh saliency with local adaptive patches for viewpoint selection”. In: *Signal Processing: Image Communication* 38 (2015), pp. 151–166.

[156] G Ososkov and P Goncharov. “Shallow and deep learning for image classification”. In: *Optical Memory and Neural Networks* 26.4 (2017), pp. 221–248.

[157] Hao Pan, Xiao-Fan Feng, and Scott Daly. “LCD motion blur modeling and analysis”. In: *IEEE International Conference on Image Processing 2005*. Vol. 2. IEEE. 2005, pp. II–21.

[158] Yanwei Pang, Manli Sun, Xiaoheng Jiang, and Xuelong Li. “Convolution in convolution for network in network”. In: *IEEE transactions on neural networks and learning systems* 29.5 (2017), pp. 1587–1597.

[159] M Mortara G Patanè, M Spagnuolo B Falcidieno, and J Rossignac. “Plumber: a method for a multi-scale decomposition of 3D shapes into tubular primitives and bodies”. In: ().

[160] Marius Pedersen and Jon Yngve Hardeberg. “Survey of full-reference image quality metrics”. In: (2009).

[161] Paolo Pellizzoni and Gianpaolo Savio. “Mesh Simplification by Curvature-Enhanced Quadratic Error Metrics”. In: *Contract* 2 (2020), p. v1.

[162] Peng Peng and Ze-Nian Li. “General-purpose image quality assessment based on distortion-aware decision fusion”. In: *Neurocomputing* 134 (2014), pp. 117–121.

[163] Josselin Petit and Rafał K Mantiuk. “Assessment of video tone-mapping: Are cameras’ S-shaped tone-curves good enough?” In: *Journal of Visual Communication and Image Representation* 24.7 (2013), pp. 1020–1030.

[164] Margaret H Pinson and Stephen Wolf. “A new standardized method for objectively measuring video quality”. In: *IEEE Transactions on broadcasting* 50.3 (2004), pp. 312–322.

[165] Alex Poole, Linden J Ball, and Peter Phillipsl. “In Search of Salience: A Response-time and Eye-movement Analysis of Bookmark”. In: *People and Computers XVIII-Design for Life: Proceedings of HCI 2004*. Springer Science & Business Media. 2007, p. 363.

[166] Lijun Qu and Gary W Meyer. “Perceptually guided polygon reduction”. In: *IEEE Transactions on Visualization and Computer Graphics* 14.5 (2008), pp. 1015–1029.

[167] Ralph Radach, Jukka Hyona, and Heiner Deubel. *The mind’s eye: Cognitive and applied aspects of eye movement research*. Elsevier, 2003.

[168] Yashas Rai, Patrick Le Callet, and Philippe Guillotel. “Which saliency weighting for omni directional image quality assessment?” In: *2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE. 2017, pp. 1–6.

[169] Ganesh Ramanarayanan, James Ferwerda, Bruce Walter, and Kavita Bala. “Visual equivalence: towards a new standard for image fidelity”. In: *ACM Transactions on Graphics (TOG)* 26.3 (2007), 76–es.

[170] Keith Rayner, Erik D Reichle, and Alexander Pollatsek. “Eye movement control in reading: An overview and model”. In: *Eye guidance in reading and scene perception* (1998), pp. 243–268.

[171] Abdul Rehman and Zhou Wang. “Reduced-reference image quality assessment by structural similarity estimation”. In: *IEEE transactions on image processing* 21.8 (2012), pp. 3378–3389.

[172] Flávio Ribeiro, Dinei Florêncio, Cha Zhang, and Michael Seltzer. “CrowdMOS: An approach for crowdsourcing mean opinion score studies”. In: *2011 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE. 2011, pp. 2416–2419.

[173] Bernice E Rogowitz and Holly E Rushmeier. “Are image quality metrics adequate to evaluate the quality of geometric objects?” In: *Human Vision and Electronic*

Imaging VI. Vol. 4299. International Society for Optics and Photonics. 2001, pp. 340–348.

[174] Michael Rubinstein, Diego Gutierrez, Olga Sorkine, and Ariel Shamir. "A comparative study of image retargeting". In: *ACM SIGGRAPH Asia 2010 papers*. 2010, pp. 1–10.

[175] Holly E Rushmeier, Bernice E Rogowitz, and Christine Piatko. "Perceptual issues in substituting texture for geometry". In: *Human Vision and Electronic Imaging V*. Vol. 3959. International Society for Optics and Photonics. 2000, pp. 372–383.

[176] Linda Ryan, David Tormey, and Perry Share. "Cultural Barriers to the Transition from Product to Product Service in the Medical Device Industry". In: *International Journal of Service Science, Management, Engineering, and Technology (IJSSMET)* 5.2 (2014), pp. 36–50.

[177] Mehdi Salehpour and Alireza Behrad. "3D face reconstruction by KLT feature extraction and model consistency match refining and growing". In: *2012 6th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT)*. IEEE. 2012, pp. 297–302.

[178] Dario D Salvucci and Joseph H Goldberg. "Identifying fixations and saccades in eye-tracking protocols". In: *Proceedings of the 2000 symposium on Eye tracking research & applications*. 2000, pp. 71–78.

[179] BT Series. "Methodology for the subjective assessment of the quality of television pictures". In: *Recommendation ITU-R BT* (2012), pp. 500–13.

[180] Ana Serrano, Vincent Sitzmann, Jaime Ruiz-Borau, Gordon Wetzstein, Diego Gutierrez, and Belen Masia. "Movie editing and cognitive event segmentation in virtual reality video". In: *ACM Transactions on Graphics (TOG)* 36.4 (2017), pp. 1–12.

[181] Feng Shao, Shanbo Gu, Gangyi Jang, and Mei Yu. "A novel no-reference stereoscopic image quality assessment method". In: *2012 Symposium on Photonics and Optoelectronics*. IEEE. 2012, pp. 1–4.

[182] Hamid R Sheikh, Alan C Bovik, and Lawrence Cormack. “No-reference quality assessment using natural scene statistics: JPEG2000”. In: *IEEE Transactions on image processing* 14.11 (2005), pp. 1918–1927.

[183] Hamid R Sheikh, Muhammad F Sabir, and Alan C Bovik. “A statistical evaluation of recent full reference image quality assessment algorithms”. In: *IEEE Transactions on image processing* 15.11 (2006), pp. 3440–3451.

[184] P. Shilane and T. Funkhouser. “Distinctive regions of 3D surfaces”. In: *ACM Trans. Graph.* 26.2 (2007), p. 7.

[185] Aleksandr Shnayderman, Alexander Gusev, and Ahmet M Eskicioglu. “An SVD-based grayscale image quality measure for local and global assessment”. In: *IEEE transactions on Image Processing* 15.2 (2006), pp. 422–429.

[186] Elizabeth Shtrom, George Leifman, and Ayellet Tal. “Saliency detection in large point sets”. In: *Proceedings of the IEEE international conference on computer vision*. 2013, pp. 3591–3598.

[187] Samuel Silva, Beatriz Sousa Santos, Carlos Ferreira, and Joaquim Madeira. “A perceptual data repository for polygonal meshes”. In: *2009 Second International Conference in Visualisation*. IEEE. 2009, pp. 207–212.

[188] Luis A da Silva Cruz, Emil Dumić, Evangelos Alexiou, Joao Prazeres, Rafael Duarte, Manuela Pereira, Antonio Pinheiro, and Touradj Ebrahimi. “Point cloud quality evaluation: Towards a definition for test conditions”. In: *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE. 2019, pp. 1–6.

[189] D Amnon Silverstein and Joyce E Farrell. “Efficient method for paired comparison”. In: *Journal of Electronic Imaging* 10.2 (2001), pp. 394–398.

[190] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. “Deep inside convolutional networks: Visualising image classification models and saliency maps”. In: (2014).

[191] Ivan Sipiran and Benjamin Bustos. “Key-components: detection of salient regions on 3D meshes”. In: *The Visual Computer* 29.12 (2013), pp. 1319–1332.

[192] R. Song, Y. Liu, R.R. Martin, and P.L. Rosin. “Mesh saliency via spectral processing”. In: *ACM Transactions on Graphics (TOG)* 33.1 (2014), p. 6.

[193] Ran Song, Yonghuai Liu, and Paul L Rosin. “Mesh saliency via weakly supervised classification-for-saliency CNN”. In: *IEEE transactions on visualization and computer graphics* 27.1 (2019), pp. 151–164.

[194] Ran Song, Wei Zhang, Yitian Zhao, and Yonghuai Liu. “Unsupervised Multi-view CNN for Salient View Selection of 3D Objects and Scenes”. In: *European Conference on Computer Vision*. Springer. 2020, pp. 454–470.

[195] Olga Sorkine, Daniel Cohen-Or, and Sivan Toledo. “High-pass quantization for mesh encoding.” In: *Symposium on Geometry Processing*. Vol. 42. Citeseer. 2003, p. 3.

[196] Alexis D Souchet, Stéphanie Philippe, Domitile Lourdeaux, and Laure Leroy. “Measuring visual fatigue and cognitive load via eye tracking while learning with virtual reality head-mounted displays: A review”. In: *International Journal of Human–Computer Interaction* 38.9 (2022), pp. 801–824.

[197] Bernhard Spanlang, Jean-Marie Normand, David Borland, Konstantina Kilteni, Elias Giannopoulos, Ausiàs Pomés, Mar González-Franco, Daniel Perez-Marcos, Jorge Arroyo-Palacios, Xavi Navarro Muncunill, et al. “How to build an embodiment lab: achieving body representation illusions in virtual reality”. In: *Frontiers in Robotics and AI* 1 (2014), p. 9.

[198] Mikhail Startsev and Michael Dorr. “360-aware saliency estimation with conventional image saliency predictors”. In: *Signal Processing: Image Communication* 69 (2018), pp. 43–52.

[199] S.L. Stoev and W. Straßer. “A case study on automatic camera placement and motion for visualizing historical data”. In: 2002, pp. 545–548.

[200] Hang Su, Subhransu Maji, Evangelos Kalogerakis, and Erik Learned-Miller. “Multi-view convolutional neural networks for 3D shape recognition”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2015, pp. 945–953.

[201] Honglei Su, Zhengfang Duanmu, Wentao Liu, Qi Liu, and Zhou Wang. “Perceptual quality assessment of 3D point clouds”. In: *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE. 2019, pp. 3182–3186.

[202] Pratik Suchde, Thibault Jacquemin, and Oleg Davydov. “Point cloud generation for meshfree methods: An overview”. In: *Archives of Computational Methods in Engineering* 30.2 (2023), pp. 889–915.

[203] Wen Sun, Qingmin Liao, Jing-Hao Xue, and Fei Zhou. “SPSIM: A superpixel-based similarity index for full-reference image quality assessment”. In: *IEEE Transactions on Image Processing* 27.9 (2018), pp. 4232–4244.

[204] Pingping Tao, Junjie Cao, Shuhua Li, Xiuping Liu, and Ligang Liu. “Mesh saliency via ranking unsalient patches in a descriptor space”. In: *Computers & Graphics* 46 (2015), pp. 264–274.

[205] Flora Ponjou Tasse, Jiri Kosinka, and Neil Dodgson. “Cluster-based point set saliency”. In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 163–171.

[206] Gabriel Taubin. “A signal processing approach to fair surface design”. In: *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*. 1995, pp. 351–358.

[207] Michela Testolina, Evgeniy Upenik, João Ascenso, Fernando Pereira, and Touradj Ebrahimi. “Performance Evaluation of Objective Image Quality Metrics on Conventional and Learning-Based Compression Artifacts”. In: *13th International Conference on Quality of Multimedia Experience*. CONF. 2021.

[208] Federico Tombari, Samuele Salti, and Luigi Di Stefano. “Unique signatures of histograms for local surface description”. In: *European conference on computer vision*. Springer. 2010, pp. 356–369.

[209] Fakhri Torkhani, Kai Wang, and Jean-Marc Chassery. “A curvature-tensor-based perceptual quality metric for 3D triangular meshes”. In: *Machine Graphics & Vision* 23.1 (2014), pp. 59–82.

[210] J.K. Tsotsos, S.M. Culhane, W.Y.K. Wai, Y. Lai, N. Davis, and F. Nuflo. “Modeling visual attention via selective tuning”. In: *Artificial intelligence* 78.1–2 (1995), pp. 507–545.

[211] Ennél Van Eeden and Wilson Chow. *Perspectives from the global entertainment & media outlook 2018–2022*. 2018.

[212] Libor Váša and Jan Rus. “Dihedral angle mesh error: a fast perception correlated distortion measure for fixed connectivity triangle meshes”. In: *Computer Graphics Forum*. Vol. 31. 5. Wiley Online Library. 2012, pp. 1715–1724.

[213] Vladimir Vovk. “Kernel ridge regression”. In: *Empirical inference*. Springer, 2013, pp. 105–116.

[214] Cuong T Vu, Eric C Larson, and Damon M Chandler. “Visual fixation patterns when judging image quality: Effects of distortion type, amount, and subject experience”. In: *2008 IEEE Southwest Symposium on Image Analysis and Interpretation*. IEEE. 2008, pp. 73–76.

[215] Hanli Wang, Jie Fu, Weisi Lin, Sudeng Hu, C-C Jay Kuo, and Lingxuan Zuo. “Image quality assessment based on local linear information and distortion-specific compensation”. In: *IEEE Transactions on image processing* 26.2 (2016), pp. 915–926.

[216] Kai Wang, Fakhri Torkhani, and Annick Montanvert. “A fast roughness-based approach to the assessment of 3D mesh visual quality”. In: *Computers & Graphics* 36.7 (2012), pp. 808–818.

[217] Qing Wang, Lin Xu, Qiang Chen, and Quansen Sun. “Import of distortion on saliency applied to image quality assessment”. In: *2014 IEEE International Conference on Image Processing (ICIP)*. IEEE. 2014, pp. 1165–1169.

[218] Weiming Wang, Haiyuan Chao, Jing Tong, Zhouwang Yang, Xin Tong, Hang Li, Xiuping Liu, and Ligang Liu. “Saliency-preserving slicing optimization for effective 3D printing”. In: *Computer Graphics Forum*. Vol. 34. 6. Wiley Online Library. 2015, pp. 148–160.

[219] Zhou Wang and Alan C Bovik. “Mean squared error: Love it or leave it? A new look at signal fidelity measures”. In: *IEEE signal processing magazine* 26.1 (2009), pp. 98–117.

[220] Zhou Wang and Alan C Bovik. “Modern image quality assessment”. In: *Synthesis Lectures on Image, Video, and Multimedia Processing* 2.1 (2006), pp. 1–156.

[221] Zhou Wang and Alan C Bovik. “Reduced-and no-reference image quality assessment”. In: *IEEE Signal Processing Magazine* 28.6 (2011), pp. 29–40.

[222] Zhou Wang, Alan C Bovik, and Brian L Evan. “Blind measurement of blocking artifacts in images”. In: *Proceedings 2000 International Conference on Image Processing (Cat. No. 00CH37101)*. Vol. 3. Ieee. 2000, pp. 981–984.

[223] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. “Image quality assessment: from error visibility to structural similarity”. In: *IEEE Transactions on Image Processing* 13.4 (2004), pp. 600–612.

[224] Benjamin Watson, Alinda Friedman, and Aaron McGaffey. “Measuring and predicting visual fidelity”. In: *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*. 2001, pp. 213–220.

[225] Yang Wen, Ying Li, Xiaohua Zhang, Wuzhen Shi, Lin Wang, and Jiawei Chen. “A weighted full-reference image quality assessment based on visual saliency”. In: *Journal of Visual Communication and Image Representation* 43 (2017), pp. 119–126.

[226] Heino Widdel. “Operational problems in analysing eye movements”. In: *Advances in psychology*. Vol. 22. Elsevier, 1984, pp. 21–29.

[227] Alexandre Willème, Saeed Mahmoudpour, Irene Viola, Karel Fliegel, Jakub Pospíšil, Touradj Ebrahimi, Peter Schelkens, Antonin Descampe, and Benoit Macq. "Overview of the JPEG XS core coding system subjective evaluations". In: *Applications of Digital Image Processing XLI*. Vol. 10752. International Society for Optics and Photonics. 2018, p. 107521M.

[228] Nathaniel Williams, David Luebke, Jonathan D Cohen, Michael Kelley, and Brenden Schubert. "Perceptually guided simplification of lit, textured meshes". In: *Proceedings of the 2003 symposium on Interactive 3D graphics*. 2003, pp. 113–121.

[229] Jian-Hua Wu, Shi-Min Hu, Chiew-Lan Tai, and Jia-Guang Sun. "An effective feature-preserving mesh simplification scheme based on face constriction". In: *Proceedings Ninth Pacific Conference on Computer Graphics and Applications. Pacific Graphics 2001*. IEEE. 2001, pp. 12–21.

[230] Jinjian Wu, Weisi Lin, Guangming Shi, and Anmin Liu. "Reduced-reference image quality assessment with visual information fidelity". In: *IEEE Transactions on Multimedia* 15.7 (2013), pp. 1700–1705.

[231] Long Xu, Weisi Lin, and C-C Jay Kuo. *Visual quality assessment by machine learning*. Springer, 2015.

[232] Shaoping Xu, Shunliang Jiang, and Weidong Min. "No-reference/blind image quality assessment: a survey". In: *IETE Technical Review* 34.3 (2017), pp. 223–245.

[233] Bincheng Yang and Hongwei Li. "A Visual Attention Model Based on Eye Tracking in 3D Scene Maps". In: *ISPRS International Journal of Geo-Information* 10.10 (2021), p. 664.

[234] Wei-Zu Yang, Liang-Chang Yu, Po-Chou Chen, and Tai-Liang Chen. "The design of multimedia web-based phone and billing system with freeware over the VoIP network". In: *IEEE International Conference on Sensor Networks, Ubiquitous, and Trustworthy Computing (SUTC'06)*. Vol. 1. IEEE. 2006, 4–pp.

[235] Junfeng Yao, Hongming Zhang, Hanhui Zhang, and Qingqing Chen. “R&D of a parameterized method for 3D virtual human body based on anthropometry”. In: *Age* 20.30 (2008), pp. 30–40.

[236] Hector Yee, Sumanita Pattanaik, and Donald P Greenberg. “Spatiotemporal sensitivity and visual attention for efficient rendering of dynamic environments”. In: *ACM Transactions on Graphics (TOG)* 20.1 (2001), pp. 39–65.

[237] Zeynep Cipiloglu Yildiz, Abdullah Bulbul, and Tolga Capin. “Modeling Human Perception of 3D Scenes”. In: *Intelligent Scene Modeling and Human-Computer Interaction*. Springer, 2021, pp. 67–88.

[238] Zeynep Cipiloglu Yildiz, A Cengiz Oztireli, and Tolga Capin. “A machine learning framework for full-reference 3D shape quality assessment”. In: *The Visual Computer* 36.1 (2020), pp. 127–139.

[239] Shivanthan Yohanandan, Andy Song, Adrian G Dyer, and Dacheng Tao. “Saliency preservation in low-resolution grayscale images”. In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 235–251.

[240] Daesub Yoon and N Hari Narayanan. “Mental imagery in problem solving: An eye tracking study”. In: *Proceedings of the 2004 symposium on Eye tracking research & applications*. 2004, pp. 77–84.

[241] Yuan Yuan, Qun Guo, and Xiaoqiang Lu. “Image quality assessment: a sparse learning way”. In: *Neurocomputing* 159 (2015), pp. 227–241.

[242] Gregory Zelinsky and David Sheinberg. “Why some search tasks take longer than others: Using eye movements to redefine reaction times”. In: *Studies in visual information processing*. Vol. 6. Elsevier, 1995, pp. 325–336.

[243] Yibing Zhan, Rong Zhang, and Qian Wu. “A structural variation classification model for image quality assessment”. In: *IEEE Transactions on Multimedia* 19.8 (2017), pp. 1837–1847.

[244] Lin Zhang, Ying Shen, and Hongyu Li. “VSI: A visual saliency-induced index for perceptual image quality assessment”. In: *IEEE Transactions on Image processing* 23.10 (2014), pp. 4270–4281.

[245] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. “The Unreasonable Effectiveness of Deep Features as a Perceptual Metric”. In: *CVPR*. 2018.

[246] Wei Zhang, Ali Borji, Zhou Wang, Patrick Le Callet, and Hantao Liu. “The application of visual saliency models in objective image quality assessment: A statistical evaluation”. In: *IEEE transactions on neural networks and learning systems* 27.6 (2015), pp. 1266–1278.

[247] Wei Zhang and Hantao Liu. “Toward a reliable collection of eye-tracking data for image quality research: challenges, solutions, and applications”. In: *IEEE Transactions on Image Processing* 26.5 (2017), pp. 2424–2437.

[248] Wei Zhang, Ralph R Martin, and Hantao Liu. “A saliency dispersion measure for improving saliency-based image quality metrics”. In: *IEEE Trans. Circuits Syst. Video Technol.* 28.6 (June 2018), pp. 1462–1466.

[249] Wei Zhang, Juan V Talens-Noguera, and Hantao Liu. “The quest for the integration of visual saliency models in objective image quality assessment: A distraction power compensated combination strategy”. In: *2015 IEEE International Conference on Image Processing (ICIP)*. Quebec City, QC, Canada: IEEE, Sept. 2015.

[250] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. “Object detectors emerge in deep scene cnns”. In: *arXiv preprint arXiv:1412.6856* (2014).

[251] Wujie Zhou, Gangyi Jiang, Mei Yu, Feng Shao, and Zongju Peng. “Reduced-reference stereoscopic image quality assessment based on view and disparity zero-watermarks”. In: *Signal Processing: Image Communication* 29.1 (2014), pp. 167–176.

[252] Jianqing Zhu, Huanqiang Zeng, Xin Jin, Yongzhao Du, Lixin Zheng, and Canhui Cai. “Joint horizontal and vertical deep learning feature for vehicle re-identification”. In: *Science China Information Sciences* 62.9 (2019), pp. 1–2.

[253] Qing Zhu, Junqiao Zhao, Zhiqiang Du, and Yeting Zhang. “Quantitative analysis of discrete 3D geometrical detail levels based on perceptual metric”. In: *Computers & Graphics* 34.1 (2010), pp. 55–65.

Appendices

Appendix A

Subjective Study of 3D Mesh Quality Scores in Virtual Reality

A.1 Distortion by Shape

A.2 Distortion by Type, Location and Levels

Table A.1: Comparison of MOS scores for both VR and desktop settings organised based on shape (Armadillo, Venus, Dyno and Rocker-Arm) and distortion type/location.

Distortion by shape	Category	VR MOS	Desktop MOS
Armadillo	Noise Uniform Low	0.4213	0.4333
Armadillo	Noise Uniform Medium	0.4694	0.4867
Armadillo	Noise Uniform High	0.4894	0.4950
Armadillo	Noise Rough Low	0.5149	0.4250
Armadillo	Noise Rough Medium	0.5340	0.4667
Armadillo	Noise Rough High	0.5915	0.6250
Armadillo	Noise Intermediate Low	0.4532	0.4917
Armadillo	Noise Intermediate Medium	0.5723	0.6750
Armadillo	Noise Intermediate High	0.6574	0.6917
Armadillo	Noise Smooth Low	0.4894	0.5500
Armadillo	Noise Smooth Medium	0.5574	0.6250
Armadillo	Noise Smooth High	0.6553	0.7000
Armadillo	Taubin Uniform Low	0.5170	0.4167
Armadillo	Taubin Uniform Medium	0.5894	0.4750
Armadillo	Taubin Uniform High	0.6149	0.6250
Armadillo	Taubin Rough Low	0.6383	0.4167
Armadillo	Taubin Rough Medium	0.6766	0.5583
Armadillo	Taubin Rough High	0.6787	0.7167
Armadillo	Taubin Intermediate Low	0.5845	0.6000
Armadillo	Taubin Intermediate Medium	0.6468	0.6167
Armadillo	Taubin Intermediate High	0.6617	0.7083

Distortion by shape	Category	VR MOS	Desktop MOS
Venus	**Noise Uniform Low**	**0.4489**	**0.4250**
Venus	**Noise Uniform Medium**	**0.4809**	**0.4667**
Venus	**Noise Uniform High**	**0.5021**	**0.5000**
Venus	**Noise Rough Low**	**0.5851**	**0.4917**
Venus	**Noise Rough Medium**	**0.6255**	**0.5083**
Venus	**NoiseRough High**	**0.6638**	**0.5917**
Venus	**Noise Intermediate Low**	**0.6106**	**0.4917**
Venus	**Noise Intermediate Medium**	**0.6426**	**0.5000**
Venus	**Noise Intermediate High**	**0.6723**	**0.5333**
Venus	**Noise Smooth Low**	**0.4872**	**0.4417**
Venus	**Noise Smooth Medium**	**0.5362**	**0.4417**
Venus	**Noise Smooth High**	**0.5553**	**0.5083**
Venus	**Taubin Uniform Low**	**0.5298**	**0.4167**
Venus	**Taubin Uniform Medium**	**0.5596**	**0.4583**
Venus	**Taubin Uniform High**	**0.5745**	**0.5917**
Venus	**Taubin Rough Low**	**0.6021**	**0.4500**
Venus	**Taubin Rough Medium**	**0.6128**	**0.6333**
Venus	**Taubin Rough High**	**0.6383**	**0.6417**
Venus	**Taubin Intermediate Low**	**0.6085**	**0.5583**
Venus	**Taubin Intermediate Medium**	**0.6191**	**0.6750**
Venus	**Taubin Intermediate High**	**0.6298**	**0.7583**

Distortion by shape	Category	VR MOS	Desktop MOS
Dyno	**Noise Uniform Low**	**0.4128**	**0.4417**
Dyno	**Noise Uniform Medium**	**0.4745**	**0.4667**
Dyno	**Noise Uniform High**	**0.5532**	**0.4833**
Dyno	**Noise Rough Low**	**0.5979**	**0.4667**
Dyno	**Noise Rough Medium**	**0.6511**	**0.5083**
Dyno	**Noise Rough High**	**0.6830**	**0.6000**
Dyno	**Noise Intermediate Low**	**0.6255**	**0.4583**
Dyno	**Noise Intermediate Medium**	**0.6277**	**0.4833**
Dyno	**Noise Intermediate High**	**0.6702**	**0.5417**
Dyno	**Noise Smooth Low**	**0.5574**	**0.4250**
Dyno	**Noise Smooth Medium**	**0.6447**	**0.4583**
Dyno	**Nois e Smooth High**	**0.6447**	**0.4667**
Dyno	**Taubin Uniform Low**	**0.6170**	**0.5333**
Dyno	**Taubin Uniform Medium**	**0.6383**	**0.6000**
Dyno	**Taubin Uniform High**	**0.6489**	**0.6750**
Dyno	**Taubin Rough Low**	**0.6362**	**0.5667**
Dyno	**Taubin Rough Medium**	**0.6511**	**0.6667**
Dyno	**Taubin Rough High**	**0.6766**	**0.7250**
Dyno	**Taubin Intermediate Low**	**0.6283**	**0.5750**
Dyno	**Taubin Intermediate Medium**	**0.6404**	**0.6333**
Dyno	**Taubin Intermediate High**	**0.6898**	**0.7917**

Distortion by shape	Category	VR MOS	Desktop MOS
Rocker-Arm	**Noise Uniform Low**	**0.5461**	**0.4417**
Rocker-Arm	**Noise Uniform Medium**	**0.5661**	**0.4667**
Rocker-Arm	**Noise Uniform High**	**0.5661**	**0.4917**
Rocker-Arm	**Noise Rough Low**	**0.6298**	**0.4250**
Rocker-Arm	**Noise Rough Medium**	**0.6298**	**0.4750**
Rocker-Arm	**Noise Rough High**	**0.6745**	**0.4833**
Rocker-Arm	**Noise Intermediate Low**	**0.6298**	**0.4333**
Rocker-Arm	**Noise Intermediate Medium**	**0.6723**	**0.4583**
Rocker-Arm	**Noise Intermediate High**	**0.6915**	**0.5667**
Rocker-Arm	**Noise Smooth Low**	**0.5661**	**0.4833**
Rocker-Arm	**Noise Smooth Medium**	**0.5661**	**0.5000**
Rocker-Arm	**Noise Smooth High**	**0.6723**	**0.5333**
Rocker-Arm	**Taubin Uniform Low**	**0.6553**	**0.4250**
Rocker-Arm	**Taubin Uniform Medium**	**0.6532**	**0.4583**
Rocker-Arm	**Taubin Uniform High**	**0.6723**	**0.5583**
Rocker-Arm	**Taubin Rough Low**	**0.6340**	**0.5917**
Rocker-Arm	**Taubin Rough Medium**	**0.6447**	**0.6333**
Rocker-Arm	**Taubin Rough High**	**0.6596**	**0.7667**
Rocker-Arm	**Taubin Intermediate Low**	**0.6660**	**0.5917**
Rocker-Arm	**Taubin Intermediate Medium**	**0.6745**	**0.6750**
Rocker-Arm	**Taubin Intermediate High**	**0.7068**	**0.7000**

Table A.2: Comparison of MOS scores for both VR and desktop settings organised based on distortion Type, Location and Levels (Noise Uniform, Noise Rough, Noise Intermediate, Noise Smooth, Taubin Uniform, Taubin Rough and Taubin Intermediate)

Distortion by Type and Location	Distortion Levels	VR MOS	Desktop MOS
Noise Uniform	Armadillo-Low	0.4213	0.4333
Noise Uniform	Armadillo-Medium	0.4694	0.4867
Noise Uniform	Armadillo-High	0.4894	0.4950
Noise Uniform	Venus-Low	0.4489	0.4250
Noise Uniform	Venus -Medium	0.4809	0.4667
Noise Uniform	Venus -High	0.5021	0.5000
Noise Uniform	Dinosaur-Low	0.4128	0.4417
Noise Uniform	Dinosaur-Medium	0.4745	0.4667
Noise Uniform	Dinosaur-High	0.5332	0.4833
Noise Uniform	RockerArm-Low	0.4681	0.4417
Noise Uniform	RockerArm-Medium	0.4702	0.4833
Noise Uniform	RockerArm-High	0.5364	0.4917

Distorton by Type and Location	Distortion Levels	VR MOS	Desktop MOS
Noise Rough	Armadillo-Low	0.5149	0.4250
Noise Rough	Armadillo-Medium	0.5340	0.4667
Noise Rough	Armadillo-High	0.5915	0.6228
Noise Rough	Venus-Low	0.5851	0.4917
Noise Rough	Venus -Medium	0.6255	0.5083
Noise Rough	Venus -High	0.6638	0.5917
Noise Rough	Dinosaur-Low	0.5979	0.4667
Noise Rough	Dinosaur-Medium	0.6511	0.5083
Noise Rough	Dinosaur-High	0.6830	0.6000
Noise Rough	RockerArm-Low	0.6298	0.4250
Noise Rough	RockerArm-Medium	0.6598	0.4750
Noise Rough	RockerArm-High	0.6745	0.4833

Distorton by Type and Location	Distortion Levels	VR MOS	Desktop MOS
Noise Intermediate	Armadillo-Low	0.5484	0.4917
Noise Intermediate	Armadillo-Medium	0.5723	0.6243
Noise Intermediate	Armadillo-High	0.6574	0.6543
Noise Intermediate	Venus-Low	0.6106	0.4917
Noise Intermediate	Venus -Medium	0.6426	0.5000
Noise Intermediate	Venus -High	0.6723	0.5333
Noise Intermediate	Dinosaur-Low	0.6255	0.4583
Noise Intermediate	Dinosaur-Medium	0.6277	0.4833
Noise Intermediate	Dinosaur-High	0.6702	0.5417
Noise Intermediate	RockerArm-Low	0.6298	0.4333
Noise Intermediate	RockerArm-Medium	0.6723	0.4583
Noise Intermediate	RockerArm-High	0.6915	0.5667

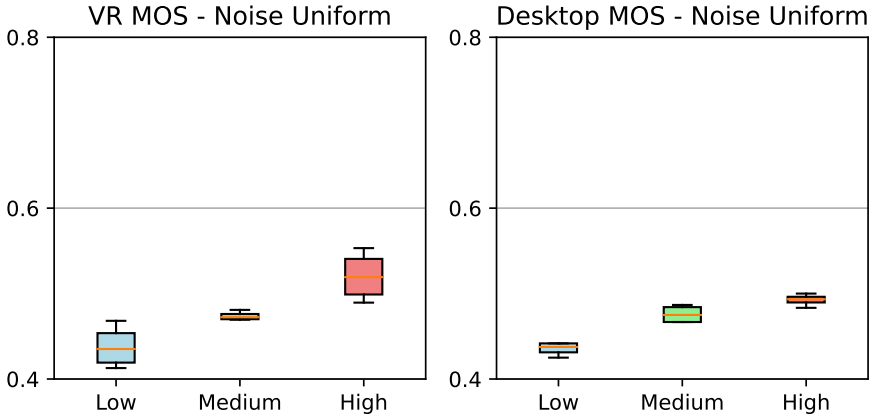
Distorton by Type and Location	Distortion Levels	VR MOS	Desktop MOS
Noise Smooth	Armadillo-Low	0.4894	0.5500
Noise Smooth	Armadillo-Medium	0.5574	0.6225
Noise Smooth	Armadillo-High	0.6553	0.6625
Noise Smooth	Venus-Low	0.4872	0.4417
Noise Smooth	Venus -Medium	0.5362	0.4617
Noise Smooth	Venus -High	0.5553	0.5083
Noise Smooth	Dinosaur-Low	0.5574	0.4250
Noise Smooth	Dinosaur-Medium	0.6247	0.4583
Noise Smooth	Dinosaur-High	0.6447	0.4667
Noise Smooth	RockerArm-Low	0.4830	0.4833
Noise Smooth	RockerArm-Medium	0.5553	0.5000
Noise Smooth	RockerArm-High	0.6723	0.5333

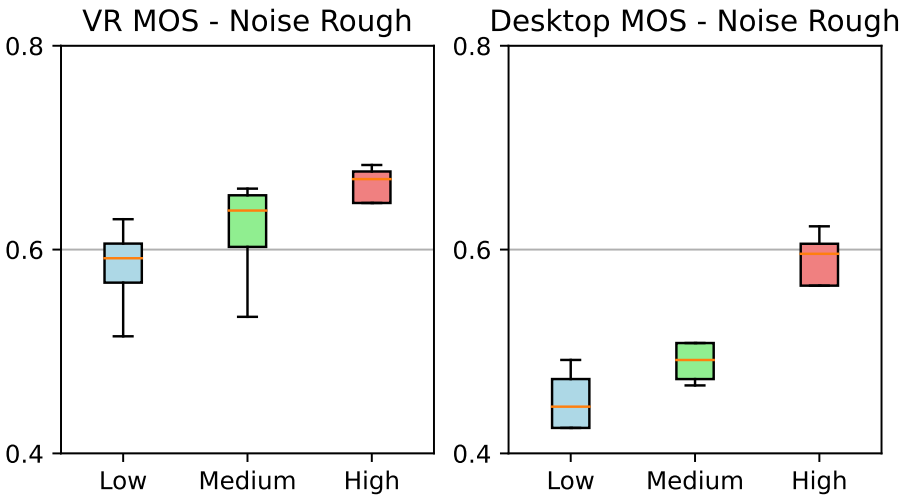
Distorton by Type and Location	Distortion Levels	VR MOS	Desktop MOS
Taubin Uniform	Armadillo-Low	0.5170	0.4167
Taubin Uniform	Armadillo-Medium	0.5894	0.4750
Taubin Uniform	Armadillo-High	0.6149	0.6250
Taubin Uniform	Venus-Low	0.5298	0.4167
Taubin Uniform	Venus -Medium	0.5596	0.4583
Taubin Uniform	Venus -High	0.5745	0.5917
Taubin Uniform	Dinosaur-Low	0.6170	0.5333
Taubin Uniform	Dinosaur-Medium	0.6383	0.6000
Taubin Uniform	Dinosaur-High	0.6489	0.6750
Taubin Uniform	RockerArm-Low	0.6532	0.4250
Taubin Uniform	RockerArm-Medium	0.6553	0.4583
Taubin Uniform	RockerArm-High	0.6723	0.5583

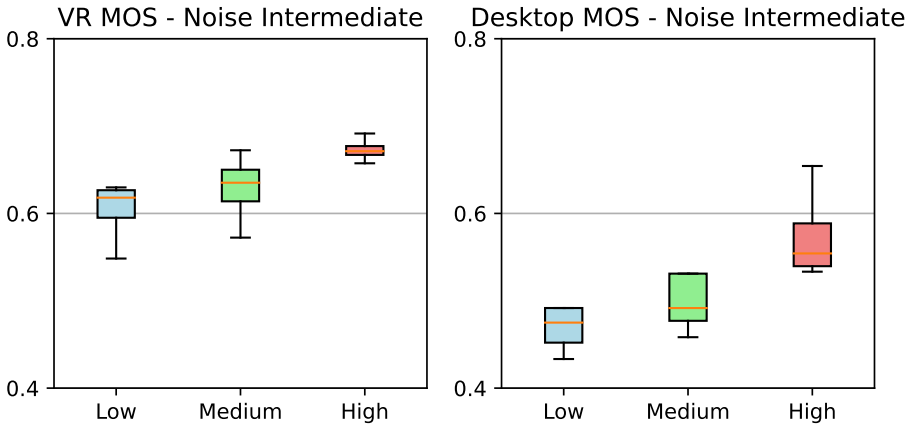
Distorton by Type and Location	Distortion Levels	VR MOS	Desktop MOS
Taubin Rough	Armadillo-Low	0.6383	0.4167
Taubin Rough	Armadillo-Medium	0.6766	0.5583
Taubin Rough	Armadillo-High	0.6787	0.7167
Taubin Rough	Venus-Low	0.6021	0.4500
Taubin Rough	Venus -Medium	0.6128	0.6333
Taubin Rough	Venus -High	0.6383	0.6417
Taubin Rough	Dinosaur-Low	0.6362	0.5667
Taubin Rough	Dinosaur-Medium	0.6511	0.6667
Taubin Rough	Dinosaur-High	0.6766	0.7250
Taubin Rough	RockerArm-Low	0.6340	0.5917
Taubin Rough	RockerArm-Medium	0.6447	0.6333
Taubin Rough	RockerArm-High	0.6596	0.7667

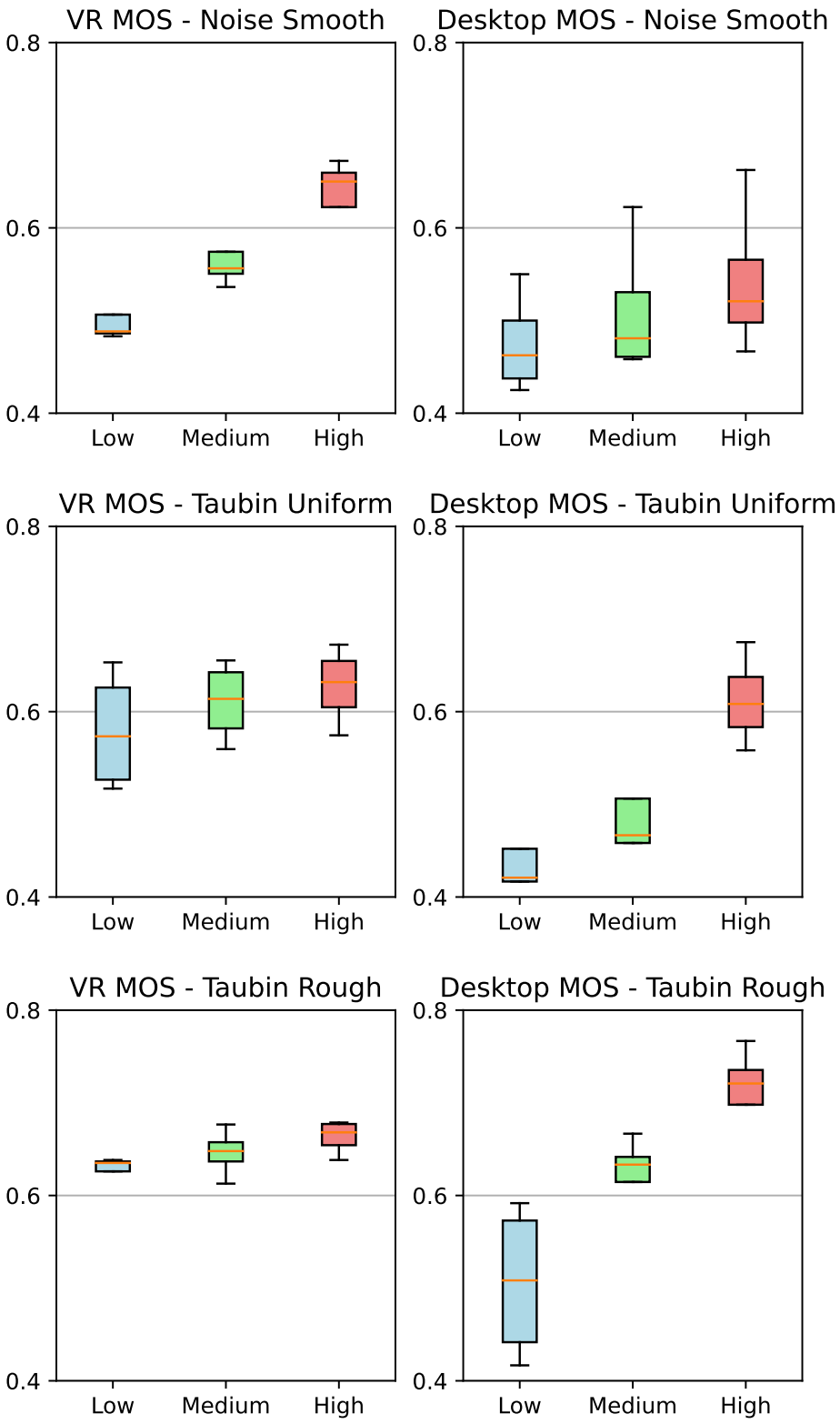
Distortion by Type and Location	Distortion Levels	VR MOS	Desktop MOS
Taubin Intermediate	**Armadillo-Low**	**0.6468**	**0.6000**
Taubin Intermediate	**Armadillo-Medium**	**0.6617**	**0.6083**
Taubin Intermediate	**Armadillo-High**	**0.6745**	**0.6167**
Taubin Intermediate	**Venus-Low**	**0.6085**	**0.5583**
Taubin Intermediate	**Venus -Medium**	**0.6191**	**0.6750**
Taubin Intermediate	**Venus -High**	**0.6298**	**0.7583**
Taubin Intermediate	**Dinosaur-Low**	**0.6298**	**0.5750**
Taubin Intermediate	**Dinosaur-Medium**	**0.6383**	**0.6333**
Taubin Intermediate	**Dinosaur-High**	**0.6404**	**0.7917**
Taubin Intermediate	**RockerArm-Low**	**0.6468**	**0.5917**
Taubin Intermediate	**RockerArm-Medium**	**0.6745**	**0.6750**
Taubin Intermediate	**RockerArm-High**	**0.6660**	**0.7000**

The figures below compare MOS scores averaged over all the shapes for both VR and desktop settings for each type of location with changing level of distortion strength.








VR MOS - Noise Smooth
Desktop MOS - Noise Smooth
VR MOS - Taubin Uniform
Desktop MOS - Taubin Uniform
VR MOS - Taubin Rough
Desktop MOS - Taubin Rough
0.8
0.6
0.4
Low
Medium
High

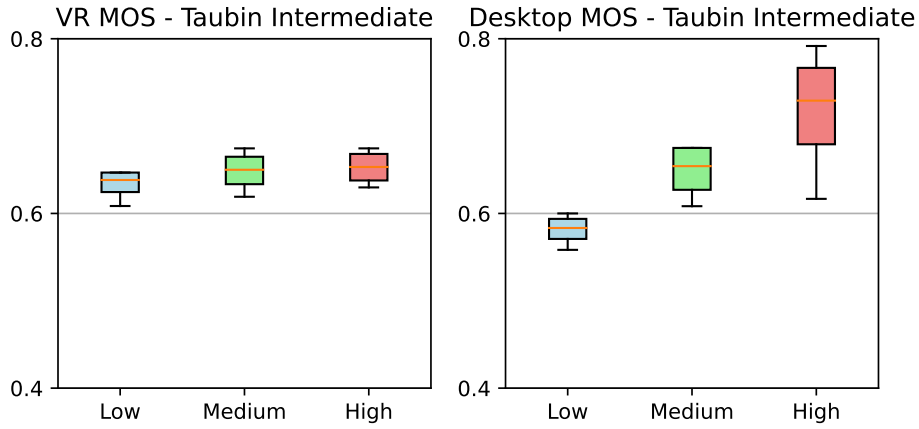


Figure A1: Comparison of MOS scores averaged over all the shapes for both VR and desktop settings for each type of location with changing level of distortion strength.

Appendix B

Learning to Predict 3D Mesh Saliency

The following meshes were taken from the Stanford repository found at:

B.1 Dataset

`http://graphics.stanford.edu/data/3Dscanrep/`

Armadillo, Bunny, Dragon, happy

`http://graphics.cs.williams.edu/data/meshes.xml#2`

Head

`http://graphics.cs.williams.edu/data/meshes.xml#2`

Teapot

The following meshes were taken from the SHREC 2011 shape retrieval contest dataset found at:

`http://www.itl.nist.gov/iad/vug/sharp/contest/2011/NonRigid/data.html`

Ant and Falling

The following mesh was taken from Keenan's 3D model repository which can be

found at:

`http://www.cs.cmu.edu/~kmcrane/Projects/ModelRepository/`

Spot

The following meshes were taken from the AIM@SHAPE-VISIONAIR shape repository which can be found at:

`http://visionair.ge.imati.cnr.it/ontologies/shapes`

- Bulldog model is provided courtesy of VCG-ISTI
- Frog model is provided courtesy of Frank-terHaar
- Kitten model is provided courtesy of Frank-terHaar
- Red circular box model is provided courtesy of INRIA
- Chair model is provided courtesy of IMATI
- Cup model is provided courtesy of MPII
- Sheep model is provided courtesy of Frank-terHaar
- Ramesses model is provided courtesy of IMATI
- Gargoyle model is provided courtesy of VCG-ISTI
- Raptor model is provided courtesy of INRIA
- Grog model is provided courtesy of VCG-ISTI

B.2 Related Saliency Models and Geometric Features

Our learning-based method for mesh saliency prediction builds on top of existing geometric features and saliency models. Existing saliency models are mostly based on geometric characteristics, and some also take global/local geometric information into account.

B.2.1 Geometric Characteristics for Mesh Saliency

One of the mesh saliency measuring methods has been used by Lee et al. [115]. It works based on the concept that salient regions are likely to be the areas that express unfamiliar geometric characteristics compared to the surrounding regions. Their approach first calculates mean curvature on mesh vertices. To perform multiscale analysis, they apply Gaussian filters of different scales to the mean curvatures. The centre-surround analysis is then achieved via Differences of Gaussian. Finally, multiscale results are aggregated to form the mesh saliency map.

Song et al. [192] propose a methodology for mesh saliency, perceptually based on the importance of a surrounding area on a 3D surface mesh. This methodology incorporates global context by use of the spectral attributes of the mesh, whereas in contrast, most existing measures are restricted to local geometric cues. Moreover, Song's method [192] takes a set of meshes as a group of meshes simplified to different degrees and calculates the saliency map for each scale by calculating the saliency of the spectral mesh for each scale. The saliency maps of the scale are then combined to generate a final saliency map.

B.2.2 Geometric Feature Extraction

This section discusses geometric features used to obtain a good feature representation of the meshes. We compute different types of features: Mesh SIFT and SHOT.

B.2.2.1 Mesh SIFT

The scale-invariant feature transform (SIFT) is a computer vision feature recognition algorithm for detecting and defining local features in images [130]. It converts image data into scale-invariant coordinates relative to local characteristics. An important aspect of SIFT is that it creates large numbers of features that cover the image over a large range

of sizes and locations. Mesh SIFT extends image-based SIFT descriptors to 3D meshes developed by [49]. Here's a simplified overview of these multi-step algorithms:

- **Saliency Map Calculation:** The saliency $S(v)$ at a vertex v can be computed based on curvature:

$$S(v) = |H(v) - \text{avg}(H(N(v)))|$$

where $H(v)$ is the mean curvature at vertex v, $N(v)$ is the set of vertices in the neighborhood of v, and $\text{avg}(H(N(v)))$ is the average curvature of this neighborhood.

- **Key Point Detection:** The keypoints K can be determined based on saliency:

$$K = \{v \in M \mid S(v) > T\}$$

where T is a saliency threshold, and M is the set of all vertices in the mesh.

- **Scale-space extrema detection:** Let $G(.,.,\sigma)$ denote a 3D Gaussian function with standard deviation σ. Then for each keypoint k at each scale σ, compute the blurred curvature $H(k,\sigma) = G(k.x, k.y, \sigma)H(k)$. Then find the scale σ at which $H(k,\sigma)$ is maximised.

- **Orientation assignment:** Assign an orientation to each keypoint based on the gradient of the mesh. Let $\nabla H(k)$ denote the gradient of the mesh at keypoint k. Then the orientation $O(k)$ of the keypoint can be defined as the direction of $\nabla H(k)$.

- **Descriptor calculation:** For each keypoint, calculate a descriptor that captures the local shape. This could be a histogram $H(k,\theta)$ of the distribution of orientations θ in the local neighborhood of k, weighted by the saliency

$$H(k,\theta) = \sum_{v \in N(k)} S(v)\delta(\theta - O(v))$$

B.2.2.2 SHOT Descriptor

According to [208], the signature and histogram of orientation (SHOT) 3D descriptor can be classified as histograms or signatures. It produces a local representation that is efficient, descriptive, and robust to noise and clutter and variations in point density. When working with triangular meshes, consider the vertices as points in 3D data. The normals at each vertex can be calculated based on adjacent triangles. Calculate the normals at each vertex of a triangular mesh by using adjacent triangles as points in the 3D data. There are some steps that the SHOT descriptor should follow: Establishing Local Reference Frames (LRFs), computing histograms and concatenating the histograms. In the case of using SHOT in mesh saliency first calculate the SHOT descriptor for each vertex. Let's represent the SHOT descriptor for a vertex V_i as $SHOT(V_i)$. then calculate the dissimilarity between these descriptors, which could use a distance metric such as Euclidean distance. using Euclidean distance, the dissimilarity D between two vertices V_i and V_j could be represented as $D(V_i, V_j) = \|SHOT(V_i) - SHOT(V_j)\|$. after that generate the saliency map. A simple way to generate a saliency score S for a vertex V_i would be to sum the dissimilarities between V_i and all other vertices in its neighbourhood N. This could be represented as $S(V_i) = \sum_{V_j \in N} D(V_i, V_j)$.

B.3 Implementation Details for 3D Mesh Saliency

In this section, we gave more details to predict 3D mesh saliency with ground truth.

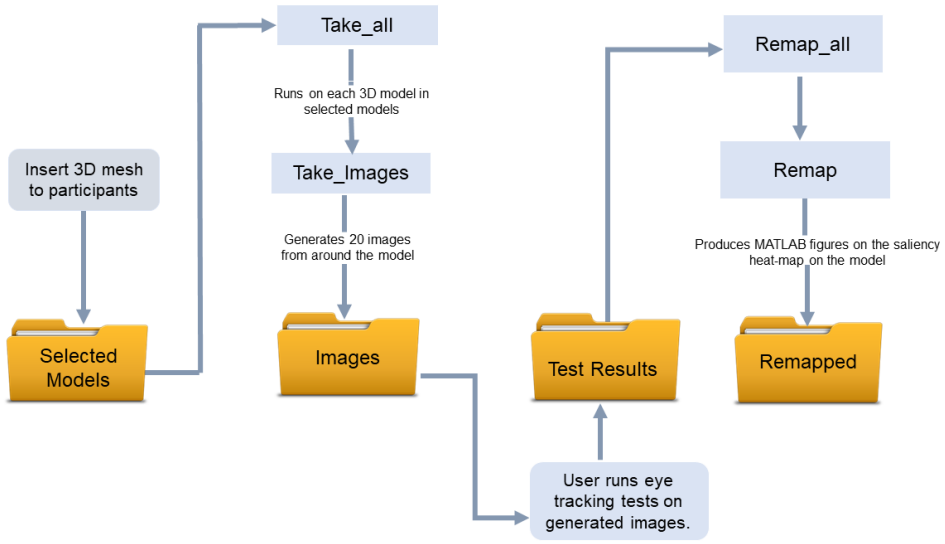


Figure B1: Illustrate data flow for eye tracking stimuli and remapping scripts.

This experiment uses toolbox graphs from MathWorks to plot the 3D mesh shapes. These toolboxes will help to read any file format related to an image. We used OBJ.format, so we placed all the 3D models in *Selected_Models* folder. The script *Take_all* use all the 3D models in the *Selected_Models* folder. then, the *Take_images* generates 20 images, and each shape has 20 views. These images will then be taken to run the eye-tracking experiment, and therefore the results of these experiments should be placed in the *Test_Results* folder. Then, *Remap_all* can decision remap on every model in the *Selected_Models* folder. *Remap* which can output a MATLAB figure of every model with a saliency heat-map and a comparison figure with the heat-map before and after normalising.

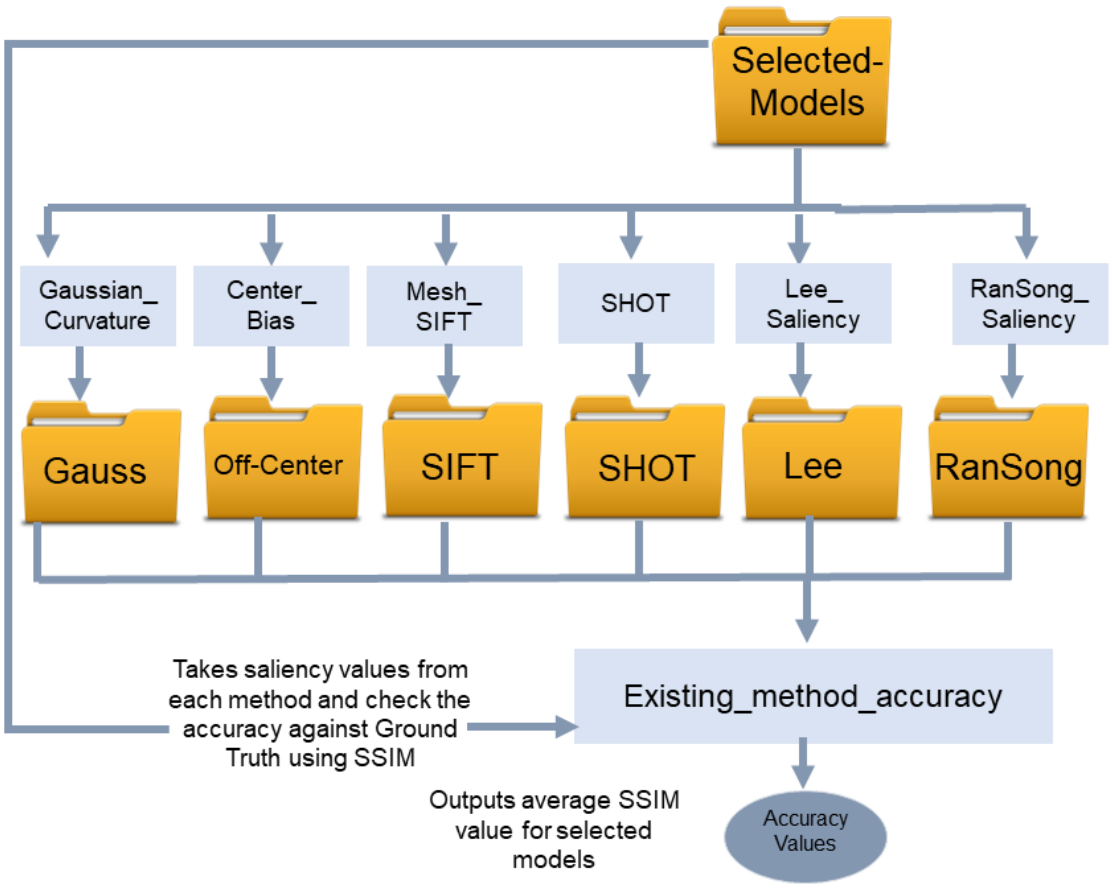


Figure B2: Illustrate data flow for existing method and evaluating scripts..

This Figure shows the flowchart of how we generate the saliency value from each method. The *Existing_method_accuracy* script takes a file of model names, and for every method there on the file. The two maps being approved into the *SSIM_mesh_ helper* are going to be the methods being tested and the ground truth data from the eye-tracking experiment. Later called *SSIM_for_3D_Saliency* script which runs the modified SSIM script that works on 3D meshes.

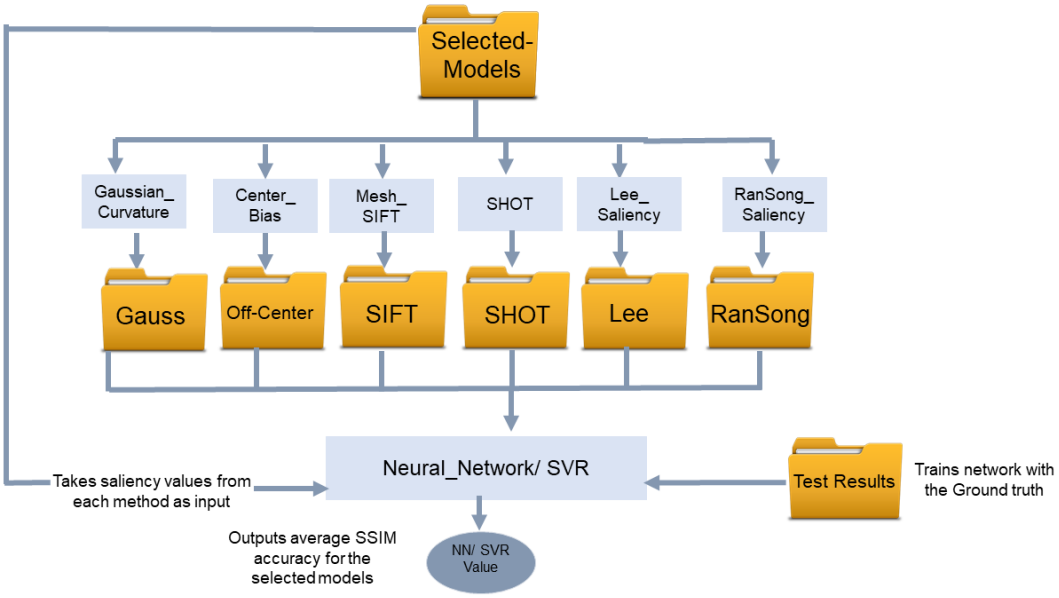


Figure B3: Illustrate data flow for existing method and evaluating scripts..

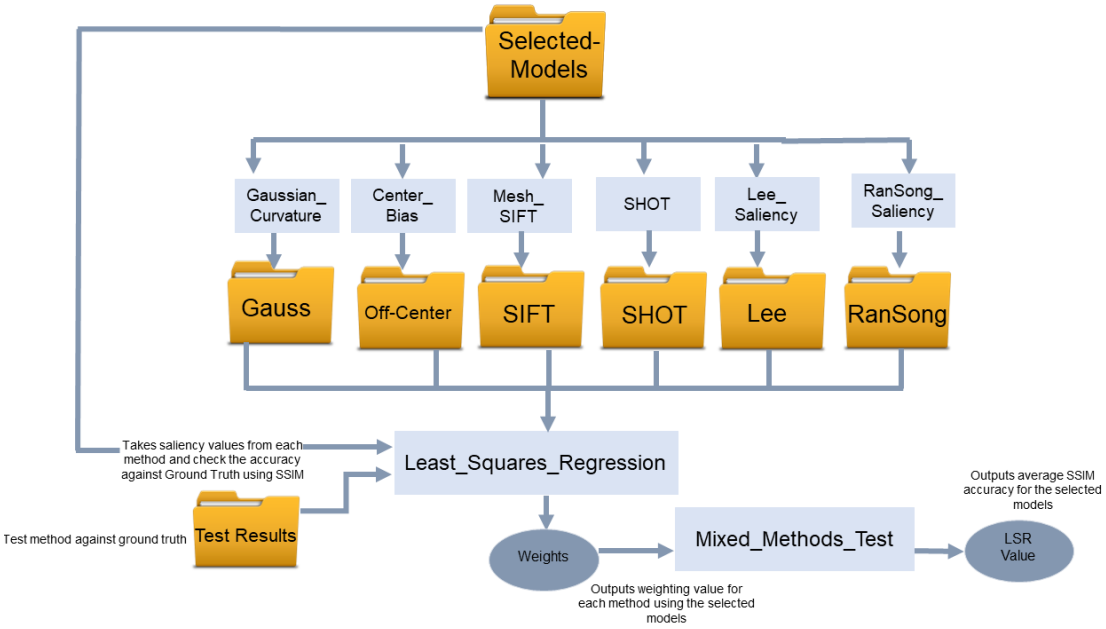


Figure B4: Illustrate data flow for existing method and evaluating scripts.

The *Least_Square_Regression* the least squares regression method feeds it the dependent relative data in being the eye-tracking experiment and the independent data being the method saliency maps of the selected models. Then sign the weighting for every method and an intercept that minimises the square error for the eye-tracking data. The *Least_Square_Regression* script then outputs a weight array with one value corresponding to each passed-in method. The *Least_Square_Regression* fit is applied to the data to calculate the weights array. To do this, all existing method saliency maps are loaded as the independent variables, and the ground truth is loaded as the dependent variable. The weight can then be approved to the *mixed_method_test* script, which combines each salience map linearly scaled with the weight.