



HAL
open science

Visual quality of rendered 3D meshes with color attributes: Subjective and objective evaluation

Yana Nehme

► **To cite this version:**

Yana Nehme. Visual quality of rendered 3D meshes with color attributes: Subjective and objective evaluation. Other [cs.OH]. Université de Lyon, 2021. English. NNT : 2021LYSEI086 . tel-03670835

HAL Id: tel-03670835

<https://theses.hal.science/tel-03670835v1>

Submitted on 17 May 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



N°d'ordre NNT : 2021LYSEI086

THESE de DOCTORAT **DE L'UNIVERSITE DE LYON**
opérée au sein de
INSA Lyon

Ecole Doctorale N° 512
INFORMATIQUE ET MATHEMATIQUES DE LYON

Spécialité / discipline de doctorat :
Informatique

Soutenue publiquement le 03/12/2021, par :
Yana NEHME

Visual Quality of Rendered 3D Meshes with Color Attributes: Subjective and Objective Evaluation

Devant le jury composé de :

MORIN Luce	Professeur	INSA Rennes	Rapporteure
SMOLIC Aljosa	Professeur	Trinity College Dublin	Rapporteur
ALLIEZ Pierre	Directeur de Recherche	INRIA	Examineur
MANTIUK Rafal	Reader	University of Cambridge	Examineur
EGLIN Véronique	Professeur	INSA Lyon	Examinatrice
LAVOUE Guillaume	Professeur	Ecole Centrale de Lyon	Directeur de thèse
LE CALLET Patrick	Professeur	Université De Nantes	Co-directeur de thèse
DUPONT Florent	Professeur	Université Lyon 1	Co-directeur de thèse
FARRUGIA Jean-Philippe	Maître de conférences	IUT Lyon 1	Invité

Département FEDORA – INSA Lyon - Ecoles Doctorales

SIGLE	ECOLE DOCTORALE	NOM ET COORDONNEES DU RESPONSABLE
CHIMIE	CHIMIE DE LYON https://www.edchimie-lyon.fr Sec. : Renée EL MELHEM Bât. Blaise PASCAL, 3e étage secretariat@edchimie-lyon.fr	M. Stéphane DANIELE C2P2-CPE LYON-UMR 5265 Bâtiment F308, BP 2077 43 Boulevard du 11 novembre 1918 69616 Villeurbanne directeur@edchimie-lyon.fr
E.E.A.	ÉLECTRONIQUE, ÉLECTROTECHNIQUE, AUTOMATIQUE https://edeea.universite-lyon.fr Sec. : Stéphanie CAUVIN Bâtiment Direction INSA Lyon Tél : 04.72.43.71.70 secretariat.edeea@insa-lyon.fr	M. Philippe DELACHARTRE INSA LYON Laboratoire CREATIS Bâtiment Blaise Pascal, 7 avenue Jean Capelle 69621 Villeurbanne CEDEX Tél : 04.72.43.88.63 philippe.delachartre@insa-lyon.fr
E2M2	ÉVOLUTION, ÉCOSYSTÈME, MICROBIOLOGIE, MODÉLISATION http://e2m2.universite-lyon.fr Sec. : Sylvie ROBERJOT Bât. Atrium, UCB Lyon 1 Tél : 04.72.44.83.62 secretariat.e2m2@univ-lyon1.fr	M. Philippe NORMAND Université Claude Bernard Lyon 1 UMR 5557 Lab. d'Ecologie Microbienne Bâtiment Mendel 43, boulevard du 11 Novembre 1918 69 622 Villeurbanne CEDEX philippe.normand@univ-lyon1.fr
EDISS	INTERDISCIPLINAIRE SCIENCES-SANTÉ http://ediss.universite-lyon.fr Sec. : Sylvie ROBERJOT Bât. Atrium, UCB Lyon 1 Tél : 04.72.44.83.62 secretariat.ediss@univ-lyon1.fr	Mme Sylvie RICARD-BLUM Institut de Chimie et Biochimie Moléculaires et Supramoléculaires (ICBMS) - UMR 5246 CNRS - Université Lyon 1 Bâtiment Raulin - 2ème étage Nord 43 Boulevard du 11 novembre 1918 69622 Villeurbanne Cedex Tél : +33(0)4 72 44 82 32 sylvie.ricard-blum@univ-lyon1.fr
INFOMATHS	INFORMATIQUE ET MATHÉMATIQUES http://edinfomaths.universite-lyon.fr Sec. : Renée EL MELHEM Bât. Blaise PASCAL, 3e étage Tél : 04.72.43.80.46 infomaths@univ-lyon1.fr	M. Hamamache KHEDDOUCI Université Claude Bernard Lyon 1 Bât. Nautibus 43, Boulevard du 11 novembre 1918 69 622 Villeurbanne Cedex France Tél : 04.72.44.83.69 hamamache.kheddouci@univ-lyon1.fr
Matériaux	MATÉRIAUX DE LYON http://ed34.universite-lyon.fr Sec. : Yann DE ORDENANA Tél : 04.72.18.62.44 yann.de-ordenana@ec-lyon.fr	M. Stéphane BENAYOUN Ecole Centrale de Lyon Laboratoire LTDS 36 avenue Guy de Collongue 69134 Ecully CEDEX Tél : 04.72.18.64.37 stephane.benayoun@ec-lyon.fr
MEGA	MÉCANIQUE, ÉNERGÉTIQUE, GÉNIE CIVIL, ACOUSTIQUE http://edmega.universite-lyon.fr Sec. : Stéphanie CAUVIN Tél : 04.72.43.71.70 Bâtiment Direction INSA Lyon mega@insa-lyon.fr	M. Jocelyn BONJOUR INSA Lyon Laboratoire CETHIL Bâtiment Sadi-Carnot 9, rue de la Physique 69621 Villeurbanne CEDEX jocelyn.bonjour@insa-lyon.fr
ScSo	ScSo* https://edsciencessociales.universite-lyon.fr Sec. : Mélina FAVETON INSA : J.Y. TOUSSAINT Tél : 04.78.69.77.79 melina.faveton@univ-lyon2.fr	M. Christian MONTES Université Lumière Lyon 2 86 Rue Pasteur 69365 Lyon CEDEX 07 christian.montes@univ-lyon2.fr

*ScSo : Histoire, Géographie, Aménagement, Urbanisme, Archéologie, Science politique, Sociologie, Anthropologie

Abstract

As technological advances and capabilities in the field of computer graphics grow day by day, the need to master the visualization and processing of 3D data increases at an equal pace. Indeed, the development of modeling software and acquisition devices makes 3D graphics rich and realistic: complex models enriched with various appearance attributes. The way this 3D content is consumed is also evolving from standard screens to Virtual and Mixed Reality (VR/MR). However, the size and complexity of these rich 3D models often make their interactive visualization problematic. This is particularly the case in immersive environments and online applications. Thus, to avoid latency and rendering issues, diverse processing operations, including simplification and compression, are usually applied, resulting in a loss of quality in the final rendering. Therefore, both subjective studies and objective metrics are needed to predict this visual loss and to assess the quality as perceived by human observers.

In this thesis, we address the aforementioned challenges. We conduct an extensive study to determine the best subjective quality assessment methodology to adopt for assessing the visual quality of 3D graphics, especially in VR. We establish two quality assessment datasets composed of meshes with vertex colors and textured meshes, respectively. The former is produced in VR and the latter in crowdsourcing. To the best of our knowledge, these are the largest datasets for meshes with color attributes to date. Moreover, we provide an in-depth analysis of the influence of source model characteristics, distortion interactions, viewpoints and animations on the perceived quality of 3D meshes. Leveraging our two established datasets, we propose two data-driven perceptual metrics for quality assessment of 3D graphics with color attributes. The first metric is model-based while the second is an image-based metric that employs convolutional neural networks. Our metrics demonstrate state-of-the-art results on our two datasets. Lastly, we investigate how incorporating visual attention into our perceptual quality metric improves the predicted quality. The datasets and the source code of the metrics are publicly available.

Keywords: Computer Graphics, Visual Quality Assessment, Subjective Quality Evaluation, Objective Quality Evaluation, 3D Graphics, 3D Meshes, Color Attributes, Textures, Vertex Colors, Subjective Methodologies, Datasets, Perceptual Metrics, CNN, Visual Attention, Virtual Reality, Crowdsourcing.

Résumé

Au fur et à mesure que le progrès technologique et les capacités dans le domaine de l'informatique graphique augmentent de jour en jour, le besoin de maîtriser la visualisation et le traitement des données 3D augmente au même rythme. En effet, le développement des logiciels de modélisation et des dispositifs d'acquisition rend les graphiques 3D riches et réalistes: des modèles complexes enrichis de divers attributs d'apparence. Le mode de consommation du contenu 3D évolue également, passant des écrans standards à la Réalité Virtuelle et Mixte (VR/MR). Cependant, la taille et la complexité des riches modèles 3D rendent souvent leur visualisation interactive problématique. C'est notamment le cas dans les environnements immersifs et les applications en ligne. Ainsi, pour éviter les problèmes de latence et de rendu, diverses opérations de traitement, dont la simplification et la compression, sont généralement appliquées aux modèles 3D, entraînant une perte de qualité dans le rendu final. Par conséquent, des études subjectives et des mesures objectives sont nécessaires pour prédire cette perte visuelle et évaluer la qualité telle que perçue par les observateurs humains.

Dans cette thèse, nous abordons les défis de l'évaluation de la qualité visuelle des graphiques 3D rendus. Nous sommes particulièrement intéressés par les maillages 3D avec des attributs de couleur, soit sous la forme de cartes de texture ou de couleurs par sommet. À cette fin, des méthodes subjectives et objectives d'évaluation de la qualité ont été proposées et les facteurs d'influence sous-jacents ont été étudiés et discutés.

Dans le domaine de l'évaluation subjective de la qualité, nous avons abordé le problème du manque de consensus sur les méthodologies appropriées à l'évaluation de la qualité des graphiques 3D. Nous avons comparé trois des méthodologies subjectives les plus répandues en traitement d'images, présentant des références cachées (méthode ACR-HR) et des références explicites (méthodes DSIS et SAMVIQ). Nous avons évalué leurs performances sur un ensemble de données de 80 modèles 3D colorés, altérés par diverses distorsions dans un contexte de réalité virtuelle. Contrairement à l'évaluation de la qualité des images et des vidéos naturelles, les résultats montrent que la présence d'une référence explicite est nécessaire pour l'évaluation de la qualité des graphiques 3D, puisque les gens ont moins de connaissances préalables sur la qualité de ces données que sur celle des images naturelles. DSIS semble être la méthode la plus appropriée pour évaluer la qualité des graphiques 3D. Elle est la méthode la plus précise et surtout la plus efficace en termes de temps. Nous recommandons d'utiliser des groupes d'au moins 24 observateurs pour la méthodologie DSIS. Cette étude constitue un premier pas vers la standardisation d'une méthodologie d'évaluation de qualité des graphiques 3D en réalité virtuelle.

Nous fournissons à la communauté deux bases de données publiques d'évaluation de la

qualité des maillages avec des attributs de couleur. La première base de données comprend 480 maillages avec des couleurs par sommet, générés à partir de 5 modèles sources associés chacun à 3 points de vue et 2 courtes animations. Il s'agit de la première base de données publique produit en VR pour de telles données. Cet ensemble de données nous a permis de tirer des conclusions intéressantes concernant l'influence des points de vue et des animations sur les scores de qualité et leur intervalles de confiance. La seconde base de données est la plus grande base de données d'évaluation de la qualité des maillages texturés à ce jour. Elle comprend 55 modèles sources altérés par des combinaisons de 5 types de distorsions appliquées sur la géométrie et la texture des maillages. Au total, plus de 343k stimuli dégradés ont été générés, dont 3000 sont associés à des scores subjectifs dérivés d'une expérience subjective menée en crowdsourcing et le reste à des scores subjectifs prédits. En s'appuyant sur cette base de données, nous présentons des analyses approfondies sur l'impact de chaque distorsion ainsi que celui de leurs combinaisons sur la qualité perçue. Nous évaluons également l'influence de la complexité de la géométrie, de la couleur et des coutures de la texture sur la perception des distorsions. Concernant la caractérisation du contenu 3D, nous proposons trois mesures, basées sur l'information spatiale et la complexité de l'attention visuelle, pour caractériser quantitativement la complexité géométrique, colorimétrique et sémantique des modèles 3D.

Enfin, nous étudions la fiabilité des expériences de crowdsourcing (CS) pour évaluer la qualité des graphiques 3D et si elles peuvent atteindre la précision des tests en laboratoire. Les résultats ont montré que, dans des conditions contrôlées et avec une approche appropriée de sélection et de filtrage des participants, une expérience en CS basée sur la méthode DSIS peut être aussi précise qu'une expérience réalisée en laboratoire.

Concernant l'évaluation objective de la qualité, la plupart des métriques dans la littérature n'évaluent que les distorsions géométriques. Lorsqu'il s'agit de maillages avec des informations de couleur, peu de travaux ont été publiés. Par conséquent, nous proposons deux nouvelles métriques d'évaluation de la qualité perceptuelle qui prennent en compte à la fois les attributs de géométrie et de couleur.

La première métrique que nous proposons est une métrique à référence complète basée modèle (model-based full-reference), développée pour les maillages 3D avec des couleurs par sommet et fonctionnant entièrement sur le domaine du maillage. Elle incorpore des caractéristiques géométriques et colorimétriques pertinentes sur le plan perceptuel. L'ensemble optimal de caractéristiques est sélectionné par régression logistique et tests de validation croisée. Une version adaptée de cette métrique est également proposée pour les nuages de points colorés. C'est la première métrique de qualité des nuages de points qui prend en compte à la fois la géométrie et la couleur. Nous étendons notre métrique en combinant ses caractéristiques géométriques et couleurs avec une mesure de la complexité de l'attention visuelle, basée sur la dispersion de la saillance visuelle. Nous montrons que l'incorporation de l'attention visuelle dans notre métrique améliore la prédiction de la qualité visuelle.

La deuxième métrique que nous présentons est une métrique de qualité à référence complète basée image (image-based full-reference) qui utilise des réseaux de neurones convolutifs. Elle est calculée sur des instantanés rendus des modèles 3D et utilise l'architecture AlexNet avec des poids linéaires appris par-dessus. Le réseau est alimenté par des patches de

référence et des patches dégradés. La qualité globale du modèle est dérivée de la moyenne des qualités des patches locaux.

Nos métriques produisent des résultats de pointe sur nos deux bases de données et surpassent les autres métriques de qualité basées images.

Les bases de données et le code source des métriques sont accessibles au public.

Mots-clés: Informatique Graphique, Évaluation de la Qualité Visuelle, Évaluation de la Qualité Subjective, Évaluation de la Qualité Objective, Graphiques 3D, Maillages 3D, Attributs de Couleur, Textures, Couleurs par Sommets, Méthodologies Subjectives, Bases de Données, Métriques Perceptuelles, Réseaux de Neurones Convolutifs, Attention Visuelle, Réalité Virtuelle, Crowdsourcing.

Remerciements

Par les lignes suivantes, je tiens à remercier toutes les personnes qui ont contribué directement ou indirectement à cette thèse en espérant de n'oublier personne.

Je tiens à remercier dans un premier temps mes rapporteurs, Luce Morin et Aljosa Smolic, d'avoir accepté ce rôle. Merci pour votre relecture attentive et vos retours très positifs. Je remercie Pierre Alliez, Rafal Mantiuk et Véronique Eglin d'avoir été les examinateurs de ma thèse.

Un grand merci pour mon directeur de thèse Guillaume Lavoué. Merci de m'avoir donné l'opportunité de faire cette thèse et de m'avoir fait confiance. J'ai beaucoup apprécié ton encadrement, et en particulier ta pédagogie. Tu m'as appris ce qu'est la vraie recherche. Tu as toujours été disponible malgré ton emploi du temps, et j'en suis très reconnaissante. Merci également de m'avoir donné l'opportunité d'enseigner durant ma thèse Je remercie mes co-directeurs Florent Dupont, Jean-Philippe Farrugia et Patrick Le Callet. Vous m'avez été d'une grande aide à de nombreuses occasions, notamment Jean-Philippe lors du développement de nos expériences en réalité virtuelle, Florent pour ta bienveillance et ton appréciation de mon travail lorsque je le sous-estimais, et Patrick pour toutes tes idées innovantes et ton point de vue perspicace qui m'ont permis de me dépasser et de viser plus loin. Merci Guillaume, Patrick, Florent et Jean-Philippe pour ces 3 années de collaboration Je suis honorée d'avoir pu travailler avec vous. Nous avons formé une super équipe et c'est la raison pour laquelle nous avons pu accomplir autant de travail et de contributions comme l'ont souligné les rapporteurs. Merci également pour tous les bons moments que nous avons partagés en dehors du travail, en particulier lors de nos réunions PISCO avec nos partenaires nantais (LS2N) et niçois (INRIA).

J'aimerais ensuite remercier les membres du LIRIS pour leur accueil et l'ambiance très agréable au laboratoire. Merci Isabelle et Catherine de l'équipe administrative pour votre travail et nos discussions. Merci à tous mes collègues, en particulier à mes super co-bureaux Jocelyn et Seddik. J'étais chanceuse de partager le même bureau avec vous pendant ces années. Merci pour tous nos rires, nos discussions, nos moments inoubliables, nos partages de nourriture au bureau (à commencer par les fameuses tartes aux pralines, jusqu'aux navets). Merci Jocelyn pour ton calme, ton optimisme ("ça pourrait être pire"), et toutes tes conseils sur Git ; tu étais toujours prêt à m'aider. Merci Seddik pour ta joie de vivre, pour les moments (plutôt les journées) passés à nous raconter n'importe quoi, pour ta culture infinie, et pour nous avoir toujours tenus informés de ce qui se passait dans le monde. Je remercie également Sandra pour nos discussions pendant les pauses déjeuner et notre partage des détresses et des enthousiasmes. Merci Méghane et Juba pour toutes nos soirées.

Merci à tous mes amis au Liban et en France pour le temps que nous avons passé ensemble, pour toutes nos discussions, nos voyages, nos balades, nos soirées, et tous les moments inoubliables qui ont rendu ces 3 années de thèse moins difficiles et stressantes. Un merci tout particulier à Noor, Judy, Fatima, Mirella et Sami. Mille merci Simona d'avoir fait partie de ce périple dans ses moindres détails, de m'avoir épaulé et soutenu sans faille. J'espère avoir pu, à mon tour, te soutenir dans ta thèse.

Enfin, rien ne pourra exprimer ma gratitude envers ma famille surtout mes parents Hanna et Gracia, ma sœur Léa et mon frère Massaad. Vous êtes et serez toujours présent dans mon esprit et mon âme. Je vous remercie pour la liberté et la confiance que vous m'avez accordées, vos prières et votre soutien infinie. Je suis vraiment reconnaissante de vous avoir. I LOVE YOU BEYOND WORDS. Je vous dédie mon travail.

Table of Contents

Abstract	i
Résumé	ii
Remerciements	v
Table of Contents	x
Lists of Figures	xvi
Lists of Tables	xviii
Introduction	1
1 Related Work	5
1.1 Subjective quality assessment	6
1.1.1 Experimental methodologies for subjective quality assessment of im- ages and videos	6
1.1.2 Subjective quality assessment of 3D graphics and resulting data sets .	6
1.1.3 Comparison of subjective methodologies	9
1.1.4 Subjective quality assessment in crowdsourcing	11
1.2 Objective quality assessment	12
1.2.1 Model-based quality metrics	13
1.2.2 Image-based quality metrics	15
1.2.3 Comparison of the two approaches	16
1.3 Conclusion	17
I Subjective Quality Assessment	18
2 Comparison of Subjective Methods for Quality Assessment of 3D Graphics in Virtual Reality	19
2.1 Experimental methodologies	20
2.2 Dataset generation	23
2.2.1 3D source model selection	23
2.2.2 Distortions	23
2.3 Virtual Environment and apparatus	26
2.3.1 Rendering	26
	vii

2.3.2	Rating interface	27
2.4	Experimental procedure	29
2.4.1	Subjective experiment 1	29
2.4.2	Subjective experiment 2	29
2.5	Participants and training	30
2.5.1	Training	30
2.5.2	Duration	30
2.5.3	Participants	31
2.6	Results of Experiment 1 and influence of explicit reference	31
2.6.1	Observers screening and data processing	31
2.6.2	Resulting MOSs and DMOSs	33
2.6.3	Consistency across subject groups	35
2.6.4	Accuracy of quality scores	36
2.6.5	Confidence intervals	39
2.7	Results of Experiment 2 and methods comparison	41
2.7.1	Observers screening and data processing	41
2.7.2	Resulting MOSs	41
2.7.3	Accuracy and time-effort	43
2.7.4	Confidence intervals	44
2.8	Discussion and recommendations	46
2.9	Conclusion	47
3	Subjective Quality Assessment of 3D Meshes with Vertex Colors in Virtual Reality	48
3.1	Dataset generation	49
3.2	Experimental environment and apparatus	51
3.3	Participants and training	52
3.3.1	Training	52
3.3.2	Creation of test sessions	52
3.3.3	Participants	53
3.3.4	Duration	53
3.4	Results and analyzes	53
3.4.1	Observers screening and data processing	53
3.4.2	Observers' agreement	54
3.4.3	Factors influencing subjective opinions	56
3.5	Discussion and recommendations	64
3.6	Conclusion	65
4	Exploring Crowdsourcing for Subjective Quality Assessment of 3D Graphics	67
4.1	Dataset	68
4.2	Lab experiment	68
4.3	Crowdsourcing experiment	68
4.3.1	Experimental environment	69
4.3.2	Participants and test sessions	70
4.4	Results and comparison of experiments	71
4.4.1	Participants screening	71
4.4.2	Resulting MOSs	72
4.4.3	Confidence intervals	74

4.4.4	Accuracy of quality scores	74
4.4.5	Participants' agreement and consistency	76
4.4.6	Content ambiguity	77
4.5	Discussion	78
4.6	Conclusion	80
5	Subjective Quality Assessment of a Large-Scale Textured 3D Mesh Dataset in Crowd-sourcing	81
5.1	Dataset generation	82
5.1.1	3D source model selection	82
5.1.2	Content characterization	84
5.1.3	Distortions	88
5.2	Subjective experiment	93
5.2.1	Test stimuli selection	93
5.2.2	Rendering	94
5.2.3	Experimental environment	95
5.2.4	Creation of test sessions	95
5.2.5	Participants and training	96
5.3	Results and analyzes	97
5.3.1	Participants screening	98
5.3.2	Resulting MOSs and annotating the whole dataset	98
5.3.3	Influence of each distortion on perceived quality	99
5.3.4	Influence of distortion interactions on perceived quality	101
5.3.5	Influence of content characteristics on perceived quality	107
5.4	Application: Rate-Distortion control	115
5.5	Discussion	116
5.6	Conclusion	117
II	Objective Quality Assessment	118
6	A Model-Based Perceptual Quality Metric for 3D Meshes with Vertex Colors	119
6.1	Toward a quality assessment metric for colored 3D meshes	120
6.1.1	Correspondence between meshes	120
6.1.2	Neighborhood Computation	121
6.1.3	Perceptually relevant features	121
6.1.4	Global perceptual quality score	123
6.2	Results and evaluation	124
6.2.1	Dataset	124
6.2.2	Performance evaluation measures	125
6.2.3	Single feature prediction performance	126
6.2.4	Toward an Optimal Combination of features	127
6.2.5	Performance evaluation and comparisons	127
6.2.6	Recommended weights	130
6.2.7	Validation on a dataset of textured 3D meshes	132
6.3	Integration of visual attention complexity	134
6.3.1	The visual attention complexity measure	135

6.3.2	Toward an optimal combination of features	136
6.3.3	Performance evaluation and comparisons	136
6.3.4	Recommended weights	137
6.3.5	Performance evaluation per quality range	138
6.4	Integration of 3D model viewpoints	139
6.5	Conclusion	141
7	An Image-Based Perceptual Quality Metric for 3D Graphics Based on CNN	143
7.1	LPIPS Overview	144
7.2	Toward a CNN-based quality metric for 3D graphics	146
7.2.1	Performance of LPIPS on our dataset	146
7.2.2	Our approach	146
7.2.3	Training on the textured 3D mesh dataset	147
7.3	Results and Evaluation	148
7.3.1	Performance evaluation on the test set of textured meshes	148
7.3.2	Performance evaluation on a dataset of colored 3D meshes	149
7.4	View-independent approach	150
7.5	Conclusion	151
 Conclusion		 152
 Publications		 156
 References		 157
 Appendix A		 171
A.1	Chapter 2	171
A.2	Chapter 3	179
A.3	Chapter 4	181
A.4	Chapter 5	182
A.5	Chapter 6	186

List of Figures

1	Top: The 3D scan come to the rescue of Notre-Dame cathedral. Below: The evolution of the Tomb Raider video game and a comparison with a scene from the latest movie (released in 2018).	1
2	The 3D textured mesh and the point cloud of a the same 3D graphic model. The Figure is reprinted from [1].	3
2.1	Illustration and timeline of the three subjective quality assessment methodologies explored in this study.	21
2.2	Illustration of the 3D graphic source models.	24
2.3	Some examples of distorted models. Acronyms refer to Distortion Type_Strength.	25
2.4	The experimental environments of the three methodologies implemented.	28
2.5	Comparison of the G1 and G2 mean scores of the ACR-HR and DSIS tests, for all the stimuli. G1 subjects did the ACR-HR session 1 st followed by the DSIS session, while G2 subjects did the DSIS session 1 st and then the ACR-HR session. For a given distortion strength, the dots are horizontally spaced apart to avoid overlapping.	34
2.6	Boxplots of MOSs obtained by the two groups of subjects involved in the ACR-HR and DSIS tests.	35
2.7	p-values computed between the rating scores of the two subject groups computed for all stimuli and for both methodologies. Red color indicates a significant difference between the scores of G1 and G2.	36
2.8	Variation of the accuracy according to the number of subjects for both methodologies and both groups (G1 subjects did the ACR-HR session 1 st followed by the DSIS session, while G2 subjects did the DSIS session 1 st and then the ACR-HR session). The accuracy (y-axis) is defined as the percentage of pairs of stimuli whose qualities were assessed as statistically different. Curves represent mean values of these percentages and areas around curves represent 2.5th - 97.5th percentiles.	37
2.9	Inter-rater reliability, of each group, in the DSIS and ACR-HR tests.	38
2.10	Intra-rater reliability among the ACR-HR and DSIS tests for the 2 groups.	38
2.11	Number of observers required in an ACR-HR test to achieve the same accuracy as a DSIS test.	39
2.13	Evolution of confidence interval (CI) widths for the ACR-HR and DSIS methodologies as a function of the number of observers involved in G1 and G2. G1 subjects did the ACR-HR session 1 st followed by the DSIS session, while G2 subjects did the DSIS session 1 st and then the ACR-HR session.	40

2.14	Boxplots of MOSs obtained for the tested methods.	42
2.15	Comparison of the mean scores of the ACR-HR, DSIS, and SAMVIQ tests, for all the stimuli. For a given distortion strength, the dots are horizontally spaced apart to avoid overlapping.	42
2.16	Variation of the accuracy according to the number of subjects (a) and time-effort (b) for the tested methodologies. The accuracy (y-axis) is defined as the percentage of pairs of stimuli whose qualities were assessed as statistically different. Curves represent mean values of these percentages and areas around curves represent 2.5th - 97.5th percentiles.	43
2.17	(a) Boxplots of CIs obtained for the tested methods, (b) Comparison of normalized CIs of the tested methods as a function of normalized (D)MOSs.	45
2.18	Evolution of CI widths as a function of the number of observers for the tested methodologies	46
2.19	Accuracy of the DSIS method according to the number of subjects.	47
3.1	Illustration of the 3D graphic source models and their selected viewpoints, respectively. Acronyms refer to Model_Viewpoint. The “ <i>Dancing Drummer</i> ” model is used only in training.	50
3.2	Some examples of distorted models displayed in the selected viewpoints. Acronyms refer to Model_Dist-Type_Dist-Strength_Viewpoint.	51
3.3	Bias b_s of each subject involved in our subjective experiment, and its confidence interval.	55
3.4	Inconsistency v_s of each subject involved in our subjective experiment, and its confidence interval.	55
3.5	Boxplot of subjects’ inconsistency in relation to their familiarity with VR.	56
3.6	The mean opinion scores of all the stimuli, associated with their confidence intervals. For a given distortion strength, the dots are horizontally spaced apart to avoid overlapping.	57
3.7	The variation of the MOSs of different (a) source models and (b) distortions depending on the viewpoints.	58
3.8	Boxplots of MOSs obtained for the combination of the viewpoint with different factors. Mean values are displayed as circles.	59
3.9	Visual example of the interaction between between the viewpoints of the <i>Fish</i> and different distortion types (Animation: R).	60
3.10	The visual quality of <i>Ari</i> highly distorted under its different viewpoints (Animation: R).	61
3.11	Boxplots of MOSs obtained for the combination of the animation with different factors. Mean values are displayed as circles.	62
3.12	Boxplots of CIs obtained for different factors or combination of factors. Mean values are displayed as circles. In (a), <i>Cham.</i> and <i>Sam.</i> refer to Chameleon and Samurai, respectively.	62
3.13	The variation of the CIs length of the <i>Samurai</i> depending on the viewpoint.	63
3.14	The mean opinion scores associated with the CIs for a subset of models displayed in <i>viewpoint 1</i> and animated with a rotation and a zoom.	64
3.15	Content ambiguity a_c of each source model associated with the different HRTs, and its confidence interval.	65

4.1	The graphical interface based on the DSIS method of our CS experiment. . .	69
4.2	Comparison of the mean opinion scores of the lab and CS experiments, for all the stimuli. Results are grouped by source models and distortion types. .	73
4.3	Score distributions for the lab and CS experiments.	73
4.4	Boxplots of MOSs obtained for the lab and CS experiments.	73
4.5	Variation of the width of Confidence Intervals (CIs) according to the number of ratings per stimulus in the CS and lab experiments. ns refers to not statistically significant ($p\text{-value} \geq 0.05$).	75
4.6	Confidence Intervals (CIs) as a function of MOSs obtained in the CS and lab experiments. The evolution of CIs is assessed according to the number of ratings per stimulus.	75
4.7	Variation of the accuracy according to the number of ratings per stimulus for the CS and lab experiments. The accuracy (y-axis) is defined as the percentage of pairs of stimuli whose qualities were assessed as statistically different. Curves represent mean values of these percentages and areas around curves represent 2.5th - 97.5th percentiles.	76
4.8	Boxplots of participants' inconsistency obtained in the CS and lab experiments.	77
4.9	Content ambiguity of each source model associated its CI (calculated as described in [2]), for the CS and lab experiments.	78
4.10	Venn diagram summarizing unreliable participants detected by the golden units inspection and ITU-R BT.500-13 screening procedure and those with high inconsistency and bias computed by the MLE model (v_s and b_s of Eq. 3.3). s^{GU} denotes the score assigned to the golden unit GU	79
4.11	Results of the questionnaire asked at the end of the CS experiment.	80
5.1	The 3D graphic source models constituting our database.	83
5.2	Characterization of the geometry, color, and semantic of textured 3D models.	86
5.3	(a) Geometry and color spatial information and (b) the visual attention complexity for our source models.	87
5.4	Some examples of distorted models. Acronyms refer to $LoD_{simpL} qp qt T_S T_Q$.	92
5.5	(a) Selection of the test stimuli by uniformly sampling the plane formed by 2 pseudo-MOSs. The black dots refer to the pseudo-MOS values of all stimuli in the dataset, while the blue dots refer to those selected for the subjective study. (b) The pseudo-MOS _{HDRVDP} distribution of the 3000 test stimuli.	94
5.6	The graphical interface of our subjective experiment.	96
5.7	Distribution of raw scores collected during the subjective experiment.	99
5.8	MOS ditribution of the 3000 test stimuli.	99
5.9	Pseudo-MOS distribution of the 343750 stimuli in the dataset.	99
5.10	Boxplots of MOSs obtained for the quantization of the (a) vertices' positions qp and (b) texture coordinates qt . Mean values are displayed as circles.	100
5.11	(a) Boxplots of MOSs obtained for the LoD simplification LoD_{simpL} . (b) Boxplots of MOSs obtained for the LoD simplification, but restricted to the least quantized models ($qp = 11$ & $qt = 10$). Mean values are displayed as circles.	100

5.12	Boxplots of MOSs obtained for the texture (a) compression T_Q and (b) sub-sampling T_S . Mean values are displayed as circles.	101
5.13	Boxplots of MOSs illustrating the interaction between the LoD simplification LoD_{simpL} and the quantization of the model's positions qp	103
5.14	Visual example illustrating the interaction between the LoD simplification and the position quantization regarding the perceived quality. The 2 nd row shows a rendering of the stimuli without the texture as well as a zoom on the chair back showing the topology of the mesh after distortion. Acronyms refer to the following combination of distortion parameters: $LoD_{simpL} qp qt T_S T_Q$.	103
5.15	Boxplots of MOSs illustrating the interaction between the geometry qp and texture coordinate qt quantization.	104
5.16	Visual example illustrating the interaction between the geometry and texture coordinate quantization regarding the perceived quality. Acronyms refer to the following combination of distortion parameters: $LoD_{simpL} qp qt T_S T_Q$	104
5.17	Boxplots of MOSs illustrating the interaction between the texture compression T_Q and sub-sampling T_S	105
5.18	Boxplots of MOSs illustrating the interaction between the texture coordinate quantization levels qt and the texture sub-sampling T_S	106
5.19	Visual example illustrating the interaction between the sub-sampling of the texture and its coordinates quantization regarding the perceived quality. Acronyms refer to the following combination of distortion parameters: $LoD_{simpL} qp qt T_S T_Q$.	106
5.20	MOSs of different models with different geometric SI_{Geo} and color SI_{Col} characteristics and having undergone the same distortions ($LoD_{simpL} qp qt T_S T_Q$).	107
5.21	Clusters of Source models grouped by geometric and color characteristics. . .	107
5.22	Boxplots of the MOSs illustrating the interaction between the (a) geometric SI_{Geo} and (b) color characteristics SI_{Col} of the models and the quantization of the position of their vertices qp	108
5.23	Visual examples illustrating the impact of the geometric complexity of the model on the perception of degradations generated by the geometry quantization. Acronyms refer to the following combination of distortion parameters: $LoD_{simpL} qp qt T_S T_Q$	109
5.24	Visual examples illustrating the impact of the color complexity of the model on the perception of degradations generated by the geometry quantization. Acronyms refer to the following combination of distortion parameters: $LoD_{simpL} qp qt T_S T_Q$.	109
5.25	Boxplots of the MOSs illustrating the interaction between the quantization of the texture coordinates qt and (a) the model color SI_{Col} and (b) geometric SI_{Geo} characteristics.	111
5.26	Visual examples illustrating the impact of the color characteristics of the model on the perception of degradations generated by the quantization of the texture coordinates. Acronyms refer to the following combination of distortion parameters: $LoD_{simpL} qp qt T_S T_Q$	111
5.27	Visual examples illustrating the impact of UV map quantization on the perceived quality of textured 3D meshes. Models are presented with their texture seams highlighted and their texture map. Acronyms refer to the following combination of distortion parameters: $LoD_{simpL} qp qt T_S T_Q$	112

5.28	Boxplots of the MOSs illustrating the interaction between the quantization of the texture coordinates qt and the amount of vertices present on texture seams V_{Tseams}	113
5.29	Visual examples illustrating the impact of the texture seams and the perception of degradations generated by the quantization of the texture coordinates. Models are presented with their texture map (left) and their UV map (right). Vertices present on texture seams are highlighted in red. Acronyms refer to the following combination of distortion parameters: $LoD_{simpL} qp qt Ts TQ$	114
6.1	Overview of the proposed metric $CMDM$	121
6.2	Performance comparison of several metrics in two cross-validation tests. Mean performance evaluation measures are reported. Error bars indicate the standard deviation over the test sets.	128
6.3	Scatter plots of subjective scores versus objective metric values for the dataset of meshes with vertex colors. Each point represents one stimulus. The fitted logistic function is displayed in black.	131
6.4	Scatter plots of subjective scores versus objective metric values for the LIRIS Textured Mesh Database [3]. Each point represents one stimulus. The fitted logistic function is displayed in black. Plots (e), (f), (g) and (h) are reprinted from [3].	133
6.5	Overview of the $CMDM-VAC$ metric.	134
6.6	Saliency data and corresponding VAC scores. In (a) and (b), the rendered viewpoint of a 3D object is on the left, and its corresponding masked saliency information is on the right.	135
6.7	Performance evaluation of $CMDM-VAC$ in two cross-validation tests. The average correlations over the test sets are reported. Error bars indicate the standard deviation over the test sets.	137
6.8	Scatter plots of subjective scores versus the $CMDM-VAC$ values for the dataset of meshes with vertex colors. Each point represents one stimulus. The fitted logistic function is displayed in black.	138
6.9	Performance evaluation of several metrics according to the quality range of stimuli.	139
7.1	To compute a distance d_0 between two patches x and x_0 , given a network F : we first compute deep embeddings, normalize the activations in the channel dimension, scale each channel by vector ω , and take the ℓ_2 distance. We then average across spatial dimension and sum across all layers. This figure is reprinted from [4].	144
7.2	Training on 2AFC: a small network G is trained to predict perceptual judgment h from distance pair (d_0, d_1) . This figure is reprinted from [4].	146
7.3	Training on quality assessment task: the overall quality \hat{Q} of an image is estimated by averaging the qualities d_i of its local patches.	147
A.1	Different parts of a boxplot.	171
A.2	Snapshots of the stimuli from the dataset used to compare the subjective methodologies. Acronyms refer to Distortion Type_Strength.	176

A.3	MOSs and confidence intervals obtained in the DSIS tests for the two groups of participants.	177
A.4	DMOSs and confidence intervals obtained in the ACR-HR tests for the two groups of participants.	178
A.5	DMOSs and confidence intervals obtained in the SAMVIQ tests.	179
A.6	MOSs and CIs of the 480 stimuli from the dataset of meshes with vertex colors. For a given distortion strength, the dots are horizontally spaced apart to avoid overlapping.	180
A.7	Mean confidence intervals of each source model for the CS and lab experiments.	181
A.8	Optimal distortion parameters for each source model, providing the best possible visual quality for a given size requirement. Models are sorted by geometric complexity SI_{Geo}	183
A.9	Optimal distortion parameters for each source model, providing the best possible visual quality for a given size requirement. Models are sorted by color complexity SI_{Col}	184
A.10	Optimal distortion parameters for each source model, providing the best possible visual quality for a given size requirement. Models are sorted by the amount of vertices present on texture seams V_{Tseams}	185
A.11	Camera positions regularly sampled around a 3D object.	186

List of Tables

1.1	Public quality assessment datasets for 3D meshes and point clouds.	10
2.1	Characteristics of the 3D graphic source models	24
2.2	Details on the distortions applied to each source model.	26
2.3	Experimental details of the tested methodologies.	32
2.4	(D)MOS Correlation matrices for ACR-HR, DSIS, and SAMVIQ.	42
3.1	Agreement and inconsistency of subjects familiar with VR and those with no VR experience.	56
5.1	List of source models, their number of vertices and semantic category.	84
6.1	Performance of individual features.	126
6.2	Performance of geometry-based features after removal of the color quantization distortion.	126
6.3	Performance comparison of several metrics in a cross-validation test among source models	129
6.4	Performance comparison of several metrics in a cross-validation test among distortion types	129
6.5	Weights and importance of the optimal subset of features for <i>CMDM</i>	130
6.6	Performance comparison of different metrics on the dataset of 80 meshes with vertex colors.	130
6.7	Performance comparison of different metrics on the LIRIS Textured dataset [3]. For metrics marked with a *, the values are reprinted from [3].	132
6.8	Performance comparison of <i>CMDM-VAC</i> on two test sets. For metrics marked with a *, the values are reported from section 6.2.5.	137
6.9	Weights and importance of the optimal subset of features for <i>CMDM-VAC</i>	138
6.10	Performance comparison of different metrics in to 2 scenarios.	140
6.11	Performance evolution of different metrics before and after integrating the viewpoint	140
6.12	Performance comparison of different metrics on the pairs of stimuli significantly affected by the viewpoints.	141
7.1	Performance comparison of different metrics on the test set of our textured 3D mesh dataset, described in Chapter 5.	149
7.2	Performance comparison of different metrics on the dataset of meshes with vertex colors, described in Chapter 3, for scenario (1). For metrics marked with a *, the values are reprinted from Table 6.6 in Chapter 6.	150

7.3	Performance comparison of different metrics on the dataset of meshes with vertex colors described in Chapter 3, for different viewpoints (scenario 2). For metrics marked with a *, the values are reprinted from Table 6.10b in Chapter 6.	150
7.4	Performance comparison of different metrics on the test set of our textured 3D mesh dataset, when several viewpoints are considered per stimulus. . . .	151

Introduction

The Cathedral of Notre-Dame de Paris was engulfed in flames 2 years ago (april 2019). Fortunately, thanks to a meticulously precise 3D scan of the building (accurate within 5 millimeters)¹, captured in 2015, the cathedral will be restored to its former glory. Recent animation movies and video games feature huge (several square kilometers), highly detailed 3D scenes that mimic the real world. These examples, illustrated in Figure 1, are among many that highlight both the interest and common use of three-dimensional (3D) graphical data in many fields of industry including digital entertainment, computational engineering, cultural heritage, automotive and healthcare.

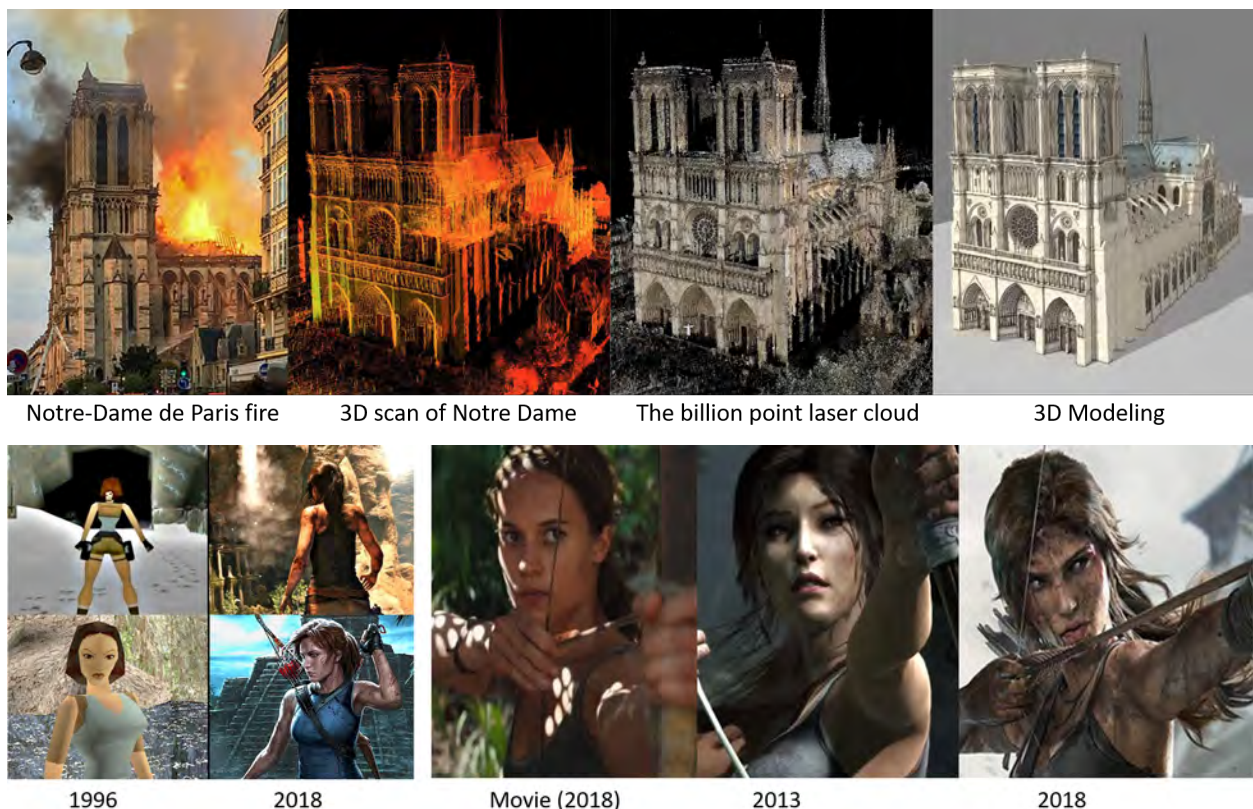


Figure 1: Top: The 3D scan come to the rescue of Notre-Dame cathedral. Below: The evolution of the Tomb Raider video game and a comparison with a scene from the latest movie (released in 2018).

¹<https://www.nationalgeographic.com/adventure/article/150622-andrew-tallon-notre-dame-cathedral-laser-scan-art-history-medieval-gothic>

The use of 3D data is also growing for the general public with the proliferation of acquisition technologies (3D scanners, 360° cameras, MRI, etc.), intuitive 3D modeling tools, 3D printers, and affordable virtual and mixed reality devices (Oculus Rift, HTC Vive, Microsoft HoloLens, etc.). All of these technologies make the size and complexity of 3D data explode. As in the examples above, the resulting 3D scenes are huge and extremely detailed: they contain several million geometric primitives, associated with a wide range of appearance attributes, intended to reproduce a realistic material appearance, as well as animation data.

The way of consuming and visualizing 3D content is also evolving from standard screens to extended reality (i.e. Augmented, Mixed and Virtual Reality AR/VR/MR), which is considered as the potential next computing platform with an estimate market of \$80B by 2025². The great advantage of extended reality technologies and Head-Mounted Displays (HMD) is that they provide 6 Degrees of Freedom (6DoF) allowing realistic human interactions and a high level of immersion. However, the visualization and interaction with 6DoF of large and complex 3D scenes remains an unsolved issue to date due to two major challenges: (1) the potential complexity of a 3D scene that can be displayed on a VR/MR HMD is substantially smaller than that on a standard screen, because the GPU must generate 4 times more images (to ensure two images per frame and a sufficient frame-rate to prevent motion sickness); (2) the strong latency problems that may occur while streaming the 3D scene on the display device. This problem is growing as more and more online VR/MR applications (VR video games, virtual museums, virtual classes, telepresence, etc.) consider 3D data stored on remote servers.

Thus, creating the illusion of immersion for a HMD results in very heavy rendering workloads: low latency, high frame-rate, and high visual quality are all needed.

To adapt the complexity of the 3D content for HMDs (notably for lightweight devices like Google Cardboard and Samsung Gear VR) and to avoid latency due to transmission, diverse processing operations, including simplification and compression, are inevitable. These operations reduce the amount of data (reduce the Level of Details (LoD) and the size of 3D data, respectively) and by extension the costs in processing, storage, and transmission. However, such operations are lossy and result in visual degradations that affect the perceptual quality of the 3D scene and, in turn, the user's Quality Of Experience (QoE). It is therefore essential to define measures to accurately assess the impact of these distortions in order to find the right compromise between visual quality and data size/LoD. For this purpose, quality assessment methodologies are required.

Quality assessment methods can be classified as subjective, or objective. Both are essential for assessing the perceptual quality of degraded models. Objective quality metrics rely on algorithms that attempt to predict the quality of degraded content as perceived by the human. Predicting visual quality is a very complex task due to the complexity of the Human Visual System (HVS). Subjective quality evaluation is based on subjective experiments in which a group of human subjects are asked to assess/rate the visual quality of test data. These experiments provide ground-truth datasets of crucial importance for understanding human psychological behavior and for training and benchmarking objective quality metrics.

²<https://www.goldmansachs.com/insights/pages/technology-driving-innovation-folder/virtual-and-augmented-reality/report.pdf>

In this thesis, we address open questions arising in the field of 3D graphics quality assessment. Polygonal meshes and point clouds are popular methods to represent 3D graphics. Figure 2 shows an example of these 3D representations. A point cloud is defined as a set of points that extends in the 3D space, determined by their x , y and z coordinates. Each point can be associated with a color attribute. The main advantages of point clouds lie in their simple data representation and fast acquisition (often obtained from 3D laser scanners and LiDAR technologies).

In contrast, a polygonal mesh explicitly represents a 3D surface. A 3D mesh is expressed by vertices (points), edges (connectors between vertices) and polygons (faces formed by edges and vertices). It is characterized by its connectivity which describes the relationship between the vertices. Connectivity (neighborhood information) is not available for point clouds as each point is simply expressed by its coordinates. The color attributes used for 3D meshes can be in the form of vertex colors (a color is assigned to each vertex) or texture maps (a 2D image is mapped onto the 3D surface). Although 3D mesh representations require more memory and storage space than point cloud representations, they have other benefits. They are better suited for geometry processing operations, since they define an explicit surface. They offer better visual quality and easier integration into computer graphics pipelines.

As a result, these 2 representations differ in several respects: mainly in the way they are rendered, and the nature of the processing operations and distortions they undergo.

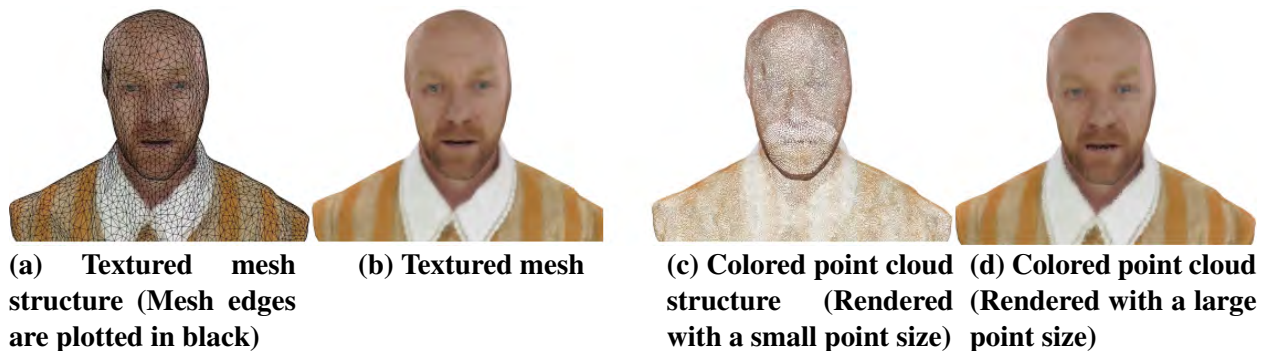


Figure 2: The 3D textured mesh and the point cloud of a the same 3D graphic model. The Figure is reprinted from [1].

In this thesis, we focus on 3D meshes with color attributes, either in the form of vertex colors or texture maps. We are interested in both subjective and objective quality evaluation. The main distinctive property of the thesis lies in the fact that we will consider immersive environments (VR), and rich 3D data associated with color attributes. Our contributions can be summarized in the following points:

- We determine the best subjective experimental methodology for evaluating the perceived quality of 3D graphics, especially in VR. We provide recommendations regarding the minimum number of participants required for such experiments.
- We provide a quality assessment dataset of meshes with vertex colors, which contains 480 stimuli generated from 5 source models each associated with 3 viewpoints and 2 short animations. It is the first public dataset produced in VR for such data.

- We provide another quality assessment dataset of textured 3D meshes. The dataset includes 55 source models corrupted by combinations of 5 types of distortions applied on the geometry and texture of the meshes. In total, more than 343k distorted stimuli were generated, of which 3000 are associated with subjective scores derived from a subjective experiment conducted in crowdsourcing and the rest with predicted subjective scores.
- Leveraging our 2 established datasets, we present in-depth analyzes on the influence of source model characteristics, distortion interactions, viewpoints and animations on the perceived quality of 3D graphics.
- We investigate the reliability of crowdsourcing experiments to assess the quality of 3D graphics and whether they can achieve the accuracy of laboratory tests.
- We propose the first model-based metric for quality assessment of 3D meshes with vertex colors. This metric incorporates geometry-based and color-based features. The metric demonstrates good results and stability on two different datasets. An adapted version of this metric was also proposed for colored point clouds.
- We propose an image-based quality metric for 3D graphics that employs convolutional neural networks. This metric shows state-of-the-art results on our two datasets.
- We investigate how visual attention can improve the performance of perceptual quality metrics.
- The datasets³ and source codes of the metrics^{4,5} are made publicly available online.

Our contributions are presented in detail in the rest of the manuscript: Chapter 1 reviews previous work on subjective and objective quality assessment of 3D graphics. Part I presents our contributions in the field of subjective evaluation, while Part II describes the proposed objective quality metrics. The general conclusion and perspectives are outlined at the end.

³<https://yananehme.github.io/datasets/>

⁴<https://github.com/MEPP-team/MEPP2>

⁵<https://github.com/MEPP-team/PCQM>

Chapter 1

Related Work

Multimedia content (text, image, audio, video, 3D graphic, etc.) are popular data used in our daily lives. For each type of these data, there is a wide variance in the storage size, resolution, complexity, data structure and compression techniques used to store and manipulate them. All of these factors affect the perceived quality of the data and therefore the user's Quality Of Experience (QoE). Thus, it is crucial to evaluate and understand the visual quality, as perceived by human observers.

The perceptual quality can be assessed using subjective studies and objective metrics. Objective metrics consist in algorithms designed to automatically predict the visual quality loss (i.e. the level of annoyance of visual artifacts). On the other hand, subjective studies, aka. user studies, involve inviting a group of human subjects to assess the visual quality of test data. These subjective experiments provide the most reliable way to create ground-truth datasets useful for understanding human psychological behavior (perception of multimedia content) as well as for benchmarking and tuning objective quality metrics.

In this thesis, we are interested in 3D data which are used in many domains (see the general introduction) and therefore subject to several processing operations such as compression (to reduce the size of large 3D data), simplification (to reduce the level of details of the 3D content), watermarking (to protect the intellectual property), noise addition during transmission processes [5–8]. All these operations are lossy and introduce artifacts/distortions which often alter the visual quality of the rendered data and thus the QoE. In this chapter, we review previous work on subjective and objective quality assessment of graphical 3D content. We provide an overview of existing datasets and metrics for predicting the visual impact of distortions applied on such data. Note that, 3D data can be represented in different ways (3D meshes, point clouds, voxels), with and without appearance attributes (color, texture, material, etc.). In the context of this thesis, we are specifically interested in **3D meshes with colors attributes**, either in the form of texture maps or vertex colors.

The rest of this chapter is organized as follows: in the first part (section 1.1), the subjective evaluation studies are detailed and the resulting datasets are presented, whilst in a second part (section 1.2), the notable objective quality metrics for 3D meshes are described. Finally, our contributions are outlined in section 1.3.

1.1 Subjective quality assessment

In this section, we first review popular methodologies for subjective quality assessment of natural images and videos, and then focus on existing subjective tests conducted with 3D graphics. We also provide an overview of existing datasets for 3D graphics. Finally, we discuss relevant works that compare subjective methodologies and experimental setups.

1.1.1 Experimental methodologies for subjective quality assessment of images and videos

Several methodologies for evaluating the subjective quality of 2D images/videos exist in the literature and have been standardized by the International Telecommunication Union (ITU) [9–11]. Among these methodologies, four are notably used nowadays: the Absolute Category Rating (ACR) method, the Double Stimulus Impairment Scale (DSIS) method, the Subjective Assessment Methodology for Video Quality (SAMVIQ) and the pairwise comparison (PC) method. The ACR method consists of presenting each impaired sequence individually to the observers and then asking them to rate its quality on a five-level quality scale. In the DSIS method, the reference content is presented first, followed by the same content impaired. The observer is asked to rate, on a five-level impairment scale, the degradation of the second sequence compared to its reference. These methods are classified as categorical rating methods since they use a discrete scale [12]. They are dominant in video quality assessment tests [10, 11]. Furthermore, the ACR with Hidden Reference method - which consists of presenting the impaired stimuli as well as the original unimpaired stimuli (references) to the subjects without informing them of the presence of the latter - is notably used by the Video Quality Experts Group (VQEG) [13]. The Pairwise Comparison (PC) method is an alternative method in which two distorted images/videos are displayed, side by side, and the observer has to choose the one having the highest quality. The fourth method is SAMVIQ. It differs from the others in several aspects. SAMVIQ uses a continuous quality scale ranged from 0 to 100. In addition, it is based on a multi-stimuli with random access approach [9, 14]: test sequences are presented one at a time but the observer is able to review each sequence and modify the quality score multiple times.

1.1.2 Subjective quality assessment of 3D graphics and resulting data sets

Subjective quality tests involving 3D models were initially introduced on meshes, more precisely on geometry-only models, to assess the artifacts induced by the simplification [15–18], smoothing [19], noise addition [19, 20], watermarking [21, 22] and vertices position quantization [23, 24]. Váša et al. [25] and Torkhani et al. [26] carried out subjective experiments to assess the perceptual quality of 3D dynamic meshes. Little work considered meshes with color attributes. Pan et al. [17] conducted a subjective experiment on textured meshes to assess the perceptual interactions between the geometry and color information. However, they considered only geometry and texture sub-sampling distortions. In 2016, Guo et al. [3] carried out a subjective experiment to assess the influence of light-

ing, shape and texture content on the perception of artifacts. They also provided a dataset of textured 3D meshes evaluated using two rendering protocols. Gutiérrez et al. [27] used this dataset to evaluate the perception of geometry and texture distortions in Mixed Reality (MR) scenarios. They also analyzed the impact of environment lighting conditions on the perceived quality of 3D objects in MR. Zerman et al. [1] conducted a subjective study to compare textured meshes and colored point clouds for a Volumetric Video (VV) compression scenario utilizing the appropriate state-of-the-art compression techniques for each 3D representation. For this purpose, they built a database and collected user opinion scores for subjective quality assessment of the compressed VV. The datasets provided by [3] and [1] are further detailed in Table 1.1.

The first subjective evaluation study for point clouds reported in the literature was conducted in 2014 [28] in an effort to assess the visual quality at different geometric resolutions, and different levels of noise introduced to both geometry and color. Since then, many experiments have been conducted on colored and colorless point clouds and have addressed different issues. Alexiou et al. [29,30] assessed the quality of geometry-only point cloud models corrupted by parametrized distortions (Gaussian noise and Octree-pruning) that simulate position errors from sensor inaccuracies and compression artifacts. In the same vein, Su et al. [31] evaluated the impact of similar distortions on the quality of a large set of colored point clouds. Torlig et al. [32] evaluated the visual quality of voxelized colored point clouds in subjective experiments that were performed in two intercontinental laboratories. A large-scale subjective quality evaluation experiment was conducted in order to assess the influence of MPEG codecs on point clouds [33]. Da Silva Cruz et al. [34] reported the first study aiming at defining test conditions for both small- and large-scale point clouds. Javaheri et al. [35] studied the impact of different rendering options on perceived quality of static point clouds subject to geometry-only compression artifacts. Very recently, Liu et al. [36] established the largest point cloud quality assessment dataset to date (2021) and conducted a fairly large subjective experiment campaign to collect subjective ratings. This dataset is described below in Table 1.1.

Regarding graphics applications requiring localized information on the distortion visibility, the local marking of visible distortions is commonly used [37–39]. In such subjective experiments, observers manually mark the visible local artifacts in impaired images.

Experimental environment and apparatus

No specific standards or recommendations exist for subjective experiments conducted on 3D content. Researchers cited in the above works have adapted existing image/video methodologies, while considering different ways to display the 3D models to observers (e.g., 2D still images, animated videos, interactive scenes) and different test equipment (e.g., 2D screen, Augmented Reality AR and Virtual Reality VR headsets).

Lavoué et al. [19], Corsini et al. [22] and Torkhani et al. [26] considered single stimulus protocols, derived from the ACR method, to assess the quality of impaired 3D meshes. The observers were able to freely interact with the 3D models and then had to rate the

visibility of the distortions between 0 (invisible) and 10. Zerman et al. [1] considered the ACR with Hidden Reference (ACR-HR) methodology to assess the perceptual differences between two VV representation formats. A passive approach was chosen for viewer interaction with the VV in order to reduce inter-subject variations, as stated by the authors: a rendered version of each VV was shown to the participants on a LCD display. Subramanyam et al. [40] also used the ACR-HR method to assess the quality of digital humans represented as dynamic point clouds. The experiment was conducted in Virtual Reality (VR), in two different viewing conditions enabling 3 and 6 degrees of freedom (DoF). Gutiérrez et al. [27] implemented the ACR-HR method in a MR environment. The observers were asked to freely explore the displayed 3D models.

Few subjective studies on 3D content have been conducted using the pairwise comparison method (PC). Guo et al. [3] opted for this method to produce their ground-truth dataset of textured 3D meshes. They animated each object in the dataset with a low-speed rotation and generated videos that were displayed to observers during the test. The same experimental procedure was adopted by Vanhoey et al. [41] to investigate the impact of light-material interactions on the perception of geometric distortion of 3D models. Christaki et al. [24] subjectively assessed the perceived quality of 3D meshes subjected to different 3D mesh compression codecs. The experiment was conducted in a Virtual Reality setting, and the content was viewed freely as a combination of natural navigation (i.e. physical movement in the real-world) and user interaction.

Despite the above work, the majority of existing experiments implement the double stimulus methodology (derived from DSIS), using various modalities to display 3D objects. Watson et al. [15] et Filip et al. [42] used still screenshots/images to evaluate mesh simplification and material distortions, while Lavoué et al. [20] and Silva et al. [18] considered free-viewpoint interactions for evaluating 3D meshes subject to noise addition and simplification. Torlig et al. [32] also used an interactive platform that was developed to display their point clouds voxelized in real-time. In AR and VR scenarios to assess the quality of point cloud models, Alexious et al. [30,43] let observers interact with the virtual stimuli with 6 DoF by physical movements in the real-world. Other authors controlled the interaction between the observer and the 3D object by using animation: Rogowitz et al. [16] and Pan et al. [17] animated their meshes with a slow rotation. Similarly, in the large-scale subjective experiment conducted on colored point clouds [36], observers can only rotate the models. Javaheri et al. [35] generated 2D rendered videos of the original and decoded point clouds. They mentioned that the use of DSIS allowed to mitigate the impact of the density of original point clouds, as well as the impact of acquisition artifacts. Da Silva Cruz et al. [34] and Su et al. [31] also employed a passive inspection protocol with defined camera paths to capture the models under evaluation.

It seems that most researchers intuitively felt that rating the absolute quality of a 3D graphical model (i.e., without the reference nearby) might be a difficult task for a naive non-expert observer.

We denote that only in [24,27,30,40,43], the experiments were conducted in an immersive environment (either in VR, AR or MR). So far, very few attempts have been made to understand the impact of display devices (2D screen, VR/AR/MR headsets) on the perceived quality of 3D content. A study reported in [44], compared the results of experiments [29] and [30] conducted on points clouds in a desktop and AR setup, respectively. The study

revealed different rating trends under the usage of different display devices as a function of the degradation type being evaluated.

Resulting datasets

Unfortunately, among the works presented above, very few have publicly released their datasets. Table 1.1 outlines the publicly available subjective quality datasets for 3D content. For 3D meshes, the available datasets of mean opinion scores concern mostly geometry-only content [18–20, 23, 25, 26, 45] and are all rather small (see the first 7 rows in Table 1.1). The only public datasets involving 3D meshes with color information are provided by Guo et al. [3] and from Zerman et al. [1], and contain respectively 136 and 24 distorted stimuli. For both datasets, the color information is provided as texture maps. For colored point clouds, the largest available datasets are those provided mainly by Alexiou et al. [33], Su et al. [31], Zerman et al. [1], and Liu et al. [36]. The latter established a massive dataset with more than 24k distorted point clouds (see Table 1.1 last row). This is the largest to date. Note that, the dataset reported in [1] actually contains both meshes and point clouds, for a total of 152 stimuli that were rated in the same subjective test (Table 1.1, row 14).

All these datasets were generated through experiments conducted on screen, except [30] which was produced in AR.

1.1.3 Comparison of subjective methodologies

Several works evaluate and compare the performance of the subjective quality assessment methodologies described above in section 1.1.1. The majority of these comparisons were performed on natural images/videos. Péchard et al. [48] evaluated the impact of the video resolution on the behavior of both ACR and SAMVIQ methods. They found that, for a given number of observers, SAMVIQ is more accurate especially when the video resolution increases. They also stated that the accuracy of the methods depends on the number of observers: 22 observers are required in ACR test to obtain the same accuracy than SAMVIQ with 15 observers. Nevertheless, SAMVIQ is considerably more time-consuming than ACR (or DSIS). Contrary to what the ITU-R BT.500-13 [11] recommends regarding the minimum number of subjects required for ACR (15 observers), VQEG [13] and Brunnström et al. [49] recommended the use of at least 24 observers.

Moving to double stimulus methods (such DSIS), the main difference between them and single stimulus methods (such ACR) is the presence of explicit references. According to [10], DSIS ratings are less biased compared to ACR ratings. Indeed due to the presence of explicit references, subjects are able to detect shape and color impairments that they may miss with the ACR method. In addition, in DSIS, the scores are not influenced by the subjects opinion of the content. Surprisingly, Mantiuk et al. [12] denoted that for the images and distortions used in their study, there was “no evidence that the double stimulus method is more accurate than the single stimulus method”. They also demonstrated that since the Pair Comparison (PC) methodology is straightforward, it tends to be the most accurate from the 4 tested methods (single stimulus, double stimulus, forced

Table 1.1: Public quality assessment datasets for 3D meshes and point clouds.

Dataset	3D representation	Static or Dynamic	Attributes	Inspection	Methodology	Distortion types	# Distorted stimuli	# Total Participants
LIRIS / EPFL DB [19]	Mesh	Static	Colorless	Interactive	Single stimulus	Noise addition Smoothing	84	12
LIRIS Masking DB [20]	Mesh	Static	Colorless	Interactive	Double stimulus	Noise addition	24	11
IEETA simplification DB [18]	Mesh	Static	Colorless	Interactive	Double stimulus	Simplification	30	65
UWB #1 DB [23]	Mesh	Static	Colorless	Passive	2AFC	Compression	63	69
RG-PCD DB [45]	Mesh	Static	Colorless	Passive	Double stimulus	Octree-pruning	24	126
UWB #2 DB [25]	Mesh	Dynamic	Colorless	Passive	Multiple stimulus	Compression Noise addition	36	37 ~ 49
3D Mesh Animation Quality DB [26]	Mesh	Dynamic	Colorless	<ul style="list-style-type: none"> Passive Interactive 	Single stimulus	Noise addition Compression Transmission error	276	<ul style="list-style-type: none"> 16 25
LIRIS Textured Mesh DB [3]	Mesh	Static	Texture maps	Passive (Generated videos)	Pair Comparison	<ul style="list-style-type: none"> On geometry: Compression Simplification Smoothing On texture: Compression sub-sampling 	136	101 (Exp.1) 20 (Exp.2)
G-PCD DB [29,30]	Point Cloud	Static	Colorless	<ul style="list-style-type: none"> Interactive Interactive in AR 	<ul style="list-style-type: none"> Single & Double Stimulus Double Stimulus 	Noise addition Octree-pruning	40	<ul style="list-style-type: none"> 28 24
M-PCCD DB [33]	Point Cloud	Static	Colorless	Interactive	Double Stimulus & Pair Comparison	Compression	244	80
Su et al. DB [31]	Point Cloud	Static	Colorless	Passive (Generated videos)	Double Stimulus	Compression Noise addition Octree-pruning	740	60
IST Rendering Point Clouds DB [35]	Point Cloud	Static	Colorless & Colored	Passive (Generated videos)	Double Stimulus	Compression	54	60
Volumetric Video Quality #1 DB [46]	Point Cloud	Dynamic	Colored	Passive (Generated videos)	Double Stimulus & Pair Comparison	Compression	32	19
Volumetric Video Quality #2 DB [1]	<ul style="list-style-type: none"> Point Cloud Mesh 	Dynamic	<ul style="list-style-type: none"> Colored Texture maps 	Passive (Generated videos)	Single Stimulus	Compression	<ul style="list-style-type: none"> 128 24 	23
SJTU-PCQA DB [47]	Point Cloud	Static	Colored	Interactive	Single Stimulus	Compression Noise addition Scaling	420	64
Liu et al. DB [36]	Point Cloud	Static	Colored	Interactive	Double Stimulus	Compression Noise addition Transmission error	<ul style="list-style-type: none"> 1020 (with MOS) 23732 (with Pseudo-MOS) 	160

choice pairwise comparison, and similarity judgments methods). However, despite the simplicity of the task in PC tests, it may become tedious if all sequences need to be tested (PC requires $\frac{n(n-1)}{2}$ trials to assess n sequences while ACR requires $n+1$ trials and DSIS n trials). Tominaga et al. [50] compared eight subjective quality assessment methods for mobile videos. They denoted that ACR, DSIS and SAMVIQ are the most reliable among the tested methods and showed that ACR is the most suitable method for quality assessment of mobile video services, in terms of total assessment time and ease of evaluation. In 2014, Kawano et al. [51] investigated the performance of ACR, DSIS and Double Stimulus Continuous Quality Scale (DSCQS) for assessing the quality of 2D and 3D Videos. They found that, ACR is the most time efficient and DSIS is the most stable. In terms of the discrimination ability, they stated that DSIS outperforms the others for low quality-videos, while DSCQS is better for high-quality videos.

Recently, Alexiou et al. [52] extended their work [29] on quality assessment of point clouds by comparing the results of an ACR test and a DSIS test, in which subjects were able to interact with the point clouds viewed on screen. They found that, the DSIS method is more consistent in identifying the level of impairments. Singla et al. [53] evaluated the performance of the DSIS and the Modified Absolute Category Rating (M-ACR) methods for omnidirectional (360°) videos using an Oculus Rift. They denoted that M-ACR is statistically slightly more reliable than DSIS since DSIS resulted in larger confidence intervals. Adhikarla et al. [54] evaluated the quality of dense light fields. They first experimented ACR but found that this method is not sensitive to subtle but noticeable degradation of quality. In addition, participants found the rating task difficult. They therefore opted for PC with a two-alternative-forced-choice, as this method is more sensitive to impairment. Subjective studies reported in [33,46] implemented both DSIS and PC to evaluate the quality of point clouds. Indeed, according to the authors, PC is not suitable for evaluating large differences in quality; therefore, they used DSIS to capture large differences introduced by distortions, and PC when the stimuli are of nearly equal quality.

1.1.4 Subjective quality assessment in crowdsourcing

Subjective quality assessment experiments (such as all of the works mentioned above) have traditionally been conducted in laboratories (lab) in a controlled environment and with high-end equipment. In recent years, CrowdSourcing (CS) experiments have become very popular, especially with the development of the internet and the growing trend of machine learning. They provide alternatives to laboratory (lab) experiments in certain cases, particularly during the COVID-19 pandemic, where participants could not be physically present in the laboratory to carry out tests. CS has been exploited in several fields and for different types of media, such as the evaluation and annotation of images, videos, audio, speeches and documents.

CS and lab studies differ considerably in several aspects. (1) A task is performed by an unspecific internet crowd in the former rather than a specific group of people in the latter. Thus, CS enables researchers to access a much larger and more diverse subject pool and build generalized datasets representative of real-life scenarios. (2) Experiments conducted in a lab environment typically last around 20-30 minutes [11], while CS experiments should be kept as short as possible. Indeed, previous works [55–57] pointed out that a CS task

should last 5-10 minutes to avoid participants' boredom, frustration and decreased attention, leading to unreliable behavior and results. (3) Regarding time-effort, CS experiments are dramatically less time-consuming than lab tests, especially when evaluating large datasets. (4) Last but not least, lab experiments allow better control of the study setup, while CS experiments are carried out in uncontrolled test environments (different viewing conditions that affect participants' perception of quality, e.g. lighting, bandwidth constraints, display device, distance between the participant and the viewing screen, etc.).

As a result, CS imposes several challenges to overcome compared to similar lab tests, notably those related to the lack of control over the participants' environment and the trustworthiness of the participants since they are not supervised in these tests. A thorough overview of these concerns can be found in [58, 59]. To detect and deal with malicious/unreliable participants, several mechanisms have been proposed over the years [58, 60, 61]. Despite these challenges, CS studies are still capable of producing accurate and reliable results if the experiment framework has been properly designed [62, 63].

Regarding the experimental methodologies used in crowdsourcing, most CS studies have used the pairwise comparison (PC) method as it is straightforward: the task of choosing one of the two stimuli is simpler than rating them on a discrete or continuous scale [62, 64, 65]. Other works have adopted the Absolute Category Rating (ACR) method [56, 63]. Available crowdsourcing frameworks already implement these methods and offer the possibility of modifying them to fit the needs of the study. A detailed overview of these frameworks and an evaluation of them is provided in [66].

1.2 Objective quality assessment

Since subjective quality assessment tests can be very time consuming and tedious, objective quality assessment metrics are critically needed to automatically predict the level of annoyance/distortion caused by the processing operations cited in the introduction of this chapter. As this thesis focuses on 3D meshes with color attributes, we will review, in this section, the state-of-the-art of objective quality assessment metrics for 3D meshes. Moreover, we will address the visual impact of distortions applied on the 3D meshes themselves (e.g., distortions introduced by compression, simplification or filtering); and we will not cover the visibility prediction of artifacts introduced during the rendering process (e.g., by global illumination approximation) or after rendering (e.g., by tone mapping). A dataset has been recently introduced that focuses on these types of image artifacts [39].

Simple geometric measures, such as Hausdorff distance [67], Mean Squared Error (MSE), Root Mean Squared (RMS) error [68] and Peak Signal-to-Noise Ratio (PSNR), are only weakly correlated with the human vision since they are based on pure geometric distances and ignore perceptual information [69]. Hence many perceptually driven visual quality metrics have been proposed. Most popular perceptual quality metrics (for images and 3D content) are based on top-down approaches. They treat the Human Visual System (HVS) as a black box and try to identify changes in content features induced by distortions to estimate perceived quality. In contrast, other metrics are based on bottom-up approaches. They rely on computational models of the Human Visual System (HVS) by modeling its relevant

components. With the rise of machine learning a third category of quality metrics emerged. These metrics are based on a purely data-driven approach, and do not rely on any explicit model. The field of image quality assessment has shown many successful uses of machine learning, particularly Convolutional Neural Networks (CNN) [4, 70–75]. Readers can refer to [76] for a comprehensive study determining the underlying reasons why deep features are good image quality predictor and may perform better than traditional (top-down and bottom-up) approaches.

3D mesh quality metrics can be classified as: model-based metrics and image-based metrics. The former operates on the 3D mesh domain (on the 3D model itself and its attributes like texture maps) while the latter predict quality using 2D snapshots of the rendered 3D model, on which Image Quality Metrics (IQMs) are computed. The rest of this section details these two approaches.

1.2.1 Model-based quality metrics

Many Mesh Visual Quality (MVQ) metrics have been proposed in the literature. These metrics are mostly Full-Reference (FR), meaning that the distorted model is compared to its reference. FR metrics are mainly used to guide/drive compression, simplification and watermarking of meshes [19, 22, 23, 77]. Most of them follow the classical approach (top-down) used in image quality assessment: local feature differences between the reference and distorted meshes are computed at vertex level, and then pooled over the entire 3D model to obtain a global quality score.

Most of the existing metrics evaluate only geometric distortions, i.e. they rely on geometric characteristics of the mesh without considering its appearance attributes. In [78], the authors proposed combining the RMS geometric distance between corresponding vertices with the RMS distance of their Laplacian coordinates which reflect the degree of surface smoothness. Bian et al. [79] proposed a Strain Energy Field-based measure (SEF) based on the energy introduced by a specific mesh distortion: the more the mesh is deformed, the greater the probability of perceiving the difference between the reference and distorted meshes. Lavoué et al. [19, 80] proposed a metric, called MSDM2, inspired by the well-known image quality metric SSIM [81]. The authors extended the SSIM to 3D meshes by using the mesh curvature as an alternative for the pixel intensities. This metric is adapted for meshes with different connectivities. Torkhani et al. [82] also proposed a metric based on local differences in curvature statistics. They included a visual masking model to their metric. In [23], the authors considered the dihedral angle differences between the compared meshes to devise their metric DAME. The above metrics consider local variations at the vertices or edges. Corsini et al. [22] proceeded differently. They computed one global roughness value per 3D model considering dihedral angles and variance of the geometric Laplacian and then derive a simple global roughness difference. In a similar approach, Wang et al. [83] proposed a metric called FMPD based on global roughness computed using the Gaussian curvature. A survey [84] detailed these works and showed that MSDM2 [80], DAME [23] and FMPD [83] are excellent predictors of visual quality.

Besides these works on global visual quality assessment (top-down approaches adapted for

supra-threshold distortions), Nader et al. [85] introduced a bottom-up visibility threshold predictor for 3D meshes. Guo et al. [86] also studied the local visibility of geometric artifacts and showed that curvature could be a good predictor of distortion visibility.

Several works used machine learning techniques in assessing the quality of 3D meshes. Lavoué et al. [87] optimized the weights of several mesh descriptors using multi-linear regression. Chetouani et al. [88] proposed a quality measure based on the fusion of selected features using the Support Vector Regression (SVR). In [89], a machine learning-based approach for evaluating the quality of 3D meshes is proposed, in which crowdsourced data is used while learning the parameters of a distance metric.

Moving to 3D dynamic meshes, Váša et al. [25] proposed a metric, called STED, based on the comparison of mesh edge lengths and vertex displacements between two animations. Torkhani et al. [26] devised a quality metric for 3D dynamic meshes which is a combination of spatial and temporal features. In more recent work, Yildiz et al. [90] developed a bottom-up approach incorporating both the spatial and temporal sensitivity of the HVS to predict the visibility of local distortions on the mesh surface.

For some use cases, the reference is not available. Therefore, No-Reference (NR) quality assessment metrics are needed. Unlike FR metrics, few NR quality metrics for 3D meshes have been proposed in the literature. These metrics are based on data-driven approaches (machine learning). Abouelaziz et al. [91] proposed a NR metrics that relies on the mean curvature features and the General Regression Neural Network (GRNN) for quality prediction. The NR metric proposed in [92] (BMQI) is based on the visual saliency and the Support Vector Regression (SVR), while that proposed in [93] is based on dihedral angles and SVR. Abouelaziz et al. [94] also used Convolutional Neural Networks (CNN) to assess the quality of 3D meshes. The CNN was fed with perceptual hand-crafted features (dihedral angles) extracted from the 3D mesh and presented as 2D patches.

All the works presented above consider only the geometry of the mesh, and therefore only evaluate geometric distortions. Regarding 3D content with color or material information, little work has been published. For meshes with diffuse texture, Pan et al. [17] derived from the results of a subjective experiment a quantitative metric that approximates perceptual quality based on texture and geometry (wireframe) resolution. Tian et al. [77] and Guo et al. [3] proposed metrics based on a weighted combination of a global distance on the geometry and a global distance on the texture image. Tian et al. [77] combined the MSE computed on the mesh vertices with that computed on the texture pixels, while Guo et al. [3] linearly combined MSDM2 [80] (mesh quality) and SSIM [81] (image quality) metrics. These metrics combine errors computed on different domains (3D mesh and texture image). To the best of our knowledge, to date, there is no model-based quality metric for 3D meshes with colors attributes that works entirely on the mesh domain.

3D data can also be represented using point clouds. The field of point cloud quality assessment is still an emerging field. Simplest metrics include point-to-point and point-to-plane distances [95]. For each point of the content under evaluation, its closest point from the reference content is computed using nearest neighbor search. The point-to-point distance refers to the distance between those two points, while the point-to-plane distance refers to the projection of the distance vector along an average normal vector. These simple distances show good correlation results with subjective opinions for simple test content

(e.g. one single type of degradation, such as in [46]). However, they report poor results for most of subjective datasets [31, 52]. Very recently, Alexiou et al. [96] proposed a metric based on differences of normal orientations and Meynet et al. [97] proposed a metric integrating the curvature information. Both metrics demonstrated improved performance. Surprisingly, whereas several subjective studies involved colored point clouds [31,33–35,46], few attempts have been made to create a quality metric that takes into account both geometry and color attributes: PCQM [98] based on a linear combination of geometry-based and color-based features, GraphSIM [99] which uses graph signal gradient as a quality index, and NR-PCQA [36] based on the sparse CNN.

As can be seen, most existing model-based quality metrics ignore the visual saliency of 3D models, yet finding salient regions (regions that attract the attention of observers) has become a useful tool for many applications such as mesh simplification [100] and segmentation [101], and quality control of VR videos (360 videos) [102,103].

1.2.2 Image-based quality metrics

To evaluate the quality of 3D content, several authors considered Image Quality Metrics (IQMs) computed on rendered snapshots. This approach can be efficient since the field of image quality assessment is highly developed [104] and many successful IQMs have been introduced: Sarnoff VDM [105], SSIM [81] (and its derivatives), VIF [106], FSIM [107], HDR-VDP2 [108,109], iCID [110], BLINDS [111], GMSD [112], [70,113], DeepSIM [72], LPIPS [4], WaDIQaM [73], NIMA [74], PieAPP [114], etc.

The image-based approach was first used to drive perceptually-based tasks, such as mesh simplification. Qu and Meyer [115] considered the visual masking properties of 2D texture maps, Sarnoff VDM [105], to drive simplification and remeshing of textured meshes. Menzel and Guthe [116] considered 2D models of the Contrast Sensitivity Function (CSF) to drive the Level of Detail simplification (LoD) of 3D meshes. Zhu et al. [117] studied the relationship between the viewing distance and the perceptibility of model details to optimize the LoD design of complex 3D building facades. They used VDP [118] and SSIM [81]. Caillaud et al. [119] used SSIM [81] to optimize textured mesh transmission. Lindstrom and Turk [120] considered a view-independent approach to evaluate the impact of simplification on 3D models. They used the RMS error computed on snapshots taken from different viewpoints (different camera positions regularly sampled on a bounding sphere). In the field of point cloud quality assessment, several authors also computed 2D image metrics on a set of snapshots around the point clouds and reported correct correlations with subjective scores [31,33].

Recently, several authors have started to exploit Convolutional Neural Networks (CNN) to assess the quality of 3D meshes using an image-based approach. The existing works considered geometry-only meshes (without color attributes). In [121], the CNN was fed with 2D rendered images of the 3D mesh generated by rotating the object. Abouelaziz et al. [122] devised a quality metric for 3D meshes by extracting feature vectors from 3 different CNN models and combining them using an extension of the Compact Bi-linear Pooling (CMP). The authors used a patch-selection strategy based on mesh saliency to give more importance to perceptual relevant (attractive) regions. In fact, not all regions of the 3D model image receive the same level of attention from observers. In general, distortions

in the salient regions (regions that attract the attention of observers) are assumed to be the most disturbing.

Note that in the field of image quality assessment, many deep learning-based methods have successfully integrated saliency [73, 123, 124].

1.2.3 Comparison of the two approaches

Few works benchmarked the performance of image-based metrics and compared them to model-based approaches for quality assessment of 3D models. As reported in [120], the advantage of image-based metrics over model-based metrics is their natural ability to handle the multimodal nature of data (geometry and color or texture information, normals), as well as their natural incorporation of the complex rendering pipeline (computation of light material interactions and rasterization). In [16], the authors conducted two subjective quality assessment experiments: the first involved 2D static images of simplified 3D objects, while the second was performed on animated sequences of these objects in rotation. The results showed that lighting conditions have a strong influence on perceived quality and that observers perceive the quality of still images and animations differently. The authors concluded that the quality of 3D objects cannot be correctly predicted by the quality of static 2D projections. Cleju and Saupe [125] found that image-based metrics generally perform better than model-based metrics, however the SSIM performance is more sensitive to the 3D model type. In 2016 [69], a more exhaustive/comprehensive study compared the results of the most efficient image metrics (at the time) to those of model-based metrics on datasets of geometry-only meshes. The authors considered different lighting conditions, rendering protocols, different ways of combining the image metric values and several datasets containing different 3D models and types of distortion. They found that the performance of image-based metrics greatly depends on distortions and contents: image-based metrics perform very well in evaluating the quality of different versions of a same object under a single type of distortion. However, they are less accurate in differentiating and ranking different distortions, or distortions applied to different 3D models. This finding is in line with the results reported in [3].

The main benefit of using image-based metrics to evaluate the visual quality of 3D objects is the natural handling of the complex interactions between the different mesh properties involved in the appearance, which avoids the problem of how to combine and weight them. On the other hand, these metrics pose other problems/limitations: (1) it is necessary to know in advance the final rendering of the stimuli, such as the lighting conditions, the displayed viewpoint (since they operate on 2D rendered snapshots). (2) Using them in a view-independent approach introduces new parameters such as the choice of the 2D views, and the pooling of quality scores obtained from different views into a single global score. (3) Finally, image-based metrics may not be practical for driving processing operations (e.g. mesh simplification). Model-based metrics may be better suited for these operations since they operate on the mesh domain, i.e. the same representation space as mesh processing algorithms. It is thus possible to control processing operations as well globally (on the whole mesh) as locally (at the vertex level).

There are no works that compare model-based and image-based metrics on datasets of 3D content with color attributes.

1.3 Conclusion

In this chapter, we presented an overview of existing works on subjective and objective quality assessment of graphical 3D content. Despite recent progress, there are still several limitations to overcome, especially when it comes to 3D graphics with color attributes. Here is a summary of these limitations as well as the positioning of our work in relation to them.

In the field of subjective evaluation, there is a lack of consensus on the best methodology to adopt for evaluating the quality of 3D graphics (section 1.1.3). We will address this problem in Chapter 2. Based on subsection 1.1.2, there are only two public quality assessment datasets for textured meshes and none for meshes with vertex colors. These datasets are rather small and were generated through experiments conducted on screen. Thence, we will produce two new datasets: a dataset of meshes with vertex colors produced in VR (presented in Chapter 3) and another large-scale dataset of textured meshes produced in crowdsourcing (presented in Chapter 5).

As discussed in subsection 1.1.2, researchers have adopted different ways to display the 3D models to observers in subjective tests. However, no attempt has been made to explore how the selection of these models, distortions, viewpoints and movements affect their perceived quality. These factors are relevant and of high importance in case of 6DoF interactions. Leveraging our 2 established datasets, we will address this problem: we will conduct in-depth analyzes in chapters 3 and 5 to study the influence of these factors on the visual quality of 3D graphics. Last but not least, since crowdsourcing (CS) was mainly used for 2D image and video quality assessment tasks (section 1.1.4), we will investigate in Chapter 4 the reliability of CS tests to assess the quality of 3D graphics.

Moving to objective evaluation, most metrics in the literature evaluate only geometric distortions. When it comes to meshes with color information (either in the form of texture or vertex-colors), little work has been published (see section 1.2.1). Therefore, we will develop, in Chapters 6 and 7, two novel perceptual quality assessment metrics that take into account both geometry and color attributes.

Part I

Subjective Quality Assessment

Chapter 2

Comparison of Subjective Methods for Quality Assessment of 3D Graphics in Virtual Reality

Designing a subjective experiment and selecting its experimental methodology are not trivial tasks because we must ensure that such an experiment yields valid, reliable and replicable results. In the past years, the International Telecommunication Union (ITU) and the Video Quality Experts Group (VQEG) have defined several methodological guidelines for 2D image and video subjective quality assessment tests [9–11,13] (see Chapter 1, section 1.1.1). No similar standards exist for quality evaluation of 3D graphics: in the field of computer graphics, most researchers have intuitively used existing methodologies for images and videos to assess the quality of 3D graphics (see Chapter 1, section 1.1.2). However, there is no evidence that these methods are valid / accurate for assessing the quality of such data. In fact, no comparison of subjective methodologies has been made for 3D graphics and therefore, there is no consensus on the best methodology to adopt for assessing the visual quality of these data, especially in a virtual or mixed reality environment. A particular open question is whether or not a reference is necessary to assess the quality of 3D graphics.

In this chapter, we attempt to make a first step toward standardizing a methodology for assessing the quality of 3D graphics. For that, we aim to answer three main questions:

- Are the subjective methodologies defined for 2D imaging quality assessment tests accurate for assessing the quality of 3D graphics?
- Is the presence of an explicit reference necessary in 3D graphics quality assessment, due to the lack of prior human knowledge about 3D graphics data compared to natural images/videos?
- What is the best methodology to assess the quality of 3D graphics?

In this regard, we selected three of the most prominent subjective methodologies in the field of image processing, involving hidden and explicit references. We evaluated their

performance on a dataset of high-quality colored 3D models, impaired with various distortions, throughout two psycho-visual experiments. We considered a Virtual Reality (VR) context for our subjective experiments since VR user studies offer the most ecological and realistic use cases for visualizing 3D content and are in high demand.

Our study allowed us to draw interesting conclusions about (1) the importance of the presence of explicit references to assess the quality of 3D graphics, (2) the most suitable methodology in terms of accuracy and time effort, and last but not least, (3) the minimum number of participants required for such experiments.

This chapter is organized as follows: We first describe, in section 2.1, the selected methodologies, then, in sections 2.2, 2.3 and 2.4, we detail the dataset as well as the experimental procedure used throughout our study. In sections 2.6 and 2.7, we analyze the results of our subjective experiments. Finally, concluding remarks and recommendations are outlined in sections 2.8 and 2.9.

2.1 Experimental methodologies

As mentioned previously, the purpose of our study is to determine the impact of the explicit reference on the quality assessment of 3D graphics and more generally to determine the best experimental methodology to adopt for such data. Thus, we selected 3 widely known test methods, for one of which the reference is hidden and for the two others the reference is explicit. The selected methodologies are presented below and illustrated in Figure 2.1

- **Absolute Category Rating with Hidden Reference (ACR-HR)**: also known as single stimulus categorical rating, in which the impaired stimuli are presented one at a time in addition to the original unimpaired stimuli (references), without informing the subjects of their presence. The observers are asked to evaluate the quality of the stimulus shown using a five-level scale, where the discrete levels correspond to bad, poor, fair, good, and excellent. Note that some methods favor continuous rather than categorical scales to avoid quantizing errors [11]. According to ITU-R BT.500-13, the presentation time for the stimulus should be ~ 10 s. It may be reduced or increased according to the content of the test sequence [10]. In our pilot study (pretests), we found that 6s presentation is sufficient to assess the quality of the presented 3D model.
- **Double Stimulus Impairment Scale (DSIS)**: also called Degradation Category Rating (DCR), in which the viewer sees an unimpaired reference model, then the same model impaired. Following that, the subject is asked to rate the impairment of the second stimulus in relation to the reference [10] using the following five-level impairment scale: Imperceptible, Perceptible but not annoying, Slightly annoying, Annoying, Very annoying. Similarly to ACR-HR, 10s presentation time is recommended per stimulus (~ 20 s / pair). However, this implementation slows-down the experiment too much since it requires at least twice as much time as ACR-HR method. The total duration of the experiment affects the efficiency of the experimental method especially in virtual reality (VR) where most subjects are potentially not used to the

CHAPTER 2. COMPARISON OF SUBJECTIVE METHODS FOR QUALITY ASSESSMENT OF 3D GRAPHICS IN VIRTUAL REALITY

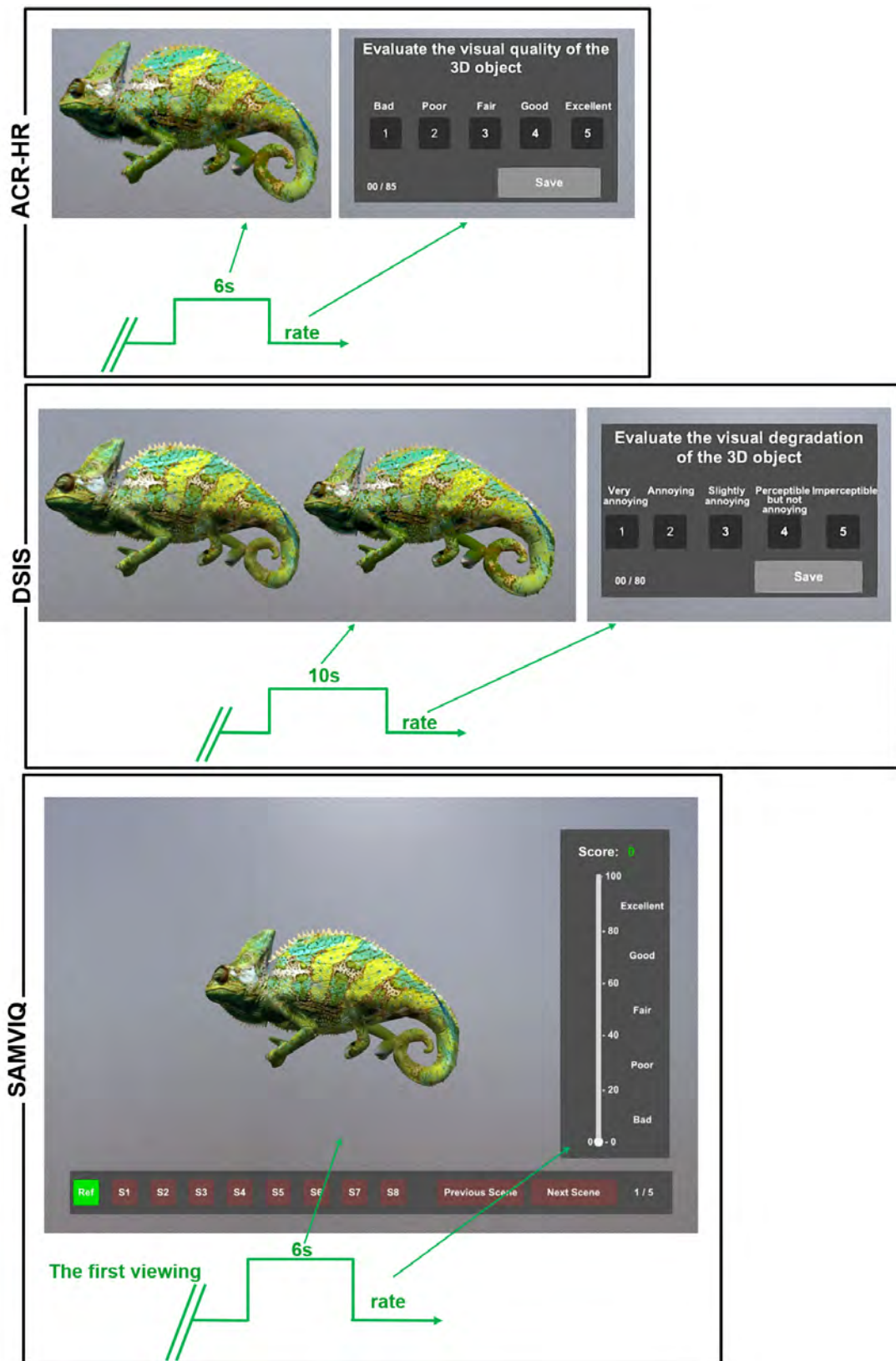


Figure 2.1: Illustration and timeline of the three subjective quality assessment methodologies explored in this study.

VR headset and tend to exhibit symptoms of cybersickness both during and after the experience [126]. To avoid this issue, 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. In addition, using simultaneous presentation makes it easier for subjects to assess differences between stimuli [10]. Note that this “simultaneous” version of DSIS is what is preferred in most subjective tests involving 3D content [8, 17, 34, 127]. For DSIS, we increased the presentation time to 10s, since comparing to ACR-HR, 6s are not sufficient to observe the 2 stimuli displayed in the scene and assess their impairments.

- **The Subjective Assessment Methodology for Video Quality (SAMVIQ)** is a multiple stimuli assessment methodology [9]. This means that each stimulus can be seen and assessed as many times as the observers want: subjects are allowed to access the stimuli and adjust their scores, as appropriate. The SAMVIQ test is divided into *scenes*. The content of a *scene* has to be homogeneous: a *scene* presents an explicit reference model as well as several versions of the same model with different impairments. Each stimulus is presented on its own and rated using a continuous quality scale. The continuous scale is graded from 0 to 100 (typically represented by a slider) and divided into five equal portions: Bad (0 to 20), Poor (20 to 40), Fair (40 to 60), Good (60 to 80), Excellent (80 to 100). The associated terms categorizing the different levels correspond to the basic ITU-R BT.500-13 five-level quality scale and are included for general guidance [128]. For SAMVIQ, the subjects are asked to assess the overall quality of the stimuli. Each stimulus (including the explicit reference) must be fully viewed at least once and then, the observer rates it. During the first viewing, all the other stimuli access buttons including the sliding rating scale are disabled. Once the current stimulus has been graded, the subject can access the previous rated stimuli to adjust their scores if needed. The latest score of each stimulus is recorded. Note that, the number of distorted stimuli is limited to ten per scene to avoid boredom and fatigue. All the stimuli of the current scene must be scored before the assessor can proceed to the next *scene* or visit the previous *scene*. To finish the test, all the stimuli of all the *scenes* must be scored. This method is functionally similar to single stimulus method (e.g. ACR-HR) with random access, nevertheless a subject can view the explicit reference whenever he wants and compare it directly to the impaired stimulus. This makes SAMVIQ similar to methods that use explicit references (e.g. DSIS). Following the ITU-R BT.1788 recommendations [9], the maximum presentation time for a stimulus is in the range of 10 to 15 s. Since SAMVIQ method provides a global score, like single stimulus methods, we set the presentation time to 6s, as for the ACR-HR method.

Test methods with hidden references (e.g. ACR-HR) better simulate real-life consumption of visual data, while methods with explicit references (DSIS, SAMVIQ) are commonly used for their high discriminating power: the presence of a reference usually facilitates the identification of differences. Additionally, these methods naturally eliminate biases from observers’ personal opinion of the content (whether they like or dislike the object), unlike methods with hidden references.

2.2 Dataset generation

The first step in conducting a subjective experiment is to prepare the dataset that will be rated by the participants. In our study, we generated a dataset of 80 meshes with diffuse color information created from five reference/source objects, on which we applied geometry and color distortions.

2.2.1 3D source model selection

To build our dataset of colored 3D graphics, we selected 5 high-resolution triangle meshes, each having diffuse color information represented by vertex colors (no texture mapping). These models are considered to be of “good” or “excellent” quality. They were chosen so as to ensure a variety of shapes and colors. Table 2.1 details the characteristics of the models, while Figure 2.2 illustrates them. Note that, the sixth model (*Dancing Drummer*) is not part of the dataset. It was used at the training stage of the experiments.

2.2.2 Distortions

The source models presented above have been corrupted by the following four types of distortion applied on geometry and color. These selected distortions represent common simplification and compression operations typically used in 3D model modeling and post-processing.

- Uniform geometric quantization (QGeo): applied on geometry. This is a very common process for lossy compression.
- Uniform LAB color quantization (QCol): applied on vertex colors. This is inspired by the usual 2D image compression processes.
- “Color-ignorant” simplification (SGeo): a surface simplification algorithm that takes into account geometry only. It consists of iterative edge collapse operations driven by the quadric error metrics [5].
- “Color-aware” simplification (SCol): a surface simplification algorithm that takes into account both geometry and color. It consists of iterative vertex removal operations, driven by a combination of (1) a geometry metric: the area loss caused by the removal; and (2) a color metric: the LAB distance between the color of the vertex to be removed and its interpolation after removal [129].

Each distortion was applied with four different strengths, adjusted manually in order to span the whole range of visual quality from imperceptible levels to high levels of impairment: we generated a large set of distortions and then we selected a sub-set of them spanning the desired visual quality (as is typically the case in subjective image quality studies [130]). Figure 2.3 illustrates some visual examples, while all details about the distortion parameters are provided in Table 2.2.



Figure 2.2: Illustration of the 3D graphic source models.

Models	#Vertices	Geometry complexity	Color characteristics	Semantic category	Created using
Aix	686061	Plane with small details	Mono-color	Art	3D scanning
Ari	645492	Intermediate	Cool & light colors	Human statues	3D scanning
Chameleon	588441	High & sharp edges	Cool & dull colors	Animal	Modeling software
Fish	216578	Low & sharp edges	Cool & warm colors	Animal	Modeling software
Samurai	449997	High	warm colors	Human statues	3D scanning
Dancing Drummer	1335436	Intermediate/High	Cool colors	Human statues	3D scanning

Table 2.1: Characteristics of the 3D graphic source models



Figure 2.3: Some examples of distorted models. Acronyms refer to Distortion Type_Strength.

CHAPTER 2. COMPARISON OF SUBJECTIVE METHODS FOR QUALITY ASSESSMENT OF 3D GRAPHICS IN VIRTUAL REALITY

Distortion type	Distortion strength	Aix	Ari	Chameleon	Fish	Samurai
QGeo	1	10 bits	10 bits	9 bits	9 bits	10 bits
	2	9 bits	9 bits	8 bits	8 bits	9 bits
	3	8 bits	8 bits	7 bits	7 bits	8 bits
	4	7 bits	7 bits	6 bits	6 bits	7 bits
QCol	1	(L=5, A=4, B=4) bits	(L=5, A=4, B=4) bits	(L=4, A=3, B=3) bits	(L=5, A=5, B=5) bits	(L=4, A=3, B=3) bits
	2	(L=4, A=3, B=3) bits	(L=4, A=3, B=3) bits	(L=3, A=2, B=2) bits	(L=4, A=3, B=3) bits	(L=4, A=2, B=2) bits
	3	(L=3, A=2, B=2) bits	(L=2, A=3, B=3) bits	(L=2, A=2, B=2) bits	(L=3, A=2, B=2) bits	(L=3, A=2, B=2) bits
	4	(L=2, A=2, B=2) bits	(L=3, A=3, B=3) bits	(L=2, A=1, B=1) bits	(L=2, A=2, B=2) bits	(L=2, A=2, B=2) bits
SGeo	1	50% removed	30% removed	50% removed	31% removed	24% removed
	2	75% removed	50% removed	75% removed	50% removed	50% removed
	3	88% removed	75% removed	87% removed	77% removed	75% removed
	4	94% removed	87% removed	92% removed	88% removed	88% removed
SCol	1	71% removed	50% removed	67% removed	77% removed	66% removed
	2	87% removed	64% removed	83% removed	79% removed	80% removed
	3	94% removed	88% removed	92% removed	87% removed	90% removed
	4	98% removed	94% removed	95% removed	96% removed	96% removed

Table 2.2: Details on the distortions applied to each source model.

Thus, we generated 80 distorted models: 5 source models \times 4 distortion types \times 4 strengths. The snapshots of all the distorted models are provide in the appendix (Figure A.2 in section A.1).

2.3 Virtual Environment and apparatus

As mentioned in the introduction of this chapter, we chose to conduct the subjective experiments in a Virtual Environment (VE) since VR is becoming a popular way of consuming and visualizing 3D content. Moreover, it offers the most realistic use cases for such data. Thus, we used the HTC Vive Pro headset¹, a high-end virtual reality headset with a resolution of 1440 x 1600 pixels per eye (2880 x 1600 pixels combined), a field of view of 110 degrees and a refresh rate of 90 Hz. Note that, the HTC Vive Pro was used in the fixed position mode with its default color calibration.

2.3.1 Rendering

A crucial question when designing subjective experiments, especially those in virtual environments (VE), is how to display 3D models to observers and whether to use static or dynamic scenes. No standardized procedures yet exist for subjective studies involving 3D content. Indeed, the existing studies for such data implement different ways to display the models to the observers: still images, free interaction or animations. Rogowitz et al. [16] proved that the perceived degradation of still images may not be adequate to evaluate the perceived degradation of the equivalent 3D model, since still images may mask both artifacts and the effect of light and shading. Thus, it is important that the object moves.

¹<https://www.vive.com>

Following this approach, some researchers [22,127] allowed subjects to interact freely with the model by rotating and zooming it (real-time interaction). However, the problem of allowing free interaction is the cognitive overload which may alter human judgments. Therefore, we decided to control the interaction between the subject and the 3D object. Inspired by the principle of pseudo-videos and as in Guo et al. [3], we used animations. Thus, we selected, for each source model of our dataset, one viewpoint that we animated with a slow rotation of 15 degrees around the vertical axis in clockwise and then in counterclockwise directions (i.e. total rotation of 30 degrees). The viewpoint selected for each model was chosen perceptually so that it covers most of the shape, color and semantic information. Note that, the animation we generate does not involve nonrigid transformations of the objects.

The dynamic stimuli were rendered in a virtual scene (using a perspective projection) at a viewing distance fixed to 3 meters from the observer and rotate in real-time. Note that, in DSIS test, the reference and the distorted model were specifically oriented in order to show exactly the same vertices of the 2 models at the same time. Stimuli size is approximately 37 degrees of visual angle. Their material type complies with the Lambertian reflectance model (diffuse surfaces). The apparent brightness of such a surface to an observer is the same regardless of the observer's angle of view/position in the scene.

The stimuli are visualized in a neutral room (light gray walls) without shadows and under a directional light (all the vertices are illuminated as though the light was always from the same direction. It simulates the sun). We aimed to design a neutral room so that the experimental environment does not influence the users' perception of the stimulus.

2.3.2 Rating interface

We opted to ensure in our tests a user's Quality Of Experience (QoE) in a fully immersive environment. So, we integrated the rating interface in the VE of our experiments. This interface is adapted to each methodology (see Figure 2.1): for the ACR-HR and DSIS tests, a rating billboard (five level discrete scale board) was displayed after the presentation time of each stimulus / pair of stimuli and the stimuli were not shown during that time. For the SAMVIQ test, we implemented a vertical slider directly on the right side of the stimuli (continuous scale from 0 to 100).

There is no time limit to vote. The same neutral room (light gray walls) utilized to show the stimuli was used in the rating environment. To vote in the ACR-HR and DSIS tests, the subject selects and saves the score using the trigger of the HTC Vive controller, while in the SAMVIQ test, the subject selects the score (drag the slider to the chosen position) using the pad of the controller and switches between stimuli and *scenes* using its trigger. As in [131], to facilitate the interaction with the rating panel, we attached a raycast beam to the controller.

Figure 2.4 illustrates the experimental environment for each of the methodologies. The experiments were developed in Unity3D using c# scripting.

CHAPTER 2. COMPARISON OF SUBJECTIVE METHODS FOR QUALITY ASSESSMENT OF 3D GRAPHICS IN VIRTUAL REALITY

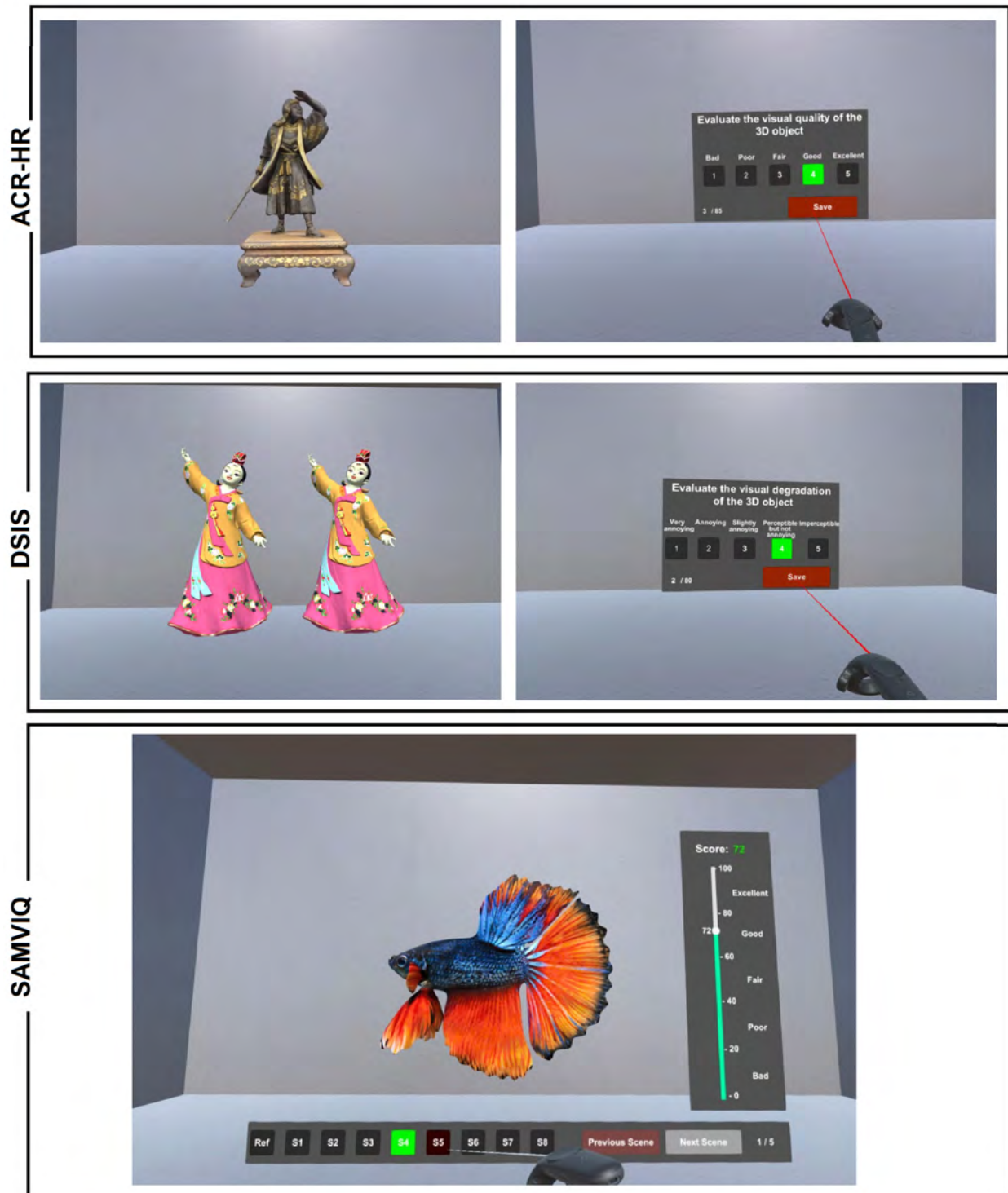


Figure 2.4: The experimental environments of the three methodologies implemented.

2.4 Experimental procedure

To address the questions raised in the introduction regarding the necessity of an explicit reference and the best methodology to evaluate the quality of 3D graphics, we designed two subjective experiments.

2.4.1 Subjective experiment 1

The goal of this experiment is to evaluate the impact of explicit references on the user quality assessment. For this purpose, we selected the ACR-HR and DSIS methods, since these 2 methods are similar in almost every aspect: each stimulus (for ACR-HR) or pair of stimuli (for DSIS) is presented once; the observer was not able to review the objects; the rating is based on a five-level discrete scale. The only difference between these 2 methods is related to the presence or not of an explicit reference: for ACR-HR, references are hidden, while for DSIS, references and distorted stimuli are displayed simultaneously side by side.

Thence, we divided our experiment into 2 sessions, one for each methodology i.e. one session consisted of presenting the stimuli using ACR-HR and the other session presented them using DSIS. In addition, in order to study whether a methodology has an influence over the other and if the order of the methodologies matters, we divided the subjects into 2 groups (G1 and G2). G1 refers to the participants who completed the ACR-HR test before DSIS and G2 refers to those who passed the DSIS session first then the ACR-HR session. None of these sessions took place on the same day in order to reduce the learning effect between stimuli. Thus, these two sessions were held at least two days apart. In each session, observers see all the stimuli of our dataset (80 stimuli, see section 2.2). The stimuli were displayed in a random order (3D models, distortions types and levels all mixed).

2.4.2 Subjective experiment 2

The purpose of this experiment is to evaluate an additional relevant methodology, in order to determine the best (the most suitable) methodology to adopt for quality assessment of 3D graphics. We chose SAMVIQ method [9]. Indeed, the SAMVIQ test differs greatly from a DSIS or ACR-HR test in several aspects: each stimulus can be seen and assessed as many times as the observers want; continuous scale from 0 to 100. Thus, by comparing the results of this experiment with those of the first experiment, we will be able to determine which method is the most adapted to evaluate the quality of 3D content. In addition, the SAMVIQ method has been of interest to several computer graphics researchers because of its high accuracy in assessing image/video quality. For instance in 2018, 3GPP² decided, in a study of VR media services [132], to start from the SAMVIQ method, since there is no existing standardized approach for subjective quality assessment in immersive environments. In this context, we conducted this experiment to investigate the performance

²The 3rd Generation Partnership Project, <https://www.3gpp.org/about-3gpp>

of this method for assessing the visual quality of 3D graphics.

Our experiment is organized such that each *scene* presents one source model and its distortions. Thus the experiment consists of 5 *scenes* (since our dataset is composed of 5 source models). In the SAMVIQ test, the maximum number of stimuli (condition) in each *scene* is limited to 10. Therefore, we divided the experiment into 2 sessions. Each session contains the 5 reference models (i.e. 5 scenes), each corrupted by 2 types of distortions, applied with 4 strengths ((Reference + 8 impaired stimuli) / scene). Thus we ensure that both sessions cover the entire quality range and have a balanced representation of visual qualities. We note that these two sessions occurred at least two days apart and for both sessions, the presentation order of the scenes was randomized across viewers.

2.5 Participants and training

2.5.1 Training

As recommended in ITU-R BT.500-13 [11], both experiments (both sessions in each experiment) started with a training in which observers could familiarize themselves with the virtual environment and the task. We selected a training 3D model not included in our original test set: “Dancing Drummer” (see Figure 2.2) and generated 11 distorted models that span the whole range of distortions. At the beginning of each session, the training models are shown in the appropriate way (single or pairwise) and with the same time (6s or 10s) adopted in the upcoming session. After each stimulus, the assigned score is marked on the corresponding rating interface (rating billboard or slider) for 5s. Note that for experiment 2 (SAMVIQ test), we insisted in the training on the possibility of switching between stimuli to correct the scores if needed. We added a practice trials stage at the end of the training: we displayed 2 extra stimuli and asked the subject to rate the quality (or the impairment according to the session). The results of these stimuli were not recorded. This stage was used to allow the observers to familiarize themselves with the experiment, to focus appropriately and to ensure that observers fully understand the task of the experiment.

2.5.2 Duration

No session took longer than 30 minutes to avoid fatigue and boredom. The total duration for the ACR-HR session was 18 minutes (informed consent/instructions + 11 training stimuli \times (6s display + 5s Rating) + 85 Test stimuli \times (6s display + \sim 4s rating)) and 23 minutes for DSIS session (informed consent/instructions + 11 training stimuli \times (10s display + 5s Rating) + 80 Test stimuli \times (10s display + \sim 4s rating)).

For the SAMVIQ test (composed of 2 sessions), each session lasted approximately 20 minutes to 30 minutes. It depended on the subject and how many times they viewed each test stimulus.

2.5.3 Participants

As mentioned previously for experiment 1, stimuli are rated by 2 groups of subjects. Thus, we involve in this experiment 30 subjects that we divide into 2 groups of 15: 27 males and 3 females, aged between 19 and 45. For experiment 2, a total of 17 subjects participated in the experiment: 4 females and 13 males, aged between 22 and 31. Participants, for both experiments, were students, interns, PhD students, engineers and professionals from the University of Lyon and LIRIS laboratory. They were naive about the purpose of the experiments. All observers reported a normal or corrected to normal vision. No participant did both experiments. In order to avoid the effect of the temporal sequencing factor, the order of stimuli was randomly generated so that each participant views the stimuli in a different order.

Table 2.3 summarizes all the experimental details of the three methodologies tested and highlights the main differences between them.

2.6 Results of Experiment 1 and influence of explicit reference

The following section presents the results of Experiment 1 described above and evaluates the impact of the session order as well as the impact of explicit references on the users' quality assessment of 3D graphics.

2.6.1 Observers screening and data processing

Before starting any analysis, participants were screened using the ITU-R BT.500-13 recommendation [11]. By applying this procedure on the collected scores of Experiment 1, we did not find any outlier participant from group 1 (G1). However, one subject from group 2 (G2) was rejected by reason of reporting implausible scores in the DSIS session (the first session for G2).

After outlier removal, the first step to analyze the results is to compute the mean score for each stimulus. For ACR-HR method, it is advised to compute the difference scores between hidden references and test stimuli instead of using directly the raw rating results. Indeed, studies show that subjects tend to assign a different quality scale for each object [12]: their rating is influenced by their opinion of the content (whether they like or dislike the object). Therefore, assessing differences in quality allows to take into account this variability in the use of the rating scale:

$$d_i^j = s_i^{ref(j)} - s_i^j \quad (2.1)$$

s_i^j refers to the score assigned by observer i to stimulus j . $ref(j)$ is the reference of stimulus j . The difference scores for the reference stimuli ($d_i^{ref(j)} = 0$) are removed from the collected data of the ACR-HR sessions of groups G1 and G2. Finally, we computed

Table 2.3: Experimental details of the tested methodologies.

	Experiment 1		Experiment 2
	<i>ACR-HR</i>	<i>DSIS</i>	<i>SAMVIQ</i>
Explicit reference	No	Yes	Yes
Quality scale	Bad to excellent	Very annoying to imperceptible	Bad to excellent
Scale type	Discrete five-level likert scale	Discrete five-level impairment scale	Continuous quality scale from 0 to 100 (represented by a slider)
Voting	Global quality of test stimuli, including hidden references	Difference between a test stimulus and its reference, simultaneously shown	Global quality of test stimuli, including explicit references
Presentation of the stimulus	Once	Once	Multiple times (random access approach)
Possibility to change the vote	No	No	Yes
Stimulus presentation time	6 sec	10 sec	6 sec
Session duration	18 min	23 min	40-60 min (divided into 2 sessions)
Subjects involved	G1: 15 G2: 15	G1: 15 G2: 15	17
Display	VR headset* (the HTC Vive Pro)	VR headset* (the HTC Vive Pro)	VR headset* (the HTC Vive Pro)

(*) 3D meshes were loaded into the VR scene and rotated in real-time.

the Difference Mean Opinion Score (DMOS) of each stimulus for both groups:

$$DMOS_j = \frac{1}{N} \sum_{i=1}^N d_i^j \quad (2.2)$$

N denotes the remaining subjects after screening observers i.e, $N=15$ for G1 and $N=14$ for G2.

For DSIS, we don't need to compute the DMOS since DSIS is based on the comparison between references and test models. Hence, we can directly use the rating results and compute the Mean Opinion Score (MOS).

$$MOS_j = \frac{1}{N} \sum_{i=1}^N s_i^j \quad (2.3)$$

Moving to quantitative analyzes, statistical tests are affected by the dependencies between samples. In our experiment, two groups of observers (G1 and G2) rated the same stimuli. The only difference between the two groups was the order of the ACR-HR and DSIS sessions. Thus, the raw rating scores of the 2 groups are independent and therefore, we could have used unpaired two-sample t-tests to quantitatively assess whether there are differences between the scores of G1 and G2. However before using a parametric test, it is important to make sure that the data follow a normal distribution. We applied several normality tests, on the rating scores; such as Shapiro-Wilk's test, Lilliefors's test, Anderson-Darling's test. All these tests ascertained that the distribution of our data is not-normal ($p\text{-value} \ll 0.05$). Hence, for our quantitative analyzes, we have opted for the unpaired two-samples Wilcoxon test (also known as Wilcoxon rank-sum test or Mann-Whitney test). It is a non-parametric alternative to the unpaired two-samples t-test.

2.6.2 Resulting MOSs and DMOSs

First, as recommended by VQEG [13], we computed the pairwise (D)MOS correlation coefficient for the 2 groups of subjects (G1 and G2), for each method (ACR-HR and DSIS). The (Pearson, Spearman rank order) correlation coefficient between G1's and G2's (D)MOSs are (0.95, 0.92) for ACR-HR and (0.97, 0.94) for DSIS. Correlation values between subjects of the two groups are relatively high for both methods. Nevertheless, G1 and G2 subjects seem slightly more correlated with the DSIS method.

Additionally, we explored the Intraclass Correlation Coefficient (ICC Type (A,k) coefficients for two-way random effects model [133] that analyzes the absolute agreement among (D)MOSs attributed to the stimuli by the two groups of subjects. The estimated ICC(A,k) for ACR-HR and DSIS are 0.89 and 0.96, respectively.

Obviously, correlation coefficients do not state everything, so we illustrate, in Figure 2.5, the results of ACR-HR and DSIS tests for all stimuli, averaged over all screened observers. To ensure a better readability in interpreting the results, we show the MOSs (instead of the DMOSs) for ACR-HR test. Note that DMOSs are used in all the statistical tests presented in the next sections. We provide, in the appendix (section A.1), the (D)MOSs and their confidence intervals plotted separately for each subject group and for each method. The results are also present as box plots in Figure 2.6.

As expected, MOSs decrease as distortion strengths increase. For the DSIS method (Figure 2.5.a), we can notice a strong consistency between the two groups and a good use of the entire rating scale. Indeed, observers of both groups showed almost the same behavior for each stimulus and their rating scores reached the scale limits.

For the ACR-HR method (Figure 2.5.b), we can notice some differences between the rating scores of the two groups. In fact, observers of G1 tend to downrate the reference stimuli (strength = 0), i.e. the scores given by G2 observers to almost all the references, except the *Chameleon*, exceed those of G1 observers. As a consequence, the used amplitude of the rating scale is reduced (see Figure 2.6.b). The specificity of the *Chameleon* model will be discussed in the next section. Moreover, we note that G2 observers were able to detect some distortions that G1 observers missed, notably the color distortions: e.g. *Aix*, *Ari* and *Samurai* corrupted by *QCol* distortion (row 2) with high strength (strengths ≥ 3) obtained better scores by G1 than by G2.

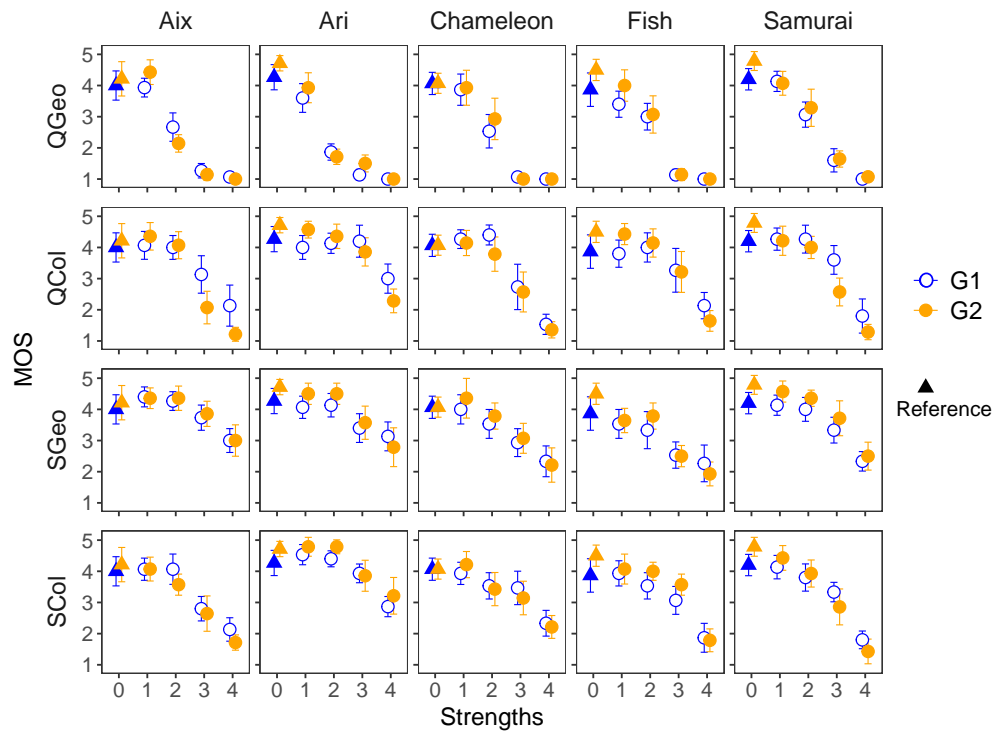
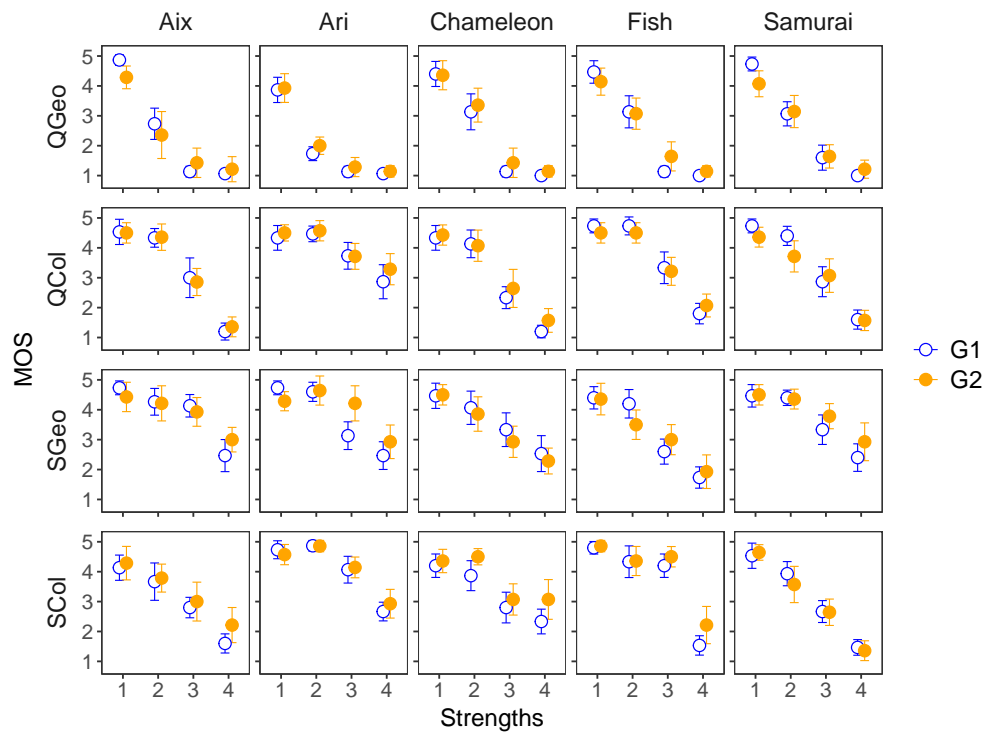


Figure 2.5: Comparison of the G1 and G2 mean scores of the ACR-HR and DSIS tests, for all the stimuli. G1 subjects did the ACR-HR session 1st followed by the DSIS session, while G2 subjects did the DSIS session 1st and then the ACR-HR session. For a given distortion strength, the dots are horizontally spaced apart to avoid overlapping.

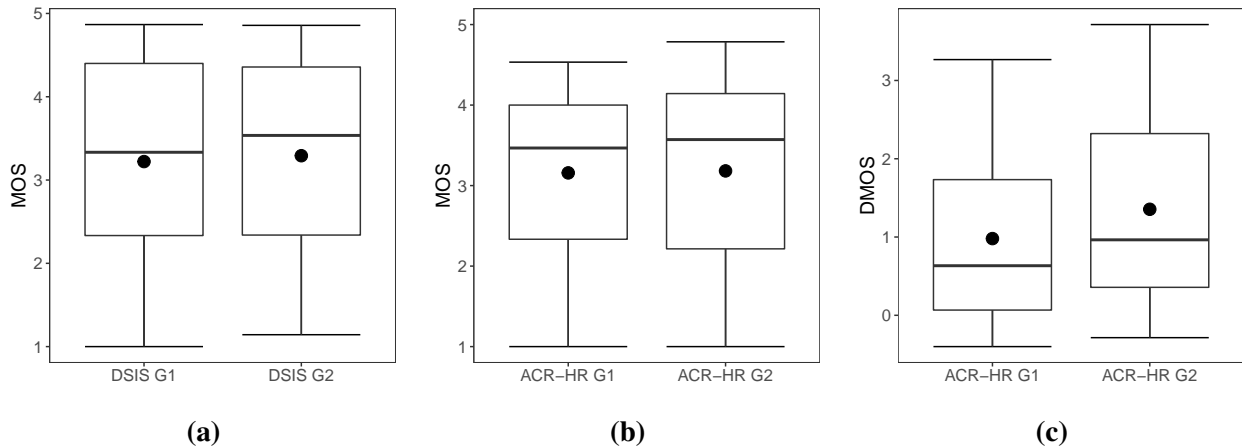


Figure 2.6: Boxplots of MOSs obtained by the two groups of subjects involved in the ACR-HR and DSIS tests.

These first results reveal some differences in the performance of the methodologies. In next sections, we assess whether these differences are statistically significant and we attempt to provide explanations for their causes.

2.6.3 Consistency across subject groups

To assess whether, for a given methodology, there are significant differences in rating scores between the two groups of observers, we conducted for each stimulus the unpaired two-samples Wilcoxon test (see section 2.6.1) on the raw scores s_i^j (for DSIS) and on the differential scores d_i^j (for ACR-HR) of the 2 groups. The null hypothesis (H0) is that, for a given stimulus, the rating scores of G1 observers are equal to those of G2 observers at the 95% confidence level. The alternative hypothesis (H1) is that the scores of G1 are greater (or lesser) than the scores of G2. The p-values are presented in Figure 2.7. The red boxes (p-value < 0.05) indicate that the corresponding stimuli have been rated significantly different by the two groups of subjects.

For the ACR-HR method, we noticed that the scores of the two groups are not consistent (i.e. differ significantly) for 12 stimuli, out of 80; especially for the LAB quantization (*QCol*) of all the models, except the *Chameleon*. This is coherent with the results observed in the previous section (section 2.6.2). Our hypothesis is that this is due to the absence of explicit references. Indeed for G1 observers, as they did the ACR-HR test first, the assessment was absolute. Thus, it was difficult for them to detect the distortions of some models, especially the color impairments for *Samurai* and *Ari*. The reason is that, for statues like *Ari* and *Samurai*, observers have no prior knowledge of the exact color of these models. This is not the case for G2 observers since they had already seen the references during their first session (DSIS session). Hence, they were able to detect the distortions (even the color distortions) that G1 observers might have missed. For the *Chameleon*, there is no significant difference between the scores of the 2 groups. We believe that this is related to the fact that people have strong prior knowledge about this model: the chameleon/iguana is an animal known worldwide and everyone has an idea

		ACR-HR																			
		Aix				Ari				Chameleon				Fish				Samurai			
		1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
QGeo		0.31	0.07	0.34	0.34	0.71	0.03	0.85	0.12	0.83	0.43	0.78	1	0.84	0.27	0.13	0.1	0.03	0.35	0.12	0.03
QCol		0.94	0.66	0.02	0.06	0.96	0.84	0.03	0.005	0.87	0.09	0.79	0.47	0.67	0.19	0.38	0.04	0.12	0.02	0.001	0.003
SGeo		0.47	0.58	0.91	0.65	0.89	0.94	0.63	0.05	0.18	0.48	0.7	0.75	0.18	0.63	0.23	0.14	0.78	0.37	0.8	0.15
SCol		0.46	0.14	0.38	0.06	0.72	0.96	0.35	0.96	0.37	0.98	0.51	0.71	0.28	0.59	0.78	0.25	0.27	0.23	0.02	0.01

		DSIS																			
		Aix				Ari				Chameleon				Fish				Samurai			
		1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
QGeo		0.01	0.15	0.51	0.96	0.96	0.19	0.55	0.54	1	0.57	0.51	0.15	0.27	0.93	0.07	0.15	0.02	0.98	0.79	0.15
QCol		0.66	0.74	0.91	0.37	0.75	0.48	1	0.26	0.88	0.94	0.6	0.15	0.33	0.24	0.77	0.32	0.09	0.05	0.51	0.9
SGeo		0.48	0.96	0.53	0.09	0.04	0.38	0.01	0.18	0.92	0.58	0.33	0.85	0.82	0.05	0.22	0.81	1	0.96	0.17	0.25
SCol		0.44	0.93	0.59	0.14	0.4	0.97	0.98	0.31	0.59	0.08	0.37	0.11	0.71	0.94	0.29	0.09	1	0.48	1	0.44

Figure 2.7: p-values computed between the rating scores of the two subject groups computed for all stimuli and for both methodologies. Red color indicates a significant difference between the scores of G1 and G2.

about its characteristics of shape, color and geometry.

For the DSIS method, we observe, for certain models and distortion types notably the color quantization (*QCol*), a better consistency/agreement among the subjects of the two groups. This confirms the fact that the presence of the reference makes the DSIS methodology more consistent across the subject groups and independent of the sessions order. The absence of explicit reference in the ACR-HR method makes it more difficult for observers to assess certain distortions (e.g. color quantization), especially when they do not have prior knowledge of the models.

2.6.4 Accuracy of quality scores

As stated by Mantiuk et al. [12]: “A more accurate method should reduce randomness in answers, making the pair of compared conditions more distinctive. A more accurate method should result in more pairs of images whose quality can be said to be different under a statistical test.”

To assess the accuracy of the tested methodologies, we thus computed the number of pairs of stimuli rated significantly different by G1 and G2 subjects. For this task, we conducted unpaired two-samples Wilcoxon tests between rating scores of each possible pairs of stimuli. We conducted $80 \times 79 / 2 = 3160$ tests. The α levels used here is 0.05.

In order to study the behavior of this accuracy according to the number of subjects, we repeated these tests for different numbers of subjects and assessed the evolution of the number of pairs of stimuli significantly different. For each number N of subjects, we considered all possible combinations (without repetition) (with $3 \leq N \leq 15$ for G1 and $3 \leq N \leq 14$ for G2) and averaged the number of pairs significantly different over all these combinations of observers. Results are shown in Figure 2.8. The numbers of pairs of stimuli on the y-axis are given in percentages of the total number (i.e., 3160).

From Figure 2.8.a, it can be noticed that, for the DSIS method, the accuracy does not evolve much from G1 to G2. Hence, double stimulus methodology seems, once again,

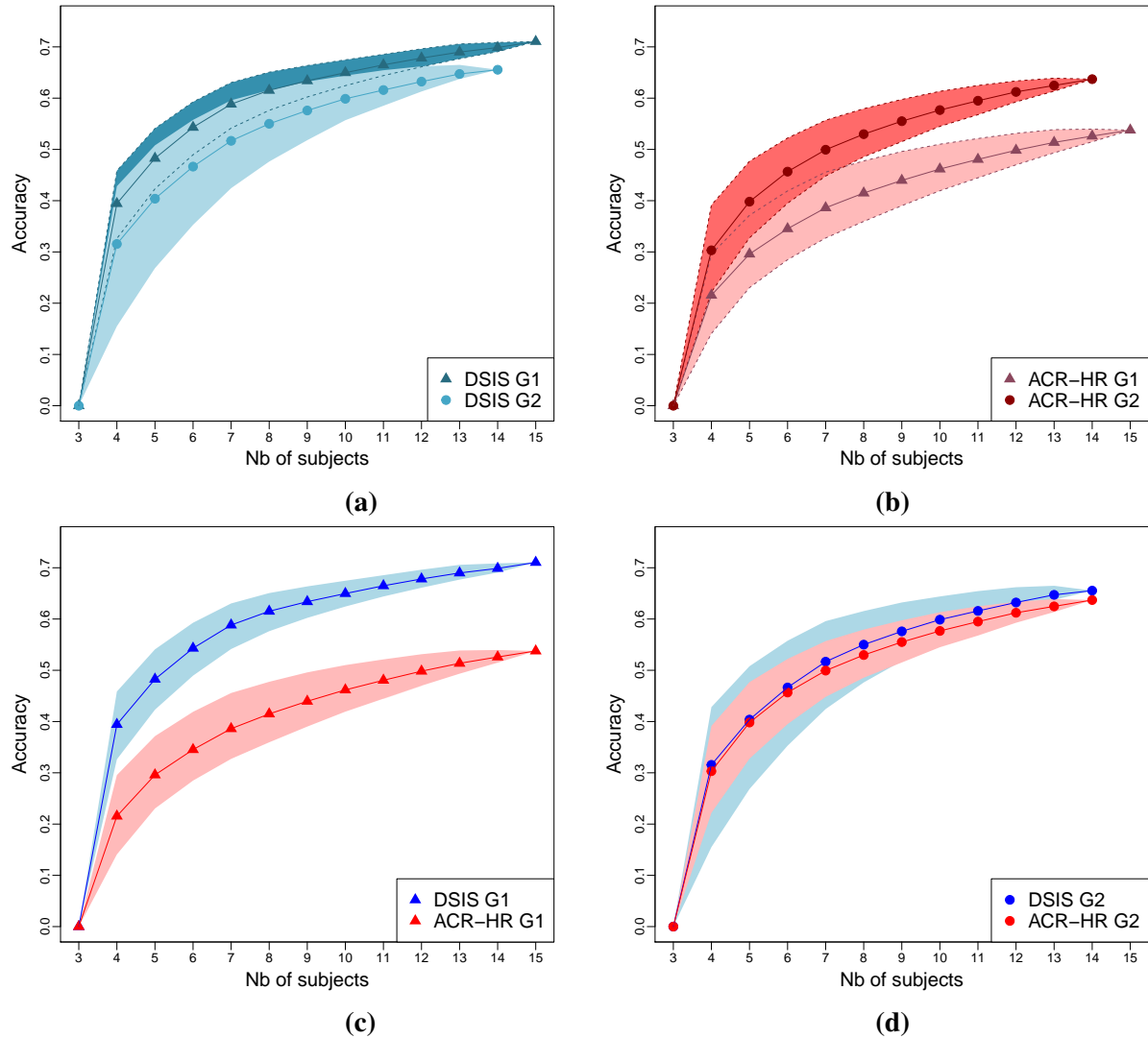


Figure 2.8: Variation of the accuracy according to the number of subjects for both methodologies and both groups (G1 subjects did the ACR-HR session 1st followed by the DSIS session, while G2 subjects did the DSIS session 1st and then the ACR-HR session). The accuracy (y-axis) is defined as the percentage of pairs of stimuli whose qualities were assessed as statistically different. Curves represent mean values of these percentages and areas around curves represent 2.5th - 97.5th percentiles.

stable and independent of the sessions order. However, this is not the case of the ACR-HR method since the accuracy undergoes a large increase for G2 compared to G1 (Figure 2.8.b). This demonstrates anew that the method without explicit reference is not consistent across the subject groups. G2 subjects - who completed the ACR-HR test in the 2nd session - were more familiar with the stimuli than G1 subjects since they had already seen the models and their references in the 1st session (the DSIS test). Therefore, they were capable of distinguishing/detecting the degradations/loss in the visual quality of the stimuli more easily than G1 observers.

Beyond this better consistency observed for DSIS, Figures 2.8.c and 2.8.d clearly show that the DSIS method is more accurate than the ACR-HR method. This is valid even for G2,

for which the ACR-HR test was conducted after the DSIS one. This finding corroborates previous results by Kawano et al. [51] obtained for stereoscopic 3D Videos and Alexiou et al. [127] obtained for point clouds. However, this result is inconsistent with comparative studies conducted with images and videos, including omnidirectional videos [53], in which Modified ACR (M-ACR) was slightly more reliable than DSIS. Our hypothesis is that people have more prior knowledge about the quality of 2D (natural) images/videos than that of 3D graphics, and therefore the presence of references does not seem necessary to assess the quality of these data.

The accuracy of a method is also related to the agreement between raters, as an accurate method is intended to reduce random scores. Thus, we assessed, using the ICC (type (A,k) coefficients for two-way random effects models), the inter-rater reliability of each group of observers in the DSIS and ACR-HR tests: i.e. for a given method, the scores of a group of raters were compared to each other to evaluate how close the raters were in terms of their scores. Results, shown in Figure 2.9, denote that the degree of agreement among raters is higher in DSIS tests than in ACR-HR tests. Moreover, subjects' agreement increased during the second session for both methods, yet this increase is larger for ACR-HR. Moreover, we assessed the intra-rater reliability among the ACR-HR and DSIS tests for the two groups of subjects: i.e. we evaluated, for each participant, the degree of consistency between their scores in the 2 tests. Results, reported in 2.10 show that the consistency between DSIS and ACR scores is higher for G2 observers than for G1 observers. This is coherent with the results observed in Figures 2.8.c and 2.8.d. For a better understating of boxplots, such as those in Figure 2.10, the reader can refer to Figure A.1 in the appendix.

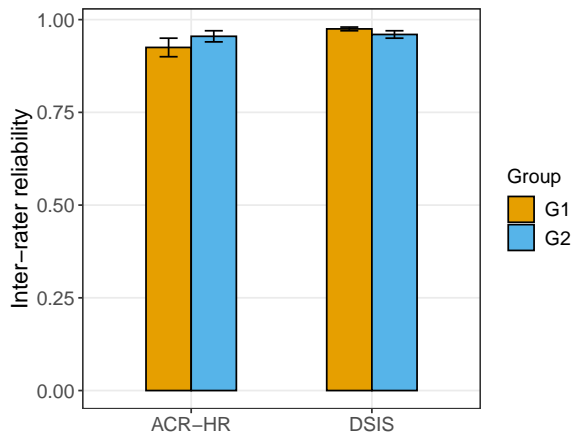


Figure 2.9: Inter-rater reliability, of each group, in the DSIS and ACR-HR tests.

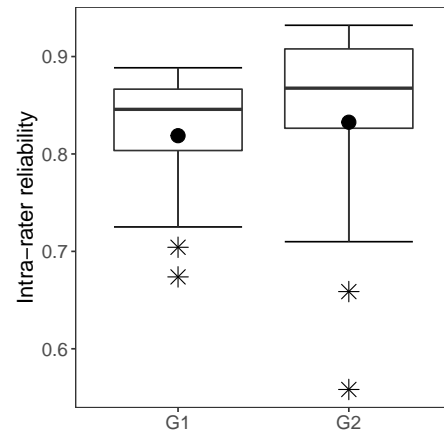


Figure 2.10: Intra-rater reliability among the ACR-HR and DSIS tests for the 2 groups.

Lastly, we determined the number of subjects required in an ACR-HR test to obtain the same accuracy as a DSIS test. We compared the accuracy of the ACR-HR test of G1 to that of DSIS test of G2, since in these tests (the first sessions of the experiment for each group) the models were unknown to the participants. As can be seen in Figure 2.11, ACR-

HR requires almost twice as many subjects as DSIS. For instance, for a given number of observers unfamiliar with the test stimuli, ACR-HR requires minimum 15 observers to get a discrimination with an overall level of 54% while DSIS requires only 8 observers.

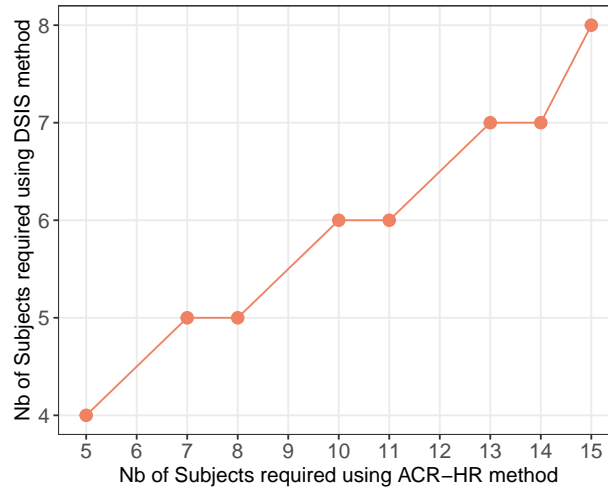


Figure 2.11: Number of observers required in an ACR-HR test to achieve the same accuracy as a DSIS test.

2.6.5 Confidence intervals

Another way to evaluate the accuracy of the methodologies is to compute the 95% confidence intervals of the obtained (D)MOSs. We thus computed these 95% confidence intervals (CIs) for both groups and methodologies, in order to determine the “true” mean score (i.e. the interval in which the (D)MOSs will reside if we have an ∞ number of observers) [11]. Then, we assessed the evolution of the CIs width for both methodologies according to the number of subjects.

The curves of Figure 2.13 were obtained by averaging the CIs width over all the possible combinations of subjects. Note that for a given source model and type of distortion, we averaged the CIs widths over the four strengths of the distortion. We can observe that the width of CIs increases as the sample size decreases. For G1 (see Figure 2.13.a), we notice that, for most stimuli, the CIs of the ACR-HR experiment are much larger than those obtained by the DSIS experiment, implying a strong dispersion of G1 scores in the ACR-HR test. This disagreement/dispersion is due to the fact that source models were unknown for G1 subjects. This disagreement is not so apparent for G2 where the CI widths given by the ACR-HR method are closer to those given by the DSIS method (see Figure 2.13.b). These results confirm once again that DSIS is more accurate than ACR-HR, regardless the group.

We illustrate in Figures 2.13.c and 2.13.d that there is almost no difference between CIs of G1 and G2 involved in DSIS, while for ACR-HR, CIs of G1 are mostly superior to those of G2, except for the *Chameleon* model. As explained in section 2.6.3, the observers’ strong prior knowledge of the characteristics of this animal increases their accuracy, even without the presence of the explicit reference.

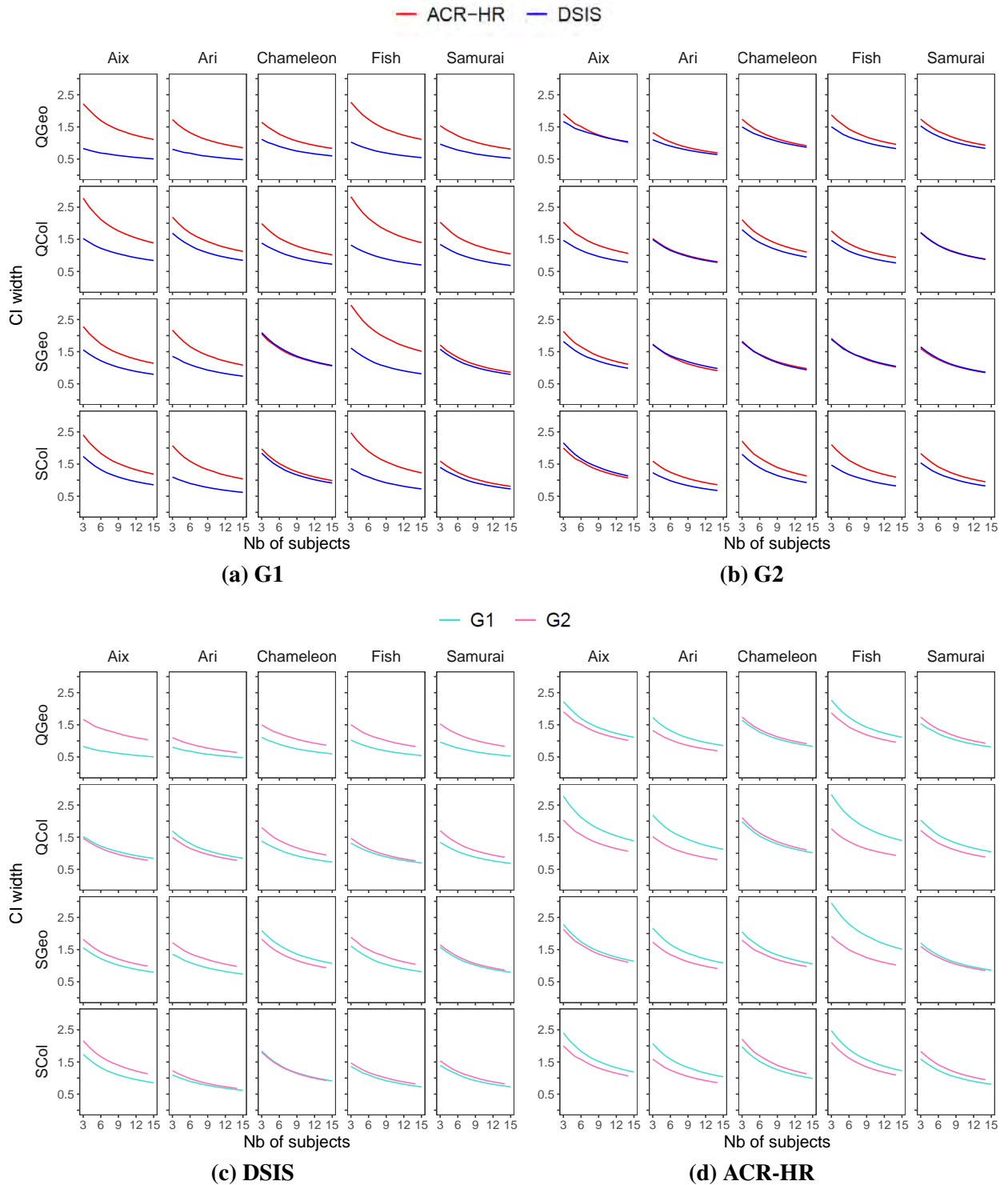


Figure 2.13: Evolution of confidence interval (CI) widths for the ACR-HR and DSIS methodologies as a function of the number of observers involved in G1 and G2. G1 subjects did the ACR-HR session 1st followed by the DSIS session, while G2 subjects did the DSIS session 1st and then the ACR-HR session.

2.7 Results of Experiment 2 and methods comparison

According to the results of Experiment 1, the presence of an explicit reference seems to be a necessity to improve not only the accuracy of the method but also to obtain lower confidence intervals. In this section, we analyze the results of Experiment 2 (the SAMVIQ test). We compare the performance of the SAMVIQ method with that of ACR-HR and DSIS in terms of accuracy and time-effort, in order to find the best (the most suitable) methodology to adopt for subjective quality assessment of 3D graphics.

The following analyzes and comparisons were carried out using the ACR-HR scores of G1 and the DSIS scores of G2. We chose these scores because observers of G1 and G2 first performed the ACR-HR and DSIS sessions, respectively and therefore the models were unknown for these subjects, as for the subjects of the SAMVIQ experiment.

2.7.1 Observers screening and data processing

The SAMVIQ screening procedure differs from that described in Recommendation ITU-R BT.500-13. The SAMVIQ rejection criteria is based on a correlation of individual scores against corresponding mean scores from all the observers [9]. By applying this procedure, we found 2 outliers. Hence, only the scores of the remaining 15 subjects will be used in our subsequent analyzes.

In order to facilitate the comparison between the ACR-HR, DSIS and SAMVIQ methods, we converted the SAMVIQ ratings (ranged from 0 to 100) to the scale of the ACR-HR and DSIS methods (1 to 5 scale) as proposed in [134,135]:

$$S'_{1-5} = \frac{S_{0-100} - 10}{20} + 1 \quad (2.4)$$

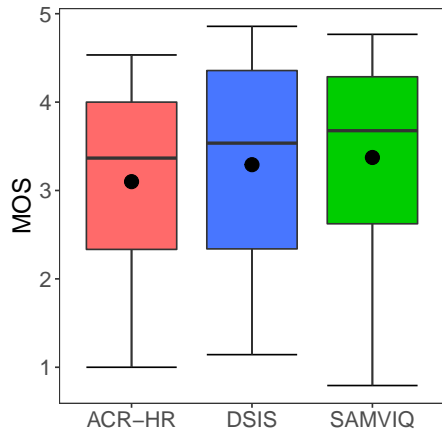
This mapping is done so that the labels on the SAMVIQ scale align with those of the ACR-HR scale (see Figure 2.1).

For quantitative analyzes, we assessed, for ACR-HR and SAMVIQ, the difference scores between references and test stimuli instead of using directly the raw scores (as explained in section 2.6.1).

2.7.2 Resulting MOSs

In this section, we compare the ACR-HR, DSIS and SAMVIQ MOSs for all the stimuli. Figures 2.14 and 2.15 show the results. Note that, for ACR-HR and SAMVIQ, we present the MOSs (instead of the DMOSs) for a better legibility of the results. We provide, in the appendix (Figures A.3.b, A.4.a and A.5), the CIs associated with the (D)MOSs of the 3 methodologies separated.

The 3 methodologies show almost the same behavior. Indeed, the pairwise (D)MOS correlation analysis indicated that the correlations of (D)MOS between pairs of tested methods are high, similarly to what was obtained by Tominaga et al. [50]. Table 2.4 summarizes



	DSIS-ACR-HR	SAMVIQ-ACR-HR	SAMVIQ-DSIS
Pearson correlation	0.943	0.937	0.946
Spearman rank order correlation	0.885	0.883	0.906

Table 2.4: (D)MOS Correlation matrices for ACR-HR, DSIS, and SAMVIQ.

Figure 2.14: Boxplots of MOSs obtained for the tested methods.

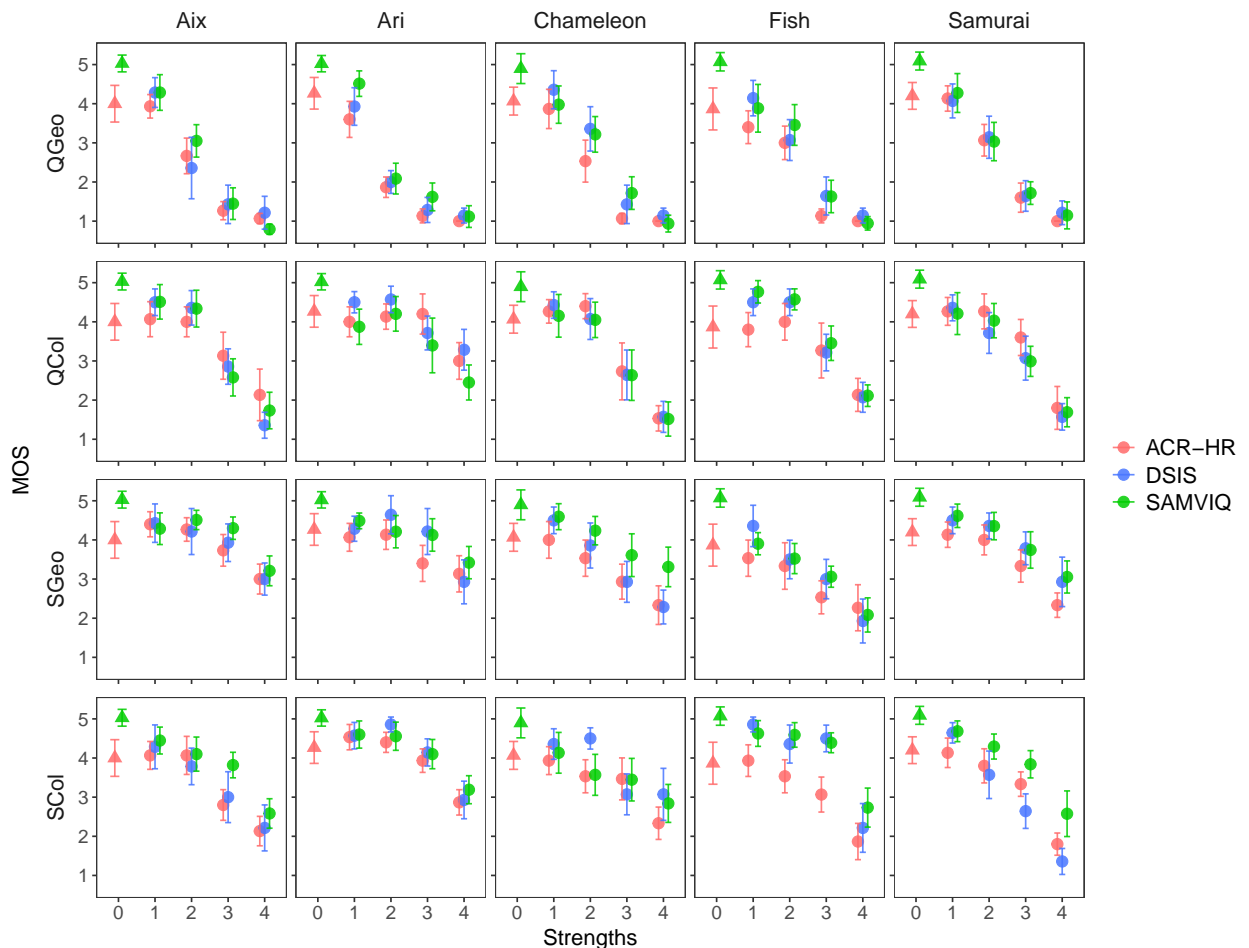


Figure 2.15: Comparison of the mean scores of the ACR-HR, DSIS, and SAMVIQ tests, for all the stimuli. For a given distortion strength, the dots are horizontally spaced apart to avoid overlapping.

the results. Nevertheless, as can be seen in Table 2.4 and Figure 2.15, the results of SAMVIQ seem slightly more correlated with those of DSIS than with those of ACR-HR: there is better consistency between the subjects of the SAMVIQ test and those of the DSIS test, notably for *Aix* and *Fish* color quantized with low strengths ($QCol$, strengths ≤ 2), *Samurai* geometrically simplified ($SGeo$), and *Fish* color simplified ($SCol$). Concerning the scores attributed to the reference models (strength = 0), SAMVIQ does not seem to downgrade them, like ACR-HR does, since references are explicit in SAMVIQ. Moreover, Figures 2.15 and 2.14, show a difference in the use of the quality scale of each method. For a continuous scale (i.e. the SAMVIQ scale), subjects tend to avoid extremities and thus tend to use a smaller range of values, overall. This is known as the “Saturation Effect” [135,136]. This effect is less visible for DSIS, since it uses a discrete categorical scale: no possible variations around best and worst qualities.

2.7.3 Accuracy and time-effort

First, we study the inter-rater reliability of SAMVIQ method. The degree of agreement among SAMVIQ raters is almost the same as that of DSIS raters ($ICC(A,k)=0.96$), and higher than that of ACR-HR raters ($ICC(A,k)=0.93$). We then investigated the accuracy (defined in section 2.6.4) of SAMVIQ method and compare it to that of DSIS and ACR-HR: we computed the percentage of pairs of stimuli rated significantly different among all possible stimuli pairs and assessed its evolution according to the number of subjects. Note that, the non-transformed ratings of SAMVIQ were used for SAMVIQ to compute its accuracy (no same-scale mapping required). Figure 2.16.a shows the results.

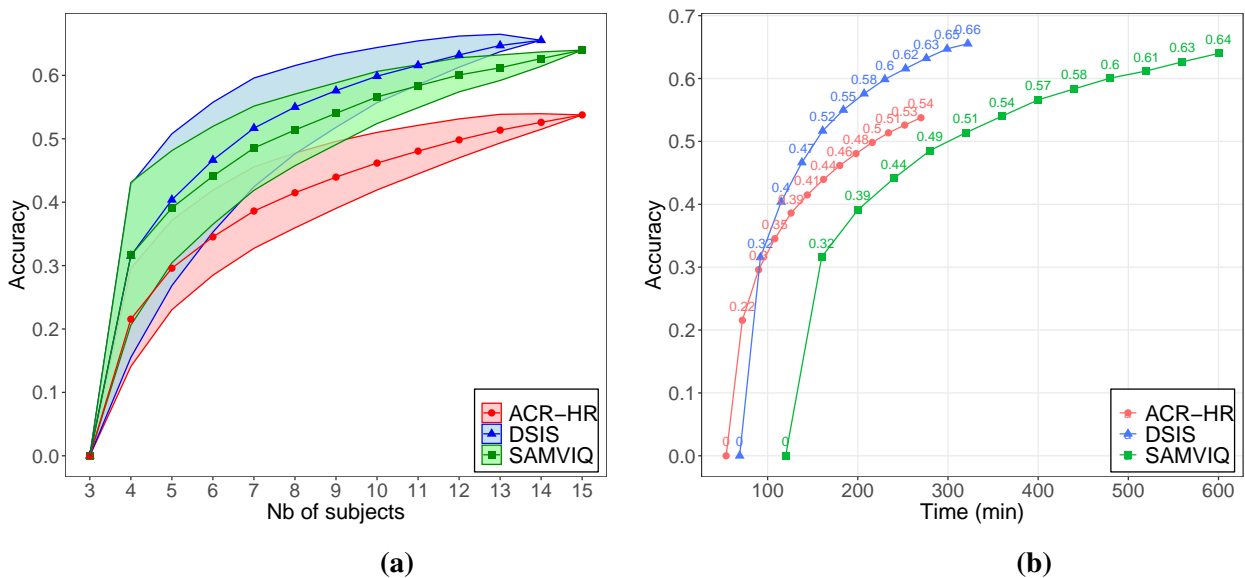


Figure 2.16: Variation of the accuracy according to the number of subjects (a) and time-effort (b) for the tested methodologies. The accuracy (y-axis) is defined as the percentage of pairs of stimuli whose qualities were assessed as statistically different. Curves represent mean values of these percentages and areas around curves represent 2.5th - 97.5th percentiles.

The accuracy of the ACR-HR is smaller than that of SAMVIQ and DSIS. DSIS is slightly

more accurate than SAMVIQ. Our hypothesis is that the detection of visual quality losses of stimuli is easier and more obvious with the DSIS method than with the SAMVIQ method. In fact, we believe that the task of DSIS is simpler (more straightforward) than that of SAMVIQ: the reference and the test stimulus are simultaneously displayed side by side in the scene and the subject is clearly asked to assess the impairments compared to the reference. However, SAMVIQ tends to be more complex since it uses a multi-stimuli with random access approach. Tominaga et al. [50] assessed the ease of evaluation of different methods and found that SAMVIQ, which has many grades on its quality scale, is more difficult than ACR-HR and DSIS.

To determinate the best methodology in subjective quality assessment tests, it is important to consider not only the accuracy of the methods, but also the time that observers need to complete the experiment. Ultimately, even less accurate methods may lead to smaller confident intervals (higher discrimination ability) if more data are collected [12]. Indeed, subjects may have difficulty maintaining their attentiveness throughout a long experiment because of fatigue and boredom. This could skew the results of the experience. Thus, we compared the time-effort of each methodology: we determined the required time for these methods to reach a certain accuracy level. To do so, we multiplied the number of observers, used in abscissa of Figure 2.16.a, by the total time of each test session required to assess the whole dataset (80 distorted models + 5 references): 18 min for the ACR-HR session, 23 min for the DSIS session and 40 min for the SAMVIQ sessions. Note that for SAMVIQ, time may vary depending on how many times the subject viewed the stimuli. We considered the fastest scenario (≈ 20 min to assess 45 objects). Results are presented in Figure 2.16.b. DSIS is the most time-efficient method. SAMVIQ takes almost twice as long as DSIS to achieve the same accuracy: SAMVIQ requires minimum 600 min to get a discriminative power of 65% while DSIS requires only 300 min. Thus, SAMVIQ is considerably more time-consuming than DSIS (and ACR-HR).

2.7.4 Confidence intervals

In this section, we evaluate the results of the subjective methodologies tested in terms of the dispersion of individual ratings (the standard deviation of subjective scores). Thus, we compared the 95% CIs of the MOSs (for DSIS) and the DMOSs (for ACR-HR and SAMVIQ) among the methods.

To do this, we normalized the (D)MOSs values, as in [137], so that 0 means the lowest quality and 1 means the highest quality:

$$nMOS_i = \frac{MOS_i - \min\{MOS_1 \dots MOS_N\}}{\max\{MOS_1 \dots MOS_N\} - \min\{MOS_1 \dots MOS_N\}} \quad (2.5)$$

$$nDMOS_i = \frac{DMOS_i - \max\{DMOS_1 \dots DMOS_N\}}{\min\{DMOS_1 \dots DMOS_N\} - \max\{DMOS_1 \dots DMOS_N\}} \quad (2.6)$$

where i is the index of the stimulus and N is the total number of stimuli.

We also normalized the CIs by expressing them as a percentage of the scale range. Figure

2.17 shows the boxplots of CIs, as well as the CIs in relation to their (D)MOSs, for the 3 methodologies tested.

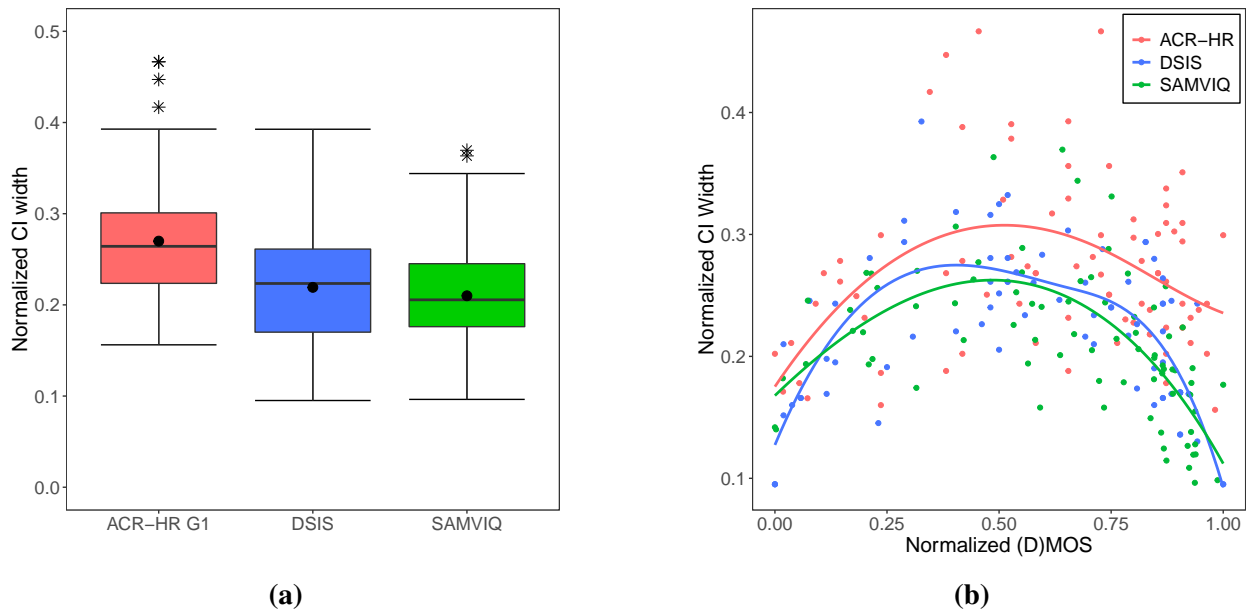


Figure 2.17: (a) Boxplots of CIs obtained for the tested methods, (b) Comparison of normalized CIs of the tested methods as a function of normalized (D)MOSs.

The CIs of ACR-HR are significantly larger than those of the other methods, implying that ACR-HR has strong dispersion between the scores of the observers. These results are consistent with those presented in section 2.6.5. It is interesting to notice that, overall, SAMVIQ CIs are smaller than those of DSIS. Still, this difference is slight, since we performed the t-test with a significance level of 5% between the CIs of DSIS and SAMVIQ and found no significant difference between the CIs of these two methods. We believe that SAMVIQ provides smaller CIs due to the subject’s ability to review stimuli and adjust scores. Note that, despite the slightly smaller CIs of SAMVIQ, DSIS produced more accurate results (see section 2.7.3), because the amplitude/range of the SAMVIQ rating scale actually used by the subjects is reduced/limited compared to that of DSIS.(i.e. the subjects did not use the whole scale in the SAMVIQ test) (see Figures 2.15 and 2.14).

We can observe, in Figure 2.17.b, that the CIs of SAMVIQ tend to be larger than those of DSIS on the extreme values of MOSs ($n(D)MOS \approx 0$ or 1). This is due to the fact that for DSIS (5-level discrete scale), there is no possible variations around best and worst qualities. However, for SAMVIQ (continuous scale), subjects tend to avoid extremities, since the choice of worst and top scores is not limited to 0 and 100 only (“Saturation Effect” presented in section 2.7.2). We can also notice that ACR-HR CIs approach those of SAMVIQ and DSIS for the worst MOS values. However for the high MOS values, the dispersion of ACR-HR scores remains high because the observers have not seen the references and therefore have no prior knowledge of the best possible quality of stimuli.

Finally, we assessed the evolution of the confidence intervals according to the number

of subjects, for the 3 methodologies, as described in section 2.6.5. Results, presented in Figure 2.18, highlight the findings of this section and section 2.7.3.

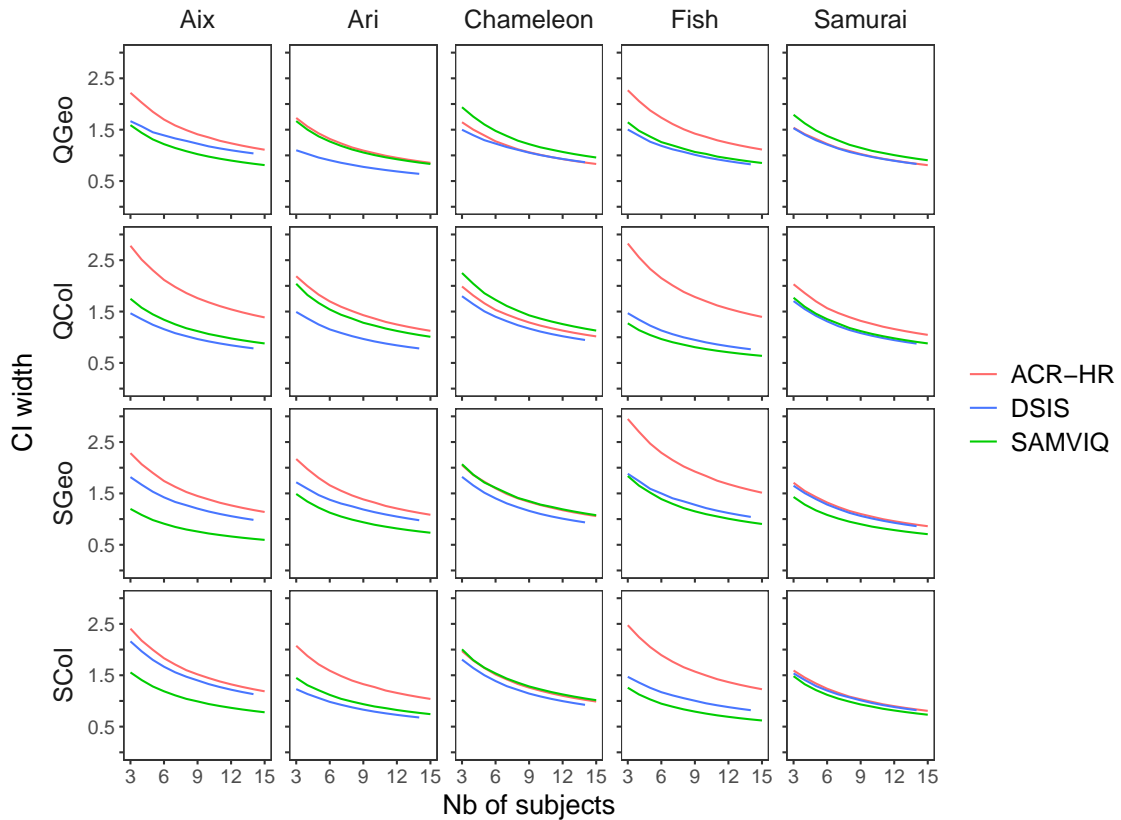


Figure 2.18: Evolution of CI widths as a function of the number of observers for the tested methodologies

2.8 Discussion and recommendations

This section summarizes the results obtained in this chapter. We found that, for the quality assessment of 3D graphics, the ACR-HR method has a poor accuracy and large CIs compared to those of DSIS and SAMVIQ, and thus requires more subjects. In fact, in ACR-HR, the assessment is absolute (absence of explicit references) and therefore observers, who had never seen the reference models before, are not able to detect all distortions, especially color impairments. Thus, they tend to be less discriminating than those who are familiar with the test stimuli. This is not the case for the DSIS and SAMVIQ methods since they present explicit references. These two methods showed almost the same performance in terms of accuracy and agreement among individual ratings (CIs). DSIS appears to be slightly more accurate, while SAMVIQ offers slightly less dispersion in subjective ratings. In regards to the time-effort, DSIS shows a great advantage. It is the most time-efficient, whereas SAMVIQ is considerably the most time-consuming: SAMVIQ takes twice as long as DSIS to achieve the same accuracy. Furthermore, the observers' task in SAMVIQ experiment is more difficult than that of DSIS (and ACR-HR).

Based on our results, we recommend the use of DSIS for the quality assessment of 3D graphics. We have also attempted to make recommendations about the required number of observers for this methodology. For this purpose, we aggregated the DSIS test's scores of the 2 groups of subjects (G1 and G2) involved in Experiment 1 and thus obtained 30 subjects. This aggregation is possible since we previously demonstrated that DSIS scores are consistent among the two groups. We recomputed the accuracy (as in sections 2.6.4 and 2.7.3) according to the number of observers.

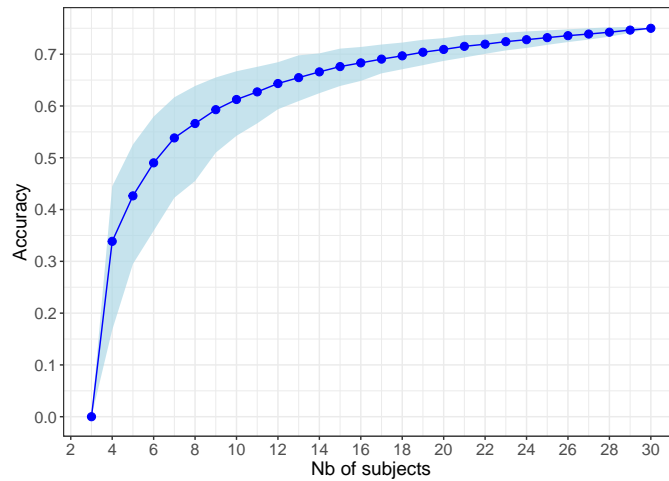


Figure 2.19: Accuracy of the DSIS method according to the number of subjects.

From Figure 2.19, we observe that at least 19 test subjects are required to be able to discriminate 70% of all possible pairs of stimuli. With 15 observers, the recommended number by the ITU-R BT.500-13 [11], we obtain an accuracy of 67%. However, with 25 subjects the discrimination increases to 73% and reaches 75% with 30 subjects. As a conclusion, and with regard to the shape of the curve, 24 subjects seem to be a good compromise.

2.9 Conclusion

In this Chapter, we designed two psycho-visual experiments that compare the performance of three of the most prominent subjective methodologies, with and without explicit references (ACR-HR, DSIS and SAMVIQ), for assessing the quality of 3D graphics in a VR environment. Results assert that the presence of an explicit reference is necessary to improve the accuracy and the stability of the method. This conclusion is not consistent with recent comparative studies conducted with images and videos. We believe that this is due to the fact that people have less prior knowledge about 3D graphics quality than about that of (natural) images.

DSIS seems to be the most suitable method to assess the quality of 3D graphics. It is the most accurate and mainly the most time-efficient. We recommend to use groups of at least 24 observers for the DSIS methodology. The only data representation used in this work is 3D meshes, however, we believe that our results remain valid for other 3D representations, such as point clouds.

Chapter 3

Subjective Quality Assessment of 3D Meshes with Vertex Colors in Virtual Reality

Subjective quality assessment experiments, and thus the resulting datasets, are of primary importance for assessing the Quality of user Experience (QoE), understanding human behavior in evaluating perceived quality, benchmarking and training objective metrics. Unfortunately, for works involving 3D meshes with color information (either in the form of texture or vertex-colors), only few have publicly released their datasets [1,3] (see section 1.1.2 of Chapter 1). Therefore, for this type of data, there is a lack of both subjective datasets and objective metrics, resulting in a lack of insight into how color and geometry distortions affect quality. Another factor that has not yet been explored (since, as discussed previously in section 1.1.2, almost all subjective experiments were conducted on screen), and which is relevant in the case of 6 Degrees of Freedom (DoF) interactions in immersive environments, is how the viewpoint and movement of 3D models affect their perceived quality.

In this chapter, we address the problem of subjective quality assessment of 3D meshes with vertex colors. To this end, we conducted a subjective experiment in a Virtual Reality (VR) environment that involved 480 animated colored meshes. The stimuli were displayed in 3 different viewpoints that we animated with 2 short movements. A total of 11520 quality judgments (24 score per stimulus) were collected. The resulting dataset allowed us to analyze the factors that influence the perceived quality of 3D graphics: we evaluate not only the visual impact of color and geometry distortions on the appearance of such data, but also the impact of source models, animations and viewpoints.

The key contributions of this work can be summarized as:

- We provide the community with a ground truth dataset of 480 meshes with vertex colors, each rated by 24 subjects. This is the largest dataset for this kind of data, and the first based on vertex color representation. It is also the first public dataset¹ produced in VR for colored 3D content.

¹<https://yananehme.github.io/datasets/>

- We provide an in-depth analysis of the effects of source models, distortions, viewpoints and movements on both quality scores and their confidence intervals. Our findings provide insights for the design of subjective studies and objective metrics for 3D content.

This chapter is organized as follows: we begin by describing our dataset and how we generated it (section 3.1). Then, in sections 3.2 and 3.3, we detail the subjective experiment. Section 3.4 provides the results, while sections 3.5 and 3.6 present concluding remarks and recommendations.

3.1 Dataset generation

To build our dataset, we extended the dataset used in Chapter 2 (section 2.2). This dataset contains 80 meshes with vertex color information generated from 5 source models (“Aix”, “Ari”, “Chameleon”, “Fish”, “Samurai”) and corrupted by 4 types of geometry and color distortion that represent common simplification and compression operations: uniform geometric quantization (QGeo), uniform LAB color quantization (QCol), simplifications that take into account either the geometry only (SGeo) or both geometry and color (SCol). Each distortion was applied with 4 different strengths that cover the whole range of visual quality from imperceptible to high levels of impairment. Full details on model characteristics and distortion parameters are provided in Tables 2.1 and 2.2 of Chapter 2.

As explained in Chapter 2 (section 2.3.1), in order to adequately assess the visual quality of 3D content, it is important that the objects move so that observers can see the dynamic effects of shading on the shape. Moreover, it is important for the observers to see the whole object and not to focus on one single viewpoint. Therefore, we selected, for each model, 3 viewpoints that we animated with 2 short movements. These 6 combinations of viewpoints and movements can be considered to be the hypothetical rendering trajectories (HRTs). HRTs is a concept introduced in [138] for free-viewpoint videos which represents the dimension of the object under test related to the interactivity part such as the camera configurations, viewpoints and trajectories.

For our experiment, we have perceptually chosen and adjusted the viewpoints of each model, so that *viewpoint 1* represents the one which contains the most geometry, color and semantic information. *Viewpoint 2* and *viewpoint 3* cover the remaining semantically relevant parts of the model. Figure 3.1 illustrates the selected viewpoints of each source model, while Figure 3.2 shows some visual examples of distortions from these viewpoints. For each viewpoint of a given stimulus, we applied 2 types of animation:

- Slow rotation (R) of 15 degrees around the vertical axis in a clockwise and then in a counterclockwise direction.
- Slow zoom in (Z) of 0.75 meters, followed by a zoom out.

As evoked in the introduction of this chapter, the generation of different stimulus orientations and animations will allow us to explore later the impact of animations and viewpoints



Figure 3.1: Illustration of the 3D graphic source models and their selected viewpoints, respectively. Acronyms refer to Model_Viewpoint. The “*Dancing Drummer*” model is used only in training.

(HRTs) on the perceived quality of 3D objects. Note that, the animations we generate do not involve non-rigid transformations of the objects.

Our dataset thus contains 480 dynamic stimuli: 5 source models \times 4 distortion types \times 4 strengths \times 3 viewpoints \times 2 animation types.

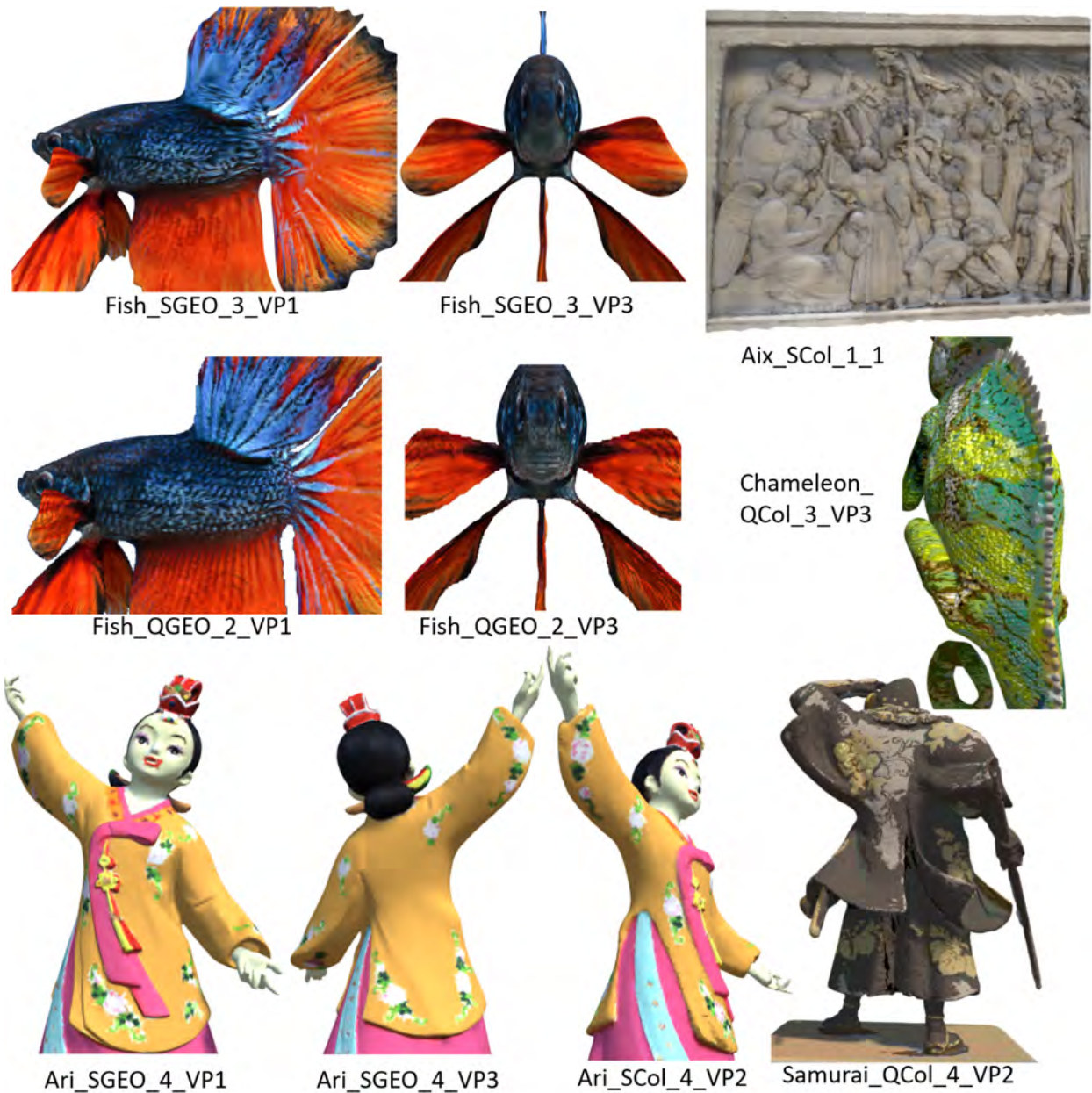


Figure 3.2: Some examples of distorted models displayed in the selected viewpoints. Acronyms refer to Model_Dist-Type_Dist-Strength_Viewpoint.

3.2 Experimental environment and apparatus

The objective of our experiment is to produce a reliable ground truth of subjective opinions for our dataset of 480 stimuli. So, we selected the Double Stimulus Impairment Scale (DSIS) methodology, as the subjective rating method: the observer sees the source model (a.k.a. reference model) and the same model impaired, simultaneously, side by side, for 10s and rates the impairment of the second stimulus in relation to the reference using a five-level impairment scale [10]. As demonstrated in Chapter 2, this method is most stable and most accurate for assessing the quality of 3D graphics (since it presents an explicit reference).

We chose to conduct the experiment in VR using the HTC Vive Pro headset, since VR offer the most ecological and realistic use cases for such data and are in high demand. We opted for the same VR experimental environment as the DSIS test of Chapter 2, illustrated in Figure 2.4 (2nd row). The reference and the distorted model were rendered in a virtual scene, side by side, at a viewing distance fixed to 3 meters from the observer, under a given viewpoint and type of animation. Note that these 2 dynamic stimuli were specifically oriented in order to show exactly the same vertices of the 2 models at the same time. Their size is approximately 37 degrees of visual angle. Their material type complies with the Lambertian reflectance model. The stimuli are visualized in a neutral virtual room (light gray walls) without shadows and under a directional light. The rating billboard was integrated in the virtual environment of the experiment and was displayed after the presentation of each pair of stimuli. There was no time limit to vote and the stimuli were not shown during that time. To vote, the subject selected and saved the score using the trigger of the HTC Vive controller.

3.3 Participants and training

3.3.1 Training

We started the experiment with a training. We proceeded as the training described in the previous Chapter (section 2.5.1). We selected a training model not included in our original dataset: “Dancing Drummer” (Figure 3.1) and generated 11 distorted models that span the whole range of distortions. Each training stimulus is displayed for 10s, then the rating panel is displayed for 5s. An example score assigned to this distortion is highlighted. We added a practice trial stage at the end of the training, in which we displayed 2 extra stimuli and asked the participant to rate their impairment.

3.3.2 Creation of test sessions

In order to maintain a sufficient level of attention, we decided to limit the number of stimuli rated per participant to 160 stimuli out of 480. Our objective is to select 160 objects per participant in the most relevant way and without producing bias in the results. So, we decided to show each participant all the source models corrupted by all the distortion types and levels. Each model will be displayed under one viewpoint in both rotation and zoom animations.

According to the recommendations of Chapter 2 (section 2.8) about the required number of observers for assessing the quality of 3D data using the DSIS method, each stimulus must be rated by at least 24 observers. Thus, a minimum of 72 participants ($480 \times 24 / 160$) is required.

With this in mind, we have developed an algorithm that creates a set of 72 batches of 160 stimuli each respecting the constraints related to: (1) the selection of stimuli for each participant, i.e. each batch must contain 5 source models \times 4 distortion types \times 4 strengths

$\times 1$ viewpoint $\times 2$ animations; and (2) the minimal number of subjective scores required per stimulus, i.e. each stimulus must be rated 24 times (= present in 24 batches).

3.3.3 Participants

A total of 72 participants took part in the experiment and they were remunerated. Participants were aged between 18 and 55. The majority were students from the University of Nantes, University of Lyon and LIRIS laboratory, while the rest were workers and professionals in different occupations. 49 males and 23 females, 45 of whom had already tried (or were familiar with) a VR headset, they were naive about the purpose of the experiments. All observers had a normal or corrected to normal vision.

3.3.4 Duration

To avoid fatigue, boredom and cyber sickness, we divided the 160 stimuli into 2 sessions of 23 min each (informed consent/instructions + 11 training stimuli \times (10s display + 5s rating) + 80 test stimuli \times (10s display + ~ 4 s rating)). None of these sessions took place on the same day in order to prevent any learning effect between stimuli. Thus, these two sessions were held at least two days apart. The stimuli were displayed in a random order (source models, distortion types and levels and animations all mixed) to each participant. Each stimulus was presented once. Participants were not able to replay the stimuli.

3.4 Results and analyzes

This section analyzes and discusses the results of our subjective experiment. First, we evaluate the agreement between the subjects. We also study their bias and inconsistency during the test. Then, we analyze the impact of main factors such as the source models, the viewpoints and the animations on the obtained opinion scores and their accuracy.

3.4.1 Observers screening and data processing

Before starting any analysis, participants were screened using the ITU-R BT.500-13 recommendation [11]. Applying this procedure on our data, we did not find any outlier participants.

A common way to analyze the opinion scores of a DSIS test is to compute the Mean Opinion Score (MOS) of each stimulus.

$$MOS_e = \frac{1}{N} \sum_{s=1}^N X_{e,s} \quad (3.1)$$

$X_{e,s}$ refers to the raw score assigned by subject s to the stimulus e . N denotes the total number of subjects.

To better understand the influence of subject and source variability on the opinion scores, we used the recovery model based on Maximum Likelihood Estimation (MLE) recently introduced by Li et al. [2]. This approach recovers subjective quality scores from noisy raw measurements, by jointly estimating the subjective quality of impaired stimuli (true score), the bias and inconsistency of subjects, and the ambiguity of the visual content all together.

$$X_{e,s} = x_e + B_{e,s} + A_{e,s} \quad (3.2)$$

$$B_{e,s} \sim N(b_s, v_s^2) \quad (3.3)$$

$$A_{e,s} \sim N(0, a_c^2) \quad (3.4)$$

$X_{e,s}$ is the raw opinion score. x_e is the (true) quality score of the stimulus e . $B_{e,s}$ is the noise factor of subject s on rating stimulus e , it follows a Gaussian distribution in which the mean b_s represents the subject's bias, and the variance v_s^2 represents the subject's inconsistency. The factor $A_{e,s}$ refers to the source c that corresponds to the stimulus e . Its parameter a_c^2 represents the ambiguity related to c . The estimation of each parameter (x_e , b_s , v_s , a_c) is associated with a 95% confidence interval (calculated as described in [2, 139]).

The MLE model improves classical MOS calculation by removing the uncertainty from subjects and contents. In our case, the recovered MOSs (x_e in eq. 3.2) remain close to the classical MOSs (from eq. 3.1) (0.998 Spearman correlation). However, the bias, inconsistency, and content ambiguity values obtained constitute valuable information for further analysis (see the following section).

The recovered MOSs (x_e instead of MOS_e) are considered as the ground truth quality scores of our dataset.

3.4.2 Observers' agreement

Before analyzing the results of our subjective experiment, it is important to evaluate the agreement between the subjects and whether they maintained their attentiveness during the test. To do so, we consider two types of indicators: (1) the correlations between subjects' ratings, and (2) the bias b_s and inconsistency v_s from the MLE model.

First, as in [31], we computed the Pearson Linear Correlation Coefficient (PLCC) and the Spearman Rank Order Correlation Coefficient (SROCC) between the scores of each observer and the recovered MOSs of the stimuli rated by this observer. We then averaged the correlations over all the subjects. The (mean, standard deviation) of PLCC and SROCC are (0.85, 0.055) and (0.81, 0.063) respectively. The mean of the 2 correlations is high, while the standard deviation is rather low, which indicates a good agreement between the

subjects.

We then further explored the internal consistency of the subject data as proposed by [59, 140]. For each stimulus, we randomly divided the subjects who rated it into two equal size groups (12 observers per group) and calculated the SROCC between the recovered MOSs of the 2 groups. After repeating the split 500 times, the range of correlations was found to be between 0.915 and 0.944 with a mean and a median value of 0.929. Hence, there is a high degree of inter-subject agreement despite the immersive viewing environment.

Moving to the second type of indicators: the subject's bias and inconsistency computed by the MLE model. Bias (shown in Figure 3.3) reflects the sensitivity of the subject to impairments. It is a systematic error generated by the subject throughout the experiment (i.e. picky/expert participants tend to be biased toward lower scores). Inconsistency (shown in Figure 3.4), a.k.a random error, points out the inattentive subjects that give random scores or subjects showing absent-mindedness for a part of the test.

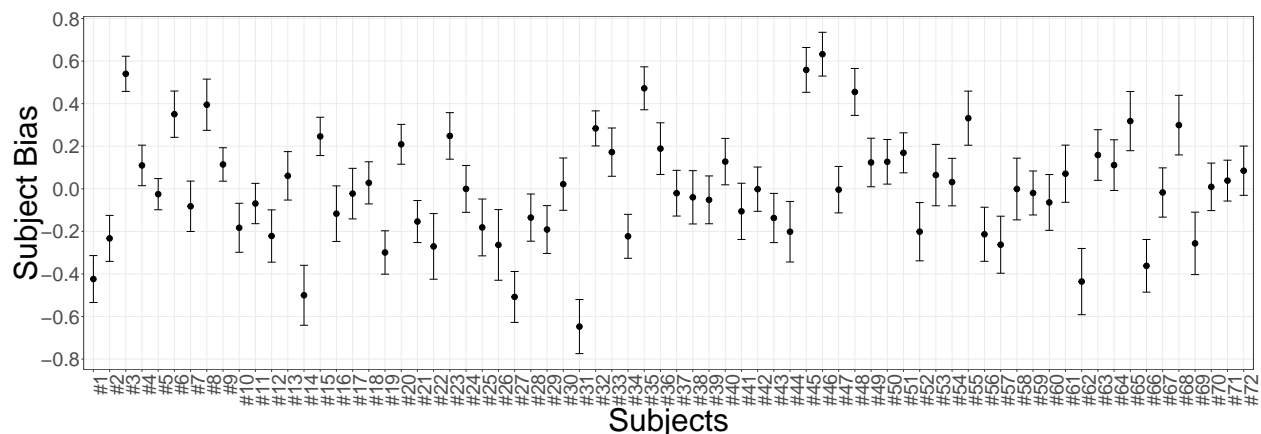


Figure 3.3: Bias b_s of each subject involved in our subjective experiment, and its confidence interval.

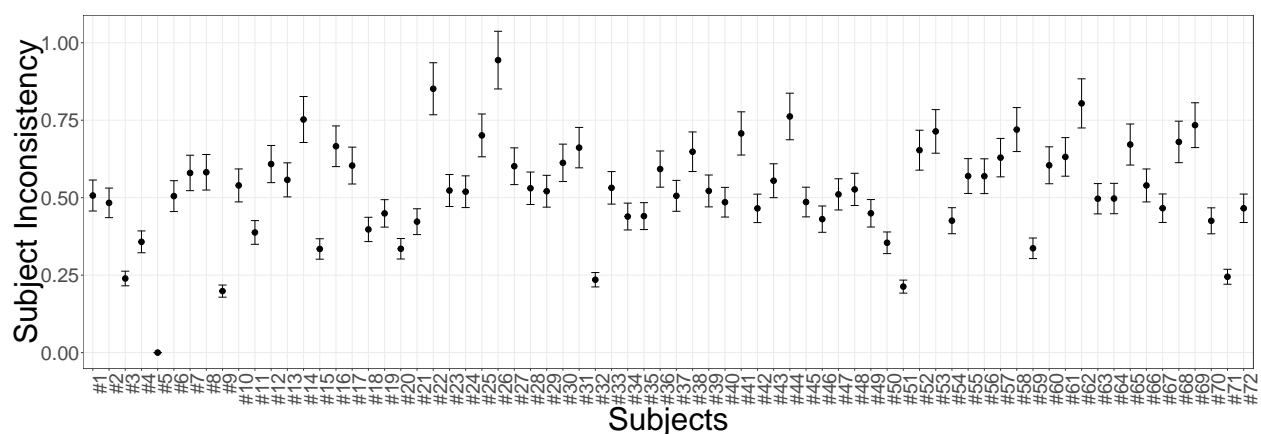


Figure 3.4: Inconsistency v_s of each subject involved in our subjective experiment, and its confidence interval.

Figures 3.3 and 3.4 show that the range of bias and inconsistency values is within those of image/video experiments [2, 139, 141]. These figures reported no implausible bias or inconsistency values, nor any loose confidence intervals, which means that subjects maintained attentiveness throughout the test and were sensitive to impairments. This is coherent with the results obtained using the ITU-R BT.500-13’s outlier detection method (section 3.4.1).

Finally, we assess whether previous VR experience influences subjects’ judgments. Thus, we divided the observers into 2 groups: those who are familiar with VR (45 subjects) and those who have never tried a VR headset (27 subjects). For each group, we computed the correlations between subjects’ ratings and MOS. We then averaged the correlations over the subjects of each group. Furthermore, we assessed the inconsistency v_s of the 2 groups. Table 3.1 summarizes the results, while Figure 3.5 shows the boxplots of subjects’s inconsistency in relation to their familiarity with VR.

Results show that there is no significant difference in the behavior of observers with no VR experience and those familiar with VR. We believe this is due to the fact that the task given to the participants is rather simple: observe and then vote using the trigger of the HTC Vive controller. As can be seen, no VR expertise is required, since there is no manipulation of the objects. Results also point out that our training stage was well-designed.

(Mean, SD)	PLCC	SROCC	Inconsistency
Subjects unfamiliar with VR	(0.845, 0.057)	(0.811, 0.065)	(0.536, 0.156)
Subjects familiar with VR	(0.845, 0.056)	(0.813, 0.062)	(0.517, 0.165)

Table 3.1: Agreement and inconsistency of subjects familiar with VR and those with no VR experience.

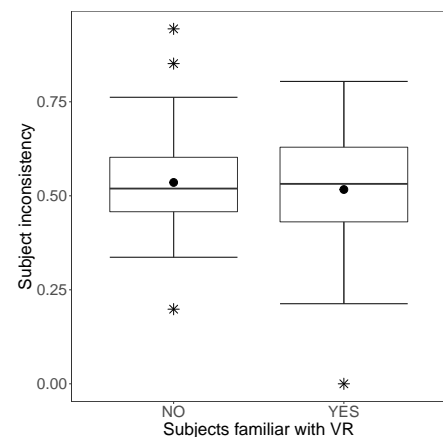


Figure 3.5: Boxplot of subjects’ inconsistency in relation to their familiarity with VR.

3.4.3 Factors influencing subjective opinions

Our objective is to provide a deep and evidence-based understanding of the factors that influencing subjective opinions. We quantitatively evaluate the effects of source models, distortions, viewpoints and movement on the mean opinion scores (MOSs) and their Confidence Intervals (CIs). Note that, the classic MOSs and CIs (MOS_e , Eq. 3.1) are used in this analysis instead of the recovered MOSs and their corresponding CIs (x_e , Eq. 3.2) obtained by the MLE model, since the latter are recovered from the influence of the source models (content ambiguity of Eq. 3.4) which we believe is an important factor to consider in our study.

Resulting MOSs and CIs

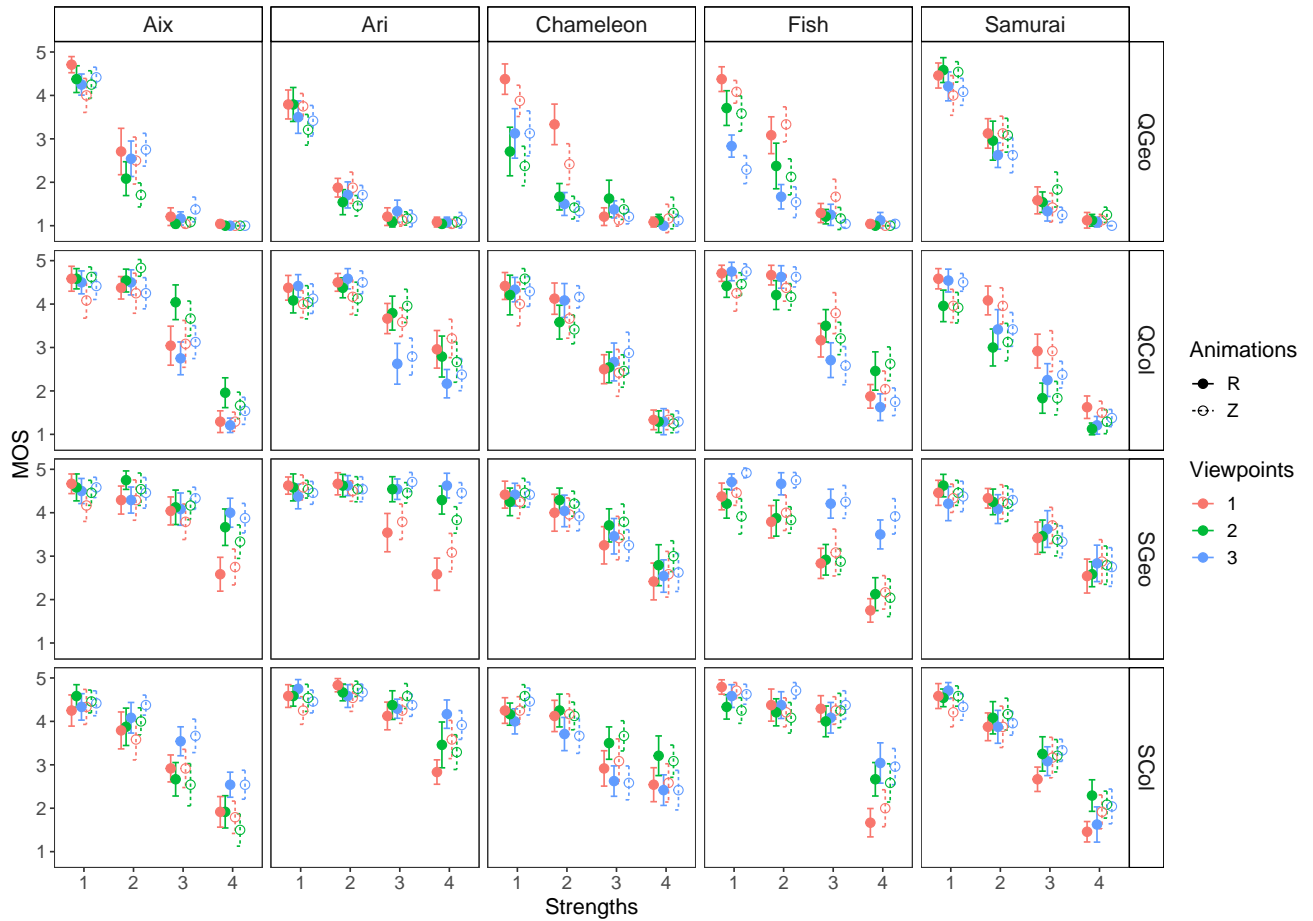


Figure 3.6: The mean opinion scores of all the stimuli, associated with their confidence intervals. For a given distortion strength, the dots are horizontally spaced apart to avoid overlapping.

Figure 3.6 shows the MOSs and confidence intervals for all the stimuli, averaged over all the observers. As expected, on the whole, MOSs decrease as distortion strengths increase. We notice that observers’ behavior was virtually the same for the stimuli whether they were rotating or zooming in/out. However, we can observe that the effect of the viewpoints is strongly related to the source model and the distortion type. For instance, the 3 viewpoints of the Chameleon geometrically simplified (Chameleon_SGeo) obtained roughly the same score, unlike the Fish geometrically simplified (Fish_SGeo) whose *viewpoint 3* was rated better than the other 2 viewpoints (see Figure 3.7.a). Regarding the influence of the distortion type, the geometric simplification of the Fish (Fish_SGeo) is more visible on *viewpoint 1* and *viewpoint 2* than on *viewpoint 3*, while its “Color-aware” simplification (Fish_SCol) has almost the same impact on the 3 viewpoints (see Figure 3.7.b)

For better readability in interpreting the influence of the viewpoints, we provide Figure A.6 in the appendix which shows the results of the 2 animations plotted separately.

We also notice variations in confidence interval length depending on the source content (i.e. the Chameleon’s CIs are globally larger than those of Ari). In addition, it seems that, overall, *viewpoint 3* provides smaller CIs than *viewpoint 1*.

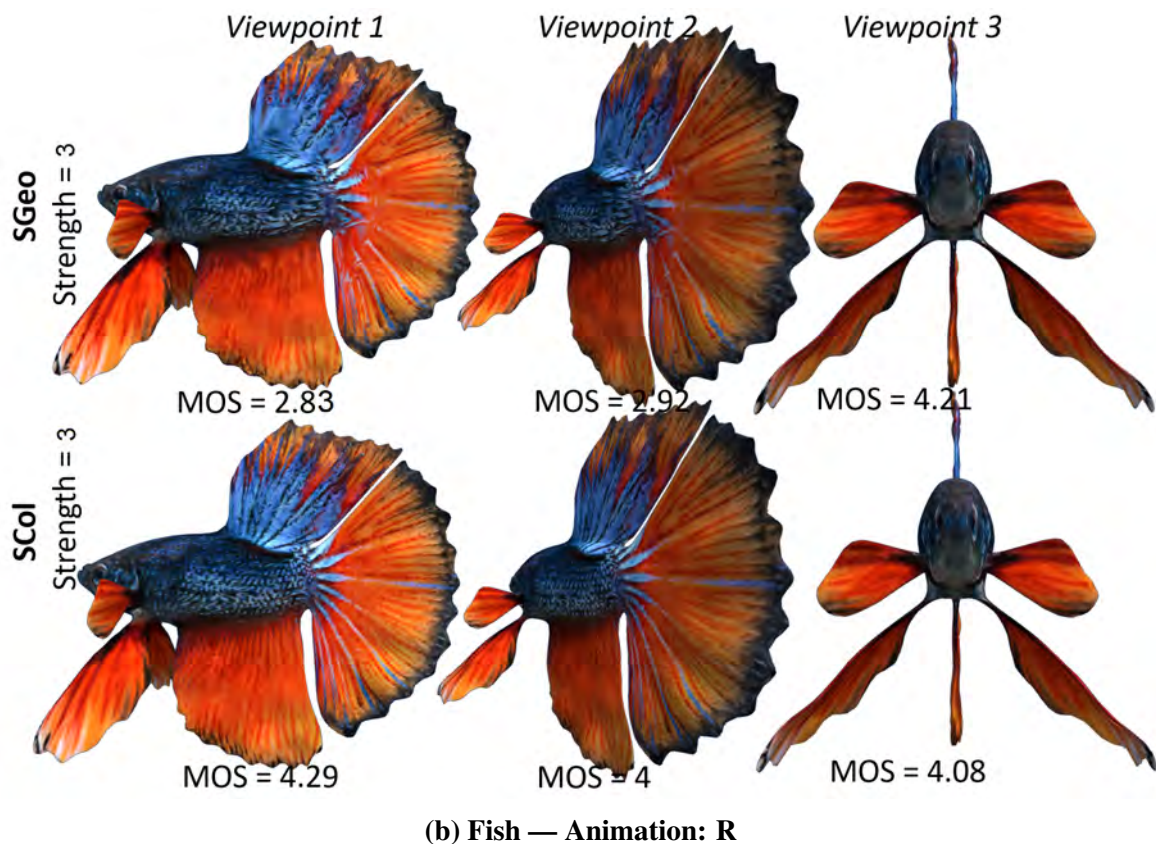
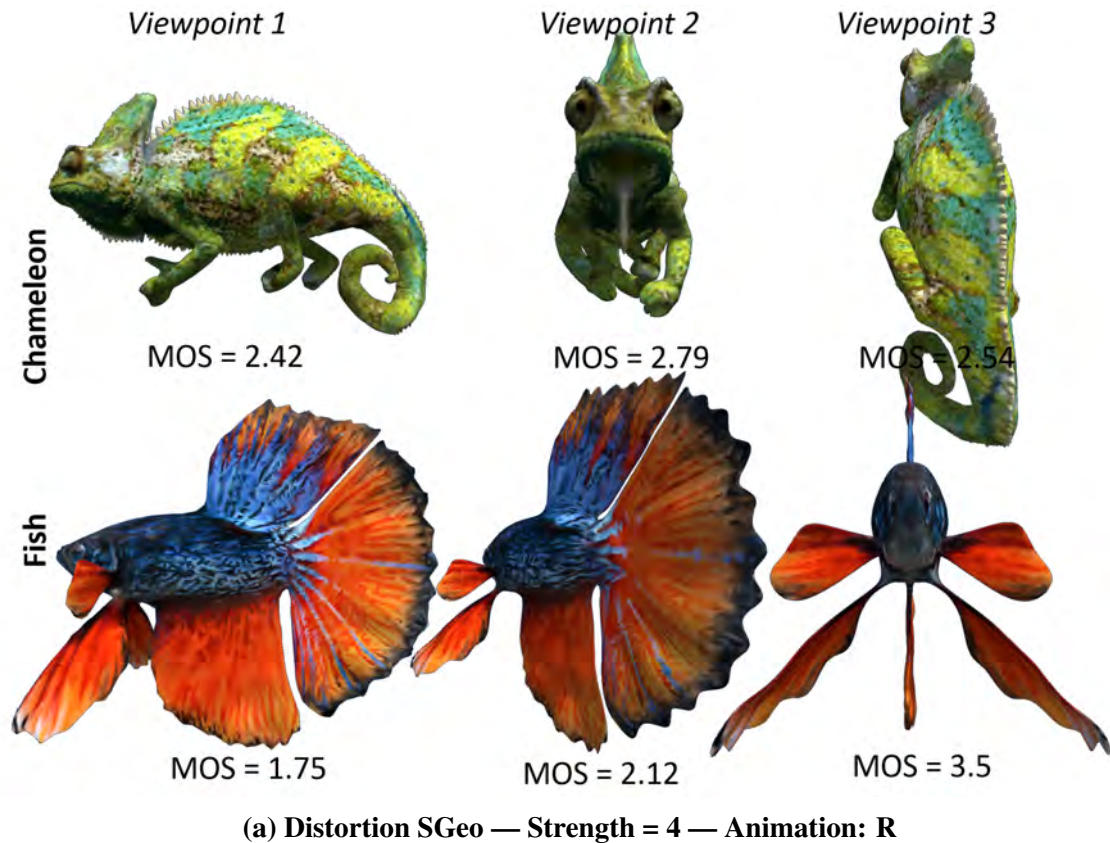


Figure 3.7: The variation of the MOSs of different (a) source models and (b) distortions depending on the viewpoints.

All these factors and phenomena are analyzed quantitatively in the following.

Influence on MOSs

We ran a Multivariate Analysis of Variance (ANOVA: Source models \times distortion types \times distortion strengths \times viewpoints \times animation types) on the rating scores of the observers. Figure 3.8 summarizes the most important results using boxplots of MOSs.

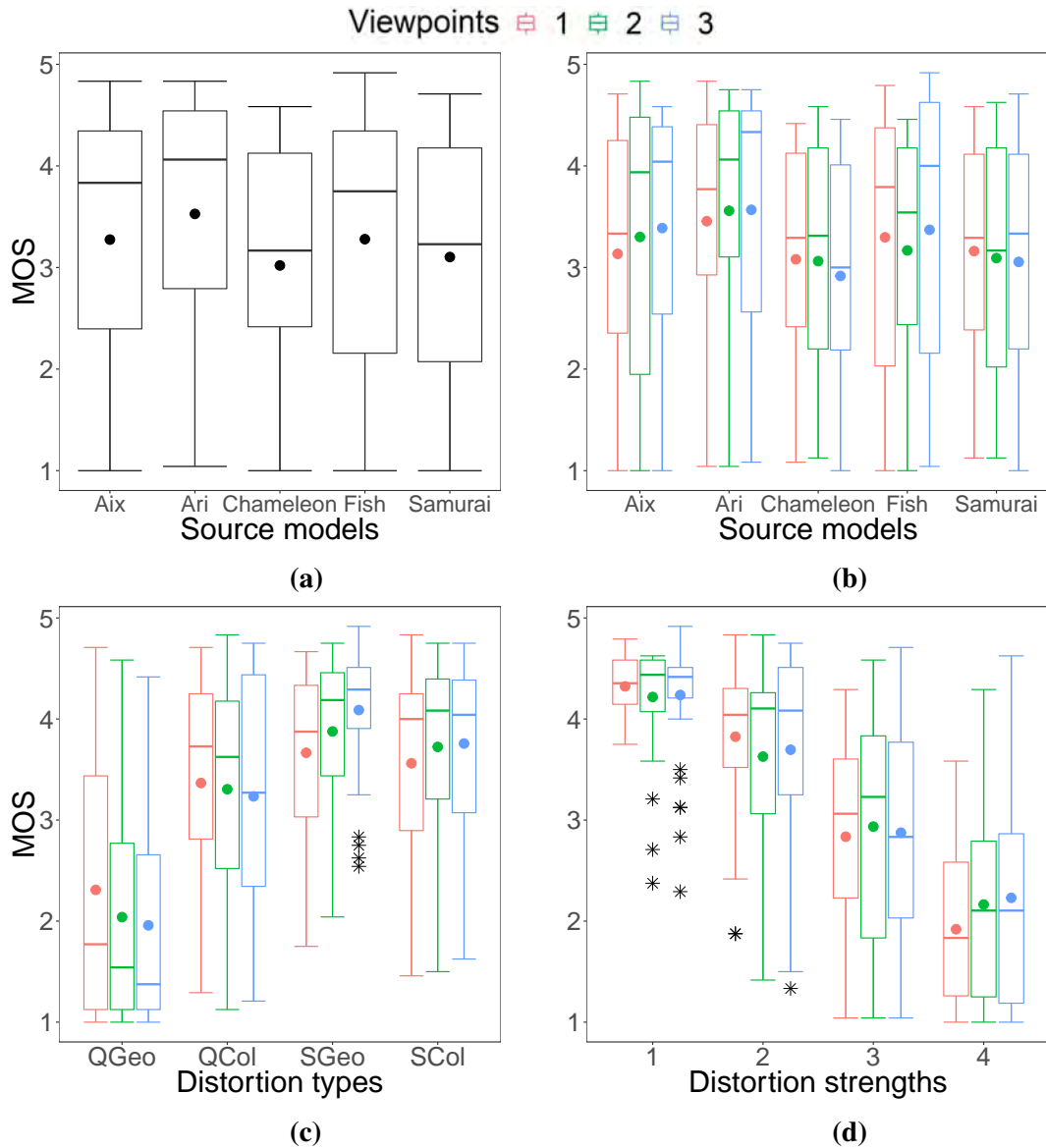


Figure 3.8: Boxplots of MOSs obtained for the combination of the viewpoint with different factors. Mean values are displayed as circles.

- Source models, distortion types and strengths: as expected, the ANOVA test shows that these 3 factors are the most significant factor variables (the corresponding p-values $\ll 0.0001$) (see Figure 3.8.a).

- **Viewpoints:** there are no significant differences in the subjective scores associated with the 3 selected viewpoints (p-value=0.189). However, a significant interaction effect was found between the viewpoint and the source content (p-value $\ll 0.0001$) (see Figure 3.8.b). This effect appears in Figure 3.6 and in the visual example Figure 3.7.a where for some models (e.g., Chameleon, Samurai) we observe the same observers' behavior for the 3 viewpoints, contrary to other models (e.g., Aix, Fish, Ari) where the perceived quality varies according to the viewpoints.

Figure 3.8.c witnesses the strong interaction between the viewpoint and the distortion type observed earlier in Figure 3.7.b (p-value $\ll 0.0001$). It shows that *viewpoint 1* is much more sensitive to geometric simplification (SGeo) than *viewpoint 3*. This effect is reversed for geometric quantization (QGeo), since *viewpoint 3* got the lowest average scores. Our hypothesis is that the geometry and silhouette alterations caused by QGeo are masked by rich colors and details of *viewpoint 1* (*viewpoint 1* is much richer in details than *viewpoints 2* and *3*). This is not the case for geometric simplification (SGeo), which markedly degrades colors and is thus more visible on *viewpoint 1*. Figure 3.9 illustrates a visual example: observers were able to detect SGeo distortion when the Fish was shown in *viewpoints 1*. This distortion is not as apparent/visible when the Fish was displayed in *viewpoint 3* and is therefore harder to detect, resulting in a higher MOS. For QGeo, we clearly observe the opposite effect.

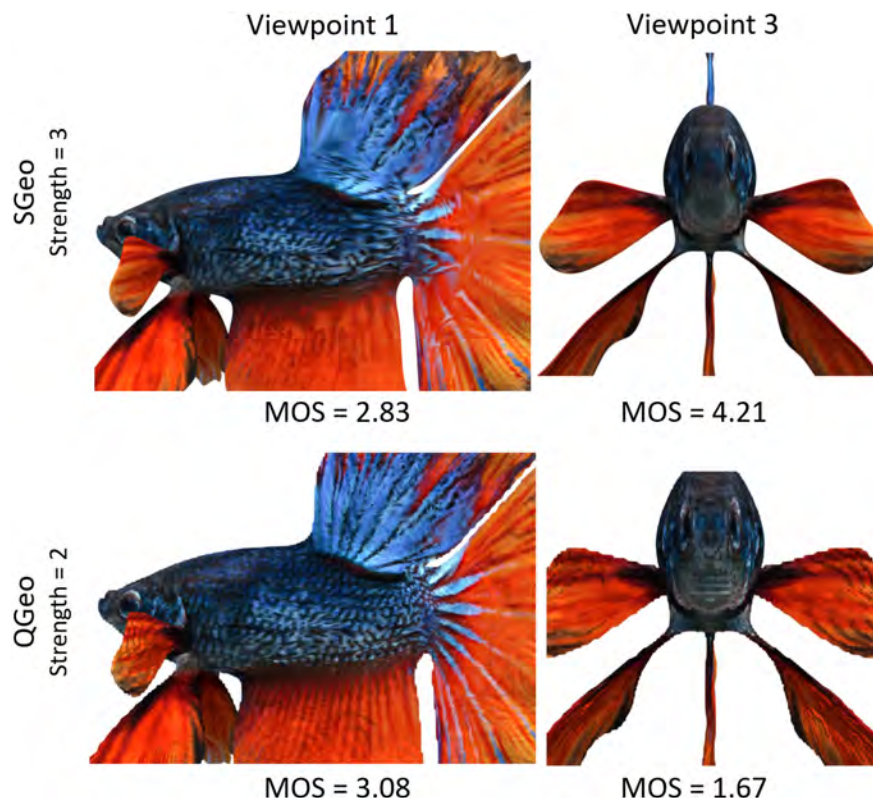


Figure 3.9: Visual example of the interaction between between the viewpoints of the *Fish* and different distortion types (Animation: R).

Figure 3.8.d shows that a significant interaction exists between the viewpoint and the dis-

tortion strength (with a p-value $\ll 0.0001$). Indeed, stimuli with high strength of impairment (strength=4) obtained better scores when displayed in *viewpoints 2* and *3* than in *viewpoint 1*. This is due to the fact that *viewpoint 1* covers most of the shape and carries the most information and details. Thus, strong distortions are particularly destructive for this viewpoint. This effect is obviously less visible for high quality stimuli (strength = 1). Figures 3.6 and 3.10 show a concrete example: for Ari geometrically simplified (SGeo) and shown in *viewpoints 2* and *3*, as distortion forces increase, MOS values remain almost stable. These 2 viewpoints show the side and the back of the statue, respectively. These areas are almost flat and contain very few geometric details, especially the shape of the back (*viewpoint 3*). Therefore, simplifying these regions, even with high strength, will not introduce any markedly visible distortions to the model. This is not the case of Ari displayed in *viewpoint 1*, since *viewpoint 1* contains more salient features and details such as the face.

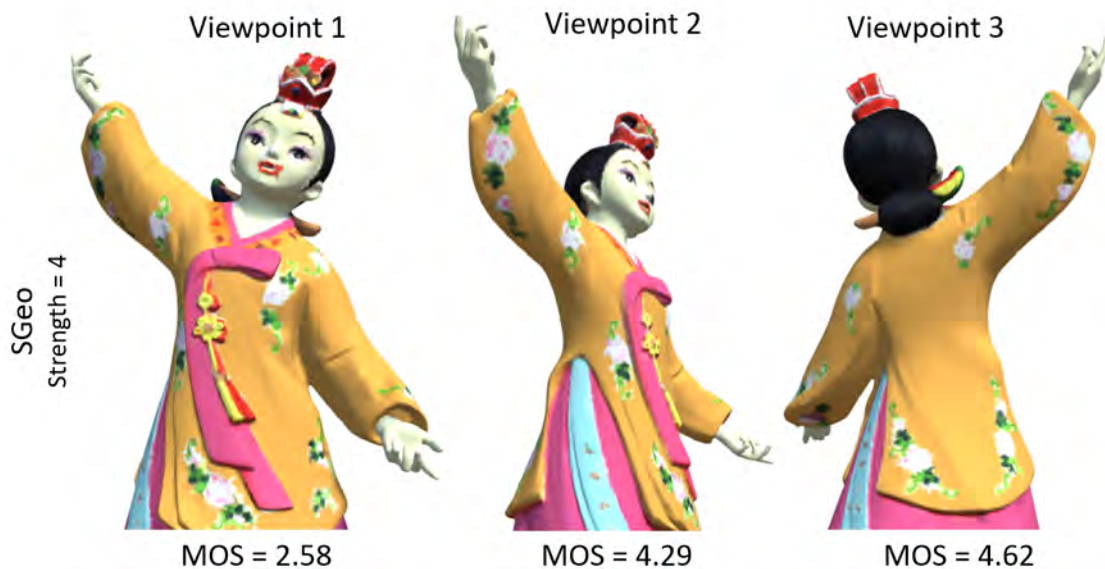


Figure 3.10: The visual quality of Ari highly distorted under its different viewpoints (Animation: R).

- **Animations:** according to the ANOVA test, the animation itself does not affect significantly the perceived quality (p-value = 0.165). However, a significant interaction was found between the animation and the distortion strength (p-value < 0.0001). Indeed, as seen in Figure 3.11.a, weak distortions (strength = 1) are easier to detect in zoom than in rotation, since by zooming in, the observer can see more details and low-level features. Stimuli with high distortion (strength=4) obtained roughly the same score in both animations. Moreover, there is a slight interaction (with a p-value=0.09) effect between the animation and the viewpoint (Figure 3.11.b). The interaction between these 2 factors will be discussed below since it has much more influence on CIs than on MOSs.

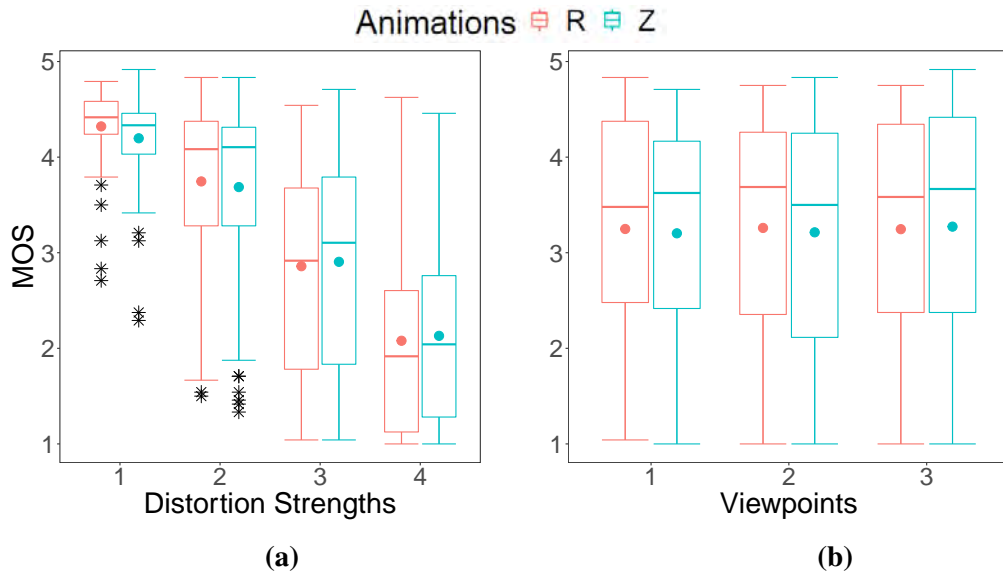


Figure 3.11: Boxplots of MOSs obtained for the combination of the animation with different factors. Mean values are displayed as circles.

Influence on CIs

This time, we ran the ANOVA test on the 95% confidence intervals of the MOSs. This allows us to evaluate the impact of the factors on the dispersion of individual ratings. As above, Figure 3.12 summarizes the most important results using boxplots of CIs.

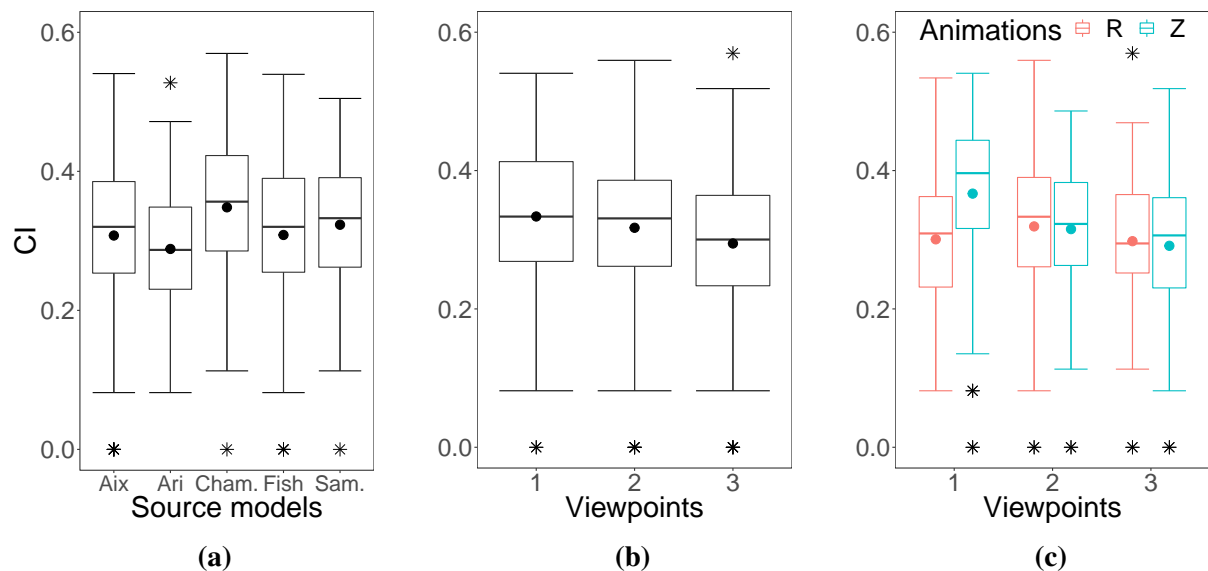


Figure 3.12: Boxplots of CIs obtained for different factors or combination of factors. Mean values are displayed as circles. In (a), *Cham.* and *Sam.* refer to Chameleon and Samurai, respectively.

- **Source models:** When looking at Figure 3.12.a, it appears obvious that the source models influence the agreement among the subjects ($p\text{-value}=0.0016$). Indeed, selection of source

models is not a trivial task: some contents tend to be more difficult to rate than others. This phenomenon is represented in the MLE model by the ambiguity of the content a_c (see Eq. 3.4 and Figure 3.15). Overall, the chameleon tends to be the source associated with the highest content ambiguity (subjects disagree). A reasonable explanation for this is that the chameleon model carries more geometric and color details than all the other models: it has a high average curvature, sharp edges, diversity of colors, and many small details to reflect its skin tone and geometrical characteristics (see Figure 3.1 and Table 2.1).

- Viewpoints: It is interesting to observe that the viewpoint has a significant impact on CIs (p -values=0.0035), unlike that on MOSs. Figure 3.12.b shows that the CIs of *viewpoint 1* are larger than those of the other viewpoints. Figure 3.13 shows a visual example. The fact that *viewpoint 1* contains more details/information on color and geometry than the others implies that this viewpoint results in higher dispersion between the observers' scores. This ties in with the observation on the source model (the previous paragraph): the more geometric and color information/details there are, the greater the dispersion between subjects' ratings.



Figure 3.13: The variation of the CIs length of the *Samurai* depending on the viewpoint.

- Animations: Overall, the CIs of the stimuli in rotation are smaller than those in zoom. Still this difference is moderate (p -value=0.052). The impact of this factor is emphasized when considering the interaction between animations and viewpoints (p -value=0.0019). Indeed, Figure 3.12.c shows that models animated with zoom tend to be more ambiguous (result in larger CIs) than those that rotate, notably when the models are displayed in *viewpoint 1*, which is the viewpoint that covers most of the shape and carries the most information. This effect can be observed in Figure 3.14, notably for the Samurai, Aix and Chameleon models. This can be explained by the fact that while zooming, especially in

viewpoint 1, the observer can see more details and low-level features, which makes the task of evaluating differences between the reference/source and the impairment stimuli more difficult than the other HRTs. This proves once again that the more information/details there are, the greater the dispersion between observers' scores.

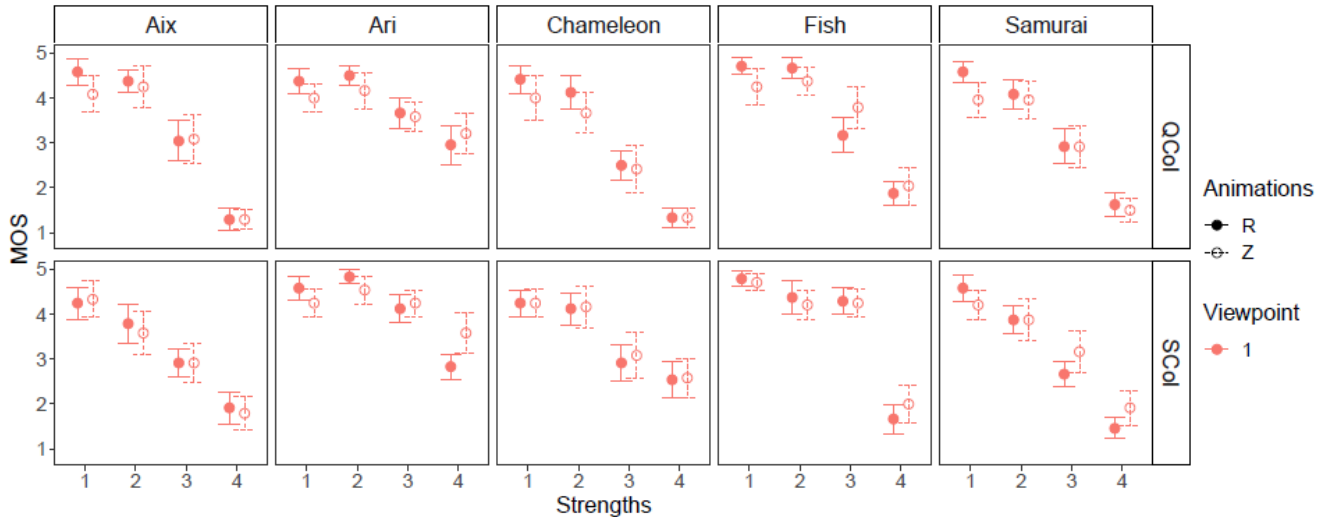


Figure 3.14: The mean opinion scores associated with the CIs for a subset of models displayed in *viewpoint 1* and animated with a rotation and a zoom.

We analyzed the ambiguity of our source contents (a_c of Eq. 3.4), obtained by the MLE model, for each viewpoint and animation. The results, provided in Figure 3.15, are consistent with Figure 3.12.c: sources with high confidence intervals are also associated with high ambiguity values. Figure 3.15 also shows clearly the interaction between the animation and the viewpoint.

3.5 Discussion and recommendations

This section synthesizes the findings of this chapter. We observed that, for given distortions, the viewpoint has a significant impact on the MOSs: Complex masking effects occur when considering the interaction between viewpoint and distortion. The geometric simplification alters the most informative viewpoint (the one richest in colors and details) as this distortion strongly degrades the colors. This effect is reversed for the geometric quantization which is less visible on this viewpoint because the rich colors and details can mask the alteration of the geometry caused by the quantization. The viewpoint also interacts with the distortion strength. Strong distortions are particularly detrimental for viewpoints rich in details. Furthermore, this factor affects the CIs of the ratings: the most informative viewpoint tends to produce the largest confidence intervals, especially when combined with a zoom animation.

Although animation, by itself, has a moderate impact on subjects' opinions and CIs, the impact of this factor is emphasized when it interacts with other factors: (1) the distortion strength for subjective scores (MOSs) as weak distortions are easier to detect in zoom than

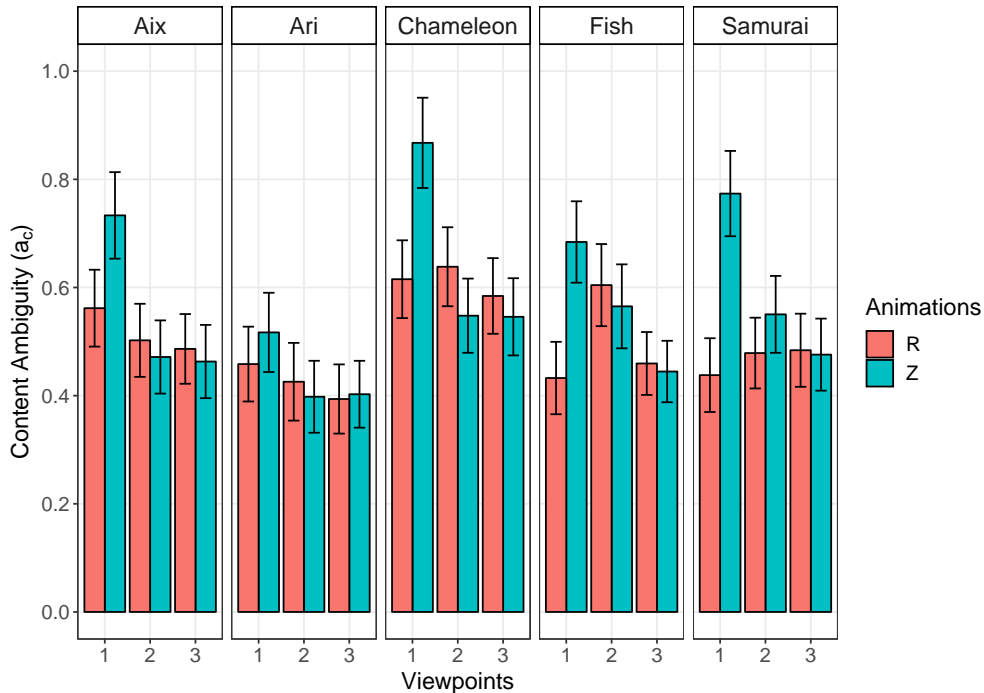


Figure 3.15: Content ambiguity a_c of each source model associated with the different HRTs, and its confidence interval.

in rotation, and (2) the viewpoint for CIs as models animated with zoom tend to be more ambiguous than those that rotate especially when they are displayed in the viewpoint that carries the most color and geometric details.

Our results also show that the ambiguity of the source model (i.e. dispersion of subject ratings) is potentially related to its geometric and color complexity: models with more details are the most difficult to rate (larger CIs and higher ambiguity values). The more visible the content's information/complexity, the higher the ambiguity and dispersion between observer scores (CIs).

Our findings suggest recommendations for the design of an objective quality assessment metric for 3D Graphics. First, since the models and their distortions and strengths have crucial importance on perceived quality, the metric must be able to adapt to the models, and to their shapes and colors. Moreover, it must be able to detect different distortions applied on both geometry and color map. We develop such a metric in Chapter 6. Considering short animation as a non-influential factor, it is ineligible for integration in the quality metric. However, since viewpoint has an impact, albeit moderate, it may be useful to take it into account in the metric. These proposals are investigated in Chapter 6.

3.6 Conclusion

In this chapter, we designed and produced a large subjectively-rated dataset of colored 3D meshes. This dataset is composed of 480 dynamic stimuli and obtained through a subjective study based on the DSIS method, in a virtual reality environment. The stimuli were

generated from 5 source models subjected to geometry and color distortions. Each stimulus was associated with 6 HRTs: combinations of 3 viewpoints and 2 animations. A total of 11520 quality judgments were collected. The dataset is made publicly available online². This study allowed us to draw interesting conclusions regarding the influence of source models, distortions, viewpoints and animations on both the quality scores and their confidence intervals. Further studies should be carried out so that the results of our experiment can be generalized to a non-VR scenario. We have undertaken this in the following chapters.

²<https://yananehme.github.io/datasets/>

Chapter 4

Exploring Crowdsourcing for Subjective Quality Assessment of 3D Graphics

Like many fields, our subjective studies have been impacted by the COVID-19 pandemic. Crowdsourcing was therefore the alternative solution.

Subjective quality assessment experiments have been traditionally conducted in laboratories (lab), in a controlled environment and with high-end equipment. Our previous subjective studies, presented in Chapters 2 and 3, belong to this category. In recent years, CrowdSourcing (CS) experiments have gained quite a lot of popularity, especially with the development of the internet and the growing trend of machine learning. In fact, crowdsourcing tests are relatively fast and are therefore more practical for evaluating large data sets, which require weeks of laboratory evaluations. In addition, CS has become even more prevalent during the COVID-19 pandemic (from 2019 to present), where participants could not be physically present in the laboratory to carry out tests.

CS and lab studies differ considerably in several aspects, e.g. the diversity of the participants, the duration and setup of the tests. As a result, CS imposes several challenges to overcome compared to similar lab tests, notably those related to the lack of control over the participants' environment and the reliability of the participants since the latter are not supervised in CS tests (section 1.1.4 of Chapter 1). Despite these challenges, CS studies are still capable of producing accurate and reliable results for different types of media if the experiment framework has been properly designed [62,63].

Before conducting large subjective quality assessment experiments in CS, we wanted to investigate:

- The reliability of CS studies to assess the quality of 3D graphics.
- Whether a CS test can achieve the accuracy of a laboratory test.

For this purpose, we designed a CS experiment that replicates as much as possible the lab experiment presented in Chapter 2 and conducted in VR. Thus, we used the same instructions, the same dataset of 3D models and the same subjective evaluation methodology: the Double Stimulus Impairment Scale (DSIS) method, which is, to the best of our knowledge, one of its first uses in CS studies. We then compared the results of the CS experiment to

the previously collected lab results and showed to what extent and under which conditions a CS study can be as accurate as a laboratory study.

This chapter is organized as follows: Sections 4.1 and 4.2 briefly recall the dataset and the lab experiment. The design of the CS experiment is detailed in section 4.3, while its results are presented in section 4.4.

4.1 Dataset

The 3D graphic stimuli used in the CS study were previously used in the laboratory based subjective experiment, reported in Chapter 2. This dataset contains 80 animated meshes with diffuse color information. It was generated from 5 source models (“Aix”, “Ari”, “Chameleon”, “Fish”, “Samurai”), corrupted by 4 types of geometry and color distortion that represent common simplification and compression operations: uniform geometric quantization (QGeo), uniform LAB color quantization (QCol), simplifications that take into account either the geometry only (SGeo) or both geometry and color (SCol). Each distortion was applied with 4 different strengths that cover the whole range of visual quality from imperceptible to high levels of impairment. The source models and some examples of distorted stimuli are illustrated in Figures 2.2 and 2.3 respectively, while full details on model characteristics and distortion parameters are provided in Tables 2.1 and 2.2.

4.2 Lab experiment

We have previously conducted several lab experiments using this dataset to assess the perceived quality of 3D graphics in VR (see Chapter 2). The experiment we consider in this chapter is the one based on the DSIS method, because as shown in Chapter 2 this method provides the best results in terms of stability and accuracy. The experiment was conducted in a VR environment using the HTC Vive Pro headset. Stimuli, animated with a slow rotation, were rendered for 10 seconds in a virtual scene at a fixed distance from the observer and rotated in real time. Please refer to section 2.3 of Chapter 2 for more details. The experimental environment of the lab test is illustrated in Figure 2.4 (2nd row). The study involved 30 participants: students and professionals at the University of Lyon (see sections 2.5.3 and 2.8). Each participant evaluated the whole dataset (80 stimuli) in one session that lasted about 23 minutes.

4.3 Crowdsourcing experiment

We conducted a crowdsourcing (CS) experiment in which participants were recruited to assess the quality of this dataset of 3D graphics. Since our goal is to compare the results of the CS experiment to previously collected Lab results, we replicated the lab experiment, described in section 4.2, as much as possible. This section provides details on the CS experiment.

4.3.1 Experimental environment

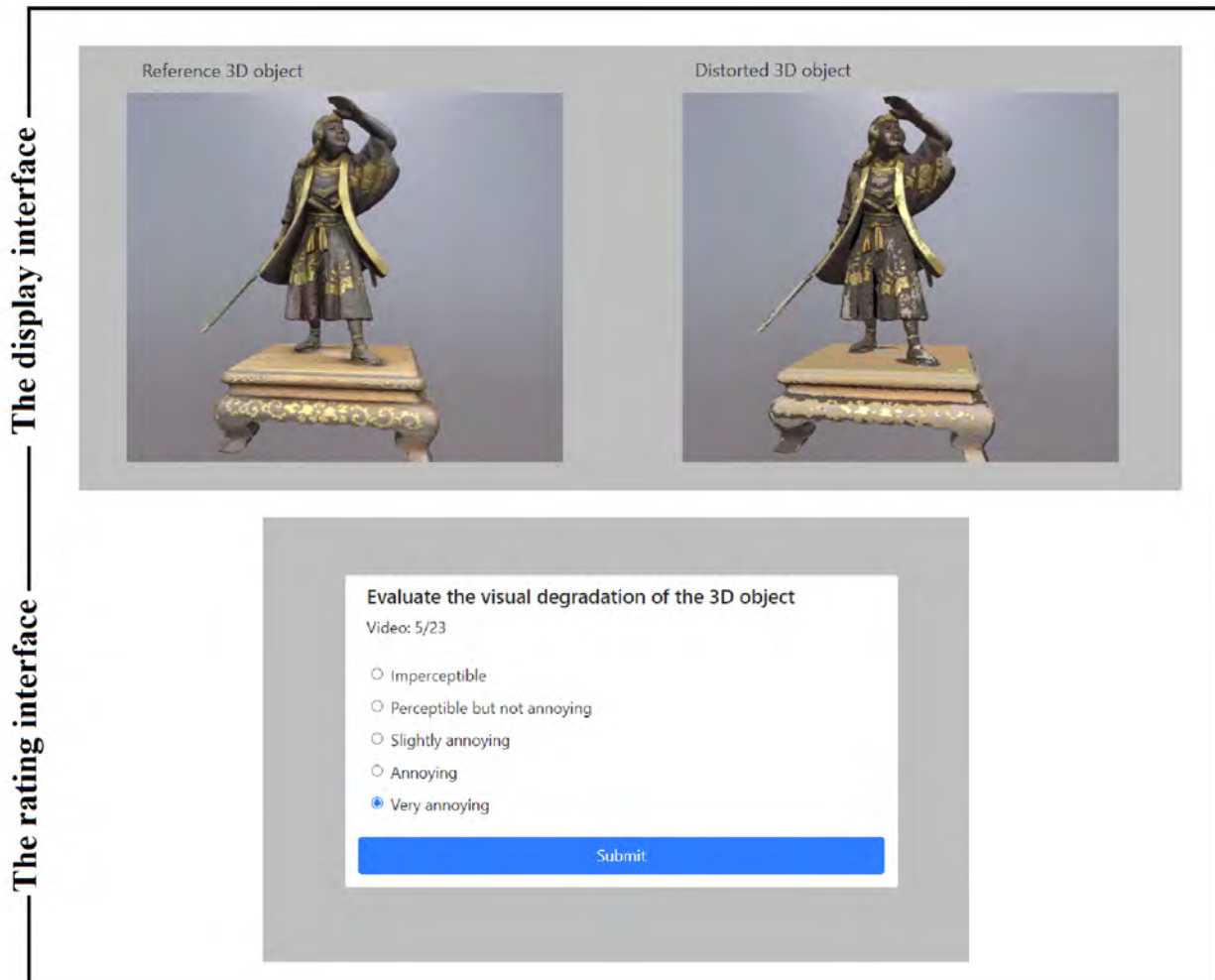


Figure 4.1: The graphical interface based on the DSIS method of our CS experiment.

In order to be able to compare the results of the CS and lab studies, we opted for the same setup as implemented in the lab experiment (section 2.3.1): the stimuli were animated with a slow rotation (of 15° around the vertical axis in clockwise and then in counterclockwise directions) and displayed in a neutral room (light gray walls) under the same viewpoints as those used in the lab experiment. Their material type complied with the lambertian reflectance model, and they were visualized without shadows and under a directional light.

The subjective testing methodology used in CS is also the same as the lab experiment, namely DSIS, in which observers see the reference model and the same model impaired, simultaneously, side by side, for 10 sec and rate the impairment of the second stimulus in relation to the reference using a five-level impairment scale, displayed after the presentation of each pair of stimuli. Thence, we generated videos of the final rendered dynamic stimuli. The videos were all in 650 x 550 resolution (so that the videos of the reference and degraded models fit simultaneously on a screen with a minimum resolution of 1920 x 1080) with a frame rate of 30 fps and encoded using H.264 encoder (mp4 container) at

a bitrate of 5 Mbps to ensure imperceptibility of compression artifacts. The duration of the videos is 10 sec, which corresponds to the display time of the models in the lab experiment.

To conduct the CS experiment, a web-based platform was developed suitable for presenting videos according to the DSIS method. It does not require the participants to install any software on their device (except a web browser with an MPEG-4 decoder). The platform first checks the screen resolution of the participants (must be equal to or higher than 1920 x 1080) as well as the page zoom level and ask them to keep the full screen mode until the end of the experiment. Otherwise, they are not allowed to proceed in the test. Once the device's compatibility has been verified, the test instructions are displayed to the participants. At the bottom of this page is a progress bar showing the status of the loading process of all the video pairs that will be used in the test. When the loading is completed a start button appears leading to the test. In this way, the videos of the reference and distorted models are ensured to be played simultaneously without any latency or unintended interruptions. Figure 4.1 illustrates the graphical interface of our CS experiment.

The experiment starts with a training, during which observers familiarize themselves with the task [11]. We selected the same training model used in the training of the lab experiment as well as its distorted versions which cover the whole range of distortions. After displaying each pair of training videos for 10 sec, the rating interface is displayed for 5 sec and an example score assigned to this distortion is highlighted. Once the training is completed the actual test starts.

The videos of the stimuli are displayed in a random order (3D models, distortion types and levels all mixed) to each participant. Each video/stimulus is presented once; participants are not able to replay the videos. Moreover, participants are not able to provide their score unless the videos have been played completely. There is no time limit for voting and videos of the stimuli are not shown during that time. At the end of the experience, participants will receive unique codes allowing them to get their remuneration.

4.3.2 Participants and test sessions

CS experiments should be kept as short as possible to keep participants motivated and to avoid unreliable results [55–57]. Therefore, we divided our dataset of 80 stimuli into 4 groups, called playlists (i.e. 20 stimuli/playlist), and each participant evaluates one playlist of the dataset. The stimuli were distributed evenly across the playlists so that each playlist contained the 5 source models and the 4 distortion types and strengths. Furthermore, as this database encloses the subjective scores obtained in the lab experiment, we opted to distribute the stimuli between the playlists so as to have the same Mean Opinion Scores (MOSs) distribution in these playlists.

Since participants in CS cannot be supervised, it is important to ensure the quality of annotations by detecting unreliable and malicious participants. To do so, there are 2 common/popular strategies [142]. The first one, known as gold standard, is the insertion of dummy stimuli or stimuli with trusted annotations, while the second strategy (a.k.a consistency question) is to collect multiple scores for repeated stimuli, which allows to assess

participant’s consistency. Inspired by these strategies and as in [59], we injected 3 trapping stimuli, that we called golden units, into the playlists of our test. They consisted of (1) a very poor quality stimulus (high level of impairment), (2) a very high quality stimulus (no impairment, the reference is compared to itself) and (3) a stimulus displayed twice to assess the participant’s consistency (coherence of his/her scores). Participants who fail to answer the golden units correctly are considered outliers and their scores are rejected (more details in section 4.4.1).

Thus, the test session of our CS experiment is constituted of 23 pairs of videos to rate (1 playlist) and lasts about 10 minutes: informed consent + loading videos + instructions + 6 training stimuli \times (10s video length + 5s Rating) + 23 test stimuli \times (10s video length + \sim 4s Rating).

We ran our experiment until each playlist was fully rated by 60 participants, which required approximately 12 hours. A total of 240 participants took part in this study: 148 males and 92 females. They were from 33 different countries and aged between 18 and 68. All participants were naive about the purpose of the experiment. Note that, participants who started the experiment and did not complete it or who left after reading the instructions (101 participants) were discarded.

The recruiting process of the participants was performed using Prolific¹, an internet marketplace that provides tens of thousands of trusted participants. Only participants having a high reliability score (score based on how well they did in past studies) and an adequate number of duly completed jobs (number that reflects their familiarity with the platform) on Prolific were admitted to the experiment.

4.4 Results and comparison of experiments

This section presents the results of the crowdsourcing experiment and compares them to the lab results in terms of accuracy, confidence intervals, and participants’ agreement. For the subsequent analyses, subjective scores ranging from very annoying to imperceptible are mapped on a discrete numerical scale from 1 to 5. Note that the scores assigned to the golden units are only taken into account in participants screening and are not considered in the rest of the analyzes.

4.4.1 Participants screening

Before starting any analysis, it is necessary to identify and remove outliers which could affect the accuracy of the results.

The participants of the lab experiment were filtered according to the ITU-R BT.500-13 recommendation [11]. We found 1 outlier (i.e. 29/30 subjects remain).

For the CS study, as recommended in [58], participants were screened by combining: (1) the ITU-R BT.500-13 screening procedure, which revealed 7 outliers among the 240 par-

¹<https://www.prolific.co/>

ticipants and (2) the golden units (trapping stimuli) analysis: we found that 4 participants incorrectly rated the very poor quality stimulus (i.e. its distortion is rated as imperceptible or perceptible but not annoying), 2 participants misjudged the very high quality stimulus (i.e. its distortion is considered very annoying or annoying; however, this stimulus is not degraded. It is the reference), and lastly, 1 participant gave inconsistent scores to the third golden unit, called G (i.e. $|s_i^{G_{rep1}} - s_i^{G_{rep2}}| \geq 3$, where s_i^G denotes the score assigned by participant i to stimulus G , shown twice $rep1$ and $rep2$). As a result, a total of 7 participants failed to rate the golden units, 3 of which were detected by the ITU-R BT.500-13 screening procedure. Thus, 11 participants were rejected (ITU-R outliers \cup Golden units outliers), i.e. 229/240 subjects remain.

After outliers removal, each stimulus is rated by 29 participants in the lab test and by at least 56 participants in the CS test. Only the scores of the screened participants will be used in the subsequent analyzes.

4.4.2 Resulting MOSs

In order to analyze the results of the CS experiment, we computed for each stimulus the Mean Opinion Score (MOS), rating scores averaged over all participants (Eq. 2.3). We compared them to those obtained from the lab experiment. Figures 4.2, 4.3 and 4.4 illustrate the results.

Figure 4.2 shows that the MOS results of the CS experiment strongly correlate with those of the lab test. Indeed, the (Pearson, Spearman's rank) correlation coefficients between CS's and lab's MOSs are (0.975, 0.954). Overall, no significant difference was found between the MOS means of the 2 studies (p-values = 0.191).

For both the lab and CS experiments, MOSs decrease as distortions strengths increase. However, we can notice some differences in the distribution of the scores and the use of the rating scale in the 2 experiments: the amplitude of the rating scale actually used by the participants in CS is reduced compared to that used in the lab experiment (see Figures 4.2 and 4.4). Overall, the stimuli rated in the lab scored lower than those in CS (see Figures 4.2 and 4.3). This effect is more visible for high strength distortions (strengths ≥ 3), meaning that the lab test participants were able to detect some distortions that the CS participants missed. Indeed, the lab test was conducted in a VR environment. Therefore, we believe that detecting visual quality losses of stimuli is easier in VR than on a 2D screen (CS), since VR headsets provide a bigger/wider Field of View (FoV) than a desktop setup and so the size in terms of visual angle of objects in VR is considerably larger than on screen (the stimuli size is approximately 37 and 18 degrees of visual angle in VR and on-screen, respectively).

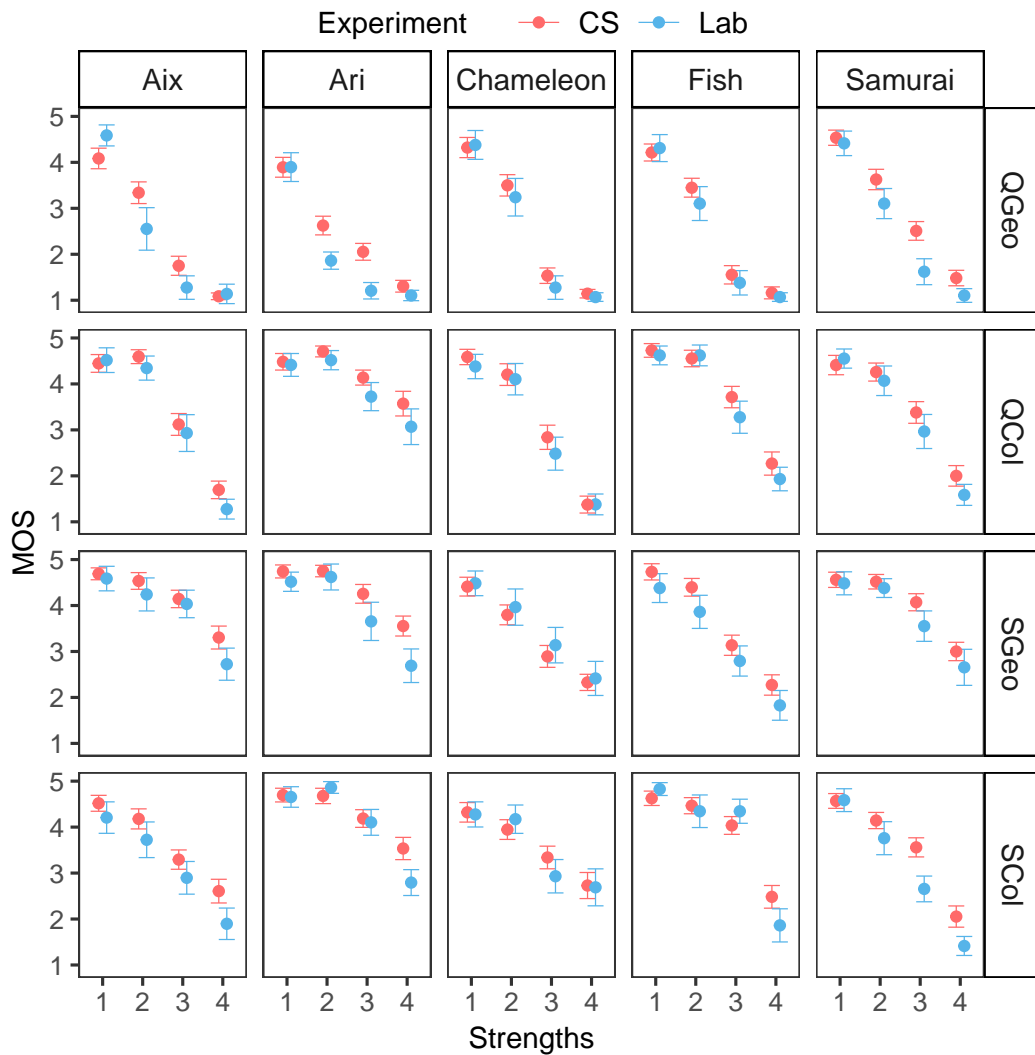


Figure 4.2: Comparison of the mean opinion scores of the lab and CS experiments, for all the stimuli. Results are grouped by source models and distortion types.

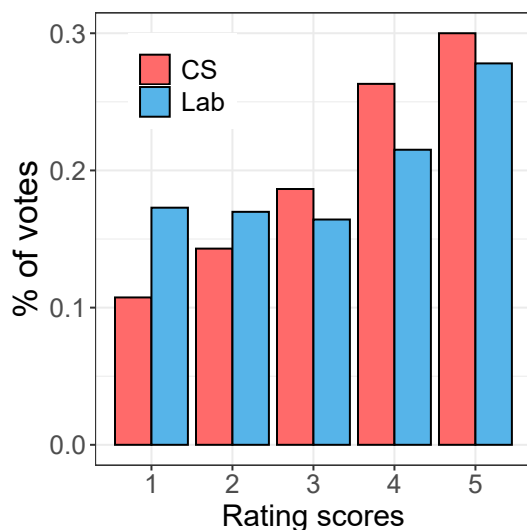


Figure 4.3: Score distributions for the lab and CS experiments.

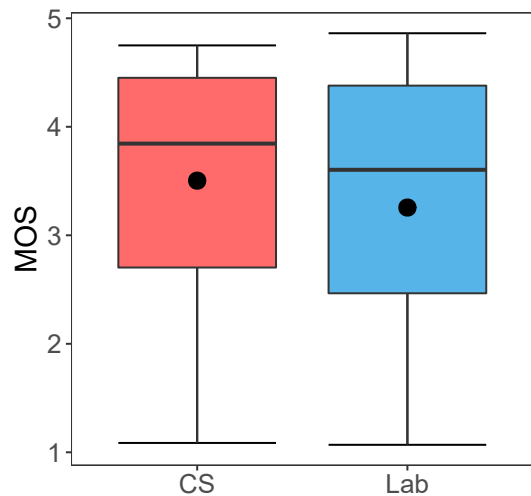


Figure 4.4: Boxplots of MOSs obtained for the lab and CS experiments.

4.4.3 Confidence intervals

Since participants in a CS experiment are not supervised, the evaluation of confidence intervals (CIs), i.e. the dispersion of individual scores, is particularly important. We therefore calculated the 95% CIs of the MOSs for the CS study and compared them to those obtained in the lab test. We assessed the evolution of the CIs width according to the number of ratings collected per stimulus (which is indirectly related to the number of participants involved in the test). Thus for each stimulus, we considered all possible combinations (without repetition) of N ratings and averaged the width of the CIs over all these ratings combinations. $N \in [1,29]$ and $[1,56]$ for the lab and CS test, respectively. Results are shown in Figures 4.5 and 4.6.

We can observe in Figure 4.5 that, for a given number of ratings, the CIs of the lab test are overall slightly larger than those of the CS test. Indeed, Figure 4.6 shows that the CIs of the lab test tend to be larger than those of the CS test in the range quality between *Annoying* and *Perceptible but not annoying* (i.e. $\text{MOS} \in [2,4]$), while the opposite is observed for high quality stimuli having a $\text{MOS} \approx 5$. The overall difference in the width of the CIs of the lab and CS tests is not significant according to the results of the t-tests (at a level of significance of 5%). Participants' agreement is further explored in section 4.4.5.

4.4.4 Accuracy of quality scores

We investigate, in this section, the accuracy (discrimination ability) of the CS test and compare it to that of the lab test. A more accurate test should yield in a larger number of pairs of stimuli whose quality can be considered different in a statistical test. As in Chapter 2, section 2.6.4, we conducted two-samples Wilcoxon tests between rating scores of each possible pairs of stimuli and computed the percentage of pairs of stimuli rated significantly different ($p\text{-value} < 0.05$) among all the possible pairs ($80 \times 79 / 2 = 3160$ pairs). Similar to the previous subsection, we evaluated the evolution of accuracy as a function of the number of ratings per stimulus: for a given stimulus and number of ratings N , we averaged the number of pairs of stimuli rated significantly different over the possible combinations of ratings. Figure 4.7 shows the results.

The ratings of the lab test are slightly more accurate than those of the CS test. For instance, at least 17 rating scores per stimulus (equivalent to 17 participants) are needed in the lab experiment to achieve accuracy with an overall level of 70%, whereas this number increases to 19 ratings in the CS experiment (corresponding to at least 76 participants in our actual CS test setup). Although the CIs tend to be smaller for the CS test than for the lab test (yet this difference is not significant, see section 4.4.3), the lab test produces slightly more accurate results, as the lab participants used the rating scale better (as shown in section 4.4.2).

Regarding the overall trend of the curves and despite this slight difference in the accuracy, the DSIS method offers stable performance in both lab and CS experiments.

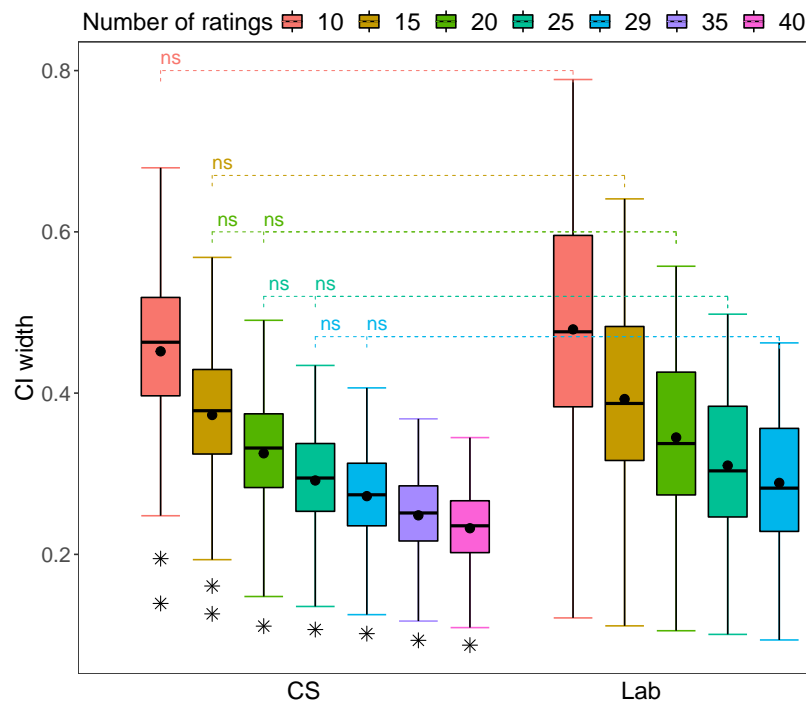


Figure 4.5: Variation of the width of Confidence Intervals (CIs) according to the number of ratings per stimulus in the CS and lab experiments. ns refers to not statistically significant ($p\text{-value} \geq 0.05$).

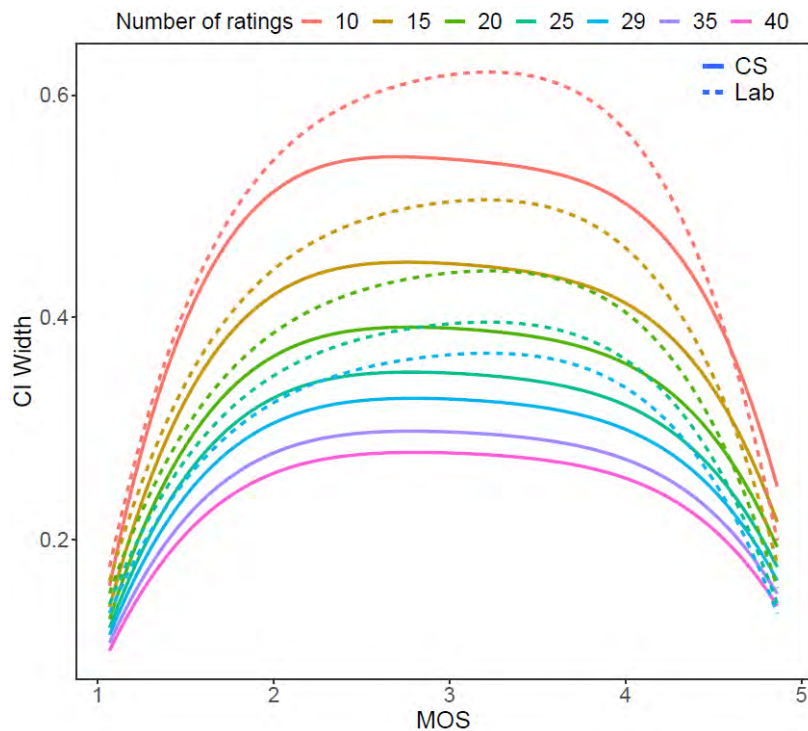


Figure 4.6: Confidence Intervals (CIs) as a function of MOSs obtained in the CS and lab experiments. The evolution of CIs is assessed according to the number of ratings per stimulus.

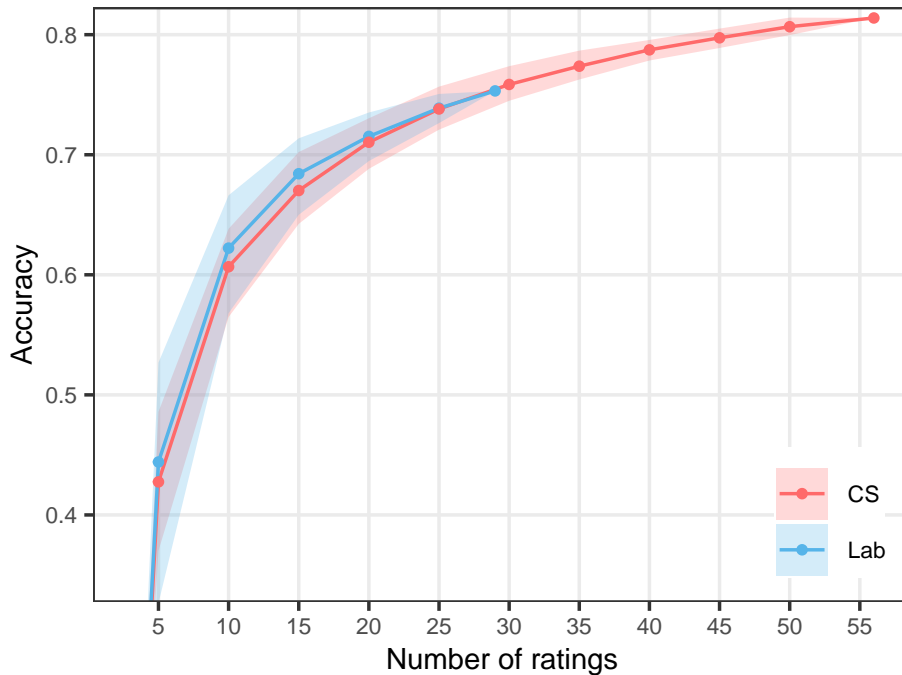


Figure 4.7: Variation of the accuracy according to the number of ratings per stimulus for the CS and lab experiments. The accuracy (y-axis) is defined as the percentage of pairs of stimuli whose qualities were assessed as statistically different. Curves represent mean values of these percentages and areas around curves represent 2.5th - 97.5th percentiles.

4.4.5 Participants' agreement and consistency

In this section, we evaluate the agreements between the participants. To do this, we evaluated the internal consistency of participants' data, as proposed by [59]. For each stimulus, we randomly split the participants who rated it into two equal size groups and computed the Spearman's rank correlation between the MOSs of the 2 groups. After 500 splits, the range of correlations was between 0.94 and 0.978 (with a mean of 0.963) in the CS experiment, and between 0.898 and 0.965 (with a mean of 0.934) in the lab experiment. Results show a high degree of inter-subject agreement in both experiments. This agreement is slightly lower in the lab experiment, possibly due to its more complex immersive viewing environment. This is in line with the CIs being slightly larger in the lab test (see section 4.4.3).

We then explored the inconsistency v_s of the participants, computed by the Maximum Likelihood Estimation (MLE) model [2] (described in Chapter 3, section 3.4.1). It flags inattentive participants who give random scores or participants who are distracted during a part of the test. The inconsistency of the CS participants cannot be directly compared to that of the lab participants, because each CS participant scored only 20 stimuli from the dataset (1 playlist out of 4) whereas one lab participant rated the entire dataset (80 stimuli). For that, we computed the inconsistency by playlist and not on the whole database: a lab participant is considered as four different participants, each belonging to a playlist and evaluating only the stimuli of that playlist.

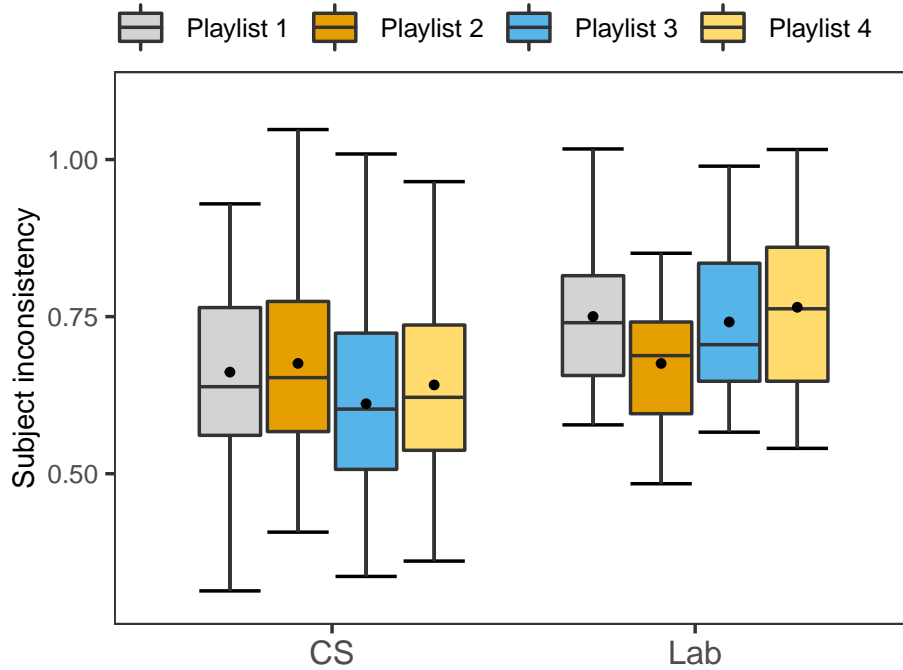


Figure 4.8: Boxplots of participants' inconsistency obtained in the CS and lab experiments.

Results, presented in Figure 4.8, show that overall the range of participant inconsistency reported in a the lab test is within that of the CS test. It is also interesting to notice that, the average inconsistency of lab participants is higher than that of CS participants, regardless of the playlists (and subsequently the stimuli). This is in line with the internal consistency being higher in the CS test.

4.4.6 Content ambiguity

As stated in [2] and explained in Chapter 3, some contents tend to be more difficult to rate than others. This is known as “Content ambiguity” and can be estimated by the MLE model (Eq. 3.4). Therefrom, we computed the ambiguity values of the 5 source models that constitute our dataset. Since the content ambiguity obtained by the MLE model depends on the participants [2], we considered the same number of ratings per stimulus ($N = 29$) for the lab and CS experiments in order to be able to compare their results. Thus, for the CS experiment, we randomly selected 100 combinations of 29 ratings for each stimulus. We averaged the ambiguity values over these combinations of ratings.

Figure 4.9 shows that all the source models are more ambiguous in the lab experiment than in the CS experiment. This may be because in VR, due to the larger FoV (as explained in section 4.4.2), participants can see more details and low-level features, which makes the task of assessing the differences between the reference and the distorted stimulus more difficult. The *Chameleon* model is associated with the highest content ambiguity in both experiments.

Content ambiguity is related to the dispersion of subjective scores (CIs). Therefore, we av-

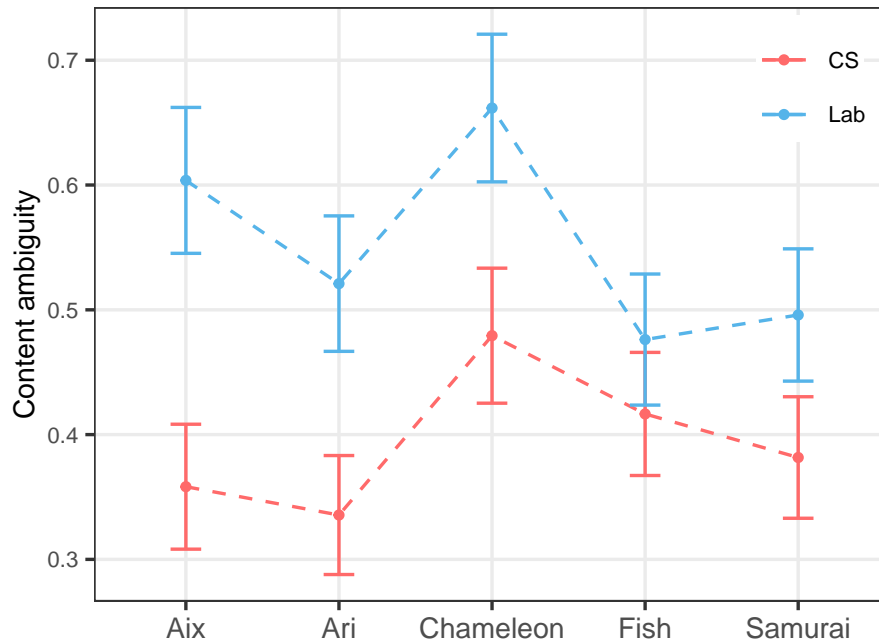


Figure 4.9: Content ambiguity of each source model associated its CI (calculated as described in [2]), for the CS and lab experiments.

eraged the CIs of the stimuli over the models. For both experiments, we obtained roughly the same shape of curves as in Figure 4.9 (see Figure A.7 in the appendix), meaning that models with high CIs are also associated with high ambiguity values. This is consistent with the results obtained in Chapter 3 section 3.4.3 (Figures 3.12.a and 3.15).

4.5 Discussion

In this chapter, we designed a subjective experiment for assessing the quality of 3D graphics in crowdsourcing (CS). The experiment was based on the DSIS method. Since in CS experiment, participants are unsupervised, we sought to impose controls on several aspects: (1) we used the Prolific platform which is more selective in recruiting and filtering participants than other similar platforms such as Mturk, Microworkers, etc. (2) We pre-screened participants during the recruiting process based on their reliability score on this platform. (3) We tried to control the viewer’s environment as much as possible (e.g. minimum screen resolution required, maintain a full screen mode, limited interactions to rating). Finally, (4) we combined different screening strategies to identify outliers. The results of this experiment were compared to results collected from a previous lab experiment conducted in a VR environment.

Results showed that under controlled conditions and with a precise/proper participant screening/filtering approach, a CS experiment can be as accurate as a lab experiment. Indeed, we obtained a high correlation between the mean opinion scores of the 2 experiments as well as a good agreement between the participants’ ratings (CIs). This agreement was slightly

lower in the lab test which can be due to the VR environment. We also observed that the amplitude of rating scale used in CS is smaller compared to that used in the lab test. Our findings corroborate previous studies, by [59,63,65], which also achieved high correlation with the lab results, and furthermore, they highlight the quality of the Prolific participants involved in our CS experiment, their reliability and seriousness despite the fact that the participants were not supervised. It is worth mentioning that the golden units approach we implemented seems to work well. In fact, as can be seen in Figure 4.10, all the outliers detected by the inspection of golden units (section 4.4.1) were found to have a high inconsistency according to the MLE model [2].

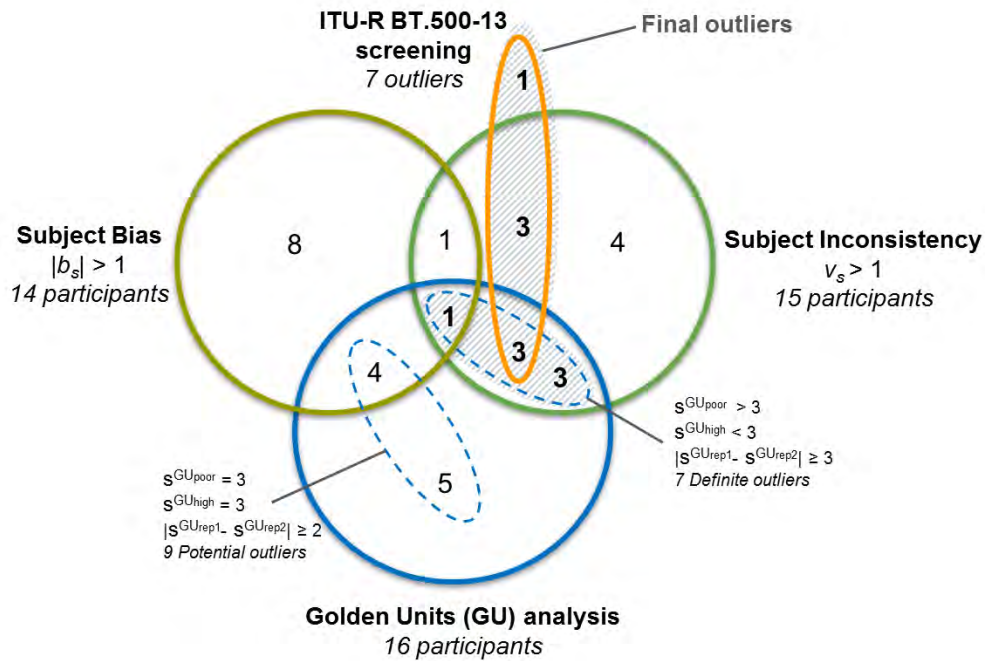


Figure 4.10: Venn diagram summarizing unreliable participants detected by the golden units inspection and ITU-R BT.500-13 screening procedure and those with high inconsistency and bias computed by the MLE model (v_s and b_s of Eq. 3.3). s^{GU} denotes the score assigned to the golden unit GU .

Regarding the experimental methodology, DSIS seems stable and adapted for evaluating the quality of 3D graphics in crowdsourcing. We believe that this is due to the fact that this method presents explicit references which simplifies the task and makes it easier: according to a short questionnaire asked at the end of our CS experiment, 89% of the participants found the experiment’s instructions clear and 94% rated the work as easy. The results of the questionnaire are presented in Figure 4.11.

In term of time-effort, it took us 12 hours to collect all the data in the CS experiment (5520 quality judgments collected), compared to 36 hours in the lab experiment (2400 quality judgments collected). Crowdsourcing is quite faster: as mentioned in the introduction of this chapter, CS is frequently used to evaluate large datasets, which requires weeks of lab evaluations. However, building and designing our CS experiment tool was a time-

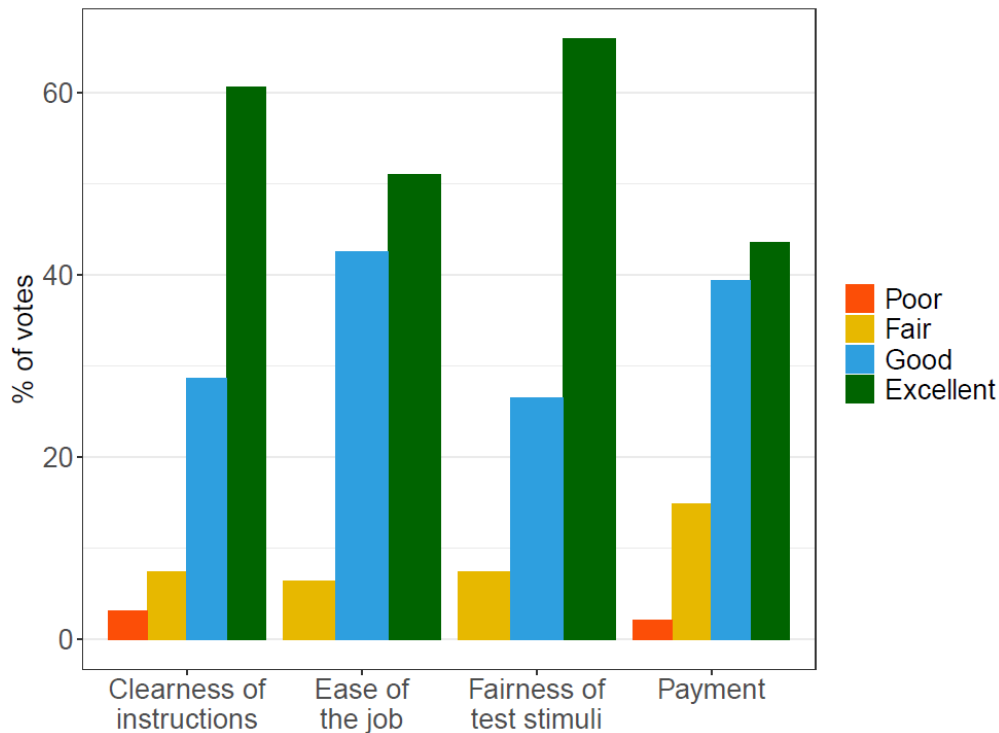


Figure 4.11: Results of the questionnaire asked at the end of the CS experiment.

intensive task and required significant software development effort (both technical and of conceptual challenges). Furthermore, as stated earlier, the CS experiment must be short to keep participants accurate and consistent, making our CS test require additional programming effort to implement a tool that evenly divides the dataset into batches/playlists and assigns a different batch to each participant.

4.6 Conclusion

In this chapter, we investigated how crowdsourcing could be used for quality assessment of 3D graphics. CS experiments based on the DSIS method appear to be an appropriate way not only to quickly collect a large amount of realistic quality judgments, but also to provide (when combined with appropriate participant screening strategies) accurate and reproducible results. Based on these promising results, we conducted a large-scale quality assessment experiment in CS, which resulted in the largest dataset of textured 3D meshes to date. The dataset and the CS experiment are presented in the next Chapter (Chapter 5). We expect our findings to help the scientific community when designing subjective quality assessment studies in CS. Further experiments are still needed to find the best compromise between the number of stimuli, the duration of the test and the accuracy of the results. Also, it would be very interesting to repeat the lab experiment, this time using a desktop setup, to clearly isolate the effects of VR.

Chapter 5

Subjective Quality Assessment of a Large-Scale Textured 3D Mesh Dataset in Crowdsourcing

Development of deep learning-based quality metrics for 3D meshes relies on the richness of available datasets. As we discussed in Chapter 1 (section 1.1.2 and Table 1.1), there is a lack of large-scale 3D meshes datasets, especially those with color attributes. The existing datasets are rather small: they only have a few hundreds of distorted meshes, which is not sufficient to drive deep learning metrics. The difficulty of establishing large-scale quality assessment datasets for 3D meshes comes from the difficulties of (1) obtaining enough source models, (2) ensuring diversity of color, geometric and semantic characteristics and (3) annotating the entire (large-scale) dataset via subjective experiment alone.

In this chapter, we produce the largest quality assessment dataset of textured 3D meshes to date. 55 source models were carefully selected and characterized. Each model was corrupted by combinations of 5 real-world distortions, related to compression and simplification, applied on the geometry and texture of the meshes. In total, 343750 distorted meshes are generated. To address the aforementioned challenges, we devised a set of measures to quantitatively characterize the geometric, color, and semantic complexity of 3D models. A large-scale subjective experiment was conducted in crowdsourcing to annotate a subset of 3000 stimuli. The quality scores of the remaining stimuli were predicted using a quality metric trained and tested on the subset of annotated stimuli. This dataset allowed us to analyze the impact of the distortions and model characteristics on the perceived quality of textured meshes.

The main contributions of this chapter are as follows:

- We provide a large-scale dataset of textured meshes, with more than 343k stimuli, of which 3000 are associated with mean opinion scores (MOSs) derived from subjective experience and the rest with predicted quality scores (pseudo-MOSs). To the best of our knowledge, it is the largest quality assessment dataset of textured 3D meshes to date.
- We propose three measures, based on spatial information and visual attention com-

plexity, to quantitatively characterize the geometric, color and semantic complexity of 3D models.

- We provide an in-depth analysis on the effect of each distortion and their combinations on the perceived quality of textured meshes. We also determine which distortions affect the quality scores the most.
- We evaluate the influence of the complexity of the model’s geometry, color and texture seams on the perception of distortions.

This chapter is organized as follows: section 5.1 describes our dataset and the set of measures we propose for 3D content characterization. Section 5.2 details the subjective experiment and the process adopted to select the subset of test stimuli. Section 5.3 provides the results, while section 5.4 presents a use case of the dataset. Sections 5.5 and 5.6 present concluding remarks.

5.1 Dataset generation

We produced a large-scale textured meshes quality assessment dataset with 343750 distorted meshes derived from 55 source models each with 6250 distorted versions. Distortions represent combination of Level of Details (LoD) simplification, and texture and geometry compression. The dataset covers a wide range of geometric, color and semantic characteristics. Indeed, each source model has been carefully selected and characterized as will be shown in subsection 5.1.2 To the best of our knowledge, it is the largest quality assessment dataset for textured 3D meshes at present.

5.1.1 3D source model selection

We have collected 55 textured 3D models from SketchFab¹. The selection was done manually and carefully to collect high quality textured meshes with creative commons licenses (released with permissive licenses) so that anyone using this dataset can publicly share his results. Table 5.1 lists the models, their number of vertices and semantic category, while Figure 5.1 illustrates them.

Some models required cleaning to repair topological and geometrical defects (zero-area triangles, non-manifold geometry, holes, etc.). Furthermore, we converted all the models to a unique format: the meshes are provided as OBJ (+ the material file), and the textures as JPEG images of 2K resolution (normalized texture size: 2048×2048). The textures encode surface colors (i.e. diffuse map); other information such as surface normals, roughness, and ambient occlusion are ignored/discarded. For models with multiple texture images, these were baked into one single image.

¹<https://sketchfab.com/features/free-3d-models>



Figure 5.1: The 3D graphic source models constituting our database.

Table 5.1: List of source models, their number of vertices and semantic category.

Model ID	#Vertices	Semantic category	Model ID	#Vertices	Semantic category	Model ID	#Vertices	Semantic category
#1	357364	Animal statue	#20	185416	Sculpture	#39	100890	Sculpture
#2	123189	Human statue	#21	149906	Apparel	#40	124936	Plant
#3	395490	Furniture	#22	74662	Animal	#41	74819	Animal
#4	99984	Machine	#23	20144	Plant	#42	150498	Decoration
#5	198683	Decoration	#24	297988	Decoration	#43	62989	Decoration
#6	258490	Human statue	#25	151002	Sculpture	#44	36204	Animal
#7	483746	Human statue	#26	98763	Machine	#45	650778	Sculpture
#8	250723	Sculpture	#27	77027	Mean of transport	#46	4999	Food
#9	189633	Animal	#28	306933	Animal skeleton	#47	110819	Book
#10	109929	Machine	#29	119038	Food	#48	56902	Decoration
#11	134472	Electronic device	#30	114145	Decoration	#49	92223	Mean of transport
#12	17560	Musical instrument	#31	16803	Musical instrument	#50	260670	Decoration
#13	75950	Animal skeleton	#32	358684	Sculpture	#51	60710	Mean of transport
#14	209609	Animal	#33	155931	Animal statue	#52	635206	Animal statue
#15	77924	Animal statue	#34	486850	Building	#53	150000	Decoration
#16	669346	Animal	#35	150006	Bust	#54	299976	Animal
#17	393652	Animal	#36	249439	Mean of transport	#55	125002	Animal
#18	27611	Food	#37	130016	Sculpture			
#19	308944	Sculpture	#38	304435	Bust			

5.1.2 Content characterization

Our goal is to create a diverse, realistic, and challenging dataset for objective quality metrics (especially those based on deep-learning approaches), able to show the utility of these metrics in real world use cases. However, creating a high-quality and diverse dataset is not a trivial task since its models must cover a wide range of color, geometric and semantic characteristics. Moreover, these characteristics must be assessed by quantifiable criteria and measures.

The characterization of 3D graphical models is not as straightforward/simple as it seems due to the multimodal nature of these data (geometry, color/texture and material information).

In the field of images and videos, the content characterization is usually based on the Spatial perceptual Information (SI) [10]. Indeed, SI was demonstrated to be suitable for image characterization and content classification in image and video quality assessment applications [143]. For instance, a number of video quality metrics use SI or closely related measures for this purpose [144]. As detailed in ITU-P910 [10] and in Eq. 5.1 the SI computation is based on the Sobel filter. It is an indicator of edge energy, i.e. it emphasizes regions of high spatial frequency that correspond to edges. Images (I) are converted to gray-scale, then filtered with horizontal and vertical Sobel kernels. The standard deviation (std) over the pixels of the Sobel-filtered images is finally computed.

$$SI = std_{space}[Sobel(I)] \quad (5.1)$$

Since there is no standard on how to characterize 3D content and inspired by the SI of images, we proposed two new measures to characterize the color and the geometry of textured 3D models. We also introduced a third measure, based on the visual attention complexity (presented later in this section), to characterize the models regarding the se-

semantic aspect.

Our approach is based on rendered images. First, we rendered the objects under their main viewpoint shown in Figure 5.1. It is a viewpoint, perceptually chosen for each model, that covers the most geometric, color and semantic information. Next, we compute the three proposed measures on the rendered images of the models as described in the following.

A. Geometric characterization

To enhance the geometry of the model, we considered a rendering that only takes into account the shading of the model without any of its texture information. Then, we computed the SI of the rendered snapshot. We thus obtain SI_{Geo} an objective indicator/measure that characterizes the geometry of the 3D model (assess its geometric complexity). This process is represented in Figure 5.2.a.

B. Color characterization

To assess the textural/color characteristics, we considered this time a rendering without any shading; just the colors from the texture are displayed. We applied the Sobel filter on the rendered snapshot and removed the silhouette detected by the filter because it rather reflects geometry information than color information, and is already taken into account by SI_{Geo} . The color characterization is thus defined by the SI_{Col} measure computed on the obtained filtered image. This approach is illustrated in Figure 5.2.b.

C. Semantic characterization

Besides determining the semantic categories of the models as we did in Table 5.1, we consider an objective indicator that quantitatively assesses their semantic complexity. Semantic information cannot be characterized by the SI. We therefore employed the Visual Attention Complexity (VAC) measure, recently proposed in [145], which is perfectly adapted for this task.

The VAC is an indicator linked to the visual saliency information. It is related to the semantics of the object. It consists in evaluating the dispersion of salient regions of the rendered 3D model. Rendered images of 3D models associated with low VAC scores indicate that there are highly salient regions that attract human visual attention (i.e., focused gaze behavior), while images with high VAC scores indicate that the saliency is diffused and not focused on one region (i.e., overall gazing behavior).

The VAC is computed as follows and illustrated in Figure 5.2.c: First, we render the 3D model with shading and texture attributes. The model is displayed in its main viewpoint. Once the snapshot of the final rendered object is generated, we compute the saliency map -which represents the probability of gazing at a given pixel- using the “Salicon” computational model as recommended in [145]. Finally, we compute a conditional entropy on the saliency map. Thus, we obtain the VAC_{score} which characterizes the visual attention

complexity of the 3D model and is closely linked to a semantic value. The detailed computation of the VAC can be found in Chapter 6 (section 6.3.1) and in [145].

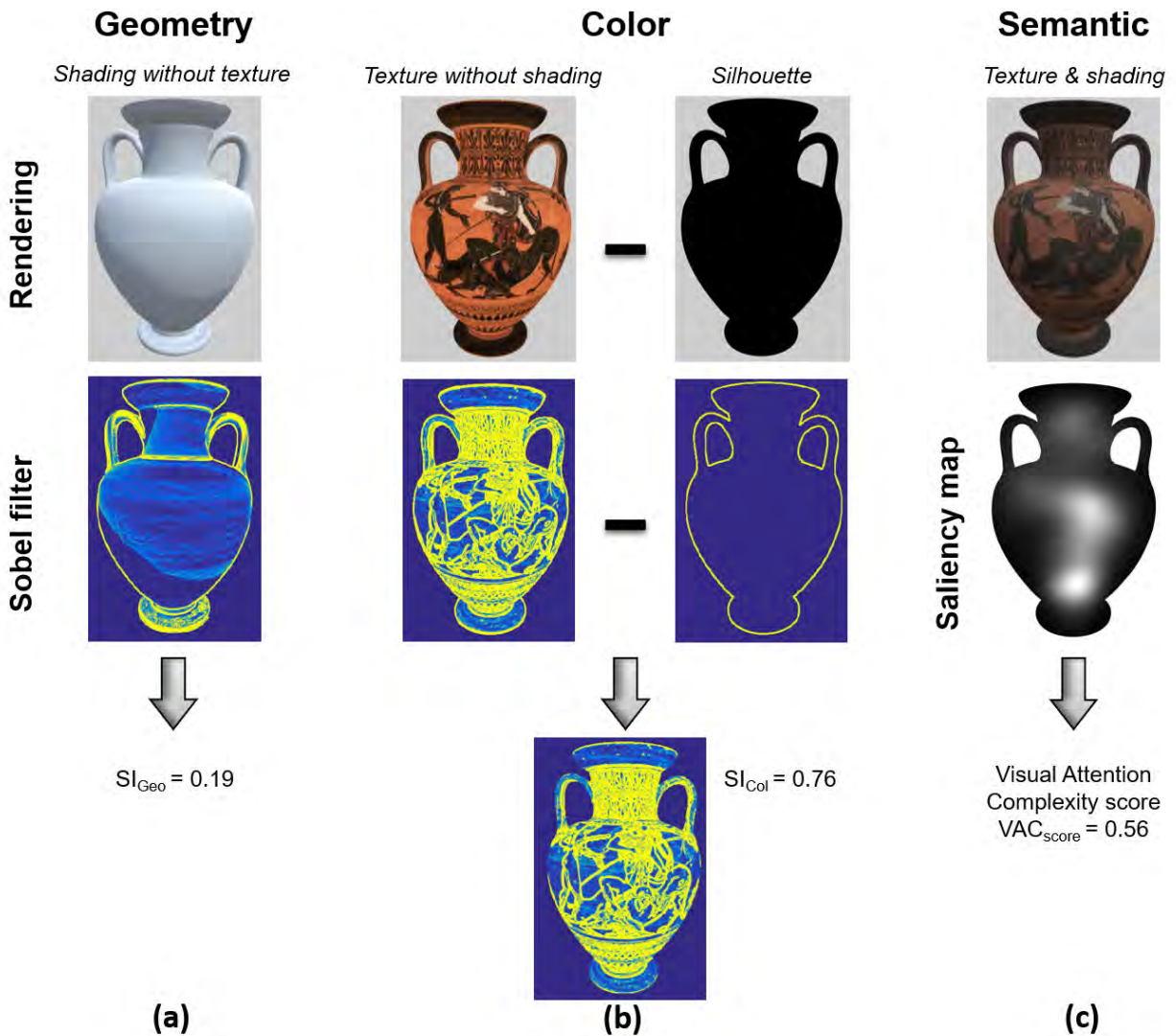
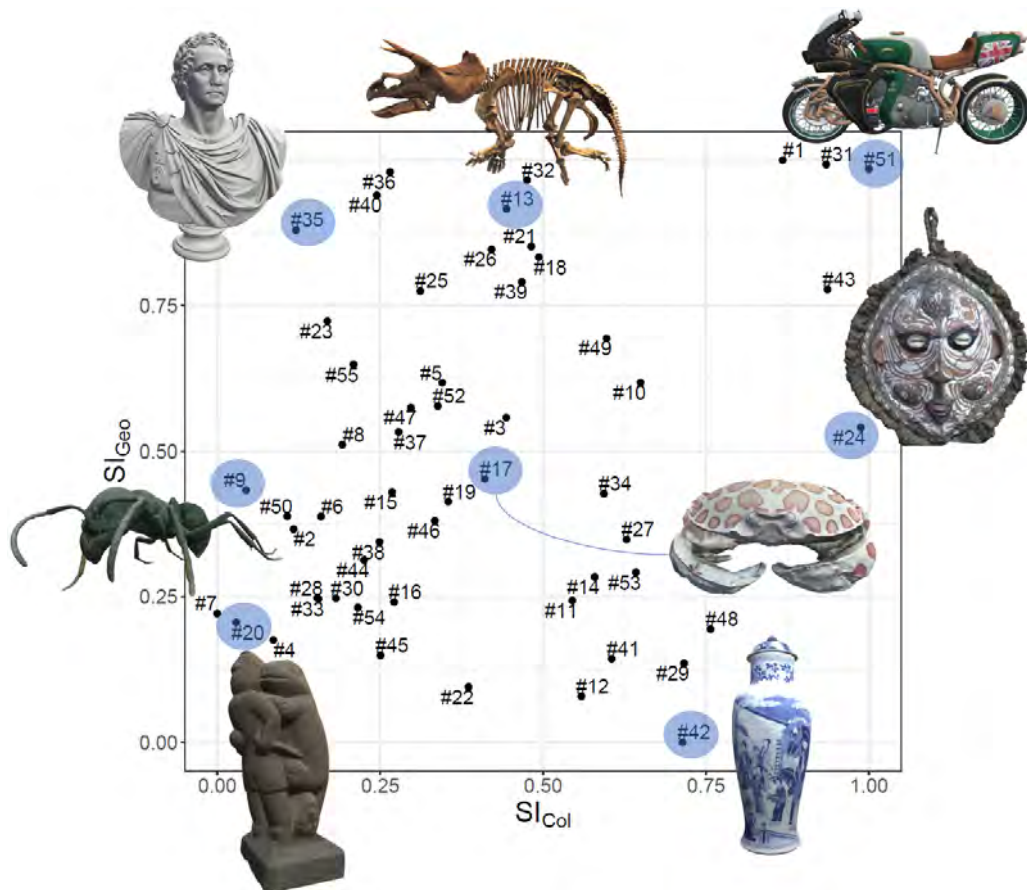


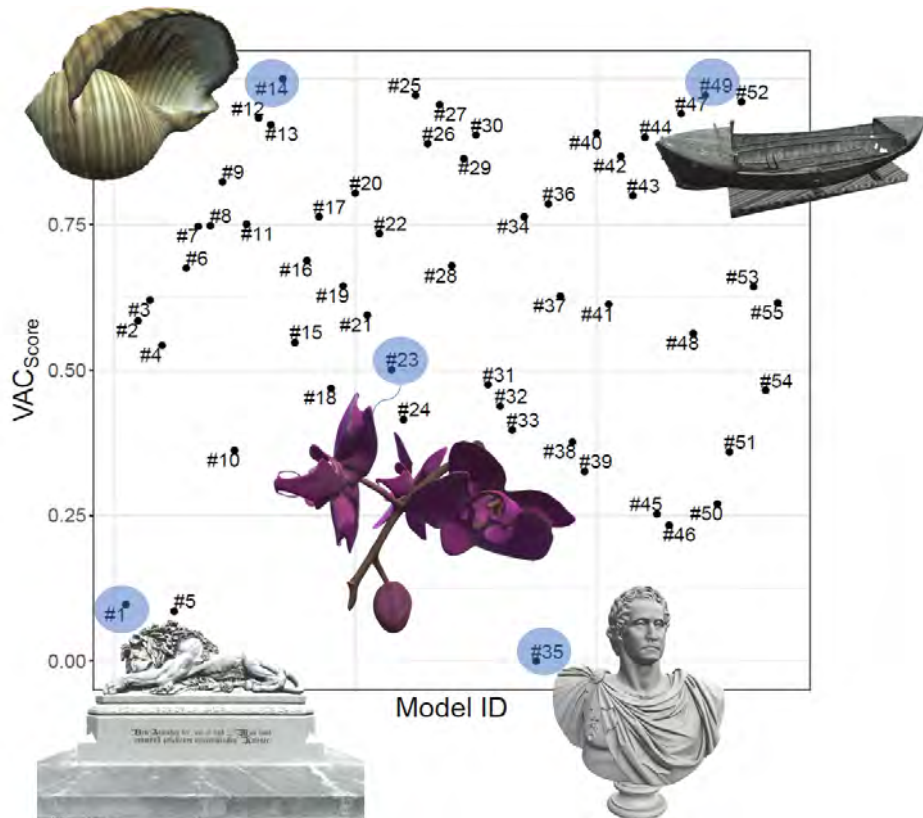
Figure 5.2: Characterization of the geometry, color, and semantic of textured 3D models.

As a result, we have proposed three measures to quantitatively characterize the geometric, color, and semantic complexity of 3D models. We applied these measures on our 55 selected models. The measures were computed on snapshots of the models having the same resolution. The obtained results are shown in Figure 5.3.

As can be seen, the source models of our dataset have various geometric, color and semantic characteristics. Figure 5.3.a shows a good distribution on the SI_{Geo}/SI_{Col} plane: Along $SI_{Geo} = 0$ axis (at the bottom of the plot) are found the models having low geometric information (such as models #42, #22, #41 and #20). Near the top of the plot are found models with rich geometric and topological information (such as #35, #13, #26 and #51). Along $SI_{Col} = 0$ axis (at the left edge of the plot) are found models with minimal color details (such as models #9, #20, #23 and #35). Near the right edge of the plot are found models with the most color details (such as #42, #24, #43 and #51). Thus, models



(a)



(b)

Figure 5.3: (a) Geometry and color spatial information and (b) the visual attention complexity for our source models.

located at the right-bottom corner of the SI_{Geo}/SI_{Col} plane, such as the model #42, present a rich texture with very low geometric complexity. On the contrary, the model #35 in the top-left corner of the plot is monochrome but has sharp edges and many small geometric details to depict its hair and face.

Regarding the semantic characteristics, Figure 5.3.b shows that our models cover a large range of visual attention complexity. For example, model #1 and #35 have a low VAC_{score} indicating that these models contain highly salient regions that are the epigraph (the writing more generally) and the face, respectively. The visual attention of the participants will probably be focused on these regions. On the other hand, there are no particularly salient regions on models #14 and #49. These models exhibit low visual attention complexity. Participants will probably have an overall gazing behavior.

The set of measures we proposed reveals the particularity/peculiarity of 3D models in terms of geometry, color and semantics. These measures are extremely fast to compute and work consistently for both coarse and dense 3D data (regardless of the 3D data representation and the number of vertices). The main drawback of these indicators is that they are view-dependent (depend on the selected viewpoint of the 3D model). This problem can be overcome by computing these measures on several snapshots taken from different viewpoints of the model.

5.1.3 Distortions

From the 55 source models, we created 343750 distorted versions generated from combinations of 5 real-world distortions. The distortions come from lossy compression and simplification algorithms applied on the geometry and on the texture.

- **Level of Detail (LoD) simplification:** a surface simplification algorithm based on iterative edge collapse and quadric error metric. This algorithm takes into account both geometry and texture and preserves UV parametrization [146]. We generated 10 levels of simplification ($LoD_{simpL} \in [L1, L10]$) by uniformly reducing the number of mesh faces so that the mesh of the most degraded level ($LoD_{simpL} = L10$) has around 2000 faces (regardless of the source model). Thus, $\Delta_{simpL} = (NbFaces_0 - NbFaces_{min})/10$ where Δ_{simpL} is the number of faces removed at each LoD_{simpL} level, $NbFaces_0$ is the number of faces of the source model, and $NbFaces_{min} = 2000$.
- **Quantization:** we uniformly quantized the position of the vertices as well as the coordinates of the texture using Draco², an open-source library for compressing and decompressing 3D geometric meshes and point clouds. To generate the compressed meshes, we considered 5 levels for each attribute:
 - The quantization bits for the position attribute qp range from 7 to 11 bits ($qp \in [7, 11]$). It represents a sub-sampling of the geometry.
 - The quantization bits for the texture coordinates attribute (a.k.a UV map) qt are between 6 and 10 bits ($qt \in [6, 10]$). The model UV map represents the

²<https://github.com/google/draco>

parametrization defined to map the texture data onto the model surface. Quantizing the texture coordinates is a subsampling of the UV map.










- **Texture map sub-sampling:** we reduced the size of the textures by resampling using the Lanczos filter which is a low pass filter. We generated 5 texture sizes (T_S): 2048×2048 (the original size), 1440×1440 , 1024×1024 , 712×712 , 512×512 .
- **Texture compression:** we used the JPEG compression which is the most commonly used algorithm for lossy 2D image compression. We selected 5 texture qualities (T_Q) obtained by varying the compression level: 90 (the best quality considered but the least effective compression), 75, 50, 25, 10 (the lowest texture quality and the highest compression).

We note that for each distortion type, the degradation levels were selected to cover a range of high, medium and low quality of distorted meshes.

By combining/mixing all geometry and texture distortions, we obtained $10 LoD_{simpl} \times 5 qp \times 5 qt \times 5 T_S \times 5 T_Q = 6250$ distortions per model, a.k.a. Hypothetical Reference Circuits (HRCs) according to [147]. HRCs denote the processing operations applied to the source models to obtain the testset.

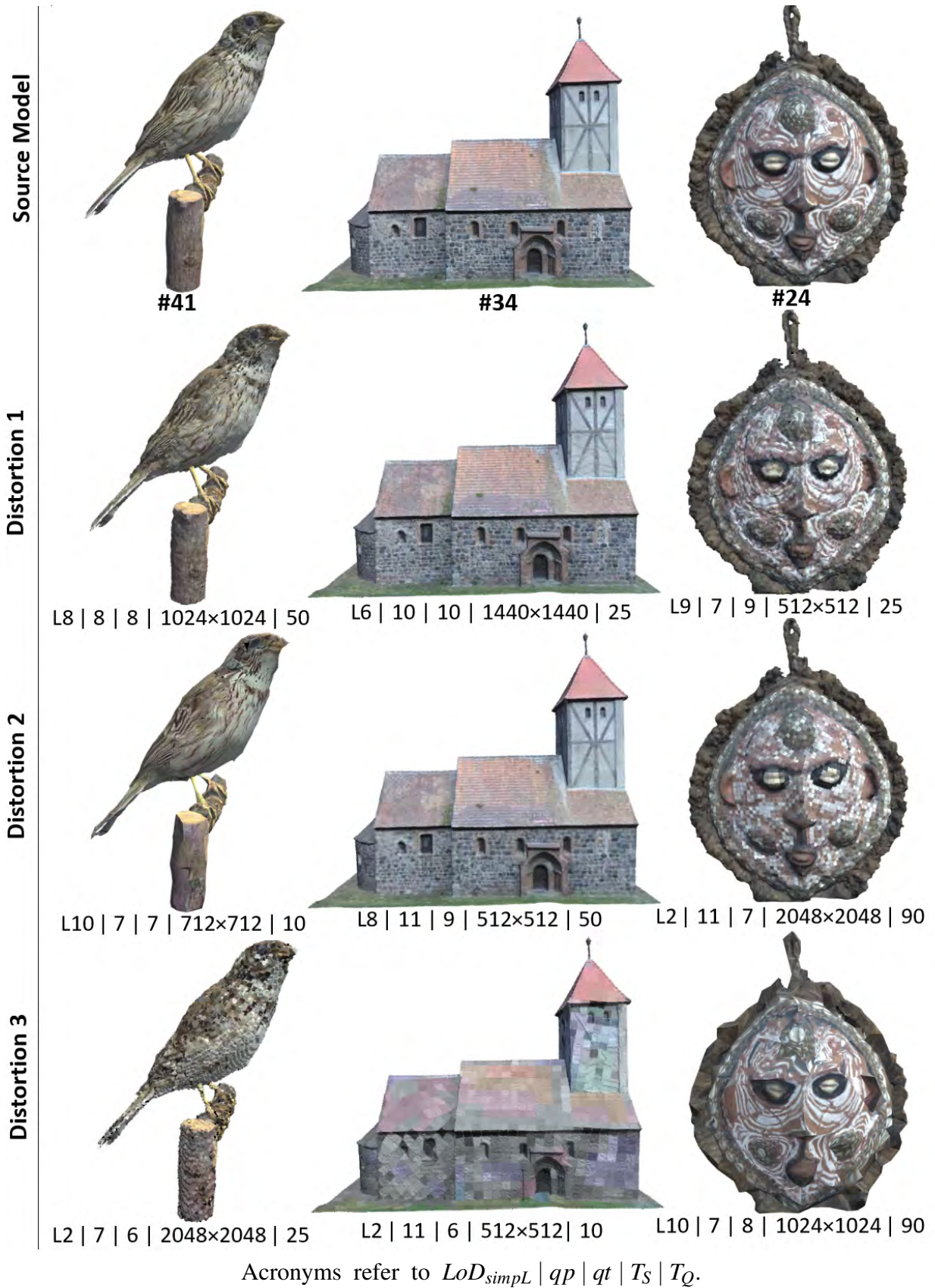
Each HRC is associate with a size (in Bytes), resulting from the compression of the source model with the corresponding distortion parameters (using JPEG for the texture and Draco encoder for the connectivity, geometry and UV maps). Thus, the size of a stimuli (in Bytes) is equal to the sum of the size of its compressed texture and its compressed 3D model.

Our dataset thus contains 343 750 distorted stimuli (55 source models \times 6250 HRCs) that span a great diversity in visual contents and distortions. Figure 5.4 shows some examples of distorted stimuli along with their distortion parameters.

Source Model	#18	#35	#47
Distortion 1	 L3 11 8 2048×2048 90	 L8 11 9 2048×2048 90	 L4 10 9 1440×1440 25
Distortion 2	 L6 7 9 512×512 10	 L1 9 7 1440×1440 10	 L7 10 7 512×512 90
Distortion 3	 L10 11 8 512×512 90	 L4 10 6 512×512 25	 L10 8 8 712×712 75

Acronyms refer to $LoD_{simpL} | qp | qt | T_S | T_Q$.

CHAPTER 5. SUBJECTIVE QUALITY ASSESSMENT OF A LARGE-SCALE TEXTURED 3D MESH DATASET IN CROWDSOURCING



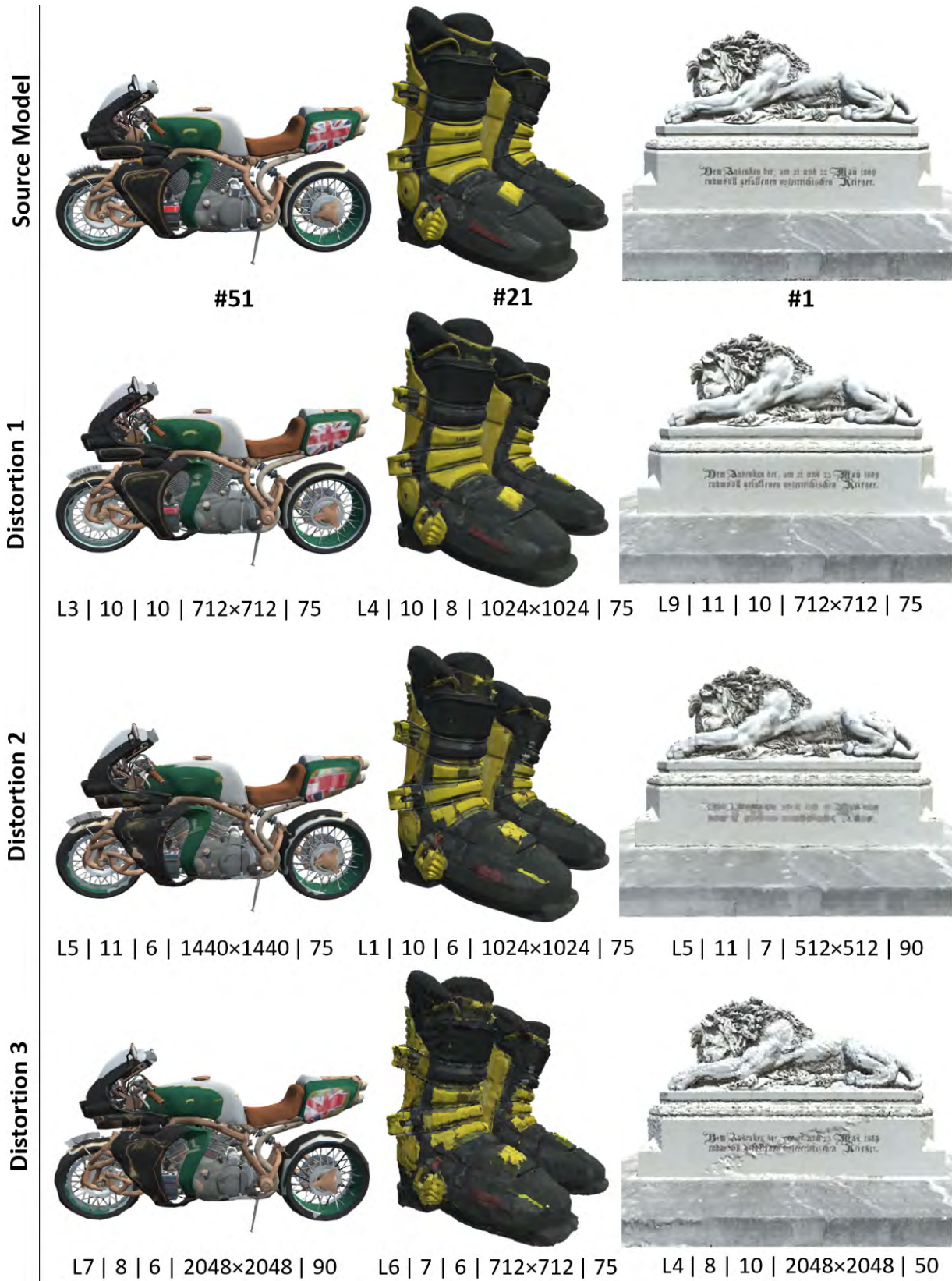


Figure 5.4: Some examples of distorted models. Acronyms refer to $LoD_{simpl} | qp | qt | T_S | T_Q$.

5.2 Subjective experiment

We conducted a very large-scale crowdsourced subjective quality assessment experiment, wherein 4513 participants were involved to rate the perceived quality of a subset of 3000 textured 3D meshes carefully selected. This section describes our extensive online subjective study.

5.2.1 Test stimuli selection

As we noted in section 5.1.3, our dataset contains more than 343k stimuli. Participants cannot be asked to rate the quality of such a large amount of data. We therefore had to select a subset of stimuli to rate in the subjective experiment. This was not a trivial task since the selected subset had to contain all the source models, as well as a large variety/diversity of HRCs (combinations of distortions created). In addition, the subset must evenly/equitably cover the entire range of quality (from imperceptible to very annoying distortions) to have a balanced representation of the visual quality. Least but not least, we want this dataset to be challenging for objective quality metrics.

We selected 3000 stimuli from 343750 (which represents about 0.9% of the total dataset). To do so, we developed the following approach based on several selection criteria and constraints.

In order to get a subset of stimuli that evenly covers the entire quality range, we predicted the MOS (Mean Opinion Score) of all the 343750 stimuli, using existing objective quality metrics that we calibrated on an existing/previous subjectively-rated dataset. Here are the details: we used our previous dataset of meshes with vertex colors, presented in Chapter 3). We fitted two logistic regression models (mapping functions) between the MOSs of this dataset and the following two objective quality metrics: (1) HDR-VDP2 [108] since HDR-VDP2 provided the best performance among the Image Quality Metrics (IQMs) tested on this dataset (see section 6.2.6 of Chapter 6) and (2) LPIPS (Learned Perceptual Image Patch Similarity) [4] since it is a commonly used IQM based on pre-trained CNN representations with many successful applications [148, 149]. Using two metrics (instead of one) will allow us to sample stimuli for which the metrics do not agree on their quality, resulting in a more challenging subset of stimuli (as explained a little later in this section). Next, to annotate our dataset of textured 3D meshes, we computed HDR-VDP2 and LPIPS on snapshots of the stimuli rendered from their main viewpoints (defined in section 5.1.2), and then predicted their MOS using the regression models. As in [150, 151], we refer to the predicted MOSs as Pseudo-MOSs. Thus, we obtained 2 pseudo-MOS values per stimulus (pseudo-MOS_{HDRVDP} and pseudo-MOS_{LPIPS}).

To ensure a good and equitable coverage of the whole visual quality range and to get a subset of challenging stimuli, we regularly sampled the plane formed by the 2 pseudo-MOSs, as shown in Figure 5.5: considering HDR-VDP2 as the pivot metric, we selected in each quality range the same number of stimuli by applying uniform sampling in this area. The sampling was not done randomly. It respects several constraints. Indeed, the 3000 test stimuli are selected so as to ensure an equal/even distribution between: (1) all

source models (e.g., as many degraded stimuli for model ID_i as for model ID_j) and (2) all levels of each distortion (e.g, almost as many stimuli are selected with a $qp = 7$ as those with a $qp = 8$).

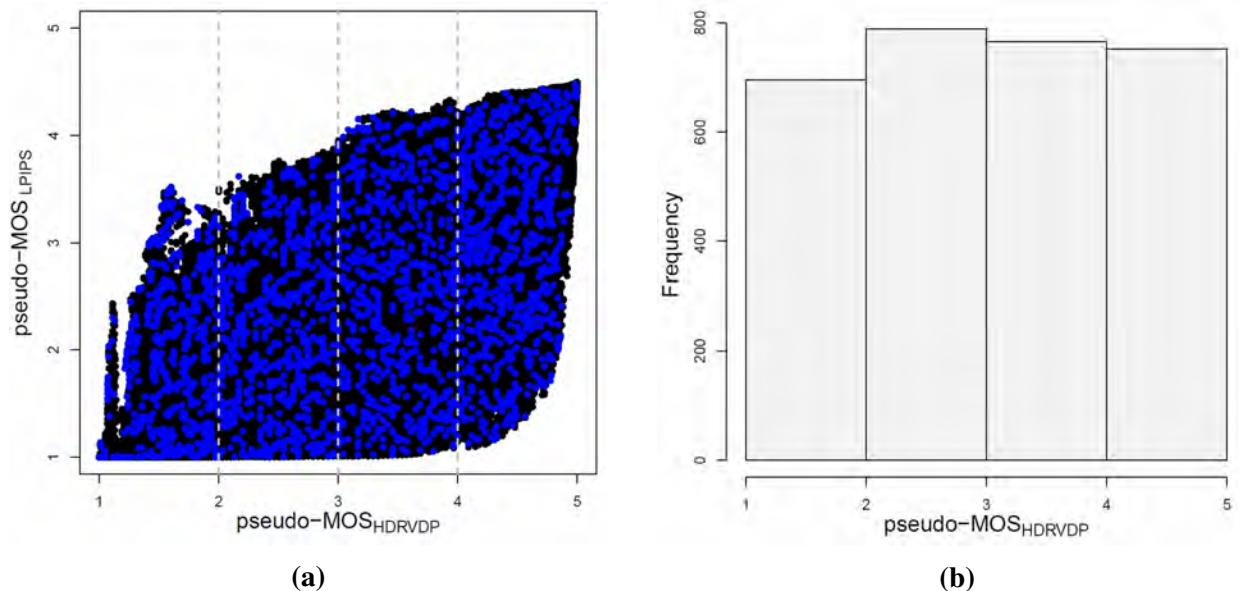


Figure 5.5: (a) Selection of the test stimuli by uniformly sampling the plane formed by 2 pseudo-MOSs. The black dots refer to the pseudo-MOS values of all stimuli in the dataset, while the blue dots refer to those selected for the subjective study. (b) The pseudo-MOS_{HDRVDP} distribution of the 3000 test stimuli.

5.2.2 Rendering

To adequately assess the visual quality of 3D content, it is important that the object moves [16] so that the observer can see the whole object and not focus on one single viewpoint. Chapters 2 and 3 provide further explanation.

Thus to avoid manual selection of multiple relevant viewpoints for each model, we animated all models in our dataset with a full rotation (360 degrees) around their vertical axis.

We kept the same setting as our preceding experiments (described in the previous chapters): we rendered the dynamic test stimuli, without shadows and under a directional light, in a neutral room (light gray walls) at a distance fixed to 3 meters from the camera. Their material type complied with the lambertian reflectance model.

Since the experiment is conducted in crowdsourcing, we adopted the same framework/setting as the CS experiment in Chapter 4: we generated videos of the final rendered dynamic stimuli (rotating stimuli) in order to limit the participant’s interactions with the 3D objects since we do not have full control over the participant’s test environment. The only interaction required by the participant is the selection of the score when rating.

The videos were all in 650×550 resolution (so that the videos of the source and degraded models fit simultaneously on a screen with a minimum resolution of 1920×1080) with a

frame rate of 30 fps and encoded using H.264 encoder (mp4 container) at a bitrate of 5 Mbps to ensure imperceptibility of compression artifacts. Videos are 8 seconds long, which is the time it takes for models to complete the full rotation.

5.2.3 Experimental environment

To design and implement our experiment, we adopted the same setup and platform used to implement the crowdsourcing (CS) experiment of the previous chapter (Chapter 4, section 4.3), which has shown its effectiveness: indeed, the latter achieved the accuracy of a lab test and produced reliable and reproducible results (see section 4.4).

Thus, we designed the experiment based on the DSIS method, in which observers see the source model/reference and the same model impaired side by side and rate the impairment of the second stimulus in relation to the reference using a five-level impairment scale, displayed after the presentation of each pair of stimuli. This method has proven to be the most stable and accurate for evaluating the quality of 3D graphics (Chapter 2).

To conduct the experiment, the web-based platform developed in Chapter 4 was used. It is suitable for presenting videos according to the DSIS method. Only a web browser with an MPEG-4 decoder is required to run the experiment; no other software needs to be installed. The platform first checks the compatibility of the participant's device: minimum screen resolution of 1920×1080 , page zoom level, maintain full screen mode throughout the experiment. The test instructions are then displayed to the participant with a progress bar, at the bottom of this page, showing the status of the loading process of all the video pairs that will be used in the test. This way, the videos of the source and distorted models will be played simultaneously during the test, without any latency or unintended interruptions. When the loading is completed a start button appears leading to the test.

Figure 5.6 illustrates the graphical interface of our subjective experiment.

The pairs of videos of the distorted 3D objects are displayed in a random order to each participant. Participants cannot replay the videos or give their score until the videos have been played completely. There is no time limit for voting and videos of the stimuli are not shown during that time. At the end of the experiment, participants will receive unique codes allowing them to get their remuneration.

5.2.4 Creation of test sessions

As explained in Chapter 4, a CS experiment should be kept as short as possible. We therefore divided our subset of 3000 test stimuli into 100 playlists. Each participant rates one playlist, i.e., only 30 stimuli. This way, we stay within the ranges/margins of stimuli to rate and duration of the experiment in Chapter 4, which represents a pilot study for this large-scale experiment. The test stimuli were fairly distributed among the playlists so that each playlist contains a maximum diversity of source models (a source model is repeated a maximum of 2 times in a playlist). Additionally, each playlist spans the full range of distortions and all playlists have nearly the same pseudo-MOS distribution.

As in chapter 4, we injected 3 Golden Units (GU, a.k.a trapping stimuli) into each playlist

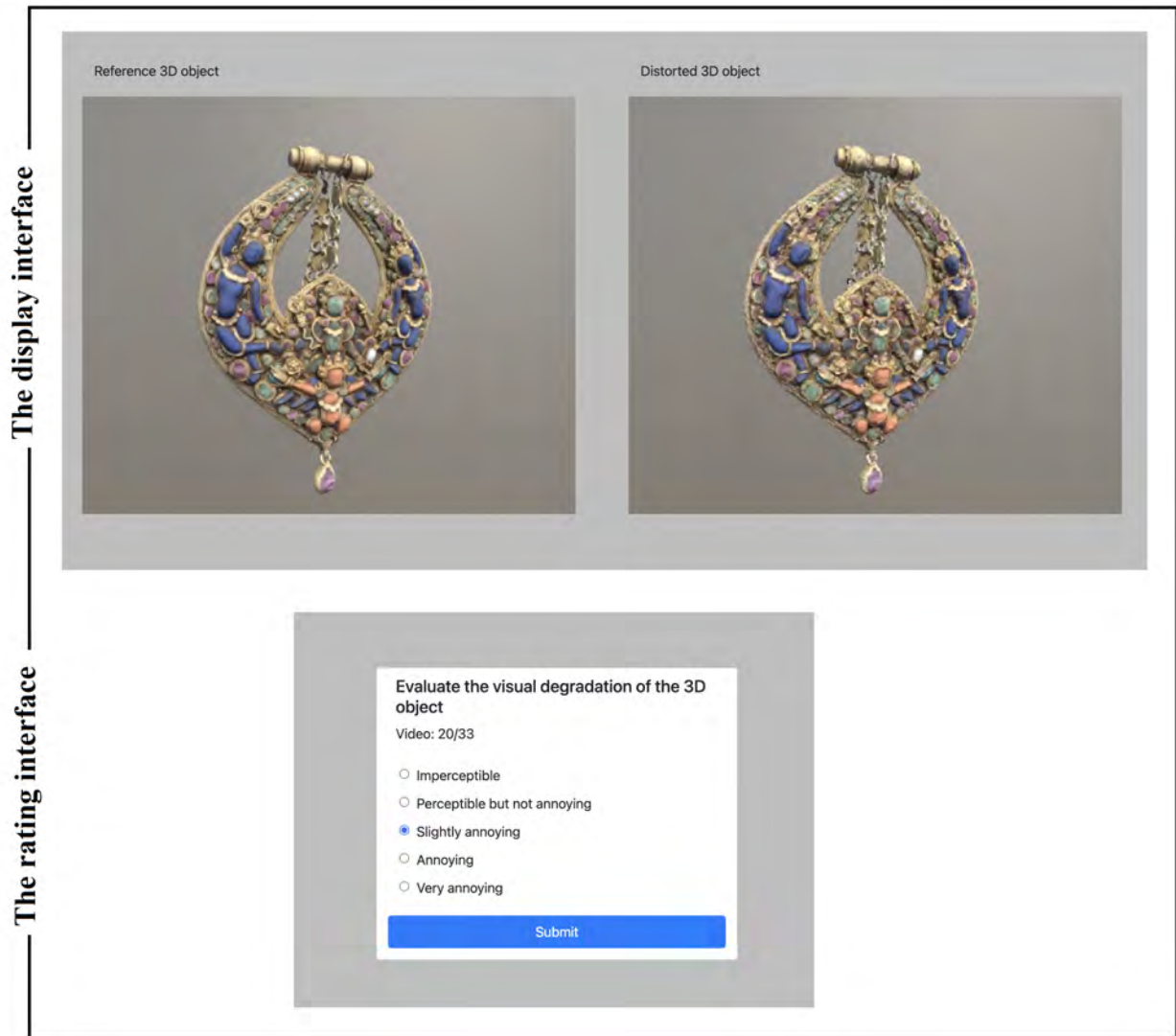


Figure 5.6: The graphical interface of our subjective experiment.

to facilitate detection of unreliable participants later. The golden units (GUs) included (1) a very poor quality stimulus, (2) a very high quality stimulus and (3) a stimulus displayed twice to assess the participant’s consistency (coherence of his/her scores). Participants who fail to answer correctly the golden units are considered outliers and their scores are discarded.

5.2.5 Participants and training

Training

The experiment started with training. In order to familiarize participants with the task and the rating scale, we selected 5 stimuli not included in the 3000 stimuli test and all referring to the same source model. Each stimulus represented one level of the five-level scale of the DSIS method. After displaying each pair of training videos for 8 sec, the

rating interface is displayed for 5 sec and the proposed score assigned to this distortion is highlighted. Once the training is completed the actual test began.

Duration

The test session of our experiment consisted of 33 pairs of videos to rate (1 playlist of 30 stimuli and 3 golden units) and lasted about 10-12 minutes: informed consent + loading videos + instructions + 5 training stimuli \times (8s video length + 5s Rating) + 33 test stimuli \times (8s video length + \sim 4s Rating).

Participants

We ran our experiment until each playlist was fully rated by 45 participants. To fix the number of participants per playlist, we referred to the CS experiment of the previous chapter (Chapter 4), but we also conducted another pilot experiment with 30 stimuli (selected from this dataset) rated by 60 participants and we assessed the evolution of the confidence intervals according to the number of participants (as we did in section 4.4.3).

It took us about 5 days to collect all the data: 148929 quality judgments were collected. A total of 4513 participants took part in the experiment: 2501 males and 2012 females. They were from 67 different countries and aged between 18 and 80. All participants were naive about the purpose of the experiment. Participants who started the experiment and did not complete it or who left after reading the instructions (1659 participants) were rejected.

The recruiting process of the participants was conducted using Prolific³, as the results of the CS experiment of chapter 4 highlight the reliability and seriousness of the Prolific participants. Only participants having a high reliability score (score based on how well they did in past studies) and an adequate number of duly completed jobs on Prolific (number that reflects their familiarity with the platform) were admitted to the experiment.

5.3 Results and analyzes

This section presents the results of our subjective experiment. We analyze the influence of the different types of distortions and their interactions on the subjective scores and thus on the perceived quality. We also evaluate the impact of model characteristics on the perception of distortions.

For the subsequent analyses, subjective scores ranging from very annoying to imperceptible are mapped on a discrete numerical scale from 1 to 5. Note that the scores assigned to the golden units are only taken into account in participants screening and are not considered in the rest of the analyzes.

³<https://www.prolific.co/>

5.3.1 Participants screening

To identify outliers, participants were filtered using the screening strategy described in Chapter 4 section 4.4.1: we combined (1) the ITU-R BT.500-13 screening procedure [11], which detected 159 outliers and (2) the golden units (GU) analysis, which revealed 110 outliers distributed as follows:

- 24 participants rated the distortion of the very poor quality stimulus (GU_{poor}) as imperceptible or perceptible but not annoying ($s_i^{GU_{poor}} \geq 4$, where $s_i^{GU_{poor}}$ denotes the score assigned by participant i to GU_{poor}).
- 39 participants rated the very good quality GU (GU_{high}) as annoying or very annoying ($s_i^{GU_{high}} \leq 2$).
- 32 participant gave inconsistent scores to the third GU showed twice ($|s_i^{GU_{rep1}} - s_i^{GU_{rep2}}| \geq 3$).
- 7 participants rated $s_i^{GU_{poor}} = 3$ & $s_i^{GU_{high}} = 3$.
- 8 participants scored ($s_i^{GU_{high}} = 3 \mid s_i^{GU_{poor}} = 3$) & $|s_i^{GU_{rep1}} - s_i^{GU_{rep2}}| = 2$.

Of the participants who failed to evaluate the golden units, 14 were also detected by the ITU-R BT.500-13 screening procedure. As a result, 255 out of 4513 participants were rejected (5.6%). Only the scores of the remaining participants will be used in the following analyzes.

5.3.2 Resulting MOSs and annotating the whole dataset

Our subjective experiment involved only 3000 stimuli out of 343750. We computed for each of the 3000 test stimuli the Mean Opinion Score (MOS) defined as the average of the rating scores of all participants (Eq. 2.3).

To annotate the remaining stimuli of the dataset, we used the Graphic-LPIPS metric, described in Chapter 7, to predict their MOS (referred to as pseudo-MOS). Indeed, Graphic-LPIPS was adapted to quality assessment tasks, and was trained and tested on our 3000 test stimuli (correlation of 0.86 for a test set of 600 stimuli). Refer to section 7.2 for more details. Figures 5.7 and 5.8 show the distribution of participants' raw scores and the distribution of MOSs obtained for the test stimuli, respectively, while Figure 5.9 presents the distribution of pseudo-MOSs for all stimuli of the data set.

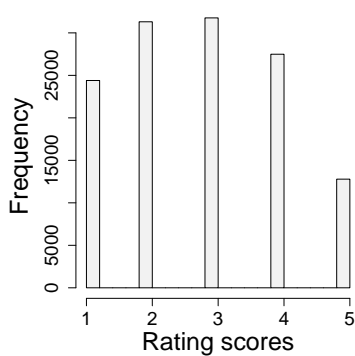


Figure 5.7: Distribution of raw scores collected during the subjective experiment.

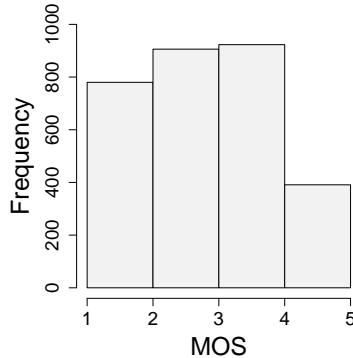


Figure 5.8: MOS distribution of the 3000 test stimuli.

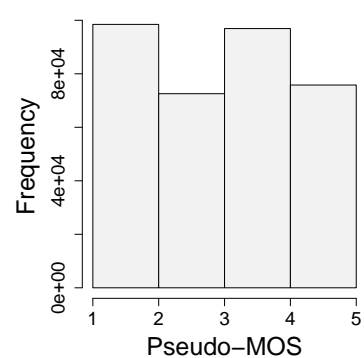


Figure 5.9: Pseudo-MOS distribution of the 343750 stimuli in the dataset.

5.3.3 Influence of each distortion on perceived quality

The perceptual quality of textured 3D content depends on both the geometry and color distortions [35]. Since the distortions in our dataset are of different natures (e.g. quantization, sub-sampling) and affect different aspects of the 3D model (geometry or color), we believe that their impact on the perceived quality is therefore very different. In this section, we provide an in-depth analysis of the effect of each distortion on the perceived quality. We also determine which distortions affect the quality scores the most. To do so, we ran a Multivariate Analysis of Variance (ANOVA: $LoD_{simpL} \times qp \times qt \times T_Q \times T_S$) on the quality score of the entire dataset. The most important results are presented below.

Influence of the geometry and texture coordinate quantization

Figures 5.10.a and 5.10.b show the impact of the quantization parameters on the visual quality of 3D models. As expected, quantizing the model position or texture coordinates with too few bits can seriously degrade model quality. The advantage of using fewer quantization bits is the size reduction of the compressed files, however the resulting visual quality is vastly different from that of the original source model. Therefore, choosing the optimal/correct quantization parameters for an application depends on the intended quality as well as the size constraint. This will be discussed further in section 5.4.

Influence of the LoD simplification

When looking at Figure 5.11.a, it appears that the most simplified stimuli ($L7$, $L8$, $L9$) rated slightly better than the less simplified stimuli ($L1$, $L2$, $L3$). This is counter-intuitive and led us to think that simplifying the models with high strength did not introduce markedly visible impairments/degradations. This is not strictly true: it is actually highly dependent on the geometry quantization level. In fact if we consider only the subset of the least quantized models ($qp = 11$ & $qt = 10$), we see that the MOS logically decreases

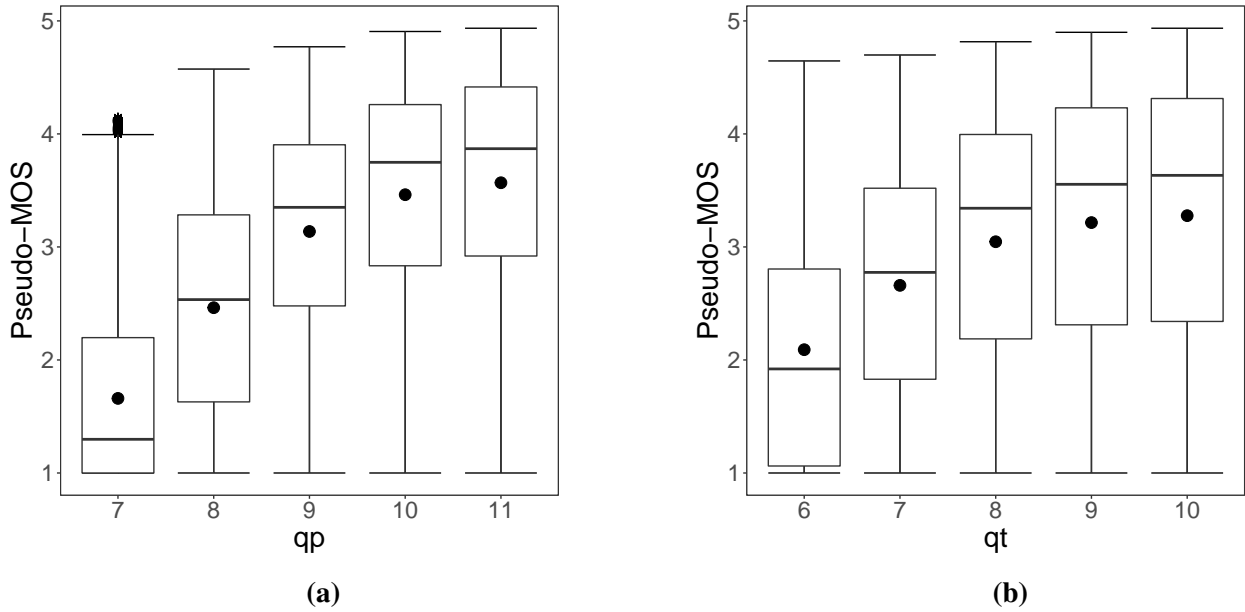


Figure 5.10: Boxplots of MOSs obtained for the quantization of the (a) vertices' positions qp and (b) texture coordinates qt . Mean values are displayed as circles.

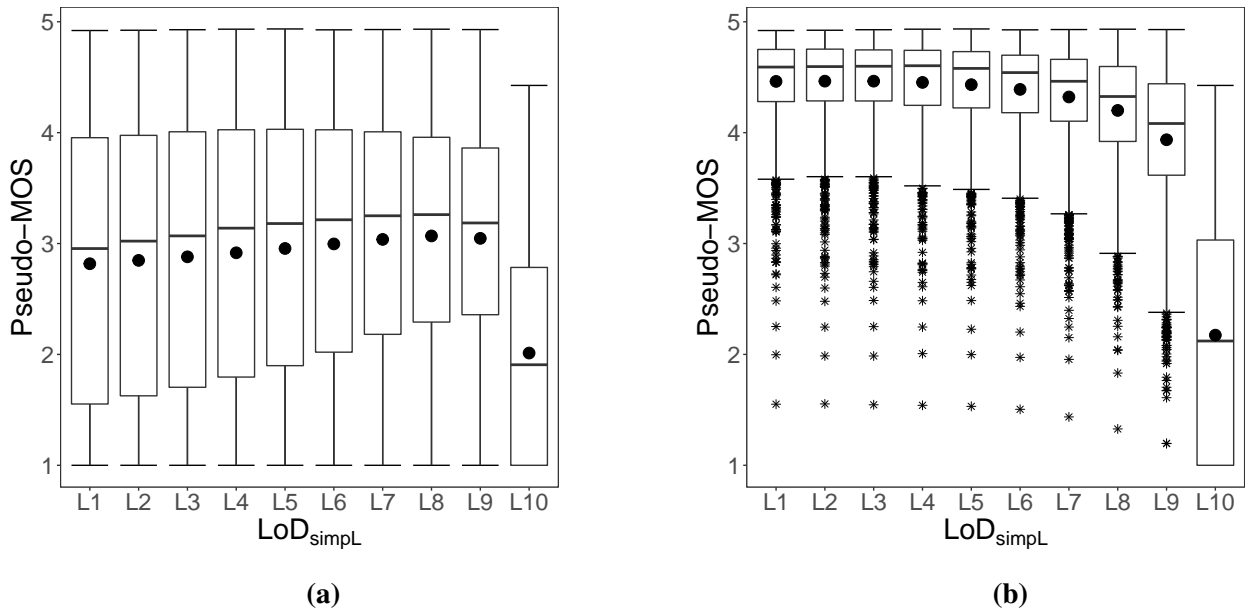


Figure 5.11: (a) Boxplots of MOSs obtained for the LoD simplification LoD_{simplL} . (b) Boxplots of MOSs obtained for the LoD simplification, but restricted to the least quantized models ($qp = 11$ & $qt = 10$). Mean values are displayed as circles.

as the simplification level increases (see Figure 5.11.b). There is thus a significant interaction between the geometry quantization of the model and its levels of detail (p -value $\ll 0.0001$). Subsection 5.3.4 details this point.

Note that for the most simplified level $L10$, it is a bit peculiar: for $L10$, the models are brutally/roughly simplified to have about 2000 faces. This is very degrading (regardless of the qp and qt values), especially for dense models with the highest number of vertices.

Influence of the texture compression and sub-sampling

According to the ANOVA test, the 2 distortions applied to the texture map (the JPEG compression T_Q and the sub-sampling T_S of the texture map) affect significantly the perceived quality (p-value $\ll 0.0001$). However, looking at Figures 5.12.a and 5.12.b, the analysis of the impact of these distortions on the MOS is not obvious as that of quantization. Figure 5.12.a shows that for $T_Q \geq 50$, the increase of the texture quality does not seem to affect the overall perceived quality.

For texture sub-sampling, Figure 5.12.b shows that increasing the texture resolution T_S more than 712×712 did not influence the perceived quality. We believe this is due to the fact that participants were not able to see/detect the difference in quality between the high resolution textures (1024×1024 , 1440×1440 , 2048×2048) since the resolution of the stimulus videos shown in the experiment was 650×550 . Thus for our visualization conditions, we can see that we can push the JPEG compression level and sub-sample the texture heavily without impacting the overall quality of the model. This allows to drastically reduce the size of the compressed data.

The impact of T_S is emphasized when considering its interaction with T_Q and qt ; see subsection 5.3.4..

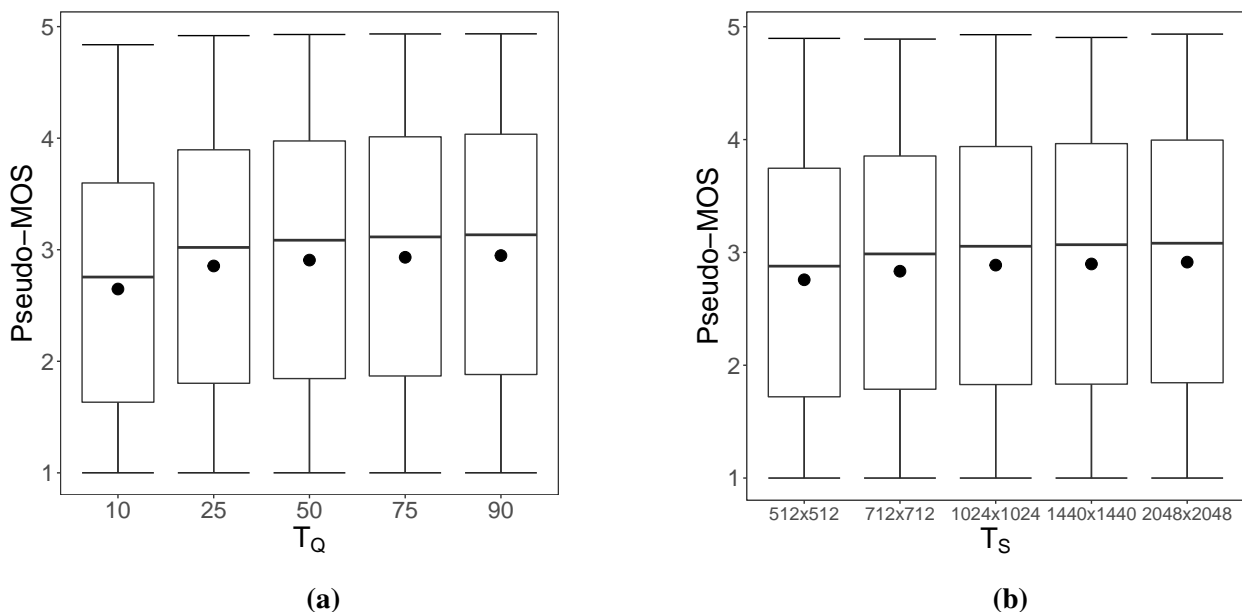


Figure 5.12: Boxplots of MOSs obtained for the texture (a) compression T_Q and (b) sub-sampling T_S . Mean values are displayed as circles.

5.3.4 Influence of distortion interactions on perceived quality

Based on the results of the previous section, we believe that the impact of the combinations of the different types of distortions differ from the cumulative impact of each distortion applied alone. In this section, we study the effect of distortion interactions on the perceived quality of textured 3D meshes.

Interaction of LoD simplification and position quantization

The perception of the LoD simplification is strongly related to (dependent on) the quantization of the model's positions (significant interaction with a p-value $\ll 0.0001$). Figure 5.13 illustrates this interaction: when quantizing model's positions with too few bits ($qp = 7$, $qp = 8$), the MOS increases as the simplification level increases (i.e., the number of vertices decreases). This effect is reversed for less quantized models ($qp = 10$, $qp = 11$) as the MOS decreases when the level of simplification increases. Overall, the local geometry alteration (local contrast alteration) caused by a strong quantization is more visible on dense meshes ($LoD_{simpL} = L1$) than on coarse meshes ($LoD_{simpL} = L9$). Figure 5.14 illustrates a visual example in which we can see that the effect of the quantization of the vertex positions is much more visible on the dense model ($LoD_{simpL} = L1$). Figure 5.14.c also shows that simplifying the mesh with a high strength then quantizing it is like applying a low-pass filter to the quantized mesh of Figure 5.14.b. Indeed, the frequency of artifacts created by geometry quantization is higher on a dense mesh than on a simplified mesh; this is what makes the artifacts more visible. This effect is related to the Contrast Sensitivity Function (CSF) which describes the visibility threshold with respect to the spatial frequency. In other words, CSF is the ability to detect subtle differences in shading and patterns. As the density of a mesh increases from a very low value, it would be slightly easier for the human visual system to notice the local contrast alteration on the mesh surface [85].

Thus, the effect we observe is a mix between the CSF and the masking effects due to the rendering and shading.

Interaction of geometry and texture coordinate quantization

It is interesting to observe that the perception of the distortion qt introduced to the UV map of a 3D model (the mapping between the model surface/geometry and the texture) is influenced/affected by the quantization of the vertex positions qp . Figure 5.15 shows the interaction between these 2 factors. We can observe that for low values of qp the improvement brought by increasing the quantization bits of the texture coordinates qt did not compensate the degradations generated by low qp and thus did not improve the MOSs much. Figure 5.16 shows 2 degraded versions of the bird (Model #33), both quantified with $qp = 6$. However, one stimulus has a higher qt ($qt = 10$) than the other ($qt = 6$). Both stimuli scored $MOS = 1$ (the lowest possible score); yet, the stimulus with $qt = 10$ (less quantized texture coordinates) shows less degradation (see bird's eye and beak). This may be due to the discrete categorical scale used in the DSIS method (five-level impairment scale) that does not allow for possible variations around best and worst qualities. We call this the "scale saturation effect".

Furthermore, looking at Figure 5.15, it seems that the quantization of the model positions (qp) has more impact on the visual quality than the quantization of the UV map (qt): for low values of qp , we obtain a low MOS whatever the value of qt . Hence, we believe that for a given level of LoD_{simpL} , T_S and T_Q , the quality Q of a textured 3D object can be represented by a multiplicative model as follows: $Q = Q_{qp}^\alpha \cdot Q_{qt}^\beta$, where potentially $\alpha > \beta$.

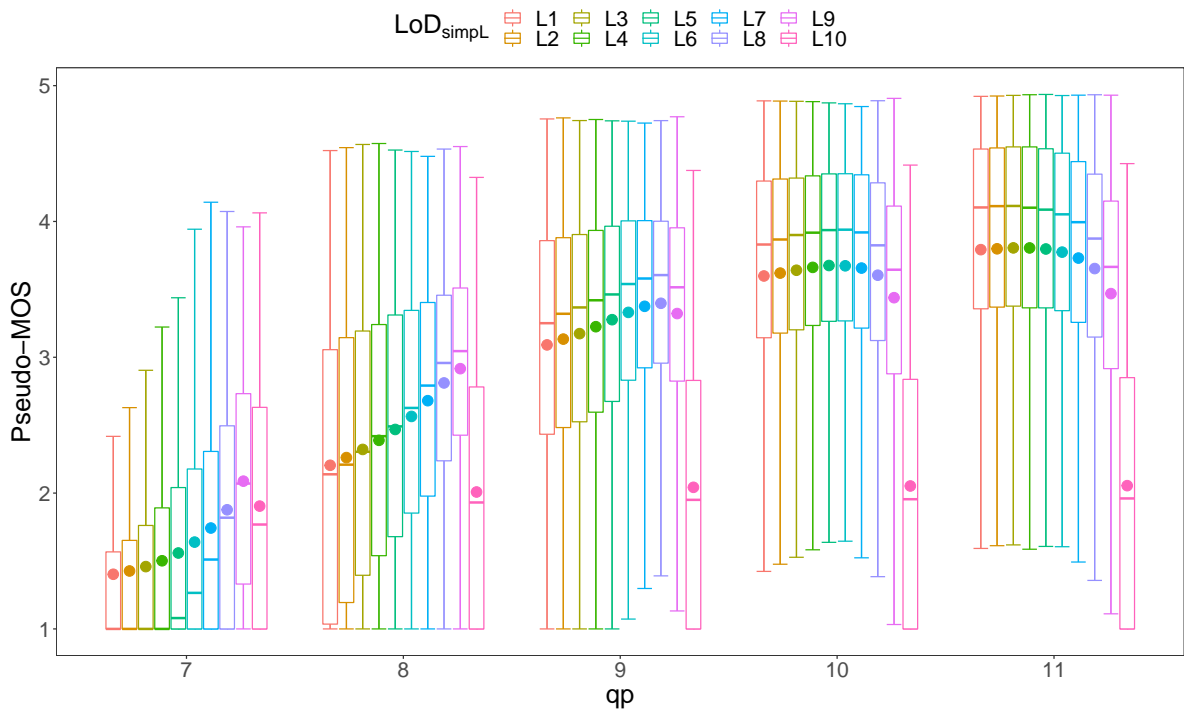


Figure 5.13: Boxplots of MOSs illustrating the interaction between the LoD simplification LoD_{simpL} and the quantization of the model's positions qp .

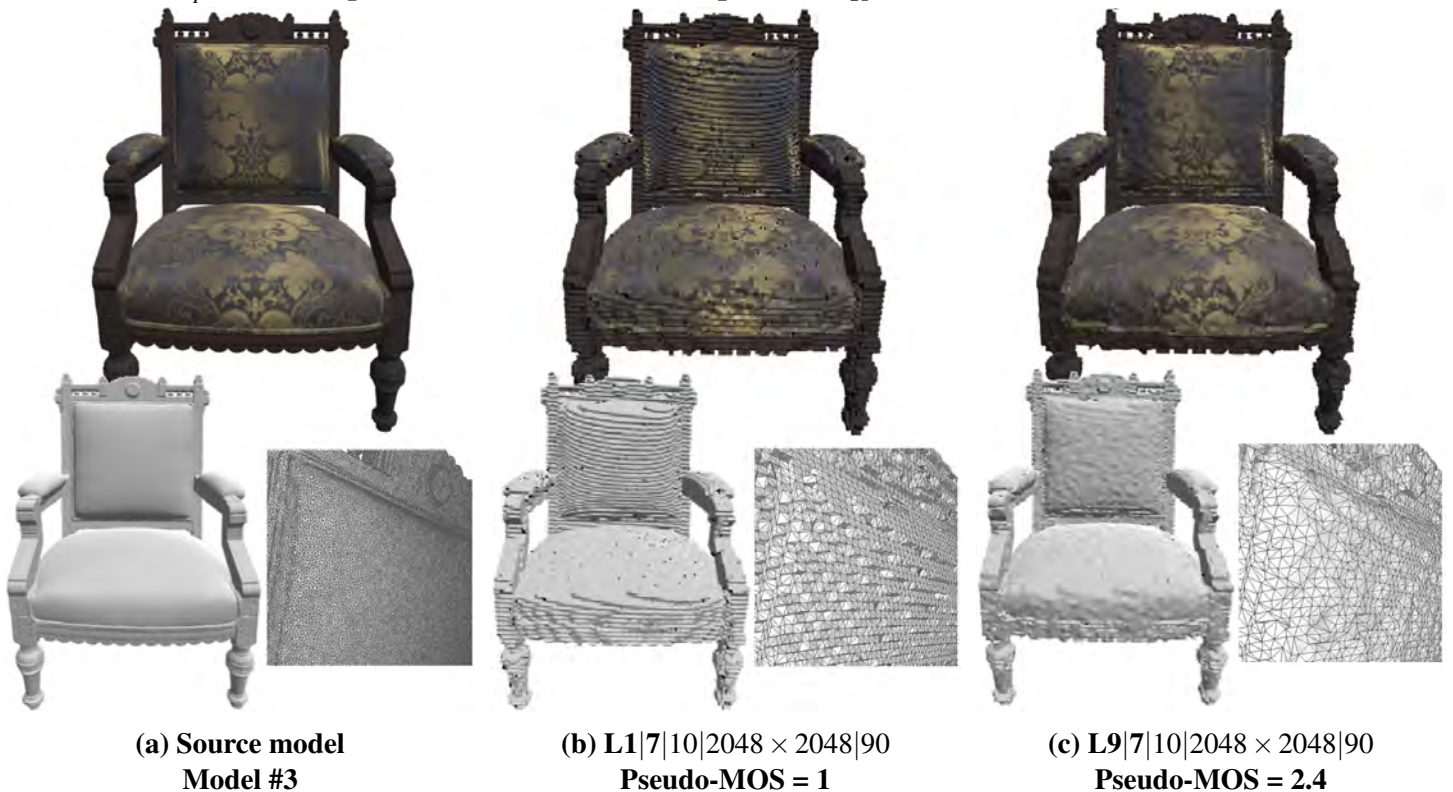


Figure 5.14: Visual example illustrating the interaction between the LoD simplification and the position quantization regarding the perceived quality. The 2nd row shows a rendering of the stimuli without the texture as well as a zoom on the chair back showing the topology of the mesh after distortion. Acronyms refer to the following combination of distortion parameters: $LoD_{simpL}|qp|qt|T_s|T_Q$.

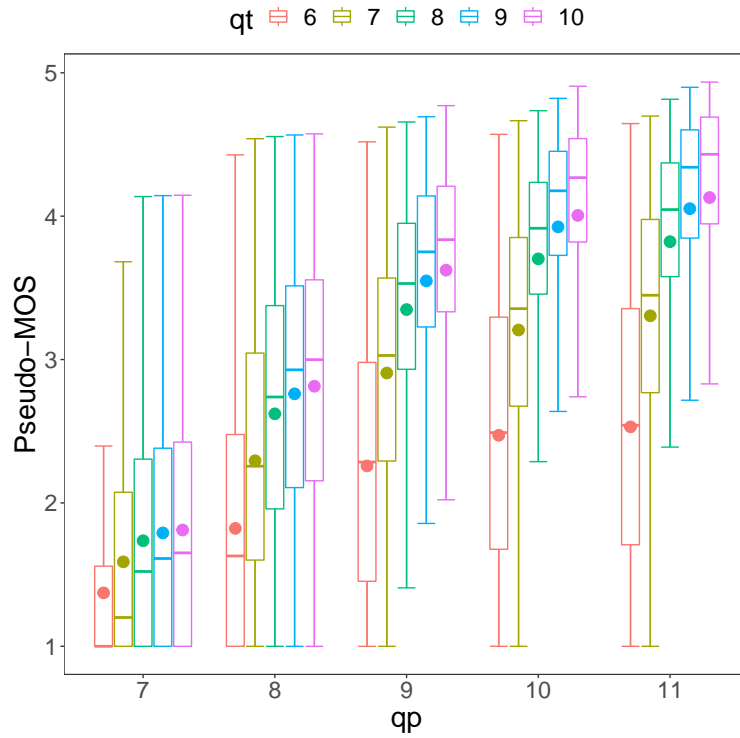


Figure 5.15: Boxplots of MOSs illustrating the interaction between the geometry qp and texture coordinate qt quantization.

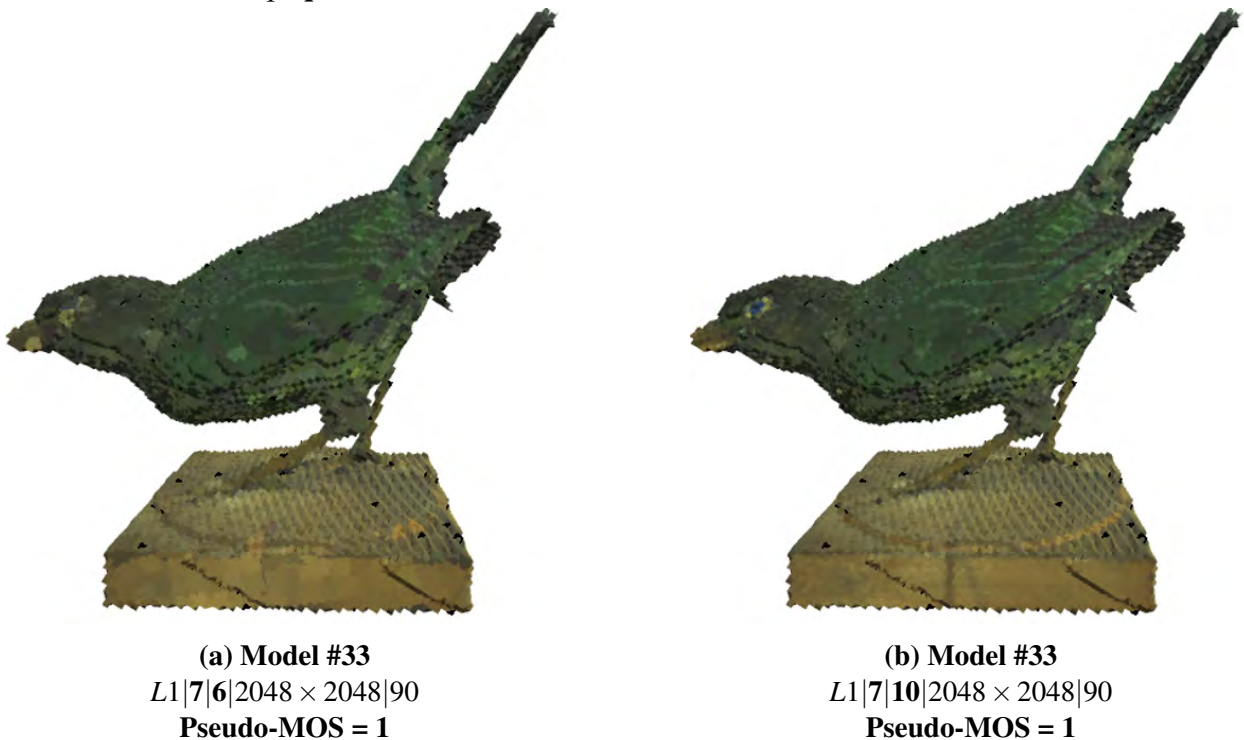


Figure 5.16: Visual example illustrating the interaction between the geometry and texture coordinate quantization regarding the perceived quality. Acronyms refer to the following combination of distortion parameters: $LoD_{simpL}|qp|qt|T_S|T_Q$.

Interaction of the texture compression and sub-sampling

There is a significant interaction (p-value $\ll 0.0001$) between the compression and sub-sampling applied on the texture image. Figure 5.17 shows its impact on the perceived quality. For the lowest texture quality ($T_Q = 10$), the MOS increases as the texture size T_S increases. Overall, compression artifacts are less visible on larger textures. The reason is that the blocking artifacts caused by the JPEG compression are bigger/larger on screen for the smaller textures.

We can also notice that stimuli with medium or low compressed textures ($T_Q \geq 50$) obtained almost the same MOSs regardless the texture size. This is coherent with what we observed in Figure 5.12.a (subsection 5.3.3). We believe that the compression standards/thresholds are not the same for a “natural” image and a “texture” image because their visualization is not the same. Indeed, a texture is mapped on the 3D object (which is then rendered) which makes the perception of compression artifacts less obvious than on a natural 2D image directly displayed on the screen. Thus, for a given perceived quality, we may push the compression level of a texture more than that of a natural image.

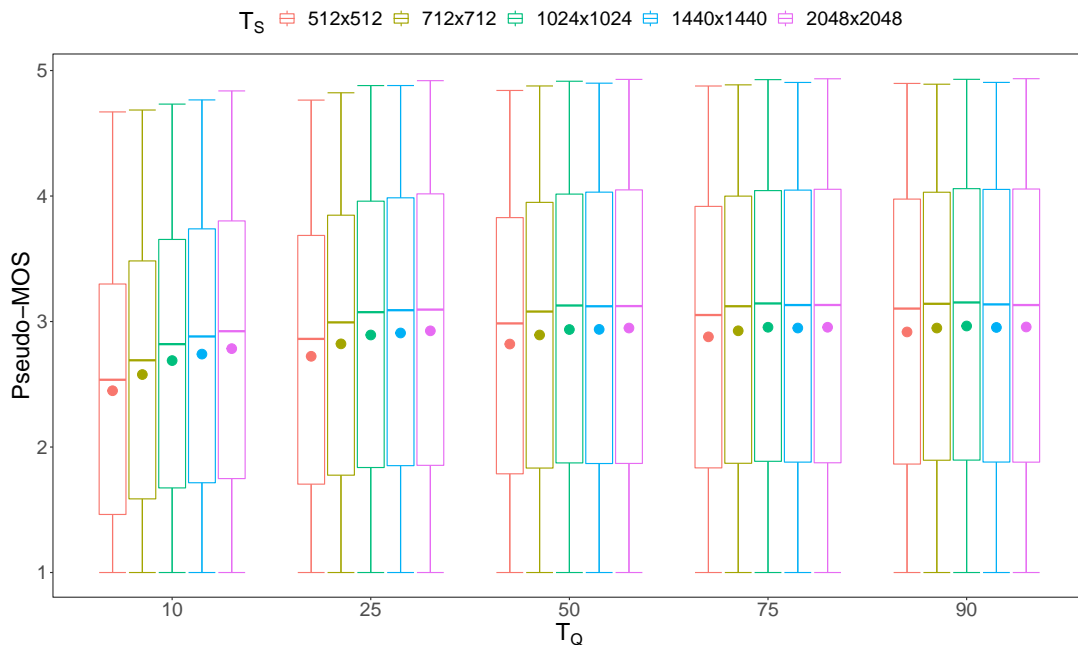


Figure 5.17: Boxplots of MOSs illustrating the interaction between the texture compression T_Q and sub-sampling T_S .

Interaction of texture coordinate quantization and texture sub-sampling

The impact of the texture sub-sampling is strongly related to the mapping of the texture on the model surface. In fact, quantizing the texture coordinates with few bits ($qt = 7$, $qt = 8$) generates a “tiling effect” as illustrated in Figures 5.18 and 5.19. According these figures, this effect is less visible on small textures: For instance, for $qt = 6$, stimuli with a texture of size 512×512 scored better than those with a texture of size 2048×2048 . This is due to the fact that sub-sampling the texture (reducing its size) reduces the high frequency information within the texture (this is a resampling using a low pass filter). Thus, the texture is smoothed, which decreases the tiling effect and subsequently increases the MOS.

qt and T_S are thus linked/related. These 2 parameters must be set with respect to each other: e.g., for low qt values (UV map highly quantized), the size of the texture T_S must be decreased.

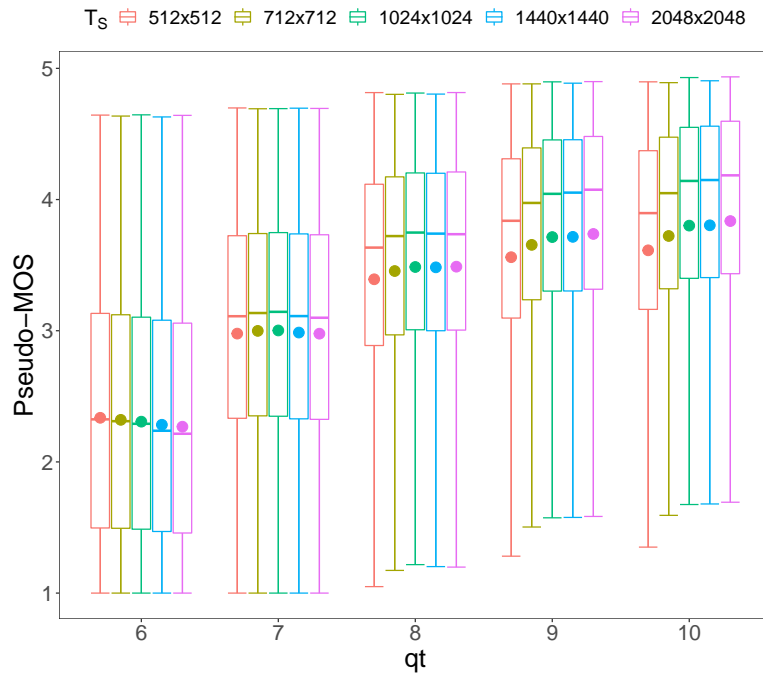


Figure 5.18: Boxplots of MOSs illustrating the interaction between the texture coordinate quantization levels qt and the texture sub-sampling T_S .

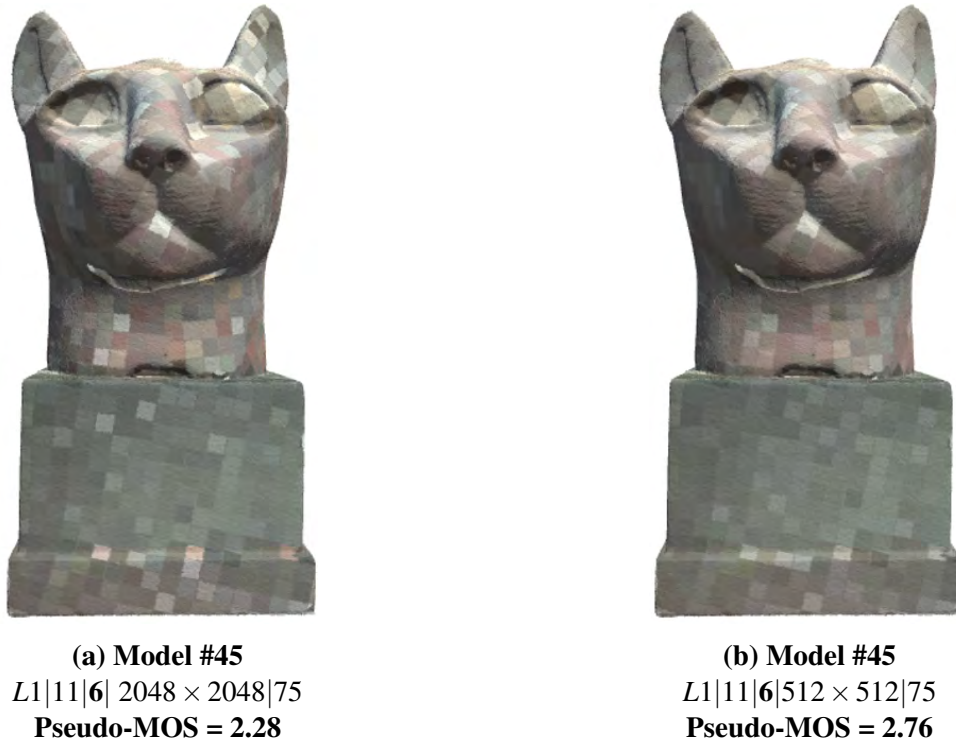


Figure 5.19: Visual example illustrating the interaction between the sub-sampling of the texture and its coordinates quantization regarding the perceived quality. Acronyms refer to the following combination of distortion parameters: $LoD_{simpL}|qp|qt|T_S|T_Q$.

5.3.5 Influence of content characteristics on perceived quality

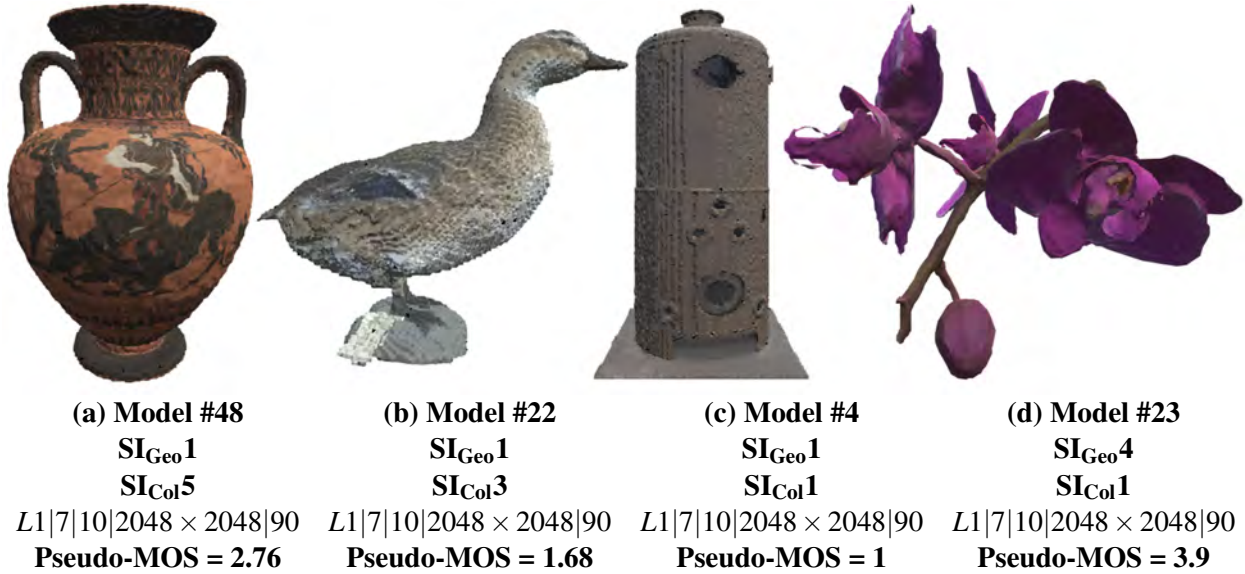


Figure 5.20: MOSs of different models with different geometric SI_{Geo} and color SI_{Col} characteristics and having undergone the same distortions ($LoD_{simpL}|qp|qt|T_S|T_Q$).

Figure 5.20 shows that for the same distortion parameters (same LoD_{simpL} , qp , qt , T_S and T_Q), the perceived quality is not the same: we obtained different ranges of MOS. This is because the content has a concealing effect on the perception of the distortions, which is consistent with the characteristics of the human visual system [152]. Thus, the impact of distortions on the quality of 3D models is highly content dependent.

In this section, we evaluate the influence of content characteristics on the perception of distortions and thus on quality. To do so, we use the set of 3D content characterization measures we developed in Section 5.1.2 (SI_{Geo} and SI_{Col}). We group our 55 models into clusters of 11 models based on their geometric and color complexity. Thus, the first cluster “ $SI_{Geo}1$ ” contains the first 11 models with the least complex geometry (lowest SI_{Geo} values), while “ $SI_{Geo}5$ ” designates the 11 models with the most geometric detail (highest SI_{Geo} values). “ $SI_{Col}1$ ” denotes the 11 source models with the least color detail while “ $SI_{Col}5$ ” refers to the models with the richest texture. Our clusters are well dispersed in the SI_{Geo}/SI_{Col} plane (cover a large range) as illustrated in Figure 5.21 which is an histogram representation of the Figure 5.3.a.

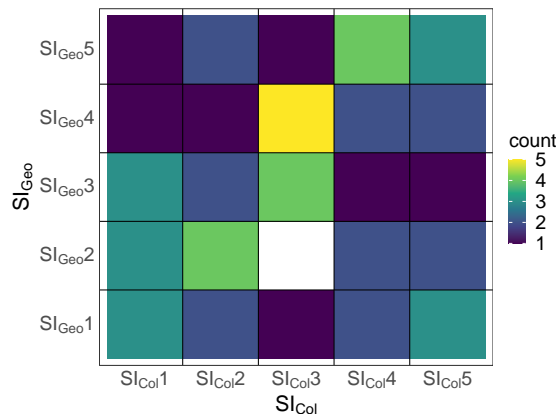


Figure 5.21: Clusters of Source models grouped by geometric and color characteristics.

Influence of geometry and color complexity on the perception of position quantization

To evaluate the influence of the model characteristics on the perception of degradations generated by the quantization of the position of its vertices qp , we fixed the levels of all other distortions at their best levels (giving the best quality), i.e. we considered only the subset of stimuli having: $LoD_{simpL} \in \{L1, L2, L3\}$ & $qt \in \{9, 10\}$ & $T_Q \in \{75, 90\}$ & $T_S \in \{1440 \times 1440, 2048 \times 2048\}$.

To assess the impact of geometry information, we eliminated the stimuli with rich textures (SI_{Col4} and SI_{Col5}) in order to dissociate the influence of geometry and color and to avoid a possible masking effect of one on the other. According to ANOVA, a significant interaction exists between the geometric complexity of the model and the visual impact of the position quantization (p-value $\ll 0.0001$). Figure 5.22a shows that the geometric information can mask the geometry alteration caused by the quantization of the vertices position: For the same quantization level qp , meshes with complex geometry ($\in \{SI_{Geo4}, SI_{Geo5}\}$, e.g., Model #40 in Figure 5.23) obtained higher MOSs than those with less complex geometry ($\in \{SI_{Geo1}, SI_{Geo2}\}$, e.g., Model #45 in Figure 5.23).

Regarding the impact of color information, the results presented in Figure 5.22b show that for the same level of quantization, models with rich texture (e.g. Model #48 in Figure 5.24) were judged to be of higher quality than those having simpler texture (e.g. Model #20 in Figure 5.24). These results corroborate those observed for point clouds and reported in [153].

Thus, we can say that the color and geometry mask the geometric degradations of a quantized 3D model.

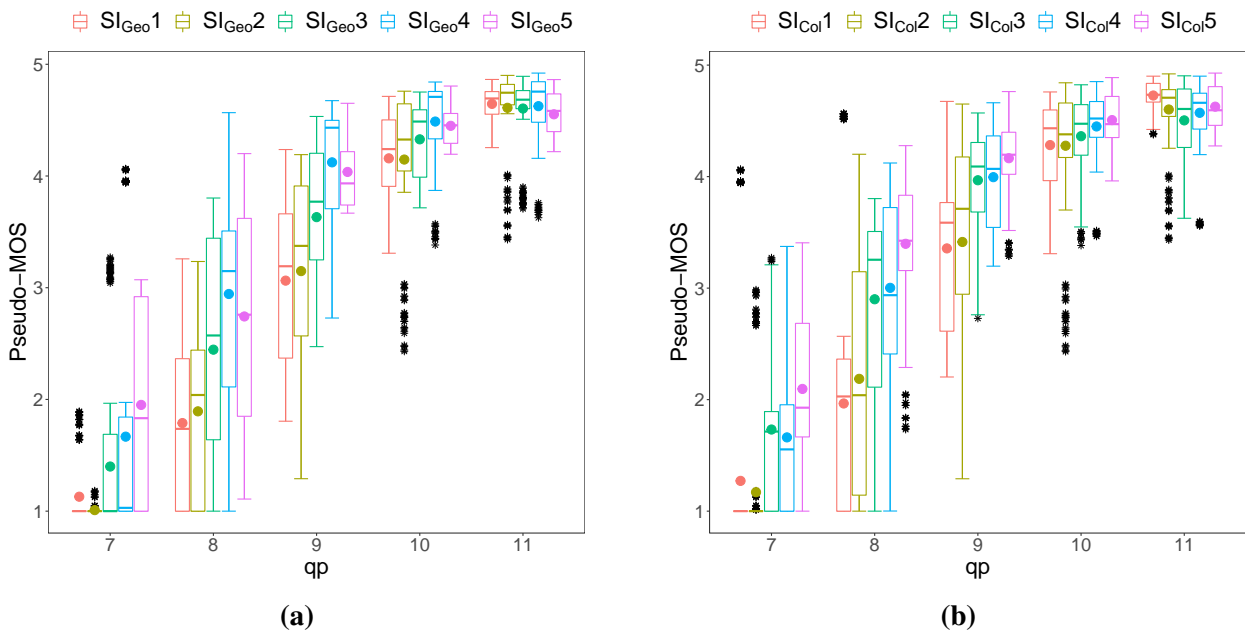


Figure 5.22: Boxplots of the MOSs illustrating the interaction between the (a) geometric SI_{Geo} and (b) color characteristics SI_{Col} of the models and the quantization of the position of their vertices qp .

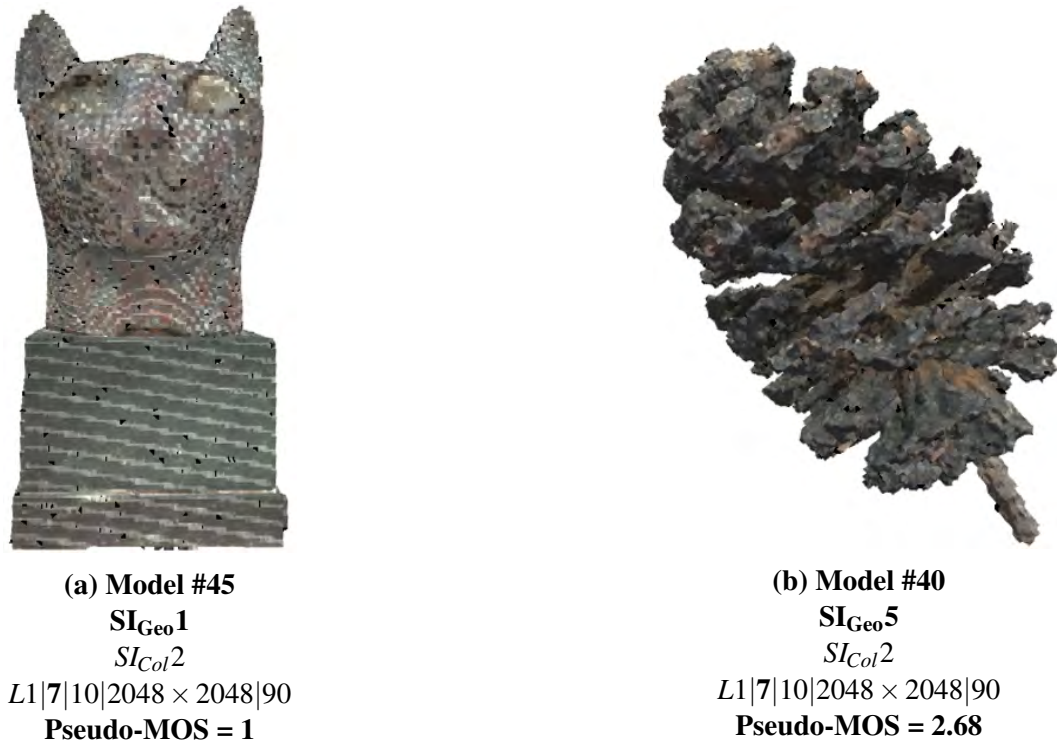


Figure 5.23: Visual examples illustrating the impact of the geometric complexity of the model on the perception of degradations generated by the geometry quantization. Acronyms refer to the following combination of distortion parameters: $LoD_{simpL}|qp|qt|T_s|T_Q$.



Figure 5.24: Visual examples illustrating the impact of the color complexity of the model on the perception of degradations generated by the geometry quantization. Acronyms refer to the following combination of distortion parameters: $LoD_{simpL}|qp|qt|T_s|T_Q$.

Influence of geometry and color complexity on the perception of texture coordinates quantization

Following the same approach described in the previous section, we varied qt and set the levels of all other distortions at their best levels ($LoD_{simpL} \in \{L1, L2, L3\}$ & $qp \in \{10, 11\}$ & $T_Q \in \{75, 90\}$ & $T_S \in \{1440 \times 1440, 2048 \times 2048\}$) in order to evaluate whether the model's characteristics can mask the impairments caused by quantizing its UV map. Figure 5.25a illustrates the impact of color characteristics while Figure 5.25b that of geometric characteristics.

Figures 5.25a and 5.26 clearly show that models with simple, especially monochromatic, textures (such as Model #50) are less sensitive to the UV map quantization than those with colorful and detail-rich textures (such as Model #24).

Looking at figure 5.25b, we realize that the interaction between the geometry of the model and the quantization of the UV map is more complex to evaluate, yet this interaction is significant (p-value $\ll 0.0001$). Indeed, for low values of qt , the MOS decreases progressively while passing from $SI_{Geo}1$ to $SI_{Geo}3$, then rises for $SI_{Geo}4$ and $SI_{Geo}5$. To better understand this behavior, we looked at visual examples of models $\in \{SI_{Geo}4, SI_{Geo}5\}$, reported in Figure 5.27. We noticed that the MOS values are not systematically high for all these models. It depends on the models, specifically their texture atlas and the quality of the surface parameterization: i.e., the fragmentation of the texture atlas (the texture seams of the UV map) and the quality of the atlas packing. In fact, quantization artifacts are clearly more visible on models whose texture atlas is highly fragmented (high number of texture seams) and/or not efficiently packed (see Figure 5.27 1st row). In contrast, UV quantization artifacts are less visible for models having (1) homogeneous/uniform texture colors or (2) less fragmented textures (low chart count, low number of texture seams), as can be seen in Figure 5.27 2nd row).

Thus, the impact of UV quantization on the visual quality depends not only on the geometric and color complexity of the model but also on the amount of texture seams (the level of fragmentation of its texture atlas). The effect of texture seams is studied in more detail in the following paragraph.

Influence of texture seams on the perception of texture coordinates quantization

According to the observations of the previous subsection 5.3.5, the perception of the UV map quantization, is affected by the amount of texture seams. Thus, in order to quantitatively assess/characterize the amount of texture seams, we computed for each source model the percentage of vertices belonging to texture seams, which we denote by V_{Tseams} . As for SI_{Geo} and SI_{Col} , we grouped our source models into 5 clusters of 11 models each: $V_{Tseams}1$ contains the first 11 models with the lowest percentage of vertices on texture seams while $V_{Tseams}5$ contains those with the highest percentage of vertices on seams. We then evaluate the impact of this factor on the artifact caused by the UV map quantization. As before, we set the levels of all other distortions at their best levels ($LoD_{simpL} \in \{L1, L2, L3\}$ & $qp \in \{10, 11\}$ & $T_Q \in \{75, 90\}$ & $T_S \in \{1440 \times 1440, 2048 \times 2048\}$).

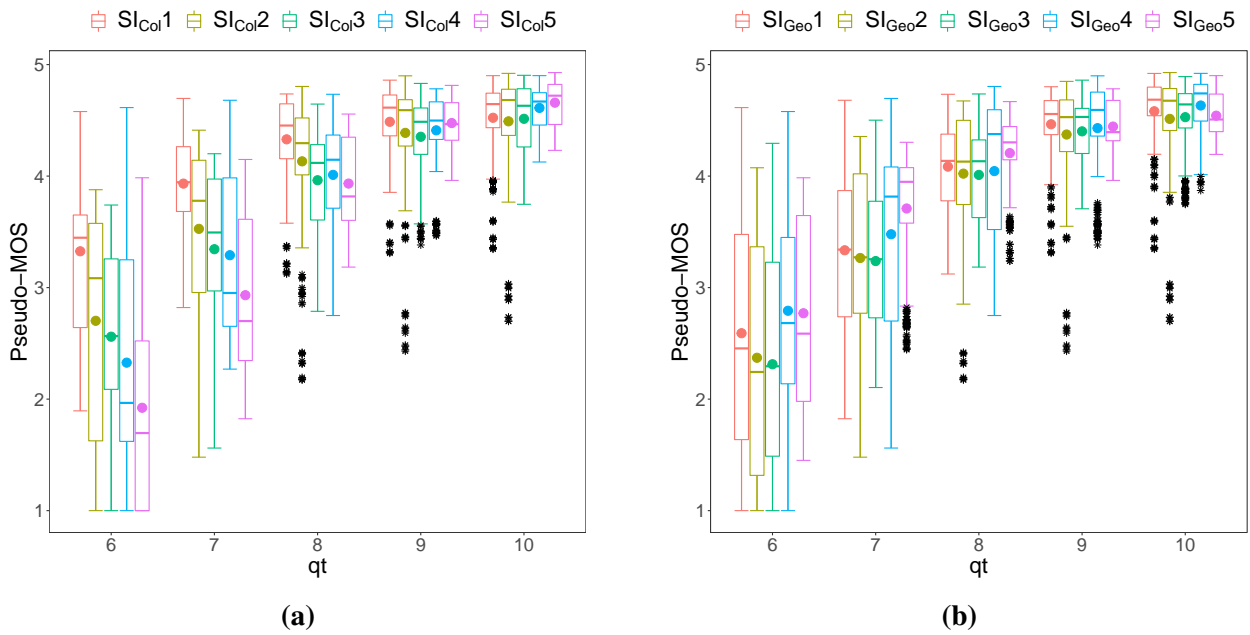


Figure 5.25: Boxplots of the MOSs illustrating the interaction between the quantization of the texture coordinates qt and (a) the model color SI_{Col} and (b) geometric SI_{Geo} characteristics.

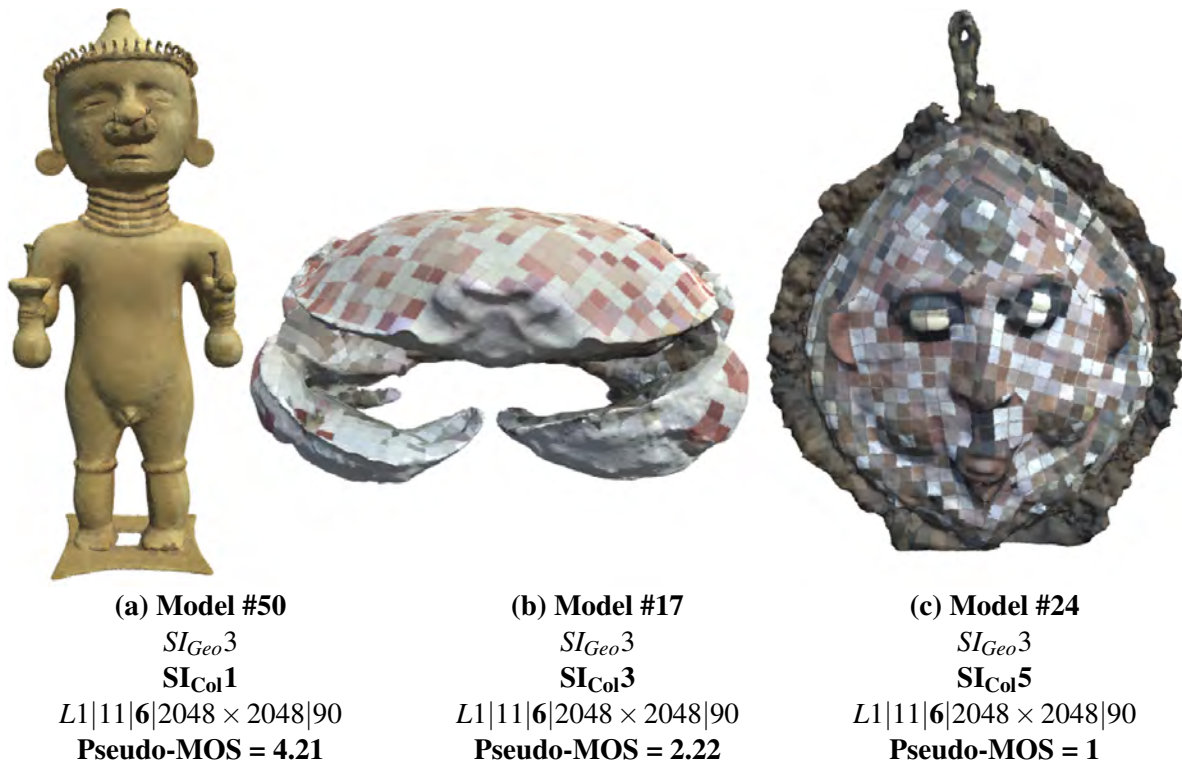


Figure 5.26: Visual examples illustrating the impact of the color characteristics of the model on the perception of degradations generated by the quantization of the texture coordinates. Acronyms refer to the following combination of distortion parameters: $LoD_{simpL}|qp|qt|T_S|T_Q$.

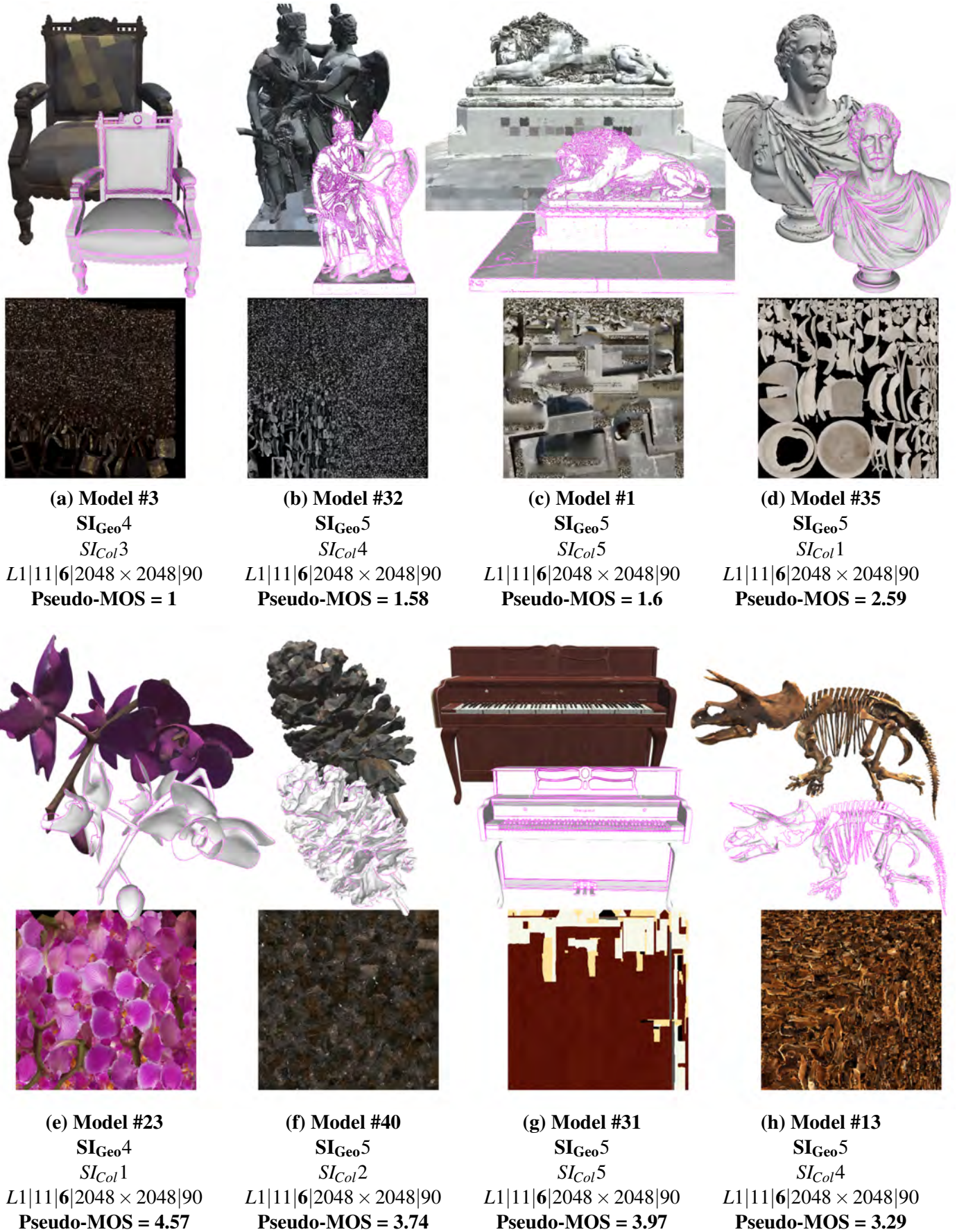


Figure 5.27: Visual examples illustrating the impact of UV map quantization on the perceived quality of textured 3D meshes. Models are presented with their texture seams highlighted and their texture map. Acronyms refer to the following combination of distortion

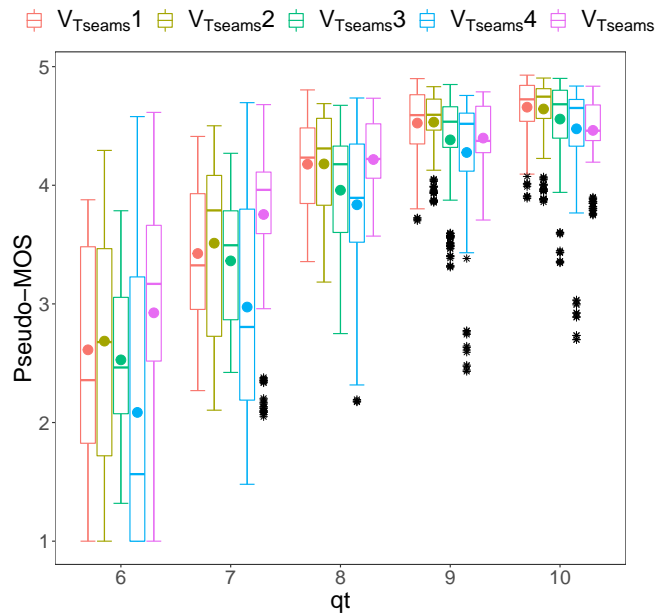


Figure 5.28: Boxplots of the MOSs illustrating the interaction between the quantization of the texture coordinates qt and the amount of vertices present on texture seams V_{Tseams} .

Results, reported in Figure 5.28, show that overall the quality decreases as the number of vertices on texture seams increases (except for $V_{Tseams5}$). In fact, the UV map quantization causes more severe artifacts at texture seams compared to region without seams, which can be detrimental to models with a large number of vertices located at the seams (models with highly fragmented texture atlas). Figure 5.29 shows an example: the left part of the bust’s jacket (Model #38) exhibits more seams than the right part, the same is true for its left eye. This makes the quantization artifacts more visible on the left part of the bust than on the rest of it. The presence of these artifacts affected the overall perceived quality of the bust and led to its quality score being lower than that of other models with less vertices located on texture seams (e.g. Models #54 and #45). These models have less fragmented texture atlas.

As for the models belonging to $V_{Tseams5}$ getting a better score than the others, this actually depends on the models and their texture atlas (as explained in the previous subsection 5.3.5). Models #13 and #31, shown in Figure 5.27, belong to $V_{Tseams5}$ and have a relatively high quality score. This is related to the fact that their texture color is uniform/homogeneous, making quantization artifacts less severe than those on a color-rich texture. It may also be related to the fact that these models have very tight packed atlases. V_{Tseams} remains a simple measure that gives a global idea on the texture seams but does not allow to finely characterize the UV map and the texture atlas. The reader is referred to Maggiordomo et al. [154] for a more comprehensive analysis of the issues related to the surface parametrization. The authors also proposed a set of quality measures that characterize the quality of the surface parametrization and texture atlases, notably the “UV Occupancy” measure that assesses the quality of the atlas packing and the “Atlas Crumbliness and Solidity” measures that capture the severity of texture seams in a given UV map.

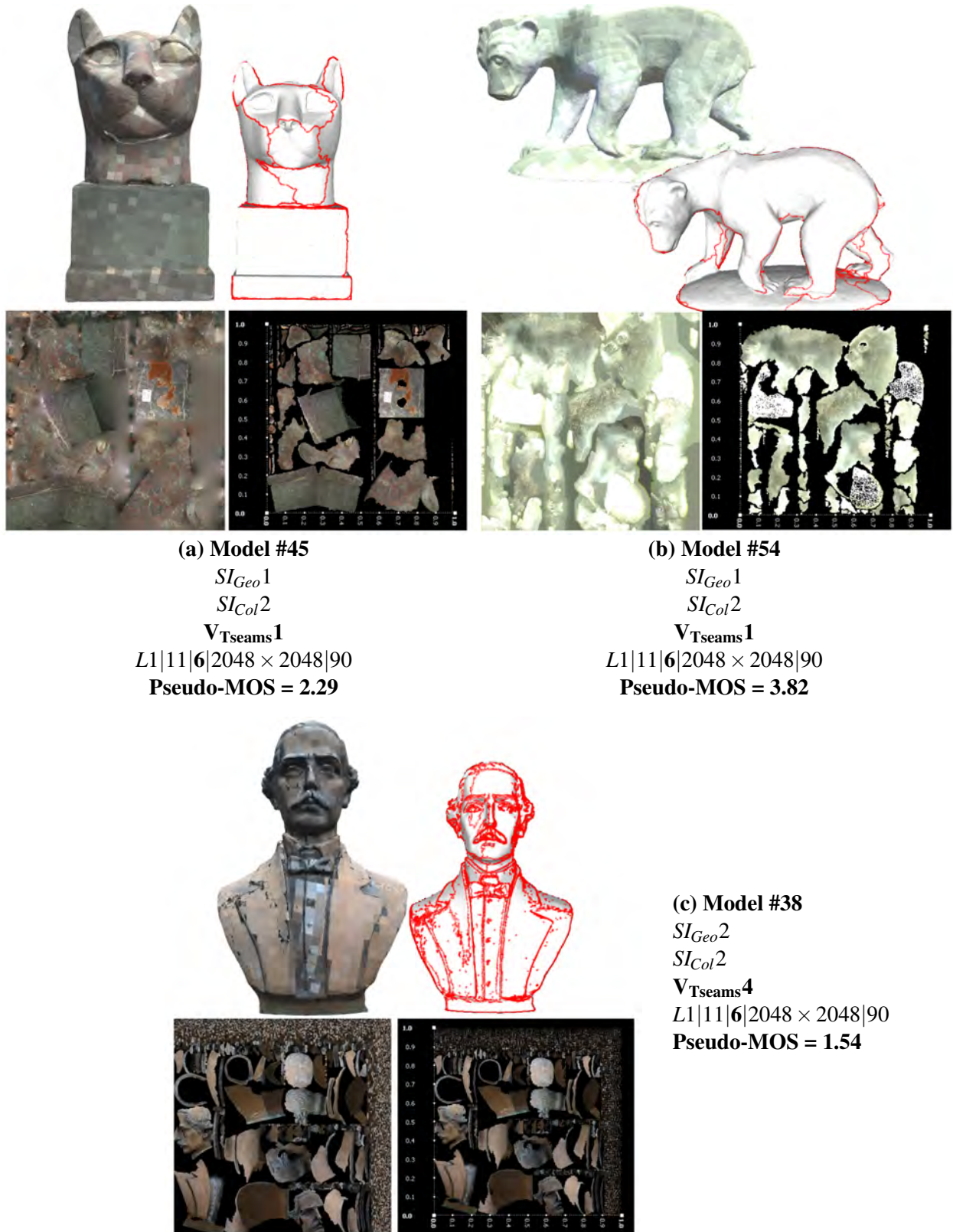


Figure 5.29: Visual examples illustrating the impact of the texture seams and the perception of degradations generated by the quantization of the texture coordinates. Models are presented with their texture map (left) and their UV map (right). Vertices present on texture seams are highlighted in red. Acronyms refer to the following combination of distortion parameters: $LoD_{simpL}|qp|qt|T_s|T_Q$.

5.4 Application: Rate-Distortion control

In Rate-Distortion (RD) optimization, the encoder settings are determined by maximizing a reconstruction quality measure subject to a constraint on the bitrate. One of the main challenges of RD optimization is to define a quality measure that can be computed with low computational cost and which correlates well with the perceptual quality. While several quality measures that fulfill these two criteria have been developed for images and videos, no such one exists for 3D textured meshes.

Using our annotated dataset of over 343k stimuli, we conducted a preliminary qualitative analysis to observe the optimal compression setting of our models under the constraint of a target bitrate. The goal was to find for each of our 55 source models the parameters of distortion providing the best possible visual quality for a given size requirement. This study is possible because each of our stimuli is associated with a quality score and a size (in KiloBytes), resulting from the compression of the source model with the corresponding parameters (using JPEG for the texture and Draco encoder for the connectivity, geometry and UV maps). Thus, the size of a stimuli (in KB) is equal to the sum of the size of its compressed texture and its compressed 3D model. The results are reported in Tables A.8, A.9 et A.10 provided in the appendix (section A.4). In each cell, we can find the combination of distortion parameters needed to achieve the highest possible quality for a given size range. The cell color indicates the range of quality we can achieve within the given size range. Our source models are sorted in ascending order according to their geometric complexity SI_{Geo} (Table A.8), their color complexity SI_{Col} (Table A.9), and the amount of vertices on the texture seams V_{Tseams} (Table A.10). This will allow us to evaluate the evolution of the minimum size needed to obtain a good quality of the reconstructed model according to its characteristics.

The results show that globally complex models, and in particular those with a large number of texture seams, are more difficult to compress as they require more quantization bits, especially finer LoD ($LoD_{simpL} < L5$) than simpler models (mainly $LoD_{simpL} \in \{L7, L8, L9\}$) to ensure a maximum quality after decompression ($MOS \in [4, 5]$). This results in an increase in the size of the compressed file and texture (and therefore higher bitrates).

These tables can be used to predict the optimal distortion parameters for a new model given a size requirement based on its characteristics: we first compute the SI_{Geo} , SI_{Col} , and V_{Tseams} measures for the new model, and then by looking at which clusters (SI_{Geo}^i , SI_{Col}^j , V_{Tseams}^k) its characteristics belong to (or more precisely, which model in our database has the closest characteristics), we can determine the combination of distortion levels that yields the best possible quality with this size requirement.

As a future work, it would be interesting to devise, using our dataset, an analytical perceptual rate-distortion model capable of maximizing the visual quality of the reconstructed textured meshes subjected to a target bitrate.

5.5 Discussion

This section synthesizes the work carried out this chapter. First, we selected 55 source models. We used the measures we proposed in section 5.1.2 to show that these models cover a wide range of geometric, color, and semantic complexity. Next, we generated 343750 distorted version of these models using combinations of Level of Details (LoD) simplification, geometry and texture coordinates quantization, and texture compression and sub-sampling.

3000 stimuli were rated in a CS subjective experiment. This subset was selected to equitably cover the entire quality range, and to be challenging for objective quality metrics. To do so, we predicted the MOS of all the stimuli of the dataset, using 2 existing objective quality metrics that we calibrated on our previous dataset of meshes with vertex colors. The quality scores of the stimuli not included in the subjective experiment were predicted using the quality metric proposed in Chapter 7.

As the distortions in our dataset are of different natures (quantization, sub-sampling, etc.) and affect different aspects of the 3D model (geometry or color), their impact on the perceived quality is very different. We found that (1) the influence of the LoD simplification is highly dependent on the geometry quantization level. Indeed, quantization artifacts are less visible on coarse meshes. (2) The quantization of the model geometry has more impact on the visual quality than the quantization of its texture coordinates. (3) Considering the interaction between the texture compression and sub-sampling, we showed that the compression of a texture map can be pushed further than that of a natural 2D image. This is because a texture is mapped on a 3D model which is then rendered, while a natural image is directly displayed on the screen. (4) For UV maps highly quantized, the size of the texture must be decreased in order to reduce the tiling effect (artifacts caused by the texture coordinates quantization).

Regarding the impact of the model geometry and color complexity on the perception of distortions, we observed that (1) Both color and geometry can mask the geometric degradations of a quantized 3D model. (2) Models with simple/monochromatic textures are less sensitive to the UV map quantization than those with rich-detail textures. However, (3) the impact of the UV quantization on the visual quality depends not only on the geometric, and color complexity of the model but also on the amount of texture seams and the quality of the texture atlas packing: quantization artifacts are clearly more visible on models whose texture atlas is highly fragmented and/or not efficiently packed.

Texture seams are, to some extent, necessary in any textured mesh. However, the presence of a large number of redundant texture seams caused by the fragmentation of the UV atlas is detrimental and poses constraints on processing operation performed over the mesh data structure, such as hindering compression to some extent. Indeed, quantizing the UV map with too few bits can seriously degraded models exhibiting a large number of texture seams.

5.6 Conclusion

We produced the largest textured meshes quality assessment dataset at present, with more than 343k distorted meshes derived from 55 source models corrupted by combinations of 5 real-world distortions, related to compression and simplification, applied on the geometry and texture. A carefully selected subset of 3000 stimuli were annotated in a large-scale CS quality assessment experiment, wherein 4513 participants were involved and more than 148k quality judgments were collected. The quality scores of the remaining distorted stimuli in the dataset were predicted using a quality metric based on CNN, adapted for this kind of data. This dataset allowed us to draw interesting conclusions regarding the impact of each distortion as well as that of their combinations on the perceived quality. We also evaluated the influence of the complexity of the geometry, color and texture seams on the perception of distortions. Regarding the characterization of 3D content, we proposed three measures, based on spatial information and visual attention complexity, to quantitatively characterize the geometric, color and semantic complexity of 3D models.

The dataset of textured meshes along with the subjective scores (MOSs and Pseudo-MOSs) will be made publicly available online to support further studies on 3D graphics quality assessment, including understanding human behavior in assessing perceived quality and facilitating the creation and evaluation of objective quality measures for 3D content.

This dataset as well as the previous one with meshes with vertex colors, served us to develop two new perceptual quality assessment metrics for 3D meshes with color attributes. These metrics are presented in the following chapters.

Part II

Objective Quality Assessment

Chapter 6

A Model-Based Perceptual Quality Metric for 3D Meshes with Vertex Colors

Most metrics in the literature were designed for 3D content without appearance attributes; they evaluate only geometric distortions. When it comes to 3D content with color information, very few works have been published (see Chapter 1, section 1.2). Actually, constructing a quality assessment metric for this kind of data is no trivial task. The main reasons are: (1) the multimodal nature of the data (it is still unclear how color and geometry artifacts affect perceived quality) and (2) the complex processing pipeline that constructs the final rendered image from the 3D content (computation of light-material interactions, viewpoint selection, and rasterization).

In this chapter, we address the problem of objective quality assessment of 3D meshes with vertex colors. We proposed the first quality metric for such data that operates entirely on the mesh domain. This metric, called *Color Mesh Distortion Measure (CMDM)*, incorporates perceptually-relevant geometry and color features and is based on a data-driven approach that overcomes the aforementioned challenges.

Despite the fact that most existing model-based metrics (metrics operating on the 3D model itself) ignore the visual saliency of 3D models (see Chapter 1, section 1.2), we believe that this factor has a crucial influence on perceived quality and that the appropriate combination of this information with the geometric and appearance attributes of the 3D model can improve the prediction of its perceived quality. Based on this hypothesis/assumption, we devised an extension of *CMDM* by combining its geometry and color features with a new perceptual feature motivated by visual saliency: the Visual Attention Complexity (VAC) feature. We call this metric *CMDM-VAC*.

The key contributions of this work can be summarized as:

- We evaluate individually the performance of a set of perceptually-relevant geometry-based and color-based features for predicting the perceived visual quality of colored meshes.
- We provide a perceptually-validated metric for measuring the quality of meshes with vertex colors. To the best of our knowledge, our proposed metric is the first attempt to integrate both geometry and color information for quality assessment of such data.

The source code of the metric is made publicly available¹ on the MESH Processing Platform (MEPP) to support further research in this area.

- We demonstrate that incorporating a visual attention complexity measure into perceptual quality metrics of 3D content improves their performance.
- We investigate how knowledge of the viewpoint of 3D models may improve the results of objective quality metrics.

This chapter is organized as follows: section 6.1 describes the proposed metric. Section 6.2 evaluates its performance on two datasets and compares it to state-of-the-art image and mesh quality metrics. We provide in section 6.3 an approach to integrate visual saliency into quality metrics. The study on integrating the viewpoint into objective metrics is presented in section 6.4 along with its results. Concluding remarks are presented in section 6.5.

6.1 Toward a quality assessment metric for colored 3D meshes

The metric we propose is a full-reference multiscale metric that operates entirely on the mesh domain, at vertex level: It is based on curvature and color statistics computed on local corresponding neighborhoods from the original and distorted models. The metric is largely inspired by the MSDM2 frameworks from which we take the curvature features and the neighborhood correspondence mechanisms [80]. To address the color-related aspects of our metric, we consider the features introduced in the 2D image-difference framework of Lissner et al. [155]. We refer to our metric as *Color Mesh Distortion Measure (CMDM)*.

Our framework is as follows: for given distorted M_{dist} and reference M_{ref} meshes, we first establish a correspondence between M_{dist} and M_{ref} (see section 6.1.1). Then for each scale h_i , we define a spherical neighborhood around each vertex v of M_{dist} (see section 6.1.2) and compute a set of local geometry-based and color-based features over the points belonging to the neighborhood of v and their corresponding points on M_{ref} (see section 6.1.3). Local single-scale feature values are pooled into global multiscale features f_j . Finally, *CMDM* is defined as a linear combination of an optimal subset of features determined through logistic regression (see section 6.1.4). An overview of the proposed metric is shown in Figure 6.1.

6.1.1 Correspondence between meshes

The first objective is to establish a correspondence between the meshes being compared (M_{dist} and M_{ref}). Thus, we match each vertex v of the distorted mesh M_{dist} with its nearest 3D point \hat{v} on the surface of the reference mesh M_{ref} using a fast asymmetric projection (as in MSDM2, we consider the AABB tree structure from CGAL [156]). Then, for each projected 3D point (\hat{v}), we compute its curvature and color using barycentric

¹<https://github.com/MEPP-team/MEPP2>

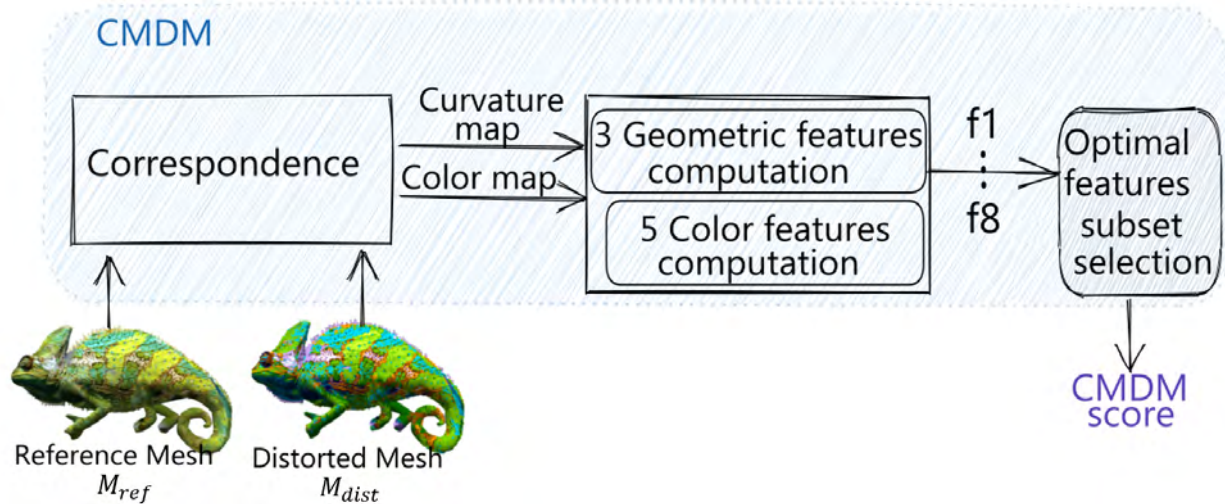


Figure 6.1: Overview of the proposed metric *CMDM*.

interpolation from vertices of the triangle to which it belongs. This way, each vertex from M_{dist} has a corresponding point on M_{ref} (with a curvature and a color value).

The correspondence is scale-independent: it takes place once, only at the beginning of the process. Nevertheless, the curvature and color values of \hat{v} are updated for each scale h_i as these values depend on the considered neighborhood which varies for each scale. This is explained in the following subsections.

6.1.2 Neighborhood Computation

As stated above, the features used in our metric are not computed globally on the entire mesh but locally at multiple scales over spherical neighborhoods around each vertex. Thus as in [80], we define, for each scale h , a neighborhood $N(v, h)$ of radius h around each vertex v of M_{dist} as the connected set of vertices belonging to the sphere with center v and radius h . We also add to this neighborhood the intersections between this sphere and the edges of M_{dist} . The curvature and color values of the intersection points are interpolated. Then, we consider for the set of points belonging to $N(v, h)$ their projected 3D points on M_{ref} (corresponding neighborhood of \hat{v}). Features are computed by considering curvature and color statistics over $N(v, h) \in M_{dist}$ and $N(\hat{v}, h) \in M_{ref}$.

6.1.3 Perceptually relevant features

For each scale h , the following 8 features are computed locally for each vertex v of M_{dist} over the points belonging to its spherical neighborhood $N(v, h)$ and to the neighborhood $N(\hat{v}, h)$ of its corresponding vertex \hat{v} on M_{ref} .

A. Geometry-based features

These features are based on mean curvature information defined at multiple scales. To compute curvature, we adopted the method developed by Alliez et al. [157], which evaluates the curvature tensor on a geodesic neighborhood around each vertex. This method is interesting and robust because it avoids the problem of sensitivity to connectivity (M_{dist} and M_{ref} do not necessarily share the same connectivity nor the same level of details). Note that, we used a radius $r = \frac{h}{3}$ for the computation of the curvature as a good compromise between small radii which capture tiny details and larger radii which provide strong smoothing effects.

As in [80], we consider the following geometry features:

$$\text{Curvature comparison } f_1^h(v) = \frac{\|\overline{C}_v^h - \overline{C}_{\hat{v}}^h\|}{\max(\overline{C}_v^h, \overline{C}_{\hat{v}}^h) + k} \quad (6.1)$$

$$\text{Curvature contrast } f_2^h(v) = \frac{\|\sigma_{C_v^h} - \sigma_{C_{\hat{v}}^h}\|}{\max(\sigma_{C_v^h}, \sigma_{C_{\hat{v}}^h}) + k} \quad (6.2)$$

$$\text{Curvature structure } f_3^h(v) = \frac{\|\sigma_{C_v^h} \sigma_{C_{\hat{v}}^h} - \sigma_{C_v^h C_{\hat{v}}^h}\|}{\sigma_{C_v^h} \sigma_{C_{\hat{v}}^h} + k} \quad (6.3)$$

where k is a constant to avoid instability when denominators are close to zero ($k = 1$ as in [80]). \overline{C}_v^h and $\overline{C}_{\hat{v}}^h$ are Gaussian-weighted averages of curvature over the points belonging to the neighborhoods $N(v, h)$ and $N(\hat{v}, h)$, respectively. Similarly, $\sigma_{C_v^h}$, $\sigma_{C_{\hat{v}}^h}$ and $\sigma_{C_v^h C_{\hat{v}}^h}$ are Gaussian-weighted standard deviations and covariance of curvature over these neighborhoods.

B. Color-based features

To compute the color features, we first transform the RGB color values of each vertex of the meshes being compared (M_{dist} and M_{ref}) into the perceptually uniform color space LAB200HL. Lissener et al. [158] recommended working in this color space since there is little cross contamination between the color attributes (lightness, chroma, hue). Each vertex v has of a lightness and two chromatic values (L_v , a_v , b_v). The chroma of the vertex v is defined as follows: $Ch_v = \sqrt{a_v^2 + b_v^2}$.

We have adapted the 2D image features, proposed by [155], to the 3D meshes. These features do not only take into account the luminance but also the chroma and hue components to better assess the chromatic distortions.

$$\text{Lightness comparison} \quad f_4^h(v) = \frac{1}{c_1(\overline{L}_v^h - \overline{L}_{\hat{v}}^h)^2 + 1} \quad (6.4)$$

$$\text{Lightness contrast} \quad f_5^h(v) = \frac{\sigma_{L_v^h} \sigma_{L_{\hat{v}}^h} + c_2}{\sigma_{L_v^h}^2 + \sigma_{L_{\hat{v}}^h}^2 + c_2} \quad (6.5)$$

$$\text{Lightness structure} \quad f_6^h(v) = \frac{\sigma_{L_v^h L_{\hat{v}}^h} + c_3}{\sigma_{L_v^h} \sigma_{L_{\hat{v}}^h} + c_3} \quad (6.6)$$

$$\text{Chroma comparison} \quad f_7^h(v) = \frac{1}{c_4(\overline{Ch}_v^h - \overline{Ch}_{\hat{v}}^h)^2 + 1} \quad (6.7)$$

$$\text{Hue comparison} \quad f_8^h(v) = \frac{1}{c_5 \overline{\Delta H}_{v\hat{v}}^h + 1} \quad (6.8)$$

where \overline{L}_v^h , $\overline{L}_{\hat{v}}^h$, \overline{Ch}_v^h and $\overline{Ch}_{\hat{v}}^h$ denote the Gaussian-weighted averages of Lightness and Chroma computed respectively over the set of points belonging to $N(v, h)$ and $N(\hat{v}, h)$. $\sigma_{L_v^h}$, $\sigma_{L_{\hat{v}}^h}$ and $\sigma_{L_v^h L_{\hat{v}}^h}$ are Gaussian-weighted standard deviations and covariance of lightness in the mentioned neighborhoods. The term $\overline{\Delta H}_{v\hat{v}}^h$ refers to the Gaussian-weighted average hue difference between $N(v, h)$ and $N(\hat{v}, h)$. It is defined as follows:

$$\Delta H_{v\hat{v}} = \sqrt{(a_v - a_{\hat{v}})^2 + (b_v - b_{\hat{v}})^2 - (Ch_v - Ch_{\hat{v}})^2} \quad (6.9)$$

The constants c_1 , c_2 , c_3 , c_4 and c_5 were set respectively to 0.002, 0.1, 0.1, 0.002 and 0.008 as in [155].

We invert the scaling of the color-based features so that they are consistent with curvature-based features (i.e. each color feature $f_j^h(v) = 1 - f_j^h(v)$). This way, a value of 0 of a geometry/color-based feature means that there is no local geometric/color distortion around vertex v . All features $\in [0, 1]$.

6.1.4 Global perceptual quality score

The set of local geometric and color features, presented in the subsection above, is computed for each vertex of the distorted mesh M_{dist} and for each scale h_i . The local multi-scale measure of a feature $f_j(v)$ is simply the average of its single-scale values.

$$f_j(v) = \frac{1}{n} \sum_{i=1}^n f_j^{h_i}(v) \quad (6.10)$$

where n is the number of scales used.

We aim to obtain a global score of visual distortion according to each feature (f_j). So,

we average the local values of each feature over all the vertices of M_{dist} .

$$f_j = \frac{1}{|M_{dist}|} \sum_{v \in M_{dist}} f_j(v) \quad (6.11)$$

where $|M_{dist}|$ is the number of vertices of the distorted mesh. The features f_j are all within the range $[0, 1]$.

Our metric is then defined as a combination of the features f_j . However, choosing the best combination model is a crucial problem. For prediction of the color-image differences, the authors in [155] used a factorial combination model, while Meynet et al. [98] considered a linear model for their point cloud quality metric. In our case, we chose to consider a linear model: (1) to make the optimization easier and (2) because we tried nonlinear models such as Minkowski pooling, which did not provide better performance. Thus, the global multiscale distortion (*GMD*) score is computed as follows:

$$GMD_{M_{dist} \rightarrow M_{ref}} = \sum_{j \in S} w_j f_j \quad (6.12)$$

S is the set of feature indexes of our linear model. w_j weights the contribution of each feature to the overall distortion prediction. $GMD_{M_{dist} \rightarrow M_{ref}}$ evaluates the distortion of the distorted model regarding the reference model. In order to strengthen the robustness of our method and to obtain a symmetric measure, we also compute $GMD_{M_{ref} \rightarrow M_{dist}}$. We retain the average as the final distortion measure *CMDM*.

$$CMDM = \frac{GMD_{M_{dist} \rightarrow M_{ref}} + GMD_{M_{ref} \rightarrow M_{dist}}}{2} \quad (6.13)$$

As in [159], the optimal subset of features of *CMDM* and their corresponding weights w_j are obtained through an optimization computed by logistic regression. The optimization is based on cross-validation, using the ground truth dataset of 3D meshes with vertex colors described in Chapter 3. The optimization is detailed later in section 6.2.4.

6.2 Results and evaluation

In this section, we evaluate the performance of our metric and compare it to state-of-the-art approaches, including 2D Image Quality Metrics (IQMs). To train and test the metric, we used the ground truth dataset of meshes with vertex colors, described in Chapter 3. We also validate our metric on a dataset from [3], composed of distorted textured meshes.

6.2.1 Dataset

The dataset used to train and test our metric is produced from a subjective study, based on the Double Stimulus Impairment Scale (DSIS) method and conducted in a VR. It is composed of 480 dynamic meshes with vertex colors. The stimuli were generated from

5 source models (“Aix”, “Ari”, “Chameleon”, “Fish”, “Samurai”) subjected to 4 types of distortion, each applied with 4 strengths. The selected distortions are: uniform quantization applied on either (1) geometry (QGeo) or (2) color (QCol), simplification algorithms that take into account either (3) the geometry only (SGeo) or (4) both geometry and color (SCol). Each stimulus was displayed in 3 viewpoints and animated with 2 short movements. Each stimulus is assigned a subjective quality score (MOS). For more details on this dataset, refer to Chapter 3, section 3.1.

In this section, we do not take into account the influence of viewpoints or that of animations. Thus, for a given stimulus, we averaged its MOSs over the 3 viewpoints and the 2 animations. Thus, the dataset used is composed of 80 stimuli.

6.2.2 Performance evaluation measures

In order to evaluate the performance of objective quality metrics, we compare the quality scores predicted by these metrics to subjective ground truth data. The standard performance evaluation measure consists in computing the Pearson Linear Correlation Coefficient (*PLCC*) and the Spearman Rank Order Correlation Coefficient (*SROCC*) between the metric predictions and subjective scores (MOSs). These indices measure, respectively, the accuracy and the monotonicity of the predictions. Note that, the Pearson correlation (*PLCC*) is computed after a logistic regression which provides a non-linear mapping between the objective and subjective scores. This allows the evaluation to take into account the saturation effects associated with human senses.

However, the correlations ignore the uncertainty of the subjective scores. Therefore, as a complementary evaluation of the performance of objective metrics, we implement the framework recently proposed by Krasula et al. [160]. This framework consists in determining the classification abilities of the metrics according to two scenarios:

- (A) Different vs. Similar: this analysis assesses how well can the metric distinguish between significantly different and similar pairs of stimuli. The first step consists in determining the pairs of stimuli in the dataset rated significantly different. To do so, we conduct a statistical test (t-test) on the raw subjective scores. Then for each pair of stimuli (i, j) , we compute the absolute difference of the predicted scores ($|\Delta_{\text{PredictedScores}}(i, j)|$) and measure how well these values are able to correctly classify the pairs of stimuli.
- (B) Better vs. Worse: this analysis is performed on the significantly different pairs only. The significantly different pairs (i, j) are divided into 2 groups: i better than j ($(\Delta_{\text{MOS}}(i, j) > 0)$) and i worse than j ($(\Delta_{\text{MOS}}(i, j) < 0)$). We then measure, according to $\Delta_{\text{PredictedScores}}(i, j)$ values, how well the metric is able to predict this classification.

As can be seen, these scenarios take into account the uncertainty of the subjective scores. Both scenarios refer to a binary classification problem (different/similar and better/worse). As a performance indicator, we consider the Receiver Operating Characteristic (ROC) and, more precisely, the Area Under the Curve (AUC) values. The AUC is a direct indicator of the performance/ability of the classifiers (1.0 corresponds to a perfect classification, 0.5

corresponds to a random one). In what follows, we refer to the AUC values by AUC_{DS} and AUC_{BW} for scenarios (A) and (B), respectively.

6.2.3 Single feature prediction performance

We evaluated the prediction performance of each feature implemented in the proposed metric. Table 6.1 shows the correlations of the individual features with the MOSs, as well as their classification abilities.

Table 6.1: Performance of individual features.

Feature	Id	$PLCC$	$SROCC$	AUC_{DS}	AUC_{BW}
Curvature comparison	f_1	0.5	0.44	0.6	0.75
Curvature contrast	f_2	0.45	0.43	0.59	0.73
Curvature structure	f_3	0.3	0.32	0.53	0.67
Lightness comparison	f_4	0.58	0.69	0.69	0.83
Lightness contrast	f_5	0.7	0.71	0.7	0.87
Lightness structure	f_6	0.68	0.71	0.69	0.87
Chroma comparison	f_7	0.38	0.59	0.64	0.78
Hue comparison	f_8	0.33	0.43	0.6	0.71

Overall, the best features are those based on the lightness information (especially f_5 , f_6). They correlate well with the subjective scores and provide a good performance in identifying the significantly different stimuli (AUC_{DS}) as well as the stimuli of better quality (AUC_{BW}). For the geometry-based features, f_1 and f_2 perform better than f_3 . However, this does not necessarily indicate the ineffectiveness of f_3 . Finally, regarding the chromatic features, chroma comparison f_7 performs slightly better than hue comparison f_8 . Note that, the geometry-based features are penalized by the color quantization (QCol), since this type of distortion is applied only on the vertex colors and does not affect the model geometry at all. Removal of this distortion improves their performance, notably with respect to correlations. $PLCC$ and $SROCC$ increase to 0.7 for f_1 and f_2 . This is shown in Table 6.2.

Table 6.2: Performance of geometry-based features after removal of the color quantization distortion.

Feature	Id	$PLCC$	$SROCC$	AUC_{DS}	AUC_{BW}
Curvature comparison	f_1	0.78	0.752	0.647	0.902
Curvature contrast	f_2	0.723	0.736	0.616	0.882
Curvature structure	f_3	0.502	0.558	0.53	0.789

6.2.4 Toward an Optimal Combination of features

Our metric contains 8 different features f_j . In this 8 dimensional space, some features are obviously more significant than others. Also, features may be redundant with one another, and if all the features are taken into account, this could potentially lead to an overfitting. Therefore, in the same vein as [159], we conducted two Leave-One Out Cross-Validation tests (LOOCV) on the dataset, described in subsection 6.2.1, to select an optimal subset of features. Each cross-validation test divides the dataset into a training set that serves to optimize feature weights (w_j in Eq. 6.12) using linear regression and a test used for testing the obtained metric.

1. We split the training and test sets according to the source models. Given that there are 5 sources in our dataset, we leave 1 source model and its distortions out for testing, while the remaining stimuli (4 models \times 16 distorted stimuli) are used for training. Thus after 5 folds, each source model is used as a test set.
2. Similar to test (1), but we divide the dataset according to the distortion types (regardless of the models). We train the metric on 3 distortion types out of 4 (5 models \times 12 distorted stimuli) and test on the fourth type. Thus after 4 folds, each distortion type is used once for testing.

These 2 types of LOOCV tests provide a good measure of the robustness of our metric. We exhaustively searched through all possible combinations of features (255 combinations), and selected the feature-subset that generates the best average performance of *CMDM* over all the test sets (9 folds) in terms of the mean of *PLCC* and *SROCC*. We obtained that the final model of our metric is composed of only 4 features: Curvature contrast (f_2), Lightness contrast (f_5) and structure (f_6) and chroma comparison (f_7). The optimal features found are consistent with the results of the single feature performance (see subsection 6.2.3).

6.2.5 Performance evaluation and comparisons

In this subsection, we present the results of the cross-validation tests, described in the previous subsection. As an ablation study, we compare our metric with two of its versions trained with different subsets of features: *CMDM_Geo* that takes into account only the geometry features and *CMDM_Col* based only on color features. As a baseline, we also include results of a classical color distance *D_LAB*, which is the average of the color difference (in the LAB2000HL color space) computed symmetrically between the reference and the distorted model. Finally, we compare our metric with 3 state-of-the-art full-reference image quality metrics (IQMs): *SSIM* [81], *HDR-VDP2* [108], *iCID* [110]. To apply these IQMs, we generate for each 3D object in our dataset, a set of 18 snapshots taken from different viewpoints (camera positions regularly sampled, as shown in Figure A.11 in the appendix). The global quality score of a stimulus, given by an IQM, is the average of the objective scores over all its snapshots. The parameters of the IQMs were set as follows: For *SSIM*, we considered a local window of size 11×11 pixels. For the resolution used

for *HDR-VDP2*, we considered 33.7 pixels per degree, which corresponds to the following experimental setting: stimuli presented on a calibrated 23" LCD display (resolution 1920×1080 pixel) at a constant viewing distance of 0.5m. The peak sensitivity parameter of *HDR-VDP2* was set to 2.4 and the selected output from this metric was the quality prediction Q . For the *iCID* metric, we considered the default parameters: equal weight of lightness, chroma, and hue, and use of chroma contrast and chroma structure.

Figure 6.2 compares the overall performance of the tested metrics for the 2 cross-validation scenarios presented in 6.2.4. Tables 6.3 and 6.4 detail the results of each test set.

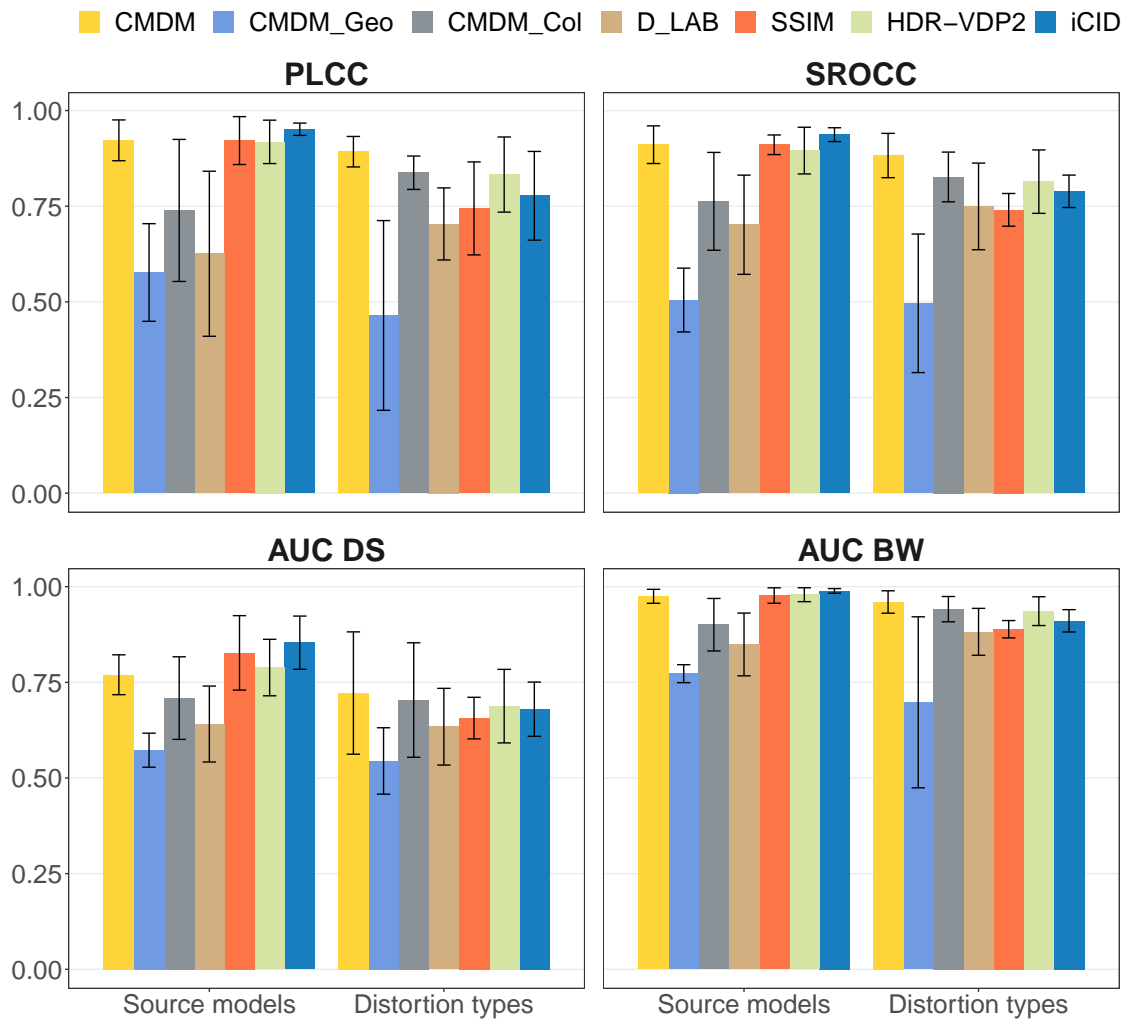


Figure 6.2: Performance comparison of several metrics in two cross-validation tests. Mean performance evaluation measures are reported. Error bars indicate the standard deviation over the test sets.

For the LOOCV test according to the source models, Figure 6.2 demonstrates that *CMDM* outperforms the tested model-based metrics. It shows almost the same performance as the IQMs in terms of correlations and detection of better quality stimuli (AUC_{BW}). IQMs provide better results in identifying the significantly different pairs of stimuli (AUC_{DS}). We believe that this is primarily related to the advantage of IQMs over our metric and other model-based metrics regarding their natural incorporation/knowledge of the entire rendering

Table 6.3: Performance comparison of several metrics in a cross-validation test among source models

	Aix				Ari				Chameleon				Fish				Samurai			
	PLCC	SROCC	AUC _{DS}	AUC _{BW}	PLCC	SROCC	AUC _{DS}	AUC _{BW}	PLCC	SROCC	AUC _{DS}	AUC _{BW}	PLCC	SROCC	AUC _{DS}	AUC _{BW}	PLCC	SROCC	AUC _{DS}	AUC _{BW}
CMDM	0.958	0.956	0.783	0.982	0.96	0.91	0.823	0.986	0.83	0.83	0.692	0.943	0.93	0.914	0.805	0.987	0.933	0.944	0.746	0.976
CMDM_Geo	0.53	0.621	0.562	0.79	0.68	0.468	0.577	0.788	0.457	0.474	0.504	0.76	0.554	0.554	0.622	0.788	0.462	0.407	0.598	0.737
CMDM_Col	0.778	0.791	0.793	0.913	0.491	0.553	0.633	0.799	0.764	0.788	0.631	0.914	0.941	0.903	0.866	0.99	0.76	0.779	0.631	0.887
D.LAB	0.791	0.826	0.77	0.924	0.282	0.497	0.523	0.737	0.776	0.747	0.609	0.897	0.734	0.779	0.713	0.9	0.546	0.659	0.59	0.787
SSIM	0.896	0.909	0.782	0.959	0.973	0.932	0.924	0.993	0.823	0.868	0.683	0.951	0.959	0.929	0.9	0.992	0.957	0.915	0.846	0.989
HDR-VDP2	0.893	0.853	0.728	0.958	0.976	0.947	0.877	0.998	0.849	0.818	0.727	0.963	0.895	0.897	0.751	0.981	0.978	0.962	0.86	0.995
iCID	0.958	0.932	0.849	0.983	0.953	0.929	0.85	0.989	0.924	0.921	0.743	0.986	0.954	0.935	0.912	0.988	0.966	0.968	0.914	0.999

Table 6.4: Performance comparison of several metrics in a cross-validation test among distortion types

	QGeo				QCol				SGeo				SCol			
	PLCC	SROCC	AUC _{DS}	AUC _{BW}	PLCC	SROCC	AUC _{DS}	AUC _{BW}	PLCC	SROCC	AUC _{DS}	AUC _{BW}	PLCC	SROCC	AUC _{DS}	AUC _{BW}
CMDM	0.882	0.825	0.537	0.933	0.917	0.924	0.893	0.973	0.93	0.94	0.871	0.995	0.841	0.841	0.641	0.939
CMDM_Geo	0.686	0.481	0.597	0.8	0.121	0.288	0.457	0.373	0.596	0.73	0.638	0.874	0.455	0.486	0.486	0.745
CMDM_Col	0.787	0.758	0.493	0.904	0.821	0.845	0.772	0.943	0.889	0.908	0.838	0.984	0.853	0.795	0.712	0.934
D.LAB	0.653	0.677	0.501	0.826	0.799	0.851	0.72	0.932	0.765	0.841	0.702	0.938	0.598	0.629	0.613	0.832
SSIM	0.875	0.794	0.709	0.903	0.792	0.708	0.696	0.896	0.722	0.756	0.623	0.901	0.588	0.704	0.598	0.855
HDR-VDP2	0.946	0.938	0.805	0.987	0.882	0.78	0.724	0.939	0.736	0.762	0.59	0.9	0.767	0.777	0.632	0.918
iCID	0.88	0.838	0.632	0.926	0.86	0.81	0.785	0.939	0.738	0.761	0.645	0.906	0.631	0.747	0.657	0.872

pipeline. Indeed, IQMs operate on snapshots that consider the same rendering, apparent brightness and lighting conditions as those seen by participants during the subjective test. On the contrary, our metric only considers 3D data, without any knowledge of the rendering conditions.

Considering the LOOCV test among the distortions, we notice that our metric performs better than other metrics, including IQMs. The color-based version of our metric (*CMDM_Col*) also produces good results. IQMs show a significant decrease in performance, compared to the LOOCV based on source models. These observations corroborate previous results by Lavoué et al. [69]: image-based metrics perform very well when evaluating the quality of different versions of a single source, yet they are less accurate when differentiating/ranking distortions applied on different sources. More details are provided in Tables 6.3 and 6.4.

From Table 6.3, it can be seen that our metrics and the IQMs demonstrate a good stability over the models. The performance (especially the correlations) of the *D_LAB* and *CMDM_Col* metrics drop dramatically for the model “Ari”. We also notice that the quality of the “Chameleon” model was the hardest to predict, since almost all the metrics (except *D_LAB* and *CMDM_Col*) exhibit a poorer performance on this model than that on the other models. This is coherent with our findings in section 3.4.3 of Chapter 3, which showed the “Chameleon” model was the most difficult to rate among all source models. It had the largest confidence intervals and the highest content ambiguity. As explained in Chapter 3 (sections 3.4.3 and 3.5), we believe this is related to the fact that this model carries more geometric and color details than all the other models.

When considering each distortion type separately (Table 6.4), several observations can be made. First, our metric performs very well on 3 types of distortion out of 4: For QCol, SCol and SGeo, it outperforms significantly the other metrics, and particularly the IQMs. However, our metric shows a poor performance when distinguish between similar and different pairs corrupted by geometric quantization (QGeo). For this distortion, *HDR-VDP2*

performs significantly better in terms of correlations and classification abilities. *CMDM* seems to underestimate the impact of geometric quantization (QGeo), which is particularly harmful for such high-resolution models. We believe that this is due to the fact that this distortion superimposes the vertices of the stimulus, meaning that we cannot know or control exactly which vertex color is taken into account in Unity’s import and render pipelines. This case points out an advantage for image-based quality metrics and highlights the importance of taking the rendering into account when evaluating visual quality.

6.2.6 Recommended weights

To provide the final recommended model of our metric, we averaged the weights obtained for each training subset of the two LOOCV tests. *CMDM* is thus defined, for the three selected scales ($h_i \in \{0.003BB, 0.0045BB, 0.006BB\}$, where *BB* is the maximum length of the Axis-Aligned Bounding Box (AABB) of the stimulus), as follows:

$$CMDM_{rec} = 0.091f_2 + 0.22f_5 + 0.032f_6 + 0.656f_7 \quad (6.14)$$

In order to reveal the relative importance of each of the 4 features, we scaled the weights presented in the above equation with the standard deviation of the features. The recommended weights, as well as the importance of each feature, are reported in Table 6.5.

Table 6.5: Weights and importance of the optimal subset of features for *CMDM*.

Feature	f_2	f_5	f_6	f_7
Recommended weight	0.091	0.22	0.032	0.656
Importance	0.33	0.46	0.07	0.136

The curvature and lightness contrast features (f_2 and f_5) have the highest overall importance. It would seem that observers are particularly sensitive to artifacts that harm the contrast (both geometric and color contrasts).

We evaluate the performance of the tested metrics, including *CMDM_{rec}*, on the whole dataset of 80 stimuli, described in subsection 6.2.1 (not in cross-validation). The results are reported in Table 6.6. Figure 6.3 shows the subjective scores with respect to objective metric values.

Table 6.6: Performance comparison of different metrics on the dataset of 80 meshes with vertex colors.

	<i>PLCC</i>	<i>SROCC</i>	<i>AUC_{DS}</i>	<i>AUC_{BW}</i>
<i>CMDM_{rec}</i>	0.913	0.9	0.782	0.968
<i>CMDM_{Geo}</i>	0.501	0.437	0.604	0.749
<i>CMDM_{Col}</i>	0.745	0.746	0.732	0.893
<i>D.LAB</i>	0.55	0.603	0.651	0.805
<i>SSIM</i>	0.797	0.799	0.716	0.912
<i>HDR-VDP2</i>	0.853	0.84	0.703	0.944
<i>iCID</i>	0.825	0.83	0.747	0.924

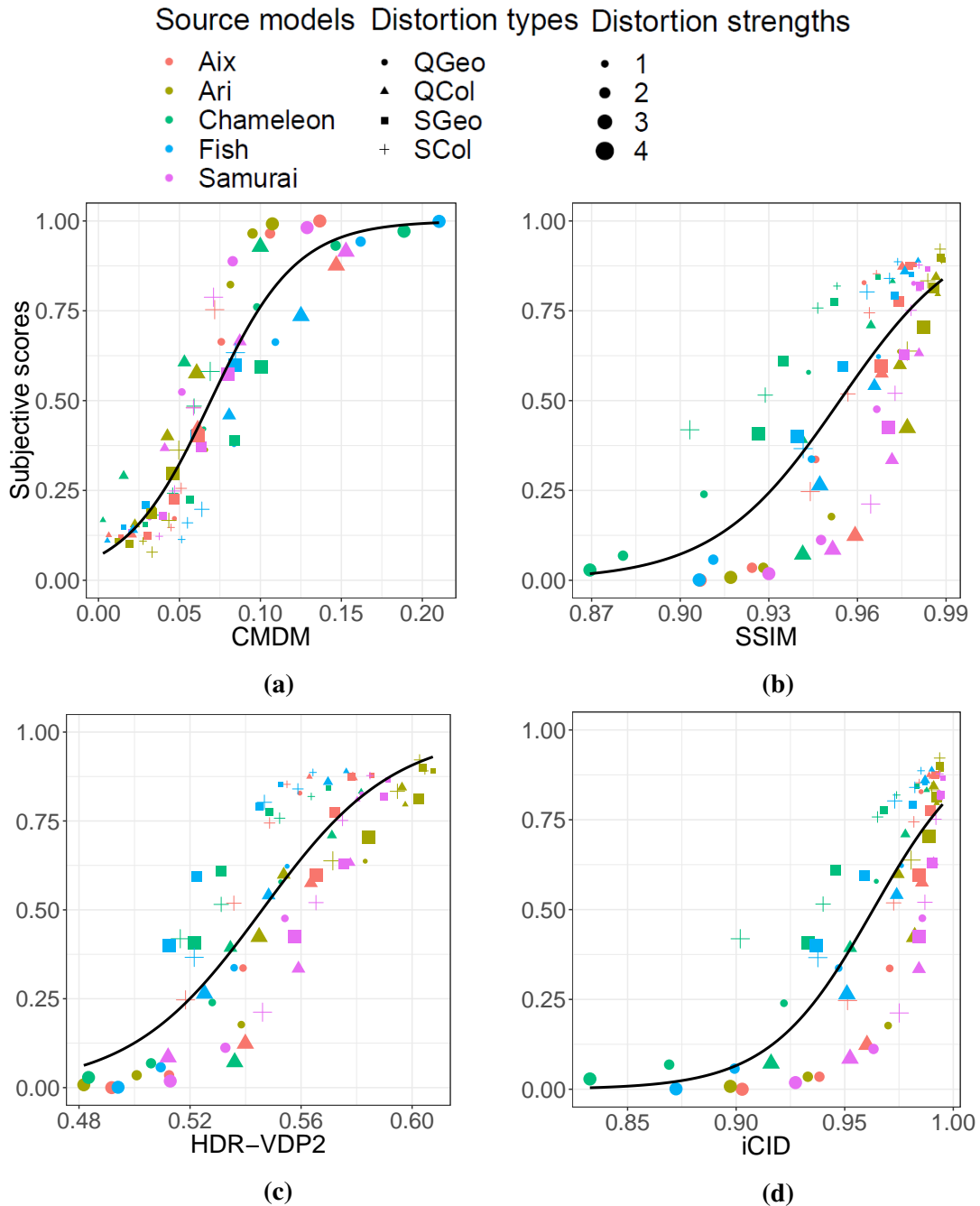


Figure 6.3: Scatter plots of subjective scores versus objective metric values for the dataset of meshes with vertex colors. Each point represents one stimulus. The fitted logistic function is displayed in black.

CMDM performs notably better than the others in terms of correlations. Moreover, the AUC values reflect its good classification abilities in both Different vs. Similar and Better vs. Worse analyses. This shows the good robustness of our metric: *CMDM* is able to differentiate and rank stimuli from different sources and different distortions.

6.2.7 Validation on a dataset of textured 3D meshes

To evaluate the robustness of our recommended metric (eq. 6.14) and to verify that it did not just learn the distortions that are specific to our dataset (subsection 6.2.1), we tested $CMDM_{rec}$ on a new dataset. Only few subject-rated datasets of 3D models with color attributes are available to the scientific community [3] [1]. We consider the LIRIS Textured Mesh Database [3], produced from a subjective study based on a pairwise comparison method. This dataset is composed of 136 textured meshes, obtained from 5 source models subjected to distortions applied to texture and geometry. Indeed, the authors generated 20 distorted versions of each source. They also selected a model (the ‘‘Dwarf’’) among the 5, and associated it with 36 mixed distortions (combination of geometry and texture distortions). To evaluate the robustness of our approach, we selected the most difficult subset, namely the one containing mixed distortions. The source model of this subset is a scan of a dwarf statue that has been reconstructed into a textured mesh of 250004 vertices. Distortions are combinations of 3 geometry distortions (geometric quantization, simplification, smoothing), each applied with 2 strengths, and 2 texture distortions (JPEG compression, sub-sampling), each applied with 3 strengths. As can be seen, these compound distortions differ significantly from the distortions of the dataset used to train $CMDM$. Before applying our metric, we transferred the texture color information into vertex colors (we assigned to each vertex a color chosen from the texture map corresponding to this vertex).

The results are summarized in Table 6.7. We included results of the IQMs presented previously, as well as the results obtained by Guo et al. [3] for different metrics: three metrics applied on rendered videos of the stimuli (the Discrete Cosinus Transform-based (DCT) metric [161], the $PSNR$ and the $MS-SSIM$ [162] applied on all frames and averaged) and three metrics directly applied on textured meshes (FQM [77] which combines the MSE computed on the mesh vertices and that computed on the texture pixels, CM_1 and CM_2 [3] both defined as a linear combination of mesh quality and texture quality). Note that, Table 6.7 shows only the correlation measures since subjective scores are derived from a paired-comparison experiment and are therefore not associated with CIs. The subjective scores with respect to the values of the objective metrics are illustrated in Figure 6.4.

Table 6.7: Performance comparison of different metrics on the LIRIS Textured dataset [3]. For metrics marked with a *, the values are reprinted from [3].

	PLCC	SROCC
$CMDM_{rec}$	0.862	0.872
SSIM	0.624	0.657
HDR-VDP2	0.83	0.844
iCID	0.502	0.552
Video-DCT*	0.32	0.50
Video-PSNR*	0.33	0.58
Video-MS-SSIM*	0.67	0.66
FQM*	0.64	0.66
CM_1 *	0.74	0.77
CM_2 *	0.80	0.85

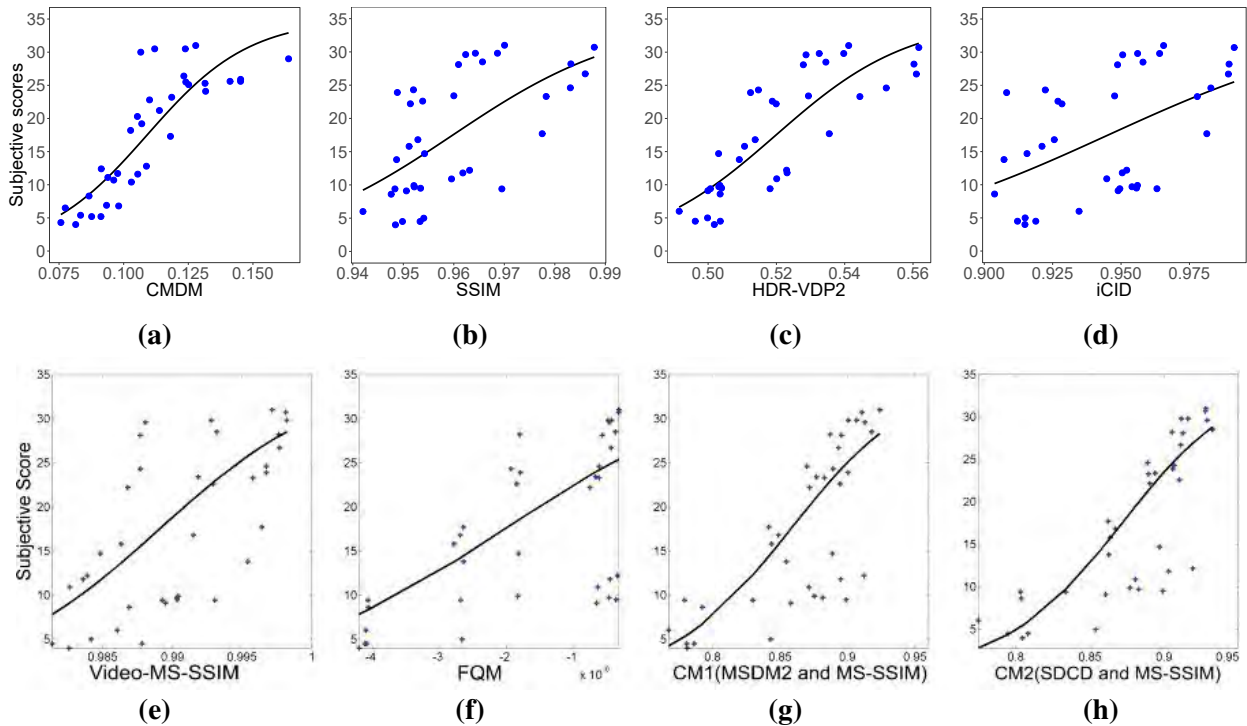


Figure 6.4: Scatter plots of subjective scores versus objective metric values for the LIRIS Textured Mesh Database [3]. Each point represents one stimulus. The fitted logistic function is displayed in black. Plots (e), (f), (g) and (h) are reprinted from [3].

Our metric provides the best results, although it was trained on a different dataset presenting different source models, different distortions and even a different color representation. *SSIM* and *iCID* show poor performance. They may be affected by the fact that the snapshots used do not have the same rendering and lighting conditions as those of the experiment. This shows the advantage of rendering independent metrics (model-based metrics) in cases where the rendering conditions are not controlled or known. Note that, the results of *SSIM* computed using snapshots of the stimuli are consistent with those reported in [3], which are computed on the rendered videos used in the subjective test.

Our metric also outperforms CM_2 , which represents the state-of-the-art of textured mesh quality assessment, and which was learned on this dataset. CM_2 is a global combination of texture and mesh distortion metrics (optimal combination of *MS-SSIM* and the Standard Deviation of Curvature Difference *SDCD* quality index). This tends to validate the fact that operating fully on the mesh domain (like our metric) ensures a better performance than combining errors computed on different domains (i.e., 3D mesh and texture image). These results also confirm the robustness of our metric compared to IQMs.

6.3 Integration of visual attention complexity

Visual attention is an important feature of the human visual system. It describes the human attention distribution for a given scene [163] and allows the identification of perceptually salient regions (regions that attract the attention of observers). Saliency information (mostly in the form of saliency maps) has been used to improve the performance of image/video quality metrics [73, 102, 113, 124]. To make better use of saliency in IQMs, Zhang et al. [164] proposed a saliency dispersion measure (using computational saliency) that takes into account the *image content* (content-dependent, semantic). In fact, when saliency is spread throughout the scene, incorporating saliency in an IQM is less likely to benefit image quality prediction [165, 166], as different observers tend to look at different parts of the image which may give a low weight to some regions with high distortion. A recent work [145] introduced the Visual Attention Complexity (VAC) measure as an adaptation of the saliency dispersion measure to 3D content. Their view-based approach offers the possibility to associate a VAC score to each viewpoint of a 3D object.

As stated in the introduction of this chapter, the incorporation of the visual attention complexity into (model-based) quality metrics for 3D models has not been exploited yet. Therefore, we propose an extension of *CMDM* by integrating the Visual Attention Complexity (VAC) indicator to its geometry and color features. We refer to this metric as *CMDM-VAC*. An overview of the proposed metric is shown in Figure 6.5.

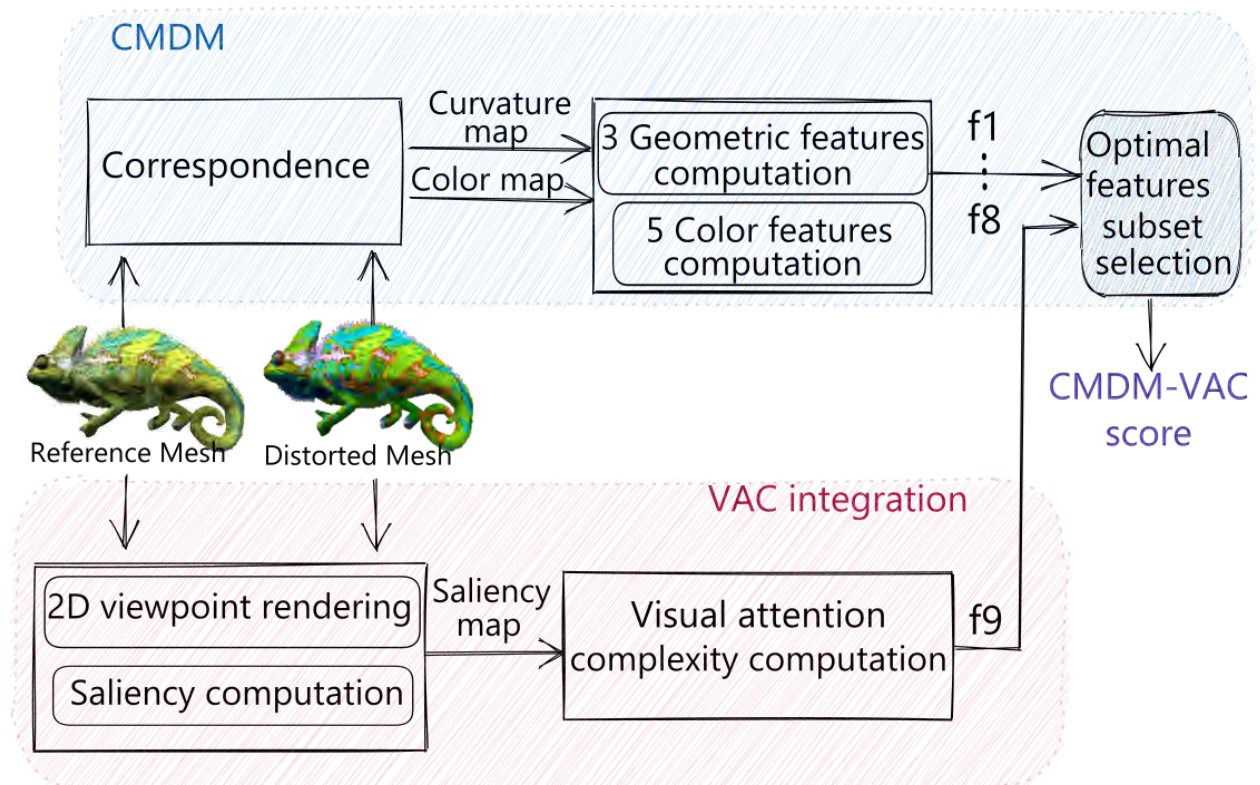


Figure 6.5: Overview of the *CMDM-VAC* metric.

6.3.1 The visual attention complexity measure

The Visual Attention Complexity (VAC) measure [145] perceptually characterizes a 3D content based on its saliency dispersion, as shown in Figure 6.6. We distinguish viewpoints with low VAC scores indicating focused 3D contents and those with high VAC scores indicating exploratory 3D contents, as detailed later in this section.

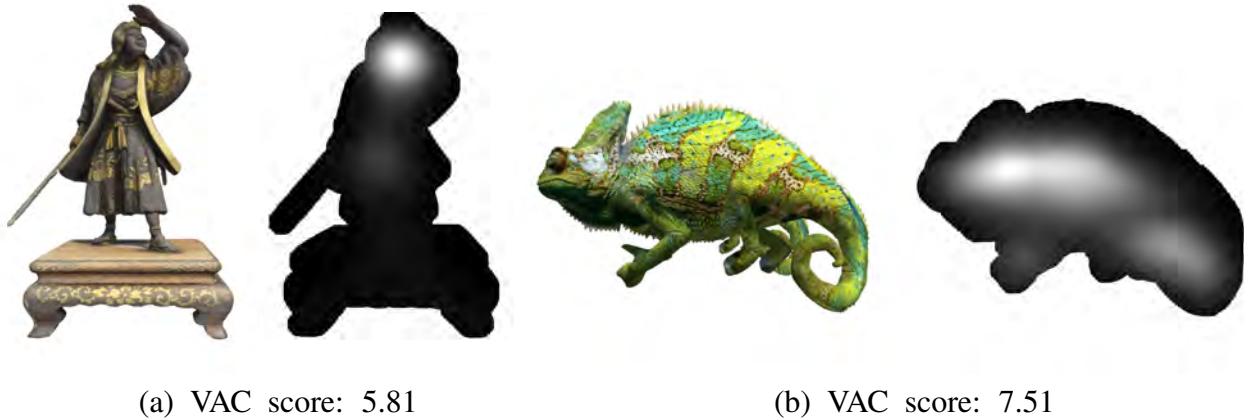


Figure 6.6: Saliency data and corresponding VAC scores. In (a) and (b), the rendered viewpoint of a 3D object is on the left, and its corresponding masked saliency information is on the right.

The VAC is computed as follows: once the viewpoints of the 3D object are generated, a computational saliency model is used to compute saliency maps. We used the Sali-con [167] computational model as recommended in [145]. The saliency information on the visible surface of the 3D object was considered with a border enlargement (using a morphological operation with a disk diameter equal to 1° of visual angle) to take into account the gazing uncertainty. Since the saliency map represents the probability of gazing at a given pixel, a normalization is applied to the saliency map so that the pixel values $\in [0,1]$. Finally, a conditional entropy is computed on the normalized saliency information, i.e., zero probability pixels are not included. Thus, the VAC score of a rendered 3D object viewpoint is defined as follows:

$$VAC\ score = - \sum_{i=1}^n p_i \log_2 p_i \quad (6.15)$$

with $p_i = h_i/K$, where h_i is the histogram for the intensity value i in the masked saliency map S (i.e., the visible surface of the 3D object), and K is the total number of pixels in S .

Low VAC score values indicate that there are strongly salient regions on the visible surface of the rendered 3D object. This is what we call focused viewpoints (see Figure 6.6.a). On the contrary, when the VAC scores are high, the saliency is diffused (i.e., overall gazing behavior). We refer to these as exploratory viewpoints (see Figure 6.6.b).

The VAC score is incorporated as a perceptual feature in *CMDM*. To ensure consistency with geometry-based and color-based features (presented earlier in section 6.1.3), we nor-

malized the VAC feature as follows:

$$\text{VAC comparison } f_9 = \frac{|VAC_{dist} - VAC_{ref}|}{VAC_{ref}} \quad (6.16)$$

All features are then within the range $[0, 1]$. The VAC feature was included in the global distortion score as the 9th feature, noted as f_9 in the linear combination of Eq. 6.12.

6.3.2 Toward an optimal combination of features

We trained and tested our metric on the dataset of 80 meshes with vertex colors, presented in subsection 6.2.1. In the same vein as *CMDM*, we performed the 2 LOOCV tests, described in section 6.2.4, to select the optimal subset of features for *CMDM-VAC*. We recall that in the first LOOCV test, we split the dataset into training and test sets based on the source models, while in the second LOOCV test, the splitting is made based on the distortion types.

As this time we have 9 different features, there are 511 possible combinations of features. Hence, we exhaustively searched through all possible combinations and selected the one that generates the best average performance of *CMDM-VAC* over all the test sets in terms of the mean of Pearson (PLCC) and Spearman (SROCC) correlations. The best model of *CMDM-VAC* that we finally obtain is composed of 5 features: Curvature contrast (f_2), Lightness contrast (f_5) and structure (f_6), hue comparison (f_8) and VAC comparison (f_9). Interestingly, the visual attention complexity feature was selected among the optimal subset of features indicating the importance of the saliency dispersion on the visual quality of artifacts. This is consistent with the single feature performance analysis described in section 6.2.3, which evaluates the performance of each feature individually in terms of correlation with MOSs: the (PLCC, SROCC) of f_9 with the MOSs are (0.61, 0.7) which ranks f_9 among the best features. The performances of the other features (geometry and color based features) are presented previously in Table 6.1.

6.3.3 Performance evaluation and comparisons

We report, in Figure 6.7, the average performance of *CMDM-VAC* over the cross-validation test sets. For comparison purposes, we included the results of *CMDM* and the IQMs obtained earlier in section 6.2.5.

In both LOOCV tests, *CMDM-VAC* performs better than *CMDM*. Considering the LOOCV test among the distortions, *CMDM-VAC* and *CMDM* outperform the IQMs showing that the latter have difficulties in ranking distortions applied on different sources [69]. The most noticeable improvements of *CMDM-VAC* were observed for the 2 test sets presented in Table 6.8.

As indicated in section 6.2.5, *CMDM* exhibits a poor performance when assessing the quality of stimuli geometrically quantized (QGeo) due to the fact that this distortion superimposes the vertices of the stimulus, which results in not knowing the exact vertex color taken into account in the rendering pipeline. Table 6.8 shows that integrating the

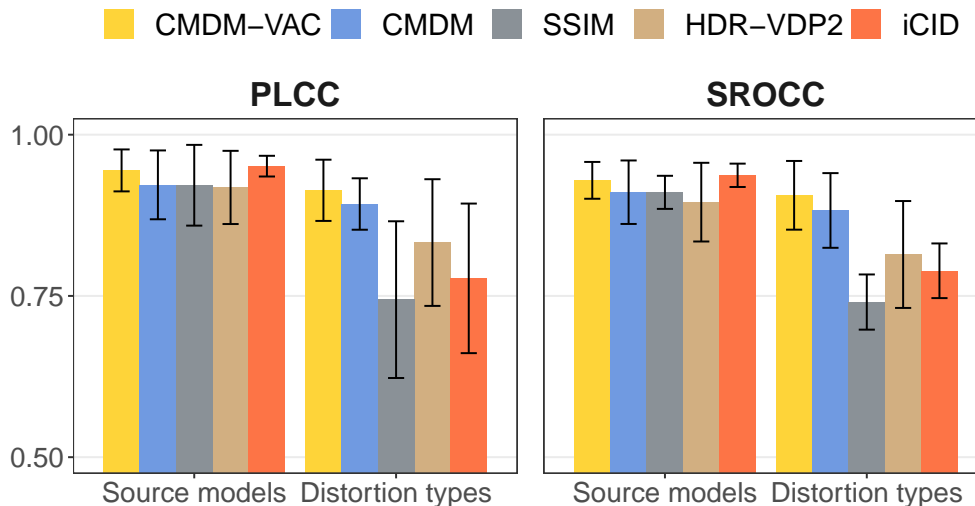


Figure 6.7: Performance evaluation of *CMDM-VAC* in two cross-validation tests. The average correlations over the test sets are reported. Error bars indicate the standard deviation over the test sets.

Table 6.8: Performance comparison of *CMDM-VAC* on two test sets. For metrics marked with a *, the values are reported from section 6.2.5.

	QGeo		Chameleon	
	<i>PLCC</i>	<i>SROCC</i>	<i>PLCC</i>	<i>SROCC</i>
CMDM-VAC	0.926	0.864	0.89	0.89
CMDM*	0.882	0.825	0.83	0.83
SSIM*	0.875	0.794	0.823	0.868
HDR-VDP2*	0.946	0.938	0.849	0.818
iCID*	0.88	0.838	0.924	0.921

VAC considerably improves the results. Indeed, the VAC is computed on a snapshot of the rendered stimulus and therefore naturally incorporates the entire rendering conditions. This once again proves the importance of integrating the rendering into the model-based quality metrics.

Regarding the second test set, it denoted in section 6.2.5 that among all the source models, the metrics perform less well on the model having the most geometry and color information: the “Chameleon”. Table 6.8 shows that visual saliency is potentially an important cue for a more effective quality assessment of complex and rich models.

6.3.4 Recommended weights

To provide the final model of *CMDM-VAC*, we averaged the weights obtained for each training set of the 2 LOOCV tests. The recommended weights, as well as the importance of each feature (defined as the weight scaled with the standard deviation of the feature), are reported in Table 6.9.

Table 6.9: Weights and importance of the optimal subset of features for *CMDM-VAC*.

Features	f_2	f_5	f_6	f_8	f_9
Recommended weights	0.092	0.202	0.028	0.182	0.496
Importance	0.293	0.359	0.051	0.165	0.131

The (PLCC, SROCC) computed over the whole dataset (80 stimuli) for *CMDM-VAC* and *CMDM* are (0.936, 0.922) and (0.913, 0.9), respectively. The performance improvement after the integration of the VAC is statistically significant (with a p-value=0.0094, obtained by a statistical test on the logistic regression residual of the 2 metrics). Figure 6.8 shows the *CMDM-VAC* values for the dataset of meshes with vertex colors compared to the subjective scores.

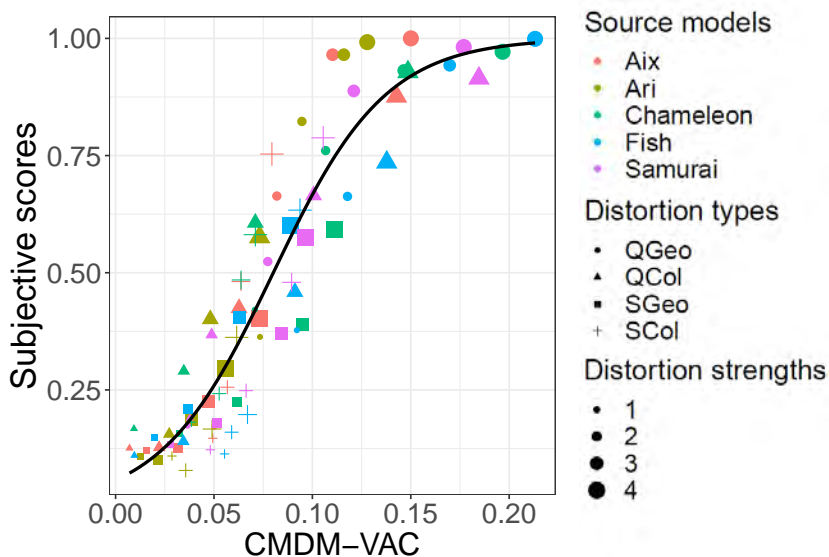


Figure 6.8: Scatter plots of subjective scores versus the *CMDM-VAC* values for the dataset of meshes with vertex colors. Each point represents one stimulus. The fitted logistic function is displayed in black

6.3.5 Performance evaluation per quality range

To assess whether *CMDM* and *CMDM-VAC* are vulnerable to the quality range of stimuli, we divided the dataset into 2 groups: (1) low-quality stimuli having MOSs $\in [1,3[$ and (2) good quality stimuli having MOSs $\in [3,5]$. Results are shown in Figure 6.9.

Among the IQMs, *HDR-VDP2* seems the best choice for estimating the perceived quality, especially for near-threshold distortions (MOSs > 3). Unlike IQMs, *CMDM-VAC* performs slightly better on low-quality stimuli than on higher quality ones. Moreover, the overall performance of our metrics seems to be more stable over quality ranges than that of IQMs.

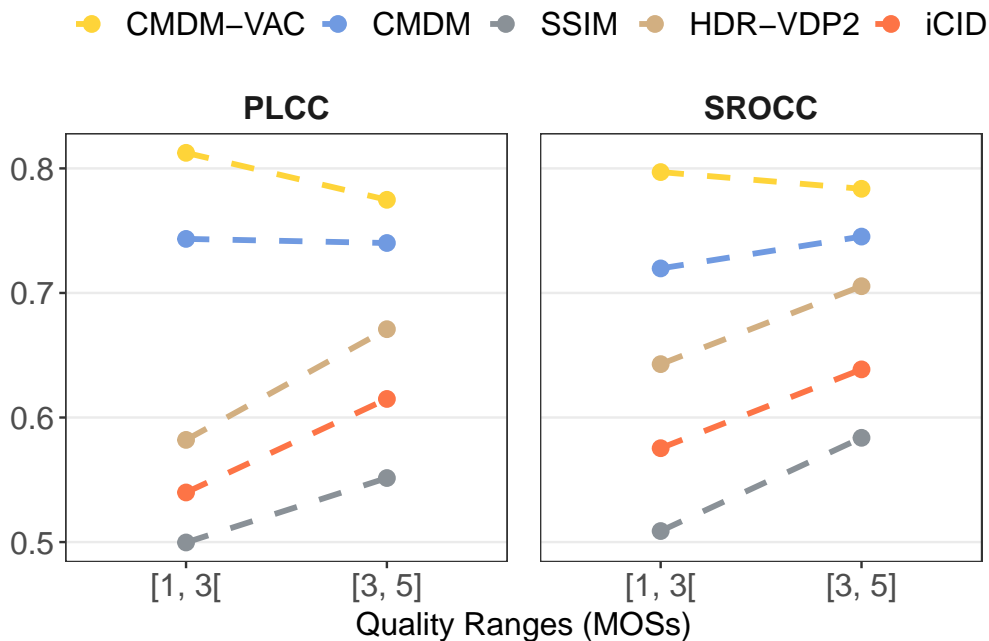


Figure 6.9: Performance evaluation of several metrics according to the quality range of stimuli.

6.4 Integration of 3D model viewpoints

According to the findings of Chapter 3 (see sections 3.4.3 and 3.5), the interaction between the distortions and the viewpoints of the stimuli has a significant impact on the quality perceived by the observers. Thus, we hypothesized that incorporating the viewpoint into *CMDM* should improve its results. Indeed, the invisible parts of the 3D model do not contribute to its visual appearance. Given a stimulus and a camera position, we determined, in a preprocessing step, which vertices are visible and which vertices are occluded by other faces of the mesh (using ray-vertex intersections). Thus, our objective metric is now computed only over the visible vertices. We can redefine Equation 6.11 as follows:

$$f_j = \frac{1}{|M'_{dist}|} \sum_{v \in M_{dist}} f_j(v) \Psi(v) \quad (6.17)$$

where Ψ is a function that returns 0 or 1 according to the visibility of vertex v and $|M'_{dist}|$ is the number of visible vertices of the distorted mesh.

To evaluate the performance of the new version of the metric $CMDM_{vis}$, we used this time a subset of 240 stimuli from the dataset of meshes with vertex colors described in section 6.2.1. Indeed, for a given stimulus, we considered its 3 viewpoints and averaged the MOSs of the 2 animations.

We tested the performance of the metric with and without integrating the visibility on these 240 stimuli. Similarly for the IQMs, we considered 2 scenarios: (1) without taking the visibility into account, so we computed the IQMs on multiple snapshots taken from different viewpoints and (2) computing the IQMs directly on the snapshot taken from the real view-

point displayed to the observer (IQM_{vis}). Tables 6.10a and 6.10b show the performance of the metrics in these 2 scenarios, while Table 6.11 shows the improvement/evolution of the metrics' performance after incorporating the viewpoint. The improvement is defined as the difference in the evaluation measures (correlations and AUCs) computed before and after integrating the viewpoint (e.g. for a given metric M : $\Delta PLCC = PLCC_{M_{vis}} - PLCC_M$).

Table 6.10: Performance comparison of different metrics in to 2 scenarios.

(a) Without integrating the viewpoint					(b) When integrating the viewpoint				
	<i>PLCC</i>	<i>SROCC</i>	<i>AUC_{DS}</i>	<i>AUC_{BW}</i>		<i>PLCC</i>	<i>SROCC</i>	<i>AUC_{DS}</i>	<i>AUC_{BW}</i>
CMDM	0.886	0.871	0.756	0.967	CMDM _{vis}	0.886	0.866	0.755	0.967
SSIM	0.773	0.768	0.697	0.915	SSIM _{vis}	0.791	0.798	0.722	0.927
HDR-VDP2	0.827	0.808	0.714	0.942	HDR-VDP2 _{vis}	0.805	0.826	0.661	0.943
iCID	0.8	0.8	0.727	0.927	iCID _{vis}	0.857	0.871	0.776	0.957

Table 6.11: Performance evolution of different metrics before and after integrating the viewpoint

	$\Delta PLCC$	$\Delta SROCC$	ΔAUC_{DS}	ΔAUC_{BW}
CMDM	0	-0.005	-0.001	0
SSIM	0.018	0.03	0.025	0.012
HDR-VDP2	-0.022	0.018	-0.053	0.001
iCID	0.057	0.071	0.049	0.03

Both versions of all metrics showed roughly the same performance in terms of correlations and classification abilities (no significant performance improvement). Our hypothesis is that this lack of improvement is due to the fact that only few stimuli of the dataset were rated significantly different across their different viewpoints. This led us to conduct a more precise study: we identified the stimuli with viewpoints associated with significantly different subjective scores. We found out that the viewpoint had a significant influence only on 88 pairs of stimuli.

Thus, instead of considering all the possible pairs of stimuli ($240 \times 239 / 2$), we compared each stimulus separately according to its 3 viewpoints VP (e.g. *Fish_SGeo_4_VP1* vs. *Fish_SGeo_4_VP2*, *Fish_SGeo_4_VP1* vs. *Fish_SGeo_4_VP3* and *Fish_SGeo_4_VP2* vs. *Fish_SGeo_4_VP3*). This limited the study to 240 pairs of stimuli ($80 \text{ stimuli} \times 3 \text{ possible combinations of viewpoint pairs}$), 88 of which were significantly impacted by the viewpoints (36%). The results are shown in Table 6.12. Note that, only the AUC values are reported since this study is based on pairs of stimuli and therefore correlations cannot be computed.

Without integrating the viewpoint information, the AUC values of all the metrics are equal to 0.5. Including the viewpoint slightly improved the results. Still, this improvement is low, except for *HDR-VDP2_{vis}*, which showed a good ability to recognize the stimulus of higher quality in a pair of stimuli.

This study takes the first step toward integrating the knowledge of the viewpoint into 3D content quality metrics. The fact that the IQMs exhibited a relatively poor performance, even though they were computed directly on the displayed rendered viewpoints, shows that

Table 6.12: Performance comparison of different metrics on the pairs of stimuli significantly affected by the viewpoints.

	$CMDM_{vis}$	$SSIM_{vis}$	$HDR-VDP2_{vis}$	$iCID_{vis}$
AUC_{DS}	0.602	0.56	0.561	0.58
AUC_{BW}	0.58	0.66	0.8	0.668

it is considerably difficult to distinguish the perceived quality of different viewpoints of the same 3D model. Further work is still needed to produce efficient metrics for this difficult scenario. In particular, we hypothesize that classical pooling should be replaced by more sophisticated pooling. It may also be useful to consider visual attention models. A starting point could be to test the performance of *CMDM-VAC* after incorporating the viewpoint.

6.5 Conclusion

In this chapter, we developed a perceptually-validated full-reference metric *CMDM* for evaluating the quality of colored 3D meshes. It is a data-driven metric that operates entirely on the mesh domain. To achieve this, we adapted a set of perceptually-relevant geometry-based and color-based features. We showed how to select an optimal subset of features using cross-validation tests and logistic regression.

We evaluated the performance of our metric, as well as that of a number of image quality metrics (IQMs), on two datasets: a dataset of meshes with vertex colors and another with textured meshes. *CMDM* provides good results in terms of correlations and classification abilities. It also demonstrates a good stability: *CMDM* is able to differentiate and rank stimuli from different sources and different distortions, unlike IQMs which perform very well when assessing the quality of different versions of a single source, but are less accurate when ranking distortions applied on different sources. Last but not least, we demonstrate that our metric can also be used for textured meshes. The code of *CMDM* was made publicly available online ².

In the second part of the chapter, we extended *CMDM* by combining its geometry and color features with the Visual Attention Complexity (VAC) measure based on the visual saliency dispersion. Integrating the VAC improved the overall performance of *CMDM* especially when assessing the quality of geometrically quantized stimuli.

Thus, including the visual attention complexity feature seems promising for improving the prediction of the visual quality of 3D content. It would be interesting to explore how the VAC can improve other perceptual quality metrics.

The last part of this chapter investigated the relevance of incorporating the viewpoint (i.e., the visible parts) of the 3D model into objective quality metrics. Further studies are still needed to effectively incorporate visibility information into quality metrics.

Finally, we would like to mention that we participate in the development of a version of *CMDM* for point clouds [98], which we do not detail in this manuscript. This metric,

²<https://github.com/MEPP-team/MEPP2>

called *Point Cloud Quality Metric (PCQM)*, considers the same initial collection of color and geometric features as *CMDM*. However, the computation of these features and the optimal combination of them differ between *CMDM* and *PCQM*, as the mesh representation differs considerably from the point cloud representation in several aspects, such as the way the data are rendered and the nature of commonly applied processing operations (and thus distortions). For more details about *PCQM*, the reader can refer to [98]. The source code of *PCQM* is also available online³.

³<https://github.com/MEPP-team/PCQM>

Chapter 7

An Image-Based Perceptual Quality Metric for 3D Graphics Based on CNN

Over the last years, Convolutional Neural Networks (CNNs) have successfully rivaled traditional Image Quality Metrics (IQMs based on bottom-up and top-down approaches, see Chapter 1 section 1.2). One reason is that they are purely data-driven: they allow for an end-to-end feature learning based on raw input data without any hand-crafted features or other types of prior domain knowledge about the human visual system or image statistics.

As discussed in the related work chapter (section 1.2.1), there is a lack of quality metrics for textured 3D meshes. Given our large-scale dataset of textured mesh described in Chapter 5 (more than 340k distorted stimuli of which 3000 stimuli are associated with subjective scores obtained from a subjective experiment), we envisaged to create an end-to-end deep-learning quality metric for such data. Given the time we had left, we could not start from scratch. Thus, we thought of taking an existing metric that employs a deep neural network for image quality assessment, and adapting it to textured meshes, then training it using our dataset. We selected the Learned Perceptual Image Patch Similarity (LPIPS) metric, proposed by et Zhang al. in 2017 [4].

LPIPS is a full-reference perceptual quality metric based on a pre-trained CNN. It evaluates the distance between two image patches. The higher the LPIPS values, the more different the patches are. Conversely, lower LPIPS values mean that the patches are more similar. LPIPS is purely data-driven. The choice of LPIPS was motivated by its many successful applications [148,149], the simplicity of the approach and the fact that it is based on an in-depth study across different architectures. In addition, LPIPS was not only trained to assess image patches similarity on parametrized distortions, but was also generalized for (validated on) real-world use cases (e.g, tasks of superresolution, frame interpolation, and image deblurring). Last but not least, the code of the metric is publicly available, as are the scripts and instructions for training and testing the metric.

The rest of this chapter is organized as follows: section 7.1 provides an overview of LPIPS. We describe in section 7.2 our approach and how the network was adapted and trained for our quality assessment tasks. Sections 7.3 and 7.4 present the results of the proposed metric over 2 datasets. Finally, limitations and concluding remarks are outlined in section 7.5.

7.1 LPIPS Overview

LPIPS was designed and trained to address the problem of “perceptual similarity” between two images/patches (x and x_0) without the need to directly fit a function to human judgments since this function may be intractable due to the high-order image structure and context dependency (many *different senses of similarity* exist) [168].

Authors employed a Siamese network for feature extraction of the inputs x and x_0 . In Siamese networks, the inputs are processed in parallel by two networks F sharing their synaptic connection weights. Feature extraction is followed by a feature fusion step. The features are then pooled to a global distance $d(x, x_0)$, designated as d_0 and ranged in $[0, 1]$, which estimates the similarity level between x and x_0 (the value of 0 indicates that the patches are similar).

LPIPS operates on image patches rather than full images. Thus, the inputs to the neural network (x and x_0) are 64×64 sized patches. As mentioned in [4], there are three reasons for this: (1) the space of full images is extremely large, which makes it much harder to cover a reasonable portion of the domain with judgments (even 64×64 color patches represent a 12k-dimensional space), (2) by choosing a smaller patch size, the focus is on lower-level aspects of similarity, which mitigate the effect of the *different sense of similarity* that can be influenced by the high-level semantics [168], and (3) modern methods for image synthesis train deep networks with patch-based losses [169].

The proposed network of LPIPS is illustrated in Figure 7.1 and detailed in the following.

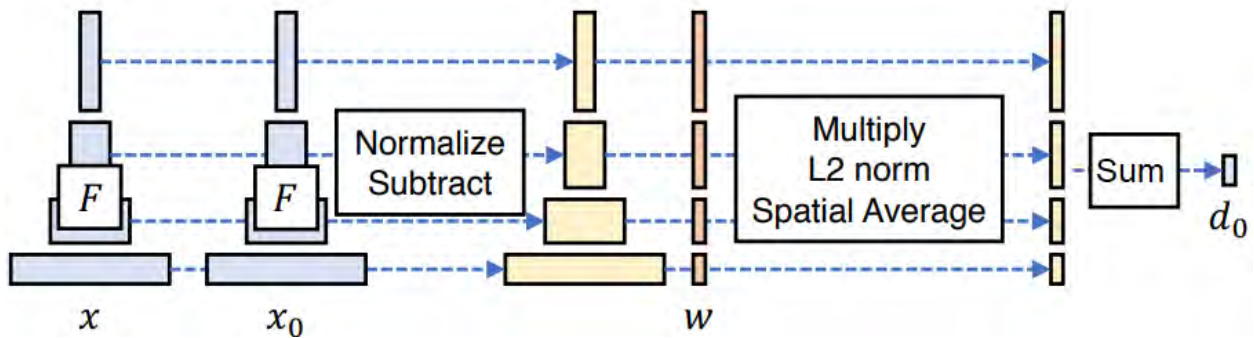


Figure 7.1: To compute a distance d_0 between two patches x and x_0 , given a network F : we first compute deep embeddings, normalize the activations in the channel dimension, scale each channel by vector ω , and take the ℓ_2 distance. We then average across spatial dimension and sum across all layers. This figure is reprinted from [4].

The authors [4] proposed the possibility of using either the SqueezeNet [170] (known to be extremely lightweight; 2.8MB), AlexNet [171] (which closely match the architecture of the human visual cortex [172]) or VGG [173] architecture (known for its successful application for various computer vision tasks [174]) as basis F for the proposed networks. In fact, they compared the three architectures and found that the 3 provide nearly the same performance and work well for perceptual similarity tasks. Five convolutional layers L were used from the VGG (58.9 MB) and the AlexNet (9.1 MB) networks. For the

SqueezeNet network (2.8MB), the first convolutional layer and 6 subsequent fire modules were used.

In order to better match low-level human judgments, the authors calibrated the existing networks F by adding a learned linear layer/weight ω (which does a 1×1 convolution) on top of each channel: a total of 1152, 1472 and 2240 weights ω were added to AlexNet, VGG and SqueezeNet, respectively (since the convolutional layers contain (64, 192, 384, 256, 256 and 256) channels for AlexNet, (64, 128, 256, 512 and 512) channels for VGG and (64, 128, 256, 384, 384, 512 and 512) channels for SqueezeNet).

Eq. 7.1 and Figure 7.1 show how to compute the distance between the reference x and distorted x_0 patches. Feature stacks are extracted from L layers and normalized in the channel dimension. We refer to them as $\hat{y}^l, \hat{y}_0^l \in \mathfrak{R}^{H^l \times W^l \times C^l}$, where C^l , $H^l \times W^l$ denote respectively the number of channel and the Height \times Width of the plane for layer l .

\hat{y}_0 and \hat{y} are scaled channel-wise by the weight vector $\omega^l \in \mathfrak{R}^{C^l}$. Then, the ℓ_2 distance is computed (over the channels). Finally, we average across spatial dimension ($H^l \times W^l$) and sum across all layers (L).

$$d(x, x_0) = \sum_l^L \frac{1}{H^l W^l} \sum_{h,w}^{H^l, W^l} \left\| \omega^l \odot \left(\hat{y}_{hw}^l - \hat{y}_0^l \right) \right\|_2^2 \quad (7.1)$$

The metric was trained and tested using a large-scale perceptual similarity dataset (BAPPS) containing over 180k distorted patches of natural images obtained from parameterized distortions (e.g., random noise, blurring, spatial shifts) and real algorithms (e.g., superresolution, colorization, frame interpolation). Over 484k subjective judgments were collected. The dataset contains two types of perceptual judgments: Two Alternative Forced Choice (2AFC; participants were asked which of two distorted patches (x_0, x_1) is more similar to the reference x) and Just Noticeable Differences (JND; participants were asked whether two patches -one reference x and one distorted x_0 - are the same or different).

The metric was trained on 80% of the 2AFC data. Given 2 distances $d(x, x_0)$ and $d(x, x_1)$ (designated as d_0 and d_1), and the subjective score $h \in [0, 1]$ (0: all participants preferred x_0 , 1: all participants preferred x_1), a small network G on top was trained to map (d_0, d_1) to h . The architecture of G uses two 32-channel FC-ReLu layers, followed by a 1-channel FC layer and a sigmoid. The network G and the loss function are illustrated in Figure 7.2. The training consists of 5 epochs at initial learning rate 10^{-4} , 5 epochs with linear decay, and a batch size of 50.

The authors considered several configurations for training: (1) keeping pre-trained network weights F fixed, and learning the linear weights w (1152, 1472 and 2240 parameters are learned for AlexNet, VGG and SqueezeNet respectively); (2) initializing F from a pre-trained classification model, and allowing all its weights to be fine-tuned; (3) initializing F from scratch using random Gaussian weights and train it entirely on the 2AFC judgments. The first configuration provided the best results as it represents a *perceptual calibration* of some parameters in the feature space. Learning linear weights on top of the AlexNet model achieves state-of-the-art results on the real algorithms test set.

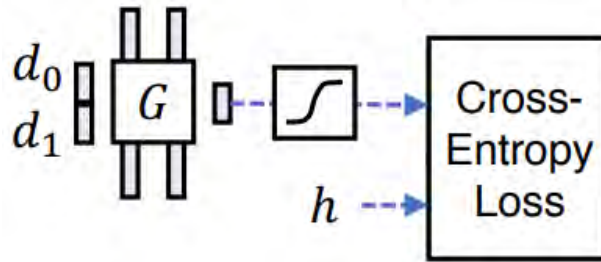


Figure 7.2: Training on 2AFC: a small network G is trained to predict perceptual judgment h from distance pair (d_0, d_1) . This figure is reprinted from [4].

7.2 Toward a CNN-based quality metric for 3D graphics

We aim to devise a new data-driven metric for evaluating the perceptual quality of textured 3D meshes using our large-scale textured mesh dataset described in Chapter 5. This dataset contains 3000 stimuli generated from 55 textured source models corrupted by various distortions applied to the geometry and texture. The stimuli were animated with full rotation around their vertical axis and evaluated in a subjective experiment based on the DSIS protocol. Thus, each stimulus is associated with a Mean Opinion Score (MOS) reflecting its overall perceived quality (see Chapter 5, section 5.3.2).

7.2.1 Performance of LPIPS on our dataset

We evaluated the performance of LPIPS on our dataset using snapshots of the stimuli rendered in their main viewpoint (shown in Figure 5.1), to which we assigned the MOS values. In fact, we assumed that this viewpoint has the most perceptual impact as it was perceptually chosen to cover the most geometric, color and semantic information (see Chapter 5, section 5.1.2). We employed the AlexNet LPIPS model associated with the configuration 1 described previously in section 7.1, as it has the best performance [4]. The Spearman Rank Order Correlation Coefficient obtained was 0.69. A recent study [76] showed that LPIPS is very good for patch-based similarity but has room for improvement in terms of overall correlation with MOS, especially on novel distortions that are not represented in the original training set of LPIPS. Indeed, LPIPS is a perceptual similarity metric, trained for 2AFC and JND tasks (not quality assessment tasks) and thus it may not capture how differently humans perceive the transitions in distortion levels. Furthermore, LPIPS was trained on natural images and not on 3D graphics images.

7.2.2 Our approach

The metric we propose is an extension of LPIPS adapted for 3D graphics and quality assessment tasks based on DSIS. Indeed, LPIPS computes distances per patches, while the MOS represents the overall quality of the stimulus. Therefore, we (1) adapted the network G to suit the MOS scores and (2) modified the global loss function so the optimization (the loss computation) is done per image (instead of patch-based losses). In other words,

we first combine distances computed for patches of the same image, then make the mapping to the MOS score.

In order to adapt LPIPS for our quality assessment task, we modified the network G (Figure 7.2) to better match subjective quality judgments (MOS). Similarly to in [70,73], we opted for the “pooling by simple averaging” approach, shown in Eq. 7.2 and Figure 7.3: Given a quality annotated distorted image I (having a quality score MOS_I) and a reference image I' , we randomly sampled patches of I . LPIPS estimates the quality locally (the distance $d(x_i^r, x_i)$, designated as d_i , is computed per patch x_i). We assume identical perceptual importance of every image region (i.e. image patch). Thus, the estimated overall image quality \hat{Q} of I is derived by simply averaging local patch qualities (distances).

$$\hat{Q}_I = \frac{1}{N_p} \sum_i^{N_p} d(x_i^r, x_i) \quad (7.2)$$

where N_p denotes the number of patches sampled from I . x_i refer to a patch from the distorted image I , while x_i^r is its corresponding on the reference image I' .

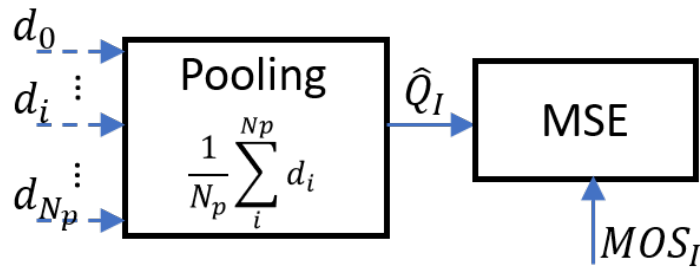


Figure 7.3: Training on quality assessment task: the overall quality \hat{Q} of an image is estimated by averaging the qualities d_i of its local patches.

The Mean Square Error (MSE) is used as the minimization criterion. The loss function is then:

$$E_I = (\hat{Q}_I - MOS_I)^2 \quad (7.3)$$

where E_I is the loss over an image.

7.2.3 Training on the textured 3D mesh dataset

As mentioned earlier, our model employs the pre-trained AlexNet network associated with the configuration 1 of LPIPS which consists of keeping the AlexNet network weights fixed and learning/optimizing only the linear weights on top ω .

To train our model, we considered for each stimulus of our dataset a snapshot taken from its main viewpoint. The snapshot size, 650×550 , is the video resolution of the stimuli seen by the participants in the subjective experiment. As discussed earlier in this section, we assigned to this snapshot the MOS obtained for the stimulus. Thus, we have 3000 annotated images representing our 3000 degraded stimuli. We divided (patchified) these images into small overlapping patches of size 64×64 . We removed patches containing

less than 65% stimulus information (i.e., the percentage of background pixels in the patch is greater than 35%). We got an average of 60 patches per stimulus.

We trained the network on randomly sampled patches of the stimulus images. As the distances computed for patches of the same image are combined for the calculation/optimization of the weights ω of the linear layers of the network (Eq. 7.2 and Eq. 7.3), we can not treat each patch as a separate sample (in other words, the patches of the same image can not be distributed over different batches). Thus, each batch was made to contain N_I images, each represented by N_p randomly sampled patches, resulting in a batch size of $N_I \times N_p$ patches. The backpropagated error is the average loss (E_I of Eq. 7.3) over the images in a batch.

During training, patches are randomly sampled every epoch to ensure that as many different image patches as possible are used in training. For validation, the N_p random patches for each validation image are only sampled once at the beginning of training in order to avoid noise in the validation loss.

N_I , N_p , along with the learning rate and the number of epochs are hyperparameters determined by tuning the metric.

80 % of the stimuli in the dataset (about 2400) are used for training and 20 % for testing. The dataset is randomly split by source model. This guarantees that no 3D model is used for both training and testing (no distorted or undistorted versions of a model used in testing or validation has been seen by the network during training). 44 source models out of 55 were included in the training while the rest were used for testing.

We refer to the version of LPIPS adapted for 3D Graphics as *Graphic-LPIPS*.

7.3 Results and Evaluation

We tuned the proposed network to optimize its performance on the test set by varying the following hyperparameters: the batch size (N_I , N_p), the learning rate, the number of epochs at the initial learning rate and the number of epochs at the decreasing learning rate. The final model we retained, and used in the subsequent performance evaluation, was trained for 10 epochs (5 epochs at initial learning rate 10^{-4} and 5 epochs with linear decay). Each batch contained $N_I = 4$ images (stimuli), each represented by $N_p = 150$ randomly sampled patches.

7.3.1 Performance evaluation on the test set of textured meshes

Table 7.1 summarizes the performance of our metric in comparison to other state-of the art Image Quality Metrics (IQMs) on the test set selected from our textured mesh dataset (around 600 stimuli obtained from 11 models), in terms of Pearson Linear Correlation (PLCC) and the Spearman Rank Order Correlation (SROCC). As for our metric, the IQMs were also computed on the snapshots taken from the main viewpoint of the stimuli, which allows for a fair comparison. Note that, PLCC is computed after a logistic regression which provides a non-linear mapping between the objective and subjective scores.

Table 7.1: Performance comparison of different metrics on the test set of our textured 3D mesh dataset, described in Chapter 5.

	Graphic-LPIPS	LPIPS	SSIM	HDR-VDP2	iCID
PLCC	0.85	0.7	0.64	0.68	0.61
SROCC	0.86	0.69	0.63	0.68	0.6

The proposed approach shows a much better correlation with MOSs than that of IQMs. We believe that the poor performance of the IQMs is due to the fact that our stimuli are derived from 11 different source models and are corrupted by a combination of 5 different types of distortions applied to both the geometry and the texture of the mesh. Indeed, as observed in section 6.2.5 of Chapter 6 and as reported in [69], IQMs are less accurate in differentiating and ranking the visibility of distortion artifacts between different 3D models.

Furthermore, we noticed that IQMs exhibit poorer performance on this dataset than the previous dataset of meshes with vertex colors presented in Chapter 3. The IQMs results on the latter dataset are reported in Table 6.6. This shows that our dataset of textured meshes is therefore globally more challenging for the quality metrics. We believe that this is related to (1) the process of selecting the 3000 stimuli, which samples a lot of stimuli for which two quality metrics did not agree (see section 5.2.1 of Chapter 5) and (2) the large variability of distortion combinations (mixed distortions) present in this dataset (whereas in the dataset of meshes with vertex colors, the distortions were not mixed/combined).

7.3.2 Performance evaluation on a dataset of colored 3D meshes

We evaluated the performance of our metric on the dataset of meshes with vertex colors, described in Chapter 3 (section 3.1), to assess its robustness. This dataset is composed of 480 dynamic stimuli, generated from 5 source models subjected to geometry and color distortions. Each stimulus was displayed in 3 viewpoints and animated with 2 short movements. The dataset was obtained through a subjective study in VR based on the DSIS method. Thus, each stimulus is assigned a MOS value.

We tested the performance of our metric according to 2 scenarios:

1. Viewpoints not taken into account: for a given stimulus, we averaged its MOSs over the different viewpoints and animations. Thus, the dataset used in this scenario is of 80 stimuli. Graphic-LPIPS was computed on the *viewpoint 1* of the stimuli (shown in Figure 3.1) which represents the main viewpoint as indicated in section 3.1.
2. Viewpoints taken into account: as each stimulus in this dataset was rated in 3 different viewpoints (illustrated in Figure 3.1), we computed Graphic-LPIPS on snapshots taken from each viewpoint displayed to the observer (i.e. total of 240 stimuli/snapshots). We considered for a given stimulus its 3 viewpoints and averaged the MOSs of the 2 animations.

Table 7.2 shows the results of scenario (1), while Table 7.3 reports those of scenario (2).

We included the results obtained in Chapter 6 for the IQMs and CMDM metric computed on this dataset. The computation of these metrics in each of these scenarios is detailed in sections 6.2.5 and 6.4 (scenario 2) respectively.

Table 7.2: Performance comparison of different metrics on the dataset of meshes with vertex colors, described in Chapter 3, for scenario (1). For metrics marked with a *, the values are reprinted from Table 6.6 in Chapter 6.

	Graphic-LPIPS	CMDM*	SSIM*	HDR-VDP2*	iCID*
PLCC	0.88	0.91	0.8	0.85	0.82
SROCC	0.88	0.9	0.8	0.84	0.83

Table 7.3: Performance comparison of different metrics on the dataset of meshes with vertex colors described in Chapter 3, for different viewpoints (scenario 2). For metrics marked with a *, the values are reprinted from Table 6.10b in Chapter 6.

	Graphic-LPIPS	CMDM*	SSIM*	HDR-VDP2*	iCID*
PLCC	0.89	0.89	0.79	0.81	0.86
SROCC	0.88	0.87	0.8	0.83	0.87

Although the proposed metric was trained on a different dataset with different models and different distortions and even different color representation (textures and not vertex colors), its performance in both scenarios is comparable to that of CMDM which was learned on this dataset. This shows the good robustness of our metric and validates its ability to differentiate and rank stimuli from different source models and different distortions. Moreover, Table 7.3 shows that our metric can be computed on different viewpoints of the 3D object (even if it is not necessarily the main viewpoint) and still provide good results in correlation with MOSs.

7.4 View-independent approach

To avoid manual selection of a relevant/main viewpoint for each 3D model (limitation of view-dependent approaches), we considered another training scenario for our metric, using a set of snapshots of the model taken from different viewpoints. This seems relevant to us, especially since all the stimuli in the database of textured meshes were animated with a full rotation (360 degrees) during the subjective test. Thus, we generated for each stimulus 4 snapshots taken from 4 camera positions regularly sampled on its bounding box. The snapshots were patchified (divided into patches) and fed to the network. We did not modify the network architecture proposed in the previous section nor the training and test sets. We trained the network on N_p randomly sampled patches of N_I stimulus images/snapshots as described in subsection 7.2.3. The results on the test set of textured meshes (presented in subsection 7.3.1) are reported in Table 7.4. For the IQMs, the global

quality score of a stimulus is the average of the IQM values computed on its 4 snapshots.

Table 7.4: Performance comparison of different metrics on the test set of our textured 3D mesh dataset, when several viewpoints are considered per stimulus.

	Graphic-LPIPS	SSIM	HDR-VDP2	iCID
PLCC	0.83	0.67	0.69	0.67
SROCC	0.85	0.69	0.69	0.68

Comparing Tables 7.1 and 7.4, we observe that the performance of our metric decreases slightly when considering a view-independent approach. This indicates that our manual choice/selection of the main viewpoint of the 3D models is indeed relevant and helps the network. It also indicates that the perceptual pooling is not uniform: some parts/viewpoints of the objects have a stronger influence on the overall quality perceived by the observer. This effect depends on the metrics, as the performance of SSIM and iCID improves when considering multiple viewpoints per stimulus, while that of HDR-VDP remains stable. All these results show that the perceptual pooling mechanism is complex for 3D objects and therefore requires the integration of a visual attention model into the network in order to learn it.

7.5 Conclusion

In this chapter, we proposed an image-based perceptual quality metric for 3D graphics based on CNN. The metric can be seen as an extension of LPIPS. It is computed on rendered snapshots of the 3D models. It employs a Siamese network fed with reference patches and distorted patches. We employed the AlexNet architecture with learning linear weights on top. The overall quality of the model is derived by averaging local patch qualities. The metric outperformed other image-based quality metrics (IMQs) in terms of correlations with subjective scores on our textured mesh dataset. It also demonstrates state-of-the-art results on our dataset of meshes with vertex colors.

We believe that our metric has still room for improvement, especially regarding the pooling of local patch qualities. Indeed, the current version of the metric assumes that all patches of the image have the same perceptual impact, which is not completely consistent with our Human Visual System (HVS). Neither the local quality nor the relative importance of local qualities are uniformly distributed over an image. Thus, it could be interesting to use a patch-wise weighting approach to account for the influence of the patch on the global quality estimate. A learned visual attention model could be used to estimate the impact of each patch on the global perceived quality.

We would also like to test our metric on new datasets, such as the LIRIS Textured Mesh dataset [3] and the Volumetric Video Quality datasets (for both meshes and point clouds) [1], to assess its generalization ability across different datasets with different types of distortion and even different 3D data representations.

Conclusion

This thesis addressed the challenges of evaluating the visual quality of rendered 3D graphics. We were specifically interested in 3D meshes with color attributes, either in the form of texture maps or vertex colors. To this end, subjective and objective quality assessment methods were proposed and underlying influencing factors were investigated and discussed. Our main contributions are summarized below along with limitations, concluding remarks and perspectives. They are grouped into two broad themes following the parts of the manuscript.

- **Subjective quality assessment**

We conducted an extensive study comparing three of the most prominent subjective methodologies in image processing, involving hidden references (ACR-HR method) and explicit references (DSIS and SAMVIQ methods). We evaluated their performance on a dataset of 80 colored 3D models, impaired with various distortions in a Virtual Reality (VR) context. We found that the ACR-HR method is less discriminating than methods with explicit references. Unlike the quality assessment of natural images and videos, we believe that the presence of an explicit reference is necessary for the quality assessment of 3D graphics, as people have less prior knowledge about the quality of these data than about the quality of natural images. DSIS and SAMVIQ exhibited almost the same performance in terms of accuracy and user agreement. However, DSIS showed a great advantage in term of time-effort. Based on these results, we advocate that DSIS is the most suitable method for evaluating the quality of 3D graphics. Finally, we recommended using groups of at least 24 observers for DSIS tests.

This study makes the first step toward standardizing a methodology for assessing the quality of 3D graphics in virtual reality. The only data representation used was 3D meshes, however, we believe that our results remain valid for other 3D representations, such as point clouds. Further work is still needed to confirm this presumption. It could also be interesting to evaluate the impact of the display devices (2D screen, VR/MR headset) on the perceived visual quality of 3D graphics.

Based on the results of the above study, we extended the previous dataset by associating 3 viewpoints and 2 animations to each stimulus, and conducted a larger subjective experiment in VR. More than 11k quality judgments were collected. The generated dataset of 480 animated meshes along with the subjective scores are publicly available¹. It is the first dataset based on vertex color representation and also the first public dataset produced in VR for 3D content with color attributes.

The resulting dataset allowed us to evaluate the factors that most influence the perceived

¹<https://yananehme.github.io/datasets/>

quality of 3D content in VR. Here are 3 of our main findings: (1) complex masking effects occur when considering the interaction between viewpoint and distortion; (2) the viewpoint also affects the Confidence Intervals (CIs) of the ratings: the most informative viewpoint tends to produce the largest CIs, especially when combined with a zoom animation; and (3) the animation, by itself, has a moderate impact on ratings and CIs, however the more visible the content information (as in zoom animation), the higher its content ambiguity.

These findings helped us design our objective quality metric for meshes with vertex colors (*CMDM*) described a little while later. Further studies are needed to generalize these results to a non-VR scenario. This brings us to our third subjective study, this time conducted on a 2D screen and in crowdsourcing.

The previous experiments were conducted in the lab in a controlled environment and with high-end VR headsets. Because of the COVID-19 pandemic, lab experiments were no longer possible, so we opted for CrowdSourcing (CS). Before conducting a large subjective quality assessment experiment in CS, we investigated whether a CS test can achieve the accuracy of a lab test for 3D graphics. For this purpose, we designed a CS experiment that replicates as much as possible the above mentioned lab experiments using the same dataset of 3D models and the same experimental methodology.

Results showed that under controlled conditions and with a proper participant screening approach, a CS experiment based on the DSIS method can be as accurate as a lab experiment. It is worth mentioning that CS is quite faster to evaluate large dataset, yet the most time intensive task is building and designing the CS experimental tool (user-friendly tool, control viewer environment, add screening test, etc.).

It would be interesting to repeat the lab experiment, this time using a desktop setup, to clearly isolate the effects of VR.

We produced the largest (to date) textured meshes quality assessment dataset, which includes 55 source models and more than 343k distorted meshes generated from combinations of 5 types of distortions (related to compression and simplification) applied on the geometry and texture of the meshes. The geometric, color, and semantic complexity of the source models was quantitatively characterized using 3 new measures based on spatial information and visual attention complexity. A carefully selected subset of 3000 stimuli were annotated in a large-scale CS quality assessment experiment, wherein more than 148k quality judgments were collected. The quality scores of the remaining stimuli were predicted using a quality metric based on deep learning.

This dataset allowed us to draw interesting conclusions regarding the impact of each distortion and their combinations on the perceived quality of textured meshes. We found a strong perceptual interaction between the geometry quantization of the mesh and its level of details. The geometry quantization affects the quality scores more than the texture coordinates (UV map) quantization. Regarding the texture distortions, decreasing the texture size reduces the perception of artifacts caused by strong UV quantization (the tiling effect). We also showed that the quality level of the JPEG compression algorithm, applied to the texture, can be reduced to very low values (50) while maintaining the quality of the final rendered image. Furthermore, we evaluated the influence of the complexity of the geometry, color and texture seams on the perception of distortions. We observed that both color and geometry can mask the geometric degradations of a quantized 3D model.

Models with monochromatic textures are less sensitive to UV map quantization; however, the impact of the UV quantization on the visual quality depends also on the amount of texture seams: quantization artifacts are clearly more visible on models exhibiting a large number of texture seams. The dataset will be made publicly available online.

We showed an application for rate-distortion control and optimization, as a real-world use case for this dataset. We aim to devise, in a near future, an analytical perceptual rate-distortion model capable of maximizing the visual quality of the reconstructed textured meshes subjected to a target bitrate.

- **Objective quality assessment**

Leveraging our 2 established datasets (resp. composed of meshes with vertex colors and textured meshes), we proposed 2 data-driven metrics for quality assessment of 3D graphics with color attributes.

The first metric, called *CMDM*, is a full reference metric for quality assessment of 3D meshes with vertex colors that operates entirely on the mesh domain. It incorporates perceptually-relevant geometry-based and color-based features. The optimal set of features was selected through logistic regression and cross-validation tests. We used our dataset of meshes with vertex colors to evaluate the performance of *CMDM* and benchmark it with state-of-the-art Image Quality Metrics. Moreover, to assess the robustness of *CMDM*, we tested it on an existing dataset of textured meshes corrupted with compound distortions that differ considerably from those used to train it. Our metric provides state-of-the-art results in terms of correlations and classification abilities for the two datasets. It also demonstrates a better stability than IQMs. Indeed, *CMDM* is able to differentiate and rank stimuli from different sources and different distortions, unlike IQMs which perform very well when assessing the quality of different versions of a single source, but are less accurate when ranking distortions applied on different sources. A version of *CMDM* adapted for colored point clouds was also developed. It is the first Point Cloud Quality Metric (PCQM) that takes into account both geometry and color. We publicly released the source codes of *CMDM*² and *PCQM*³.

We extended *CMDM* by combining its geometry and color features with the Visual Attention Complexity (VAC) measure, based on the visual saliency dispersion. We showed that incorporating a VAC measure into our perceptual quality metric improves the prediction of visual quality.

We also studied how the knowledge of the visible parts (the viewpoint) of the 3D model can improve the results of quality metrics. Further studies are still needed to effectively incorporate visibility information into quality metrics.

The second metric we introduced, called *Graphic-LPIPS*, is an image-based full-reference quality metric based on Convolutional Neural Networks (CNNs). This metric is an extension of *LPIPS*, adapted to 3D graphics and quality assessment tasks. It is computed on rendered snapshots of the 3D models and employs the AlexNet architecture with learning linear weights on top. The overall quality of the model is derived by averaging local patch qualities. *Graphic-LPIPS* outperformed other IMQs in terms of correlations with subjective

²<https://github.com/MEPP-team/MEPP2>

³<https://github.com/MEPP-team/PCQM>

scores on our textured mesh dataset, which proved to be more challenging for quality metrics than other existing datasets. *Graphic-LPIPS* also provides as good results as *CMDM* on our dataset of meshes with vertex colors. The source code will be available online. Since some parts of the 3D objects have a stronger impact on the overall perceived quality than others, we believe that an important improvement to the metric would be to use a patch-wise weighting approach using a learned visual attention model that estimates the impact of each patch on the overall perceived quality.

• Perspectives

The quality assessment of 3D content with appearance attributes can still be considered in its early stages. Despite our work, as well as recent progress, there are still several limitations to overcome. We have indicated above, at the end of the paragraph of each contribution, how each of our research can be extended in the (near) future. Below is a brief summary of what we perceive as ultimate goals of our work.

1. Take into account, in a quality metric, the animation and rendering parameters such as lighting, materials, appearance attributes, including not only color but also metalness, roughness and normals information, etc.
2. Develop a version of the metric that operates in real-time and thus able to evaluate and drive the choices of the levels of details in the viewport during the interaction in VR.
3. Associate a visual attention model with the metric capable of predicting where the observer is looking.
4. Integrate all these tools into an online system, taking into account the network capacity.

Thus, the final tool will allow interactive visualization of rich 3D data in immersive environments, with a high quality of user experience.

Publications

• International journals

Yana Nehmé, Florent Dupont, Jean-Philippe Farrugia, Patrick Le Callet, and Guillaume Lavoué. Visual Quality of 3D Meshes with Diffuse Colors in Virtual Reality: Subjective and Objective Evaluation, In *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, Vol. 27, No. 3, March 2021. doi: 10.1109/TVCG.2020.3036153 - **Replicability Stamp**

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)*, Vol. 18, No. 1, Jan. 2021. doi: 10.1145/3427931

• International conferences

Yana Nehmé, Patrick Le Callet, Florent Dupont, Jean-Philippe Farrugia, and Guillaume Lavoué. Exploring Crowdsourcing for Subjective Quality Assessment of 3D Graphics, In *IEEE International Workshop on Multimedia Signal Processing (MMSP)*, Oct. 2021.

Yana Nehmé, Mona Abid, Guillaume Lavoué, Matthieu Perreira Da Silva, and Patrick Le Callet. CMDM-VAC: Improving a Perceptual Quality Metric for 3D Graphics by Integrating a Visual Attention Complexity Measure. In *IEEE International Conference on Image Processing (ICIP)*, Sept. 2021. doi: 10.1109/ICIP42928.2021.9506662.

Gabriel Meynet, **Yana Nehmé**, Julie Digne, and Guillaume Lavoué. *PCQM: A Full-Reference Quality Metric for Colored 3D Point Clouds*. In *IEEE International Conference on Quality of Multimedia Experience (QoMEX)*, June. 2020. doi: 10.1109/QoMEX48832.2020.9123147. - **Best Student Paper Award**

Yana Nehmé, Jean-Philippe Farrugia, Florent Dupont, Patrick Le Callet, and Guillaume Lavoué. Comparison of Subjective Methods, With and Without Reference, for Quality Assessment of 3D Graphics, *ACM Symposium on Applied Perception (SAP)*, Sept. 2019. doi: 10.1145/3343036.3352493.

References

- [1] E. Zerman, C. Ozcinar, P. Gao, and A. Smolic, “Textured Mesh vs Coloured Point Cloud : A Subjective Study for Volumetric Video Compression,” *Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*, pp. 1–6, 2020.
- [2] Z. Li and C. Bampis, “Recover subjective quality scores from noisy measurements,” *Proceeding of the 2017 Data Compression Conference (DCC), IEEE*, pp. 52–61, 04 2017.
- [3] J. Guo, V. Vidal, I. Cheng, A. Basu, A. Baskurt, and G. Lavoué, “Subjective and Objective Visual Quality Assessment of Textured 3D Meshes,” *ACM Transactions on Applied Perception*, vol. 14, pp. 1–20, 10 2016.
- [4] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, “The unreasonable effectiveness of deep features as a perceptual metric,” *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 586–595, 2018.
- [5] M. Garland and P. S. Heckbert, “Surface simplification using quadric error metrics,” *ACM Siggraph*, pp. 209–216, 1997.
- [6] P. Alliez and C. Gotsman, *Recent Advances in Compression of 3D Meshes*, pp. 3–26. Springer, 01 2005.
- [7] K. Wang, G. Lavoué, F. Denis, and A. Baskurt, “A comprehensive survey on three-dimensional mesh watermarking,” *IEEE Transactions on Multimedia*, vol. 10, no. 8, pp. 1513–1527, 2008.
- [8] G. Lavoué and R. Mantiuk, “Quality assessment in computer graphics,” *Visual Signal Quality Assessment: Quality of Experience (QoE)*, pp. 243–286, 2015.
- [9] ITU-R BT.1788, “Methodology for the subjective assessment of video quality in multimedia applications,” *International Telecommunication Union*, 2007.
- [10] ITU-T P.910, “ Subjective video quality assessment methods for multimedia applications,” *International Telecommunication Union*, 2008.
- [11] ITU-R BT.500-13, “Methodology for the subjective assessment of the quality of television pictures BT Series Broadcasting service,” *International Telecommunication Union*, 2012.
- [12] R. K. Mantiuk, A. Tomaszewska, and R. Mantiuk, “Comparison of Four Subjective Methods for Image Quality Assessment,” *Computer Graphics Forum*, vol. 31, pp. 2478–2491, dec 2012.

-
- [13] VQEG, “Multimedia test plan 1.19,” 2007.
- [14] Q. Huynh-Thua and M. Heath, “Examination of the samviq methodology for the subjective assessment of multimedia quality,” *Third Inter. Workshop on Video Processing and Quality Metrics for Consumer Electronics*, 01 2007.
- [15] B. Watson, “Measuring and predicting visual fidelity,” *ACM Siggraph*, pp. 213–220, 2001.
- [16] B. E. Rogowitz and H. E. Rushmeier, “Are image quality metrics adequate to evaluate the quality of geometric objects?,” *Proc SPIE 4299, Human Vision and Electronic Imaging VI*, 06 2001.
- [17] Y. Pan, I. Cheng, and A. Basu, “Quality metric for approximating subjective evaluation of 3-D objects,” *IEEE Transactions on Multimedia*, vol. 7, pp. 269–279, apr 2005.
- [18] S. Silva, B. S. Santos, C. Ferreira, and J. Madeira, “A Perceptual Data Repository for Polygonal Meshes,” *2009 Second International Conference in Visualisation*, pp. 207–212, jul 2009.
- [19] G. Lavoué, E. Drelie Gelasca, F. Dupont, A. Baskurt, and T. Ebrahimi, “Perceptually driven 3D distance metrics with application to watermarking,” *Applications of Digital Image Processing XXIX*, vol. 6312, pp. 150 – 161, 2006.
- [20] G. Lavoué, “A local roughness measure for 3d meshes and its application to visual masking,” *ACM Trans. Appl. Percept.*, vol. 5, no. 4, 2009.
- [21] E. Drelie Gelasca, T. Ebrahimi, M. Corsini, and M. Barni, “Objective evaluation of the perceptual quality of 3d watermarking,” *IEEE International Conference on Image Processing (ICIP)*, 2005.
- [22] M. Corsini, E. D. Gelasca, T. Ebrahimi, and M. Barni, “Watermarked 3-D mesh quality assessment,” *IEEE Transactions on Multimedia*, vol. 9, pp. 247–256, 2007.
- [23] L. Váša and J. Rus, “Dihedral Angle Mesh Error: a fast perception correlated distortion measure for fixed connectivity triangle meshes,” *Computer Graphics Forum*, vol. 31, no. 5, 2012.
- [24] K. Christaki, E. Christakis, and P. Drakoulis, “Subjective Visual Quality Assessment of Immersive 3D Media Compressed by Open-Source Static 3D Mesh Codecs,” *25th International Conference on MultiMedia Modeling (MMM)*, pp. 1–12, 2018.
- [25] L. Váša and V. Skala, “A perception correlated comparison method for dynamic meshes,” *IEEE transactions on visualization and computer graphics*, vol. 17, pp. 220–30, 02 2011.
- [26] F. Torkhani, K. Wang, and J.-M. Chassery, “Perceptual quality assessment of 3D dynamic meshes: Subjective and objective studies,” *Signal Processing: Image Communication*, vol. 31, pp. 185–204, Feb. 2015.

-
- [27] J. Gutiérrez, T. Vigier, and P. Le Callet, “Quality evaluation of 3d objects in mixed reality for different lighting conditions,” *Electronic Imaging*, vol. 2020, 01 2020.
- [28] J. Zhang, W. Huang, X. Zhu, and J.-N. Hwang, “A subjective quality evaluation for 3d point cloud models,” *2014 International Conference on Audio, Language and Image Processing*, pp. 827–831, 2014.
- [29] E. Alexiou and T. Ebrahimi, “On subjective and objective quality evaluation of point cloud geometry,” *2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX)*, pp. 1–3, 2017.
- [30] E. Alexiou, E. Upenik, and T. Ebrahimi, “Towards subjective quality assessment of point cloud imaging in augmented reality,” *2017 IEEE 19th International Workshop on Multimedia Signal Processing (MMSP)*, pp. 1–6, 2017.
- [31] H. Su, Z. Duanmu, W. Liu, Q. Liu, and Z. Wang, “Perceptual Quality Assessment of point Clouds,” *IEEE International Conference on Image Processing*, pp. 3182–3186, 2019.
- [32] E. M. Torlig, E. Alexiou, T. A. Fonseca, R. L. de Queiroz, and T. Ebrahimi, “A novel methodology for quality assessment of voxelized point clouds,” *SPIE Optical Engineering + Applications*, p. 18, 2018.
- [33] E. Alexiou, I. Viola, T. M. Borges, T. A. Fonseca, R. L. de Queiroz, and T. Ebrahimi, “A comprehensive study of the rate-distortion performance in mpeg point cloud compression,” *APSIPA Transactions on Signal and Information Processing*, vol. 8, p. e27, 2019.
- [34] L. A. da Silva Cruz, E. Dumić, E. Alexiou, J. Prazeres, R. Duarte, M. Pereira, A. Pinheiro, and T. Ebrahimi, “Point cloud quality evaluation : Towards a definition for test conditions,” *International Conference on Quality of Multimedia Experience*, p. 6, 2019.
- [35] A. Javaheri, C. Brites, F. Pereira, and J. Ascenso, “Point Cloud Rendering after Coding : Impacts on Subjective and Objective Quality,” *arXiv:1912.09137*, pp. 1–13, 2019.
- [36] Y. Liu, Q. Yang, Y. Xu, and L. Yang, “Point cloud quality assessment: Large-scale dataset construction and learning-based no-reference approach,” *preprint arXiv:2012.11895*, 2020.
- [37] M. Čadík, R. Herzog, R. Mantiuk, K. Myszkowski, and H.-P. Seidel, “New measurements reveal weaknesses of image quality metrics in evaluating graphics artifacts,” *ACM Transactions on Graphics (Proc. of SIGGRAPH Asia)*, vol. 31, pp. 1–10, 2012.
- [38] R. Piórkowski, R. Mantiuk, and A. Siekawa, “Automatic detection of game engine artifacts using full reference image quality metrics,” *ACM Transactions on Applied Perception*, vol. 14, pp. 14:1–14:17, Mar. 2017.

- [39] K. Wolski, D. Giunchi, N. Ye, P. Didyk, K. Myszkowski, R. Mantiuk, H.-P. Seidel, A. Steed, and R. Mantiuk, "Dataset and metrics for predicting local visible differences," *ACM Transactions on Graphics*, vol. 37, pp. 1–14, 11 2018.
- [40] S. Subramanyam, J. Li, I. Viola, and P. Cesar, "Comparing the quality of highly realistic digital humans in 3dof and 6dof: A volumetric video case study," *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 127–136, 2020.
- [41] K. Vanhoey, B. Sauvage, P. Kraemer, and G. Lavoué, "Visual quality assessment of 3D models: On the influence of light-material interaction," *ACM Transactions on Applied Perception*, vol. 15, no. 1, 2017.
- [42] J. Filip, M. Chantler, P. Green, and M. Haindl, "A psychophysically validated metric for bidirectional texture data reduction," *ACM Transactions on Graphics*, vol. 27, p. 138, 12 2008.
- [43] E. Alexiou, N. Yang, and T. Ebrahimi, "Pointxr: A toolbox for visualization and subjective evaluation of point clouds in virtual reality," *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*, pp. 1–6, 2020.
- [44] E. Alexiou and T. Ebrahimi, "Impact of visualisation strategy for subjective quality assessment of point clouds," *2018 IEEE International Conference on Multimedia Expo Workshops (ICMEW)*, pp. 1–6, 2018.
- [45] E. Alexiou, T. Ebrahimi, M. V. Bernardo, M. Pereira, A. Pinheiro, L. A. Da Silva Cruz, C. Duarte, L. G. Dmitrovic, E. Dumic, D. Matkovic, and A. Skodras, "Point cloud subjective evaluation methodology based on 2d rendering," *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*, pp. 1–6, 2018.
- [46] E. Zerman, P. Gao, C. Ozcinar, and A. Smolic, "Subjective and Objective Quality Assessment for Volumetric Video Compression," *Electronic Imaging*, vol. 2019, no. 10, pp. 323–1–323–7, 2019.
- [47] Q. Yang, H. Chen, Z. Ma, Y. Xu, R. Tang, and J. Sun, "Predicting the perceptual quality of point cloud: A 3d-to-2d projection-based exploration," *IEEE Transactions on Multimedia*, 2020.
- [48] Stephane Péchard, Romuald Pepion, Patrick Le Callet, "Suitable methodology in subjective video quality assessment: a resolution dependent paradigm," *International Workshop on Image Media Quality and its Applications, IMQA2008*, 2008.
- [49] K. Brunnström and M. Barkowsky, "Statistical quality of experience analysis: On planning the sample size and statistical significance testing," *Journal of Electronic Imaging*, vol. 27, p. 1, 09 2018.
- [50] T. Tominaga, T. Hayashi, J. Okamoto, and A. Takahashi, "Performance comparisons of subjective quality assessment methods for mobile video," *2010 Second International Workshop on Quality of Multimedia Experience (QoMEX)*, pp. 82–87, 2010.

-
- [51] T. Kawano, K. Yamagishi, and T. Hayashi, "Performance comparison of subjective assessment methods for stereoscopic 3d video quality," *IEICE Transactions on Communications*, vol. E97.B, no. 4, pp. 738–745, 2014.
- [52] E. Alexiou and T. Ebrahimi, "On the performance of metrics to predict quality in point cloud representations," *Applications of Digital Image Processing XL*, vol. 10396, pp. 282 – 297, 2017.
- [53] A. Singla, W. Robitza, and A. Raake, "Comparison of Subjective Quality Evaluation Methods for Omnidirectional Videos with DSIS and Modified ACR," *Electronic Imaging*, vol. 2018, no. 14, 2018.
- [54] V. Kiran Adhikarla, M. Vinkler, D. Sumin, R. K. Mantiuk, K. Myszkowski, H.-P. Seidel, and P. Didyk, "Towards a quality metric for dense light fields," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.
- [55] J. Redi, E. Siahaan, P. Korshunov, J. Habigt, and T. Hossfeld, "When the crowd challenges the lab: Lessons learnt from subjective studies on image aesthetic appeal," *Proceedings of the Fourth International Workshop on Crowdsourcing for Multimedia*, p. 33–38, 2015.
- [56] R. Z. Jiménez, L. F. Gallardo, and S. Möller, "Influence of number of stimuli for subjective speech quality assessment in crowdsourcing," *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*, pp. 1–6, 2018.
- [57] M. Reimann, O. Wegen, S. Pasewaldt, A. Semmo, J. Döllner, and M. Trapp, "Teaching Data-driven Video Processing via Crowdsourced Data Collection," *Eurographics 2021 - Education Papers*, 2021.
- [58] T. Hoßfeld, C. Keimel, M. Hirth, B. Gardlo, J. Habigt, K. Diepold, and P. Tran-Gia, "Best practices for qoe crowdtesting: Qoe assessment with crowdsourcing," *IEEE Transactions on Multimedia*, vol. 16, no. 2, pp. 541–558, 2014.
- [59] D. Ghadiyaram and A. C. Bovik, "Massive online crowdsourced study of subjective and objective picture quality," *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 372–387, 2016.
- [60] J. Vuurens, A. P. de Vries, and C. Eickhoff, "How much spam can you take? an analysis of crowdsourcing results to increase accuracy," *Proc. ACM SIGIR Workshop on Crowdsourcing for Information Retrieval (CIR'11)*, pp. 21–26, 2011.
- [61] J. Li, S. Ling, J. Wang, and P. Le Callet, "A probabilistic graphical model for analyzing the subjective visual quality assessment data from crowdsourcing," *Proceedings of the 28th ACM International Conference on Multimedia*, p. 3339–3347, 2020.
- [62] K.-T. Chen, C.-C. Wu, Y.-C. Chang, and C.-L. Lei, "A crowdsourcable qoe evaluation framework for multimedia content," *Proceedings of the 17th ACM International Conference on Multimedia*, p. 491–500, 2009.

- [63] F. Ribeiro, D. Florencio, and V. Nascimento, “Crowdsourcing subjective image quality evaluation,” *2011 18th IEEE International Conference on Image Processing*, pp. 3097–3100, 2011.
- [64] T.-K. Huang, C.-J. Lin, and R. C. Weng, “Ranking individuals by group comparisons,” *Proceedings of the 23rd International Conference on Machine Learning*, p. 425–432, 2006.
- [65] J. SØgaard, M. Shahid, J. Pokhrel, and K. Brunnström, “On subjective quality assessment of adaptive video streaming via crowdsourcing and laboratory based experiments,” *Multimedia Tools Appl.*, vol. 76, p. 16727–16748, Aug. 2017.
- [66] T. Hoßfeld, M. Hirth, P. Korshunov, P. Hanhart, B. Gardlo, C. Keimel, and C. Timmerer, “Survey of web-based crowdsourcing frameworks for subjective quality assessment,” *2014 IEEE 16th International Workshop on Multimedia Signal Processing (MMSP)*, pp. 1–6, 2014.
- [67] N. Aspert, D. Santa-Cruz, and T. Ebrahimi, “Mesh: measuring errors between surfaces using the hausdorff distance,” *Proceedings. IEEE International Conference on Multimedia and Expo*, vol. 1, pp. 705–708 vol.1, 2002.
- [68] P. Cignoni, C. Rocchini, and R. Scopigno, “Metro: Measuring error on simplified surfaces,” *Computer Graphics Forum*, vol. 17, 1998.
- [69] G. Lavoué, M. C. Larabi, and L. Vasa, “On the Efficiency of Image Metrics for Evaluating the Visual Quality of 3D Models,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 8, pp. 1987–1999, 2016.
- [70] L. Kang, P. Ye, Y. Li, and D. Doermann, “Convolutional neural networks for no-reference image quality assessment,” *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1733–1740, 2014.
- [71] S. A. , M. Pedersen, and S. Yu, “Image quality assessment by comparing cnn features between images,” *Journal of Imaging Science and Technology*, vol. 60, pp. 604101–6041010, 11 2016.
- [72] F. Gao, Y. Wang, P. Li, M. Tan, J. Yu, and Y. Zhu, “Deepsim: Deep similarity for image quality assessment,” *Neurocomputing*, vol. 257, pp. 104–114, 2017. Machine Learning and Signal Processing for Big Multimedia Analysis.
- [73] S. Bosse, D. Maniry, K.-R. Müller, T. Wiegand, and W. Samek, “Deep neural networks for no-reference and full-reference image quality assessment,” *IEEE Transactions on Image Processing*, vol. 27, no. 1, pp. 206–219, 2018.
- [74] H. Talebi and P. Milanfar, “Nima: Neural image assessment,” *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 3998–4011, 2018.
- [75] A. Mustafa, A. Mikhailiuk, D. A. Iliescu, V. Babbar, and R. K. Mantiuk, “Training a better loss function for image restoration,” *preprint arXiv:2103.14616*, 2021.

- [76] T. Tariq, O. T. Tursun, M. Kim, and P. Didyk, “Why are deep representations good perceptual quality features?,” *Computer Vision – ECCV 2020*, pp. 445–461, 2020.
- [77] D. Tian and G. AlRegib, “Batex3: Bit allocation for progressive transmission of textured 3-d models,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 1, pp. 23–35, 2008.
- [78] Z. Karni and C. Gotsman, “Spectral compression of mesh geometry,” *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques*, p. 279–286, 2000.
- [79] Z. Bian, S.-M. Hu, and R. Martin, “Evaluation for small visual difference between conforming meshes on strain field,” *Journal of Computer Science and Technology*, vol. 24, 01 2009.
- [80] G. Lavoué, “A Multiscale Metric for 3D Mesh Visual Quality Assessment,” *Computer Graphics Forum*, vol. 30, no. 5, pp. 1427–1437, 2011.
- [81] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, “Image quality assessment: From error visibility to structural similarity,” *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.
- [82] F. Torkhani, K. Wang, and J.-M. Chassery, “A Curvature Tensor Distance for Mesh Visual Quality Assessment,” *ICCVG 2012 - International Conference on Computer Vision and Graphics*, vol. Vol. 7594, pp. 253–263, 11 pages, Sept. 2012.
- [83] K. Wang, F. Torkhani, and A. Montanvert, “A Fast Roughness-Based Approach to the Assessment of 3D Mesh Visual Quality,” *Computers & Graphics*, 2012.
- [84] M. Corsini, M. C. Larabi, G. Lavoué, O. Petrik, L. Váša, and K. Wang, “Perceptual Metrics for Static and Dynamic Triangle Meshes,” *Computer Graphics Forum*, vol. 32, pp. 101–125, feb 2013.
- [85] G. Nader, K. Wang, F. Hétyroy-Wheeler, and F. Dupont, “Just Noticeable Distortion Profile for Flat-Shaded 3D Mesh Surfaces,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, pp. 2423–2436, Nov. 2016.
- [86] J. Guo, V. Vidal, A. Baskurt, and G. Lavou, “Evaluating the local visibility of geometric artifacts,” *ACM Symposium in Applied Perception*, 2015.
- [87] G. Lavoué, I. Cheng, and A. Basu, “Perceptual quality metrics for 3d meshes: Towards an optimal multi-attribute computational model,” *2013 IEEE International Conference on Systems, Man, and Cybernetics*, pp. 3271–3276, 2013.
- [88] A. Chetouani, “Three-dimensional mesh quality metric with reference based on a support vector regression model,” *Journal of Electronic Imaging*, vol. 27, no. 4, pp. 1 – 9, 2018.
- [89] Z. C. Yildiz, A. C. Oztireli, and T. Capin, “A machine learning framework for full-reference 3D shape quality assessment,” *Visual Computer*, vol. 36, no. 1, pp. 127–139, 2020.

-
- [90] Z. C. Yildiz and T. Capin, “A perceptual quality metric for dynamic triangle meshes,” *EURASIP Journal on Image and Video Processing*, vol. 2017, no. 12, 2017.
- [91] I. Abouelaziz, M. El Hassouni, and H. Cherifi, “A curvature based method for blind mesh visual quality assessment using a general regression neural network,” *2016 12th International Conference on Signal-Image Technology Internet-Based Systems (SITIS)*, pp. 793–797, 2016.
- [92] A. Nouri, C. Charrier, and O. Lézoray, “3D Blind Mesh Quality Assessment Index,” *IS&T International Symposium on Electronic Imaging*, Jan. 2017.
- [93] I. Abouelaziz, M. El Hassouni, and H. Cherifi, “No-reference 3d mesh quality assessment based on dihedral angles model and support vector regression,” *Image and Signal Processing*, pp. 369–377, 2016.
- [94] I. Abouelaziz, M. E. Hassouni, and H. Cherifi, “A convolutional neural network framework for blind mesh visual quality assessment,” *2017 IEEE International Conference on Image Processing (ICIP)*, pp. 755–759, 2017.
- [95] D. Tian, H. Ochimizu, C. Feng, R. Cohen, and A. Vetro, “Geometric distortion metrics for point cloud compression,” *2017 IEEE International Conference on Image Processing (ICIP)*, pp. 3460–3464, 2017.
- [96] E. Alexiou and T. Ebrahimi, “Point cloud quality assessment metric based on angular similarity,” *2018 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, 2018.
- [97] G. Meynet, J. Digne, and G. Lavoué, “Pc-msdm: A quality metric for 3d point clouds,” *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, pp. 1–3, 2019.
- [98] G. Meynet, Y. Nehmé, J. Digne, and G. Lavoué, “PCQM: A Full-Reference Quality Metric for Colored 3D Point Clouds,” *International Conference on Quality of Multimedia Experience*, 2020.
- [99] Q. Yang, Z. Ma, Y. Xu, Z. Li, and J. Sun, “Inferring point cloud quality via graph similarity,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.
- [100] Y. Zhao, Y. Liu, R. Song, and M. Zhang, “A saliency detection based method for 3d surface simplification,” *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 889–892, 2012.
- [101] X. Jiao, T. Wu, and X. Qin, “Mesh segmentation by combining mesh saliency with spectral clustering,” *J. Comput. Appl. Math.*, vol. 329, pp. 134–146, 2018.
- [102] S. Croci, S. B. Knorr, L. Goldmann, and A. Smolic, “A framework for quality control in cinematic vr based on voronoi patches and saliency,” *2017 International Conference on 3D Immersion (IC3D)*, pp. 1–8, 2017.

-
- [103] S. Croci, C. Ozcinar, E. Zerman, S. Knorr, J. Cabrera, and A. Smolic, “Visual attention-aware quality estimation framework for omnidirectional video using spherical voronoi diagram,” *Quality and User Experience*, vol. 5, 04 2020.
- [104] Z. Wang and A. C. Bovik, “Modern image quality assessment,” *Morgan & Claypool Publishers*, vol. 2, no. 1, 2006.
- [105] J. Lubin, “A visual discrimination model for imaging system design and evaluation,” *Vision Models for Target Detection and Recognition*, pp. 245–283, 1995.
- [106] H. Sheikh and A. Bovik, “Image information and visual quality,” *IEEE Transactions on Image Processing*, vol. 15, no. 2, pp. 430–444, 2006.
- [107] L. Zhang, L. Zhang, X. Mou, and D. Zhang, “Fsim: A feature similarity index for image quality assessment,” *IEEE Transactions on Image Processing*, vol. 20, no. 8, pp. 2378–2386, 2011.
- [108] R. Mantiuk, K. J. Kim, A. G. Rempel, and W. Heidrich, “Hdr-vdp-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions,” *ACM Trans. Graph.*, vol. 30, July 2011.
- [109] M. Narwaria, R. Mantiuk, M. P. D. Silva, and P. L. Callet, “HDR-VDP-2.2: a calibrated method for objective quality prediction of high-dynamic range and standard images,” *Journal of Electronic Imaging*, vol. 24, no. 1, pp. 1 – 3, 2015.
- [110] J. Preiss, F. Fernandes, and P. Urban, “Color-image quality assessment: From prediction to optimization,” *IEEE Transactions on Image Processing*, vol. 23, no. 3, pp. 1366–1378, 2014.
- [111] M. A. Saad, A. C. Bovik, and C. Charrier, “A dct statistics-based blind image quality index,” *IEEE Signal Processing Letters*, vol. 17, no. 6, pp. 583–586, 2010.
- [112] W. Xue, L. Zhang, X. Mou, and A. C. Bovik, “Gradient magnitude similarity deviation: A highly efficient perceptual image quality index,” *IEEE Transactions on Image Processing*, vol. 23, no. 2, pp. 684–695, 2014.
- [113] M. Čadík, R. Herzog, R. Mantiuk, R. Mantiuk, K. Myszkowski, and H.-P. Seidel, “Learning to predict localized distortions in rendered images,” *Computer Graphics Forum*, vol. 32, no. 7, pp. 401–410, 2013.
- [114] E. Prashnani, H. Cai, Y. Mostofi, and P. Sen, “Pieapp: Perceptual image-error assessment through pairwise preference,” *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.
- [115] L. Qu and G. W. Meyer, “Perceptually guided polygon reduction,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 14, no. 5, pp. 1015–1029, 2008.
- [116] N. Menzel and M. Guthe, “Towards perceptual simplification of models with arbitrary materials,” *Computer Graphics Forum*, vol. 29, 2010.

-
- [117] Q. Zhu, J. Zhao, Z. Du, and Y. Zhang, “Quantitative analysis of discrete 3d geometrical detail levels based on perceptual metric,” *Computers & Graphics*, vol. 34, no. 1, pp. 55–65, 2010.
- [118] S. J. Daly, “Visible differences predictor: an algorithm for the assessment of image fidelity,” *Human Vision, Visual Processing, and Digital Display III*, vol. 1666, pp. 2 – 15, 1992.
- [119] F. Caillaud, V. Vidal, F. Dupont, and G. Lavoué, “Progressive compression of arbitrary textured meshes,” *Computer Graphics Forum*, vol. 35, p. 475–484, Oct. 2016.
- [120] P. Lindstrom and G. Turk, “Image-driven simplification,” *ACM Transactions on Graphics*, vol. 19, p. 204–241, July 2000.
- [121] I. Abouelaziz, A. Chetouani, M. El Hassouni, and H. Cherifi, “A blind mesh visual quality assessment method based on convolutional neural network,” *Electronic Imaging*, vol. 2018, pp. 423–1, 01 2018.
- [122] I. Abouelaziz, A. Chetouani, M. El Hassouni, L. J. Latecki, and H. Cherifi, “No-reference mesh visual quality assessment via ensemble of convolutional neural networks and compact multi-linear pooling,” *Pattern Recognition*, vol. 100, p. 107174, 2020.
- [123] J. Li, L. Zou, J. Yan, D. Deng, T. Qu, and G. Xie, “No-reference image quality assessment using prewitt magnitude based on convolutional neural networks.,” *Signal, Image and Video Processing*, vol. 10, no. 4, pp. 609–616, 2016.
- [124] S. Jia and Y. Zhang, “Saliency-based deep convolutional neural network for no-reference image quality assessment,” *Multimedia Tools and Applications*, vol. 77, 06 2018.
- [125] I. Cleju and D. Saupe, “Evaluation of supra-threshold perceptual metrics for 3d models,” *Proceedings of the 3rd Symposium on Applied Perception in Graphics and Visualization*, p. 41–44, 2006.
- [126] J. J. LaViola, “A discussion of cybersickness in virtual environments,” *ACM SIGCHI Bulletin*, vol. 32, pp. 47–56, 2000.
- [127] E. Alexiou and T. Ebrahimi, “On the performance of metrics to predict quality in point cloud representations,” *Applications of Digital Image Processing XL*, vol. 10396, pp. 282 – 297, 2017.
- [128] EBU, “SAMVIQ - Subjective Assessment Methodology for Video Quality,” *Tech. Rep. BPN 056, European Broadcasting Union*, 2003.
- [129] H. Lee, G. Lavoué, and F. Dupont, “Rate-distortion optimization for progressive compression of 3D mesh with color attributes,” *The Visual Computer*, vol. 28, pp. 137–153, may 2012.

-
- [130] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, "Study of subjective and objective quality assessment of video," *IEEE Transactions on Image Processing*, vol. 19, pp. 1427–41, 06 2010.
- [131] G. Regal, R. Schatz, J. Schrammel, and S. Suette, "VRate: A Unity3D Asset for integrating Subjective Assessment Questionnaires in Virtual Environments," *10th International Conference on Quality of Multimedia Experience, QoMEX 2018*, pp. 1–3, 05 2018.
- [132] G. T. .-. V15.2.0, "Virtual Reality (VR) media services over 3GPP," *Technical Specification Group Services and System Aspects*, 2018.
- [133] K. McGraw and S. Wong, "Forming inferences about some intraclass correlation coefficients," *Psychological Methods*, vol. 1, pp. 30–46, 03 1996.
- [134] Stephane Péchard, Romuald Pepion, Patrick Le Callet, "Suitable methodology in subjective video quality assessment: a resolution dependent paradigm," *International Workshop on Image Media Quality and its Applications, IMQA2008*, 2008.
- [135] Q. Huynh-Thu, M. Garcia, F. Speranza, P. Corriveau, and A. Raake, "Study of rating scales for subjective quality assessment of high-definition video," *IEEE Transactions on Broadcasting*, vol. 57, no. 1, pp. 1–14, 2011.
- [136] P. Corriveau, C. Gojmerac, B. Hughes, and L. Stelmach, "All subjective scales are not created equal: The effects of context on different scales," *Signal Process.*, vol. 77, p. 1–9, Aug. 1999.
- [137] K. Kawashima, K. Yamagishi, and T. Hayashi, "Performance comparison of subjective quality assessment methods for 4k video," *IEICE Transactions*, vol. 101-B, pp. 933–945, 2018.
- [138] S. Ling, J. Gutiérrez, K. Gu, and P. Le Callet, "Prediction of the influence of navigation scan-path on perceived quality of free-viewpoint videos," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 9, no. 1, pp. 204–216, 2019.
- [139] Z. Li, C. Bampis, L. Janowski, and I. Katsavounidis, "A Simple Model for Subject Behavior in Subjective Experiments," *IS&T International Symposium on Electronic Imaging*, pp. 1 – 14, 2020.
- [140] M. Chen, Y. Jin, T. Goodall, X. Yu, and A. C. Bovik, "Study of 3d virtual reality picture quality," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 1, pp. 89–102, 2020.
- [141] P. Pérez, N. García, and . Villegas, "Subjective assessment of adaptive media playout for video streaming," *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, pp. 1–6, 2019.
- [142] A. Sorokin and D. Forsyth, "Utility data annotation with amazon mechanical turk," *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1–8, 2008.

-
- [143] H. Yu and S. Winkler, “Image complexity and spatial information,” *2013 Fifth International Workshop on Quality of Multimedia Experience (QoMEX)*, pp. 12–17, 2013.
- [144] U. Engelke, M. Kusuma, H.-J. Zepernick, and M. Caldera, “Reduced-reference metric design for objective perceptual quality assessment in wireless imaging,” *Signal Processing: Image Communication*, vol. 24, no. 7, pp. 525–547, 2009.
- [145] M. Abid, M. Perreira Da Silva, and P. Le Callet, “Perceptual characterization of 3d graphical contents based on attention complexity measures,” *QoEVMA’20: Proceedings of the 1st Workshop on Quality of Experience (QoE) in Visual Multimedia Applications*, p. 31–36, 2020.
- [146] M. Garland and P. S. Heckbert, “Simplifying surfaces with color and texture using quadric error metrics,” *Proceedings of the Conference on Visualization ’98*, p. 263–269, 1998.
- [147] VQEG, “Report on the validation of video quality models for high definition video content,” June 2010.
- [148] X. Huang, M.-Y. Liu, S. Belongi, and J. Kautz, “Multimodal unsupervised image-to-image translation,” *ECCV 2018*, vol. 11207, pp. 179–196, 2018.
- [149] C. Yang, Z. Wang, X. Zhu, C. Huang, J. Shi, and D. Lin, “Pose guided human video generation,” *ECCV 2018*, vol. 11214, pp. 204–219, 2018.
- [150] Y. Liu, Q. Yang, Y. Xu, and L. Yang, “Point cloud quality assessment: Dataset construction and learning-based no-reference approach,” *preprint arXiv:2012.11895*, 2021.
- [151] J. Wu, J. Ma, F. Liang, W. Dong, G. Shi, and W. Lin, “End-to-end blind image quality prediction with cascaded deep neural network,” *IEEE Transactions on Image Processing*, vol. 29, pp. 7414–7426, 2020.
- [152] S. Karunasekera and N. Kingsbury, “A distortion measure for blocking artifacts in images based on human visual sensitivity,” *IEEE Transactions on Image Processing*, vol. 4, no. 6, pp. 713–724, 1995.
- [153] Q. Liu, H. Yuan, R. Hamzaoui, H. Su, J. Hou, and H. Yang, “Reduced reference perceptual quality model with application to rate control for video-based point cloud compression,” *IEEE Transactions on Image Processing*, vol. 30, pp. 6623–6636, 2021.
- [154] A. Maggioridomo, F. Ponchio, P. Cignoni, and M. Tarini, “Real-world textured things: a repository of textured models generated with modern photo-reconstruction tools,” *preprint ArXiv:2004.14753*, 2020.
- [155] I. Lissner, J. Preiss, P. Urban, M. S. Lichtenauer, and P. Zollner, “Image-difference prediction: From grayscale to color,” *IEEE Transactions on Image Processing*, vol. 22, no. 2, pp. 435–446, 2013.
- [156] P. Alliez, S. Tayeb, and C. Wormser, “3D fast intersection and distance computation,” *CGAL Editorial Board edition 3.5, CGAL User and Reference Manual*, 2009.

-
- [157] P. Alliez, D. Cohen-Steiner, O. Devillers, B. Lévy, and M. Desbrun, “Anisotropic polygonal remeshing,” *ACM Trans. Graph.*, vol. 22, p. 485–493, July 2003.
- [158] I. Lissner and P. Urban, “Toward a unified color space for perception-based image processing,” *IEEE Transactions on Image Processing*, vol. 21, no. 3, pp. 1153–1168, 2012.
- [159] Y. Liu, J. Wang, S. Cho, A. Finkelstein, and S. Rusinkiewicz, “A no-reference metric for evaluating the quality of motion deblurring,” *ACM Transactions on Graphics*, vol. 32, no. 6, pp. 171–175, 2013.
- [160] L. Krasula, K. Fliegel, P. Le Callet, and M. Klíma, “On the accuracy of objective image and video quality models: New methodology for performance evaluation,” *2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)*, pp. 1–6, 2016.
- [161] F. Xiao, “Dct-based video quality evaluation,” *Final Project for EE392J*, vol. 769, 2000.
- [162] Z. Wang, E. P. Simoncelli, and A. C. Bovik, “Multiscale structural similarity for image quality assessment,” *The Thrity-Seventh Asilomar Conference on Signals, Systems Computers, 2003*, vol. 2, pp. 1398–1402, 2003.
- [163] L. Itti, C. Koch, and E. Niebur, “A model of saliency-based visual attention for rapid scene analysis,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.
- [164] W. Zhang, R. R. Martin, and H. Liu, “A saliency dispersion measure for improving saliency-based image quality metrics,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 6, pp. 1462–1466, 2018.
- [165] H. Liu and I. Heynderickx, “Visual attention in objective image quality assessment: Based on eye-tracking data,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 21, no. 7, pp. 971–982, 2011.
- [166] H. Liu, U. Engelke, J. Wang, P. Le Callet, and I. Heynderickx, “How does image content affect the added value of visual attention in objective image quality assessment?,” *IEEE Signal Processing Letters*, vol. 20, no. 4, pp. 355–358, 2013.
- [167] M. Jiang, S. Huang, J. Duan, and Q. Zhao, “Salicon: Saliency in context,” *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1072–1080, 2015.
- [168] D. L. Medin, R. L. Goldstone, and D. Gentner, “Respects for similarity,” *Psychological Review*, vol. 100, pp. 254–278, 1993.
- [169] Q. Chen and V. Koltun, “Photographic image synthesis with cascaded refinement networks,” *ICCV*, pp. 1520–1529, 10 2017.

- [170] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, “Squeezenet: Alexnet-level accuracy with 50x fewer parameters and 0.5mb model size,” *CVPR*, 2017.
- [171] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” *Advances in Neural Information Processing Systems*, vol. 25, pp. 1097–1105, 2012.
- [172] D. Yamins and J. J. DiCarlo, “Using goal-driven deep learning models to understand sensory cortex,” *Nature Neuroscience*, vol. 19, pp. 356–365, 2016.
- [173] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *ImageNet Challenge*, vol. abs/1409.1556, 2014.
- [174] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, “ImageNet Large Scale Visual Recognition Challenge,” *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

Appendix A

This appendix contains related data (tables, graphs, illustrations, etc.) that complement and clarify notions presented in the chapters of the manuscript.

The appendix is organized as follows: each section provides additional materials for a given chapter.

A.1 Chapter 2

Understanding Boxplots

In descriptive statistics, a boxplot is a visualization method for graphically depicting groups of numerical data through their quartiles. Figure A.1 illustrates the type of boxplot used in this manuscript and its different parts. The spacings between the different parts of the box indicate the degree of dispersion and skewness in the data, and show outliers. For instance, Q1 and Q3 are thresholds below which 25% and 75% of the data points fall respectively. InterQuartile Range ($IQR = Q3 - Q1$) represents how 50% of the points were dispersed.

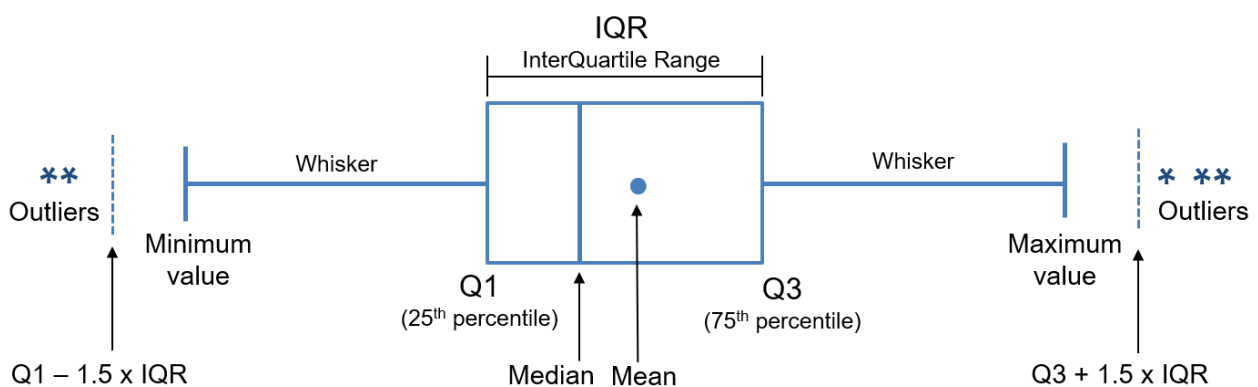


Figure A.1: Different parts of a boxplot.

Distortions

Figure A.2 presents the snapshots of the 80 distorted models generated from 5 source models (meshes with vertex colors) \times 4 distortion types (QGeo, QCol, SGeo, SCol) \times 4 distortion strengths. These stimuli constitute the dataset used to compare the subjective methodologies in Chapter 2. Acronyms in the Figure A.2 refer to Distortion Type_Strength.

Resulting MOSs and DMOSs

We provide, in Figure A.3, the MOSs OF all stimulus along with their Confidence Intervals (CIs) obtained for the two groups of participants (G1 and G2) in the DSIS tests. Figure A.4 shows the DMOSs and CIs in the ACR-HR tests for G1 and G2. We recall that G1's subjects did the ACR-HR session first followed by the DSIS session, while G2's subjects did the DSIS session first and then the ACR-HR session.

Figure A.5 presents the DMOSs and the CIs acquired for all the stimuli in the SAMVIQ test.

QGeo_1

QGeo_2

QGeo_3

QGeo_4



QCol_1

QCol_2

QCol_3

QCol_4



SGeo_1

SGeo_2

SGeo_3

SGeo_4



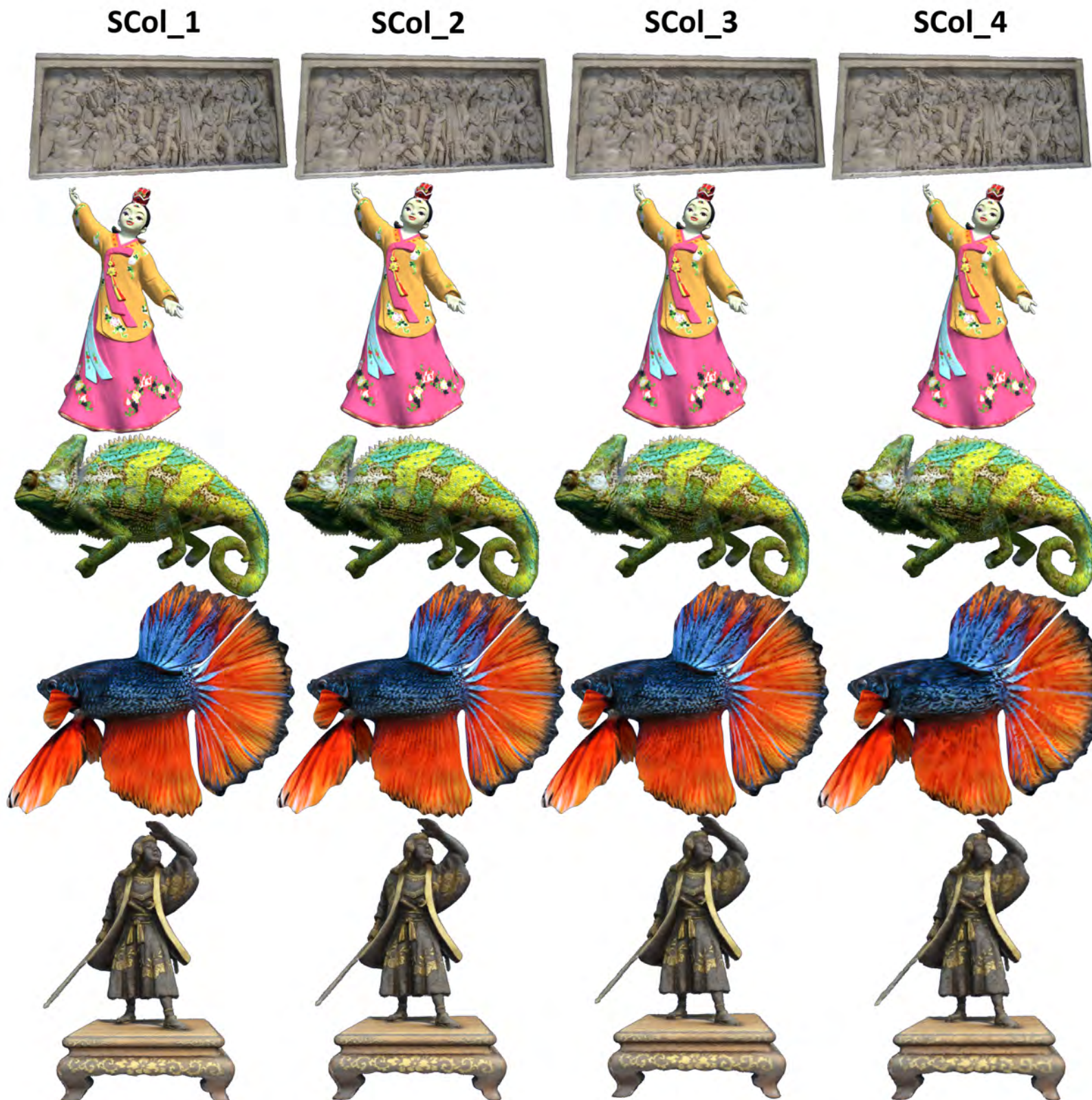
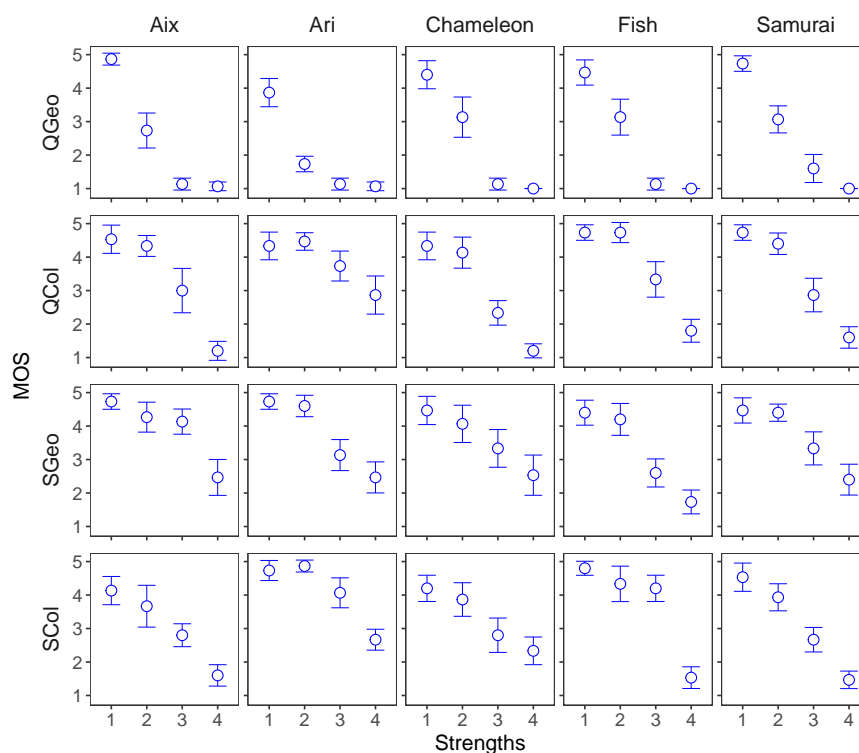
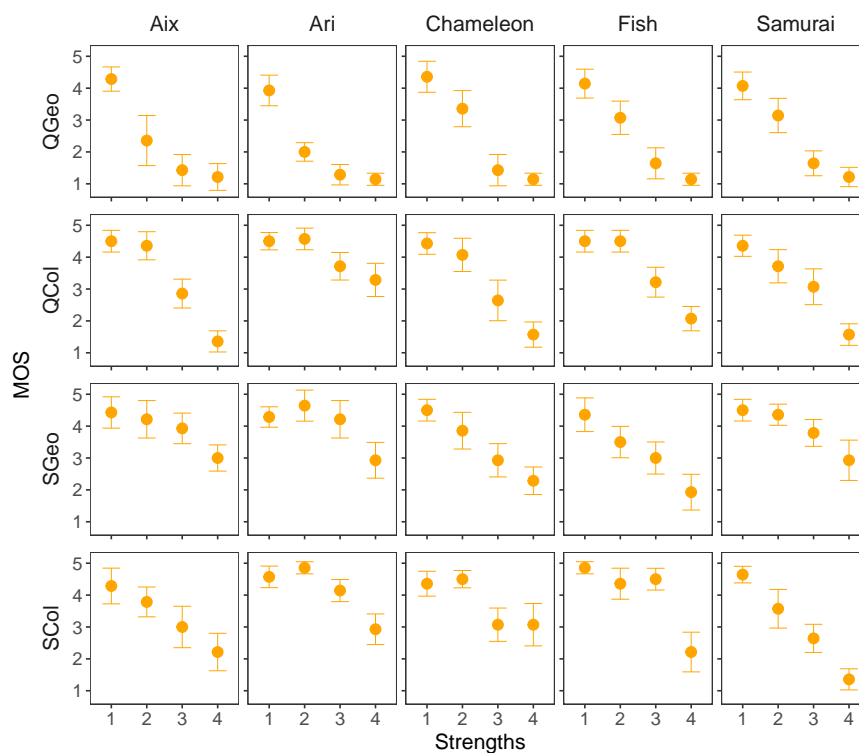


Figure A.2: Snapshots of the stimuli from the dataset used to compare the subjective methodologies. Acronyms refer to Distortion Type_Strength.

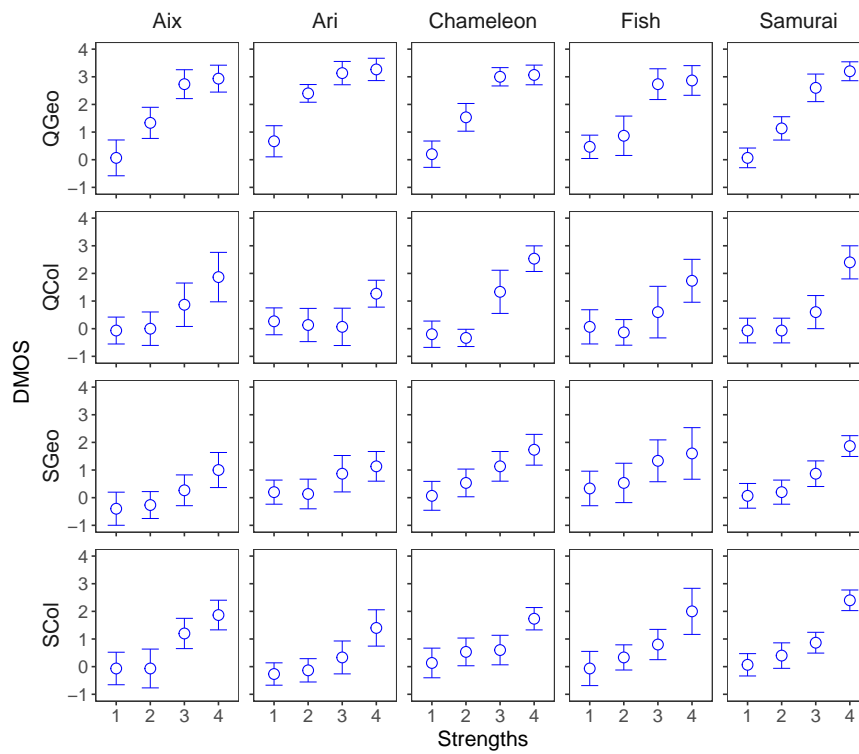


(a) G1

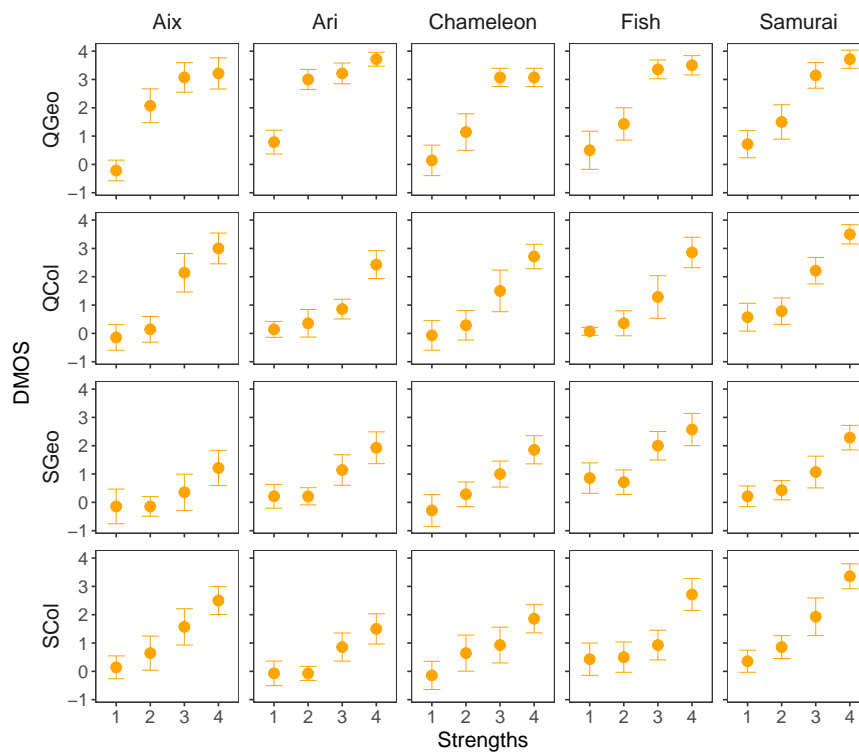


(b) G2

Figure A.3: MOSs and confidence intervals obtained in the DSIS tests for the two groups of participants.



(a) G1



(b) G2

Figure A.4: DMOSs and confidence intervals obtained in the ACR-HR tests for the two groups of participants.

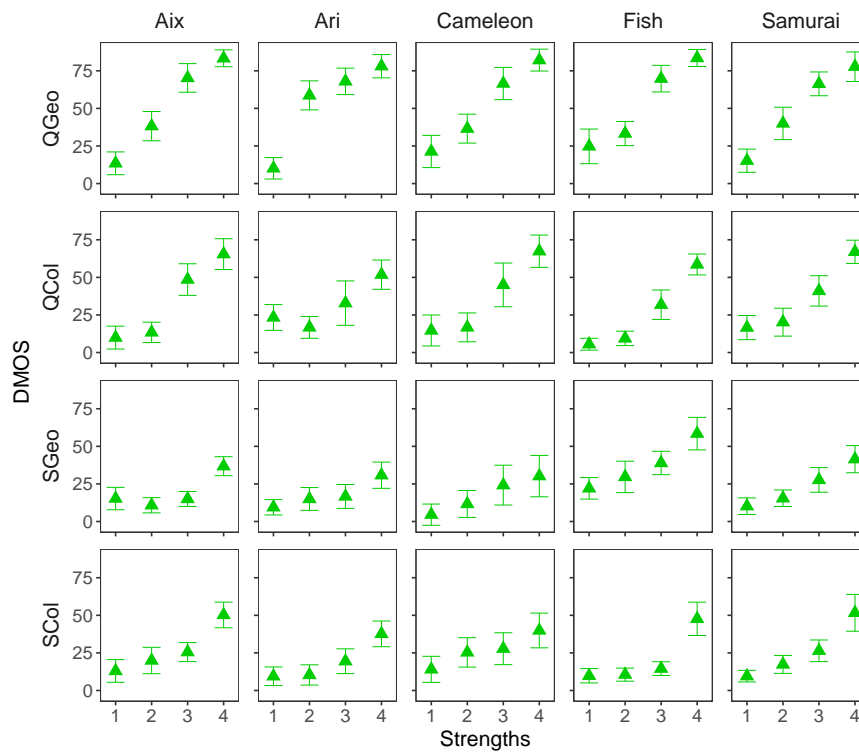
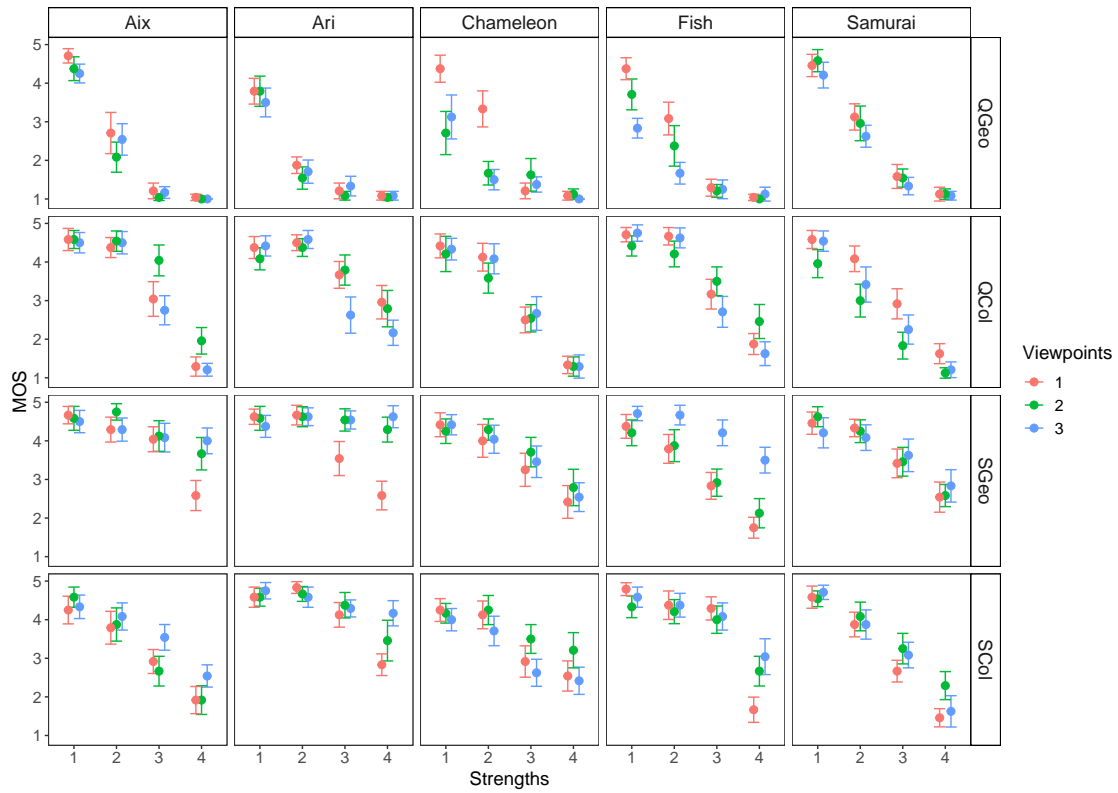


Figure A.5: DMOSs and confidence intervals obtained in the SAMVIQ tests.

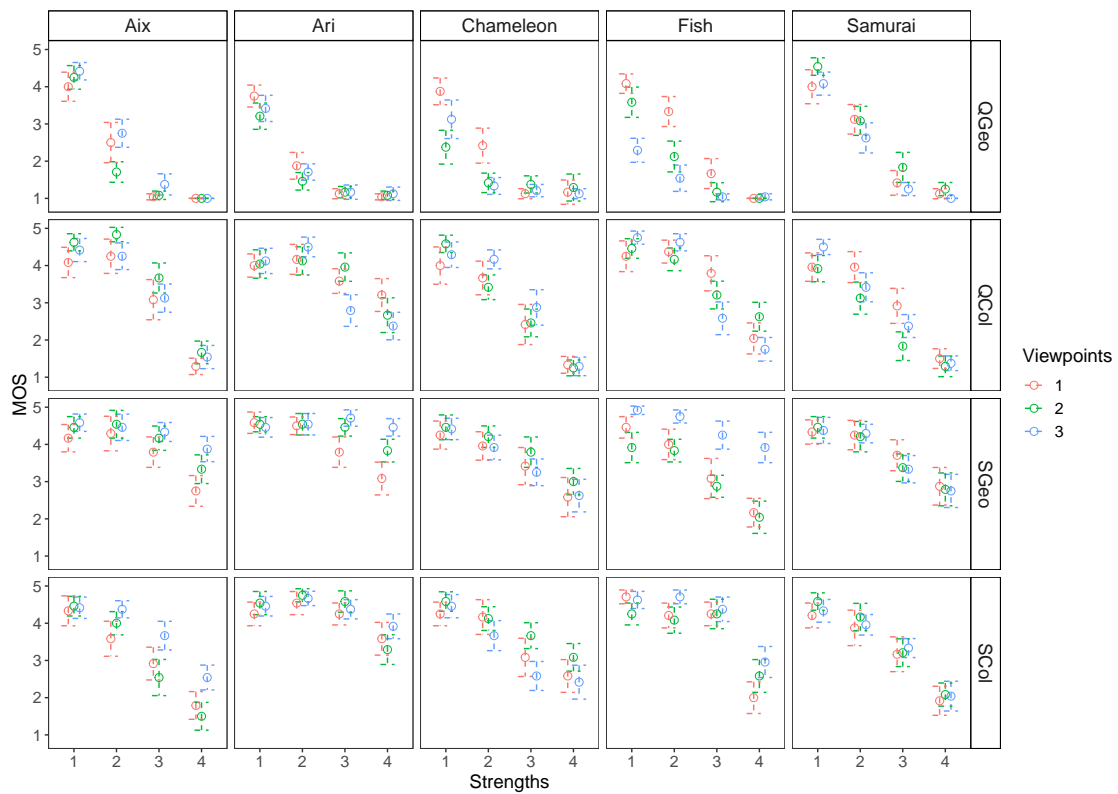
A.2 Chapter 3

Resulting MOSs and CIs

In this subsection, we present the MOSs and CIs obtained for our ground truth database of 480 animated 3D meshes with vertex colors. Figure A.6.a shows the results of the stimuli in rotation, while Figure A.6.b shows the results of those in zoom.



(a) Stimuli animated with a slow rotation (R).



(b) Stimuli animated with a slow Zoom (Z).

Figure A.6: MOSs and CIs of the 480 stimuli from the dataset of meshes with vertex colors. For a given distortion strength, the dots are horizontally spaced apart to avoid overlapping.

A.3 Chapter 4

Content ambiguity

Content ambiguity is related to the dispersion of subjective scores (CIs). Therefore for the lab and CS test, we averaged the CIs of the stimuli of the dataset, described in section 4.1, over the models. Results are shown in Figure A.7.

Note that, since we have 29 ratings per stimulus in the lab test and at least 56 in the CS test, we considered the same number of rating per stimulus ($N = 29$) for the lab and CS experiments in order to have a fair comparison. Thus, for the CS experiment, we randomly selected 100 combinations of 29 ratings for each stimulus. We averaged the CI values over these combinations of ratings.

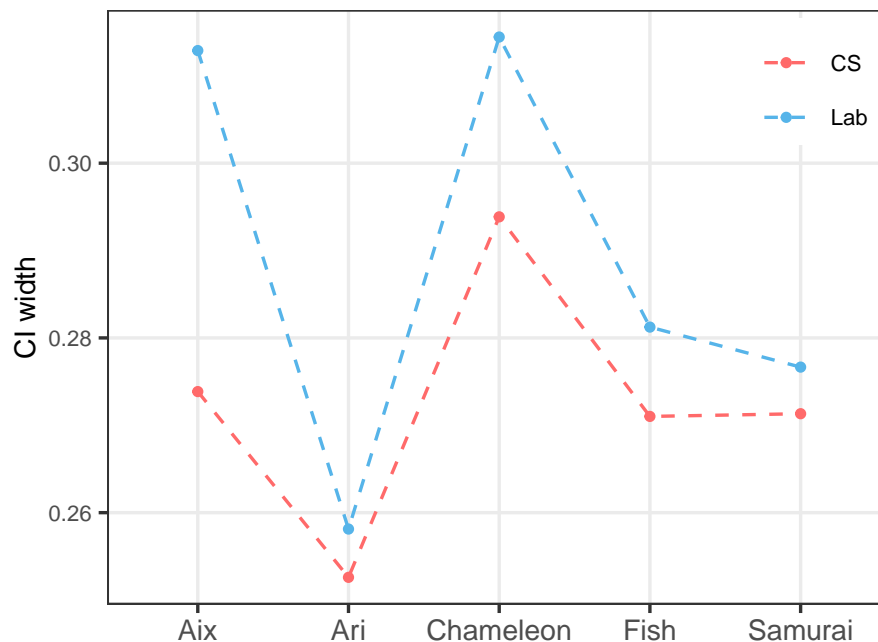


Figure A.7: Mean confidence intervals of each source model for the CS and lab experiments.

All the source models have larger CIs in the lab experiment. We obtained almost the same shape of curves as in Figure 4.9, meaning that models with high CIs are also associated with high ambiguity values.

A.4 Chapter 5

Application: Rate-Distortion control

We conducted a preliminary qualitative analysis to find the optimal compression setting for the source models of our textured meshes dataset, under the constraint of a target bitrate. The goal is to find for each of our 55 models the parameters of distortion providing the best possible visual quality for a given size requirement (in KiloBytes). The size of a stimuli (in KB) is equal to the sum of the size of its compressed texture and its compressed 3D model. The results are reported below in Tables A.8, A.9 et A.10. In each cell, we find the combination of distortion parameters needed to achieve the highest possible quality for a given size range. The cell color indicates the range of quality we can achieve within the given size range. Our source models are sorted in ascending order according to their geometric complexity SI_{Geo} (Table A.8), their color complexity SI_{Col} (Table A.9), and the amount of vertices on the texture seams V_{Tseams} (Table A.10).

	$V_{Tseams1}$	$V_{Tseams2}$	$V_{Tseams3}$	$V_{Tseams1}$	$V_{Tseams5}$
#46	L9 8 9 512x512 10	L4 10 10 9 512x512 50	L11 11 10 712x712 50	L11 11 10 1024x1024 50	L11 11 10 1024x1024 50
#18	L9 9 6 512x512 10	L8 9 8 512x512 10	L5 10 8 512x512 25	L3 10 9 712x712 25	L5 11 10 1024x1024 25
#27	L9 9 6 512x512 10	L7 11 8 712x712 25	L8 9 8 512x512 10	L7 9 9 712x712 10	L2 11 10 1024x1024 25
#26	L10 10 6 512x512 10	L9 7 6 512x512 10	L8 9 7 512x512 25	L7 10 9 512x512 125	L2 11 10 1024x1024 10
#41	L9 9 7 512x512 10	L9 9 7 512x512 10	L7 9 9 512x512 25	L5 10 10 1024x1024 10	L5 11 9 712x712 125
#30	L8 9 6 512x512 10	L6 11 9 712x712 25	L6 11 9 512x512 25	L3 11 10 1024x1024 10	L2 11 10 1024x1024 10
#25	L10 10 10 512x512 10	L10 10 10 512x512 10	L9 10 7 512x512 25	L8 10 8 512x512 25	L6 10 10 712x712 50
#13	L9 8 7 512x512 10	L9 8 7 512x512 10	L8 9 9 512x512 25	L7 10 9 512x512 10	L4 10 10 512x512 50
#29	L10 9 10 512x512 25	L10 11 9 1024x1024 25	L10 10 10 1024x1024 10	L9 11 10 1024x1024 10	L9 11 10 1024x1024 25
#1	L10 10 6 512x512 10	L10 9 8 512x512 30	L9 9 9 712x712 75	L9 11 10 1024x1024 75	L9 11 10 1024x1024 90
#3	L10 10 6 512x512 10	L10 10 10 512x512 10	L10 10 10 1024x1024 25	L9 10 8 712x712 25	L9 10 10 1024x1024 10
#5	L10 10 6 512x512 10	L9 9 8 512x512 10	L8 11 10 512x512 25	L9 9 8 712x712 25	L9 9 8 712x712 50
#8	L10 10 6 512x512 10	L10 10 10 1024x1024 10	L9 9 8 512x512 25	L7 10 10 512x512 75	L8 11 10 1024x1024 10
#19	L10 10 10 512x512 10	L10 11 10 712x712 25	L9 8 8 512x512 25	L9 9 9 712x712 50	L8 10 10 1024x1024 50
#29	L10 10 10 512x512 10	L9 8 9 712x712 10	L8 10 10 1024x1024 10	L7 10 10 1024x1024 10	L5 10 10 1024x1024 25
#39	L10 8 8 512x512 10	L10 8 9 512x512 50	L9 9 9 512x512 10	L9 11 9 512x512 50	L9 10 10 1024x1024 25
#23	L9 7 6 512x512 10	L4 11 8 512x512 10	L11 11 10 712x712 25	L11 11 10 512x512 75	L11 11 10 1024x1024 50
#16	L10 10 10 512x512 10	L10 11 10 512x512 25	L10 8 10 512x512 75	L10 8 10 1024x1024 25	L9 11 10 1024x1024 10
#6	L10 10 8 512x512 10	L10 10 9 712x712 25	L9 10 8 512x512 25	L9 11 10 512x512 75	L8 11 10 1024x1024 10
#2	L10 9 10 512x512 10	L10 10 9 712x712 25	L9 8 9 512x512 10	L8 11 10 512x512 25	L7 11 10 1024x1024 10
#49	L10 10 10 512x512 10	L9 8 7 712x712 10	L9 8 9 712x712 50	L8 10 10 1024x1024 25	L8 11 10 1024x1024 75
#21	L10 10 10 512x512 10	L9 8 8 712x712 10	L9 8 9 1024x1024 25	L7 9 10 1024x1024 25	L8 11 10 1024x1024 10
#37	L10 10 10 512x512 10	L10 10 10 1024x1024 10	L9 9 9 512x512 25	L7 11 10 712x712 25	L8 11 9 712x712 25
#36	L10 10 6 512x512 10	L10 10 10 1024x1024 10	L9 9 6 512x512 25	L8 11 10 712x712 25	L8 11 9 712x712 75
#7	L10 10 6 512x512 10	L10 10 10 1024x1024 10	L10 10 10 712x712 10	L10 10 10 512x512 75	L9 9 9 512x512 10
#32	L10 7 8 512x512 10	L10 11 10 512x512 50	L9 11 10 512x512 90	L9 11 10 1024x1024 50	L8 11 10 1024x1024 75
#40	L10 10 10 512x512 10	L9 8 6 512x512 125	L8 9 9 712x712 25	L6 9 9 712x712 25	L5 9 10 712x712 50
#22	L10 9 10 512x512 10	L9 10 10 512x512 25	L8 11 10 512x512 75	L5 11 10 1024x1024 25	L5 11 10 1024x1024 50
#43	L10 9 8 512x512 10	L9 10 9 512x512 10	L7 9 10 512x512 25	L4 10 10 512x512 25	L2 11 10 1024x1024 25
#47	L10 9 8 512x512 10	L9 8 7 512x512 125	L8 9 10 512x512 50	L6 10 9 512x512 50	L4 10 10 1024x1024 25
#5	L10 11 8 512x512 10	L10 11 8 512x512 10	L9 10 10 512x512 50	L8 11 10 512x512 50	L4 10 10 1024x1024 25
#35	L10 11 8 512x512 10	L9 8 7 512x512 10	L8 9 10 512x512 25	L7 10 9 712x712 25	L6 11 10 512x512 75
#41	L10 8 8 512x512 10	L10 10 10 512x512 10	L9 10 10 512x512 25	L6 11 10 512x512 50	L5 11 10 512x512 90
#24	L10 8 8 512x512 10	L10 11 8 512x512 50	L9 10 10 512x512 25	L6 11 10 512x512 50	L3 11 10 1024x1024 75
#28	L10 11 8 712x712 10	L10 11 8 712x712 10	L9 10 10 512x512 75	L9 11 10 512x512 75	L3 11 10 1024x1024 90
#50	L10 10 8 512x512 10	L9 8 6 512x512 10	L9 11 9 512x512 50	L8 11 10 1024x1024 25	L8 11 10 1024x1024 90
#17	L10 10 8 512x512 10	L10 10 10 1024x1024 10	L9 9 9 512x512 10	L8 11 10 1024x1024 25	L6 11 10 1024x1024 90
#10	L10 10 10 512x512 10	L9 8 9 512x512 10	L9 11 10 512x512 50	L8 11 10 1024x1024 25	L6 11 10 1024x1024 90
#4	L10 10 10 512x512 10	L9 10 10 512x512 25	L7 11 10 512x512 50	L6 11 10 512x512 90	L6 11 10 1024x1024 90
#48	L10 9 8 512x512 25	L9 8 9 512x512 25	L8 10 10 712x712 25	L5 11 10 512x512 90	L3 11 10 1024x1024 90
#33	L10 9 8 512x512 25	L9 8 9 512x512 25	L9 11 10 712x712 25	L4 11 10 1024x1024 25	L1 11 10 1024x1024 75
#9	L10 9 8 712x712 10	L9 10 8 512x512 125	L8 11 10 712x712 25	L4 11 10 1024x1024 25	L1 11 10 1024x1024 90
#45	L10 10 10 512x512 125	L10 10 10 512x512 150	L9 11 10 512x512 75	L3 11 10 1024x1024 75	L1 11 10 1024x1024 90
#64	L10 8 7 712x712 10	L10 9 10 712x712 25	L8 10 9 712x712 25	L8 11 10 1024x1024 25	L5 11 10 1024x1024 75
#11	L10 11 10 512x512 50	L9 9 9 712x712 50	L8 10 10 1024x1024 50	L7 11 10 1024x1024 75	L4 11 10 1024x1024 75
#4	L10 10 10 512x512 10	L10 10 10 712x712 25	L9 10 10 512x512 75	L8 11 10 1024x1024 50	L5 11 10 2048x2048 50
#53	L10 11 10 512x512 10	L10 11 10 512x512 150	L9 10 10 512x512 75	L8 11 10 512x512 75	L5 11 10 2048x2048 90
#34	L10 7 6 512x512 10	L10 10 10 1024x1024 10	L9 9 9 512x512 25	L9 10 10 1024x1024 25	L7 11 10 1024x1024 75
#20	L10 9 8 712x712 25	L9 8 9 512x512 50	L9 11 10 712x712 75	L8 10 10 712x712 75	L3 11 10 1024x1024 75

MOS

1-2

2-3

3-4

4-5

Figure A.10: Optimal distortion parameters for each source model, providing the best possible visual quality for a given size requirement. Models are sorted by the amount of vertices present on texture seams V_{Tseams} .

A.5 Chapter 6

Settings for image quality metric

To evaluate the performance of our proposed metric *CMDM*, we compared it to 3 state-of-the-art full-reference Image Quality Metrics (IQMs): *SSIM* [81], *HDR-VDP2* [108], *iCID* [110]. To apply these IQMs, we generated for each 3D object of the database, a set of 18 snapshots taken from different viewpoints: the camera was placed at regularly sampled positions around the vertical axis of the stimulus, as shown in Figure A.11.



Figure A.11: Camera positions regularly sampled around a 3D object.



FOLIO ADMINISTRATIF

THESE DE L'UNIVERSITE DE LYON OPEREE AU SEIN DE L'INSA LYON

NOM : NEHME
(avec précision du nom de jeune fille, le cas échéant)

DATE de SOUTENANCE : 03/12/2021

Prénoms : Yana

TITRE : Visual Quality of Rendered 3D Meshes with Color Attributes: Subjective and Objective Evaluation

NATURE : Doctorat

Numéro d'ordre : 2021LYSEI086

Ecole doctorale : Informatique et Mathématiques de Lyon (N°512)

Spécialité : Infomatique

RESUME : As technological advances and capabilities in the field of computer graphics grow day by day, the need to master the visualization and processing of 3D data increases at an equal pace. Indeed, the development of modeling software and acquisition devices makes 3D graphics rich and realistic: complex models enriched with various appearance attributes. The way this 3D content is consumed is also evolving from standard screens to Virtual and Mixed Reality (VR/MR). However, the size and complexity of these rich 3D models often make their interactive visualization problematic. This is particularly the case in immersive environments and online applications. Thus, to avoid latency and rendering issues, diverse processing operations, including simplification and compression, are usually applied, resulting in a loss of quality in the final rendering. Therefore, both subjective studies and objective metrics are needed to predict this visual loss and to assess the quality as perceived by human observers. In this thesis, we address the aforementioned challenges. We conduct an extensive study to determine the best subjective quality assessment methodology to adopt for assessing the visual quality of 3D graphics, especially in VR. We establish two quality assessment datasets composed of meshes with vertex colors and textured meshes, respectively. The former is produced in VR and the latter in crowdsourcing. To the best of our knowledge, these are the largest datasets for meshes with color attributes to date. Moreover, we provide an in-depth analysis of the influence of source model characteristics, distortion interactions, viewpoints and animations on the perceived quality of 3D meshes. Leveraging our two established datasets, we propose two data-driven perceptual metrics for quality assessment of 3D graphics with color attributes. The first metric is model-based while the second is an image-based metric that employs convolutional neural networks. Our metrics demonstrate state-of-the-art results on our two datasets. Lastly, we investigate how incorporating visual attention into our perceptual quality metric improves the predicted quality. The datasets and the source code of the metrics are publicly available.

MOTS-CLÉS : Visual Quality Assessment, Subjective Quality Evaluation, Objective Quality Evaluation, 3D Graphics, 3D Meshes, Color Attributes, Textures, Vertex Colors, Subjective Methodologies, Datasets, Perceptual Metrics, CNN, Visual Attention, Virtual Reality, Crowdsourcing.

Laboratoire (s) de recherche : Laboratoire d'Informatique en Image et Systèmes d'Information (LIRIS)

Directeur de thèse: Guillaume Lavoué

Président de jury : Pierre Alliez

Composition du jury : Pierre Alliez, Luce Morin, Aljosa Smolic, Rafal Mantiuk, Véronique Eglin, Guillaume Lavoué, Patrick Le Callet, Florent Dupont