TU Delft

Delft University of Technology

Document Version
Final published version

Citation (APA)
Zhou, X. (2026). *Human-centric quality assessment and visual attention modeling for point clouds*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:3b9f08a9-8adc-40b2-9fff-e193129d1cc0

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
In case the licence states “Dutch Copyright Act (Article 25fa)”, this publication was made available Green Open Access via the TU Delft Institutional Repository pursuant to Dutch Copyright Act (Article 25fa, the Taverne amendment). This provision does not affect copyright ownership.
Unless copyright is transferred by contract or statute, it remains with the copyright holder.

Sharing and reuse
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.

Human-Centric Quality Assessment & Visual Attention Modeling for Point Clouds

Xuemei Zhou

Human-Centric Quality Assessment and Visual Attention Modeling for Point Clouds

Human-Centric Quality Assessment and Visual Attention Modeling for Point Clouds

Human-Centric Quality Assessment and Visual Attention Modeling for Point Clouds

for the purpose of obtaining the degree of doctor
at Delft University of Technology
by the authority of the Rector Magnificus, prof. dr. ir. H.Bijl,
chair of the Board for Doctorates
to be defended publicly on
Wednesday, 4 March 2026 at 10:00 am.

by

Xuemei ZHOU

This dissertation has been approved by the promotors.

Composition of the doctoral committee:

Rector Magnificus,	chairperson
Prof. dr. P. S. Cesar Garcia,	Delft University of Technology, *promotor*
Dr. I. Viola,	Centrum Wiskunde & Informatica, *copromotor*
Independent members:	
Prof. dr. A. Hanjalic	Delft University of Technology, NL
Dr. M.W.A. Wijntjes	Delft University of Technology, NL
Prof. dr. P. Le Callet	Polytech Nantes, France
Prof. dr. M.G. Martini	King's College London, UK
Dr. Z. Yumak	Utrecht University, NL
Prof. dr. ir. A. Bozzon,	Delft University of Technology, NL, reserve member

Keywords: Virtual Reality, Point Cloud, Visual Saliency, Quality Assesment

Copyright © 2026 by X. Zhou

ISBN 978-94-6518-263-6

An electronic copy of this dissertation is available at
https://repository.tudelft.nl/.

Contents

Summary

Human-Centric Quality Assessment and Visual Attention Modeling for Point Clouds addresses a central challenge in immersive three-dimensional media: how to measure, understand, and predict human perceptual experience when interacting with high-fidelity point-cloud content in virtual and mixed-reality environments. Point clouds are increasingly adopted in extended reality (XR), telepresence, free-viewpoint video, and autonomous systems due to their flexibility in representing complex 3D scenes. However, their irregular structure, large data volume, and the coupled effects of geometric and texture distortions make perceptual quality assessment particularly challenging. Widely used image and video quality assessment metrics are known to correlate poorly with human Mean Opinion Scores (MOS) when directly applied to point clouds, highlighting the need for perceptually grounded quality models tailored to this media type.

This dissertation develops methodologies, datasets, and objective metrics that bridge signal processing techniques with human-centred evaluation in order to improve perceptual alignment in Point Cloud Quality Assessment (PCQA). Beyond metric design, the thesis also provides a human-centred experimental framework and prototype system to support subjective quality assessment and visual saliency analysis in immersive environments, enabling controlled yet valid perceptual studies.

The thesis is organized into three complementary parts. The first part focuses on objective PCQA. Novel quality metrics are proposed to capture the combined influence of geometry and texture on perceived quality. A full-reference metric, PointPCA+, and a no-reference model, M3-Unity, are introduced. These approaches employ modality-aware feature representations that explicitly account for the characteristics of point-cloud geometry and color attributes, together with carefully designed similarity measures or learning-based regression models. Extensive evaluations on public benchmark datasets demonstrate that the proposed metrics achieve improved correlation with human MOS compared to state-of-the-art methods, while also exhibiting robustness across different compression schemes and distortion types. These results confirm the importance of modality-specific modeling and perceptually motivated feature design for PCQA.

The second part of the thesis investigates human visual attention for dynamic point-cloud content in immersive environments. Eye-tracking experiments were conducted in six-degrees-of-freedom virtual reality settings to capture gaze behavior during the viewing of dynamic point clouds. Two datasets were constructed: a task-dependent dataset (QAVA-DPC), in which participants performed explicit visual tasks, and a task-free dataset (TF-DPC), designed to capture natural viewing behavior. The experimental design incorporates systematic preprocessing, stimulus normalization, and error profiling of head-mounted display eye trackers to ensure data reliability and reproducibility. These datasets enable a detailed analysis of how task demands, motion, and temporal dynamics influence visual saliency and viewing behavior in immersive 3D scenes.

The third part explores the integration of visual saliency into objective quality assessment. By incorporating both ground-truth and predicted saliency maps into PCQA pipelines, the thesis examines how attention-guided feature weighting and perceptual pooling strategies affect quality prediction performance. Experimental results show that saliency-aware approaches can improve predictive accuracy, although the magnitude of improvement depends on the underlying quality metric and pooling strategy. These findings highlight that visual attention is a valuable perceptual cue, but must be integrated in a principled and task-aware manner. Overall, the results demonstrate that attention-aware models can better prioritize perceptually relevant distortions, which is particularly beneficial for applications such as point-cloud compression, adaptive streaming, and immersive media delivery.

In addition to methodological advances, this dissertation contributes a range of resources that support reproducible research and open science for the research community. The gaze-annotated datasets for dynamic point clouds are made available to support reproducible research and facilitate further investigation of visual attention in immersive media. The thesis also documents detailed experimental protocols inspired by ITU recommendations and adapted to extended-reality settings, contributing to ongoing efforts toward standardization and open science in volumetric media research.

SAMENVATTING

Mensgerichte kwaliteitsbeoordeling en modellering van visuele aandacht voor puntwolken behandelt een centrale uitdaging binnen immersieve driedimensionale media: hoe de menselijke perceptuele ervaring kan worden gemeten, begrepen en voorspeld bij interactie met hoogwaardige puntwolk-inhoud in virtuele en gemengde-realiteitomgevingen. Puntwolken worden steeds vaker toegepast in extended reality (XR), telepresence, free-viewpoint video en autonome systemen vanwege hun flexibiliteit bij het representeren van complexe 3D-scènes. Hun onregelmatige structuur, grote datavolume en de gekoppelde effecten van geometrische en textuurvervormingen maken perceptuele kwaliteitsbeoordeling echter bijzonder uitdagend. Veelgebruikte beeld- en videokwaliteitsmaten correleren slecht met menselijke Mean Opinion Scores (MOS) wanneer zij direct op puntwolken worden toegepast, wat de noodzaak onderstreept van perceptueel onderbouwde kwaliteitsmodellen die specifiek zijn afgestemd op dit mediatype.

Dit proefschrift ontwikkelt methodologieën, datasets en objectieve maatstaven die signaalverwerkingstechnieken verbinden met mensgerichte evaluatie, met als doel een betere perceptuele afstemming in Point Cloud Quality Assessment (PCQA) te realiseren. Naast metriekontwerp biedt het proefschrift een mensgericht experimenteel raamwerk en een prototypesysteem ter ondersteuning van subjectieve kwaliteitsbeoordeling en visuele-saliëntieanalyse in immersieve omgevingen, waardoor gecontroleerde maar ecologisch valide perceptuele studies mogelijk worden gemaakt.

Het proefschrift is opgebouwd uit drie complementaire delen. Het eerste deel richt zich op objectieve PCQA. Er worden nieuwe kwaliteitsmaten voorgesteld die de gecombineerde invloed van geometrie en textuur op de waargenomen kwaliteit modelleren. Een full-reference metriek, PointPCA+, en een no-reference model, M3-Unity, worden geïntroduceerd. Deze benaderingen maken gebruik van modaliteitsbewuste feature-representaties die expliciet rekening houden met de kenmerken van puntwolk-geometrie en kleurattributen, in combinatie met zorgvuldig ontworpen similariteitsmaten of leergebaseerde regressiemodellen. Uitgebreide evaluaties op openbare benchmarkdatasets tonen aan dat de voorgestelde methoden een verbeterde correlatie met menselijke MOS bereiken ten opzichte van de stand van de techniek, en tevens robuust zijn over verschillende compressieschema's en vervormingstypen. Deze resultaten bevestigen het belang van modaliteitsspecifieke modellering en perceptueel gemotiveerd feature-ontwerp voor PCQA.

Het tweede deel van het proefschrift onderzoekt menselijke visuele aandacht bij dynamische puntwolk-inhoud in immersieve omgevingen. In zes vrijheidsgraden (6DoF) virtual-realitysettings zijn eye-trackingexperimenten uitgevoerd om kijkgedrag tijdens het bekijken van dynamische puntwolken vast te leggen. Er zijn twee datasets geconstrueerd: een taakafhankelijke dataset (QAVA-DPC), waarin deelnemers expliciete visuele taken uitvoerden, en een taakvrije dataset (TF-DPC), bedoeld om natuurlijk kijkgedrag vast te leggen. Het experimentele ontwerp omvat systematische preprocessing, stimulusnormali-

satie en foutprofilering van eye trackers in head-mounted displays om de betrouwbaarheid en reproduceerbaarheid van de data te waarborgen. Deze datasets maken een gedetailleerde analyse mogelijk van de invloed van taakeisen, beweging en temporele dynamiek op visuele saliëntie en kijkgedrag in immersieve 3D-scènes.

Het derde deel verkent de integratie van visuele saliëntie in objectieve kwaliteitsbeoordeling. Door zowel grondwaarheids- als voorspelde saliëntiekaarten te integreren in PCQA-pijplijnen, wordt onderzocht hoe aandachtgestuurde featureweging en perceptuele poolingstrategieën de prestaties van kwaliteitsvoorspelling beïnvloeden. Experimentele resultaten tonen aan dat saliëntiebewuste benaderingen de voorspellende nauwkeurigheid kunnen verbeteren, hoewel de mate van verbetering afhankelijk is van de onderliggende kwaliteitsmetriek en de toegepaste poolingstrategie. Deze bevindingen benadrukken dat visuele aandacht een waardevolle perceptuele aanwijzing is, maar op een principiële en taakbewuste manier moet worden geïntegreerd. In het algemeen laten de resultaten zien dat aandachtbewuste modellen perceptueel relevante vervormingen beter kunnen prioriteren, wat met name voordelig is voor toepassingen zoals puntwolk-compressie, adaptieve streaming en levering van immersieve media.

Naast methodologische vooruitgang levert dit proefschrift een reeks middelen die reproduceerbaar onderzoek en open science binnen de onderzoeksgemeenschap ondersteunen. De met blikdata geannoteerde datasets voor dynamische puntwolken worden openbaar beschikbaar gesteld om reproduceerbaar onderzoek te bevorderen en verder onderzoek naar visuele aandacht in immersieve media te faciliteren. Daarnaast documenteert het proefschrift gedetailleerde experimentele protocollen, geïnspireerd door ITU-aanbevelingen en aangepast aan extended-realitycontexten, waarmee wordt bijgedragen aan lopende inspanningen op het gebied van standaardisatie en open science in volumetrisch mediaonderzoek.

1

Introduction

1.1. Background and Motivation

In today's digital landscape, multimedia systems have significantly transformed the way we capture, process, visualize, and consume three-dimensional (3D) data. The evolution has moved beyond traditional 2D video formats to panoramic video and, more recently, to volumetric video. The concept of volumetric video, inspired by holograms and 3D virtual environments often depicted in science fiction, reflects a growing aspiration to replicate reality with unprecedented detail—surpassing the limitations of flat screens. A general framework for a volumetric video service is illustrated in Figure 1.1. From initial capture to final display, increasing the Quality of Experience (QoE) for the end user is our focus in this volumetric processing pipeline.

Various formats have been developed to represent 3D media, including meshes, voxels, and point clouds. Each format has its own strengths and limitations depending on the target requirements. Among them, point clouds serve as a promising representation of 3D scenes and objects. Unlike meshes or voxels, point clouds provide a direct and flexible means of capturing spatial information together with associated attributes. A point cloud is typically defined as a collection of points in 3D space, where each point can be represented by a basic six-dimensional vector (x, y, z, r, g, b) that encodes both geometric position and color information. Owing to their lightweight structure and descriptive capacity, point clouds have become fundamental in a wide range of applications, including computer-aided design, Virtual and Augmented Reality (VR/AR), autonomous driving, and remote communication. These applications demand high-fidelity 3D representations; however, distortions may be introduced at various stages of the pipeline—from raw data acquisition to rendering—across the server, network, or user end. Therefore, accurate assessment of the perceptual quality of point clouds is essential to optimize the QoE for end users. Precise and efficient objective quality metrics can be leveraged to optimize reconstruction, compression, and rendering algorithms, which in turn lead to improved overall system performance. In parallel, understanding where users focus their attention during interaction when consuming the point cloud—i.e., visual saliency in point clouds—can improve the prediction accuracy of the objective quality metrics, by taking into account the behaviors of the user.

The proliferation of 3D sensing technologies—from LiDAR to multi-view cameras—

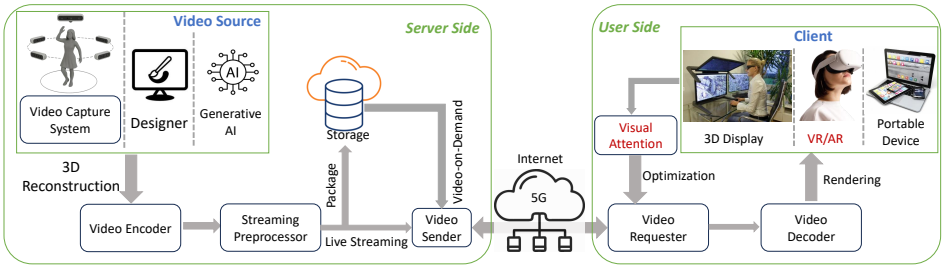


Figure 1.1: A general framework for volumetric video service.

has also made point clouds a popular format of modern immersive media systems [1]. As collections of unstructured 3D points, point clouds enable photorealistic representations of complex scenes. However, their irregular structure and massive data volumes pose significant challenges for processing, compression, and quality evaluation. Unlike 2D images or videos, where quality assessment is well-studied, point clouds introduce unique complexities:

1. **Complex Distortions**: The distortions, particularly during acquisition, compression, and transmission of point clouds, affect both geometry and texture simultaneously, and the characteristics of the distortions vary significantly across different compression algorithms, making it difficult to develop high-generalization quality assessment models. Compression artifacts, e.g., from MPEG ' s Video-based Point Cloud Compression (V-PCC) and Geometry-based Point Cloud Compression (G-PCC), or synthetic noise (simulating the acquisition/transmission noise) degrade both geometry and texture, often non-uniformly across the point cloud. The point clouds inspected in this thesis are highly realistic and dense, ranging from nearly 1 to 4 million points per frame. This sheer scale introduces another level of difficulty beyond the distortion measurement algorithmic design itself, as the time complexity of assessing distortions and their impact on quality can become prohibitive without careful computational considerations.

2. **Human Perception In 6 Degrees of Freedom (6DoF)**: In immersive VR/AR, users freely explore 3D content, making traditional image-based objective quality metrics inadequate for capturing viewpoint-dependent perceptual quality. Especially, consuming the point cloud with the Head-Mounted-Display (HMD) is a more natural way than with 2D screen-based devices, however, there is no standard methodology for collecting the ground-truth data, such as quality score and salient area. The most recent guidelines for conducting the subjective quality assessment experiment from VQEG, International Telecommunication Union (ITU) Recommendation, have only been updated to VR images so far, not to mention the protocol in VR/VR with 6DoF or eye tracking. The human behaviour and eye-tracking knowledge need to be updated accordingly. Objective Point Cloud Quality Assessment (PCQA) metrics in 6DoF need more investigation, because from a systematic point of view, the performance depends on how humans observe the point

clouds in 6DoF with HMD.

3. **Changing Structure Over Time**: Dynamic point clouds (e.g., sequences of point cloud frames) require temporal consistency in quality evaluation, compounded by varying point densities and motions. Different frames have different points, and the reference and distorted point cloud frames also have different points, which makes it more difficult to find the correspondence between the two different frames or the motion vector during the consecutive frames.

As the ultimate consumers of point clouds are humans, the Human Visual System (HVS) becomes crucial for perceptually driven quality assessment [2]. While HVS-based models are well-studied for traditional images and videos, capturing its mechanisms for point cloud processing remains a significant challenge. Processing and parameterizing these flexible 3D point cloud representations is generally non-trivial, and learning from these irregular data may lead to less effective modeling of hierarchical structures and sensitivity to noise, which, as a consequence, causes hard and costly optimization algorithmically [3]. On the other side, saliency-guided metrics have improved 2D Image/Video Quality Assessment (IQA/VQA), their extension to 3D point clouds remains nascent. Existing methods often rely on simplified assumptions (e.g., central bias) or neglect task-dependent viewing behaviors in interactive environments. This gap limits the development of perceptually optimized pipelines for point cloud compression, rendering, and quality evaluation.

Given these challenges, objective PCQA is crucial for ensuring high QoE for end-users [4]. Existing PCQA metrics can be broadly categorized into formula-based and learning-based methods. Formula-based metrics, commonly used in full-reference PCQA tasks, are comparatively easy to measure the similarity but frequently fail to align with human perception, especially concerning complex compression artifacts. For instance, PCQM [5] and GraphSIM [6] assess similarity between reference and distorted point clouds through signal processing techniques. While their hand-crafted features effectively capture structural distortions holistically, processes like curvature and color feature calculations in PCQM, and graph construction with color gradient computations in GraphSIM, can be computationally intensive. In contrast, learning-based methods leverage Deep Neural Networks (DNNs) to predict quality scores. MM-PCQA [7] employs a self-attention mechanism to extract quality-aware features by integrating 3D geometry and 2D texture information. Similarly, CLIP-PCQA [8] predicts quality based on discrete quality descriptions and score distributions, leveraging the CLIP model's philosophy. While these methods have shown promise, particularly with multi-modal feature extraction, they often face challenges such as high computational complexity, limited interpretability, and lack of intermediate outputs like local region saliency or the relative impact of geometric versus attribute distortions. These limitations hinder their effectiveness in optimizing compression algorithms and designing efficient coding strategies. To address these gaps, this thesis aims to develop a solution that combines theoretical rigor with practical applicability. For example, extending similarity measurement principles from 2D images and videos to point clouds is non-trivial: one must embed both distorted and reference point clouds into a shared space, account for the spatial impact of distortions, and capture the interplay between geometry and texture. By focusing on these aspects, our framework not only improves perceptual alignment with human vision but also provides actionable

insights for compression and rendering optimization.

HVS modeling, particularly visual saliency, plays a vital role in estimating perceived media quality [9]. Saliency-based models identify perceptually important regions, thus improving the relevance of quality prediction. Lin *et al.* [10] provide a comprehensive overview of saliency and quality models for point clouds and meshes, emphasizing user-centric and methodological perspectives. For example, Laazoufi *et al.* [11] propose a no-reference PCQA model that filters for salient points and constructs statistical features based on them. However, transferring saliency maps from a reference point cloud to its distorted version remains a major challenge. Furthermore, point cloud quality prediction is inherently more complex than its 2D counterparts because it must simultaneously consider geometric structure and visual appearance. Feature extractors often fail to capture nuanced information that significantly impacts perceived quality. To address these limitations, we incorporate visual saliency into the point cloud analysis, enabling the refinement of perceptually relevant features. By guiding feature extraction with human attention cues, our method better aligns with subjective evaluations and enhances the model' s ability to perform fine-grained quality ranking.

The primary research problem addressed in this thesis is the lack of robust, interpretable, and perceptually aligned objective metrics for evaluating point cloud quality, particularly under compression-induced distortions. Current methods either deviate from human perception or are too computationally intensive for real-world deployment. Therefore, the proposed metrics in this thesis aim to bridge the gap between mathematically derived metrics and human visual perception, especially within the context of HMD-based consumption in 6DoF environments—ultimately improving QoE in eXtended Reality (XR) applications.

From the perspective of subjective PCQA and the understanding of point cloud in XR, while prior studies have explored static saliency detection[12–14], there is currently no standard protocol for jointly evaluating perceptual quality and visual saliency in VR scenarios with 6DoF. Moreover, existing research has not addressed saliency modeling for dynamic point clouds, nor has it explored saliency-guided PCQA for temporal sequences. Beyond the development of objective metrics, this thesis also aims to conduct subjective quality assessment studies and visual saliency detection experiments within VR environments using HMDs. These efforts seek to establish principled methodologies and datasets that can serve as a foundation for future research and practical applications in immersive media.

The main objectives of this thesis are fourfold:

1. To develop an accurate and interpretable PCQA framework that generalizes well across various distortions and captures the relationship between geometry and texture.

2. To understand the role of visual saliency in point clouds, particularly in dynamic contexts where temporal information plays a critical role.

3. To integrate saliency models into PCQA metrics, improving their perceptual alignment and informativeness.

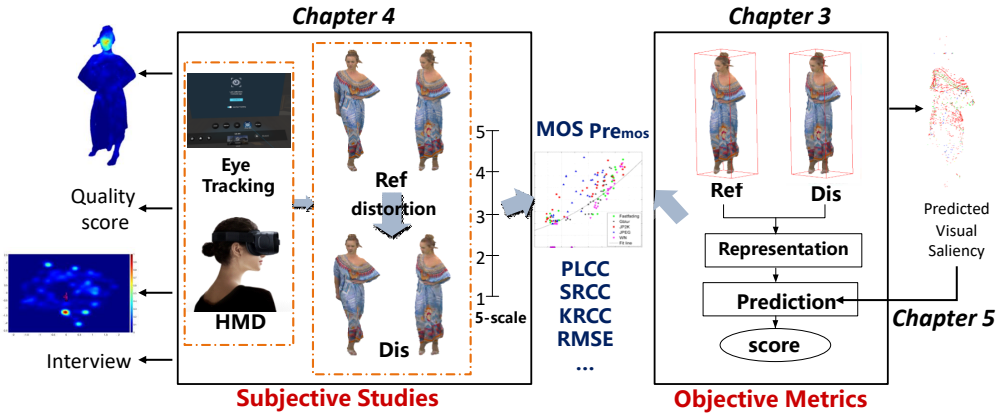


Figure 1.2: High-level illustration of PCQA explored in Chapters 3-5

4. To rigorously evaluate the proposed metrics through extensive subjective and objective experiments.

To fulfill these goals, this work adopts a hybrid methodology combining signal processing, machine learning, and user-centric evaluation. Novel metrics are proposed and validated using diverse benchmark datasets. Furthermore, this work includes the construction of new visual saliency datasets in VR with 6DoF under two distinct interaction tasks, enabling a detailed qualitative and quantitative analysis of saliency behavior over time. These contributions aim to provide valuable insights to the immersive media community and to support the development of practical, perceptually meaningful tools for assessing 3D point cloud quality.

The structure of this thesis is as follows: Chapter 1 introduces the background and motivation, outlining basic concepts around PCQA and visual saliency. Chapter 2 reviews the relevant literature. Chapter 3 presents the proposed objective PCQA metrics, highlighting the contributions of both geometric and texture-based components. Chapter 4 details the subjective experimental design used to construct visual saliency datasets for dynamic point clouds under two distinct VR task scenarios. Chapter 5 demonstrates the integration of saliency information into PCQA models for static and dynamic point clouds. An overview of the relationships among Chapters 3–5 is illustrated in Figure 1.2. Figure 1.2 provides a high-level overview of the human-centred research framework adopted in this thesis. The framework starts with subjective quality assessment of point clouds in immersive VR environments, where user studies are conducted to collect MOS as perceptual ground truth. Based on these subjective scores, objective quality metrics are then developed with the goal of achieving strong correlation with human perceptual judgments. In parallel, visual saliency is investigated as an additional perceptual factor that influences quality perception. Saliency information is obtained by analysing human behavioural responses, in particular eye-tracking data collected during immersive viewing of point cloud content. Finally, the derived saliency models are incorporated into objective PCQA frameworks to further improve their predictive performance by emphasizing

perceptually relevant regions. Finally, Chapter 6 concludes the thesis with discussions on future research directions.

1.2. Research Problems

Figure 1.1 illustrates the complete pipeline of volumetric video streaming. This thesis focuses primarily on the user side of the pipeline, highlighted in red on the right side of Figure 1.1. Specifically, we evaluate the perceptual quality of the point cloud in VR environment, refined with visual attention, encompassing two key components. First, we assess the QoE of end-users based on the perceptual quality of point clouds, utilizing both subjective and objective methods, while excluding other quality of service factors such as bandwidth, transmission, packet loss, and delay [15, 16]. Second, we investigate the detection and prediction of user visual attention–identifying where users focus when interacting with point cloud content. Accurate prediction of visual attention can inform the optimization of compression, rendering, and quality assessment algorithms. An overview of the thesis is shown in Figure 1.3. This thesis aims to answer the following key research problem:

How can the perceptual quality of 3D point clouds be accurately evaluated both subjectively and objectively?

1.2.1. Objective Quality Metrics of Point Cloud

As a starting point, we focus on objective quality assessment of point clouds, given that subjective quality studies are both time-consuming and costly. While subjective assessments provide the ground truth for validating objective metrics, they are not scalable for practical applications. Therefore, developing accurate and perceptually aligned objective metrics becomes essential.

In this thesis, we begin by adopting the full-reference paradigm, which evaluates the perceptual quality by measuring the similarity between a reference point cloud and its distorted version. This approach considers the intrinsic characteristics of point clouds —primarily geometry and texture. Building on this, we further investigate how each of these attributes individually contributes to the final perceived quality in a learning based manner, as judged by human visual perception.

This leads us to our first major research question:

R1: How to measure the perceptual quality of static point clouds under various distortion types?

To systematically address this problem, we decompose it into two sub-questions, each targeting a specific aspect of the evaluation process:

1. *How to measure the perceptual quality of the point cloud with the aid of the reference point cloud?*

2. *How to measure the individual contribution of the intrinsic point cloud attributes based on human visual perception?*

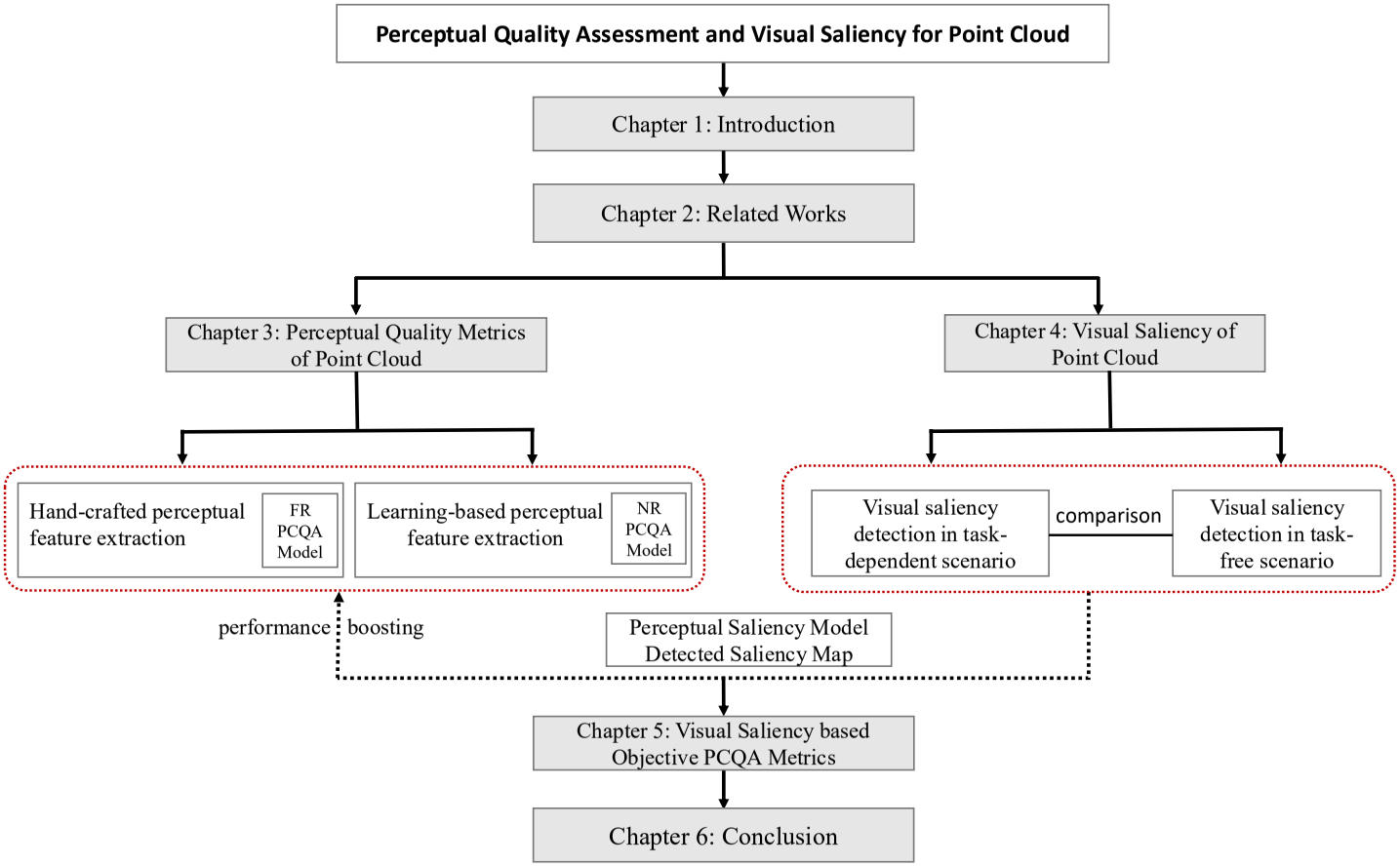


Figure 1.3: Overview of this thesis

Previous work on full reference PCQA metrics can be broadly categorized into projection-based and point-based metrics, depending on the domain in which the evaluation algorithm is applied. Given access to a reference point cloud, full-reference metrics typically emphasize feature extraction and similarity measurement design. Over time, the field has evolved from working with early, colorless, computer-generated point clouds to more complex, dense, and colorful point clouds captured via real-world sensors. This evolution has led to the inclusion of both geometric and textural features in quality metrics, providing more comprehensive and perceptually aligned assessments. In fact, combining geometry and texture has been shown to yield higher correlation with Mean Opinion Scores (MOS) from subjective studies. A common approach for comparing point clouds is to establish correspondences via k-nearest neighbors (knn) or range search methods. However, these neighborhood-based techniques can be inaccurate, especially when distortions alter local structures. To mitigate this, we propose a method that first aligns the distorted point cloud to the reference using Principal Component Analysis (PCA). We then extract features and perform comparisons within the reference point cloud' s coordinate space. This alignment ensures more reliable and meaningful feature correspondence, leading to improved quality estimation.

Furthermore, while prior studies have demonstrated that integrating both geometric and textural attributes improves the accuracy of quality prediction, it is important to consider the actual viewing context of the end user. In immersive environments, users typically experience point clouds through HMD in 6DoF. which presents a perceptual experience that differs significantly from traditional 2D screen-based viewing. In these settings, users perceive not only the 2D projections of 3D objects but also the depth and spatial presence enabled by the native 3D point cloud modality. The HVS, being a highly non-linear and context-sensitive system, processes these visual stimuli in a complex manner. Therefore, how the combination of different dimensional modalities (2D & 3D) and intrinsic attributes (geometry & texture) collectively influence perceived quality remains an open question. In this thesis, we further investigate this interaction, analyzing how these factors contribute to the final perceptual score under various distortion types, thereby enhancing the understanding of quality perception in immersive point cloud experiences.

1.2.2. Visual saliency detection of dynamic point clouds

Visual saliency has been widely used in IQA/VQA metrics [17, 18] to further enhance the performance of objective quality metrics. This is based on the premise that not all distortions contribute equally to perceptual quality [2]. We aim to investigate whether this assumption remains valid for point clouds. However, the study of visual saliency in point clouds remains limited, with minimal research dedicated to comparing visual saliency maps for point cloud quality assessment. Furthermore, 3D point clouds are inherently more suited for display in immersive environments, such as XR. Despite this, there is currently no dedicated visual saliency dataset for dynamic point clouds. To address this gap, we construct a visual saliency dataset using eye-tracking technology. This leads to our second research question:

R2: How can visual saliency in dynamic point clouds be detected and compared in immersive environments?

This research question can be further divided into two sub-research problems

1. *How to accurately detect the visual saliency of dynamic point clouds in VR?*
2. *How do visual saliency maps for dynamic point clouds vary under different task conditions?*

To effectively capture salient regions, it is essential to incorporate insights from prior research on visual saliency in images and videos. The experimental design should be carefully structured within a laboratory setting, to ensure accurate visual saliency acquisition and enable a fair comparison between the two experimental settings. While VR environments offer the advantage of enabling users to freely and immersively observe point clouds, they also introduce challenges such as HMD slippage. This can increase experimental instability, introduce additional errors, and pose a risk of data inaccuracy, particularly in eye-tracking data due to misalignment or tracking drift caused by the movement of the headset. In this thesis, we construct two datasets with two subjective experiments, one involving a quality assessment task in VR, and the other conducted in a task-free setting. In the quality assessment experiment, we implement a meticulous error-profiling process to mitigate the impact of HMD-induced tracking inaccuracies. Beyond the subjective comparison of visual saliency maps, an objective similarity metric is required to quantitatively evaluate the consistency of visual saliency across different experimental conditions.

1.2.3. Extending PCQA metrics with visual saliency

Visual saliency has various applications, and its integration into quality assessment metrics is a key objective of this research. Beyond quality assessment, the visual saliency of point clouds can be leveraged for applications such as point cloud streaming, transmission, and rendering. Previous studies have demonstrated that incorporating visual saliency enhances the performance of objective IQA/VQA metrics [19]. We aim to investigate whether this finding extends point cloud, leading to our third research question:

R3: What is the added value of incorporating visual saliency for PCQA metrics?

This research question can be further divided into two sub-research problems:

1. *How to incorporate the visual saliency map on PCQA metrics for dynamic point clouds?*
2. *How to improve the performance of PCQA metrics by 2D visual saliency map?*

We hypothesize that integrating saliency information will enhance PCQA metrics, with the key challenge being how to effectively incorporate this information. For full reference metrics, several considerations must be addressed: Is it necessary to use visual saliency maps from both the reference and distorted point clouds? Furthermore, how should the reference saliency map be applied to the distorted point cloud, given the varying number of points between them? In this thesis, we utilize both detected visual saliency maps of the reference and distorted dynamic point clouds, and then we make use of predicted 2D visual saliency to enhance the accuracy and reliability of PCQA metrics for static point clouds.

1

1.3. Contributions of this thesis

We now present an overview of the main contributions of this thesis, structured by chapter, including our experimental methodology, algorithmic developments, implementations, and corresponding results.

For objective PCQA metrics, we propose a novel paradigm that measures feature similarity in a shared space by jointly capturing geometric and textural differences. In terms of subjective quality assessment and visual saliency detection, we design a novel experimental procedure inspired by ITU standards. This procedure allows us to simultaneously capture MOSs and visual saliency information for dynamic point clouds, thereby enabling an investigation into the relationship between perceived quality and attention allocation. We conduct a follow-up experiment under different task conditions, analyzing how task demands affect the distribution of visual saliency. Finally, we extend existing objective PCQA metrics by incorporating both ground-truth and predicted saliency maps—obtained via learning-based models—to demonstrate the potential of saliency integration in enhancing objective quality predictions.

1.3.1. Objective PCQA metrics

In Chapter 3, we introduce and analyze two PCQA metrics: PointPCA+ and M3-Unity, each addressing distinct aspects of quality evaluation. PointPCA+ is a full reference metric that relies on hand-crafted features and emphasizes the importance of a common space for error-based distortion measurement. While PointPCA+ demonstrates the significance of precise distortion analysis, its high computational complexity underscores the need for more efficient PCQA metrics in real-world applications, where computational resources are often limited. PointPCA+ also outperforms the majority of existing PCQA metrics, reaching second place in Track#1 and third place in Track#3 and Track#5 of the ICIP 2023 PCVQA grand challenge [20].

On the other hand, M3-Unity is a reference-free metric that leverages learning-based features to assess point cloud quality. Beyond its strong overall performance across evaluated datasets, M3-Unity provides valuable insights into the interplay between geometry and texture in perceptual quality assessment. This analysis is particularly beneficial for downstream tasks such as point cloud compression, as it enables the prioritization of geometry or texture based on the specific requirements of the compression method. By understanding how these attributes influence perceived quality, we can optimize compression strategies to achieve better visual fidelity while maintaining efficiency.

Together, these metrics contribute to the advancement of PCQA by addressing the need for accurate distortion measurement and emphasizing the practical challenges of computational efficiency. They also provide a deeper understanding of the factors influencing perceptual quality, paving the way for more effective point cloud processing and compression techniques.

These contributions are presented in **Chapter 3** and are based on:

1. **Xuemei Zhou**, Evangelos Alexiou, Irene Viola and Pablo Cesar. 2025. PointPCA+: A full-reference Point Cloud Quality Assessment metric with PCA-based features. Signal Processing: Image Communication. [21]

2. **Xuemei Zhou**, Irene Viola, Yunlu Chen, Jiahuan Pei and Pablo Cesar. 2024. Deciphering Perceptual Quality in Colored Point Cloud: Prioritizing Geometry or Texture Distortion? Proceedings of the 32nd ACM International Conference on Multimedia (ACM MM). [22]

1.3.2. Visual Saliency of Point Clouds

In Chapter 4, we propose a comprehensive and meticulously designed experimental protocol for conducting subjective quality assessments of dynamic point clouds in VR environments, integrated with eye-tracking technology. This protocol not only ensures rigorous data collection but also results in the creation and public release of a novel dataset comprising eye-tracking data and corresponding quality scores for dynamic point clouds. This dataset serves as a foundational resource for future research on visual saliency in dynamic point clouds, enabling cross-validation with other datasets and providing valuable insights for standardization organizations.

Additionally, we conduct an initial exploration of visual saliency maps under different task conditions, offering a deeper understanding of how saliency varies across contexts. To quantify these variations, we introduce a similarity metric designed to measure differences between visual saliency maps of dynamic point clouds, taking into account multiple factors such as spatial correlation. Through both qualitative and quantitative analyses, we provide a clearer understanding of how visual saliency shifts under diverse tasks and evolves with the motion inherent in dynamic sequences.

These contributions establish a robust framework for studying visual saliency in dynamic point clouds, offering tools and insights that advance the field and support future investigations into human perception of 3D visual content. A Compressed point cloud dataset with eye-tracking and quality assessment in mixed reality that includes both eye-tracking data and quality rating scores for dynamic point clouds is constructed using a similar experimental protocol [23].
These contributions are presented in **Chapter 4** and are based on:

1. **Xuemei Zhou**, Irene Viola, Evangelos Alexiou, Jansen, Jack, and Pablo Cesar. 2023. QAVA-DPC: Eye-Tracking Based Quality Assessment and Visual Attention Dataset for Dynamic Point Cloud in 6 DoF. 2023 IEEE International Symposium on Mixed and Augmented Reality (IEEE ISMAR). [24]

2. **Xuemei Zhou**, Irene Viola, Silvia Rossi and Pablo Cesar, 2025. Comparison of Visual Saliency for Dynamic Point Cloud: Task-free vs. Task-dependent. IEEE Transactions on Visualization and Computer Graphics (IEEE TVCG). [25]

1.3.3. Visual Saliency-based Objective PCQA metrics

While visual saliency detection for dynamic point clouds has been addressed in the previous chapter, and numerous image saliency prediction algorithms exist, the integration of saliency information into PCQA remains an open question. There already exist numerous algorithms for image saliency prediction. Our contribution in this chapter is to investigate whether incorporating visual saliency maps enhances the performance of dynamic point cloud quality assessment. If a performance gain is observed, it would indicate that visual

saliency is indeed beneficial for predicting the perceptual quality of dynamic point clouds. Conversely, if no improvement is found, alternative approaches for obtaining or utilizing saliency information may be necessary. Our experimental results reveal that the effectiveness of saliency information is highly dependent on the choice of pooling methods and PCQA metrics, highlighting the need for careful consideration in its application.

These contributions advance the understanding of visual saliency's role in PCQA and provide practical methodologies for incorporating saliency information into quality assessment frameworks, paving the way for more accurate and human-aligned evaluation of point clouds.

These contributions are presented in **Chapter 5** and are based on:

1. **Xuemei Zhou**, Irene Viola, Ruihong Yin, and Pablo Cesar. 2024. Visual-Saliency Guided Multi-modal Learning for No Reference Point Cloud Quality Assessment. Proceedings of the 3rd Workshop on Quality of Experience in Visual Multimedia Applications (ACM MM Workshop). [26]

2. **Xuemei Zhou**, Irene Viola, Evangelos Alexiou, Jansen, Jack, and Pablo Cesar. 2025. Subjective and Objective Quality Assessment for Dynamic Point Cloud with Visual Attention in 6 DoF. Transactions on Multimedia Computing Communications and Applications (ACM TOMM). [27]

2

Related Works and Background

This Chapter presents the related work reviewed in this thesis. Since the thesis covers three main areas, we organize the related work accordingly: objective metrics for media content, subjective studies, and visual saliency in media content. Within each category, the discussion is further divided into specific subtopics. Additionally, we include a review of the datasets commonly used to evaluate objective PCQA metrics, as well as the standard evaluation criteria for both PCQA metrics and visual saliency maps. By outlining the current state of the field, highlighting state-of-the-art approaches, and identifying emerging trends, this chapter provides the necessary background to support the methodologies proposed in Chapters 3 to 5.

2.1. Objective Quality Assessment Studies

Objective quality assessment is a computational approach to evaluating the perceptual quality of media content, such as images, videos, and 3D media content. Unlike subjective assessments that rely on human observers, objective quality assessment employs mathematical models to estimate quality in a consistent and repeatable manner. These models are designed to be both accurate and efficient, making them suitable for real-world applications like streaming, compression, and immersive media delivery. Traditional metrics include Peak Signal-to-Noise Ratio (PSNR), which measures the pixel-wise difference between original and distorted content, and Structural Similarity Index Measure (SSIM) [28], which assesses perceived structural changes by considering luminance, contrast, and structural information. For video content, the Video Multimethod Assessment Fusion (VMAF) metric [29], developed by Netflix, combines multiple features—such as Visual Information Fidelity (VIF) and Detail Loss Metric (DLM)—using machine learning to better align with human visual perception. The design of these metrics often integrates principles from signal processing, human visual system modeling, and perceptual psychology to ensure that the assessments correlate well with human judgments.

In the context of emerging 3D media formats, such as mesh and point clouds, specialized objective quality metrics have been developed to address the unique challenges posed by these data types. The following subsections delve into the objective quality assessment methodologies specific to static and dynamic point clouds, respectively.

2.1.1. Objective Quality Assessment of Static Point Cloud

Objective PCQA metrics can be divided into point-based, projection-based, and feature-based models based on the way to process the point clouds. In addition, learning-based models are reviewed, where feature extraction and fusion are performed using machine learning techniques. Objective PCQA algorithms automatically evaluate the visual quality of point clouds as human judgments, it can be classified as Full-Reference (FR), Reduced-Reference (RR) and No-Reference (NR) based on the availability of reference information. In the following, we list PCQA metrics according to the first categorization approach.

Point-based PCQA Point-based metrics such as point-to-point, point-to-plane, and their variants [30–33] measure degradations between the original and distorted point clouds per point, mainly based on Euclidean or color space distances [34]. Alexiou *et al.* [35] propose the angular similarity of tangent planes among corresponding points, which considers neighborhood information. These metrics are computationally efficient but suffer from a crude correspondence of matching between points.

Projection-based PCQA Projection-based approaches adapt existing IQA metrics to PCQA, which is also called image-based. In the Projection-based approaches, firstly used in [36] for point clouds, the rendered models are mapped onto planar surfaces, on which conventional IQA metrics are applied to provide a quality score. The prediction accuracy of IQA metrics on 2D views of point clouds is initially examined in [37]. Yet, enabling a

large number of views in denser camera arrangements may lead to redundancies and extra computational costs, without guaranteeing performance improvements, as indicated in [38]. Excluding pixels from the views that don't belong to the effective part of the displayed model (i.e., background information), was found to improve the accuracy of the predicted quality in [38]. Moreover, the estimation of the global degradation score by incorporating importance based on the time of inspection of human subjects is proposed in the same study and is found to increase the prediction performance. Yang *et al.* [39] introduce a metric based on a weighted combination of global and local features, which are extracted from 6 orthographic texture and depth images. Wu *et al.* [40] obtain the same projections, and weight the contributions of each face based on the ratio of the size of a plane to the sum of the area of six planes on the bounding box. Moreover, they propose patch projection of both geometry and texture data, with patches obtained based on the direction of the normals, ensuring identical contours. A final score is obtained as a weighted average between quality scores from geometry and texture information. In [41], point clouds undergo translations, rotations and scaling before projection. They suggested that the principle of information content-weighted pooling provides a good framework and proposed to use of IW-SSIM on the projected views. The mentioned methods are mainly based on the projected texture information, without considering the geometry structure of the point cloud. Liu *et al.* [42] provide a PCQA model based on the principle of information content weighted structural similarity (IW-SSIM) [43]. Icosphere and a series of transformations are employed to generate viewpoints. PQA-Net [44] takes 6 orthographic projections of point clouds as inputs, features are extracted after convolution neural network blocks, and they share a distortion identification and a quality prediction module that assist in obtaining final quality scores. PQA-Net adopts a two-step strategy to train the multi-task neural network. However, the projection process and the number of viewpoints have a non-negligible impact on the final prediction accuracy; besides, how to combine the quality score on each viewpoint into a score is also not straightforward.

Feature-based PCQA Feature-based metrics consider perceptual loss from both geometry and texture properties. Viola *et al.* [45] extract the color statistics, histogram, and correlogram to assess the level of impairment and combine the color-based metrics with geometry-based metrics to form a global quality score. Alexiou *et al.* [46] employ the local distributions of point clouds to predict perceptual degradations from topology and color. Yang *et al.* [6] construct a local graph centered at resampled key points for both reference and distorted point clouds, with the color gradient on the local graph being used to measure distortions. Diniz *et al.* [47] introduced a texture descriptor based on perceptual color distance patterns, which is scale and rotation-invariant [48]. Meynet *et al.* [5] utilize an optimally-weighted linear combination of curvature-based and color-based features to evaluate visual quality. Diniz *et al.* [49] adopt the statistical information of the extracted geometry/color features and feed them into a regression model.

Learning-based PCQA Learning-based modules have also been used to extract perceptual features, machine learning techniques often used in the quality regression phase, and deep learning-based NR PCQA have gained more interest since they align with real-life applications. Zhang *et al.* [50] utilize natural scene statistics and entropy on the

quality-related geometry and color feature domains, which are projected from 3D space, and support vector regression is employed to obtain quality scores. PKT-PCQA [51] adopts a progressive knowledge transfer to convert the coarse-grained quality classification knowledge to the fine-grained quality prediction task. The key clusters are extracted based on global and local information, an attention mechanism is incorporated into the network design. Structure Guided Resampling [52] considers that HVS is highly sensitive to structure information, it first exploits the unique normal vectors of point clouds to execute regional pre-processing. Both the cognitive peculiarities of the human brain and naturalness regularity are involved in the designed quality-aware features. MM-PCQA [7] partitions point clouds into sub-models for local geometry representation and renders them into 2D projections for texture. A symmetric cross-modal attention module is used for integrating quality-aware information. IT-PCQA [53] utilizes the rich prior knowledge in images and builds a bridge between 2D and 3D perception in the field of quality assessment. A hierarchical feature encoder and a conditional discriminative network is proposed to extract latent features and minimize the domain discrepancy. GPA-net [54] proposes a graph convolution kernel, i.e., GPAConv, which attentively captures the perturbation of structure and texture, within a multi-task framework. A coordinate normalization module is utilized to stabilize the results of GPAConv under shift, scale and rotation transformations. pmBQA [55] proposes a projection-based blind quality indicator via multimodal learning by using 4 homogeneous modalities (i.e., texture, normal, depth and roughness). As a lightweight alternative, Green Learning (GL) has emerged as a promising paradigm, offering mathematically transparent and computationally efficient solutions. GL models have shown success in blind IQA/VQA tasks [56, 57], balancing interpretability with predictive accuracy. However, their simplified architectures struggle to capture the complex interplay of spatial and attribute distortions in 3D point clouds, leading to suboptimal performance in high-precision PCQA tasks.

Interested readers may refer to [34] for a more comprehensive review of the literature. In summary, point-based schemes may neglect the high-dimensional properties of point clouds and the interplay among these dimensions, thereby limiting their effectiveness. Projection-based methods often rely on 2D IQA, which may not adequately capture the intrinsic characteristics of point clouds. Feature-based schemes tend to have a high level of complexity, while the interpretability of deep learning-based methods is a drawback with training requiring a huge amount of data.

2.1.2. Objective Quality Assessment of Dynamic Point Clouds

Objective quality assessment of 2D/3D video has achieved remarkable progress in recent years. However, few specific objective quality metrics have been designed for dynamic point clouds so far. In the following, we summarize the metrics designed for dynamic point clouds, with a focus on temporal pooling strategies, the fusion of multiple quality indicators, and the incorporation of visual saliency.

Ak *et al.* [58] explore the possibility of temporal sub-sampling of the content under evaluation for objective quality evaluation without sacrificing the correlation with the subjective opinion. 30 different objective quality metrics are tested on the Vsense VVDB2 dataset, combined with temporal sub-sampling and temporal pooling methods. Results show that the performance of objective metrics for point cloud compression is minimally

affected by the temporal sub-sampling rate. Freitas *et al.* [59] investigate the added value of incorporating temporal pooling into dynamic point clouds quality assessment model using metrics designed for static point clouds. They find that the performance of temporal pooling is consistently better when a temporal variation model [60] is used. The same authors investigate the suitability of geometric-aware texture descriptors to blindly assess the quality of colored dynamic point clouds [61] on top of the same temporal pooling strategy, leading to similar conclusions.

Yang *et al.* [62] conduct a subjective user study to understand the effectiveness of different perceptual quality metrics for volumetric video, and design an objective metric called Volu-FMAF to better evaluate user perceptual quality. Volu-FMAF combined point-based and pixel-based metrics with viewpoint-related features. They further design a distortion-aware rendered image super-resolution network in a volumetric video streaming framework, which exploits the insight obtained in the user study. Damme *et al.* [63] present a thorough correlation analysis of both FR and NR objective metrics to subjective MOS with a double purpose. Additionally, they investigate how region of interest selection and weighting procedures impact accuracy to enhance it further. The study shows that the classical video quality metric VMAF is very well-suited as an objective benchmark for volumetric media streaming in terms of correlation to subjective scores, and a combination of NR features could provide a good real-time assessment. Fan *et al.* [64] propose a deep-learning-based NR volumetric video quality assessment method based on multi-view learning. They first project volumetric videos to 2D video sequences from various viewpoints. ResNet 3D is utilized to extract quality-aware features, and a quality regression module is designed to fuse the features learned from the multiple viewpoints and jointly regress them into quality scores. Marvie *et al.* [65] benchmark and calibrate several objective quality metrics on a challenging volumetric video dataset represented as textured meshes: two model-based approaches (MPEG PCC and PCQM) and one image-based approach (IBSM) for which they introduced two new features that specifically detect holes and temporal defects. For each metric, the optimal selection and combination of features are determined by logistic regression through cross-validation. The performance analysis, combined with MPEG experts' requirements, leads to recommendations on the features of most importance through learned feature weights, such as temporal pooling, integrating an attention model.

Objective quality assessment for dynamic point clouds is still in its early stages, with most existing methods adapting strategies from 2D video quality assessment. Recent studies have explored the impact of temporal pooling and sub-sampling techniques, showing that perceptual accuracy can be maintained with reduced temporal resolution. Other approaches extend static point cloud metrics by incorporating temporal variation models or multi-view projections to better capture temporal dynamics. While a few deep learning-based models have been proposed to predict quality from rendered views, they rely on pixel- and point-level features without fully addressing the unique spatiotemporal characteristics of 3D data. Visual saliency has been introduced in some models using predicted saliency maps; however, due to the lack of gaze-annotated DPC datasets, these models cannot yet leverage human attention directly. Overall, current methods demonstrate promising directions but remain limited by their reliance on 2D paradigms and the lack of perceptually grounded 3D datasets.

2.2. Subjective Quality Assessment Studies

In the development of objective PCQA metrics, subjective quality assessment datasets are the basis for their design and validation. Ground-truth ratings for visual impairments in stimuli, namely Mean Opinion Score (MOS) or Differential MOS (DMOS), are obtained through subjective quality assessment experiments [66, 67]. These scores primarily indicate the degree of degradation affecting the contents with little information about the representations of distortion. The psychological and physiological mechanisms of HVS reveal that humans cannot perceive the signal change below a certain threshold [68]. Consequently, numerous distorted contents below the threshold form an equal-quality space with combinations of various distortion types and levels [69].

Existing synthesized PCQA datasets, generated either by performing a test in a controlled laboratory environment or by mimicking it with point cloud processing algorithms, are size-limited and distortion-type and distortion-level unbalanced, often resulting from the principle when constructing the dataset [70–73]. For example, when using test methods such as Absolute Category Rating (ACR), the distortion level for a certain distortion type must be visually distinguishable to obtain meaningful MOS values and to avoid fatiguing subjects. Regarding the compression distortion of PCQA, recent research indicates a monotonic relationship with bit rate when evaluating compression quality [74, 75]. However, it's crucial to emphasize that subjective quality evaluation sometimes does not follow a monotonic behavior. This is particularly evident in cases where the balance between geometry and color quality in point cloud content is challenging to establish [76]. This complexity adds difficulty to discerning between different point cloud coding modules.

Therefore, it becomes imperative to integrate both geometry and texture distortions in assessing the perceptual quality of point clouds for PCQA. The utilization of effective geometry- and texture-related features can enhance the optimization of various algorithms associated with perception.

2.2.1. Subjective Quality Assessment of Static Point Cloud

Subjective quality assessment are widely regarded as the most reliable method to evaluate the quality of point clouds, the interested reader may refer to [34] for a detailed overview. Recently, many subjective studies have been conducted and reported in the literature to assess the performance of compression distortion in terms of visual quality. Lots of works present the subjective result for compressed point cloud, such as base point cloud compression method from MPEG [77]; octree pruning using the Point Cloud Library and projection-based method implemented in the 3DTK toolkit [78]; V-PCC and G-PCC variants [79, 80]. Later, other distortion types are introduced in the SJTU-PCQA dataset [39] to mimic the acquisition and re-sampling noise besides the compression distortions. Liu *et al.* [81] distorts the source point clouds with 4 processes to simulate real-world application scenarios and enrich the contents beyond those addressed by MPEG and JPEG. Liu *et al.* [82] construct the largest dataset so far with pseudo-quality scores to support neural network training. 31 types of impairments covering a wide range of impairments during point cloud production, compression, transmission, and presentation are included. More recently, learning-based point cloud compression techniques have been considered. AK

et al. [75] include the GeoCNN compression distortion. Lazzarotto *et al.* [83] first analyzes the impact of different configuration parameters on the performance of MPEG and JPEG Pleno compression with the aid of objective metrics.

2.2.2. Subjective Quality Assessment of Dynamic Point Clouds

Whereas subjective quality assessment of static point clouds has been explored in more detail in the literature [38], analogous research on dynamic point clouds is still a sophisticated and challenging problem, owing to numerous factors such as the evaluation methodology, rendering method, display equipment and so forth.

Zerman *et al.* [84] conduct a subjective experiment on two dynamic point clouds (VsenseVVDB) using V-PCC compression [85]. Additionally, they argue that certain geometric distortion metrics are incongruent with the expected quality. Hooft *et al.* investigate how and to what extent various aspects impact the user's QoE, via extensive subjective evaluation of volumetric 6 DoF streaming [86]. Mekuria *et al.* evaluate the subjective quality of the CWI-PCL codec performance in a realistic 3D tele-immersive system in a virtual room scenario, in which users are represented and interact as 3D avatars and/or 3D point clouds [30]. The subjective study shows that introduced prediction distortions are negligible compared with the original reconstructed point clouds. Cao *et al.* [87] study the perceptual quality of compressed 3D sequences, for both point cloud compression and mesh-based compression. They explore the impact of bit rate and observation distance on perceptual quality. Cox *et al.* [88] present VOLVQAD, a volumetric video quality assessment dataset with 376 video sequences. The volumetric video sequences are first encoded with MPEG V-PCC using 4 different avatar models and 16 quality variations, and then rendered into test videos for quality assessment using 2 different background colors and 16 different quality switching patterns with 2D display. However, these experiments are all with a desktop setting. Viola *et al.* [89] compare two different VR viewing conditions enabling 3/6 DoF, along with a desktop setting, to understand how interaction in the virtual space affects the perception of quality. Results show no statistical difference between scores given in a desktop and VR setup; however, qualitative results highlighted the added value of interactive evaluation. One limitation of the study lies in the time duration (5 seconds) of the sequences used for the evaluation, as the authors use 150 frames. Subramanyam *et al.* [74] evaluate the performance of several adaptive streaming solutions in an interactive VR experiment. They compare the performance of V-PCC with respect to CWI-PCL, using various adaptive streaming strategies. Quantitative subjective results and qualitative insights indicate that V-PCC has a more favorable performance than the CWI-PCL, especially at low bit rates. Damme *et al.* [90] conducted an in-depth subjective study on the impact of converting point clouds to meshes with varying-quality representations. Additionally, while end-users demonstrate awareness of quality switches, the effect on their perception remains limited. Gutierrez *et al.* [91] present a subjective study on dynamic point clouds using the absolute category rating methodology and considering different compression rates using the MPEG standard V-PCC. Results on users' exploration behavior show no significant differences when visualizing point clouds with different qualities, no changes in the behavior during the test session, and no correlation between exploration activity and quality assessments. Nguyen *et al.* [23] provide an open-source compressed point cloud dataset with eye tracking and quality assessment in mixed reality

including 4 dynamic point clouds. Opinion scores and eye-tracking data are collected in a user study under different experimental settings. While numerous studies have explored the subjective quality assessment of dynamic point clouds, there is a gap in research focusing on visual attention and quality assessment in VR with 6 DoF.

2.3. Visual Saliency and its application for Media Content

Owing to HVS's selectivity in responding to the most attractive features in the visual field, it's inappropriate to treat each voxel equally [2]. Saliency maps aim to capture these perceptual priorities by highlighting the regions that most influence visual attention or recognition tasks. In the context of 2D images, saliency has been widely studied and used to interpret the influence of individual pixels on classification outcomes [92–94]. This concept has recently been extended to 3D point clouds, where saliency maps help identify the importance of each point in terms of perceptual or semantic relevance [12, 14].

These developments reveal the potential of saliency for a wide range of media applications, including adversarial analysis, perceptual quality assessment, and content-aware processing. To better contextualize its role, we present a review of saliency datasets, saliency-integrated objective quality assessment metrics, and recent efforts to incorporate saliency into PCQA algorithms. Additionally, we discuss how task-specific factors can influence visual saliency patterns. Together, these perspectives offer a comprehensive understanding of how saliency can enhance the analysis and evaluation of immersive media content, with a focus on dynamic point clouds.

2.3.1. Visual Saliency Datasets

In the early stages of visual attention computation, due to the limitations of eye-tracking technologies, different collection procedures for salient points were pursued. For example, Chen *et al.* [95] investigate "Schelling points" on 3D meshes, feature points selected by people in a pure coordination game due to their salience. They designed an online experiment that asked people to select points via mouse-tracking technology on 3D surfaces that they expected would be selected by other people. This dataset is widely used as a benchmark for objective saliency detection algorithms for colorless point cloud/mesh [96, 97]. Later methods employ handcrafted descriptors [97, 98] from lower-level geometric properties to detect the point cloud/mesh saliency, but these approaches lack expressiveness and overlook real human viewing behaviors [99].

More recently, to explore the visual attention of 3D contents, eye-tracking experiments remain the main way to understand human visual behaviors. Sitzmann *et al.* [100] capture and analyze gaze and head orientation data of users exploring stereoscopic, static omnidirectional panoramas, for a total of 1,980 head and gaze trajectories for three different viewing conditions. They found the existence of a particular fixation bias, which can be used to adapt existing saliency predictors to immersive VR conditions. Nguyen *et al.* [101] introduce a large saliency dataset for 360-degree videos with a new methodology supported by psychology studies with HMD. They describe an open-source software

implementing this methodology that can generate saliency maps from any head-tracking data. Lavoué *et al.* [102] present a dataset that records the eye-movement data for rendered 3D shapes. During their experiment, 3D meshes are rendered using different materials and lighting conditions under different scenes, and the rendered videos of 3D meshes are shown on the screen for subjects to observe. Ding *et al.* [103] propose a novel 6DoF mesh saliency dataset that provides both the subject's 6DoF data and eye-movement data, and a 6DoF mesh saliency detection algorithm based on the uniqueness measure and the bias preference is developed. Abid *et al.* [104] compute the visual saliency of the point cloud considering the viewpoint from which the 3D content was seen/rendered, using an offline-computed view-based saliency map. One eye-tracking experiment on a 2D screen is conducted to verify the proposed saliency map. Alexiou *et al.* [14] conduct an eye-tracking experiment in an immersive 3D scene. A method to exploit the high-quality recorded gaze measurements is introduced based on per-session profiling, and a scheme to determine areas of fixations in a static point cloud is proposed. Nguyen *et al.* [23] propose a dataset with compressed dynamic point clouds, rating scores, and eye-tracking data with AR HMD. However, only 4 reference dynamic point clouds have an associated visual saliency map.

To the best of our knowledge, no existing dataset has been publicly released that captures visual attention in dynamic point clouds. Moreover, none of the previous subjective studies have systematically examined how task conditions or variations in XR environments influence the distribution of visual saliency in dynamic point clouds.

2.3.2. Extending IQA/VQA metrics with Visual Saliency

Recent literature in eye-tracking-based visual saliency for immersive contents has mainly focused on task-free experiments to gather visual attention maps [100, 105]; no study has been conducted to link visual attention to visual quality assessment for volumetric videos. The literature suggests that visual attention might be beneficial for understanding the process of perception of visual quality for 2D images/videos; in fact, different metrics for IQA have been extended with a computational model of visual attention [2], but the resulting gain on the metrics' performance is so far unclear. To better understand the added value of including visual attention in the design of objective metrics for 2D images, some works in the literature have taken advantage of recorded visual attention data. Lin *et al.* [106] perform two eye-tracking experiments: one with a free-looking task and one with a quality assessment task. They found a tendency that adding saliency to a metric yields a larger amount of gain in performance. The extent of the performance gain tends to depend on the specific objective metric and the image content. In addition, the gain is small for objective metrics that already show a high correlation with perceived quality for a given distortion type. Zhang *et al.* [107] propose a new methodology to eliminate the inherent bias due to the involvement of stimulus repetition. The refined methodology result in a new eye-tracking dataset with a large degree of stimulus variability. Based on ground-truth labeling, the statistical evaluation shows that the visual attention information of both the referenced and distorted scene is beneficial for IQA metrics, but the latter tends to further boost the effectiveness of integrating attention in IQA metrics. Jin *et al.* [108] utilize an eye-tracker to create foveation-compressed VR datasets and evaluate both the foveated and non-foveated objective IQA/VQA algorithms.

To better understand whether the findings regarding visual saliency and quality assessment on 2D images/videos can hold for volumetric contents, ad-hoc datasets that combine the two aspects are needed. That is one of the research gaps we aim to fill with this thesis.

2.3.3. Extending PCQA metrics with Visual Saliency

Bourbia *et al.* [109] present an NR approach that incorporates the advantage of the transformer encoder architecture and the visual saliency to predict the perceived visual quality of distorted point clouds. They project the point cloud into multi-view and weight each view with its corresponding calculated saliency map through a pointwise multiplication to detect the regions of interest, and then regress the weighted sub-images to a quality score. However, the weighted sub-images are not guaranteed to be the saliency areas correlated to the perceptual quality. RR-CAP [110] makes the first attempt to simplify reference and distorted point clouds into projected saliency maps with a downsampling operation in an RR manner. The objective quality scores of distorted point clouds are produced by combining content-oriented similarity and statistical correlation measurements based on the saliency maps. PQSM [111] introduces a 3D point cloud saliency map generating method, which integrates depth information to enhance geometric representation. Three structural descriptors capturing geometry, color, and saliency discrepancies are used to construct local neighborhoods. A saliency-based pooling strategy refines the descriptors, yielding a comprehensive quality score. Laazouf *et al.* [11] firstly compute a 3D saliency map for each distorted point cloud. Then, a threshold-based filter is used to select the most salient points. Estimates of their statistical properties (Entropy, Standard deviation, Skewness, Kurtosis, Median and Mean) form a feature vector from both geometrical and perceptual attributes. The support vector regressor is utilized to regress the feature vector as a quality score. These three non-learning metrics mainly consider one modality.

2.3.4. Task Impact on Visual Saliency

Understanding how the allocation of human visual attention changes depending on perceptual tasks offers clear benefits in developing techniques and improving the quality of experience in VR/AR. This is a complex behavior that holds great importance for the field of IQA/VQA. Specifically, task-free means that the user observes the content as naturally as possible, with fixation data from such free viewing commonly used to evaluate visual saliency. In contrast, task-dependent means that the user observes the media content to fulfill a specific task; in the case of IQA/VQA, to evaluate the visual quality. In these experiments, the MOSs (typically ranging from 1 to 5) across users serves as the ground truth for quality evaluation.

Meur *et al.* [112] carry out two eye-tracking experiments on 10 original video sequences in a free viewing and a quality assessment task, separately. The comparison between eye movements indicates that the degree of similarity between human priority maps is rather high. They observe that saliency-based distortion pooling does not significantly improve the performances of the VQA metric. Liu *et al.* [106] and Hani *et al.* [113] perform a similar experiment procedure for IQA, Liu evaluates whether and to what extent the addition of natural scene saliency is beneficial to objective quality prediction in general terms, and Hani conclude that it is not fair to compare the effect of adding

saliency in objective metrics without specifying how the saliency was measured.

In larger contexts, task effects more broadly influence visual attention in immersive environments. Hadnett-Hunter *et al.* [114] investigated free-viewing, search, and navigation tasks in interactive virtual environments and found task-specific differences in several human visual attention measures, particularly during navigation. Their findings demonstrated the potential for using attention data to dynamically adapt virtual simulations and games. Hu *et al.* [115] analyzed eye and head movements of participants performing free-viewing, visual search, saliency, and tracking tasks in 360-degree VR videos. They revealed significant task-driven differences in fixation durations, saccade amplitudes, head rotation velocities, and eye-head coordination. EHTask–a learning-based method that employs eye and head movements to recognize user tasks in VR is proposed. Their work provides meaningful insights into human visual attention under different VR tasks and guides future work on recognizing user tasks in VR. Malpica *et al.* [116] systematically examined the impact of free exploration, memory, and visual search tasks on visual behavior in immersive scenes. They reported consistent task-specific differences in eye and head movement patterns, offering practical insights for designing task-oriented immersive applications.

To the best of our knowledge, we are the first to investigate the impact of tasks on human attention deployment in the context of dynamic point clouds, building on insights from video, VR, and immersive media studies.

2.4. Dataset and Evaluation Criteria

This section presents the datasets and evaluation criteria used in this thesis. It provides a detailed description of both static and dynamic PCQA datasets. The evaluation includes performance-based criteria for objective metrics as well as visual saliency similarity measures.

2.4.1. PCQA Dataset for Static Point Cloud

Below, we list widely used datasets for the evaluation of the objective PCQA metrics. Five publicly available datasets were recruited for performance evaluation, namely, M-PCCD, SJTU, WPC, BASICS, and MJ-PCCD.

The M-PCCD [79] consists of 8 point clouds whose geometry and color are encoded using V-PCC and G-PCC variants, resulting in 232 distorted stimuli. Detailed distortion types include Octree-Lifting, Octree-RAHT, TriSoup-Lifting, TriSoup-RAHT and V-PCC. The contents in M-PCCD depict either human figures or objects. The SJTU [39] includes 9 reference point clouds with each point cloud corrupted by seven types of distortions under six levels, generating $378 = 9 \times 7 \times 6$ distorted stimuli. Detailed distortion types include Octree-based compression, Color Noise (CN), Geometric Gaussian Noise (GGN), downsampling, and combinations of the CN, GGN and downsampling. SJTU includes 5 human body models and 4 inanimate objects. The WPC [42] contains 20 reference point clouds with each point cloud degraded under five types of distortions and different levels, leading to $740 = 20 \times 37$ distorted stimuli. Detailed distortion types include Octree-LPCC, TriSoup-SPCC, V-PCC, Gaussian noise and downsampling. WPC dataset only collects objects including snacks, fruits and vegetables, etc. The Broad Quality Assessment of

Table 2.1: Overview of static point cloud quality assessment datasets

Datasets	Contents		Distortion Types		Distortion Levels	Total
M-PCCD	Humans & Inanimate Objects	8	Octree-Lifting, Octree-RAHT, TriSoup-Lifting, TriSoup-RAHT, V-PCC	5	G-PCC: 6 V-PCC: 5	232
SJTU	Humans & Inanimate Objects	9	Octree-based compression, CN, GGN, downsampling, CN + GGN, CN + downsamplin, GGN + downsampling	7	6	378
WPC	Inanimate Objects	20	Octree-LPCC, TriSoup-SPCC, V-PCC, Gaussian noise, downsampling	5	Geometry: 3 Texture:1/3/4	740
BASICS	1. Humans & Animals 2. Inanimate Objects 3. Buildings & Landscapes	75	Octree-RAHT, Octree-Predlift, V-PCC, GeoCNN	4	G-PCC: 5 V-PCC: 6	1494
MJ-PCCD	Humans & Inanimate Objects	6	G-PCC, V-PCC, JPEG Pleno standards	3	4	213

Table 2.2: Publicly available subjective quality assessment and visual attention datasets for point clouds.

Dataset	Type	Degradation	Stimuli	Time	Display	Interaction	Opinion Score	Visual Attention
VsenseVVDB [84]	Dynamic	down-sampling, V-PCC	32	6.6s	2D monitor	✗	✓	✗
VsenseVVDB2 [117]	Dynamic	Mesh: Draco+JPEG Point Clouds: G-PCC, V-PCC	28 136	10s	2D monitor	✗	✓	✗
Owlii [118]	Dynamic	Mesh: TFAN, FFmpeg Point Clouds: V-PCC, FFmpeg	20	20s	2D monitor	✗	✓	✗
Marvie et al. [65]	Dynamic	Position, Texture coordinate HEVC, Triangle holes	176	10s	2D monitor	✗	✓	✗
VOLVQAD [88]	Dynamic	V-PCC	376	10s	2D monitor	✗	✓	✗
Subramanyam et al. [119]	Dynamic	CWI-PCL, V-PCC	72	5s	HMD	✓	✓	✗
ComPEQ-MR [23]	Dynamic	V-PCC, G-PCC	52	10s	AR	✓	✓	✓
ViAtPCVR [14]	Static	Only reference	8	-	HMD	✓	✗	✓

Static Point Clouds (BASICS) [75] is used in the ICIP 2023 PCVQA grand challenge, and comprises 75 point clouds from 3 different semantic categories: (i) Humans & Animals, (ii) Inanimate Objects, and (iii) Buildings & Landscapes. Each point cloud is compressed with 3 compression methods from the MPEG standardization field, i.e., Octree-RAHT, Octree-Predlift and V-PCC; 1 learning-based algorithm, i.e., GeoCNN, at varying compression levels, resulting in 1494 processed point clouds. BASICS dataset is aimed at providing a foundation for research that supports telepresence applications, in terms of compression and quality assessment. MJ-PCCD [83] is created by compressing 6 reference point clouds from the JPEG Pleno test set at 4 different bitrates with the GPCC, VPCC, and JPEG Pleno standards, producing 213 distorted stimuli.

We provide a summary based on the content variety, types and levels of distortion, and the total number of point cloud samples in each dataset, as illustrated in Table 2.1..

2.4.2. PCQA Datasets for Dynamic Point Cloud and Visual Saliency in Point Cloud

Table 2.2 provides a comprehensive overview of publicly available subjective quality assessment and visual attention datasets for point clouds. The summary includes key attributes of each dataset, particularly those that involve dynamic point clouds or contain visual saliency information. The datasets are compared based on the type of content (dynamic or static), types of distortions applied (e.g., compression, down-sampling, or geometric degradation), the number of stimuli, the duration of each sequence, and the display setup used during the study (e.g., 2D monitor, HMD, AR).

Additionally, the table indicates whether the dataset includes user interaction during the study, whether it provides opinion scores for quality evaluation, and whether visual attention data (e.g., eye tracking or saliency maps) is available. This summary helps highlight the differences in experimental design, application scenarios, and data availability across existing datasets, offering valuable insight for future research in point cloud quality and saliency modeling.

2.4.3. Evaluation Criteria

Four commonly used evaluation criteria are used to reflect the relationship between objective scores and subjective scores [120, 121]: (1) Pearson Linear Correlation Coefficient (PLCC), which measures the linearity of prediction; (2) Spearman Rank-order Correlation Coefficient (SRCC), which measures the monotonicity of prediction; (3) Root MSE (RMSE), which measures the error of prediction. (4) Kendall rank-order correlation coefficient (KRCC), which evaluates how well the ranking of predicted scores matches the ranking of actual scores, complementing SRCC.

Higher values of PLCC, SRCC and KRCC indicate better performance in terms of correlation with human opinion, while lower RMSE indicates better consistency. A five-parametric logistic regression is adopted [122].

3

Objective Quality Metrics of Point Cloud

This chapter investigates objective quality metrics for static point clouds. While existing metrics were reviewed in the previous chapter, here we introduce novel contributions for both FR and NR PCQA. Specifically, we propose an FR metric based on PCA-derived statistical features combined with a Random Forest regression model to predict perceptual quality scores. To complement this, we also develop a learning-based NR metric, addressing the practical need for NR solutions in real-world applications where reference data is often unavailable. Furthermore, we conduct an in-depth analysis of the relative contributions of geometric and texture attributes to perceived point cloud quality. These findings not only inform the design of more perceptually aligned similarity measures but also offer valuable guidance for optimizing bit allocation in point cloud compression. Together, the proposed FR and NR frameworks provide versatile tools that can be adapted to different application scenarios. The effort of this chapter is to develop an accurate and efficient mathematical model for evaluating the perceptual quality of point clouds, enabling its integration into various algorithms throughout the point cloud processing pipeline.

This chapter is based on the following publications:

1. ***Xuemei Zhou**, Evangelos Alexiou, Irene Viola and Pablo Cesar. 2025. PointPCA+: A full-reference Point Cloud Quality Assessment metric with PCA-based features. Signal Processing: Image Communication.* [21]

2. ***Xuemei Zhou**, Irene Viola, Yunlu Chen, Jiahuan Pei and Pablo Cesar. 2024. Deciphering Perceptual Quality in Colored Point Cloud: Prioritizing Geometry or Texture Distortion? Proceedings of the 32nd ACM International Conference on Multimedia (ACM MM).* [22]

Point cloud is prevailing among the available 3D imaging formats [123]. There is a demand for effective and efficient objective PCQA metrics, to guide the design, optimization, and parameter tuning of point cloud processing pipelines. PCQA metrics have been extensively utilized in various applications, including visual tasks, restoration [124, 125], compression [126, 127], as well as for quality monitoring in various systems [74, 128–130].

PCQA has three categories, namely FR, RR, and NR, based on the availability of reference point cloud data. Compared with RR and NR, FR PCQA metric requires the fully available reference during the execution, which perceives quality as a comparison of differences between a degraded and a pristine version [10]. The FR paradigm enables comprehensive distortion analysis by leveraging all available information, whereas NR metrics, which do not require reference data, are more applicable in real-world scenarios. These two paradigms are complementary: together, they enrich the broader landscape of objective PCQA and can provide valuable tools for supporting various application needs.

From a methodological standpoint, PCQA methods are often formulated using a pipeline that includes feature extraction followed by feature regression, although alternative formulations also exist [131–134]. Researchers have utilized both hand-crafted and learning-based features, usually combined with a non-linear function or learning-based regressor [135, 136]. End-to-end schemes can significantly improve the prediction performance by fitting the ground truth well. However, the HVS mechanisms behind them are difficult to explain [137]. In addition, one of the most challenging issues of end-to-end learning-based approaches is their requirement for large amounts of labeled data for training [82, 138, 139]. The learning model may have difficulties in handling various contents and distortions if the training set is not sufficiently large or fails to adequately represent real-world contents [140]. Besides, the shortage of data may probably cause serious over-fitting problems. In practice, existing PCQA datasets remain relatively small and often exhibit imbalance in distortion types. Therefore, it is essential to analyze model performance on a per-dataset basis rather than solely pursuing high-performance metrics.

Among all the visual artifacts for point clouds, the encountered distortions can be categorized into geometric and textural distortions, which can be created by compression algorithms and other noise-generation methods. Particularly in the context of lossy compression, approaches have been devised to encode geometric coordinates or associated attributes, depending on application requirements [85]. Given the necessity of color attributes for human visualization, combining algorithms for both geometric and textural attributes is essential for holistic representation. Consequently, numerous studies have recently evaluated point cloud quality both subjectively and objectively [34]. Subjective studies investigate the quality of point clouds under different distortion types of both geometry and texture attributes or of a single attribute [141]. Objective metrics also follow a similar paradigm to predict quality. Early objective metrics primarily focused on geometric distortions. Geometric-based metrics, from a simple displacement such as point-to-point or point-to-plane [31] distances in the Euclidean space to a more complex geometric feature such as point-to-distribution [142] and density-to-density [143] distances, examine the quality only from a geometric perspective. Color-based metrics [45, 144] produce a score computed only from the color attribute.

However, most existing studies compute geometric or textural similarity between the

reference and distorted point clouds independently [5, 46, 145]. These methods typically rely on point-wise correspondences based on geometric alignment, which may not fully capture perceptual similarity.

To address these limitations, and drawing inspiration from PointPCA [46], this chapter introduces an alternative FR strategy: instead of assessing similarity from the two point clouds separately, we propose to project them into a shared feature space and measure their similarity there. This approach aims to better capture structural and perceptual differences with this projection operation. We present PointPCA+, an enhanced version of PointPCA, and provide detailed analysis based on our prior work [146], including evaluations across various distortion types and an investigation of feature importance for quality prediction. The contributions of PointPCA+ are threefold:

- We extend the PointPCA framework by performing PCA on the geometry data of the reference point cloud and transforming both the reference and distorted point clouds onto the new basis. This way we can capture differences in their shape properties effectively.

- We utilize knn algorithm to determine the neighborhood, which is faster and returns a consistent number of points, therefore further decreasing the computational cost of subsequent processing steps.

- We perform extensive experimentation on four publicly available datasets, demonstrating that PointPCA+ consistently achieves superior performance across four distinct datasets. A thorough analysis investigates the effectiveness of different handcrafted features concerning specific distortion types, aiming to discern which features are more impactful for different distortion type categories.

What's more, we would also want to understand further which distortion we should prioritize when measuring the perceptual quality of the point cloud. These mentioned objective metrics are hard to disentangle when distortions affect both attributes simultaneously, even when one attribute is not explicitly distorted (for example, distortion in geometry will affect the texture). Therefore the landscape has evolved to incorporate both geometry and texture [5, 6, 46, 145], with several approaches integrating multimodal learning. MM-PCQA [7] introduces multimodal learning for PCQA, combining uncolored point clouds and projected texture maps. MFT-PCQA [147] further improves the performance with a mediate-fusion strategy. pmBQA [55] perceives the quality by using 4 homogeneous modalities. Despite these advancements, existing metrics often overlook certain dimensionalities and fail to exploit the potential of both attributes. Besides, the role of the distortion type is ignored. Furthermore, Lazzarotto *et al.* [83] reveal that alternative trade-offs between geometry and texture can potentially provide better visual quality in a pair-wise comparison experiment. These studies shed certain light on how such interplay varies based on the distortion type as a first step towards this underexplored aspect in PCQA in a subjective manner. None of the existing metrics has explored how the geometric/textural distortion and their interplay contribute to the perceived quality of the point cloud automatically. Therefore, a more considerate design that can consider the interplay of such attributes in the HVS is needed.

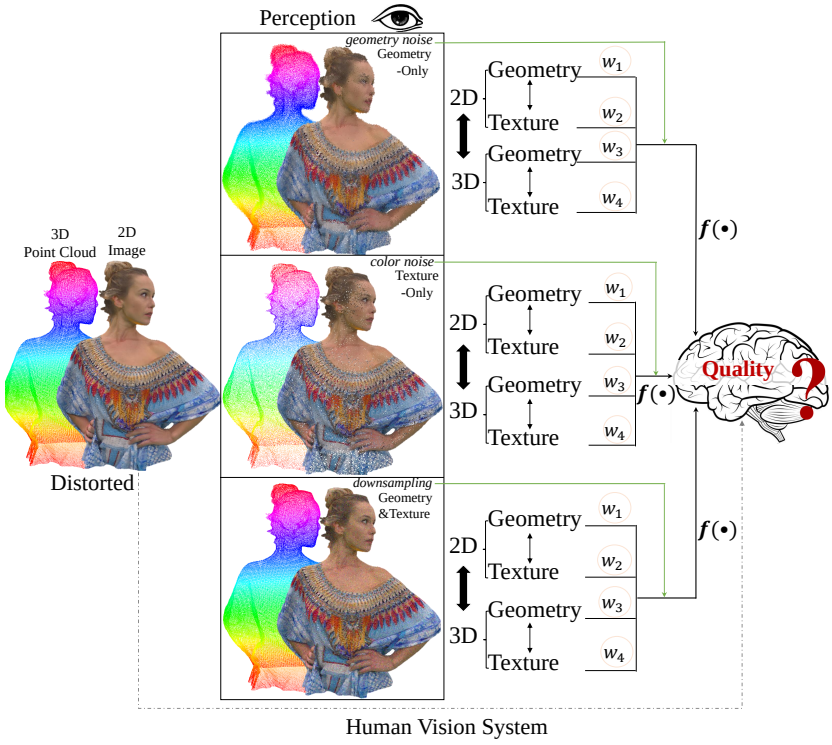


Figure 3.1: **A high level illustration of M3-Unity**. The distortion type serves as a prior in shaping the perceptual quality through the HVS. The interplay of 4 modalities in representing entangled distortion adds complexity to this process.

Understanding attributes and their interrelationships is crucial in various real-world applications. Nevertheless, the relative significance of each attribute representation as well as the interplay between them remain ambiguous in the context of PCQA, which reflects human perception preferences. As for which attribute is more important, we refer to specific distortion types. To this end, this metric, **M**ulti-**M**odality and **MU**lti-task **n**o reference qual**ity** assessment for colored point clouds, termed **M3-Unity**, investigates two attributes and their interplay for perceptual quality assessment in a NR deep learning-based way. In particular, we use additional 3D normal and multi-view projections to retain the intrinsic characteristics of the point cloud and mimic the imaging process of HVS. Additionally, we measure the relationship between geometry and texture and their interplay given a specific distortion type, as demonstrated in Figure 3.1. To summarize, our key contributions of M3-Unity are fourfold:

- We propose M3-Unity, a metric that uses 4 modalities across attributes and dimensionalities to represent the point cloud. The multi-task decoder involving distortion type classification selects the best combination among 4 modalities based on the distortion type, aiding in the regression task.

- The performance of M3-Unity and its variant demonstrates clear advantages over the state-of-the-art metrics across four datasets, showcasing substantial gains in comparison.
- We apply attention mechanism to establish inter/intra associations among patches (especially within dimensionality, we keep the spatial correspondence), yielding both local and global features, to fit the highly nonlinear property of HVS.
- We delve into the relationship between geometric and textural distortion in terms of PCQA. Extensive experiments are conducted to determine whether geometric, textural, or their interplay is prioritized under various distortion types.

3.1. PointPCA+

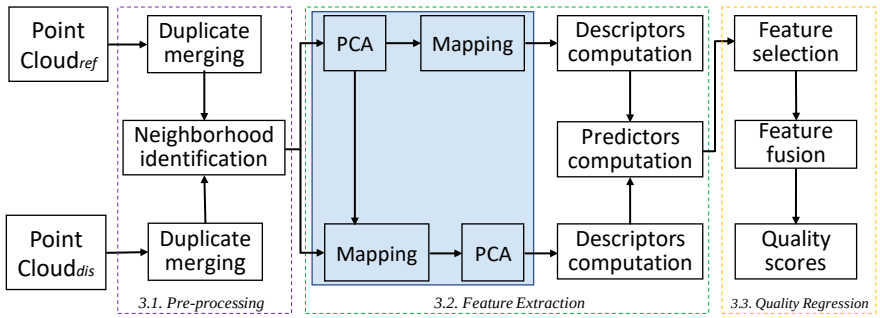


Figure 3.2: PointPCA+ architecture: both the reference and the distorted point cloud are passing through every stage to compute a quality score. Operations in the blue box are applied only to the geometry data of point clouds. Only the reference point cloud serves as a reference to identify neighborhoods.

In Figure 3.2, the PointPCA+ framework is illustrated, which is split into three modules, namely, (a) pre-processing, (b) feature extraction, and (c) quality regression which are introduced in the following subsections. Note that a FR metric typically uses either the pristine or the impaired content as a reference, or both. In our metric design, only the pristine point cloud serves as a reference.

3.1.1. Pre-processing

To ensure coherent geometry and color information without redundancies, points with identical coordinates that belong to the same point cloud are merged [31]. The color of a merged point is obtained by averaging the color of corresponding points sharing the same coordinates. For an FR PCQA metric, identifying matches between reference and distorted point clouds is crucial for comparing corresponding local properties. In our method, we use the knn algorithm to identify neighborhood pairs between two point clouds. In particular, for each point that belongs to a reference point cloud $\mathscr{A}$, we find its N nearest reference points, and its N nearest distorted points from the distorted point cloud $\mathscr{B}$, in terms of Euclidean distance.

3.1.2. Feature Extraction

To capture local perceptual quality degradations of a distorted point cloud, we compute geometry and texture descriptors based on the identified neighborhoods. Statistics based on these descriptors are subsequently calculated and serve as predictors of visual quality. Features are finally obtained via pooling over these predictors. As mentioned earlier, our method uses only the pristine point cloud as a reference to find the matches in the distorted point cloud.

Geometry descriptors Given a query point $\mathbf{p}_i$ of $\mathscr{A}$, the subscript i denotes the point index, $1 \leq i \leq |\mathscr{A}|$, and $|\mathscr{A}|$ is the cardinality. The coordinates of $\mathbf{p}_i$'s N nearest neighbors in $\mathscr{F}$ are indicated as $\mathbf{p}_n^{g,\mathscr{F}} = (x_n, y_n, z_n)^{\mathrm{T}}$, with $1 \leq n \leq N$ and $\mathscr{F} \in \{\mathscr{A}, \mathscr{B}\}$. The geometry of $\mathbf{p}_i$ is denoted as $\mathbf{p}_i^{g,\mathscr{A}}$, and the geometry of its closest neighbor in $\mathscr{B}$ is denoted as $\mathbf{p}_i^{g,\mathscr{B}}$. Initially, the covariance matrix $\Sigma_i^{\mathscr{A}}$ is computed as

$$\Sigma_i^{\mathscr{A}} = \frac{1}{N} \sum_{n=1}^{N} \left(\mathbf{p}_n^{g,\mathscr{A}} - \bar{\mathbf{p}}_i^{g,\mathscr{A}}\right) \cdot \left(\mathbf{p}_n^{g,\mathscr{A}} - \bar{\mathbf{p}}_i^{g,\mathscr{A}}\right)^{\mathrm{T}}, \tag{3.1}$$

where $\bar{\mathbf{p}}_i^{g,\mathscr{A}}$ indicates the centroid, given as

$$\bar{\mathbf{p}}_i^{g,\mathscr{A}} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{p}_n^{g,\mathscr{A}}. \tag{3.2}$$

Then, eigen-decomposition is applied to $\Sigma_i^{\mathscr{A}}$, to obtain the eigenvectors which form an orthonormal basis $\mathbf{V}^{\mathscr{A}}$ composed of eigenvectors $\mathbf{v}_m^{\mathscr{A}}$, where $m = 1, 2, 3$. Next, we map the reference and distorted neighborhoods to the new orthonormal basis, denoted as $\omega_n^{\mathscr{F}} = (\mathbf{p}_n^{g,\mathscr{F}} - \bar{\mathbf{p}}_i^{g,\mathscr{A}}) \cdot \mathbf{V}^{\mathscr{A}}$. Finally, we apply PCA to the covariance matrix of $\omega_n^{\mathscr{B}}$ and compute the eigenvectors $\mathbf{v}_m^{\mathscr{B}}$. This process is visually demonstrated in Figure 3.3, showcasing the distinction between the two bases.

The merit inherent in projecting the geometry of both the reference and distorted point clouds onto a shared orthonormal basis, established by the reference point cloud, lies in the capacity to unify the representation of geometry degradation within a common space. This enables a more precise measurement of geometry similarity within the framework of the FR PCQA paradigm.

The mapped coordinates of the reference and distorted points $\omega_n^{\mathscr{F}}$, the eigenvectors $\mathbf{v}_m^{\mathscr{F}}$ and the unit vectors $\mathbf{u}_m$, with $\mathbf{u}_1 = [1,0,0]^{\mathrm{T}}$, $\mathbf{u}_2 = [0,1,0]^{\mathrm{T}}$ and $\mathbf{u}_3 = [0,0,1]^{\mathrm{T}}$, are used to construct the geometric descriptors defined in Table 3.1.

Texture descriptors The color space is first converted from RGB to YCbCr [148]. This conversion is motivated by the fact that the human eye is more sensitive to changes in brightness than changes in color according to HVS. We denote the texture information of $\mathbf{p}_i$'s N nearest neighbors in $\mathscr{F}$ as $\mathbf{p}_n^{t,\mathscr{F}} = (Y_n, Cb_n, Cr_n)^{\mathrm{T}}$. The proposed 6 texture descriptors are defined in Table 3.1.

Explanation of descriptors Each geometry descriptor represents an interpretable shape property inside the neighborhood. Specifically, $\mathbf{e}$ denotes the error vector between the

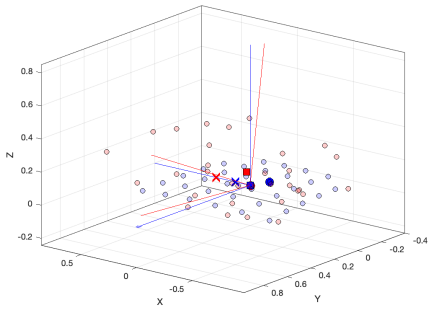


Figure 3.3: The orthonormal basis formed by both reference (*longdress*) and distorted (*longdress_Octree-Lifting_R04*) point clouds. The orange color represents the geometry around one point (2130th point) of the reference point cloud and the corresponding basis after PCA operation; the purple color represents the geometry around the matched point of the distorted point cloud and the corresponding basis after PCA operation.

(a) *longdress* (b) $\mu(\mathrm{L}^{\mathcal{A}})$: Point-PCA (c) $\mu(\mathrm{L}^{\mathcal{A}})$: Point-PCA+ (d) $\mu(\mathrm{P}^{\mathcal{A}})$: Point-PCA (e) $\mu(\mathrm{P}^{\mathcal{A}})$: Point-PCA+

Figure 3.4: The point cloud *longdress* and statistical features using mean of linearity (Figures 4.7(b)-4.11(c)), planarity (Figures 4.7(d)-4.7(e)). The amplitudes of statistical features are color-mapped, with red indicating higher and blue lower values. It can be noticed that the mean of linearity (4.7(b)) and planarity (4.7(d)) of PointPCA/PointPCA+ capture high- and low-frequency geometric regions, respectively. Additionally, PointPCA+ has lower complexity as the neighborhood is determined through *knn*.

Table 3.1: Definition of descriptors.

	Descriptor	Definition	Distance
Geometric	Error vector	$\mathbf{e} = (\omega_i^{\mathscr{B}} - \omega_i^{\mathscr{A}})$	r_α
	Error along axes	$\epsilon_m = (\omega_i^{\mathscr{B}} - \omega_i^{\mathscr{A}})^T \cdot \mathbf{u}_m$	r_β
	Error from origin	$\varepsilon^{\mathscr{F}} = \omega_i^{\mathscr{F}}$	r_α, r_β
	Mean	$\mu^{\mathscr{B}} = \frac{1}{N}\sum_n \omega_n^{\mathscr{B}}$	r_α, r_β
	Variance	$\lambda^{\mathscr{F}} = \frac{1}{N}\sum_n \left(\omega_n^{\mathscr{F}} - \mu^{\mathscr{F}}\right)^2$	r_δ
	Sum of variance	$\Sigma^{\mathscr{F}} = \sum_m \lambda_m^{\mathscr{F}}$	r_δ
	Covariance	$\Sigma = \frac{1}{N}\sum_n \left(\omega_n^{\mathscr{A}} - \mu^{\mathscr{A}}\right) \cdot \left(\omega_n^{\mathscr{B}} - \mu^{\mathscr{B}}\right)^{\mathrm{T}}$	r_γ
	Omnivariance	$\mathrm{O}^{\mathscr{F}} = \sqrt[3]{\prod_m \lambda_m^{\mathscr{F}}}$	r_λ
	Eigenentropy	$\mathrm{E}^{\mathscr{F}} = -\sum_m \lambda_m^{\mathscr{F}} \cdot \log \lambda_m^{\mathscr{F}}$	r_δ
	Anisotropy	$\mathrm{A}^{\mathscr{F}} = (\lambda_1^{\mathscr{F}} - \lambda_3^{\mathscr{F}}) / \lambda_1^{\mathscr{F}}$	r_δ
	Planarity	$\mathrm{P}^{\mathscr{F}} = (\lambda_2^{\mathscr{F}} - \lambda_3^{\mathscr{F}}) / \lambda_1^{\mathscr{F}}$	r_δ
	Linearity	$\mathrm{L}^{\mathscr{F}} = (\lambda_1^{\mathscr{F}} - \lambda_2^{\mathscr{F}}) / \lambda_1^{\mathscr{F}}$	r_δ
	Scattering	$\mathrm{S}^{\mathscr{F}} = \lambda_3^{\mathscr{F}} / \lambda_1^{\mathscr{F}}$	r_δ
	Change of curvature	$\mathrm{C}^{\mathscr{F}} = \lambda_3^{\mathscr{F}} \big/ \sum_m \lambda_m^{\mathscr{F}}$	r_δ
	Parallelity	$\mathscr{P}_m = 1 - \mathbf{u}_m \cdot \mathbf{v}_m^{\mathscr{B}}$	–
	Angular similarity	$\theta = 1 - \frac{2 \cdot \arccos(\cos(\mathbf{u}_m, \mathbf{v}_m^{\mathscr{B}}))}{\pi}$	–
Textural	Mean	$\tilde{\mu}^{\mathscr{F}} = \frac{1}{N}\sum_n \mathbf{p}_n^{t,\mathscr{F}}$	r_δ
	Variance	$\tilde{\mathbf{s}}^{\mathscr{F}} = \frac{1}{N}\sum_n \left(\mathbf{p}_n^{t,\mathscr{F}} - \tilde{\mu}^{\mathscr{F}}\right)^2$	r_δ
	Sum of variance	$\tilde{\Sigma}^{\mathscr{F}} = \sum_m \tilde{s}_m^{\mathscr{F}}$	r_δ
	Covariance	$\tilde{\Sigma} = \frac{1}{N}\sum_n \left(\mathbf{p}_n^{t,\mathscr{A}} - \tilde{\mu}^{\mathscr{A}}\right) \cdot \left(\mathbf{p}_n^{t,\mathscr{B}} - \tilde{\mu}^{\mathscr{B}}\right)^{\mathrm{T}}$	r_γ
	Omnivariance	$\tilde{\mathrm{O}}^{\mathscr{F}} = \sqrt[3]{\prod_m \tilde{s}_m^{\mathscr{F}}}$	r_δ
	Entropy	$\tilde{\mathrm{H}}^{\mathscr{F}} = -\sum_m \tilde{s}_m^{\mathscr{F}} \cdot \log \tilde{s}_m^{\mathscr{F}}$	r_δ

mapped coordinates of the reference query point and its nearest neighbor, and ϵ_m is the projected distance of the error vector across the m-th axis. The ε is used to capture the

Euclidean and projected distances of the mapped reference query point or its nearest distorted neighbor from the centroid and principal axes, respectively. $\mu^{\mathscr{B}}$, $\lambda^{\mathscr{F}}$, $\Sigma^{\mathscr{F}}$ and Σ reveal local statistics. $\mathrm{E}^{\mathscr{F}}$ provides an estimation of the space uncertainty on the projected surfaces. Additionally, $\mathscr{P}_m$ and θ_m assess the parallelity and the angular dispersion of the distorted plane. The remaining geometry descriptors explore the topology of a local region from different aspects, relying on the spatial dispersion along different principal axes. $\tilde{\mu}^{\mathscr{F}}$, $\tilde{\mathbf{s}}^{\mathscr{F}}$ and $\tilde{\Sigma}^{\mathscr{F}}$ of the YCbCr channel express the intrinsic distribution of luminance and chromatic components. $\tilde{\Sigma}$ and $\tilde{\mathrm{O}}^{\mathscr{F}}$ show the variability of color information. $\tilde{\mathbf{H}}^{\mathscr{F}}$ provides an estimation of color uncertainty of the local region. Every descriptor is computed per point $\mathbf{p}_i$.

Predictors Predictors are defined as the error samples obtained by computing a distance over descriptor values. We define different distance functions for different descriptors. We use the Euclidean distance to measure the point-to-point distances between query point pairs under the new basis

$$r_\alpha = \sqrt{\textstyle\sum_m \mathbf{d_1}^2}, \tag{3.3}$$

where $\mathbf{d_1}$ is the difference between two points. We use the absolute value to measure the point-to-plane distance, as

$$r_\beta = |\mathbf{d_2}|, \tag{3.4}$$

where $\mathbf{d_2}$ indicates the projected distance between a point and the reference axes. We use the following definition of relative difference for the covariance features

$$r_\gamma = \frac{|\mathbf{q}^{\mathscr{A}} \odot \mathbf{q}^{\mathscr{B}} - \mathbf{Q}|}{\mathbf{q}^{\mathscr{A}} \odot \mathbf{q}^{\mathscr{B}}}, \tag{3.5}$$

where $\{\mathbf{q}^{\mathscr{F}} = \lambda^{\mathscr{F}}, \mathbf{Q} = \Sigma\}$ and $\{\mathbf{q}^{\mathscr{F}} = \tilde{\mathbf{s}}^{\mathscr{F}}, \mathbf{Q} = \tilde{\Sigma}\}$, for geometry and texture attributes, respectively, $\odot$ is for element-wise product. We use the relative difference formula [145], for the remaining descriptors

$$r_\delta = \frac{|\phi^{\mathscr{A}} - \phi^{\mathscr{B}}|}{|\phi^{\mathscr{A}}| + |\phi^{\mathscr{B}}| + \varepsilon}, \tag{3.6}$$

where ε is a small constant to avoid undefined operations. Finally, the definitions of parallelity and angular similarity descriptors incorporate a distance function. For notational purposes only, we define distances r_ρ and r_θ to be identical to the definitions of $\mathscr{P}_m$ and θ_m, respectively. Table 3.1 enlists distance function(s) used per descriptor.

Features Features are defined by pooling over predictor values. Specifically, predictors $\psi_{i,j,k}$ are obtained per point $\mathbf{p}_i$, descriptor j, and distance function r_k, $k \in \{\alpha, \beta, \gamma, \delta, \rho, \theta\}$. This is done for all descriptors j in Table 3.1, using the corresponding distances r_k. Through pooling, we obtain a feature $f_{j,k}$ for every predictor:

$$f_{j,k} = \frac{1}{|\mathscr{A}|} \textstyle\sum_{i=1}^{|\mathscr{A}|} \psi_{i,j,k}. \tag{3.7}$$

(a) *longdress-Octree-Lifting-R02* (b) $L^{\mathscr{B}}$ (c) $P^{\mathscr{B}}$

Figure 3.5: The predictors of linearity (Figure 3.5(b)) and planarity (Figure 3.5(c)) of point cloud *longdress-Octree-Lifting-R02* (Figure 3.5(a)). The amplitudes of predictors are color-mapped, with red indicating higher and blue lower values. Notably, the relative differences in linearity and planarity of PointPCA+ effectively highlight regions that can form linear structures and detect bulges, such as wrinkles in clothing, in the *longdress*.

3.1.3. Quality Regression

To obtain a quality score that is well-aligned with the HVS, the Recursive Feature Elimination (RFE) algorithm is used to select the most relevant predictor set among all the proposed predictors. RFE [149] improves model accuracy, and efficiency, and reduces overfitting. Machine learning-based regression models have been extensively used to tackle the quality regression problem in the domain of quality assessment, we then use the random forest algorithm to regress the selected predictors to the final quality score.

3.1.4. Differences with PointPCA

In comparison to PointPCA, our pre-processing methodology involves solely utilizing the pristine point cloud as a reference, dispensing with the need for both pristine and distorted point clouds. Regarding geometric features, we transform the xyz values of both the pristine and distorted point clouds into the basis formed by the pristine point cloud, eliminating the necessity for separate bases for distinct point cloud sets. For textural features, we directly apply statistical functions to the color values in the YCbCr space without resorting to PCA decomposition. The decision to forego PCA on texture stems from the lack of physical significance post-decomposition of YCbCr channel values, as observed in geometry. Simultaneously, this approach aids in reducing computational complexity. Additionally, PointPCA+ adopts distinct distance functions for various descriptors, eliminating constraints on the computation of mean and standard deviation values. We illustrate the disparities between PointPCA and PointPCA+ for a specific point cloud, *longdress*, in Figure 3.4. Notably, we visualize the linearity and planarity geometric descriptors before

comparison. From Figure 3.4, the geometric dissimilarity measurements differ when using knn and $r\text{-}search$ methods. For PointPCA, we employed the $r\text{-}search$ method with $r = 0.008 \times B$, where B represents the maximum length of the bounding box of the reference point cloud. In contrast, for PointPCA+, we used the knn algorithm with $k = 81$. While for the statistical features μ, the knn algorithm with $k = 9$ was applied. Both the linearity and planarity are decreased compared with PointPCA. We also visualize the predictions corresponding to linearity and planarity in Figure 3.5. We can see that the predictors exactly describe the line and the uneven region of point cloud *longdress*.

3.1.5. Experimental Results

In this Section, we report the evaluation results of the proposed PointPCA+ metric under three public datasets with other 8 state-of-the-art metrics. Moreover, we report the performance achieved in the ICIP 2023 Point Cloud Visual Quality Assessment (PCVQA) grand challenge[1]. Specifically, the challenge consists of 5 tracks, which correspond to different use cases in which quality metrics are typically used. The first two tracks aim to assess the perceptual fidelity of distorted contents with/without respect to the originals for any level of distortion, respectively. This is the most generic and traditional set-up for quality metrics. The next two tracks focus on metrics for high-end quality with/without access to the original content. These are desirable in applications such as content production, high-quality streaming, digital twins, etc. The last track should be sensible to quality differences within different processed versions of the same point cloud content, which is suitable for optimization scenarios. We participated in Track#1 FR broad-range quality estimation, Track#3 FR high-range quality estimation, and Track#5 FR intra-reference quality estimation. Additional analysis related to cross-dataset validation and feature importance is carried out across all the aforementioned datasets. These IN-DEPTH analysis aim to demonstrate the generalizability of the proposed PointPCA+.

Three publicly available datasets were recruited for performance evaluation, namely, M-PCCD, SJTU, and WPC. For the PCVQA Grand Challenge, all submissions undergo testing on a designated test set curated by the organizers. The evaluation of performance relies on five standard criteria provided by the organizers: including PLCC, SRCC, Difference/Similar Analysis quantified by Area Under the Curve (D/S$_{AUC}$), Better/Worse Analysis quantified by Correct Classification percentage (B/W$_{CC}$) [151], and the Runtime Complexity (RC). Notably, no function is employed for score mapping. Additionally, for the assessment of performance metrics on three other commonly used datasets, the criteria including SRCC, PLCC, KRCC, and MSE are chosen.

Implementation details

We use RFE to select the best feature set among all the predictors, with the best SRCC performance on the PCVQA grand challenge test set. In the inference stage, the default configuration of scikit-learn (version 1.2.2) in Python is used. Regarding the neighborhood size for the computation of descriptors, $K = 81$ is chosen considering complexity and performance, after experimenting with $K \in \{9, 25, 49, 81, 121\}$.

[1]https://sites.google.com/view/icip2023-pcvqa-grand-challenge

Table 3.2: SRCC performance on M-PCCD, SJTU and WPC datasets

Metric	PointPCA+	PointPCA[46]	PCQM[5]	PointSSIM[145]	BitDance[150]	Plane2Plane[35]	P2Plane_MSE[34]	P2P_MSE [34]	PSNR Y[34]
M-PCCD	**0.943**±0.022	0.941±0.032	0.940±0.032	0.925±0.024	0.859±0.061	0.847±0.076	0.901±0.025	0.896±0.042	0.798±0.162
SJTU	0.865±0.064	**0.890**±0.056	0.862±0.030	0.708±0.070	0.748±0.077	0.761±0.039	0.578±0.155	0.612±0.157	0.743±0.083
WPC	0.857±0.040	**0.866**±0.036	0.749±0.036	0.465±0.059	0.451±0.054	0.454±0.069	0.452±0.065	0.563±0.071	0.614±0.061

Performance evaluation on M-PCCD, SJTU and WPC

We compare PointPCA+ with existing FR point-based quality metrics, the results are shown in Table 3.2. The best performance among these metrics is highlighted in boldface, with the second best underlined. Specifically, each dataset is split into two partitions that contain 80% and 20% of the contents for training and testing, respectively, with all the distorted versions of a specific content placed in one partition. For M-PCCD, SJTU, and WPC, we use 6/2, 7/2, and 16/4 contents for training/testing, respectively. Then, a quality prediction model is trained on the training data and tested on the corresponding testing data of the same dataset, for within-dataset validation. this process is repeated for all possible 80%-20% splits of each dataset, leading to 28, 36, and 4845 testing partitions and an equal number of corresponding quality prediction models for M-PCCD, SJTU, and WPC respectively. Finally, the average and the standard deviation of SRCC index computed across all testing splits of each dataset, are reported. From Table 3.2 we can see that PCA-based metrics are competitive with the highest SRCC on the three datasets, especially the performance of PointPCA on WPC is increased by 15.62% in terms of SRCC though the performance of PointPCA+ is a slightly lower than PointPCA (0.866 VS 0.857).

Performance evaluation on BASICS

We split BASICS into training-validation-test with 60%-20%-20% following the rules from the PCVQA grand challenge [152]. Table 3.3 to Table 3.5 show the official evaluation results of Track#1, Track#3 and Track#5, respectively.

Referencing Tables 3.4-3.5, several notable observations emerge from the competition results across all three FR tracks. 1) Despite strong performances in Track 1 and Track 5, none of the teams attained satisfactory results in Track 3 of in PCVQA, highlighting the challenges associated with fine-grained PCQA. 2) Examining PointPCA+, it is evident that the extracted features within a neighborhood size of 81, combined with the application of statistical functions (e.g., mean and variance) on geometry, may fail to capture subtle differences between two point clouds. This highlights the limitations of using statistical features to capture subtle differences.

Cross-dataset validation

To verify the generalization and robustness of the proposed PointPCA+, we conduct cross-dataset experiments among all 4 datasets. We train the model using the entire content of one dataset and then test it separately using the entire content of the other three datasets. The experimental results are shown in Table 3.9. From Table 3.9, we can draw the following observations:

1. PointPCA+ performs well in generalization and robustness, particularly when trained

Table 3.3: Track#1 (FR broad range): top 4 performance comparison on the official PCVQA grand challenge test set, evaluated by the challenge organizers. Best in bold and second best underlined. Our submission is ranked in 2nd place.

Submission	SRCC	PLCC	D/S_{AUC}	B/W_{CC}	RC(s)
KDDIUSCJoint	**0.875**	**0.917**	**0.888**	**0.970**	<u>42.80</u>
PointPCA+	<u>0.874</u>	<u>0.909</u>	<u>0.871</u>	<u>0.961</u>	1000.00
SJTU MMLAB	0.871	0.896	0.832	0.955	**8.60**
SlowHand	0.791	0.825	0.805	0.924	130.47

Table 3.4: Track#3 (FR high range): top 4 performance comparison on the official PCVQA grand challenge test set, evaluated by the challenge organizers. Best in bold and second best underlined. Our submission is ranked in 3rd place.

Submission	SRCC	PLCC	D/S_{AUC}	B/W_{CC}	RC(s)
SJTU MMLAB	**0.630**	**0.592**	**0.665**	**0.909**	**8.60**
KDDIUSCJoint	0.551	<u>0.516</u>	<u>0.642</u>	0.872	<u>42.80</u>
PointPCA+	<u>0.603</u>	0.479	0.625	<u>0.886</u>	1000.00
SlowHand	0.377	0.423	0.565	0.780	130.47

Table 3.5: Track#5 (FR intra-reference): top 4 performance comparison on the official PCVQA grand challenge test set, evaluated by the challenge organizers. Best in bold and second best underlined. Our submission is ranked in 3rd place.

Submission	D/S_{AUC}	B/W_{CC}	RC(s)
SJTU MMLAB	0.808	**0.947**	**8.60**
KDDIUSCJoint	**0.822**	0.933	<u>42.80</u>
PointPCA+	<u>0.811</u>	<u>0.938</u>	1000.00
SlowHand	0.753	0.854	130.47

Table 3.6: Cross-dataset validation among M-PCCD, SJTU, WPC and BASICS datasets. Both the training and testing used all the content among the datasets. Best in bold.

	Test															
	M-PCCD				SJTU				WPC				BASICS			
Train	PLCC	SRCC	KRCC	RMSE	PLCC	SRCC	KRCC	RMSE	PLCC	SRCC	KRCC	RMSE	PLCC	SRCC	KRCC	RMSE
M-PCCD	–	–	–	–	**0.726**	**0.725**	**0.541**	3.148	0.500	0.469	0.328	4.003	0.847	**0.777**	**0.589**	**0.899**
SJTU	**0.855**	**0.881**	**0.708**	2.578	–	–	–	–	**0.578**	**0.608**	**0.442**	**2.575**	0.732	0.717	0.523	1.946
WPC	0.731	0.878	0.703	4.767	0.610	0.604	0.435	**2.049**	–	–	–	–	**0.848**	0.726	0.536	3.675
BASICS	0.832	0.880	0.692	**1.057**	0.579	0.645	0.472	2.810	0.500	0.490	0.347	3.202	–	–	–	–

on the small M-PCCD dataset and tested on the large BASICS dataset. Combined with Table 3.2, the cross-dataset evaluation performance is even higher than certain FR PCQA metrics, for example BitDance [150], PSNR_Y [34], etc.

2. Compared with M-PCCD, SJTU and BASICS datasets, the WPC dataset has the worst performance among all the evaluation metrics. SRCC and PLCC of PointPCA+ on WPC trained on M-PCCD and BASICS achieve the same accuracy as random guessing, this may be because the contents in WPC only contain objects,

and the distortion types of WPC are more complex compared with the other three datasets.

3. PointPCA+ shows better SRCC performance on the SJTU dataset compared to the WPC dataset. This is likely because SJTU shares specific human figures with the M-PCCD dataset and includes both human and object categories for M-PCCD and BASICS, while SJTU and WPC have no overlapping content.

In conclusion, the cross-dataset performance of PointPCA+ is promising but its generalization depends on dataset composition. Training on large, diverse datasets and testing on smaller ones normally yields better results, while the reverse leads to poor generalization. However, this can not hold if there exists a domain shift (i.e., in our case, contents and distortion type) between the testing set and the training set.

Performance on individual distortion type

To further explore the effectiveness of the designed geometry and texture features for a specific distortion type, we test the performance per distortion type per dataset, with the results listed in Table 3.7. We can draw the following observations.

1. When assessing compression distortion, V-PCC emerges as the most challenging distortion type of which to evaluate perceptual quality, aligning with findings from a prior study [44]. Predicting the perceptual quality of compression distortions from G-PCC and learning-based methods proves to be more manageable, with Octree-Lifting exhibiting a slight advantage over Octree-RAHT on M-PCCD and BASICS datasets.

2. For CN and Gaussian noise distortions, CN has the poorest performance. However, prediction accuracy improves by 16.06% for CN+GGN in terms of SRCC on the SJTU dataset, indicating that geometry-related features help capture these distortions. Gaussian noise performs worst on WPC, as it affects both geometry and texture, creating a compounded distortion.

3. For downsampling distortion types, the performance of PointPCA+ is notably high on the SJTU dataset but relatively lower on the WPC dataset. This implies that downsampling on objects presents a challenge for the HVS to discern, as it may exert a masking effect on objects more prominently than on humans.

In summary, PointPCA+ exhibits proficiency in predicting compression distortions. However, its effectiveness diminishes when confronted with distortion instances primarily manifesting in color values, as observed for CN distortions on the SJTU dataset. Additionally, when the equilibrium between geometric and color is disrupted, as exemplified by Gaussian noise on the WPC dataset, PointPCA+ struggles to accurately gauge the extent of degradation.

Table 3.7: Performance comparison of PointPCA+ metrics for different distortion types per dataset. DT refers to Distortion Type and NC denotes the Number of Contents.

Dataset	DT	NC	PLCC	SRCC	KRCC	RMSE
M-PCCD	Octree-Lifting	48	0.938	0.991	0.962	0.826
	Octree-RAHT	48	0.962	0.984	0.931	0.704
	TriSoup-Lifting	48	0.951	0.965	0.879	0.597
	TriSoup_RAHT	48	0.970	0.963	0.870	0.561
	V-PCC	40	0.823	0.815	0.674	1.266
SJTU	CN	54	0.621	0.741	0.545	2.594
	CN + GGN	54	0.894	0.860	0.697	0.825
	Downsampling	54	0.969	0.944	0.848	0.562
	Downsampling+CN	54	0.946	0.937	0.788	1.371
	Downsampling+GGN	54	0.981	0.965	0.879	0.736
	GGN	54	0.955	0.958	0.848	0.810
	Octree (PCL)	54	0.975	0.965	0.879	0.797
WPC	Downsampling	60	0.802	0.795	0.594	1.177
	Octree (LPCC)	80	0.916	0.881	0.708	0.850
	Trisoup (SPCC)	240	0.987	0.923	0.818	0.594
	Gaussian noise	180	0.628	0.650	0.417	1.826
	V-PCC	180	0.737	0.756	0.566	2.054
BASICS	GeoCNN	294	0.971	0.941	0.787	0.305
	Octree-Predlift	375	0.975	0.937	0.792	0.289
	Octree-RAHT	375	0.895	0.876	0.693	0.441
	V-PCC	450	0.787	0.690	0.514	0.371

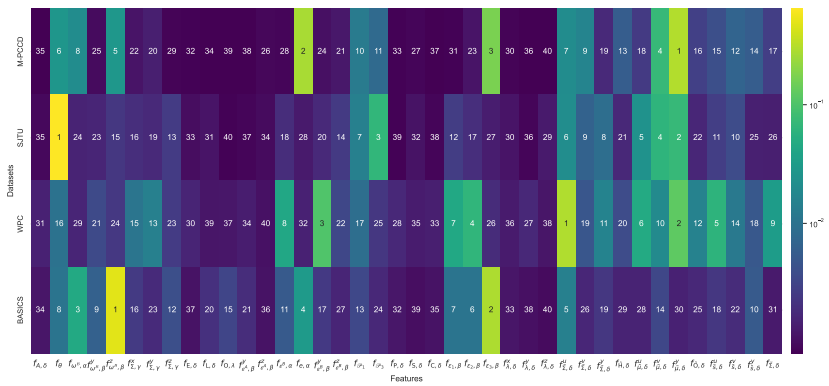
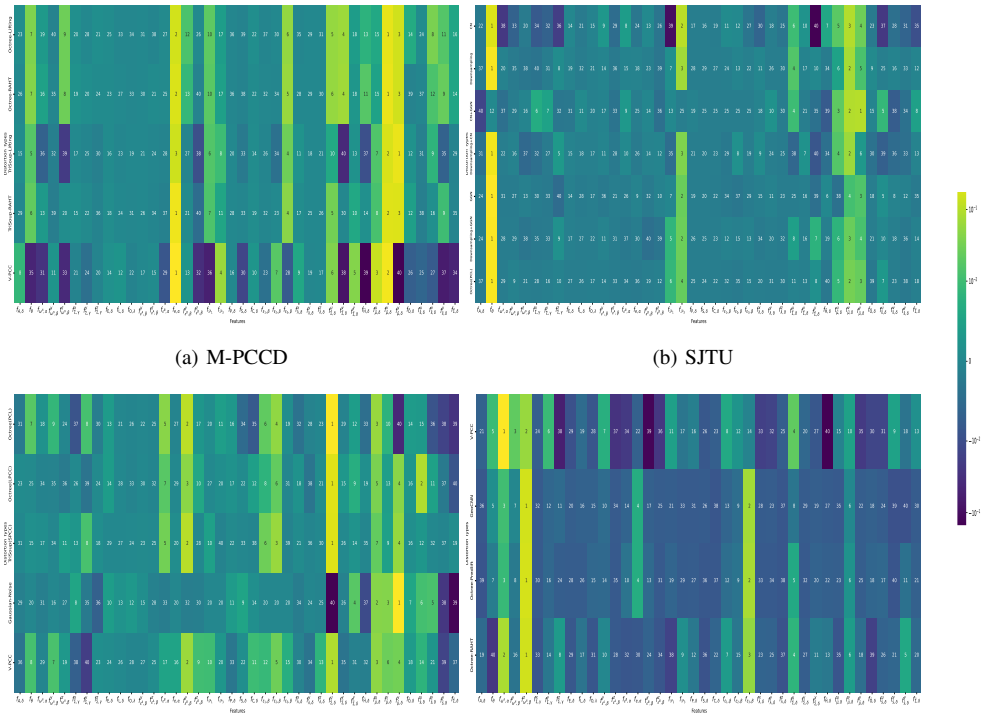


Figure 3.6: The feature importance of all the extracted geometry and texture features within pointPCA+ metric for M-PCCD, SJTU, WPC and BASICS datasets. The numbers are obtained by ranking the 40 features based on the importance score per dataset.



(a) M-PCCD (b) SJTU

(c) WPC (d) BASICS

Figure 3.7: Feature importance of PointPCA+ per distortion type for M-PCCD, SJTU, WPC, and BASICS datasets. The numbers represent ranked permutation feature importance scores across 40 features per distortion type.

Feature importance of PointPCA+ for different distortion types per dataset

We evaluated the effectiveness of the 40 selected features after RFE. The model was trained on 80% of the dataset, and feature importance was calculated on the remaining 20% for each distortion type. This was done using the permutation importance technique [153] based on MSE. Permutation importance shows how crucial a feature is for a specific model, rather than its standalone predictive value, helping to assess the generalization ability of the features across different distortion types. The feature importance of PointPCA+ across the four datasets is shown in Figure 3.6. The feature importance for each distortion type per dataset is illustrated in Figure 3.7. Combining Figure 3.6 and Figure 3.7, we can draw the following conclusions:

1. The feature importance ranking varies across different datasets, primarily due to distinctions in content and distortion types among the four datasets. The variance on the z axis ($f^{z}_{\lambda,\delta}$) within the geometric features obtained the lowest importance ranking on both M-PCCD and BASICS, ranked 29th and 38th on SJTU and WPC datasets, respectively.

2. Geometric features consistently exhibit superior feature importance rankings when compared to textural features, as observed in the top-5 features across all four datasets. In M-PCCD, SJTU, and WPC, the top-3 rankings comprise a combination of texture and geometry features. However, in BASICS, only geometry features are represented in the top-3 ranking.

3. The point-to-point distance ($f_{\mathbf{e},\alpha}$) and the mean value of v channel ($f^{v}_{\tilde{\mu},\delta}$) demonstrate strong performance across various distortion types in M-PCCD dataset. Notably, in TriSoup-Lifting, the mean value of y channel ($f^{y}_{\tilde{\mu},\delta}$) excels but performs less optimally in the case of V-PCC distortion. In SJTU dataset, the cosine similarity of the y axis (f^{y}_{θ}) outperforms other features for all distortion types, except for CN+GGN distortion. Similarly, the covariance of the u channel ($f^{u}_{\tilde{\Sigma}}$) excels for all distortion types, with the exception of Gaussian noise in WPC. Meanwhile, in BASICS, the projected distances of the distorted centroid from reference planes on the z axis ($f^{z}_{\omega^{B},\beta}$) consistently yield high rankings across all distortion types.

3.2. M3-Unity

We illustrate the proposed M3-Unity as shown in Figure 3.8. First, we preprocess the colored point cloud and extract multimodal features with 3D and 2D encoders, respectively (3.2.2). Second, we introduce the cross-attributes attentive fusion module, which captures the local and global associations at both the intra- and inter-modality perception (3.2.3). Last, we employ dual decoders to jointly learn both quality regression and distortion-type classification (3.2.4).

3.2.1. Multimodal geometry-texture input processing

A colored point cloud, denoted as $\mathscr{P}$, is a set of N 3D point elements. Each point element is assigned a 3D coordinate $\mathbf{p}^{\text{coord}} \in \mathbb{R}^3$ and an RGB color value $\mathbf{p}^{\text{RGB}} \in \mathbb{R}^3$ as features:

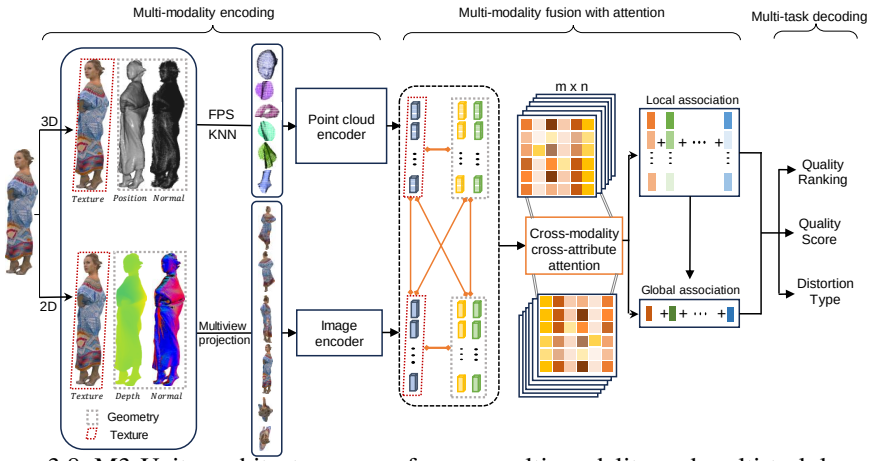


Figure 3.8: M3-Unity architecture: no-reference multi-modality and multi-task learning for PCQA.

$\mathscr{P} = \{(\mathbf{p}_i^{\text{coord}}, \mathbf{p}_i^{\text{RGB}})\}_{i=1}^{N}$. We introduce how the point cloud data is processed into multiple modalities of geometry and texture features as follows.

Processing the point cloud as 3D patches. To deal with dense point clouds of very large N with common neural architectures for point cloud encoding, we first decompose each point cloud into patches following [7, 154]. we obtain a set of $n = 6$ point cloud patches from each of the point cloud $\mathbf{P} \subset \mathscr{P}$, and each $\mathbf{P}$ is of cardinality k. To do this, we adopt Farthest Point Sampling (FPS) to obtain a set of anchor points and find the K-Nearest Neighbors (KNN) for each point. For each point cloud patch $\mathbf{P}$, we describe the geometry and texture features for each point element, such that the texture features are essentially the RGB features $\mathbf{p}^{\text{tex}} = \mathbf{p}^{\text{RGB}} \in \mathbb{R}^3$, and the geometry feature is the 3D coordinate $\mathbf{p}^{\text{coord}}$, augmented by concatenating a normal vector $\mathbf{p}^{\text{normal}}$ calculated from the original point cloud as $\mathbf{p}^{\text{geo}} = [\mathbf{p}^{\text{coord}}, \mathbf{p}^{\text{normal}}] \in \mathbb{R}^6$, i.e. $\mathbf{P} = \{(\mathbf{p}_i^{\text{geo}}, \mathbf{p}_i^{\text{tex}})\}_{i=1}^{k}$. Additionally, $\mathbf{P} \in \mathbb{P}$ where $\mathbb{P}$ is defined as the set of all 3D point patches extracted from the same point cloud.

Processing the point cloud as projected views. We further project the colored point cloud to $m = 6$ 2D views following Liu *et al.* [44], which are evenly distributed in the 3D space from the ∞ and $-\infty$ of the three Cartesian coordinate axes. For each 2D view, the color RGB values from the 3D points are ray-casted to the pixel space, and we calculate depth and normal maps from the 3D geometry, resulting in the 2D geometry feature $\mathbf{X}^{\text{geo}} \in \mathbb{R}^{H \times W \times 4}$ and the 2D texture feature $\mathbf{X}^{\text{tex}} \in \mathbb{R}^{H \times W \times 3}$, where $H \times W$ is the pixelated resolution of the 2D projections. Similarly we define $\mathbb{X}$ as the set of six projected views from a point cloud: $\mathbf{X} = [\mathbf{X}^{\text{geo}}, \mathbf{X}^{\text{tex}}] \in \mathbb{X}$.

3.2.2. Point cloud multimodal encoding

The goal of multimodal encoding is to represent 3D point cloud patches and 2D projection views as embeddings and adapt those embeddings for multimodal fusion.

For the 3D modality, we opt for PointNet++ [155] to encode each 3D point cloud patch $\mathbf{P} = \{(\mathbf{p}_i^{\text{geo}}, \mathbf{p}_i^{\text{tex}})\}_{i=1}^{k} \subset \mathbb{P}$ while separating attributes from geometry and texture:

$$\mathbf{h}_{\text{3D}}^{\text{geo}} = \textsc{PointNet++}\left(\{\mathbf{p}_i^{\text{geo}}\}_{i=1}^{k}\right); \tag{3.8}$$

$$\mathbf{h}_{\text{3D}}^{\text{tex}} = \textsc{PointNet++}\left(\{\mathbf{p}_i^{\text{tex}}\}_{i=1}^{k}\right). \tag{3.9}$$

$\mathbf{h}_{\text{3D}}^{\text{geo}} \in \mathbb{R}^d$ and $\mathbf{h}_{\text{3D}}^{\text{tex}} \in \mathbb{R}^d$ are d-dimensional embeddings of 3D geometry and texture features. Note that to encode texture feature, we still use the 3D coordinates to obtain spatial processes in the PointNet++ such as the farthest-point sampling and grouping.

For the 2D modality, we choose ResNet50 [156] as the 2D encoder that applies to the geometry and texture channels $\mathbf{X}^{\text{geo}}$ and $\mathbf{X}^{\text{tex}}$ separately of each 2D project view $\mathbf{X} \in \mathbb{X}$:

$$\mathbf{h}_{\text{2D}}^{\text{geo}} = \textsc{ResNet}\left(\mathbf{X}^{\text{geo}}\right); \tag{3.10}$$

$$\mathbf{h}_{\text{2D}}^{\text{tex}} = \textsc{ResNet}\left(\mathbf{X}^{\text{tex}}\right). \tag{3.11}$$

Likewise, $\mathbf{h}_{\text{2D}}^{\text{geo}} \in \mathbb{R}^d$ and $\mathbf{h}_{\text{2D}}^{\text{tex}} \in \mathbb{R}^d$ are encoded as d-dimensional 2D geometry and texture embeddings.

3.2.3. Cross-attribute attentive fusion

The core mechanism of attention gains popularity for capturing the associations when processing images [154, 157–159].

We employ patch attention [160, 161] to capture the local and global associations for both intra- and inter-modality features, followed by a symmetric fusion function that averages the cross-attented features to model the symmetric interaction of the source pair of features.

Symmetric intra-modality attentions. For each 3D point cloud patch $\mathbf{P} \in \mathbb{P}$, we employ intra-modality attention by applying the symmetric fusion function $\Psi^*(\cdot,\cdot)$ to encode the interrelationship of geometry and texture features. For simpler notation, we assign a random sequence for the patches and arrange the set of the features extracted features $\mathbf{h}_{\text{3D}}^{\text{geo}}$ and $\mathbf{h}_{\text{3D}}^{\text{tex}}$ for all patches in forms of matrices as $\mathbf{H}_{\text{3D}}^{\text{geo}} \in \mathbb{R}^{n\times d}$ and $\mathbf{H}_{\text{3D}}^{\text{tex}} \in \mathbb{R}^{n\times d}$.

The 3D intra-modality attentive fusion becomes

$$\mathbf{H}_{\text{3D}}^{\text{intra}} = \Psi^*(\mathbf{H}_{\text{3D}}^{\text{geo}}, \mathbf{H}_{\text{3D}}^{\text{tex}}) \in \mathbb{R}^{n\times d}. \tag{3.12}$$

$$\mathbf{h}_{\text{3D}}^{\text{intra}} = \textsc{Mean}(\mathbf{H}_{\text{3D}}^{\text{intra}}) \in \mathbb{R}^{d}, \tag{3.13}$$

where $\textsc{Mean}(\cdot)$ is the mean pooling over the sequence dimension to achieve the global feature for the entire point cloud from aggregating all patches in an attentive manner. $\Psi^*(\cdot,\cdot)$ is the symmetric fusion function based on the attention function $\Psi(\cdot,\cdot)$ such that:

$$\Psi^*(\mathbf{x},\tilde{\mathbf{x}}) = \frac{1}{2}\left(\Psi(\mathbf{x},\tilde{\mathbf{x}}) + \Psi(\tilde{\mathbf{x}},\mathbf{x})\right), \tag{3.14}$$

which assumes equal sequence dimensions $l_1 = l_2$ of the Query and Key in the transformer. And $\Psi(\cdot,\cdot)$ is the basic fusion transformer, which is computed by an attentive representation of a target modality referred to a reference modality in the multi-head self-attention.

Similarly for the 2D modality $\mathbb{X}$, we define $\mathbf{H}_{2D}^{geo} \in \mathbb{R}^{m\times d}$ and $\mathbf{H}_{2D}^{tex} \in \mathbb{R}^{m\times d}$, and the 2D intra-modality attention is

$$\mathbf{h}_{2D}^{intra} = \textsc{Mean}(\mathbf{H}_{2D}^{intra}) = \textsc{Mean}(\Psi^*(\mathbf{H}_{2D}^{geo}, \mathbf{H}_{2D}^{tex})) \in \mathbb{R}^d. \tag{3.15}$$

We clarify that the random sequence assignment would not affect the final output feature detailed, since the attention function is equivariant to the permutation of the sequence, and we will average over the sequence dimension to aggregated feature output.

Symmetric inter-modality attention. For inter-modality attentive features, we cross-attend each pair of 3D point cloud patch and 2D projection in the combinatorial set $\{\mathbf{P}, \mathbf{X}\} \in \mathbb{P} \times \mathbb{X}$. We employ the inter-modality attention by applying $\Psi^*(\cdot,\cdot)$ across 3D and 2D modalities. Note that this result can only be achieved when we have the same number of 3D patches and 2D projections $n = m$ for each point cloud. In the rest of this section, we will discard the notation of m and consistently use n for $|\mathbb{P}| = |\mathbb{X}| = 6$ to reduce confusion.

$$\begin{aligned}
\mathbf{H}_{inter}^{geo\text{-}geo} &= \Psi^*(\mathbf{H}_{3D}^{geo}, \mathbf{H}_{2D}^{geo}) \in \mathbb{R}^{n\times d} \\
\mathbf{H}_{inter}^{geo\text{-}tex} &= \Psi^*(\mathbf{H}_{3D}^{geo}, \mathbf{H}_{2D}^{tex}) \in \mathbb{R}^{n\times d} \\
\mathbf{H}_{inter}^{tex\text{-}geo} &= \Psi^*(\mathbf{H}_{3D}^{tex}, \mathbf{H}_{2D}^{geo}) \in \mathbb{R}^{n\times d} \\
\mathbf{H}_{inter}^{tex\text{-}tex} &= \Psi^*(\mathbf{H}_{3D}^{tex}, \mathbf{H}_{2D}^{tex}) \in \mathbb{R}^{n\times d}.
\end{aligned} \tag{3.16}$$

Similar to Eq. 3.13, we apply average pooling $\textsc{Mean}(\cdot)$ to obtain global inter-modality attentive features $\mathbf{h}_{inter}^{geo\text{-}geo}$, $\mathbf{h}_{inter}^{geo\text{-}tex}$, $\mathbf{h}_{inter}^{tex\text{-}geo}$, and $\mathbf{h}_{inter}^{tex\text{-}tex}$ for the entire point cloud.

Feature aggregation. We aggregate all multi-modal geometry and texture features as well as all intra- and inter-modality attentive features for the final feature encoding:

$$\begin{aligned}
\mathbf{h} = &\mathop{\mathbb{E}}_{\mathbf{P}_i\in\mathbb{P}}[\mathbf{h}_{3D,i}^{geo} + \mathbf{h}_{3D,i}^{tex}] + \mathop{\mathbb{E}}_{\mathbf{X}_j\in\mathbb{X}}[\mathbf{h}_{2D,j}^{geo} + \mathbf{h}_{2D,j}^{tex}] \\
&+ \frac{\mathbf{h}_{3D}^{intra} + \mathbf{h}_{2D}^{intra}}{2} + \frac{\mathbf{h}_{inter}^{geo\text{-}geo} + \mathbf{h}_{inter}^{geo\text{-}tex} + \mathbf{h}_{inter}^{tex\text{-}geo} + \mathbf{h}_{inter}^{tex\text{-}tex}}{4}.
\end{aligned} \tag{3.17}$$

The resulting feature $\mathbf{h}$ serves as the input to the decoder heads for final predictions, to be detailed as follows.

3.2.4. Multi-task learning with dual decoders

Dual decoders. We define dual decoders using multi-layer perception for quality regression and distortion-type classification respectively with a regression head $\psi_{regression}$ and a classification head $\psi_{classification}$, both taking the aggregated feature $\mathbf{h}$ as the input. The regression head $\psi_{regression}$ is a two-layer ReLU-MLP that outputs y the quality score:

$$y = \psi_{regression}(\mathbf{h}) \in \mathbb{R}. \tag{3.18}$$

The classification head $\psi_{classification}$ is a three-layer ReLU-MLP with a softmax activation attached to the output layer, which gives z the one-hot prediction of classification type:

$$z = \psi_{classification}(\mathbf{h}) \in \mathbb{R}^c, \tag{3.19}$$

where c is the number of types of distortions.

Learning loss. We define and jointly learn the dual decoders by a triplet learning loss $\mathscr{L}$ for a mini-batch with size of n as:

$$\mathscr{L} = \lambda_1 \mathscr{L}_{\text{mse}} + \lambda_2 \mathscr{L}_{\text{rank}} + \lambda_3 \mathscr{L}_{\text{ce}}, \tag{3.20}$$

where $\lambda_1, \lambda_2, \lambda_3 \in [0, 1]$ are importance scores used to control the proportion of each type of loss.

Specifically, we compute Mean Square Error (MSE) loss between predicted quality scores and human scores as:

$$\mathscr{L}_{\text{mse}} = \frac{1}{n} \sum_{i=1}^{n} (y_i - y_i')^2. \tag{3.21}$$

We compute ranking loss of the predicted quality scores and human scores as:

$$\begin{aligned} \mathscr{L}_{\text{rank}} &= \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} l_{ij}, \quad \text{where} \\ l_{ij} &= \max\Big(0,\ |y_i - y_j| - (-1)^{\mathbb{1}(y_i < y_j)} \cdot \big(y_i' - y_j'\big)\Big). \end{aligned} \tag{3.22}$$

Here i and j are the corresponding indexes for two point clouds in a mini-batch, and $\mathbb{1}(\cdot)$ is the indicator function.

We compute the cross-entropy loss of the predicted distortion type and the ground-truth labels:

$$\mathscr{L}_{\text{ce}} = \frac{1}{n} \sum_{i=1}^{n} -\big(z_i' \log(z_i) + (1 - z_i') \log(1 - z_i)\big) \tag{3.23}$$

3.2.5. Experimental Setup

Datasets. We employ the SJTU-PCQA [39], WPC [81], BASICS [75] and MJ-PCCD [83] datasets for validation.

Comparable methods. We selected 13 state-of-the-art PCQA metrics for comparison, which consist of 5 FR metrics: PCQM [5], GraphSIM [6], PointSSIM [145], MPED [162] and PointPCA [46]; 2 RR metrics: PCM-RR [163] and RR-CAP [110], and 6 NR metrics: 3D-NSS [164], IT-PCQA [53], VS-ResNet [165], MM-PCQA [7], ResSCNN [82] and GMS-3DQA [166].

Implementation details. The proposed M3-Unity is implemented using PyTorch [167]. We use the Adam optimizer [168] with a weight decay of 1e-4, an initial learning rate of 5e-5, and a batch size of 4. The model is trained for 100 epochs. Each point cloud patch has a cardinality k of 2048, the number of local patches and image projections both equal to 6. Projected images have a resolution of 1920×1080, and cropped image patches are 224×224. We use PointNet++ [155] as the point cloud encoder and initialize ResNet50 [156] with a pre-trained model on ImageNet [169] as the image encoder. The multi-head attention module employs 8 heads and the feed-forward dimension is 2048. MOS values are scaled between $[1, 10]$. λ_1, λ_2 and λ_3 are all set to 1. We employ k-fold cross-validation to evaluate performance [44]. We conduct 9/5/6-fold cross-validation for SJTU-PCQA, WPC and MJ-PCCD datasets, respectively, and report average scores.

Table 3.8: **Performance comparison among the proposed and the state-of-the-art Point Cloud Quality Assessment (PCQA) metrics on the 4 datasets.** Best in bold and second with underlined fonts. Please note that the state-of-the-art results were taken from the literature, often with different training strategies and splits, and not independently validated by the authors.

Category	Method	SJTU-PCQA		WPC		BASICS		MJ-PCCD	
		SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
FR	PointSSIM [145]	0.687	0.714	0.454	0.467	0.692	0.725	0.467	0.597
	PCQM [5]	0.864	0.885	0.743	0.750	0.810	0.888	0.779	0.858
	GrahSim [6]	0.878	0.845	0.583	0.616	/	/	0.758	0.844
	MPED [162]	0.898	0.915	0.620	0.618	0.761	0.835	0.735	0.811
	PointPCA [46]	0.907	0.932	0.890	0.894	0.866	0.926	0.834	0.702
RR	PCM-RR [163]	0.482	0.336	0.310	0.343	0.436	0.518	0.497	0.636
	RR-CAP [110]	0.758	0.769	0.716	0.731	0.558	0.740	0.550	0.735
NR	IT-PCQA [53]	0.630	0.580	0.568	0.561	0.310	0.302	0.658	0.807
	3D-NSS [164]	0.714	0.738	0.648	0.651	0.617	0.657	0.446	0.411
	ResSCNN [82]	0.810	0.860	0.735	0.752	0.628	0.682	0.759	0.842
	VS-ResNet [165]	0.830	0.860	0.760	0.770	0.711	0.852	0.526	0.583
	MM-PCQA [7]	0.910	0.923	0.841	0.856	0.831	0.882	0.860	0.898
	GMS-3DQA [166]	0.911	0.918	0.831	0.834	0.855	0.930	0.879	**0.936**
	M3-Unity(Proposed)	**0.947**	**0.961**	**0.900**	**0.900**	**0.872**	**0.937**	**0.903**	0.919

Table 3.9: Cross-dataset validation among 4 datasets. Both the training and testing are on the complete dataset.

	Test											
	SJTU-PCQA			WPC			BASICS			MJ-PCCD		
Train	SRCC	PLCC	RMSE	SRCC	PLCC	RMSE	SRCC	PLCC	RMSE	SRCC	PLCC	RMSE
SJTU-PCQA	–	–	–	0.444	0.473	2.020	0.537	0.671	0.794	0.457	0.701	0.835
WPC	0.821	0.841	1.314	–	–	–	0.617	0.712	0.752	0.643	0.767	0.751
BASICS	0.523	0.559	2.013	0.509	0.514	1.967	–	–	–	0.825	0.867	0.582
MJ-PCCD	0.635	0.653	1.838	0.440	0.507	1.976	0.779	0.827	0.602	–	–	–

For the BASICS dataset, we follow the 60%-20%-20% training-validation-testing split, ensuring no content overlap between training and testing sets. For FR PCQA metrics requiring no training, we assess them on the same testing sets.

3.2.6. Experimental Results

Overall Performance Results of SRCC and PLCC on four datasets for the proposed M3-Unity and other 13 PCQA metrics are shown in Table 3.8. First, M3-Unity significantly outperforms the compared metrics in terms of SRCC on all datasets. Second, compared with GMS-3DQA, which uses the projection-based grid mini-patch sampling only from image modality, the PLCC decreases by 0.017 on the MJ-PCCD. One possible reason is there are super dense/sparse point clouds in MJ-PCCD. Therefore, the projection takes effect when revealing the overlap/hole. While compared with MM-PCQA, which uses 2 modalities from 3D and 2D, M3-Unity is better across 4 datasets, that's because we utilized multi-attributes for both dimensionalities and the interplay among them. In summary, M3-Unity demonstrates robust and competitive performance across 4 benchmarks. This validates our motivation that incorporating multi-attributes in both dimensionalities and the interplay contributes to improved perceptual quality inference.

Table 3.10: Ablation study of M3-Unity on key components, i.e., distortion type, attention, and modality. The numbers in brackets denote the performance of the IB and HA, with the best performance highlighted in blue and orange, respectively.

Settings	SJTU-PCQA			WPC			BASICS			MJ-PCCD		
	SRCC	PLCC	ACC	SRCC	PLCC	ACC	SRCC	PLCC	ACC	SRCC	PLCC	ACC
M3-Unity	**0.947** (**0.933**|0.964)	**0.961** (**0.949**|0.964)	0.728 (**0.583**|**0.795**)	0.900 /	0.900 /	0.981 /	0.872 (0.867|0.889)	**0.937** (0.925|0.929)	**0.847** (0.810|0.840)	**0.903** (0.858|**0.892**)	0.919 (0.908|0.905)	**0.643** (0.545|0.552)
(Distortion Type)												
/wo DT classification	0.938 (0.930|0.963)	0.951 (0.948|0.966)	/	0.898 /	0.898 /	/	0.856 (0.860|0.872)	0.924 (0.916|0.933)	/	0.900 (**0.873**|0.883)	0.924 (**0.917**|0.920)	/
(Attention)												
/wo patch attention	0.919 (0.876|0.946)	0.941 (0.921|0.950)	0.537 (0.446|0.671)	0.849 /	0.855 /	0.969 /	0.684 (0.691|0.777)	0.733 (0.802|0.807)	0.730 (0.525|0.740)	0.846 (0.808|0.853)	0.869 (0.847|0.881)	0.590 (**0.611**|0.587)
(Modality)												
/wo 2D projection	0.914 (0.886|0.947)	0.947 (0.938|0.954)	0.595 (0.542|0.610)	0.608 /	0.638 /	0.792 /	0.770 (0.759|0.771)	0.638 (0.850|0.815)	0.610 (0.565|0.650)	0.736 (0.533|0.776)	0.812 (0.664|0.838)	0.492 (0.462|0.403)
/wo 3D point cloud	0.943 (0.900|**0.967**)	0.957 (0.941|**0.971**)	**0.773** (0.571|**0.795**)	**0.911** /	**0.912** /	**0.989** /	**0.879** (**0.872**|**0.890**)	**0.937** (**0.930**|**0.945**)	0.843 (**0.905**|**0.880**)	0.896 (0.860|0.880)	**0.931** (0.912|**0.936**)	0.624 (0.575|**0.636**)

Cross Dataset Validation To verify the generalization and robustness of M3-Unity, we conduct cross-dataset experiments among all datasets. The results are shown in Table 3.9. From Table 3.9, we can see that M3-Unity has good generalization performance, the cross-dataset performance is even higher than certain FR PCQA metrics, for example, the performance is higher than PointSSIM when training on WPC and testing on SJTU-PCQA (the SRCC of MM-PCQA [7] is 0.769, and the PLCC of CoPA [170] is 0.643) and MJ-PCCD datasets.

Time and complexity analysis We provide the parameter size by dividing the trained neural network into four parts: image encoding (70.5M), point cloud encoding (3.3M), attention (23.1M), and decoding (1.2M). M3-Unity/M3-Unity (3D Point Cloud-Only)/M3-Unity (2D Projection-Only) contain 98.1M/25.4M/97.0M parameters using approximately 37GB/30GB/14GB GPU memory with batch size 4 and has an average inference time of 0.49s/0.44s/0.04s for 1 point cloud from the SJTU dataset on A100.

Ablation Study

We conduct an ablation study on M3-Unity to examine the impact of key components for the performance. Additionally, in the context of the 4 datasets characterized by distinct content and distortion types, we categorized each dataset into Human and Animal (HA) and Inanimate Object (IO) subsets and reported the related performance. Note: WPC only includes IO.

Impacts of distortion type classification. To verify the effect of the distortion type classification module, we compare the performance with only the regression decoder. The result is in Table 3.10 (Distortion Type). Omitting the distortion type classification task causes a slight performance drop across the four datasets. Notably, the prediction accuracy (ACC) of distortion types differs considerably between the WPC and MJ-PCCD datasets. ACC measures the proportion of correct predictions out of the total. There is no discernible correlation between distortion type classification accuracy and quality prediction accuracy with the current datasets.

Impacts of the modalities. Combining 4 modalities improves visual representations compared to unimodal approaches, as shown in Table 3.10. M3-Unity generally outperforms unimodal models, except on the WPC dataset, indicating the contribution of all modalities to perceptual representations. Among the modalities, 2D texture is most crucial for most datasets. However, for the BASICS dataset, 2D geometry performed best (SRCC/PLCC of 0.849/0.911 versus 0.835/0.909). Additionally, image-based modalities are more important than point cloud-based ones, as the HVS prioritizes visual stimuli from images.

Impacts of the attention. The self-attention mechanism calculates semantic affinities between different items in a data sequence [161], i.e., we capture the local context within the point cloud, by enhancing input embedding with the support of FPS and KNN search. Upon removing the attention module, the results are presented in Table 3.10 (Attention). M3-Unity exhibits superiority in comparison to the model without attention.

Our investigation found that M3-Unity and its variants consistently perform better on HA than IO data, as measured by SRCC across all datasets, with HA data numbers equal to or greater than IO for SJTU-PCQA and MJ-PCCD datasets. Specifically, we observed that patch attention predominantly influences performance for the SJTU and BASICS datasets, whereas 2D projection assumes a pivotal role for the WPC and MJ-PCCD datasets within

Figure 3.9: **Point cloud *Unicorn* comparison between learning-based and traditional FR metrics.** The left side shows the reference *Unicorn*, while the right side displays the distorted version with geometry Gaussian noise (points randomly shifted within 0.02%).

Table 3.11: Performance comparison among the proposed metric with different variants on 4 datasets.

Settings	SJTU-PCQA			WPC			BASICS			MJ-PCCD		
	SRCC	PLCC	RMSE	SRCC	PLCC	RMSE	SRCC	PLCC	RMSE	SRCC	PLCC	RMSE
M3-Unity	**0.947**	**0.961**	0.834	**0.900**	**0.900**	**0.989**	**0.872**	**0.937**	**0.375**	**0.903**	0.919	0.643
Texture-Only	0.942	0.956	**0.675**	0.895	0.894	1.021	0.855	0.905	0.457	0.874	**0.927**	**0.413**
Geometry-Only	0.888	0.915	0.948	0.644	0.670	1.692	0.837	0.905	0.677	0.818	0.860	0.561

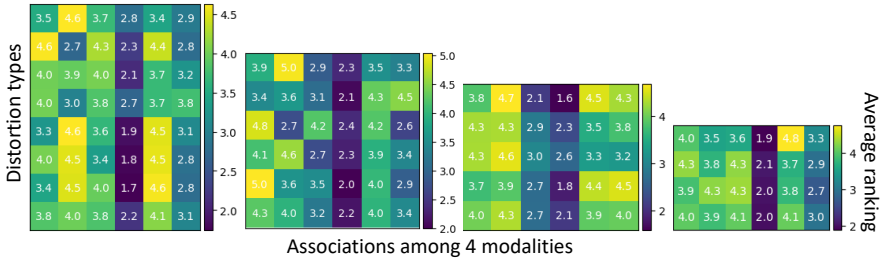


Figure 3.10: Visualization of the 6 associations' average rankings per distortion type across 4 datasets (tex2D_geo2D, tex3D_geo3D, tex2D_tex3D, tex2D_geo3D, geo2D_tex3D, geo2D_geo3D). The result is computed in the same way as described in Sec §3.2.5 *Implementation details*. Lower values indicate higher perceptual quality importance. The datasets in order from left to right are SJTU-PCQA, WPC, BASICS, and MJ-PCCD. The distortion types in order from top to down are as described in Sec §3.2.5 *datasets* and overall ranking.

the framework of M3-Unity, relative to other components. Upon further analysis, we found that excluding the patch attention component resulted in a performance drop of 9.4% for IO data and 6.2% for HA data. Similarly, when excluding the 2D projection component, the performance drop was more pronounced, with reductions of 21.8% for IO data and 9.3% for HA data. Remarkably, IO data consistently exhibited a greater decline in performance compared to HA data across the datasets, except for the BASICS

dataset, where the performance decrement was comparable for both categories.

3.3. Discussion

Generalizability of the proposed metrics At the dataset level, performance is significantly influenced by the total number of point clouds. In extensive datasets encompassing diverse content, prediction accuracy tends to be lower compared to smaller datasets with fewer variations, even when their distortion types are similar. Furthermore, distortion levels play a crucial role in impacting prediction accuracy, with fine-grained distortion proving more challenging than coarse division. In the context of point clouds, compound distortion doesn't necessarily result from the mixture of multiple distortion types, even a single distortion type can concurrently compromise texture and geometry. Hand-crafted geometry and texture features exhibit distinct strengths and weaknesses across various distortion types. Combining both types of features adaptively with distortions may enhance prediction accuracy.

Interplay between geometry and texture. To further explore which distortion representation is allocated more attention when encountering degradations, we predict the quality with geometry-only (3D position, normal point clouds, 2D depth, normal maps) and texture-only (3D texture point cloud, 2D texture map) features, separately. The performance is in Table 3.11.

In addition, we assessed the quality of the distorted point cloud by examining it from both geometry-only and texture-only perspectives in comparison to the reference one. Figure 3.9 illustrates the results obtained by the variants of M3-Unity alongside the results from FR PCQA metrics. Specifically, we use the average of norm and curvature of PointSSIM [145] as the geometry measurement, while Y_PNSR serves as the texture measurement. In the FR manner, Y_PNSR exhibits greater similarity to the reference *Unicorn* point cloud (MOS: 9.117) than geometry, underscoring the predominant role of texture-related representation in predicting the quality of the *Unicorn* point cloud. Notably, our model's prediction (Texture-Only) aligns closely with the distorted *Unicorn* point cloud (MOS: 4.591), indicating that the learning-based model consistently concludes with the FR metric. This verification underscores the significant impact of texture on geometry Gaussian noise.

Interplay among the associations. We've identified 6 association features, to understand their contributions separately, we compared their cosine similarity to the final feature map before decoding [171]. By ranking (round to one decimal place) the features based on similarity, we observed their influence on perceptual quality across distortion types and datasets, as depicted in Figure 3.10, we can draw the following observations: (1) **Mixed Distortion in Colored Point Clouds**: The most important factor for quality is the association between 2D texture and 3D geometry. Following closely is the association of geometry in both dimensionalities (SJTU-PCQA and MJ-PCCD) and texture in both dimensionalities (WPC and BASICS). The importance of the least crucial factor varies depending on the specific distortion type. (2) **Compression**: VPCC and GPCC's quality is least influenced by 3D-related association. V-PCC distorts 2D images due to its

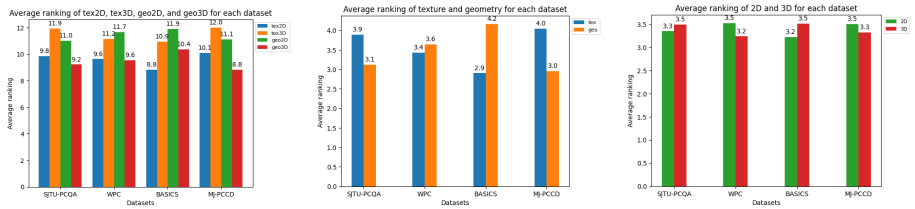


Figure 3.11: Average ranking grouped by different modality, attributes, and dimensionality. Each bar represents a ranking.

3

projection-based coding, while G-PCC follows a geometry-based coding principle, with attribute coding relying on decoded geometry, making the correlations between 3D geometry and 3D texture less effective. (3) **Relative importance grouped by modalities, attributes and dimensionality**: The average ranking of them is shown in Figure 3.11, which is accumulated based on Figure 3.10, assuming one geo2D and geo3D compose the geometry, similarly for texture, 3D and 2D. It shows that 2D texture and 3D geometry are the most influential. Additionally, geometry distortion is more pronounced than texture for SJTU-PCQA and MJ-PCCD, since GPCC and JPEG Pleno in MJ-PCCD dataset can produce super dense/sparse stimuli and with uneven point distribution; SJTU-PCQA has more types of geometric distortion. 3D distortion is more pronounced than 2D for WPC and MJ-PCCD datasets.

3.4. CONCLUSIONS

In this chapter, we address *R1: How to measure the perceptual quality of static point clouds under various distortion types?* by developing both FR and NR PCQA paradigms, with the goal of providing practical solutions for evaluating point cloud quality across a range of real-world scenarios. Whether reference point clouds are available or not, we aim to offer effective strategies for designing appropriate perceptual quality assessment methods.

We propose a PCA-based FR PCQA metric, namely PointPCA+, which relies on an enriched set of lower complexity descriptors with respect to its PointPCA predecessor. After a pre-processing step, features are extracted from both geometric and textural domains. A subset of features is selected to enhance the stability of the model, and a learning-based feature fusion based on ensemble learning is applied to the feature subset, to provide a quality score for a distorted point cloud. Our experimental results demonstrate that PointPCA+ outperforms the majority of existing PCQA metrics, reaching second place in Track#1 and third place in Track#3 and Track#5 of the ICIP 2023 PCVQA grand challenge.

Compared to other teams in the PCVQA grand challenge, PointPCA+ has the highest computational complexity, with a processing time of 1000 ms, while the lowest reported is 8.6 ms. This inefficiency stems from the point-based FR PCQA framework, especially the PCA decomposition applied to each point. One solution to reduce the computational load is to downsample the point cloud before processing, while maintaining its overall structure. Additionally, PointPCA+ mainly focus on local features, neglecting the broader

geometric and texture context of the point cloud. To improve, incorporating global features that capture more comprehensive information is essential. This will lead to a more efficient and holistic characterization of point clouds, meeting the demands of the PCQA field. These additions are poised to contribute to a more holistic and efficient characterization of the point cloud, aligning with the rigorous demands of the PCQA field.

In the proposed second metric **M3-Unity**, we introduce an NR framework designed for evaluating the quality of colored point clouds across multiple modalities and tasks. The self-attention mechanism is employed to fuse modality-related features, therefore enhancing the feature representations for quality assessment. Our framework enables a comprehensive measurement of the contributions stemming from both inter- and intra-associations, particularly concerning distinct distortion types relevant to perceptual quality assessment. In our investigations, we discovered that relying solely on 3D positional data may not suffice for accurately gauging geometric distortion, and the interplay between the attributes is crucial in understanding the overall distortion. We observed notable performance improvements by incorporating additional geometric information such as surface normals and association features. Furthermore, we draw conclusions about the prioritization of geometry/texture for point cloud quality assessment, providing valuable insights for bit allocation in point cloud compression and various high-level computer vision tasks.

In the next chapter, we present subjective quality evaluation protocols for dynamic point clouds, with a particular focus on visual saliency detection in VR environments. By collecting user feedback, we aim to better understand human perceptual behavior when interacting with dynamic point clouds in immersive settings. Through both qualitative and quantitative analyses, we aim to explore two key aspects: how distortions impact perceived quality, and how different task scenarios influence visual saliency patterns. Insights gained from these user studies will deepen our understanding of human visual behavior and inform the design of more perceptually aligned point cloud saliency map similarity metrics. Furthermore, these findings will guide the development of more effective feature extraction strategies for objective PCQA metrics, providing an opportunity to validate the relevance and effectiveness of the features introduced in this chapter.

4

Visual Saliency of Point Cloud

In the previous chapter, we presented two objective PCQA metrics for static point clouds and analyzed the contributions of geometric and textural attributes to perceived quality. Building on this foundation, the current chapter shifts focus to the subjective evaluation of dynamic point clouds, with particular emphasis on visual saliency. Our goal is to understand how visual saliency patterns evolve across different distortion levels and task conditions within immersive environments. To this end, we conducted two complementary subjective experiments. In the first experiment, participants performed a quality assessment task, enabling us to examine how variations in point cloud quality influence visual saliency. In the second experiment, the same stimuli were presented under a free-viewing condition, allowing us to investigate how task demands affect the distribution of visual saliency. We further introduce a distribution-based point cloud visual saliency map similarity metric to quantitatively compare saliency patterns. The contributions of this chapter include quantitative results comparing visual saliency under varying tasks, statistical analyses across both content and user dimensions, and qualitative insights into the factors that guide user attention and influence quality perception. This work deepens our understanding of visual saliency in dynamic point clouds and offers valuable implications for adaptive delivery and optimization in VR-based remote communication.

This chapter is based on the following publications:

1. ***Xuemei Zhou**, Irene Viola, Evangelos Alexiou, Jansen, Jack, and Pablo Cesar. 2023. QAVA-DPC: Eye-Tracking Based Quality Assessment and Visual Attention Dataset for Dynamic Point Cloud in 6 DoF. 2023 IEEE International Symposium on Mixed and Augmented Reality (IEEE ISMAR).* [24]
2. ***Xuemei Zhou**, Irene Viola, Silvia Rossi and Pablo Cesar, 2025. Comparison of Visual Saliency for Dynamic Point Cloud: Task-free vs. Task-dependent. IEEE Transactions on Visualization and Computer Graphics (IEEE TVCG).* [25]

The rapid evolution of immersive media technologies has shifted the spotlight toward 3D content, with volumetric video—particularly Dynamic Point Clouds (DPCs)–emerging as a prominent format for interactive experiences [172]. A DPC is essentially a sequence of individual point cloud frames played in succession. DPCs are increasingly used in various applications, including automotive/robotic navigation [173], medical imaging [174], virtual video conferencing [175, 176], among others. However, each point cloud frame requires a large number of points to faithfully represent the content and achieve a good QoE. Therefore, effective compression is essential before transmission, storage, rendering, and display. Quality degradation will inevitably be introduced during this end-to-end pipeline, which deteriorates the visual quality and affects the perception. Measuring and understanding these distortions–especially in 6 DoF—is still a challenge for both subjective and objective quality assessment [172].

Most existing studies on DPCs rely on 2D displays, where pre-recorded sequences are viewed along fixed trajectories [84, 117, 118]. While this simplifies experimentation, it restricts user interactivity. On the other hand, HMDs with 6DoF support provide users with immersive and photorealistic interactions. However, such immersive studies typically involve a smaller number of sequences (approximately 20) that are often static or very short in duration (e.g., around 5 seconds), due to technical limitations associated with real-time rendering [119].

Despite increased interest in point cloud processing, no dedicated visual attention datasets for DPCs have been released to date. Most prior work focuses on static point clouds [14], often using a limited number of undistorted models. There remains a notable research gap in connecting visual attention and perceived quality in the context of DPCs.

The HVS efficiently processes complex visual scenes by selectively focusing attention on salient regions—a phenomenon known as *visual saliency* or *visual attention*. This process allows for efficient scene interpretation and has been extensively studied in 2D images and videos [177–182]. Research has shown that task-driven viewing—such as during quality assessments—can significantly influence gaze behavior compared to free viewing [106, 112]. However, these findings may not directly apply to 3D DPCs due to their higher data complexity and the different nature of interaction in HMD-based environments. For instance, spatial and central fixation biases observed in 2D media [183, 184] may not translate to immersive DPC experiences.

A few efforts have been made to collect gaze data for point clouds. Alexiou *et al.* [14] conducted an eye-tracking experiment in VR under a task-based condition, while Nguyen *et al.* [23] released an open-source, task-free eye-tracking dataset for four dynamic point clouds in mixed reality using the Hololens 2. Table 2.2 provides a summary of these existing datasets. However, the impact of task-driven versus task-free conditions on visual attention in DPCs has yet to be systematically studied. Unlike in 2D, no dataset to date captures the same DPC content across different perceptual tasks in immersive environments.

To address this gap, we present a new contribution in this chapter: the creation of two complementary datasets that enable the exploration of visual attention in task-driven and task-free scenarios in immersive VR: **1. Task-Dependent Dataset (QAVA-DPC)**: This dataset includes eye-tracking and head movement data from 40 participants evaluating the visual quality of 50 DPCs (5 reference and 45 distorted versions) in a 6DoF VR setup.

Distortions were introduced using three different codecs at three quality levels. Heatmap-based saliency maps were generated for each sequence, connecting visual attention with perceived quality.

2. Task-Free Dataset (TF-DPC): This second dataset contains gaze and head movement data from 24 participants who freely explored 19 reference DPCs in the same VR environment. Five of the DPCs overlap with the task-dependent dataset, enabling direct comparison between viewing conditions.

To quantify the differences in visual saliency across conditions, we employ both Pearson's Correlation Coefficient (PCC) and a modified Earth Mover's Distance (EMD) metric [185]. Our analysis reveals significant shifts in gaze behavior based on task context. For instance, Figure 4.1 illustrates how visual attention focuses more consistently on detailed facial regions during quality assessment tasks compared to free viewing of the same content.

To conclude, our contributions are threefold and can be summarised as follows:

- We propose two new datasets. In the task-dependent scenario. We create QAVA-DPC, which contains 5 reference dynamic point clouds; each DPC is encoded by 3 codecs, with each codec configured at 3 distortion levels. Fixation maps are constructed, collected, and presented for both the reference and distorted sequences as heatmaps overlaid on top of the stimuli frames. To the best of our knowledge, this is the first time connecting visual attention and visual quality for DPC in VR. We create the other visual attention dataset for 19 original dynamic point clouds in a task-free VR experiment with 6DoF.

- We release all raw data, containing the gaze samples and movement trajectory collected in our study, along with the code to compute and compare the dynamic point cloud visual saliency maps, and the software used to perform the experiment, at the following links: `https://github.com/cwi-dis/ISMAR_PointCloud_EyeTracking`, `https://github.com/cwi-dis/TVCG2025-TaskFree_PointCloudEyeTracking`.

- We provide an in-depth analysis of the collected dataset, using quantitative measures to explore the dataset in terms of gaze and trajectory; furthermore, we use qualitative methods to draw further insights from interviews.

- We perform a cross-condition comparison of saliency maps to evaluate how perceptual tasks affect visual attention in immersive DPC environments.

These novel datasets offer valuable opportunities for developing reliable saliency models for 3D representations, which are essential for augmented and mixed reality applications [186, 187]. For instance, they can enable advancements in several areas, including saliency-guided compression [188, 189] and live reconstruction [190] for point cloud streaming, saliency-aware point cloud mixup for data augmentation [12], volume visualization [191], foveated rendering [192], point cloud transmission [190] and visual quality assessment [109, 110, 193].

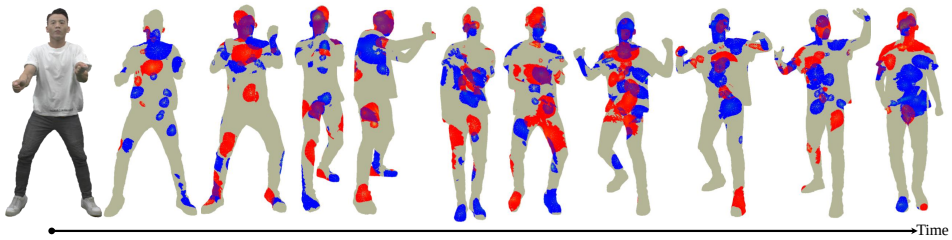


Figure 4.1: Fixation maps of *dancer* sequences with uniform temporal sampling every 30 frames. The blue regions represent task-free conditions, while the red regions indicate task-dependent conditions. Gray areas denote nonsalient regions in both conditions, and overlapping areas are shown as a blend of the two colormaps. The visualization corresponds to an average fixation map computed across two user studies, involving 40 participants in the task-dependent experiment and 24 participants in the task-free experiment, respectively.

4.1. QAVA-DPC Dataset Construction: Task-Dependent

To the best of our knowledge, no dataset has yet been released for visual attention of DPCs, which is our main contribution of the QAVA-DPC dataset. In our first study, we aim to complement existing literature by performing an experiment comparing the visual quality of several state-of-the-art compression techniques for DPC. We do so in an interactive manner, using an HMD-based VR rendering of 10s sequences from various datasets, which has not been done before in the literature in combination with eye-tracking, so we can better understand whether the findings regarding visual saliency and quality assessment on 2D images/videos can hold for volumetric contents in VR.

4.1.1. Stimuli Selection

For the creation of the dataset, we selected 5 dynamic point clouds from 3 public datasets, namely VsenseVVDB2 [117], 8i [194] and Owlii [118]. To show the diversity of dynamic point clouds, we considered the Spatial Information (SI) and Temporal Information (TI) for each content [195]. We projected the source point cloud into 4 views, which are the left, right, front, and back view, of its bounding box to apply SI and TI separately, then obtain the maximum value among the 4 views over all the first 300 frames as the final SI/TI for one sequence. The distribution of all dynamic point clouds can be seen in Fig.4.2, we finally choose *dancer*, *exercise*, *long dress*, *rafa2*, and *soldier* as the contents in our dataset. The dispersed state in SI (horizontal axes)/TI (vertical axes) shows the diversity of our contents in the spatial/temporal domain.

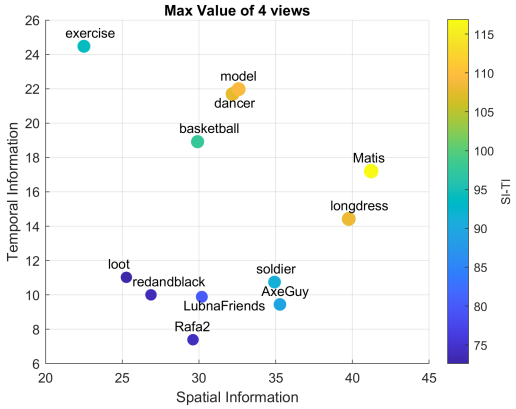


Figure 4.2: Distribution of SI and TI of 12 source dynamic point clouds from 3 datasets, the color value is computed by $\sqrt{(SI^2 + TI^2)}$

.

4.1.2. Stimuli Processing

Before conducting the subjective experiment on DPC, specific procedures are necessary due to the codec implementations. These procedures, including pre-processing, encoding, and rendering, are aimed at minimizing additional influencing factors.

Pre-processing

The sequences mentioned above are selected from different datasets, which means the resolution, position and orientations vary. The dynamic point clouds should be life-size so to create a realistic tele-immersive scenario. To do so, we normalize the dynamic point clouds to a similar bounding box. The geometry precision of *dancer* and *exercise* is voxelized from 11 to 10. The source models are processed with rotation, translation, and scaling. Additionally, since the VPCC encoder fails to deal with decimals, the coordinates of dynamic point clouds are rounded before VPCC compression. CWI-PCL encoder has specific requirements for the resolution of dynamic point clouds, so before CWI-PCL compression the coordinates go through the scaling operation.

Encoding

Distorted versions are generated using the state-of-the-art MPEG PCC reference software Test Model Category 2 Version 18 (TMC_2V-18.0) and Category 1&3 Version 14 (TMC_{13}V-14.0) from now on referred to as VPCC and GPCC [85]. We also adopt the CWI-PCL [30] codec as a comparison. GPCC is proposed mainly for the aim of compressing static point cloud, VPCC is developed for DPC compression, and CWI-PCL is mainly used to comply with real-time requirements. To compare them in a fair way, we set the GPCC encoder with Region-Aptive Hierarchical Transform (RAHT) to compress point-wise color attributes and Octree for geometry representation; the VPCC encoder with All Intra (AI) mode, which adapts intra-prediction for one frame; and the CWI-PCL

intra frame, geometry coded with octree subdivision and color attributes encoded based on JPEG.

To define the configuration parameters for the encoders, the MPEG Common Test Conditions [196] are followed. To compare different codecs and different distortion levels, we select the distortion levels that can reveal a similar low-medium-high quality range among the 3 codecs. Specifically, for GPCC we select three distortion levels, namely R02, R04, and R05, by setting *positionQuantizationScale* and QP parameters. For VPCC, we select three distortion levels, namely R01, R03, and R05 by setting different geometry QP, attribute QP, and *occupancyPrecision* parameters. For CWI-PCL, we choose three combinations of octree depth with JPEG QP parameters to match a similar quality range, by looping over octree depth from 7 to 9 and JPEG QP from 25 to 95 (step size = 10). When testing on the dataset, the above parameter settings for the three codecs yielded subjectively similarly from the perspective of the quality range. Specific parameter settings are shown in Table 4.1. Each DPC has 3 compression codecs, and each codec has 3 distortion levels, for a total of 45 distorted dynamic point clouds.

Rendering

Rendering is the process of producing a visual representation that can be consumed by users using an available display. In the case of point clouds, different rendering methods have a significant impact on perceived quality [142]. In our experiment, the point clouds were rendered using a point-based rendering approach, in which each point is directly visualized using its original color attributes. No surface reconstruction, mesh generation, or geometry-altering post-processing (e.g., normal estimation or shading based on reconstructed surfaces) was applied prior to rendering, in order to preserve the original point cloud data and avoid introducing rendering-induced artifacts.

Our experiment software is developed in Unity (version 2021.3.10.f1), exploiting the SteamVR plugin (version 1.24.7) to connect with VR headsets and controllers. CWI Point Cloud (CWIPC) supported unity package (version 0.10.0) helps us import the dynamic point clouds and playback them inside Unity [197]. A high-level diagram indicating the hardware/software dependencies is provided in Fig.4.3. Notably, a large size of DPC file might take up too much memory and cause a system hang. So we first transform the DPC data to CWIPC-supported point cloud playback format to improve the software stability. To ensure smooth playback of DPC, we take advantage of the Unity Coroutine scheme to preload each DPC into memory before the user switches to next DPC. 5 dynamic point clouds with their corresponding operation are selected in our test.

It should be noted for each sequence, we only choose the first 300 frames from the source model. The frame rate for playback is 30 frames per second, hence each video lasts for 10 seconds. We use HTC Vive Pro Eye devices with eye-tracking capabilities and Vive hand controllers for participants to interact in our experiment. To develop eye-tracking applications for the Vive Pro Eye we use the native HTC Vive SRanipal SDK. The sampling frequency (binocular) of the eye tracker is 120 HZ.

For the same stimuli, both reference and distorted versions are watertight by adjusting the point size to the average distance among its 10 nearest neighbors all over all points in the point cloud [119]. Within a DPC, we utilize the same point size for all frames. All the point clouds are rescaled to a similar size, around 1.8m in height, to mimic a realistic

Table 4.1: Parameter sets for the selected encoders

Encoders	Distortion Level		
GPCC (Octree-RAHT)	R02 (0.125, 46)	R04 (0.5, 34)	R05 (0.75, 28)
VPCC (AI)	R01 (32,42, 4)	R03 (24, 32, 4)	R05 (16, 22, 2)
CWI-PCL	R01 (7, 25)	R02 (8, 95)	R03 (9, 95)

tele-immersive scenarios. The VR scene is illuminated by a virtual lamp on the ceiling centralized above the models. The lamp is set as an area light with intensity values of 2 in Unity to simulate ordinary lighting in a room.

4.1.3. Experimental Procedure

Since there is no specific recommendation for designing subjective quality assessment experiments for DPC in VR, we have made an effort to adhere to existing ITU recommendations on images/videos [66, 198, 199] to establish an appropriate assessment methodology for DPC. In our subjective study, we opted for Absolute Category Rating with Hidden Reference (ACR-HR). Each time only a single DPC was shown to the observer; test materials included impaired DPC with randomly inserted intact Hidden Reference (HR) sequences, represented as any other test stimulus. To avoid the effects of contextual or memory comparisons, we randomly generated a playlist for each subject, and care was given to avoid displaying the same DPC model consecutively.

Before the experiment, the visual acuity and color vision of every subject was tested using Snellen [200] and Ishihara [201] charts. Each subject was informed in advance about the manner and purpose of the study as part of the informed consent procedure. At the beginning of the session, the inter-pupillary distance was measured and the headset was adjusted by the subject accordingly. Then, a training session was conducted to help familiarize the subjects with the system, including the controllers and the naming of each button to communicate more easily. One training point cloud sequence, namely *loot*, was used, which was not included in the dataset. The quality range of *loot* was similar to the quality range of the test videos, giving the subjects a sense of what they would see in the formal sessions. The subjects always started at the same location, which is 1.5 meters away from the center of the virtual room, but could move freely from there onward.

A DPC was located in the center of the virtual room, and each DPC was randomly rotated between $[0°, 360°]$ to avoid bias. During the experiment, the subjects were instructed to view each model carefully in the VR environment, by moving freely during the playback of each DPC. The subjects were also required to stand still while doing the calibration and error profiling. A fixed distance was set between the subjects and the error profiling scene, which was a circle composed of 9 eye-ball markers.

After feeling comfortable with the set-up, the participants were informed about the task that is assigned to them: “we ask you to examine a set of human DPC models, each model

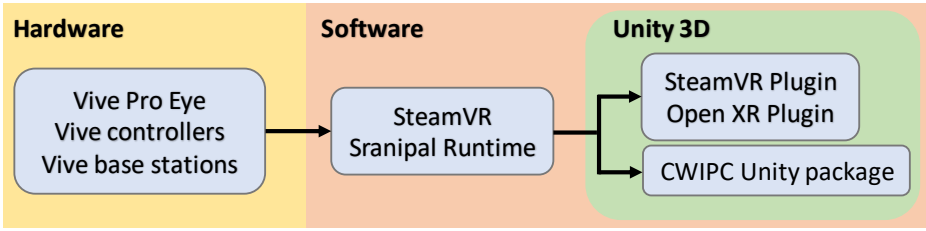


Figure 4.3: Schematic diagram with the hardware and software modules together with their inter-dependencies.

will be looped three times, each loop is last for 10 seconds; after visualization, we will ask you to rate the quality of the stimuli you are looking at, and in the same time, we will record your gaze-related data". To determine the number of loops, we referred to related papers on video quality assessment and eye-tracking-based visual saliency detection [112, 202–204]. Additionally, in [205], the effect of exposition time by repeating the same video from 1 to 4 loops was explored, concluding that more loops do not necessarily result in more unique fixation points for most videos. Hence we chose 3 loops instead of once or an unlimited number. There were two dummy objects at the beginning of each session to familiarize the subject with the testing procedure and the rating scales. For each subject, the test was split into two rounds, lasting for around 30 minutes each, with a mandatory 5-minute break in between. Before and after each round, participants were requested to fill in a Simulator Sickness Questionnaire (SSQ) on a 1-4 discrete scale (1=none to 4=severe) [206]. For every model and subject, a round was split into four consecutive steps:

1 *Calibration* is the process by which the geometric characteristics of a participant's eyes are estimated as the basis for a fully customized and accurate gaze point calculation, which is implemented to optimize the eye tracking algorithm. Calibration was done at the beginning of the experiment, and only when calibration was successful users could enter into the DPC playback stage.

2 *Inspection of models* is the step where the participants are observing DPC, while their trajectory and gaze-related information are recorded.

3 *Quality evaluation of models* requires the subject to rate DPC. The rating button was marked with labels ranging from "Poor" to "Excellent" to facilitate anchoring the rating process, and subjects could use their controllers to select and submit a score without taking off the headset.

4 *Error profiling* is issued as the last step in order to estimate the accuracy of the gaze measurements due to calibration inaccuracies, or HMD displacements. A regular circle of 9 markers at pre-defined positions in the virtual scene was presented to the user. Based on the recorded gaze measurements, the average angular error was computed per marker online. This procedure allowed us to decide whether the gaze data obtained from a certain session was accurate or not.

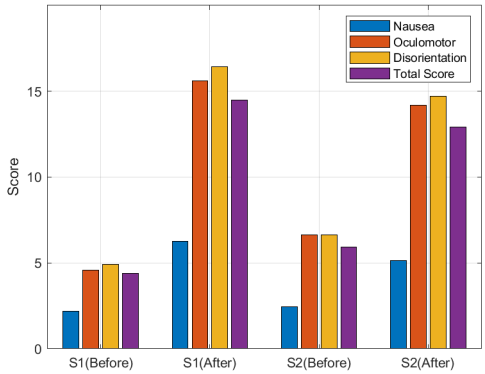


Figure 4.4: SSQ score for two test sessions

A total of 40 participants took part in the subjective tests of this study, with a diverse composition that includes 1 non-binary individual, 19 males, and 20 females. The participants' ages ranged from 20 to 34, with an average age of 26.90 and a standard deviation of 3.51. Each participant observed half of the dynamic point clouds among all stimuli, leading to 20 opinion scores per sequence.

In terms of occupation, the majority (80%) of the participants were students, ranging from bachelor to PhD levels. The remaining 20% were researchers, postdoctoral fellows, and one auditor. Regarding familiarity with VR devices, 7 participants had never experienced VR before the experiment, 26 participants had intermediate experience (using VR 1 to 3 times), and 7 of them were considered experts, having backgrounds as VR designers or researchers. Additionally, 22 out of 40 participants wore glasses during the experiment.

4.1.4. Data Processing

Processing of SSQ Data

SSQ comprises 16 symptoms which are further grouped into three different categories: Oculomotor, Nausea, and Disorientation; we also computed the total score. Fig. 4.4 suggests that simulator scores are increasing after performing the experiment. However, it can be seen that breaks help in reducing simulator sickness.

Processing of Opinion Scores

After removing the scores of the first two dummy objects, outlier detection was performed according to ITU-T Recommendations P.913 [199]. The recommended threshold values $r_1 = 0.75$ and $r_2 = 0.8$ were used. No outliers were found in our test. After outlier detection, the MOS was computed for each DPC. The associated 95% Confidence Intervals (CIs) were obtained assuming a Student's t-distribution. Additionally, the DMOS was obtained by applying HR removal, as the difference between the mean MOS of the hidden reference and the MOS of the distorted point cloud, following the procedure described in

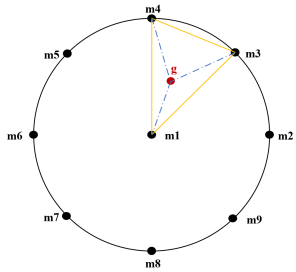


Figure 4.5: Estimation of the angular error for one gaze data

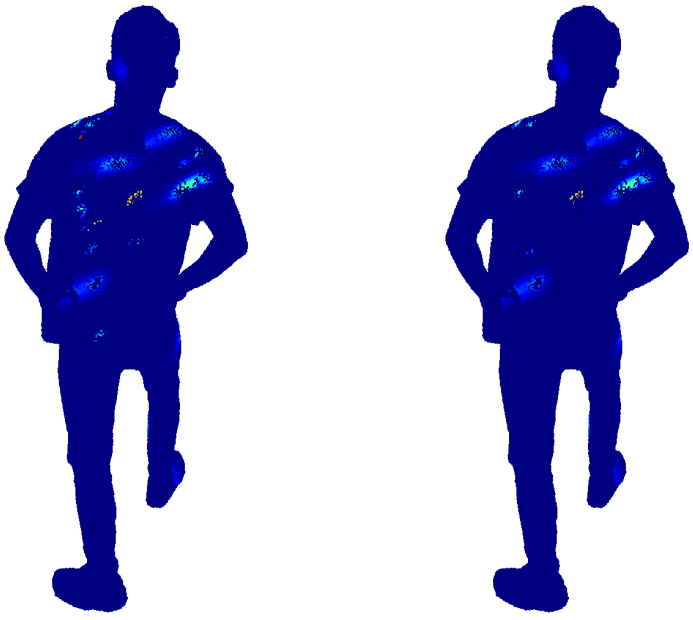
ITU-T Recommendations P.913 [199].

Processing of Gaze Data

One subject walked into the body of two dynamic point clouds in the VR environment when observing, so the corresponding gaze data was not included. We ignored the initial 400 ms gaze data of each user to avoid unintentional gaze because of the unexpected appearance of the DPC. Then, only the valid gaze samples provided by the SRanipal SDK were selected.

1 *Verify the data validity of gaze data:* A barycentric interpolation with weights equal to corresponding angular errors obtained from the profiling was applied. A threshold of 7.5° was used to discard unintentional gaze. After displaying each target, 0.8 seconds will be waited before including the samples in actual calculations. This delay accounts for the initial moments in eye-tracking data during the actual experiment, which can be influenced by factors such as calibration stabilization, participant adaptation, and gaze analysis during fully engaged periods [207]. We used GazeMetrics [208], an open-source tool for measuring the data quality of HMD-based eye trackers, to compute the angular error. Finally, we applied a compensatory weighted average angular error to each gaze sample. This was repeated for every user gaze sample to maintain high-quality estimations while avoiding discarding useful data. Fig.4.5 illustrates the estimation of angular error for gaze data in 2D, g represents the intersection between the gaze ray and the plane formed by nine markers denoted as $m1$ to $m9$. These markers were positioned at a distance of 1.25 meters relative to the camera within the VR environment.

2 *Identifying fixation points of gaze data:* Taking into account the dynamic nature of our content, we chose the Dispersion-Threshold Identification (I-DT) [209] method. I-DT leverages the fact that fixation points, owing to their reduced velocity, tend to cluster in close proximity [210]. It identifies fixation points as groups of consecutive points within a particular dispersion, or maximum separation. The I-DT algorithm requires two parameters, the dispersion threshold and the duration threshold. We set the dispersion threshold equal to 3° and the duration threshold equal to 100 ms, separately. Thus we took the average of these gaze points within the duration threshold as the fixation point.



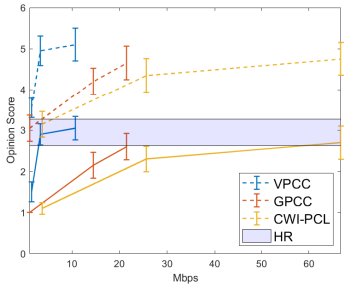
(a) Fixation points of *dancer*, front view (original fixations)

(b) Fixation points of *dancer*, front view (with filtering)

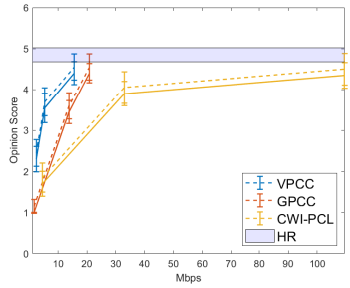
Figure 4.6: Fixation map comparison with/without filtering by DBSCAN.

3 *Mapping gaze data to DPC frames:* We proceeded by associating the filtered gaze data with the currently viewed frames and translating the gaze data (x, y, z) from world space into fixation points within that corresponding frame. As a result, we got all the gaze data in an endeavor to cover 300 frames in total. We adopted the truncated-cone-sector algorithm to assign weights to points in a given dynamic point cloud frame [14].

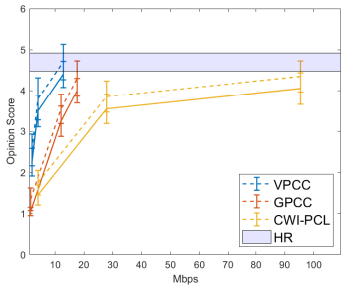
4 *Fusing multiple users' gaze data to dynamic point cloud frames:* A fixation map is the aggregation of fixation points from all users viewing the same dynamic point cloud frame at a given timestamp, which can mark the region of interest. In our experiment, unintentional observation could cause isolated fixation points on dynamic point cloud frame after mapping. Thus, it was necessary to filter out these noisy fixation points. We choose the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [211] algorithm to filter out noisy fixation points. Based on the density of fixation points on the point cloud, the DBSCAN is configured to remove the noisiest fixation points in clusters with high density at the same time be able to retain certain core fixation points in clusters with less density) [101]. Fig.4.6 illustrates the effect of filtering noisy fixations. DBSCAN requires two parameters: ϵ is the radius of the circle to be created around each data



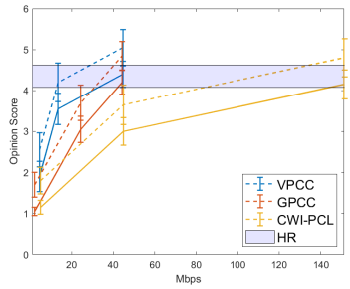
(a) *rafa2*



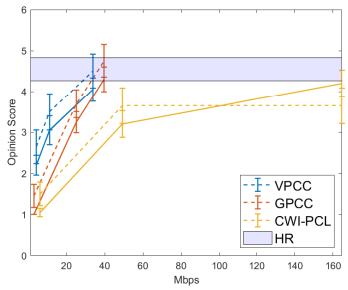
(b) *dancer*



(c) *exercise*



(d) *long dress*



(e) *soldier*

Figure 4.7: MOS (solid line) and DMOS (dashed line) against bit-rate, expressed in Mbps. HR scores are shown using a shaded purple plot.

point to check the density and θ is the minimum number of points required inside that circle for that data point to be classified as a core point. θ should increase as the point size α of a point cloud becomes small, which means a high-density point cloud. The minimum number of points is computed as

$$\theta = \left[\frac{2^7}{1 + 20 * \alpha} \right]. \tag{4.1}$$

ϵ is decided by k-distance graph) [212]. We took the average of fixation maps generated by multiple users, which is defined as

$$VS_f = \frac{1}{N}\sum_{n=1}^{N}(VS_n). \tag{4.2}$$

where VS_f is the fixation map for each dynamic point cloud frame, VS_n is the fixation map for each dynamic point cloud frame by one subject, specifically, VS_n also takes the average number of times a frame is viewed by one subject. N denotes one specific frame that has been viewed by N subjects in total. After we got the averaged fixation map for one dynamic point cloud frame, we applied the DBSCAN filtering operation to get the final fixation map.

4.1.5. Experimental Result

Analysis of Opinion Scores

Fig.4.7 shows the results of the subjective quality assessment of the contents in 6DoF viewing conditions. In particular, the MOS scores associated with the compressed contents are shown with solid lines, along with relative CIs, whereas the dashed lines represent the respective DMOS scores. The HR scores for each content are represented with a solid line to indicate the MOS, and a shaded plot for the corresponding CIs.

While evaluating the point cloud codecs, we observe that, under similar bitrates, VPCC codec exhibits the best perceptual quality, GPCC the second, and CWI-PCL is the last codec for all 5 contents. This observation is consistent with [117, 119]. From the perspective of contents, MOS and DMOS present similar trends, as the MOS for the HR contents is between 4 and 5. The only exception to the trend is *rafa2*, for which the MOS for the reference content remains at around 3. This is likely related to the reconstruction error: compared with other contents captured in more professional studio settings, the reference version of *rafa2* does not offer a satisfactory quality. The calculated DMOS is between [3, 5], due to the fact that the reference content was rated so low.

Analysis of Gaze Data

To understand how and what users explore dynamic point cloud in VR, we analyze the relationship between fixations and bitrates. Fig.4.8 represents the number of fixations of each subject on each content. It should be noted that the fixations are the filtered ones on individual dynamic point cloud instead of the raw fixations of subjects. Fig.4.9 depicts the exact number of fixations across all subjects on different bitrates. Combining Fig.4.2, 4.8, 4.9, we have the following observations:

- Subjects are more interested in the high-motion dynamic point cloud (i.e., with higher TI) compared with the low-motion one. For example, the average number of fixations on *dancer* and *long dress* is higher than *rafa2* and *soldier*, which have less TI on average.
- There is no indication that low-quality contents will receive less attention. In fact, we do not observe any particular trend regarding the number of fixations changing

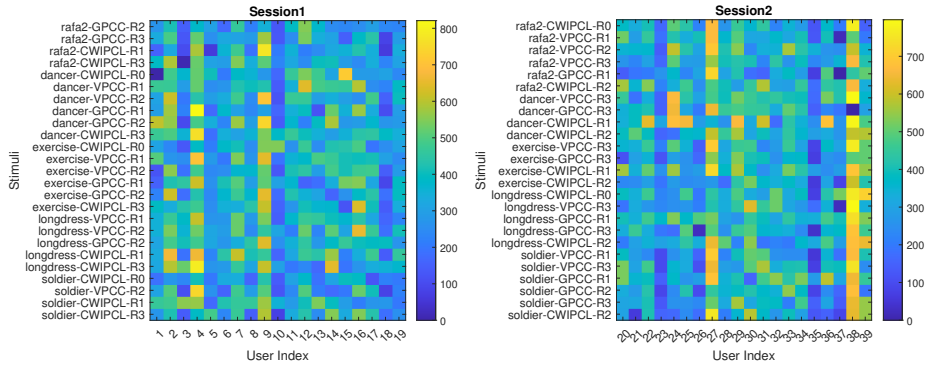


Figure 4.8: The fixations for each subject and for each content. Each row denotes the fixations on a specific content and each column denotes the fixations for each subject, respectively. $R1$ (low), $R2$ (medium), and $R3$ (high) indicate the bitrates of each codec, while $R0$ denotes the reference dynamic point cloud .

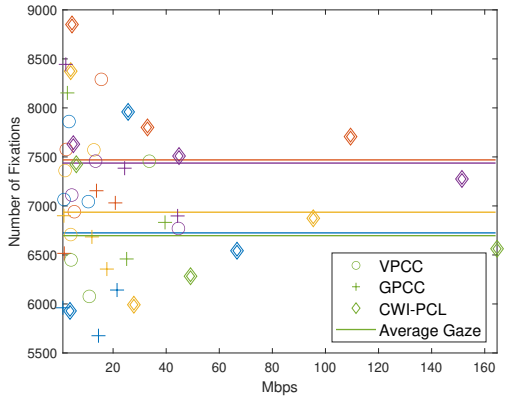


Figure 4.9: Fixations against bitrates, expressed in Mbps. The average number of fixations is expressed with a line. Each color denotes a content, specifically, *rafa2* is in blue, *dancer* is in red, *exercise* is in yellow, *long dress* is in purple and *soldier* is in green.

with varying quality levels. Visual attention for dynamic scenes should be considering both motion and quality.

- Certain subjects consistently exhibit a higher number of gaze fixations (e.g., user 27 and 38 in Fig. 4.8), possibly due to the individual differences of the participants or the accuracy of the device during the experiment.

We also explore where the subjects are looking at the dynamic point cloud in VR, and how the quality degradation will impact the visual attention in a dynamic scene. Subjects pay attention to unrealistic rendering artifacts, such as high-heeled shoes and hair of *long dress*. Fig.4.10 depicts the fixation map on these two areas. Certain frames miss the heelpiece; certain frames exhibit unnatural hair rendering. Fig.4.11 shows the fixation

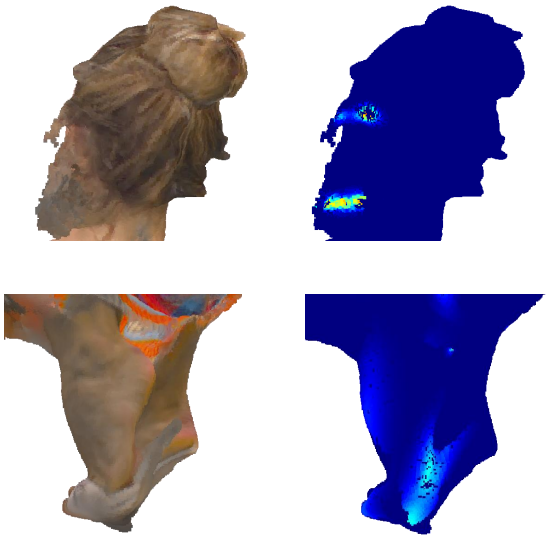


Figure 4.10: Fixation map on the hair and heal of *long dress*

map of both the reference and all distorted *long dress* point cloud frames. We can see subjects are interested in the face and the area with high motion. For all 5 contents, subjects tend to focus on the faces and the front view of the dynamic point cloud , despite the random rotation of the dynamic point cloud itself. No differences are observed for the salient area in different distortion levels. The fixation intensity on the face is consistent across all the distortion levels; the fixation intensity in high-motion areas floats with the motions; the fixation intensity on the remaining area has no pattern. This randomness may come from unintentional fixation or the random rotation of dynamic point clouds.

4.1.6. Qualitative Results

32 interview audio recordings were transcribed into texts and coded using Dovetail[1], with eight participants declining to participate. Following Maguire's guideline on thematic analysis [213], we initially reviewed and labeled the text, organized these labels into themes, and subsequently convened to establish the connection between perceptual quality and visual attention during the subjective test. Each participant is denoted as P1-P32, with the number of participants concurring with each statement indicated in parentheses. The qualitative results of this thematic analysis are presented in this section, with detailed findings discussed in the following subsections.

Factors that Impact Quality Assessment and Visual Attention

Temporal information Participants (13) pointed out that the flickering effect is the most bothersome artifact in our DPC playback scene, often leading to lower ratings. (p21:

[1] https://dovetail.com/

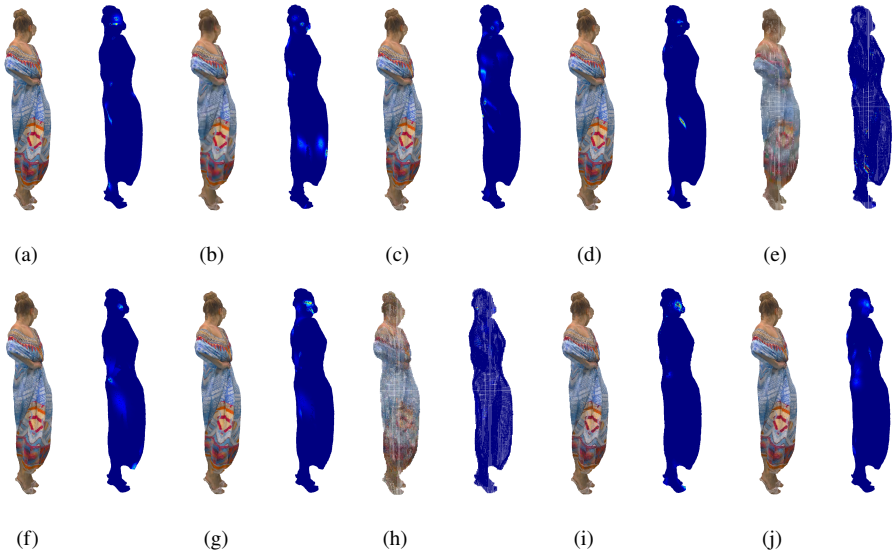


Figure 4.11: The referenced and distorted versions of point cloud *long dress* (frame 128) with corresponding visual attention maps based on the proposed processing protocol. Figures 4.11(a) is the reference version. Figures 4.11(b)-4.11(j) are the distorted version of *long dress* from low bitrate to high bitrate. Specifically, 4.11(b)-4.11(d): VPCC, 4.11(e)-4.11(g): GPCC, 4.11(h)-4.11(j): CWI-PCL.

"If it's very like vague or like with big dots, then it's ok, it's fake, but if it flickers all the time, that could be a bit annoying, it's sort of a lot to see"). Additionally, participants (20) reported that they tended to explore more during the observation of high-motion point cloud sequences.

Geometry and texture Distortions in geometry (12) and texture (11) are identified as the second and third factors influencing the subjective rating of point clouds under scrutiny. (P3: " I was observing precisely two things, the edges of the body and how distorted they are, and also some distortions inside the costume.") For DPC quality assessment, temporal information emerges as more critical than either geometric or texture distortions, with geometry and texture exhibiting nearly equal importance.

Distance impacts the quality rating Participants (8) discovered that the viewing distance can impact the subjective rating of the same content. Five of them noted that the appearance of holes is determined by the viewing distance, resulting in visually distinct point clouds even when inspecting the same sequence. This finding complements the conclusion in [90] that objects viewed from a greater distance tend to receive lower ratings compared to their closer counterparts.

Relationship between visual attention and quality assessment Participants favored the *longdress* (15), *soldier* (11), and *dancer* (9) point cloud sequences among all the contents. The two primary reasons cited were their high quality (19) and the presence of cues aiding in quality score determination (15). The characteristics of the point cloud itself influenced participants' visual attention. For instance, individuals tended to focus more on content characterized by high motion (*dancer*), realism (*longdress*, *soldier*), and intricate details (*soldier*) to facilitate the quality assessment task.

Subjective Quality Experiment Design for Dataset Construction

Procedure Participants (9) recommend streamlining the calibration and error profiling process to enhance user-friendliness, while also acknowledging the importance of maintaining data accuracy. This suggests a need for striking a balance between data precision and ease of use. Such a balanced approach is crucial for advancing the development of eye-tracking techniques, particularly in terms of calibration procedures.

Interaction with Content Participants (5) highlighted the importance of allowing individuals to determine the number of loops for each content themselves. They emphasized that for tasks requiring quality assessment, they may already have instant results in mind for certain content. Furthermore, participants expressed a desire for increased interaction between themselves and the objects within DPCs playback scenes in VR for better evaluation of the quality, underscoring the importance of customizable interaction for effective evaluation. (P4: "... I like to rotate it to whatever angle I want and then go and see it.")

Content Characteristics and Quality Assessment Task: Influencing User Interaction

Impacts of the quality assessment task for interaction in VR Participants (21) attributed most of their movement to the need to observe the front face to determine or recheck their quality score ratings. Additionally, VR cues presented in the human figures also prompted them to move extensively, enabling them to identify more cues for evaluating point cloud quality. (P5: "When I was seeing the quality, I was seeing the helmet and it had like a small thing on top, and there is a difference in the quality of that as well, and even the gun, you could see like different features on the gun. So there are more things to look at.")

Impacts of contents characteristics for interaction in VR Participants (21) expressed the view that if the quality level is easy to discern, they prefer to remain stationary until the sequence completes its loop, such as in cases of excellent and poor quality scales. However, if the quality level is relatively challenging to distinguish, individuals opt to move around for a comprehensive observation, even without engaging in random rotation operations. This observation aligns with Damme's conclusion [90] that human perception of underlying quality representations is intricately linked with the content and its geometric properties under examination.

4.2. TF-DPC Dataset Construction: Task-Free

To investigate how visual attention is deployed on dynamic point clouds and compare it with task-dependent saliency maps [24], we conducted a task-free eye-tracking experiment in a VR environment. During the experiment, we recorded the position (x, y, z coordinates) and rotation (three Euler angles around the x, y, and z axes) of the camera associated with each participant's HMD, along with timestamped data. This information was used to analyze participants' navigation movements within the physical space (i.e., the floor). Gaze-related data (gaze origin in x, y, z, and normalized gaze direction vector, the position of the point cloud frames) was collected following the same method as in [24] to generate saliency maps.

4.2.1. Materials

We select all 12 point cloud sequences from UVG-VPC dynamic point cloud dataset [214], 5 reference sequences from the QAVQ-dynamic point cloud dataset [24], and 2 sequences from the Owlii dataset [215] for the task-free eye-tracking experiment. We selected all the reference contents from the QAVA-DPC dataset as it contains task-dependent visual attention maps, thus aiding us in our purpose of comparing task-dependent and task-free viewing, and we complemented it with additional high-quality contents to provide additional saliency data. We compute the Spatial Information (SI) and Temporal Information (TI) for each content [195], by projecting the point cloud into 4 views, namely, left, right, front, and back view, of its bounding box to apply SI and TI separately, then obtain the maximum value among the 4 views over all the frames as the final SI/TI for one sequence. The distribution of all dynamic point clouds can be seen in Figure 4.2. The dispersed state in SI (horizontal axes)/TI (vertical axes) shows the diversity of our contents in the spatial/temporal domain. All the stimuli are reference quality (without any compression distortion).

4.2.2. Apparatus

To ensure that the high-level task is the only variable, we used the same apparatus as) [24], to enable a fair comparison with the other task-dependent experiment. Our experiment software is developed in Unity (version 2021.3.10.f1). The CWI point cloud unity package (version 0.10.0) is used to import and playback the dynamic point clouds) [197]. For the UVG-VPC dataset, each sequence contains 250 frames, while other sequences contain 300 frames. The frame rate is 30 frames per second, with each video being displayed 3 times. We use HTC Vive Pro Eye devices with eye-tracking capabilities and Vive hand controllers for participant interaction. The eye-tracking applications are developed using the native HTC Vive SRanipal SDK.

We ensured a watertight appearance of all the stimuli by adjusting the point size to the average distance among its 10 nearest neighbors all over all points in the point cloud [119]. They are rescaled to a similar size, around 1.8m in height, to mimic realistic tele-immersive scenarios. The VR scene is illuminated by a virtual lamp on the ceiling centered above the models. The lamp is set as an area light with intensity values of 2 in Unity to simulate ordinary lighting in a room.

4.2.3. Procedure

In this study, we use a within-subject design. To avoid the effects of contextual or memory comparisons, we randomly generated a playlist for each subject. Before the experiment, the visual acuity and color vision of every subject was tested using Snellen [200] and Ishihara [201] charts. Participants were briefed and signed a consent form prior to taking part in the study. At the beginning of the session, the inter-pupillary distance was measured and the headset was adjusted by the subject accordingly. Then, a training session was conducted to help familiarize the subjects with the system, including the controllers and the naming of each button to interact more easily. Two training sequence, namely *loot* and *redandblack*, were used, which were not included in the dataset. The subjects always started at the same location, which is 1.5 meters away from the center of the virtual room, but could move freely from there onward and ended anywhere they preferred. A stimulus was located in the center of the virtual room, and each stimulus was randomly rotated between [0°,360°] to avoid bias. During the experiment, the subjects were instructed to view each model freely in the VR environment during the playback of each sequence. The subjects were also required to stand still while doing the calibration and error profiling.

For each subject, the test was split into two rounds, lasting for around 17 minutes each, with a mandatory 5-minute break in between. Before and after each round, participants were requested to fill in a Simulator Sickness Questionnaire (SSQ) on a 1-4 discrete scale (1=none to 4=severe) [206]. For every model and subject, a round was split into three consecutive steps:

1 *Calibration* was done at the beginning of the experiment, and only when calibration was successful users could enter into the dynamic point cloud playback stage.

2 *Inspection of models* is the step where the participants are observing the dynamic point cloud naturally, while their movement trajectory and gaze-related information are recorded.

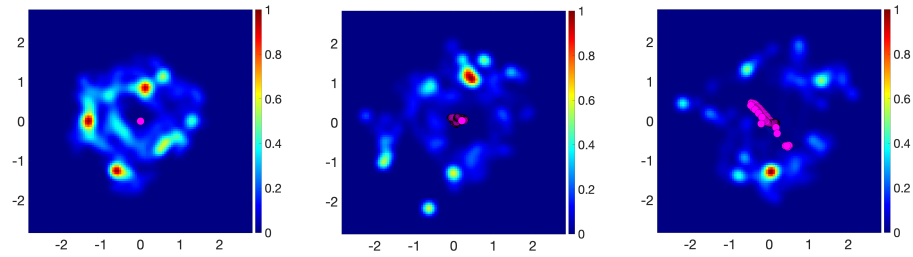
3 *Error profiling* is issued as the last step in order to estimate the accuracy of the gaze measurements due to calibration inaccuracies, or HMD displacements.

After participants finished the two rounds, they were requested to fill out the Immersive Presence Questionnaire (IPQ)[2]. Last, the researchers conducted a semi-structured interview. The interview was conducted individually in a non-VR setting, and the entire conversation was recorded for analysis purposes.

4.2.4. Participants

A total of 24 participants took part in the subjective tests of this study, with a diverse composition that includes 1 non-binary individual, 12 males, and 11 females. The participants' ages ranged from 23 to 35, with an average age of 28.33 and a standard deviation of 3.10. Each participant observed all the dynamic point cloud stimuli. In terms of occupation, the majority (66.67%) of the participants were students, ranging from master to PhD levels. The remaining 33.34% were researchers (scientist and lecturer), one landscape designer, and one accountant. Regarding familiarity with VR devices, 5 participants had

[2] https://www.igroup.org/pq/ipq/index.php



(a) rafa2 (b) CasualSpin (c) HelloGoodbye

Figure 4.12: Averaged spatial distribution over time of the main location visited by users while displaying three different content: (a) *rafa2*, (b) *CasualSpin*, and (c) *HelloGoodbye*. The centroid position of each volumetric content is represented by a sequence of pink points on the floor.

never experienced VR before the experiment, 13 participants had intermediate experience (using VR 1 to 3 times), and 6 of them were considered experts, having backgrounds as VR designers or researchers. Additionally, 17 out of 24 participants wore glasses during the experiment. No ethical approval was sought for this study, due to the absence of an established ethical review board at the institution where the research was conducted. The experimental protocol, including participant consent and data collection, was reviewed through an internal board to be compliant with current GDPR legislation. Participants consented to the collection and usage of their data at the start of the experiment, after being informed about the study.

Analysis of gaze data

4.2.5. Experiment Results

Analysis of movements on the physical space

The analysis of the movements on the physical space is based on the recorded data associated with the position and rotation of HMD collected during experiments. For the following analysis, the data was resampled at 30Hz. A general overview of the navigation behaviour of participants on the floor (plane XY) is given in Figure 4.12 for three selected contents, *rafa2*, *HelloGoodbye* and *CasualSpin*. We chose these volumetric point clouds based on their SI and TI values to investigate how the users movements change in relation with content characteristics. As shown in Figure 4.2, *rafa2* has low TI and SI, *CasualSpin* has high value of SI while *HelloGoodbye* is characterised by high TI. The volumetric content is initially placed approximately at the center of the floor plane and since the sequences are dynamic, we also represent their position over time with a trajectory of pink dots. It can be noted that the first sequence is the less dynamic since *rafa2* stays in its initial position (Figure 4.12 (a)). This brings to a more static behaviour also from the participants who mainly stay in one location without exploring the area around the content: there are indeed some strong red spots which represent the position where users spent most of their time and the shadow of the user position is quite compacted around

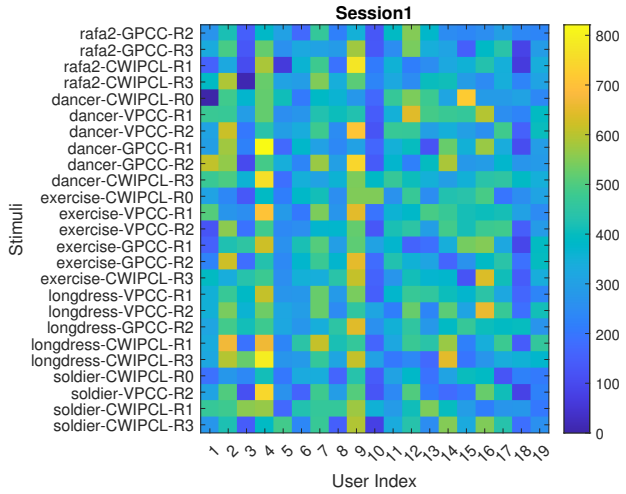


Figure 4.13: Fixations per subject and stimuli in the proposed TF-DPC dataset. Each row represents the fixations corresponding to a specific stimulus, and each column represents the fixations of an individual subject.

the content. The point cloud *CasualSpin* is instead spinning around itself. In this case, participants are more spread around the content to display it from different perspectives as shown in Figure 4.12 (b) but they are still quite compact. On the contrary, Figure 4.12 (c) shows a more dynamic exploratory behaviour from the users while displaying *HelloGoodbye*. To be noted that this sequence is also the most dynamic one since it walks back and forward. Thus, users tend to explore more while watching dynamic sequences, as already observed in [216].

To understand deeper visual exploration, we now analyze the relationship between gaze and stimuli. Following the same gaze data processing in [24], we ignored the initial 400 ms gaze data of each user to avoid unintentional gaze because of the unexpected appearance of the dynamic point cloud. Then, only the valid gaze samples provided by the native HTC Vive SRanipal SDK were selected. Each valid gaze sample was processed as follows: 1) Verify the data validity of gaze data by calculating the weighted average angular error to each gaze sample with the help of GazeMetrics [208]; 2) Identify fixation points of gaze data by dispersion threshold identification algorithm; 3) Map gaze data to dynamic point cloud frame with truncated-cone-sector algorithm [14]; 4) Fuse multiple users' gaze data to dynamic point cloud frames. After the four steps, we obtained the saliency map per frame. Each point cloud frame has a normalized fixation intensity range in [0,1] for each point, 0 meaning non-salient and 1 meaning the most salient. For the processing details, please refer to [24]. Figure 4.13 represents the number of fixations of each subject on each stimulus. Specifically, each row denotes the number of fixation points per stimulus across the different users. Blue colors indicate a low value of fixations while yellow ones indicate high values. Vertically, we can notice consistent behavior per participant across the different stimuli. For example, User 14 always has a low value of fixation, independent of the visualized volumetric stimuli, indicating a more erratic

behavior. On the contrary, User 1 appears to have more consistent fixations across the stimuli. Thus, participants tend to preserve similar gaze behavior (highly erratic or quite static) independently of the volumetric stimuli. Similar outcomes were observed also in [216]. Looking at Figure 4 per row (i.e., a single stimulus across different users), we can notice that stimuli with higher TI got more attention: *FlowerDance* and *model*, which are characterized by higher TI, present more fixations than *rafa2*. To further our analysis, in Figure 4.14, we show the saliency map (randomly selected frame 150_{th}) for these three sequences. We can see that all three sequences show fixations on semantically relevant areas, such as the face. However, in *FlowerDance*, who is in the middle of a spinning motion, and *model*, who is simply adjusting her dress, the fixation areas are smaller and more dispersed across the stimuli, as the users' attention is drawn by the motion of the dresses or any patterns on them. We further analyse and discuss gaze data in Section 4.3.1.

Analysis of SSQ and IPQ data

SSQ comprises 16 symptoms which are further grouped into three different categories: Oculomotor, Nausea, and Disorientation; we also computed the total score according to [206]. The simulator scores increased after the experiment. Specifically, the total scores rose from 6.37 to 10.33 before and after Session 1, and from 5.91 to 10.08 before and after Session 2. However, it can be seen that breaks help in reducing simulator sickness. The current version of the IPQ has three subscales (Spatial Presence, Involvement, Experienced Realism) and one additional general item not belonging to a subscale. We calculate the mean across the users for each factor. The participants experience high levels of Spatial Presence ($M_{SP} = 4.5$) and Involvement ($M_{INV} = 3.8$), whereas lower levels of Realisms ($M_{REAL} = 3.3$). The possible reason is that there is no interaction between the user and the content, as mentioned in Section 4.2.6, and there is no eye contact. The virtual room is empty for better capturing the visual attention, which normally gets a lower score for the question: "the virtual world seemed more realistic than the real world."

4.2.6. Qualitative results

22 valid interview audio recordings were transcribed into texts and coded using Dovetail[3]. Following Maguire's guideline on thematic analysis [213], we initially reviewed and labeled the text, organized these labels into themes, and subsequently convened to establish the connection between content and visual attention during the subjective test. Each participant is denoted as P1-P24, with the number of participants concurring with each statement indicated in parentheses.

Factors that Capture Visual Attention Allocation

Temporal information Participants (18) pointed out that movement is the most attractive factor in our dynamic point cloud playback scene (P21: "when you are watching a video, it's easy to follow the direction of the movements."). 11 of them interpreted the information conveyed by the content as interesting to attract their attention. However,

[3] https://dovetail.com/

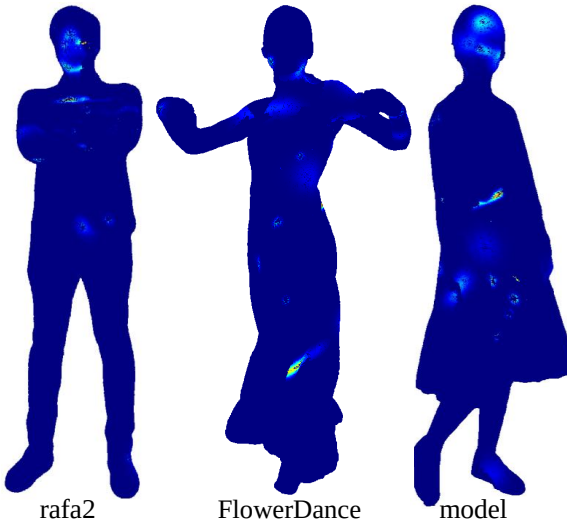


Figure 4.14: The visual saliency map of the 150th frame of the dynamic point cloud with the front view.

participants (16) also noted that high-motion sequences do not necessarily attract more attention than low-motion sequences.

Artifacts and Details Artifacts (9) and details (9) are identified as the co-second factors attracting people's attention. (P8: "what I focused on also negative things are, on the edges of the point calls often there was like rippling, sort of flickering, attracts a lot of attention, distracts me, other than that, I think eyes like faces in general, people like the expression.")

Geometry and Texture Geometry (2) and texture (7) are identified as the second and third factors influencing the subjective rating of point clouds under scrutiny. (P3: " I was observing precisely two things, the edges of the body and how distorted they are and also some distortions inside the costume.")

In terms of visual attention allocation, temporal information proves to be more crucial than either geometry or texture, with both geometry and texture showing relatively low importance. The details of the dynamic point cloud fall somewhere in between, while negative artifacts in the point cloud attract significant attention, aligning with findings from a previous study [89].

Factors Affecting Visual Attention

Participants (12) reported the realism of the content and naturalness of the action would change their attention. (P1: "I have to say there's an effect, if I see the quality is good, I usually will look closer. I will check the details. But if the quality is so poor that I can see distortion everywhere, then I will consciously, I will realize this is not real. So

I will be less interested.") Abrupt distortions of the sequence will shift attention, (P5: "The point cloud's intended focal point might end up being overlooked because the flaws draw my attention away from it, instead I focus on the imperfections."). It is worth noting that all the point clouds under test were of reference quality; that is, any impairment was derived from the acquisition itself, and was not due to any additional processing such as compression. Thus, the acquisition methods themselves can have a significant impact on visual attention. This observation aligns with Zhang's conclusion [107] that distortions always change the attended regions.

Factors Influencing User Interaction

Participants (14) attributed most of their movement to the need to observe the front face to have more understanding of the human figure. They noted that sequences showing the same human figure with only slight variations in movement and clothing, as in the UVG-VPC dataset, led to decreased movement and reduced interest. This repetition (5) and the monotonous actions (5) made the task feel not engaging and dull. Limited space (8) and cable (1) result in less movement of the participants.

Designing the Construction of a Visual Attention Dataset

Content Participants favored the *longdress* (7), *soldier* (6), and *Gymnast* (5) point cloud sequences among all the contents, describing them as both realistic and engaging. However, some participants (3) noted that there are only human figures. Additionally, they expressed a desire for more varied objects and increased interactivity, such as eye contact between themselves and the content, to enhance the immersive experience.

Display equipment for dynamic point cloud Participants (16) stated that using an HMD in VR is a better alternative to a 2D screen, as it provides greater immersion and freedom. (P7: "I think it's more intuitive if you feel more real when you see it, by 1 to 1 ratio is like your size, it's like next to you while on the screen it's like really small, you can zoom in but then the screen is not as big or you only see maybe one part of it even though it's a big screen, it's not 3D.") However, the HMD is heavy (2) and uncomfortable for prolonged use (2), while 5 participants noted that its effectiveness depends on the specific application.

4.3. Comparison between task-free and task-dependent

To explore how visual tasks impact the visual attention, we quantitatively analyze gaze statistics and saliency map similarity between task-free and task-dependent scenarios. To be noted these analyses are limited to the five shared sequences across the proposed dataset and the one presented in [24]: *rafa2* (low SI, TI), *dancer* (medium SI, high TI), *exercise* (low SI, high TI), *longdress* (high SI, medium TI), and *soldier* (medium SI, TI).

4.3.1. Comparison Consistency of Gaze

To analyze the allocation of visual attention depending on the task, we propose three measurements. We choose the total fixation number instead of other statistics of the

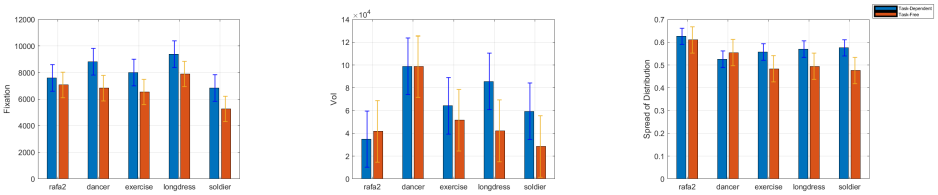


Figure 4.15: Aggregation of fixations, VoI, and the spread of the distribution across participants of task-free and task-dependent experimental scenarios for the 5 shared dynamic point clouds from both QAVQ-DPC and proposed TF-DPC datasets, separately.

gaze [217] (the mean duration or scan-path magnitude), because since the fixation is obtained through the dispersion threshold identification algorithm, the duration of consecutive gaze samples is implicitly considered. Apart from gaze behavior, our focus is on where the gaze is allocated within 3D point cloud frames. We select the Volumes of Interest (VoI) [218], which can show how many volumes have been observed by humans, and the distribution of the VoI, which can tell us how their attention is dispersed across the point cloud. VoI is computed as the total number of points whose fixation intensity is larger than zero, the spread of VoI is the average pairwise distance of the VoI within the point cloud. Figure 4.15, from left to right, shows the fixation, VoI, and the spread of VoI across participants in both a task-free and task-depended experiment. We can observe the following: 1) Fixations for all 5 sequences with variant SI and TI perform consistently. The fixation number under task-free is lower than under task-dependent conditions since people need to focus relatively more to evaluate the quality of the sequences. 2) Generally, more fixations mean larger VoI and sparser distribution of the VoI. However, this is not true for *dancer* and *rafa2* sequences.

To analyze the difference between tasks with respect to these measures of visual attention, we ran a set of analysis of variance (ANOVA) tests. We grouped all fixations by task and aggregated measures by participant for each content per frame. One-way ANOVAs indicate the overall effect of the task on these measures. The p-value is below the threshold (0.05) of significance for all the contents per measure except for the spread of distributions for *rafa2* and the RoI for *dancer*, which are 0.1641 and 0.8008, separately. In conclusion:

- Across all 5 sequences, the number of fixations is significantly different between task-free and task-dependent scenarios. Task-dependent viewers, who were evaluating the quality of the content, consistently had more fixations compared to task-free viewers, who likely scanned the content more freely. This supports the idea that task-related goals require more focused attention, leading to a higher fixation count. Sequences with higher SI and TI, such as *longdress* and *dancer*, tend to capture more attention, evidenced by the higher number of fixations. In contrast, lower SI and TI sequences like *rafa2* generally had fewer fixations, as they may not have been as visually engaging.

- There is a significant difference for most contents, with task-dependent conditions leading to larger VoIs. This suggests that when participants are given specific tasks, they distribute their attention more widely across the point cloud (multiple specific areas), perhaps because the tasks prompt them to explore more regions for relevant information. While in free-viewing, they explored generally, driven by personal curiosity or passive observation rather than the active search for specific details. *dancer* stands out as the only content where both conditions cover the same. This could mean that the nature of the *dancer* does not lead to a noticeable change in the areas participants attend to, regardless of whether they are given a task or not.
- There is a significant difference for most contents, with task-dependent conditions leading to a broader spread of attention. However, for *rafa2*, there is no significant difference between the two conditions since it lacks of a main attention area, likely due to its low SI and TI and no particularly engaging features to attract viewers' attention. As a result, people tend to look around more. The possible reason for the higher spread of VoI for *dancer* while remaining the same VoI is due to its continuous movements over time, with the dynamic dance gestures evenly capturing attention across the point cloud.

4.3.2. Comparison Consistency of Visual Saliency Map

We aim to compare the point cloud saliency map in task-free and task-dependent scenarios. Commonly used metrics for such a comparison are listed in Table 4.2. The key properties include location or distribution-based, similarity or dissimilarity measurement, sensitivity to 0 values, and consideration of spatial distance. Since the generated saliency map for dynamic point clouds uses exactly the same method in [24], which does not include an explicit fixation point on the point cloud, the location-based metrics are not applicable to our continuous point cloud saliency maps. Among the distribution-based metrics, SIM, as a similarity metric, penalizes misalignment and is sensitive to missing values and 0 values, while KL, as a dissimilarity metric, is also sensitive to 0 values. Thus, based on the recommendation for metric selection [219, 220] and the characteristics of our dynamic point cloud saliency map, i.e., the majority of the points are non-salient (i.e., fixation intensity equal to 0), we opt not to use them. EMD, as a dissimilarity, is the only metric that considers spatial distance. Herein we choose PCC to measure the similarity and adapt EMD, which is used to measure the 2D saliency map, to measure the dissimilarity.

The PCC is a statistical method to measure how correlated or dependent two variables are. In our scenario, given the visual saliency maps obtained from a task-free **F** and task-dependent **D** experiment, PCC can be defined as follows [221]:

$$PCC(\mathbf{F},\mathbf{D}) = \frac{cov(\mathbf{F},\mathbf{D})}{\sigma_{\mathbf{F}}\sigma_{\mathbf{D}}}. \tag{4.3}$$

where $cov(\cdot)$ is the covariance and σ is the standard deviation. PCC ranges from -1 to 1, with higher absolute values indicating stronger correlation between visual saliency maps. However, PCC is sensitive to outliers and only compares the magnitudes of corresponding points. This makes it unable to account for shifts in point locations or partial matches in attended areas, which are common in eye-tracking experiments due to device limitations

Table 4.2: Property of Evaluation Metrics for Image Saliency Map

	AUC	NSS	IG	SIM	KL	PCC	EMD
Location-based	✓	✓	✓				
Distribution-based				✓	✓	✓	✓
Similarity	✓	✓	✓	✓		✓	
Dissimilarity					✓		✓
Sensitive to 0 values			✓	✓	✓		
With spatial distance							✓

or participant preferences. This issue is especially noticeable in large point clouds. To address this, we propose to adapt EMD for dissimilarity measurement [222], as it better captures the distribution of attention by incorporating spatial information. EMD helps to alleviate the issues of point shifts and partial matches in large volumetric content cases. Specifically, we generate the "signature" (a feature that can represent the saliency map) by calculating a histogram of the fixation intensity at each point in 3D space. We denote a discrete, finite distribution $\mathbf{p}$ from the saliency map obtained in the task-free experiment as

$$\mathbf{p} = \{(p_1, w_1), \ldots, (p_m, w_m)\} \equiv (\mathbf{P}, \mathbf{w}) \in \mathbb{D}^{K \times m} \tag{4.4}$$

where $\mathbf{P} = [p_1, \ldots, p_m] \in \mathbb{R}^{K \times m}$ represents the signature with m points (or clusters), $w_i \geq 0$ represents the weight or fraction associated with the i-th point (or cluster) for all $i = 1, \cdots, m$. Here K is the dimension of ambient space (Euclidean space for 3D point cloud) of the points $p_i \in \mathbb{R}^K$, and m is the number of points (or clusters). The total weight of the distribution $\mathbf{p}$ is $w_\Sigma = \Sigma_{i=1}^m w_i$. Given two distributions in task-free and task-dependent scenarios as $\mathbf{p} = (\mathbf{P}, \mathbf{w}) \in \mathbb{D}^{K,m}$ and $\mathbf{q} = (\mathbf{Q}, \mathbf{u}) \in \mathbb{D}^{K,n}$. We used the following EMD [222]:

$$\mathrm{EMD}(\mathbf{p}, \mathbf{q}) = \frac{\min_{F=(f_{ij}) \in \mathscr{F}(\mathbf{p},\mathbf{q})} \mathrm{WORK}(F, \mathbf{p}, \mathbf{q})}{\min(w_\Sigma, u_\Sigma)}. \tag{4.5}$$

The EMD distance $\mathrm{EMD}(\mathbf{p}, \mathbf{q})$ between $\mathbf{p}$ and $\mathbf{q}$ is the minimum amount of work to match between distribution $\mathbf{p}$ and $\mathbf{q}$, normalized by the weight of the lighter distribution. Thus, to obtain the EMD value, we need to find the optimal flow by solving the transportation problem. The work done by a feasible flow $F \in \mathscr{F}(\mathbf{p}, \mathbf{q})$ in matching $\mathbf{p}$ and $\mathbf{q}$ is given by

$$\mathrm{WORK}(F, \mathbf{p}, \mathbf{q}) = \sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij} d_{ij}, \tag{4.6}$$

where $d_{ij} = d(p_i, q_j)$ is the "ground distance" between p_i and q_j. We consider the degree of salience and the spatial information of the point cloud jointly, the ground distance is now defined as

$$d_{ij} = \lambda |h_i - h_j| + (1 - \lambda)[(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2]^{\frac{1}{2}}, \tag{4.7}$$

where h_i is the middle value of the i_{th} bin of the histogram in $\mathbf{p}$, and (x_i, y_i, z_i) is the location of the centroid point located in i_{th} bin of $\mathbf{p}$. λ is a weight used to balance the

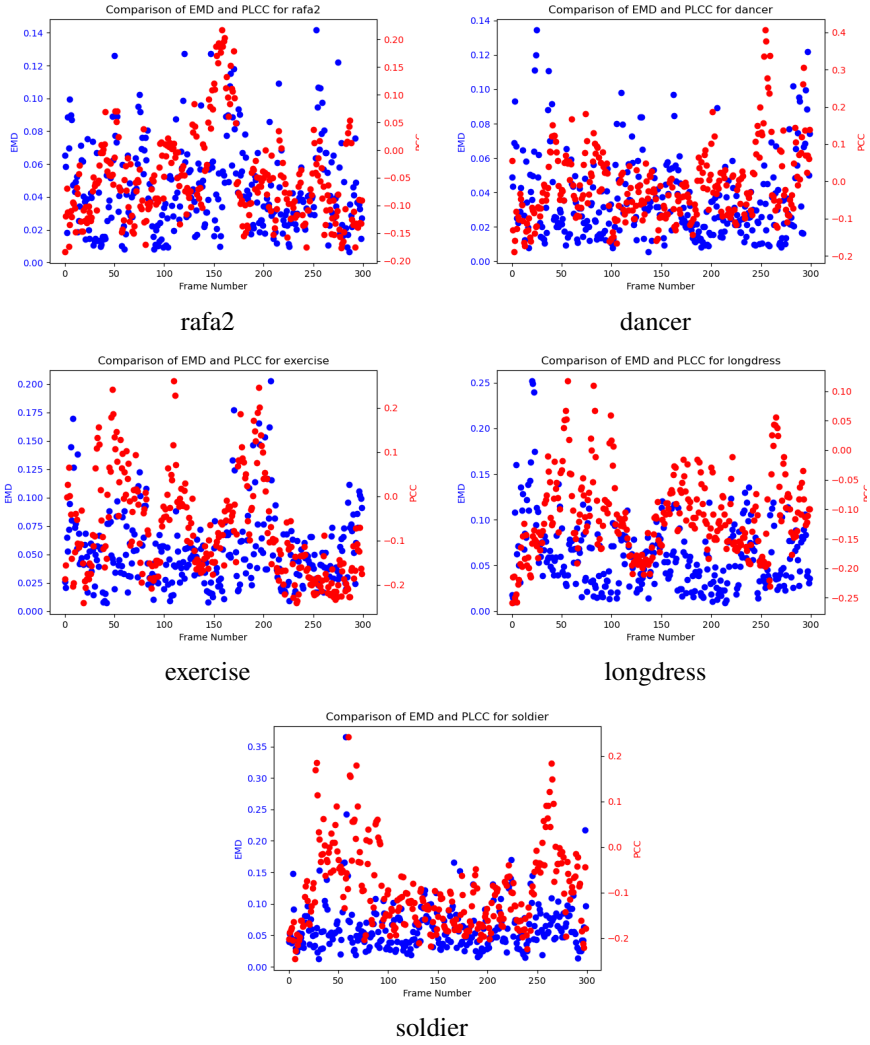


Figure 4.16: Similarity of point cloud saliency maps between task-free and task-dependent scenarios through EMD (•) and PCC (•) for the shared 5 sequences per frame, separately.

importance between spatial information and the magnitude of the fixation intensity value.

The flow F is a feasible flow between $\mathbf{p}$ and $\mathbf{q}$ iff

$$f_{ij} \geq 0 \quad i = 1, \ldots, m, \; j = 1, \ldots, n, \tag{4.1}$$

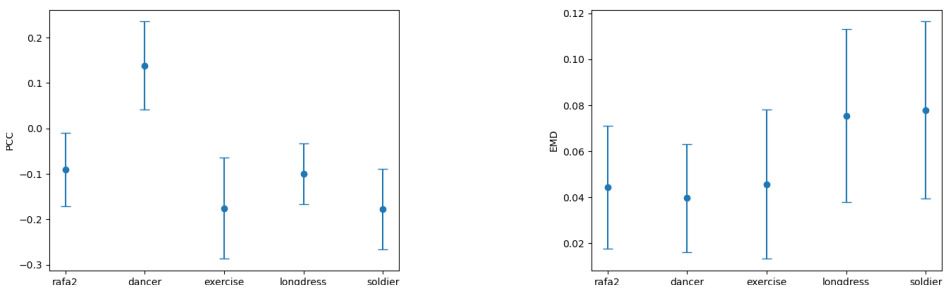
$$\sum_{j=1}^{n} f_{ij} \leq w_i \quad i = 1, \ldots, m, \tag{4.2}$$

$$\sum_{i=1}^{m} f_{ij} \leq u_j \quad j = 1, \ldots, n, \quad \text{and} \tag{4.3}$$

$$\sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij} = \min(w_\Sigma, u_\Sigma). \tag{4.4}$$

The detailed explanation for the constraints can be found in [222]. The coordinates of the distribution points are not used directly in the EMD formulation, only the ground distances d_{ij} between points are needed. A larger EMD indicates a larger difference between two distributions while an EMD of zero indicates that two distributions are the same. In this paper, we remove the points that are non-salient in both experiments before we compute the PCC and EMD to obtain an accurate measurement. The bin number of the histogram is set to 30, λ is set to 0.5.

To fairly compare similarity and dissimilarity metrics, we normalize the EMD values to [0,1] range and convert dissimilarity into similarity. This is achieved by dividing the computed EMD by the maximum possible EMD for a given histogram, assuming all the mass (i.e., salient points) starting in the leftmost bin need to be moved to the rightmost bin. The similarity score for EMD is then calculated as 1 minus the normalized EMD. Figure 4.16 compares PCC and EMD values for the shared 5 contents per frame, separately. We observe that PCC exhibits greater variance for *exercise*, *longdress*, and *soldier*, as evidenced by fluctuations in the PCC values across frames. This variability suggests that PCC is sensitive to outliers in the saliency map, leading to greater variation in visual similarity over time for these contents. In contrast, EMD demonstrates more stable and consistent behavior, with values that remain within a narrower range, indicating reduced fluctuations. This stability arises from EMD's consideration of spatial information and its partial match property. Fig 4.17 shows the average similarity across frames in task-



Average PCC Average EMD

Figure 4.17: Similarity of point cloud visual saliency maps between task-free and task-dependent for the shared 5 sequences averaged over 300 frames, separately.

free and task-dependent scenarios. Notably, *dancer* is identified as the most similar sequence by PCC, while *soldier* is the most similar according to EMD. PCC's emphasis on matching magnitudes at the same points leads to high similarity scores for *dancer*, where obvious salient regions identified by humans remain consistent over time, independently of the task.

Combining Figure 4.16 and Figure 4.17, it becomes clear that both EMD ([0, 0.35]) and PCC ([-0.25, 0.4]) exhibit low similarity values, suggesting substantial differences between task-free and task-dependent scenarios. This highlights that task-dependent scenarios in dynamic point clouds significantly alter human visual attention. EMD identifies overlapping regions of attention in both scenarios, providing a more spatially-aware similarity measure, while PCC captures sharp variations for specific content. Figure 4.18 shows saliency maps for *soldier* at the frames with maximum similarity under EMD and PCC metrics. Visually, the saliency in the 58_{th} frame appears more similar than in the 61_{th} frame, with the inset of the head showing greater overlap, particularly from the back view. This comparison further demonstrates that while both PCC and EMD have their strengths, EMD's consideration of spatial information makes it more suitable for evaluating saliency in point cloud data.

4.3.3. Summary

Quality assessment, as a high-level perceptual task, significantly influences how visual attention is deployed when evaluating dynamic point clouds in VR. As discussed in Section 4.3.1, one key observation is that participants exhibit fewer fixations in task-free conditions compared to task-dependent ones. This is evident in Figure 4.18, where task-dependent viewers focus more on specific details, such as the spotlight on the soldier' s hat. In contrast, task-free viewers typically form a general impression, primarily attending to broader features like facial expressions, rather than thoroughly exploring "less critical" details once they have grasped the overall scene.

In task-dependent conditions, the demand for precise quality evaluation prompts participants to observe the sequence more carefully. Their goal is to gather visual cues to assess the content's quality, which explains why saliency maps under the quality assessment task tend to have a larger VoI. Additionally, due to content repetition (same content with different quality level), participants in task-dependent conditions are less inclined to explore the back of the point cloud, preferring the primary areas in the front view that they deem relevant for the quality assessment task. In task-free conditions, participants generally scan the content broadly, focusing on prominent movements or artifacts. Since they are not bound by a specific objective, they tend to observe both the front and back views of the point clouds without particular focus.

The spread of the VoI, however, varies between different conditions for different reasons. In task-dependent, participants' attention is drawn to specific features from head to toe, like the spotlight on the hat, the watch on the hand, and the shoes, as shown in Figure 4.14 the frame 58 under task-dependent condition. Participants' attention is more targeted, with individual differences in strategies for assessing quality. This variability contributes to the spread of the VoI but with greater focus on elements that are crucial to quality judgment. In contrast, the task-free condition reflects a more passive viewing approach. Participants form a holistic view of the scene, only directing their gaze toward

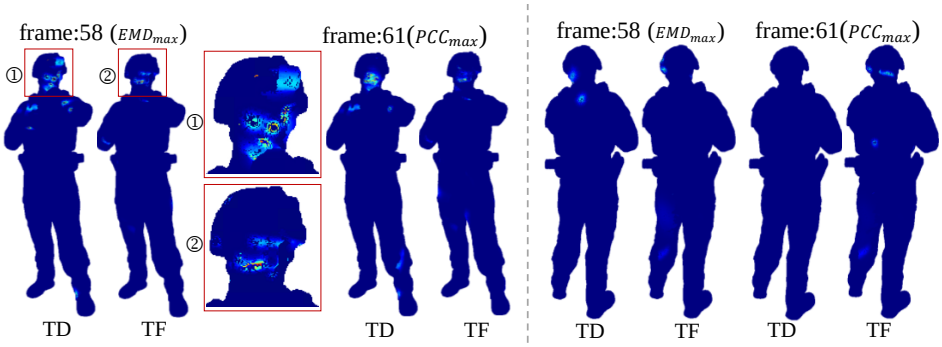


Figure 4.18: Saliency map visualization of *soldier* in frame 58 and frame 61, identified as the most similar maps using the adapted EMD and PCC metrics. The left side of the dotted line shows the front view of the *soldier*, while the right side shows the back view. TD refers to the saliency map collected under task-dependent conditions, and TF refers to task-free.

areas of movement or obvious artifacts. Without the demand to assess quality, their focus is less concentrated on specific details, and their viewing patterns reflect a broader exploration of the scene.

Movement and semantic information in the dynamic point clouds, such as facial expressions or body movements, consistently attract visual attention in both scenarios. For example, in Figure 4.19, participants frequently fixate on faces across multiple frames. Interestingly, visual attention appears to be more consistent in task-dependent conditions, especially when it comes to fine details, regardless of whether the scene has high or low TI. Participants are more likely to scrutinize these details to detect subtle distortions, which are critical for assigning quality scores. This difference in attention deployment highlights how task-driven objectives shape visual behavior, with task-dependent viewers engaging in top-down mechanisms and task-free viewers adopting a more relaxed, impressionistic approach.

4.4. Discussion

4.4.1. The influence of task for DPC visual attention

Our first quality assessment experiment, which focused on evaluating the visual quality of DPCs, likely influenced participants' attention toward specific content areas that facilitated the task: for example, areas with patterns on which distortions would easily be spotted. That does not necessarily mean that the same area would be a salient region had the test been administered with a different task or task-free. Insights from the semi-structured interviews confirmed that the identified salient regions were not only inherently attention-grabbing but also offered cues that aided participants in the quality assessment process. This dual influence highlights the contextual nature of visual attention, driven both by intrinsic content characteristics and the task's demands. Moreover, the accuracy of visual attention map predictions varies depending on the display environment and asso-

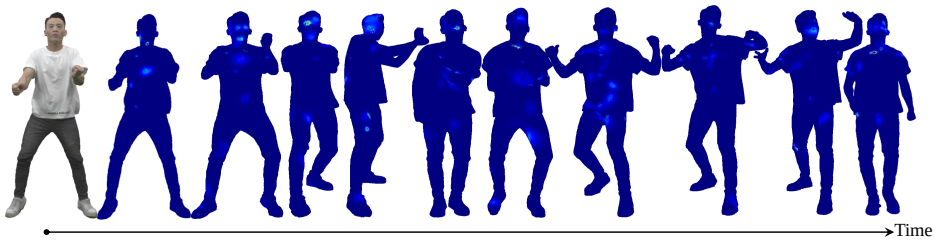


Figure 4.19: Fixation maps of *dancer* sequences with uniform temporal sampling every 30 frames under the task-dependent scenarios.

ciated tasks [223, 224]. Future research should explore how visual attention varies across different tasks or task-free scenarios to uncover more generalized patterns of saliency in DPCs and their implications for real-world applications [225].

4.4.2. Visual attention applications for DPC

The insights derived from gaze data for DPC extend far beyond quality assessment. Accurate prediction of visual attention provides a robust foundation for optimizing compression and streaming strategies. For example, saliency-driven bit allocation can enhance encoding efficiency by prioritizing the fidelity of regions that capture user attention while allocating fewer resources to non-salient areas. Additionally, integrating saliency maps with temporal motion information, such as motion vectors, enables adaptive streaming parameters that dynamically adjust to user focus, ensuring a seamless and engaging experience. Furthermore, leveraging these insights can support semantic segmentation of motion-dominant and motion-static regions, which has significant potential for applications such as object recognition and interactive XR environments.

4.4.3. Visual attention collection limitations

In the second comparison study, we collect a task-free saliency dataset for dynamic point clouds and investigate the task impact on human attention allocation. We observed that a central bias persists to some extent when viewing human faces, regardless of whether the conditions are task-free or task-dependent. However, our study is limited by the fact that TF-DPC focuses solely on human figures, excluding other immersive content types like landscapes or interactive objects. This limitation stems directly from the lack of high-quality, realistic datasets of dynamic point cloud objects, as to date, only synthetic datasets including dynamic objects are present in the literature [226, 227]. Thus, the outcomes of this study are valid only for the dynamic human category, and future work should explore broader content types. We chose a wired HMD to maintain consistency with the conditions of the previous study; however, this choice restricted physical movement due to the HMD's cable, and the device's weight and discomfort may have increased cognitive load, potentially resulting in fewer and less stable fixation points. To explore this, future studies should consider assessing pupil dilation and blink rate, reliable indicators

of cognitive load, alongside gaze amplitude and fixation patterns. These constraints may limit the ability to collect naturally viewing saliency maps and could introduce systematic biases. Using wireless HMDs, such as the HTC Vive Focus Vision, could improve ecological validity. Additionally, dynamic point clouds in high-quality XR scenarios are inherently dense, but the visual saliency regions occupy only a small portion of the content. Increasing the participant sample size in future studies would enhance statistical power and improve the generalizability of the findings.

4.4.4. Visual saliency collection under various perceptual tasks

The findings of our comparative study on the impact of high-level tasks for human visual attention deployment differ from previous research on images [228] but align with conclusions drawn from static 3D models [229]. Specifically, similarity metrics indicate lower saliency collection for static 3D models (PCC: 0.35) [229] compared to images (PCC: 0.84) [228]. While task-dependent, top-down mechanism effects on overt visual attention have been well-studied for 2D media [230], how these findings translate to dynamic point clouds remains largely unexplored. Additionally, there is evidence that traditional attention paradigms may not fully apply to newer media formats, such as panoramic videos [231]. Our findings have shown that quality assessment has a significant impact on human visual attention deployment, with both saliency maps under task-free and quality assessment tasks focusing on semantic area and movement. However, their focus differs, as mentioned in the above Section 4.3.3. A critical question that emerges from our study is whether saliency collected under task-free conditions or task-dependent conditions provides greater value for specific applications, such as point cloud quality assessment. Exploring the temporal dynamics of saliency in dynamic point clouds–how it evolves over time under varying task demands–critical for optimizing visual representations. Future research should focus on exploring the temporal dynamics of saliency across various perceptual tasks, clarifying the benefits of different saliency detection methods, and incorporating these insights into prediction models tailored to dynamic point clouds for specific applications.

4.4.5. Visual saliency collection in AR

3D visual saliency has been measured through various devices, such as the eye-tracking glasses [232], AR HMD [23], and VR HMD [24]. Understanding the differences between these devices is essential for accurately predicting saliency while accounting for factors such as spatial bias [181], center bias [183], and systematic error [233]. Nguyen [23] released saliency maps for four dynamic point clouds (namely *BlueSpin*, *CasualSquat*, *FlowerDance*, and *ReadyForWinter*) in AR, overlapping with our proposed TF-DPC dataset. Thus, using these four sequences, we were able to conduct an initial analysis of saliency maps across different devices. Notably, not every frame in the AR sequences contains fixation data, so we retained only the frames with salient areas present in both AR and VR. We computed the average VoI ratio (salient area relative to the entire point cloud across the sequence), as shown in Figure 4.20. Our findings from the comparative study indicate that the VoI in the AR condition is significantly smaller than in the VR laboratory setting, with participants primarily focusing on limited regions of the point

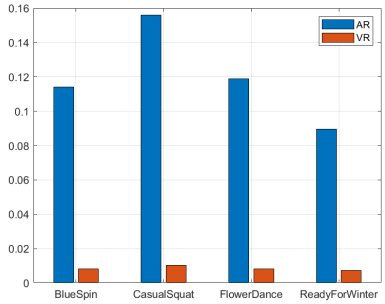


Figure 4.20: The average ratio of the VoI for the shared 4 dynamic point cloud sequences in AR and VR.

cloud. This reduction may be attributed to the HoloLens' limited field of view (about 52°) compared to a VR headset (about 110°). Furthermore, since AR blends virtual context with the real environment, users must frequently switch contexts and refocus their gaze [234], which can further reduce fixations on dynamic point clouds. Additionally, participants cannot view the entire life-sized point cloud unless they step back. Thus, the experimental protocol for saliency collection in AR requires careful consideration.

4.4.6. Evaluation metrics for the similarity of point cloud saliency maps

Several metrics exist for quantitatively measuring the similarity of 2D saliency maps, some of which can be adapted to static point cloud saliency maps with minimal adjustments. However, location-based metrics like NSS, which depend on precise fixation points, may not be directly applicable to point clouds. Human gaze fixation corresponds to a specific pixel in 2D images, but in 3D point clouds, the gaze ray may not intersect with any point in space, requiring approximation methods that introduce inaccuracies. Thus, metrics relying on fixation locations may not be suitable for point clouds unless these approximations are properly addressed. For distribution-based metrics, which compare the overall spread of attention, present a different challenge: how should we balance coverage similarity (whether the same areas are salient, regardless of magnitude) against magnitude similarity (whether the saliency levels are comparable)? Some scenes may show full spatial matches but differ in magnitude, or vice versa, making it unclear which aspect should be prioritized. This decision depends on the specific application.

Riche *et al.* [220] argue that no single metric is sufficient for evaluating saliency map similarity. The 3D nature of point clouds and the relatively small salient regions further complicate this task. For dynamic point clouds, the added dimension of time introduces variability due to motion, requiring spatial-temporal saliency distributions to be more effective in measuring similarity. Especially for human dynamic point clouds, for example, in Figure 4.1, the 151_{st} frame of *dancer* sequence, should the saliency of symmetric semantic areas (the left and right feet) be treated equivalently when we measure the similar-

ity? Incorporating metrics that consider temporal consistency and semantic relationships could help capture nuances in saliency similarity, particularly in dynamic scenarios where motion and semantic equivalency, such as symmetrical regions, play a significant role.

4.5. Conclusion

In this chapter, we presented the construction and analysis of two visual saliency datasets for DPCs in VR environments with 6DoF, each designed under distinct viewing conditions: task-dependent and task-free. These user studies were conducted to address *R2: How can visual saliency in dynamic point clouds be detected and compared in immersive environments?*

The QAVA-DPC dataset was developed for task-dependent scenarios, in which participants were asked to perform a quality assessment task while observing dynamic point clouds. A dedicated experimental protocol was designed to simultaneously capture subjective quality scores and visual attention data in a VR setting. Analysis of visual attention maps, along with insights derived from semi-structured interviews, provided a deeper understanding of human perceptual behavior in immersive environments. While prior studies have suggested that visual saliency is beneficial for applications such as image/video streaming, compression, and quality assessment, our work represents an initial step toward incorporating saliency into dynamic point cloud evaluation. Future research will focus on predicting visual attention in DPCs and developing objective quality metrics that better correlate with human perceptual responses.

In contrast, the TF-DPC dataset targets task-free viewing scenarios and includes 19 dynamic point cloud sequences. Participants were allowed to freely explore the content without any specific task, enabling the collection of natural gaze patterns and movement trajectories. This setting allows us to study how visual attention is distributed in unconstrained, immersive conditions, offering a complementary perspective to the task-driven findings.

To compare saliency under these two viewing paradigms, we conducted a detailed analysis of gaze statistics and saliency map similarities. For this purpose, we introduced a novel similarity metric based on the Earth Mover' s Distance (EMD), adapted to account for both spatial distribution and saliency intensity. This metric enables a more distribution-relevant comparison of saliency maps in dynamic point cloud settings. Our experimental findings indicate that task context—especially high-level cognitive tasks like quality assessment—significantly influences visual attention patterns. Moreover, this influence is modulated by content-specific characteristics, particularly the presence and complexity of temporal dynamics.

To support further research, we publicly release not only the QAVA-DPC and TF-DPC datasets, but also the associated experimental software, raw eye-tracking and motion data, scripts for visual attention map generation, and the implementation of our EMD-based saliency map comparison method.

While the primary focus of this chapter was to support perceptual quality assessment of DPCs through subjective experiments and saliency data collection, the next chapter shifts toward integrating this ground-truth saliency information into objective quality assessment models. Specifically, we explore how saliency-guided features can enhance the

perceptual alignment and robustness of objective PCQA metrics in dynamic 6DoF environments. In addition to leveraging the captured ground-truth visual saliency, we also investigate learning-based saliency prediction by adapting existing visual saliency models originally designed for images, aiming to further enrich the feature representation for PCQA and boost the quality prediction performance.

5

Visual Saliency-based Objective PCQA metrics

In the previous two chapters, we assessed the perceptual quality of point clouds—both dynamic and static—through subjective studies and objective PCQA models. While high prediction accuracy was achieved, further improvements are possible by incorporating visual saliency information, either collected from user experiments or predicted using computational models. This chapter explores how visual saliency can enhance objective PCQA. We first integrate ground-truth saliency maps into point-based PCQA models for dynamic point clouds. We then adapt image-based saliency prediction models for use with projection-based PCQA of static point clouds. The integration of saliency leads to improved perceptual alignment and prediction performance, albeit with increased computational complexity. The contributions of this chapter include practical strategies for applying visual saliency in PCQA, quantitative analysis of performance gains, and insights into model complexity and temporal behavior—offering guidance for future saliency-aware point cloud streaming and quality evaluation applications.

This chapter is based on the following publications:

1. ***Xuemei Zhou**, Irene Viola, Ruihong Yin, and Pablo Cesar. 2024. Visual-Saliency Guided Multimodal Learning for No Reference Point Cloud Quality Assessment. Proceedings of the 3rd Workshop on Quality of Experience in Visual Multimedia Applications. (ACM MM Workshop).* [26]
2. ***Xuemei Zhou**, Irene Viola, Evangelos Alexiou, Jansen, Jack, and Pablo Cesar. 2025. Subjective and Objective Quality Assessment for Dynamic Point Cloud with Visual Attention in 6 DoF. Transactions on Multimedia Computing Communications and Applications (ACM TOMM).* [27]

Point clouds play a crucial role in various real-world applications. Visual attention also plays a crucial role in various vision tasks, such as segmentation, localization, and registration [235]. Notably, leveraging visual attention maps to weight quality maps has shown improvements in perceptual quality prediction [236]. By connecting visual attention and visual quality for DPCs, quality allocation between salient regions and surrounding areas, saliency-aware compression and streaming, and saliency-aided objective quality metrics can be further investigated and optimized. Hence, predicting the visual quality of point clouds accurately and efficiently, in a way that correlates well with the HVS, is highly desired.

The intricate geometrical structure and densely packed points of point clouds, complete with attributes such as color, normal, and transparency, allow for detailed representations of environments, objects, and humans. While this richness of information is valuable, it presents challenges for the efficiency and accuracy of PCQA metrics. The different factors that contribute to the visual quality of point clouds are not fully understood, adding to the complexity of developing effective PCQA metrics.

In this chapter, we explore the added value of incorporating visual saliency into PCQA metrics in two distinct scenarios: dynamic point clouds and static point clouds. For dynamic point clouds, we leverage the QAVA-DPC dataset, which provides both subjective quality scores and corresponding point cloud saliency maps. These saliency maps are integrated into point-based objective PCQA models to enhance their perceptual alignment. In contrast, for static point clouds—where no comparable dataset exists—we take advantage of the maturity of image-based saliency prediction models. We adapt these learning-based saliency models to generate saliency maps for projected views of static point clouds and integrate them into projection-based PCQA models. Through this approach, we aim to systematically evaluate the potential of visual saliency to improve the performance and perceptual relevance of objective PCQA metrics across both dynamic and static point cloud scenarios. In summary, the contributions of this chapter focus on leveraging visual saliency information to enhance objective PCQA metrics. Specifically, they can be summarized as follows:

- We propose an objective PCQA metric that leverages 2D visual saliency maps projected from 3D point clouds, enabling quality assessment to prioritize perceptually important regions.
- We benchmark existing 3D objective PCQA metrics directly using 3D point cloud visual saliency maps collected from our subjective user studies, demonstrating how saliency integration can enhance perceptual alignment.

5.1. Visual saliency guided PCQA metrics for DPC

Subjective quality assessment leads to ground-truth ratings for visual impairments that appear in a stimulus. Subjective quality assessment for DPC has been explored in desktop viewing conditions [84, 117] or in immersive environments with users consuming the contents through an HMD under 6DoF [89, 119]. In the latter case, information about users' movement can be captured in addition to subjective quality ratings, to understand how users navigate and observe objects in VR space. A more accurate representation of

the user's consumption is given by gaze data, which highlights the specific areas of content being viewed with focused attention. This information aids in the creation of visual attention maps. Incorporating visual attention into quality assessment has demonstrated potential improvement for predicting the visual quality of 2D/3D image/video [2, 107]. Nonetheless, visual attention for DPC is still in its infancy, thus hindering the utilization of its outcomes in aiding visual quality assessment.

In Chapter 4, we addressed this gap by developing an eye-tracking-based Quality Assessment and Visual Attention dataset for DPCs (QAVA-DPC). This dataset includes a wide variety of content and compression distortions, utilizing MPEG standard codecs: V-PCC, G-PCC, and the MPEG reference codec (referred to as CWI-PCL).

Building upon this foundation, the present chapter benchmarks objective PCQA metrics enhanced by the integration of visual saliency information from QAVA-DPC. The inclusion of visual saliency maps offers deeper insights into human perceptual behavior in 6DoF environments—an essential component for optimizing QoE. Our approach integrates ground-truth eye-tracking data into point-based PCQA models by using the visual saliency maps as spatial weighting functions. This allows the metrics to give more importance to perceptually relevant regions of the point cloud.

It is important to note that we deliberately exclude learning-based saliency models from this benchmarking study. Since most image or video-based saliency models are not directly designed for dynamic point clouds, applying such models would introduce an additional layer of transformation complexity and possible bias. Our goal here is to keep the saliency information as accurate and reliable as possible by relying on ground-truth eye-tracking data. Therefore, saliency-guided PCQA metrics that rely on the transformation of visual saliency are excluded from this benchmarking for dynamic point clouds to maintain consistency.

Additionally, it is important to clarify that we possess visual saliency maps for both the reference and the distorted point clouds. These saliency maps are used independently to weight each point cloud when computing the corresponding PCQA scores. We do not apply any knn or explicit region correspondence matching between reference and distorted point clouds when incorporating the saliency information, in order to avoid introducing any correspondence-related artifacts.

The main contribution of this chapter lies in benchmarking several state-of-the-art PCQA metrics—originally developed for static point clouds—alongside two temporal pooling methods, on the QAVA-DPC dataset. Furthermore, we validate the performance of these metrics when enhanced with ground-truth visual saliency information. This study provides valuable insights into how saliency-guided PCQA can improve perceptual alignment in dynamic point cloud quality assessment, and offers guidance for the future integration of visual attention in immersive media quality evaluation.

5.1.1. BENCHMARKING OF OBJECTIVE QUALITY METRICS FOR DPC

Dataset and Evaluation Criteria The QAVA-DPC dataset is used to evaluate the added value of incorporating visual saliency information into dynamic point cloud quality assessment. To measure the relationship between objective quality scores and subjective ground-truth scores, we employ three commonly used evaluation criteria: PLCC, SRCC and RMSE.

Performance Evaluation of Objective Quality Metrics without Visual Attention
With the subjective scores collected in our experiments, we conduct an evaluation and comparison of existing objective metrics for the task of DPC quality assessment. Only point-based metrics are considered. We chose metrics adopted by the MPEG group, namely, point-to-point and point-to-plane with MSE and Hausdorff distances, with and without using Peak Signal to Noise Ratio (PSNR), and the weighted average color differences for Y, U, and V channels in terms of PSNR, which is defined as [37]:

$$PSNR_{YUV} = \frac{6 \cdot PSNR_Y + PSNR_U + PSNR_V}{8}. \tag{5.1}$$

Furthermore, we choose another 4 state-of-the-art metrics, namely, histY [45], pointSSIM [145], PCM_RR [163] and PCQM [5]. Since the metrics are originally designed for static contents and do not explicitly consider the temporal aspect, we apply the metrics to each frame and then pool all the 300 frames' scores as the final score for one DPC sequence. Based on the findings of [61], we choose the mean as the baseline and variation model [60] as the temporal pooling method to obtain the objective score for each DPC sequence.

Performance Evaluation of Objective Quality Metrics with Visual Attention
Since the visual saliency is point-based, we integrated the importance weight with the quality score computed by the existing metric for each point per frame, and adopted the same pooling strategy, to obtain the final quality score. The importance weight is normalized between 0-1 based on the computation result in Eq. 4.2. We adopted two ways of weighting the point-based metrics with visual saliency. The first involved considering only the salient region and excluding the unsalient region, which is defined as:

$$Q_v^1 = M \cdot VS_f, \tag{5.2}$$

where Q_v^1 is the quality score with only the salient region for one frame belonging to a DPC, and M is the quality score for the corresponding frame from the point-based metrics. The second method retained the un-salient region but assigned relatively higher weights to the salient region, termed as normalized saliency, which is defined as:

$$Q_v^2 = M \cdot (VS_f + 1), \tag{5.3}$$

where Q_v^2 is the quality score with the normalized saliency for one frame belonging to a DPC.

The results of the performance indices for the 13 objective metrics, with mean and variation as the temporal pooling methods, are presented in Tables 5.1 and 5.2. Consistent with findings in [58], we note that altering the temporal pooling method does not significantly impact high-performing quality metrics (PLCC higher than 0.5). In the original implementation, PCM_RR achieves the highest PLCC/SRCC performance with average pooling, p2point_Hausdroff demonstrates the best performance when only considering the salient region with average pooling, hist_Y achieves the highest SRCC performance after applying normalized salient weighting, and pointSSIM achieves the best PLCC/RMSE and PCQM achieves the best SRCC after normalized saliency. The significant performance increase of p2point_Hausdroff metric may be attributed to the exclusion

of the non-salient region, which helps mitigate the sensitivity to outliers. This refinement ensures that only errors within the salient region are retained. Generally, several studies report that modern video quality assessment models experience only slight improvements from incorporating saliency [237]. However, for dynamic point cloud quality assessment, metrics that focus exclusively on salient regions show a decline in performance. Conversely, metrics using a normalized weighting strategy achieve similar results to the original implementation, with performance also depending on the temporal pooling methods employed.

5.1.2. DISCUSSION

Advancing and enhancing the explanation of DPC quality assessment metrics From the experimental results of the benchmarking, we can see that there is potential for improving the performance of point-based metrics through a strategic combination of visual attention and pooling methods. However, the visual attention data available for DPCs typically covers only a small region of the dense object, limiting its efficacy in advancing DPC quality assessment metrics. Our investigation reveals that employing temporal variation pooling methods leads to decreased performance, prompting further exploration into suitable temporal pooling techniques or spatial-temporal pooling for DPC quality metrics. It is crucial to account for the intrinsic characteristics of DPCs beyond merely relying on video quality indicators. Additionally, careful consideration should be given to matching different distortion types with appropriate temporal pooling methods.

Extending 2D saliency to 3D saliency to help with PCQA Collecting ground-truth 3D saliency data for point clouds is inherently challenging due to the lack of standardized annotation protocols and the difficulty of capturing user attention in 3D space. As a result, directly training a deep learning-based 3D saliency prediction model remains a non-trivial task. A promising alternative is to leverage well-established 2D saliency models by projecting point clouds into multiple 2D views and subsequently mapping the predicted 2D saliency back onto the 3D domain.

The work by [238] highlights the importance of the number and spatial arrangement of viewpoints in this projection-based strategy. Their findings confirm that the selection and diversity of views play a critical role in accurately identifying salient regions in 3D data. Furthermore, this projection-based approach facilitates the generation of pseudo-ground-truth saliency maps for 3D point clouds, which can be used to train or fine-tune deep learning models. It also provides a foundation for applying domain adaptation techniques [239], helping to bridge the domain gap between 2D and 3D representations. These efforts not only enhance the generalization capability of saliency models across modalities but also improve the perceptual quality assessment of point clouds by emphasizing regions that align with human visual attention.

In summary, adapting 2D saliency to 3D through view-based projections offers a practical and scalable solution for 3D saliency estimation. This strategy is particularly valuable for PCQA applications in immersive environments, where identifying and preserving perceptually important regions is crucial for optimizing user experience.

Table 5.1: Performance evaluation of state-of-the-art quality metrics on QAVA-DPC with average pooling. Columns represent Original Implementation (OI), Salient-Part-Only (SPO), and Normalized Saliency Weighting (NSW). The best performance for PLCC/SRCC/RMSE is highlighted in bold and marked with red/blue/orange color respectively.

Metrics	PLCC			SRCC			RMSE		
	OI	SPO	NSW	OI	SPO	NSW	OI	SPO	NSW
p2point_MSE	0.782	0.613	0.781	0.738	0.544	0.738	0.728	0.902	0.730
p2point_MSE_PSNR	0.513	0.580	0.513	0.465	0.555	0.465	1.002	0.929	1.003
p2point_Hausdroff	0.200	**0.854**	0.434	0.240	**0.687**	0.323	1.144	**0.594**	1.052
p2point_Hausdroff_PNSR	0.521	0.661	0.364	0.146	0.641	0.204	0.997	0.856	1.088
p2plane_MSE	0.710	0.571	0.710	0.690	0.546	0.689	0.822	0.937	0.823
p2plane_MSE_PSNR	0.628	0.597	0.484	0.483	0.571	0.483	0.908	0.915	1.026
p2plane_Hausdroff	0.207	0.767	0.428	0.257	0.634	0.281	1.142	0.732	1.055
p2plane_Hausdroff_PNSR	0.514	0.699	0.532	0.163	0.617	0.183	1.005	0.816	0.994
YUV_PSNR	0.661	0.606	0.662	0.654	0.561	0.652	0.876	0.907	0.875
hist_Y	0.840	0.204	0.827	0.820	0.086	0.781	0.633	1.143	0.657
PCM_RR	**0.844**	0.058	0.444	**0.822**	0.046	0.398	**0.626**	1.166	1.046
pointSSIM	0.836	0.652	**0.836**	0.772	0.649	0.771	0.641	0.885	**0.640**
PCQM	0.813	0.334	0.833	0.758	0.256	**0.815**	0.681	1.101	0.646

Table 5.2: Performance evaluation of state-of-the-art quality metrics on QAVA-DPC with variation pooling. Columns represent OI, SPO, and NSW. The best performance for PLCC/SRCC/RMSE is highlighted in bold and marked with red/blue/orange color respectively.

Metrics	PLCC			SRCC			RMSE		
	OI	SPO	NSW	OI	SPO	NSW	OI	SPO	NSW
p2point_MSE	0.769	0.604	**0.769**	0.731	0.536	**0.730**	0.747	**0.909**	**0.751**
p2point_MSE_PSNR	0.267	0.108	0.267	0.135	0.180	0.135	1.125	1.134	1.125
p2point_Hausdroff	0.527	0.593	0.515	0.251	0.536	0.223	0.993	0.919	1.001
p2point_Hausdroff_PNSR	0.248	0.114	0.319	0.232	0.069	0.286	1.131	1.133	1.107
p2plane_MSE	0.662	0.570	0.662	0.621	0.539	0.625	0.875	0.937	0.875
p2plane_MSE_PSNR	0.210	0.103	0.210	0.119	0.134	0.119	1.141	1.134	1.142
p2plane_Hausdroff	0.453	**0.605**	0.501	0.163	0.556	0.217	1.041	**0.909**	1.010
p2plane_Hausdroff_PNSR	0.361	0.054	0.327	0.242	0.005	0.299	1.089	1.139	1.103
YUV_PSNR	0.643	0.601	0.643	0.625	**0.583**	0.625	0.894	0.912	0.894
hist_Y	0.735	0.249	0.669	0.727	0.054	0.617	0.792	1.133	0.868
PCM_RR	0.552	0.393	0.293	0.485	0.216	0.020	0.974	1.074	1.146
pointSSIM	0.705	0.594	0.707	0.710	0.539	0.714	0.828	0.940	0.825
PCQM	**0.807**	0.357	0.732	**0.738**	0.256	0.552	**0.690**	1.091	0.796

Dataset applications and prospective extensions QAVA-DPC, encompassing MOS and DMOS, users' gaze data, and our meticulously processed visual attention maps, hold significant potential as a foundational reference for the multiple aspects. First, since the dataset includes the raw data alongside the visual attention maps, it provides researchers and practitioners with valuable resources to develop and test novel algorithms for post-processing gaze data and creating visual attention maps. Additionally, QAVA-DPC facilitates the comparison of visual saliency maps across different devices (e.g., screen-based

or XR-based), highlighting the need for similarity metrics tailored to DPC. Furthermore, the dataset enables the development of objective quality metrics and visual attention prediction models for DPC without requiring resource-intensive user studies. Moreover, insights derived from the qualitative analysis and visual attention design paradigms can drive advancements in DPC-related research and applications. Finally, existing point-based objective quality metrics can be refined and tailored for DPC to explore how to incorporate visual attention and assess its added value in DPC quality assessment.

5.2. VISUAL SALIENCY GUIDED PCQA METRICS FOR STATIC POINT CLOUD

We have demonstrated that incorporating visual saliency can enhance the perceptual relevance of objective PCQA metrics. However, our analysis also exposed certain limitations when applying saliency information directly to point-based PCQA methods for DPCs. These limitations are primarily due to the sparse and localized nature of eye-tracking data, which often covers only a limited portion of the dynamic scene. Furthermore, our findings suggest that performance variations may stem from multiple intertwined factors —such as the choice of temporal pooling strategy and the use of saliency weighting— making it difficult to isolate their individual effects on metric performance.

This complexity is inherent in DPCs, where quality assessment must account for both spatial and temporal dimensions. The interplay between these factors complicates the interpretation of results and underscores the need for more efficient and flexible methods to integrate saliency into PCQA frameworks. Moreover, the high computational cost and limited alignment between 3D point-based metrics and visual saliency data further constrain their practical applicability.

Motivated by these insights, we pivot our focus to static point clouds and explore image-based NR PCQA. This direction addresses both the methodological challenges observed in the dynamic setting and the practical constraints of real-world applications, where pristine reference point clouds are often unavailable. While point-based metrics excel at evaluating geometric and topological fidelity, they are less compatible with saliency integration and are often computationally demanding. In contrast, image-based approaches allow us to harness well-established 2D IQA techniques by projecting point clouds into images. This not only reduces computational complexity but also enables more effective utilization of saliency information, laying the groundwork for perceptually aligned, reference-free quality assessment methods.

PQA-Net [44] takes 6 orthographic projections of point clouds as inputs, features are extracted after Convolution Neural Network (CNN) blocks, and they share a distortion identification and a quality prediction module that assist in obtaining final quality scores. IT-PCQA [53] utilizes the rich prior knowledge in images and builds a bridge between 2D and 3D perception in the field of quality assessment, a hierarchical feature encoder and a conditional discriminative network is proposed to extract effective latent features and minimize the domain discrepancy. pmBQA [55] proposes an image-based blind quality indicator via multi-modal learning by using four homogeneous modalities (i.e., texture, normal, depth and roughness). MM-PCQA [7] partitions point clouds into sub-models for local geometry representation and renders them into 2D projections for texture. Geometry and

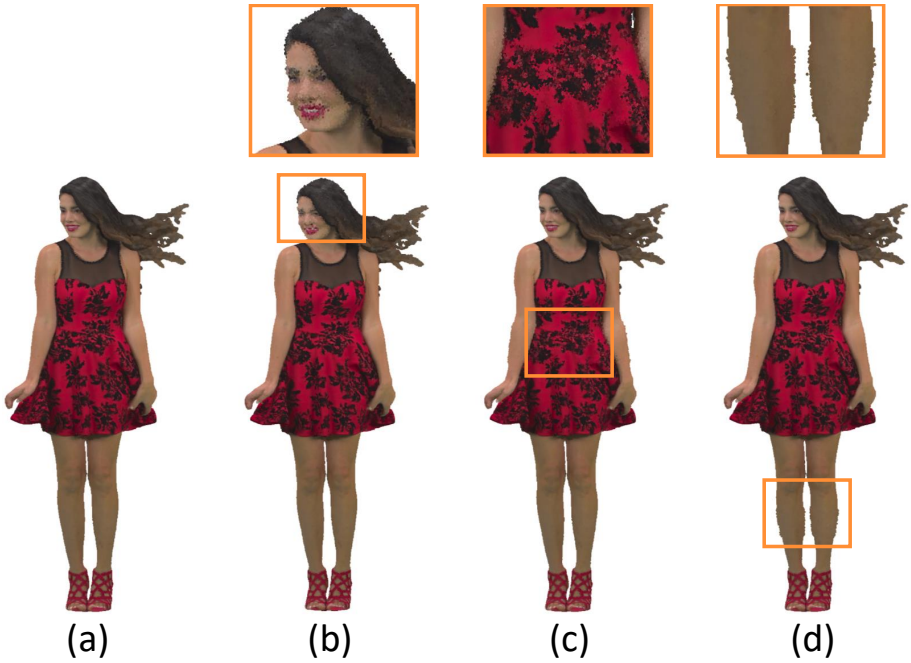


Figure 5.1: Illustration to show the perceptual impact of distortion in different areas on *redandblack* point cloud. (a) is the reference version. (b)-(d) depict the effects of introducing geometry and color Gaussian noise with equal intensity on the face, dress, and legs, respectively. Notably, (c) exhibits nearly identical perceptual quality as the reference point clouds, attributed to the chaotic background texture that effectively masks the distortion. (d) ranks second in perceptual quality, while (a) is observed to have the least favorable perceptual quality.

texture features are extracted separately using point-based and image-based neural networks. A symmetric cross-modal attention module is used for integrating quality-aware information. IT-PCQA [53] reveals the potential connection between different types of media content in the field of quality assessment. PQA-Net [44] and pmBQA [55] use the multi-task decoder and multiple modality-related features on 2D; MM-PCQA [7] proves the effectiveness of cross-modality perception for PCQA.

The aforementioned NR metrics mainly consider the projected images of the point cloud or are completed with 3D point cloud modality. However, they do not consider the impact of visual saliency on improving the prediction accuracy of media content [2]. As illustrated in Figure 5.1, the impact of distortion in different regions of the point cloud (*RedandBlack*) is evident. Recent developments have seen certain metrics incorporating visual saliency into their design paradigms [10]. Some directly extract visual saliency on the 3D point cloud [11, 240], while others employ existing saliency prediction models on 2D projections [109], subsequently re-projecting them onto the 3D point cloud. Visual

saliency is utilized either as a quality indicator or as a weight map for pooling extracted handcrafted features [111], with the aim of selecting features under the guidance of visual saliency. In contrast, our approach utilizes the saliency map from a pre-trained 2D saliency prediction model to guide the selective learning of low-level features, which are extracted by the image encoder. This aims to automatically identify visually salient areas that aid in perceptual quality prediction. Specifically, we propose incorporating depth-related priors into the 2D saliency map to inherently provide a sense of depth for point clouds. Additionally, the low-level feature maps extracted by a CNN-based image encoder, which preserve spatial information, are weighted with the refined saliency map pixel-wisely. The high-level features, which contain semantic information, are processed through a cross-modality attention mechanism to obtain local correspondence and global feature. By concatenating the corrected visual saliency with the local and global embeddings, we generate the final score through two branches: quality score regression and distortion type classification.

As shown in Figure 5.1, the perceptual quality of point clouds is dependent on distortion type since the HVS has different tolerances for different distortions, and where the distortion is located can have a huge impact on the overall quality of point clouds [17]. Thus, the proposed visual saliency guided multi-modal learning can estimate the perceptual quality of point clouds effectively and comprehensively. The main contributions of this subsection are summarized as follows:

- We propose a Visual Saliency guided multimodal NR PCQA (ViSam-PCQA) metric. Visual saliency from the pre-trained model is treated as pseudo-ground-truth, used to correct low-level features that contain learned attention and spatial information. The spatial information is crucial when weighing the visual saliency map with the feature maps from texture, depth and normal maps.
- We utilize the cross-modality attention to obtain the local correspondence among modalities and global features within the same modality, which can compensate for the stereo spatial information loss during the 3D-to-2D projection.
- Extensive experimental evaluations demonstrate that ViSam-PCQA outperforms other state-of-the-art methods. Ablation studies elucidate the distinct contributions of each component within the framework, with a particular emphasis on highlighting the crucial role played by the corrected visual saliency.

5.2.1. FRAMEWORK

The framework overview is exhibited in Figure 5.2. The point clouds are first projected into three different modalities, texture map, depth map and normal map. Then we use an image encoder θ_I to extract the low-level features and high-level features, respectively. Since the primary cues for visual attention often come from the 2D projections captured by the retina [241], depth and normal information are crucial for spatial perception and object localization. We use a pre-trained visual saliency model on the texture image and use its output to correct the low-level feature after an image encoder. At the same time, the texture image, depth image, and normal image are put into the same image encoder

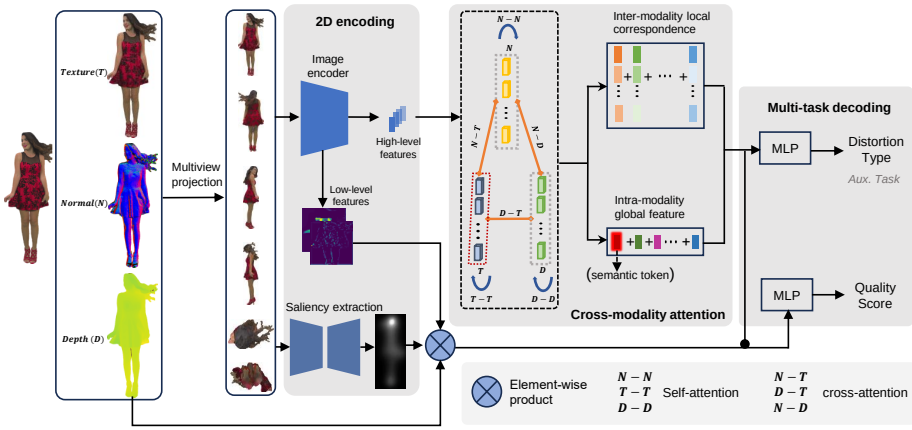


Figure 5.2: The overall framework of our proposed method.

5

to get the semantic feature. Subsequently, the semantic features are put into an intra-and-inter modality attention module to get the global and local features of the point cloud. Finally, the global and local features are concatenated to the distortion type classification branch to learn the distortion-oriented features. The corrected visual saliency with the distortion-oriented features are concatenated and decoded into the quality values via the quality regression branch.

Pre-processing Consider one point cloud denoted as $\mathscr{P} = \{p_{(1)}, p_{(2)}, ..., p_{(i)}\}_{i=1}^{N} \in R^{N\times 6}$, where each point $p_{(i)=[p_i^G, p_i^T]} = [x, y, z, G, R, B]$ indicates the geometry coordinates and the RGB color information, N stands for the number of points belonging to the point cloud. Let $\mathscr{P}$ be orthogonally projected onto M different 2D planes around the bounding box, resulting in M texture maps, $\mathscr{T}_{\Updownarrow} \in R^{H\times W\times 3}$, M depth maps, $\mathscr{D}_{\Updownarrow} \in R^{H\times W\times 1}$, and M normal maps, $\mathscr{N}_{\Updownarrow} \in R^{H\times W\times 3}$, where $m \in M = |\{\text{up, down, left, right, front, back}\}|$ and $H \times W$ denotes the resolution of m_{th} projected image after removing the background. For texture map $\mathscr{T}_{\Updownarrow}$, we calculate the 2D saliency map based on the current state-of-the-art perceptual saliency detection algorithm [242], which is defined as $\mathscr{V}_{\Updownarrow} = \{I_{i,m}\}_{i=1}^{H\times W} \in R^{H\times W\times 1}$, where $I_{i,m}$ denotes for the importance value of the i_{th} pixel from the m_{th} texture map.

Corrected Saliency Map Generation We select the CNN-based image encoder that can retain 2D spatial information [243] at the shallow layers to extract the low-level features from only the texture image. TranSalNet [242] is used to extract the salient area of the texture image, which is defined as

$$\mathrm{V}_m = \phi(\mathscr{T}_m), \tag{5.4}$$

ϕ is the pre-trained TranSalNet model, $\mathrm{V}_m \in R^{H\times W\times 1}$ is the extracted saliency map. Intuitively, in the stereo scenes, human has a preference to the area that is closer to themself [136]. So after obtaining the saliency map, the corresponding depth image is laid on the

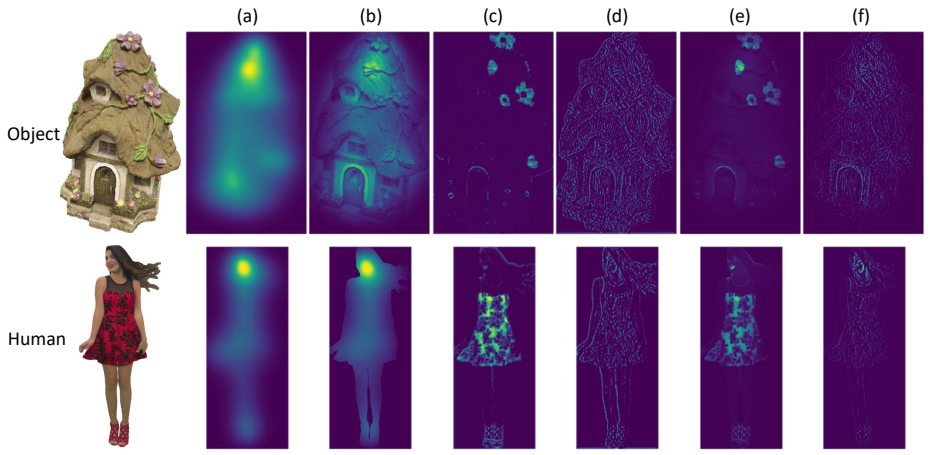


Figure 5.3: Examples of the visual saliency related operations. From (a) to (f) are the saliency map detected by TranSalNet; the depth-guided saliency map; the output of the 1st/256th channel of layer1 in ResNet; the corrected saliency map for the 1st/256th channel, respectively.

up of it, which is expressed as

$$\mathrm{V}_{dm} = \mathrm{V}_m \bigotimes \mathscr{D}_m, \tag{5.5}$$

where $\bigotimes$ is the element-wise product. Subsequently, we utilize the depth-guided saliency map V_{dm} to weight the low-level feature map of all channels. By producing the element-wise product of the pseudo-ground-truth visual saliency and the learned visual saliency through the network [244], the effect of pseudo-ground-truth saliency maps considering the HVS for intervening in the saliency maps learned by the network is achieved, resulting in the corrected saliency maps,

$$\hat{F}_V = Avg(\mathrm{V}_{dm} \bigotimes \mathrm{L}_m^C), \tag{5.6}$$

L_m^C is the shallow layer output of the CNN based image encoder with C channels, $Avg()$ is the average pooling along the feature map and multi-view channels. Figure **??** shows the initial saliency map through the pre-traind model, the depth-related saliency map, the feature map and the corrected saliency map of the first and last channel, respectively. Notably, we computed the saliency map across the entire projection and integrated this global prior with the feature maps from all channels. Each channel captures distinct salient areas based on different filters, as observed in Figure 5.3. For instance, the first channel highlights the texture on the dress as salient, while the last channel emphasizes the contour of the projected image. We enable the network to autonomously learn the allocation of importance with the global saliency prior.

Multi-modal Feature Extraction We next use the same CNN-based image encoder Ψ to extract the high level features from the texture map, depth map, and normal map,

separately, resulting in:

$$\mathrm{H}_m^K = \Psi(K_m), \tag{5.7}$$

where $k \in \{\mathscr{T}, \mathscr{D}, \mathscr{N}\}$, $\mathrm{H}_m^K \in R^d$ is a d-dimension representation.

Global Feature Aggregation and Local Feature Correspondence via Transformer Considering three distinct modalities and an image encoder handling input as image patches, we utilize the Transformer architecture to extract both global and local features. To extract the global features within each modality, the self-attention module is applied to each modality. Besides, similar to BERT [245] and ViT [246], we introduce a learnable semantic token in the self-attention module. The semantic token is shared among the multi-view projections, serving as a global feature of the whole point cloud. For the correspondence among different modalities, we use a symmetrical attention module to explore the relationship of the image patches from different modalities. Here, we take 3 modality-related features as input, and obtain the intra-attention global features F_α^k is defined as

$$\mathrm{F}_\alpha^k = \Theta(\mathrm{Z}^k, \mathrm{Z}^k), \tag{5.8}$$

$\mathrm{Z}^k = [T_s^k, H_1^k, H_2^k, ..., H_M^k] \in R^{(1+M)\times d}$, $T_s \in R^d$ is the sementic token. $\mathrm{F}_\alpha^k \in R^d$ is a d dimensional representation. The inter-modality local features among 3 different modalities F_β^a, are defined as

$$\begin{aligned} \mathrm{F}_\beta^{H^{\mathscr{T}}, H^{\mathscr{D}}} &= \Theta^*(H^{\mathscr{T}}, H^{\mathscr{D}}), \\ \mathrm{F}_\beta^{H^{\mathscr{T}}, H^{\mathscr{N}}} &= \Theta^*(H^{\mathscr{T}}, H^{\mathscr{N}}), \\ \mathrm{F}_\beta^{H^{\mathscr{N}}, H^{\mathscr{D}}} &= \Theta^*(H^{\mathscr{N}}, H^{\mathscr{D}}), \end{aligned} \tag{5.9}$$

Likewise, $\mathrm{F}_\beta^{H^{\mathscr{T}}, H^{\mathscr{D}}}$, $\mathrm{F}_\beta^{H^{\mathscr{T}}, H^{\mathscr{N}}}$ and $\mathrm{F}_\beta^{H^{\mathscr{N}}, H^{\mathscr{D}}}$ are d dimensional representation. The modality-related feature can express the local relationship among the modalities. For example, the local region of texture distortions related to facial features might exhibit a stronger association with the front view rather than the back view of texture distortion. The global feature is the feature map derived from the semantic token. The final quality embedding can be concatenated by the intra-modal global features and the inter-modal local features obtained by:

$$\hat{F}_Q = \hat{F}_g \oplus \hat{F}_l, \tag{5.10}$$

where $\oplus$ indicates the concatenation operation, and $\hat{F}_Q$ represents the final quality-aware features, the global feature $\hat{F}_g$ and local feature $\hat{F}_l$ are defined as follows:

$$\hat{F}_g = \mu(\mu([H_1^k, H_2^k, ..., H_M^k]) + \mathrm{F}_\alpha^k), \tag{5.11}$$

and

$$\hat{F}_l = \mu(H^k + \mathrm{F}_\beta^{H^{\mathscr{T}}, H^{\mathscr{D}}} + \mathrm{F}_\beta^{H^{\mathscr{T}}, H^{\mathscr{N}}} + \mathrm{F}_\beta^{H^{\mathscr{N}}, H^{\mathscr{D}}}), \tag{5.12}$$

in which μ is the mean operation along multi-view channel. The multi-task decoder consists of a Multi Layer Perception (MLP)-based classifier and regressor. The regressor and the classifier are a two- and three-layer ReLU-MLP respectively.

$$\begin{aligned} \hat{Q} &= \vec{D}(\hat{F}_Q), \\ \vec{P} &= \vec{D}(\hat{F}_Q \oplus \hat{F}_V), \\ \hat{P} &= softmax(\vec{P}), \end{aligned} \tag{5.13}$$

where $\hat{Q}$ is the predicted quality score, $\hat{P} = \{\hat{p_1}, \hat{p_2}, ..., \hat{p_E}\}$ is the predicted probability over E distortion types, and $\vec{P}$ is the output of the fully connected layers for distortion type classification before softmax.

For the quality regression task, we focus on minimizing the average prediction error of all training samples and lay importance on the ranking of the quality as [7]. Therefore, the loss function for the regression task includes two parts: MSE and ranking error, which can be derived as:

$$\mathscr{L}_1 = \frac{1}{n}\sum_{e=1}^{n}(q_e - q'_e)^2, \tag{5.14}$$

where q_e is the predicted quality scores, q'_e is the ground truth labels of the point cloud, and n is the size of the mini-batch. The rank loss can better assist the model in distinguishing the quality ranking even the point clouds in a mini-batch have similar quality levels. To this end, we use the differentiable rank function described in [247] to approximate the rank loss:

$$\begin{aligned} \mathscr{L}_2^{ij} &= \max\Big(0, \big|q_i - q_j\big| - e\big(q_i, q_j\big)\cdot\Big(q'_i - q'_j\Big)\Big), \\ e\big(q_i, q_j\big) &= \begin{cases} 1, q_i \geq q_j, \\ -1, q_i < q_j, \end{cases} \end{aligned} \tag{5.15}$$

where i and j are the corresponding indexes for two point clouds in a mini-batch and the rank loss can be derived as:

$$\mathscr{L}_2 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathscr{L}_2^{ij}, \tag{5.16}$$

cross-entropy loss $\mathscr{L}_3$ is used for distortion type classification. Then, the loss function can be calculated as the weighted sum of MSE loss, rank loss and distortion type classification loss:

$$Loss = \lambda_1\mathscr{L}_1 + \lambda_2\mathscr{L}_2 + \lambda_3\mathscr{L}_3 \tag{5.17}$$

where λ_1, λ_2 and λ_3 are used to control the proportion of the MSE loss, the rank loss and distortion type classification loss.

5.2.2. EXPERIMENTS

In this section, we utilized 3 datasets, namely SJTU, WPC and BASICS datasets. The evaluation of performance relies on three standard criteria, including SRCC, PLCC, and prediction accuracy (ACC). Additionally, the logistic regression recommended by the standardization organization [249] is used to map the dynamic range of the scores from the predicted score into the quality label range.

Table 5.3: Performance comparison with state-of-the-art approaches on the SJTU, WPC and BASICS datasets. Best in bold and second with underline. State-of-the-art results for NR-PCQA are cited from the literature, employing varied training strategies and splits, without independent validation by the authors.

Type	Modal Number	Methods	SJTU Dataset		WPC Dataset		BASICS Dataset	
			SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
FR	1	PointSSIM [145]	0.687	0.714	0.454	0.467	0.692	0.725
	1	MSE-p2po [248]	0.729	0.812	0.456	0.485	0.799	0.005
	1	PSNR-yuv [37]	0.795	0.817	0.449	0.530	0.510	0.543
	1	PCQM [5]	0.864	0.885	0.743	0.750	0.810	0.888
	1	GraphSIM [6]	0.878	0.845	0.583	0.616	0.773	0.801
	1	PointPCA [46]	0.907	0.932	0.890	0.894	0.866	0.926
NR	2	IT-PCQA [53]	0.630	0.580	0.540	0.550	0.310	0.302
	1	3D-NSS [50]	0.714	0.738	0.648	0.651	0.617	0.657
	4	pmBQA [55]	0.900	0.932	0.912	0.898	/	/
	2	MM-PCQA [7]	0.910	0.923	0.841	0.856	0.831	0.882
	1	GMS-3DQA [166]	0.911	0.918	0.831	0.834	0.855	0.930
	1	Wang's [154]	0.930	0.940	0.800	0.810	/	/
	1	PKT-PCQA [51]	0.932	0.912	0.557	0.560	/	/
	3	ViSam-PCQA (Ours)	**0.953**	**0.962**	**0.920**	**0.920**	**0.887**	**0.936**

Table 5.4: Ablation study of ViSam-PCQA for key components, i.e., corrected visual saliency, multi modalities that include both the depth and normal map, and distortion type, DT is short for Distortion Type.

Settings	SJTU Dataset		
	SRCC	PLCC	ACC
ViSam-PCQA	**0.953**	0.962	**0.762**
(*Visual Saliency*) /wo corrected saliency maps	0.951	0.962	0.751
(*Modality*) /wo depth & normal maps	0.942	0.952	0.659
(*Distortion Type*) /wo DT classification	0.950	**0.965**	/

Implementation Details

All the projections are rendered with the assistance of Open3D [250], the number of projections for each modality is naturally set to 6. Adam optimizer [251] is utilized with weight decay 1e-4, the initial learning rate is set as 5e-5, the batch size is set as 18, and the model is trained for 100 epochs. The projected images are randomly cropped into image patches at the resolution of 224×224 for all modalities and corresponding saliency maps. The ResNet50 [156] is used as the image encoder, which is initialized with the ImageNet dataset [169].

The multi-head attention module employs 8 heads and the feed-forward dimension is set as 2048. The weights λ_1, λ_2 and λ_3 for $\mathscr{L}_1$, $\mathscr{L}_2$ and $\mathscr{L}_3$ are set as 1.

For relatively small datasets, SJTU (378) and WPC (740), the k-fold cross-validation

strategy is employed to accurately estimate the performance of the proposed method. 9-fold and 5-fold cross validation is selected for SJTU and WPC, respectively. The average performance is recorded as the final result. For the BASICS dataset, we divide the dataset into train-validation-test as the ratio of 6:2:2. There is no content overlap between the training and testing sets. For the FR-PCQA methods that require no training, we simply validate them on the same testing sets and record the average performance.

Overall Performance

14 state-of-the-art PCQA methods are selected for comparison, which consist of 6 FR-PCQA and 8 NR-PCQA methods. The FR-PCQA methods include PointSSIM [145], MSE-p2point (MSE-p2p) [248], PSNR_YUV [37], PCQM [5], GraphSIM [6], and PointPCA [46], these metrics construct and evaluate on a point-to-point comparison or local neighborhood to include the structural information. The NR-PCQA methods include: GMS-3DQA [166], which takes the projections from only the texture. 3D-NSS [50], PKT-PCQA [51], and Wang's metric [154] evaluate the quality directly on the point cloud, the last two adopt multi-task learning, which includes distortion type classification, distortion level regression/classification, and quality regression, respectively. IT-PCQA [53], MM-PCQA [7], pmBQA [55] resolve the PCQA problem with more than one modality.

The results, as shown in Table 5.3, highlight ViSam-PCQA's superior performance across all evaluation criteria on both SJTU and WPC datasets, representing a significant advancement. Notably, the SRCC/PLCC witnessed an increase of 2.2%/5.2% and 0.87%/2.4% when compared with the second-best metric for SJTU and WPC datasets, respectively. Moreover, our model outperforms all FR-PCQA metrics, underscoring its ability to capture essential point cloud characteristics and align closely with the HVS. Summarizing the outcomes, several key conclusions can be drawn: 1) The incorporation of additional modalities (pmBQA) and heightened modality complexity (MM-PCQA) does not consistently result in performance enhancement, suggesting the existence of redundant information that may confound the network. 2) In contrast to models like GPA-Net and Wang's metric, which integrate two auxiliary tasks (distortion type classification and distortion degree regression), our emphasis on the quality regression task with visual saliency related features, suggests that an excessive refinement of auxiliary tasks may not necessarily bolster prediction accuracy. 3) The consistent performance observed on SJTU, WPC and BASICS datasets, despite variations in distortion types and content, underscores the robustness of ViSam-PCQA.

Ablation Study

The SJTU dataset encompasses a variety of contents, including both human figures and inanimate objects, and exhibits a broad range of distortion types. To gain a deeper understanding, we conducted ablation studies on the SJTU dataset by systematically removing key components one at a time.

Impacts of the corrected saliency maps. Quality assessment should align with human perception. Saliency maps highlight regions in an image that are perceptually more important or salient, it directs attention to parts of the point cloud that may have a more

significant visual impact. The SRCC performance has a slight increase on SJTU. Additionally, incorporating the pseudo-ground-truth saliency map with the learning process enables to capture the details and variations in quality that might be overlooked by a uniform weighting approach guided only by the quality score regression. We can see the visual saliency helps the auxiliary task, the accuracy of distortion type classification improves 1.4% on SJTU. This, in turn, can lead to a more nuanced and accurate quality evaluation.

Impacts of the modalities. Combining information from both depth and normal contributes to a more realistic visual representation of the scene [252]. The depth map provides information about the distance of each point in the point cloud from the camera, which can help in assessing the surface details and detecting discontinuities. Normal maps encode surface normals at each point, which can aid in evaluating the smoothness and geometric fidelity. Removing such information will result in an inaccurate estimation for an overall perceptual experience. All criteria performance drops (1.2%, 1.0% and 13.5% for SRCC, PLCC and ACC) for the SJTU dataset. Leveraging depth/normal maps in PCQA provides a multi-faceted approach to evaluating geometric accuracy, surface details, and visual realism.

Impacts of the distortion type classification. We assume that the distortion type classification task can facilitate the quality regression task. However, from Table 5.4 we can see a the SRCC has a slight drop for SJTU datasets after removing the auxiliary task. In summary, depth and normal modalities contribute essential geometric details to enhance the structural integrity of the point cloud. Visual saliency functions as a refinement mechanism, elevating prediction accuracy across all aspects. The efficacy of an additional distortion type classification task is contingent upon the dataset's specific characteristics. Notably, within the proposed framework, the multi-modal completion yields superior performance gains compared to the other two components.

Cross-dataset Evaluation

To gauge the generalization capability of the proposed ViSam-PCQA, cross-dataset evaluations were conducted. Our approach involves training the model on the entire dataset and testing it on all data from another dataset. The resulting performance metrics, presented in Table 5.5, demonstrate the model's ability to generalize across different datasets. Notably, ViSam-PCQA exhibits superior generalization compared to other learning-based models, for example, GPA-Net and MM-PCQA, their performance for SJTU→WPC and WPC→SJTU are 0.424/0.431 and 0.535/0.574, 0.430/0.459 and 0.769/0.778, respectively. Surprisingly, training on the WPC dataset and testing on the SJTU dataset yields even better performance than certain FR-PCQA and NR-PCQA metrics on the SJTU test set, indicating a robust generalization tendency. However, training on BASICS and testing on WPC gets the lowest performance, that's mainly because WPC only contains objects, and the BASICS dataset contains learning-based compression distortion types.

Table 5.5: Cross-dataset evaluation among SJTU, WPC and BASICS datasets. Note the model is validated on the test dataset with all the contents.

Training Dataset	Testing Dataset					
	SJTU		WPC		BASICS	
	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
SJTU	–	–	0.531	0.516	0.488	0.654
WPC	0.788	0.817	–	–	0.608	0.646
BASICS	0.577	0.591	0.393	0.391	–	–

5.2.3. Discussion

Combining multi-modal saliency: ViSam-PCQA leverages visual saliency derived from 2D projections to weight low-level features, aligning the assessment process more closely with human visual perception. This approach underscores the importance of incorporating human attention models into PCQA, suggesting that future metrics could benefit from more sophisticated models of visual attention, possibly combined with eye-tracking data or advanced saliency prediction algorithms. While 2D projections facilitate the application of image-based techniques, they inherently involve a loss of 3D structural information. Future research could investigate methods to mitigate this loss, such as incorporating 3D saliency maps or developing hybrid models that combine 2D and 3D analyses. Additionally, exploring alternative projection techniques that preserve more geometric information could enhance the fidelity of quality assessments.

Incorporating User-Centric Quality Metrics: Combining the captured visual saliency through subjective study and the objective saliency prediction model. We find that Traditional PCQA metrics often rely on objective measures that may not fully capture user perceptions of quality. Integrating subjective quality assessments, such as user studies or crowd-sourced evaluations, could provide valuable insights into the alignment between metric predictions and human judgments. This user-centric approach could inform the development of more perceptually relevant PCQA models.

5.3. Conclusion

In this chapter, we addressed *R3: What is the added value of visual saliency for PCQA metrics?* through two complementary studies: (1) the integration of user gaze data for dynamic point clouds, and (2) the use of predicted saliency maps for static point clouds via learning-based models.

We systematically evaluated existing objective PCQA metrics with and without visual saliency integration to assess saliency's utility for dynamic point cloud quality assessment. While prior studies demonstrate saliency's benefits in 2D image/video quality prediction, this work pioneers its exploration for dynamic point clouds. Our results confirm that saliency-aware metrics enhance performance, but their design demands careful consideration of: (i) metric construction principles, (ii) feature redundancy, (iii) temporal

dynamics in dynamic PCs, and (iv) effective saliency integration within the metric architecture.

To further advance this direction, directly combining saliency information into PCQA metrics for dynamic point clouds involves numerous factors that must be carefully addressed. To better harness the power of saliency information, we proposed a novel saliency-guided, multi-modal NR PCQA metric, ViSam-PCQA, designed for static point clouds. ViSam-PCQA integrates a saliency map generated from a pre-trained model into its learning pipeline to enhance quality prediction. The metric leverages multi-modal inputs—including texture, depth, and normal maps—encoded into both low- and high-level feature representations. The texture features are refined through depth-guided saliency maps to emphasize perceptually important regions, while a Transformer module extracts global and local cross-modal correspondences. Finally, quality regression and distortion type classification are performed to produce the overall quality score. Extensive evaluations on three public datasets demonstrate that ViSam-PCQA achieves substantial improvements over existing state-of-the-art methods.

Looking forward, we aim to develop advanced saliency prediction models for both dynamic and static point clouds and to further refine saliency-aware PCQA metrics. A key research direction is to design metrics in which visual saliency and quality assessment are jointly optimized, ensuring that they complement each other. Moreover, rather than treating saliency as an add-on at the final stage of the processing pipeline, future systems should aim to incorporate saliency information progressively and modularly, starting from early stages of point cloud processing, to support real-world applications.

In the next chapter, we conclude the thesis by revisiting the central research questions and summarizing the key discussions and findings presented throughout the work. Additionally, we introduce the datasets developed using the experimental protocols proposed in the previous chapters, alongside other publicly available resources created over the course of this research. These contributions aim to support future work in point cloud quality assessment, visual saliency, and related research.

6

Conclusion

Throughout this thesis, we have proposed and evaluated a series of PCQA metrics aimed at quantifying the perceptual quality of point clouds, with a particular focus on capture and the interplay between geometric structure and texture information. Our work systematically examined the individual and combined contributions of these attributes to perceptual quality. In addition, we conducted detailed studies on the visual saliency of dynamic point clouds within immersive VR/AR environments. Beginning in Chapter 3, we refined similarity measurement approaches and analyzed how geometric and textural features jointly influence quality perception. In Chapter 4, we introduced a subjective evaluation protocol tailored to dynamic point clouds in immersive scenarios and investigated the influence of task-driven conditions on visual attention distribution. Building on these insights, Chapter 5 extended the objective PCQA models by incorporating findings from the subjective experiments and leveraging the generated visual saliency maps. In this final chapter, we first revisit the research questions posed in the introductory chapter. We then reflect on the key lessons learned, summarize the contributions and resources developed through this work, and discuss their potential impact on future research. Finally, we highlight the limitations of the current study and outline promising directions for future investigation.

This chapter is based on the following publications:

1. *Minh Nguyen, Shivi Vats, **Xuemei Zhou**, Irene Viola, Pablo Cesar, Christian Timmerer, Hermann Hellwagner. 2024. ComPEQ-MR: Compressed Point Cloud Dataset with Eye Tracking and Quality Assessment in Mixed Reality. Proceedings of the 15th ACM Multimedia Systems Conference (MMSys).* [23]
2. *Marouane Tliba, **Xuemei Zhou**, Irene Viola, Pablo Cesar, Aladine Chetouani, Giuseppe Valenzise and Dufaux, Frédéric. 2024. Enhancing Immersive Experiences through 3D Point Cloud Analysis: A Novel Framework for Applying 2D Visual Saliency Models to 3D Point Clouds. In 2024 16th International Conference on Quality of Multimedia Experience (QoMEX).* [253]
3. *Guillaume Gautier, **Xuemei Zhou**, Thong Nguyen, Jack Jansen, Louis Fréneau, Marko Viitanen, Uyen Phan, Jani Käpylä, Irene Viola, Alexandre Mercat, Pablo Cesar, Jarno Vanne. 2025. UVG-CWI-DQPC: Dual-Quality Point Cloud Dataset for Volumetric Video Applications. Proceedings of the 33rd ACM International Conference on Multimedia (ACM MM).* [254]

6.1. Thesis Summary

This thesis explored the perceptual quality assessment of point clouds through a combination of objective modeling, visual saliency detection and related applications, and subjective validation within immersive environments. Our goal is to improve the alignment between objective metrics and human perception, thereby contributing to more accurate and perceptually meaningful PCQA, which can be applied to real-life applications. The main research question guiding this thesis was:

How can the perceptual quality of 3D point clouds be accurately evaluated both subjectively and objectively?

To address this overarching question, we decomposed it into three sub-questions.

To answer *R1: How can we measure the perceptual quality of static point clouds under various distortion types?* Chapter 3 introduces and evaluates two objective PCQA metrics: PointPCA+ and M3-Unity. PointPCA+ is a PCA-based metric that focuses on capturing the dominant geometric features of point clouds. It offers computational efficiency and effective handling of geometric distortions while consistently outperforming state-of-the-art solutions. This metric provides valuable insights into the design of similarity measurements, particularly given the challenges posed by differences in the number of points and distributions between reference and distorted point clouds. Building on these findings, M3-Unity adopts an integrated approach by combining both attributes and modality. This deep-learning–based method enhances sensitivity to a wider range of distortions through high-level semantic features, but it incurs higher computational costs and requires more data for training. Moreover, M3-Unity provides guidance by analyzing the relative contributions of geometric and textural attributes for different distortion types, thereby complementing and extending the capabilities of PointPCA+. Both metrics consistently outperformed traditional methods across several benchmarks.

To answer *R2: How can visual saliency in dynamic point clouds be detected and compared in immersive environments?* Chapter 4 focuses on the subjective investigation of visual saliency in dynamic point clouds within immersive VR environments featuring 6 DoF. This chapter begins by creating two specialized datasets: one for simultaneous visual saliency and quality assessment of dynamic point clouds, and another for visual saliency evaluation under free-viewing conditions without task constraints. Both datasets are developed using meticulously designed experimental protocols based on ITU recommendations for immersive media to ensure accurate collection of gaze data and mean opinion scores. Comprehensive qualitative and quantitative analyses are conducted, which mutually corroborate, revealing how task demands impact visual attention distribution and elucidating the interplay between dynamic visual saliency and point cloud distortions. These findings not only deepen our understanding of perceptual factors and human behavior from the users' perspective in VR-based PCQA but also provide a robust foundation for enhancing quality assessment and visual saliency detection methodologies in immersive environments, ultimately guiding future research in this emerging field.

To answer *R3: What is the added value of visual saliency for PCQA metrics?* Chapter 5 investigates this question by integrating visual saliency into the objective quality assessment framework. This integration is a natural step toward improving the performance of PCQA metrics, as it aligns computational evaluation more closely with human visual perception. We first applied the collected saliency maps of dynamic point clouds to existing

point-based static PCQA metrics using two different pooling strategies. However, the benchmarking results did not show consistent performance improvements. This can be attributed to the inherent complexity of dynamic point cloud quality assessment, where multiple factors influence perceptual quality, making accurate prediction more challenging than in the static case. Consequently, we shifted our focus to static point clouds and utilized learned visual saliency models—benefiting from the maturity of saliency prediction in 2D image domains. This allowed us to systematically evaluate whether saliency information contributes meaningfully to PCQA performance. The results demonstrated that the saliency-enhanced metrics consistently outperformed traditional approaches, particularly in their ability to detect subtle distortions. These findings confirm that integrating visual saliency helps bridge the gap between objective assessment and human perception. Overall, the analyses reinforce the idea that combining standard feature-based methods with perceptual attention modeling presents a promising direction for advancing PCQA methodologies, also in the context of immersive media.

6.2. Discussion

This work represents one of the earliest efforts to bridge visual saliency modeling and algorithmic quality assessment in the context of point clouds within immersive media environments. Through this multidisciplinary exploration, several key insights and lessons have emerged that hold implications for both future research and applied system development.

Visual Saliency Prediction and Application

Task-Relevant Saliency is Contextual In this thesis, visual saliency of dynamic point clouds was derived exclusively through user studies. One of the most notable findings was the significant variation in saliency distributions under different task conditions. Specifically, we observed both distinct differences and notable overlaps between free-viewing scenarios and task-oriented (quality assessment) settings. While task demands influenced users to focus on certain regions more strategically in the quality assessment scenario, some consistently salient areas were shared across both conditions, indicating intrinsic visual importance regardless of task. These results emphasize that visual saliency is not only task-dependent but also contains stable components, and thus, experimental protocols must be carefully designed to account for this variability. Recognizing the influence of task context is essential for the development of future applications, including perceptual compression, adaptive streaming, and attention-aware rendering in XR.

Visual saliency datasets are often created with different objectives and task contexts in mind—for example, memorization, quality assessment, visual search, and others. As a result, it is challenging to select a single universal saliency map suitable for all applications. Ultimately, there is no "best" saliency map, but rather the *most suitable* one for a given context. When comparing visual saliency maps generated under the same task context, it is important to adopt an appropriate evaluation approach. Point-to-point comparison metrics are often inadequate for capturing the semantic information present in 3D point cloud saliency maps—especially in VR environments, where device-related

and procedural biases are common. In such cases, distribution-based metrics offer a more reliable and robust means of evaluating 3D point cloud saliency maps in VR.

Visual Saliency for Point Cloud Streaming Predicting visual saliency in point clouds remains a challenging and resource-intensive task. Accurate modeling of the HVS is inherently complex, and current predictive methods have yet to achieve satisfactory reliability. In our work, significant effort was devoted to generating high-quality saliency maps at considerable computational and operational cost—often for only marginal performance gains. This raises a critical consideration: while visual saliency is valuable, it is not the primary objective in point cloud streaming or quality assessment. Rather, saliency should be viewed as an auxiliary tool that enhances core tasks when applied strategically.

To maximize its utility, we propose that visual saliency be treated as a hierarchical and context-aware element within the point cloud processing pipeline. Coarse estimation of salient regions can be integrated in early stages such as pre-processing or compression, enabling low-cost optimization strategies. Fine-grained and accurate saliency prediction should be reserved for specific use cases where clear performance benefits can be demonstrated. Furthermore, under varying contextual and application-specific conditions, we advocate for a broader definition of ”high-importance” regions that may combine frequency of use, motion characteristics, and visual saliency.

In summary, for visual saliency to contribute effectively without disproportionate cost, it should be embedded into the point cloud processing workflow in a modular and adaptive manner—starting from early-stage operations rather than being confined to end-stage refinement. This procedural integration enables a better cost-benefit tradeoff and ensures that saliency-driven techniques deliver maximal impact where they are most needed.

Incorporating Human-centric Paradigm into Subjective Quality Assessment

This work sits at the intersection of signal processing, perceptual psychology, and human-computer interaction. Such interdisciplinary integration was essential to formulate holistic solutions. One particularly valuable insight emerged from incorporating qualitative, user-centered analysis alongside the standard experimental process [199, 255, 256]. Existing standards typically define outlier rejection procedures for MOS based purely on statistical analysis of user ratings. However, this approach has certain limitations: it may overlook content ambiguity and fail to account for individual differences in subjective ratings [199]. By analyzing questionnaire data collected before and after the subjective study, we were able to refine outlier rejection from an alternative, more perceptually informed perspective. Furthermore, thematic analysis of participant interviews provided valuable insights into key perceptual factors—for example, identifying which distortions were most noticeable or which content areas consistently attracted visual attention. These findings directly informed subsequent stages of feature engineering and experimental design, particularly for XR applications where perceptual nuances are critical. Compared to relying solely on extensive algorithmic experiments and performance-driven feature selection, incorporating user-derived insights enabled us to craft perceptually meaningful

features. This approach helped to narrow the feature pool, saving both time and computational resources while improving the interpretability of the model.

Finally, conducting a small-scale user study at the final stage of the proposed algorithm is essential to incorporate end-user feedback and validate its effectiveness. This ensures that the algorithm ' s performance aligns with human perceptual expectations and supports practical adoption.

This contrasts with purely data-driven strategies that rely on exhaustive computational feature extraction followed by statistical ranking or factor analysis. While such methods are useful, early engagement with participants often accelerated insight generation and improved experimental robustness. This underscores the complementary power of qualitative and quantitative methods in perceptual media research.

Incorporating Human-centric Paradigm into Objective Quality Assessment

Designing objective PCQA metrics that genuinely reflect human perception requires more than mathematical optimization—this is the ultimate goal of our work. A practical and meaningful metric requires grounding in perceptual and cognitive models. A recurring observation throughout this research was that feature sets optimized solely for numerical correlation with MOS often diverged from the features that users intuitively perceived as important. However, understanding user-driven guidelines alone is not sufficient. The irregular and unstructured nature of point clouds makes it inherently challenging to design and optimize features capable of effectively capturing distortions, whether through handcrafted approaches or data-driven learning. Furthermore, the highly non-linear and complex characteristics of the HVS add yet another layer of difficulty to this process. This disconnect highlights the necessity of integrating perceptual mechanisms, such as visual saliency and HVS-inspired modeling, into the development of objective quality metrics. Such integration is essential for creating assessment methods that are not only mathematically robust but also perceptually meaningful and aligned with the expectations of end users.

Challenges and Considerations for Building Point Cloud Datasets

High-end capture systems can produce highly realistic 3D representations, but their significant cost and complex setup limit their practical deployment and widespread adoption. In contrast, consumer-grade capture systems, which typically use lower-resolution cameras and fewer capture angles, tend to produce sparser point clouds with more occlusions and noise. While quality limitations in consumer-grade data can be mitigated through techniques such as occlusion removal, point cloud densification, and accurate camera registration, the development and evaluation of such techniques rely heavily on the availability of appropriate datasets.

Another important category of point cloud datasets incorporates user behavioral data (e.g., gaze tracking or navigation patterns) to support adaptive rendering, streaming optimization, and perceptual quality analysis. These datasets are often built on top of existing

point cloud sequences, but they fundamentally require access to the original, raw captured data as a basis for secondary studies aimed at modeling user engagement and visual attention in immersive environments.

However, capturing such raw datasets is both expensive and time-consuming. Moreover, the post-processing of point cloud data typically requires multiple complex algorithms and often involves extensive manual verification to ensure data quality. Therefore, it is critical to invest effort into improving the raw data capture process itself, with the goal of automating as much of it as possible. Increased automation would enable the creation of larger-scale datasets that can support robust evaluation and benchmarking of new algorithms.

Finally, when proposing the capture of a new dataset, it is essential to clearly justify its purpose and contribution, ensuring that the effort is not merely a replication of existing work but instead provides added value to the community. This is particularly important because point cloud dataset capture remains a resource-intensive process, and meaningful advancements in dataset design can have a lasting impact on future research.

Reflections on Immersive VR and Eye Tracking

Conducting studies in immersive VR environments with integrated eye tracking introduced practical and methodological challenges that were initially underestimated. Devices such as the HTC Vive Pro Eye [257] offer promising specifications; however, their real-world usage revealed discrepancies between manufacturer-reported performance and actual reliability during extended user sessions. Issues such as HMD slippage, head motion drift, and calibration instability often affected gaze tracking accuracy and data consistency.

These challenges highlight the necessity for robust error profiling, calibration protocols, and post-processing validation when conducting eye-tracking-based studies in immersive environments. Furthermore, our experience suggests that reported accuracy metrics should be interpreted conservatively—especially in research that relies on fine-grained gaze estimation for perceptual modeling.

In summary, the discussion points raised in this chapter reflect both the conceptual depth and the practical complexity of conducting perceptual quality assessment in immersive 3D settings. They also point to promising directions for enhancing methodological rigor, user-centered design, and system-level robustness in future work.

6.3. Future Work

Future research should continue bridging the gap between subjective and objective PCQA by refining both methodological foundations and technical implementations. A key priority is the design of controlled subjective studies to systematically isolate perceptual factors influencing quality judgments. Insights derived from these experiments—both qualitative and quantitative—can guide the development of perceptually aligned computational models. These models should integrate global and structural features informed by HVS principles.

From an implementation perspective, improving the efficiency and scalability of objective metrics is essential. Future work should explore strategies such as parallel compu-

tation, model pruning, or approximation schemes to enable real-time processing of point cloud data in XR pipelines

In the context of saliency-guided PCQA, further effort is needed to adapt and extend saliency models from image and video domains to point clouds. One promising direction is the projection of point cloud saliency into 2D views, which enables the application of mature image-based quality metrics and facilitates cross-domain validation. Moreover, with an accurate saliency map available for the reference point cloud, it may be feasible to infer the saliency map of its distorted counterpart by modeling typical degradation-induced shifts in attention. While not exact, such approximations can reduce the need for costly and time-intensive eye-tracking studies.

Ultimately, the field would benefit from a unified framework that integrates perceptual attention, multimodal features, and efficient computation. To this end, we identify the following key directions:

- **Towards real-time PCQA:** Current state-of-the-art methods often rely on complex feature extraction or deep neural networks that are not optimized for real-time execution, especially on resource-constrained devices such as standalone HMDs or mobile platforms. Future work should investigate lightweight architectures, including the use of efficient backbone networks, feature distillation, or pruning strategies to reduce inference time. Moreover, exploiting parallel computing frameworks (e.g., GPU-accelerated pipelines or neural processing units) and adopting model quantization or compression can enable faster runtime without significantly sacrificing quality prediction accuracy. Integrating such real-time PCQA models into live immersive streaming or rendering systems would open up possibilities for dynamic quality control and adaptive user feedback.

- **Unified perceptual models:** PCQA is inherently multimodal, involving the interplay of geometric structure, surface texture, and visual attention. Yet, most existing PCQA models treat these modalities in isolation or rely on handcrafted heuristics for integration. A promising direction is the development of unified perceptual models that jointly learn from geometry, color, and saliency information in a coherent, data-driven manner. Such models should be designed to capture complex interactions between spatial fidelity, appearance features, and human gaze patterns. Moreover, they should be robust across a wide range of content types (e.g., human avatars, objects, scenes) and distortion types (e.g., compression, noise, rendering artifacts). Incorporating attention mechanisms, such as cross-modal transformers or saliency-guided feature fusion, may further enhance model expressiveness. The ultimate goal is to create generalizable and scalable PCQA models that not only achieve high correlation with human judgments but also serve as foundational tools for perceptual optimization in XR pipelines.

- **Standardized benchmarking:** From the subjective study perspective, there is currently no standardized methodology for conducting user studies or validating experimental results in the PCQA domain. Our work includes two user studies —one exploring the relationship between perceptual quality and visual attention, and another comparing visual attention under different tasks. These studies were designed following ITU recommendations for immersive quality assessment and

built on prior experience in eye-tracking experiments [99, 209, 258]. However, many factors in XR environments (e.g., viewing distance, rendering algorithms) may affect perceived quality and require further investigation—especially in terms of how to incorporate them into a unified experimental framework. Moreover, the algorithm used to generate dynamic point cloud saliency maps requires cross-validation against existing methods to ensure its accuracy. Establish publicly available datasets that include synchronized quality scores, eye-tracking data, and a wide range of distortions to support reproducibility and comparative evaluation, and broader adoption.

- **Cross-domain saliency transfer:** The integration of visual saliency to enhance perceptual quality prediction remains an open challenge. Saliency prediction for point clouds is highly dependent on accurate saliency modeling, which is itself underdeveloped. Current models lack robustness, particularly for dynamic point clouds. Explore transfer learning strategies to bridge 2D and 3D saliency prediction, minimizing the dependence on domain-specific training data.

- **Beyond quality assessment:** The perceived quality of point clouds is strongly influenced by the mode of content consumption, particularly in XR environments, where users typically interact with content through HMD. Different XR scenarios introduce varying degrees of perceptual masking and attention dynamics, which can significantly affect quality perception. However, our current understanding of human behavior in immersive 3D environments remains limited. This knowledge gap constrains the development of perceptually-driven PCQA methodologies that are truly aligned with end-user experience. One of the future works should focus on evaluating perceptual quality in a scenario-specific manner, taking into account user interactions, viewing behaviors, and contextual cues. Furthermore, insights from perceptual modeling can be extended to support intelligent rendering strategies, such as foveated streaming and user-adaptive quality control, enabling more efficient and immersive XR experiences.

This thesis provides an initial step toward bridging computational quality assessment and human visual perception for point clouds. Through new datasets, perceptually motivated metrics, and immersive eye-tracking studies, it contributes a structured framework for future exploration. It is our hope that these insights will support and inspire further advancements in perceptually driven 3D media technologies.

Bibliography

[1] K. W. Tesema, L. Hill, M. W. Jones, M. I. Ahmad and G. K. Tam. 'Point Cloud Completion: A Survey'. In: *IEEE Transactions on Visualization and Computer Graphics* (2023).

[2] W. Lin and C.-C. J. Kuo. 'Perceptual visual quality metrics: A survey'. In: *Journal of visual communication and image representation* 22.4 (2011), pp. 297–312.

[3] Y. Chen. *Continuity in 3D Visual Learning*. ASCI dissertation series. Yunlu Chen, 2023. URL: https://books.google.nl/books?id=MzA00AEACAAJ.

[4] A. Tious, T. Vigier and V. Ricordel. 'New challenges in point cloud visual quality assessment: a systematic review'. In: *Frontiers in Signal Processing* 4 (2024), p. 1420060.

[5] G. Meynet, Y. Nehmé, J. Digne and G. Lavoué. 'PCQM: A Full-Reference Quality Metric for Colored 3D Point Clouds'. In: *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*. 2020, pp. 1–6. DOI: 10.1109/QoMEX48832.2020.9123147.

[6] Q. Yang, Z. Ma, Y. Xu, Z. Li and J. Sun. 'Inferring Point Cloud Quality via Graph Similarity'. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020), pp. 1–1. DOI: 10.1109/TPAMI.2020.3047083.

[7] Z. Zhang, W. Sun, X. Min, Q. Zhou, J. He, Q. Wang and G. Zhai. 'MM-PCQA: Multi-Modal Learning for No-reference Point Cloud Quality Assessment'. In: *arXiv preprint arXiv:2209.00244* (2022).

[8] Y. Liu, Y. Zhang, Z. Shan and Y. Xu. 'CLIP-PCQA: Exploring Subjective-Aligned Vision-Language Modeling for Point Cloud Quality Assessment'. In: *arXiv preprint arXiv:2501.10071* (2025).

[9] Z. Wang, H. R. Sheikh, A. C. Bovik *et al.* 'Objective video quality assessment'. In: *The handbook of video databases: design and applications*. Vol. 41. Citeseer, 2003, pp. 1041–1078.

[10] W. Lin, S. Lee *et al.* 'Visual saliency and quality evaluation for 3D point clouds and meshes: An overview'. In: *APSIPA Transactions on Signal and Information Processing* 11.1 (2022).

[11] A. Laazoufi and M. El Hassouni. 'Saliency-based point cloud quality assessment method using aware features learning'. In: *2022 9th International Conference on Wireless Networks and Mobile Communications (WINCOM)*. IEEE. 2022, pp. 1–5.

[12] T. Zheng, C. Chen, J. Yuan, B. Li and K. Ren. 'Pointcloud saliency maps'. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 1598–1606.

[13] X. Ding, W. Lin, Z. Chen and X. Zhang. 'Point cloud saliency detection by local and global feature fusion'. In: *IEEE Transactions on Image Processing* 28.11 (2019), pp. 5379–5393.

[14] E. Alexiou, P. Xu and T. Ebrahimi. 'Towards modelling of visual saliency in point clouds for immersive applications'. In: *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE. 2019, pp. 4325–4329.

[15] S. Karim, H. He, A. A. Laghari and H. Madiha. 'Quality of Service (QoS): Measurements of Video Streaming'. In: *International Journal of Computer Science Issues* 16.6 (Nov. 2019). DOI: 10.5281/zenodo.3987056. URL: https://doi.org/10.5281/zenodo.3987056.

[16] T. Yamazaki. 'Quality of experience (QoE) studies: Present state and future prospect'. In: *IEICE Transactions on Communications* 104.7 (2021), pp. 716–724.

[17] L. Zhang, Y. Shen and H. Li. 'VSI: A visual saliency-induced index for perceptual image quality assessment'. In: *IEEE Transactions on Image processing* 23.10 (2014), pp. 4270–4281.

[18] W. Zhang, A. Borji, Z. Wang, P. Le Callet and H. Liu. 'The application of visual saliency models in objective image quality assessment: A statistical evaluation'. In: *IEEE transactions on neural networks and learning systems* 27.6 (2015), pp. 1266–1278.

[19] X. Zhou, Y. Zhang, N. Li, X. Wang, Y. Zhou and Y.-S. Ho. 'Projection Invariant Feature and Visual Saliency-Based Stereoscopic Omnidirectional Image Quality Assessment'. In: *IEEE Transactions on Broadcasting* 67.2 (2021), pp. 512–523. DOI: 10.1109/TBC.2021.3056231.

[20] X. Zhou, E. Alexiou, I. Viola and P. Cesar. 'PointPCA+: Extending PointPCA Objective Quality Assessment Metric'. In: *2023 IEEE International Conference on Image Processing Challenges and Workshops (ICIPCW)*. 2023, pp. 1–5. DOI: 10.1109/ICIPC59416.2023.10328338.

[21] X. Zhou, E. Alexiou, I. Viola and P. Cesar. 'PointPCA+: A full-reference Point Cloud Quality Assessment metric with PCA-based features'. In: *Signal Processing: Image Communication* 135 (2025), p. 117262. ISSN: 0923-5965. DOI: https://doi.org/10.1016/j.image.2025.117262. URL: https://www.sciencedirect.com/science/article/pii/S0923596525000098.

[22] X. Zhou, I. Viola, Y. Chen, J. Pei and P. Cesar. 'Deciphering Perceptual Quality in Colored Point Cloud: Prioritizing Geometry or Texture Distortion?' In: *Proceedings of the 32nd ACM International Conference on Multimedia*. 2024, pp. 7813–7822.

[23] M. Nguyen, S. Vats, X. Zhou, I. Viola, P. Cesar, C. Timmerer and H. Hellwagner. 'Compeq-mr: compressed point cloud dataset with eye tracking and quality assessment in mixed reality'. In: *Proceedings of the 15th ACM Multimedia systems conference*. 2024, pp. 367–373.

[24] X. Zhou, I. Viola, E. Alexiou, J. Jansen and P. Cesar. 'QAVA-DPC: eye-tracking based quality assessment and visual attention dataset for dynamic point cloud in 6 DoF'. In: *2023 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE. 2023, pp. 69–78.

[25] X. Zhou, I. Viola, S. Rossi and P. Cesar. 'Comparison of Visual Saliency for Dynamic Point Clouds: Task-free vs. Task-dependent'. In: *IEEE Transactions on Visualization and Computer Graphics* (2025), pp. 1–11. DOI: 10.1109/TVCG.2025.3549863.

[26] X. Zhou, I. Viola, R. Yin and P. Cesar. 'Visual-Saliency Guided Multi-modal Learning for No Reference Point Cloud Quality Assessment'. In: *Proceedings of the 3rd Workshop on Quality of Experience in Visual Multimedia Applications*. 2024, pp. 39–47.

[27] X. Zhou, I. Viola, E. Alexiou, J. Jansen and P. Cesar. 'Subjective and Objective Quality Assessment for Dynamic Point Cloud with Visual Attention in 6 DoF'. In: *ACM Transactions on Multimedia Computing, Communications and Applications* (2025).

[28] Z. Wang, A. C. Bovik, H. R. Sheikh and E. P. Simoncelli. 'Image quality assessment: from error visibility to structural similarity'. In: *IEEE Transactions on Image Processing* 13.4 (2004), pp. 600–612.

[29] Z. Li, K. Swanson, C. Bampis, L. Krasula and A. Aaron. 'Toward a Better Quality Metric for the Video Community'. In: *Netflix Technology Blog* (Dec. 2020). URL: https://netflixtechblog.com/toward-a-better-quality-metric-for-the-video-community-7ed94e752a30.

[30] R. Mekuria, K. Blom and P. Cesar. 'Design, Implementation, and Evaluation of a Point Cloud Codec for Tele-Immersive Video'. In: *IEEE Transactions on Circuits and Systems for Video Technology* 27.4 (2017), pp. 828–842. DOI: 10.1109/TCSVT.2016.2543039.

[31] D. Tian, H. Ochimizu, C. Feng, R. Cohen and A. Vetro. 'Geometric distortion metrics for point cloud compression'. In: *IEEE ICIP*. 2017, pp. 3460–3464.

[32] A. Javaheri, C. Brites, F. Pereira and J. Ascenso. 'Improving PSNR-Based Quality Metrics Performance For Point Cloud Geometry'. In: *2020 IEEE International Conference on Image Processing (ICIP)*. 2020, pp. 3438–3442. DOI: 10.1109/ICIP40778.2020.9191233.

[33] A. Javaheri, C. Brites, F. Pereira and J. Ascenso. 'A Generalized Hausdorff Distance Based Quality Metric for Point Cloud Geometry'. In: *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*. 2020, pp. 1–6. DOI: 10.1109/QoMEX48832.2020.9123087.

[34] E. Alexiou, Y. Nehmé, E. Zerman, I. Viola, G. Lavoué, A. Ak, A. Smolic, P. Le Callet and P. Cesar. 'Chapter 18 - Subjective and objective quality assessment for volumetric video'. In: *Immersive Video Technologies*. Ed. by G. Valenzise, M. Alain, E. Zerman and C. Ozcinar. Academic Press, 2023, pp. 501–552. ISBN: 978-0-323-91755-1. DOI: https://doi.org/10.1016/B978-0-32-391755-1.00024-9.

[35] E. Alexiou and T. Ebrahimi. 'Point Cloud Quality Assessment Metric Based on Angular Similarity'. In: *2018 IEEE International Conference on Multimedia and Expo (ICME)*. 2018, pp. 1–6. DOI: 10.1109/ICME.2018.8486512.

[36] R. L. de Queiroz and P. A. Chou. 'Motion-Compensated Compression of Dynamic Voxelized Point Clouds'. In: *IEEE Transactions on Image Processing* 26.8 (2017), pp. 3886–3895. DOI: 10.1109/TIP.2017.2707807.

[37] 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'. In: *Applications of Digital Image Processing XLI*. Vol. 10752. 2018, pp. 174–190.

[38] E. Alexiou and T. Ebrahimi. 'Exploiting user interactivity in quality assessment of point cloud imaging'. In: *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE. 2019, pp. 1–6.

[39] 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'. In: *IEEE Transactions on Multimedia* 23 (2020), pp. 3877–3891.

[40] X. Wu, Y. Zhang, C. Fan, J. Hou and S. Kwong. 'Subjective Quality Database and Objective Study of Compressed Point Clouds With 6DoF Head-Mounted Display'. In: *IEEE Transactions on Circuits and Systems for Video Technology* 31.12 (2021), pp. 4630–4644. DOI: 10.1109/TCSVT.2021.3101484.

[41] Q. Liu, H. Su, Z. Duanmu, W. Liu and Z. Wang. 'Perceptual Quality Assessment of Colored 3D Point Clouds'. In: *IEEE Transactions on Visualization and Computer Graphics* (2022).

[42] Q. Liu, H. Su, Z. Duanmu, W. Liu and Z. Wang. 'Perceptual Quality Assessment of Colored 3D Point Clouds'. In: *IEEE Transactions on Visualization and Computer Graphics* (2022), pp. 1–1. DOI: 10.1109/TVCG.2022.3167151.

[43] Z. Wang and Q. Li. 'Information Content Weighting for Perceptual Image Quality Assessment'. In: *IEEE Transactions on Image Processing* 20.5 (2011), pp. 1185–1198. DOI: 10.1109/TIP.2010.2092435.

[44] Q. Liu, H. Yuan, H. Su, H. Liu, Y. Wang, H. Yang and J. Hou. 'PQA-Net: Deep no reference point cloud quality assessment via multi-view projection'. In: *IEEE transactions on circuits and systems for video technology* 31.12 (2021), pp. 4645–4660.

[45] I. Viola, S. Subramanyam and P. Cesar. 'A Color-Based Objective Quality Metric for Point Cloud Contents'. In: *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*. 2020, pp. 1–6. DOI: 10.1109/QoMEX48832.2020.9123089.

[46] E. Alexiou, X. Zhou, I. Viola and P. Cesar. 'PointPCA: Point cloud objective quality assessment using PCA-based descriptors'. In: *EURASIP Journal on Image and Video Processing* 2024.1 (2024), p. 20.

[47] R. Diniz, P. G. Freitas and M. Farias. 'A novel point cloud quality assessment metric based on perceptual color distance patterns'. In: *Electronic Imaging* 2021.9 (2021), pp. 256–1.

[48] R. Diniz, P. G. Freitas and M. C. Farias. 'Multi-Distance Point Cloud Quality Assessment'. In: *2020 IEEE International Conference on Image Processing (ICIP)*. 2020, pp. 3443–3447. DOI: 10.1109/ICIP40778.2020.9190956.

[49] R. Diniz, P. G. Freitas and M. C. Farias. 'Point cloud quality assessment based on geometry-aware texture descriptors'. In: *Computers & Graphics* (2022).

[50] Z. Zhang, W. Sun, X. Min, T. Wang, W. Lu and G. Zhai. 'No-Reference Quality Assessment for 3D Colored Point Cloud and Mesh Models'. In: *arXiv preprint arXiv:2107.02041* (2021).

[51] Q. Liu, Y. Liu, H. Su, H. Yuan and R. Hamzaoui. 'Progressive Knowledge Transfer Based on Human Visual Perception Mechanism for Perceptual Quality Assessment of Point Clouds'. In: *arXiv preprint arXiv:2211.16646* (2022).

[52] W. Zhou, Q. Yang, Q. Jiang, G. Zhai and W. Lin. *Blind Quality Assessment of 3D Dense Point Clouds with Structure-Guided Resampling*. 2022. DOI: 10.48550/arXiv.2208.14603. arXiv: 2208.14603. URL: https://arxiv.org/abs/2208.14603.

[53] Q. Yang, Y. Liu, S. Chen, Y. Xu and J. Sun. 'No-reference point cloud quality assessment via domain adaptation'. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 21179–21188.

[54] Z. Shan, Q. Yang, R. Ye, Y. Zhang, Y. Xu, X. Xu and S. Liu. 'GPA-Net: No-Reference Point Cloud Quality Assessment with Multi-task Graph Convolutional Network'. In: *IEEE Transactions on Visualization and Computer Graphics* (2023).

[55] W. Xie, K. Wang, Y. Ju and M. Wang. 'pmBQA: Projection-based Blind Point Cloud Quality Assessment via Multimodal Learning'. In: *ACM MM*. 2023, pp. 3250–3258.

[56] Z. Mei, Y.-C. Wang and C.-C. J. Kuo. 'Blind Video Quality Assessment at the Edge'. In: *IEEE Transactions on Multimedia* (2024).

[57] Z. Mei, Y.-C. Wang and C.-C. J. Kuo. 'GSBIQA: Green Saliency-guided Blind Image Quality Assessment Method'. In: *2024 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. 2024, pp. 1–6. DOI: 10.1109/APSIPAASC63619.2025.10848779.

[58] A. Ak, E. Zerman, S. Ling, P. L. Callet and A. Smolic. 'The Effect of Temporal Sub-sampling on the Accuracy of Volumetric Video Quality Assessment'. In: *2021 Picture Coding Symposium (PCS)*. 2021, pp. 1–5. DOI: 10.1109/PCS50896.2021.9477449.

[59] P. G. Freitas, M. Gonçalves, J. Homonnai, R. Diniz and M. C. Farias. 'On the Performance of Temporal Pooling Methods for Quality Assessment of Dynamic Point Clouds'. In: *2022 14th International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE. 2022, pp. 1–6.

[60] A. Ninassi, O. Le Meur, P. Le Callet and D. Barba. 'Considering temporal variations of spatial visual distortions in video quality assessment'. In: *IEEE Journal of Selected Topics in Signal Processing* 3.2 (2009), pp. 253–265.

[61] P. G. Freitas, G. D. Lucafo, M. Gonçalves, J. Homonnai, R. Diniz and M. C. Farias. 'Comparative Evaluation of Temporal Pooling Methods for No-Reference Quality Assessment of Dynamic Point Clouds'. In: *Proceedings of the 1st Workshop on Photorealistic Image and Environment Synthesis for Multimedia Experiments*. 2022, pp. 35–41.

[62] M. Yang, D. Wu, Z. Wang, M. Hu and Y. Zhou. 'Understanding and Improving Perceptual Quality of Volumetric Video Streaming'. In: *2023 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE. 2023, pp. 1979–1984.

[63] S. Van Damme, M. T. Vega and F. De Turck. 'A full-and no-reference metrics accuracy analysis for volumetric media streaming'. In: *2021 13th International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE. 2021, pp. 225–230.

[64] Y. Fan, Z. Zhang, W. Sun, X. Min, J. Lin, G. Zhai and N. Liu. 'MV-VVQA: Multi-View Learning for No-Reference Volumetric Video Quality Assessment'. In: *2023 31st European Signal Processing Conference (EUSIPCO)*. IEEE. 2023, pp. 670–674.

[65] J.-E. Marvie, Y. Nehmé, D. Graziosi and G. Lavoué. 'Crafting the MPEG metrics for objective and perceptual quality assessment of volumetric videos'. In: *Quality and User Experience* 8.1 (2023), p. 4.

[66] ITU-R BT.500-13. *Methodology for the subjective assessment of the quality of television pictures*. International Telecommunications Union. Jan. 2012.

[67] ITU-T P.1401. *Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models*. International Telecommunication Unionn. July 2012.

[68] A. Yu and K. Grauman. 'Just noticeable differences in visual attributes'. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2015, pp. 2416–2424.

[69] Y. Tian, B. Chen, S. Wang and S. Kwong. 'Towards Thousands to One Reference: Can We Trust the Reference Image for Quality Assessment?' In: *IEEE Transactions on Multimedia* (2023).

[70] K. Ma, Z. Duanmu, Q. Wu, Z. Wang, H. Yong, H. Li and L. Zhang. 'Waterloo exploration database: New challenges for image quality assessment models'. In: *IEEE Transactions on Image Processing* 26.2 (2016), pp. 1004–1016.

[71] V. Hosu, H. Lin, T. Sziranyi and D. Saupe. 'KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment'. In: *IEEE Transactions on Image Processing* 29 (2020), pp. 4041–4056.

[72] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, M. Carli and F. Battisti. 'TID2008-a database for evaluation of full-reference visual quality assessment metrics'. In: *Advances of modern radioelectronics* 10.4 (2009), pp. 30–45.

[73] X. Zhou, I. Viola, E. Alexiou, J. Jansen and P. Cesar. 'QAVA-DPC: Eye-Tracking Based Quality Assessment and Visual Attention Dataset for Dynamic Point Cloud in 6 DoF'. In: *2023 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. Los Alamitos, CA, USA: IEEE Computer Society, Oct. 2023, pp. 69–78. DOI: 10.1109/ISMAR59233.2023.00021.

[74] S. Subramanyam, I. Viola, J. Jansen, E. Alexiou, A. Hanjalic and P. Cesar. 'Subjective QoE Evaluation of User-Centered Adaptive Streaming of Dynamic Point Clouds'. In: *QoMEX*. IEEE. 2022, pp. 1–6.

[75] A. Ak, E. Zerman, M. Quach, A. Chetouani, A. Smolic, G. Valenzise and P. Le Callet. 'BASICS: Broad quality assessment of static point clouds in a compression scenario'. In: *IEEE Transactions on Multimedia* (2024).

[76] J. Prazeres, M. Pereira and A. M. Pinheiro. 'Quality Evaluation of Point Cloud Compression Techniques'. In: *Available at SSRN 4549486* (2023).

[77] R. Mekuria, K. Blom and P. Cesar. 'Design, implementation, and evaluation of a point cloud codec for tele-immersive video'. In: *IEEE TCSVT* 27.4 (2016), pp. 828–842.

[78] 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'. In: *2019 eleventh international conference on quality of multimedia experience (QoMEX)*. IEEE. 2019, pp. 1–6.

[79] 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'. In: *APSIPA Transactions on Signal and Information Processing* 8 (2019), e27.

[80] X. Wu, Y. Zhang, C. Fan, J. Hou and S. Kwong. 'Subjective Quality Database and Objective Study of Compressed Point Clouds With 6DoF Head-Mounted Display'. In: *IEEE Transactions on Circuits and Systems for Video Technology* 31.12 (2021), pp. 4630–4644. DOI: 10.1109/TCSVT.2021.3101484.

[81] Q. Liu, H. Su, Z. Duanmu, W. Liu and Z. Wang. 'Perceptual Quality Assessment of Colored 3D Point Clouds'. In: *IEEE Transactions on Visualization and Computer Graphics* (2022), pp. 1–1. DOI: 10.1109/TVCG.2022.3167151.

[82] Y. Liu, Q. Yang, Y. Xu and L. Yang. 'Point cloud quality assessment: Dataset construction and learning-based no-reference metric'. In: *ACM Transactions on Multimedia Computing, Communications and Applications* 19.2s (2023), pp. 1–26.

[83] D. Lazzarotto, M. Testolina and T. Ebrahimi. 'Subjective performance evaluation of bitrate allocation strategies for MPEG and JPEG Pleno point cloud compression'. In: *arXiv preprint arXiv:2402.04760* (2024).

[84] E. Zerman, P. Gao, C. Ozcinar and A. Smolic. 'Subjective and objective quality assessment for volumetric video compression'. In: *Electronic Imaging* 2019.10 (2019), pp. 323–1.

[85] S. Schwarz, M. Preda, V. Baroncini, M. Budagavi, P. Cesar, P. A. Chou, R. A. Cohen, M. Krivokuća, S. Lasserre, Z. Li, J. Llach, K. Mammou, R. Mekuria, O. Nakagami, E. Siahaan, A. Tabatabai, A. M. Tourapis and V. Zakharchenko. 'Emerging MPEG Standards for Point Cloud Compression'. In: *IEEE Journal on Emerging and Selected Topics in Circuits and Systems* 9.1 (2019), pp. 133–148. DOI: 10.1109/JETCAS.2018.2885981.

[86] J. van der Hooft, M. T. Vega, C. Timmerer, A. C. Begen, F. De Turck and R. Schatz. 'Objective and Subjective QoE Evaluation for Adaptive Point Cloud Streaming'. In: *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*. 2020, pp. 1–6. DOI: 10.1109/QoMEX48832.2020.9123081.

[87] K. Cao, Y. Xu and P. Cosman. 'Visual quality of compressed mesh and point cloud sequences'. In: *IEEE access* 8 (2020), pp. 171203–171217.

[88] S. R. Cox, M. Lim and W. T. Ooi. 'VOLVQAD: An MPEG V-PCC Volumetric Video Quality Assessment Dataset'. In: *Proceedings of the 14th Conference on ACM Multimedia Systems*. 2023, pp. 357–362.

[89] I. Viola, S. Subramanyam, J. Li and P. Cesar. 'On the impact of vr assessment on the quality of experience of highly realistic digital humans: A volumetric video case study'. In: *Quality and User Experience* 7.1 (2022), p. 3.

[90] S. Van Damme, I. Mahdi, H. K. Ravuri, J. van der Hooft, F. De Turck and M. T. Vega. 'Immersive and interactive subjective quality assessment of dynamic volumetric meshes'. In: *2023 15th International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE. 2023, pp. 141–146.

[91] J. Gutiérrez, G. Dandyyeva, M. Dal Magro, C. Cortés, M. Brizzi, M. Carli and F. Battisti. 'Subjective evaluation of dynamic point clouds: Impact of compression and exploration behavior'. In: *2023 31st European Signal Processing Conference (EUSIPCO)*. IEEE. 2023, pp. 675–679.

[92] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik and A. Swami. 'The limitations of deep learning in adversarial settings'. In: *2016 IEEE European symposium on security and privacy (EuroS&P)*. IEEE. 2016, pp. 372–387.

[93] K. Simonyan, A. Vedaldi and A. Zisserman. 'Visualising image classification models and saliency maps'. In: *Deep Inside Convolutional Networks* 2 (2014), p. 2.

[94] J. T. Springenberg, A. Dosovitskiy, T. Brox and M. Riedmiller. 'Striving for simplicity: The all convolutional net'. In: *arXiv preprint arXiv:1412.6806* (2014).

[95] X. Chen, A. Saparov, B. Pang and T. Funkhouser. 'Schelling points on 3D surface meshes'. In: *ACM Transactions on Graphics (TOG)* 31.4 (2012), pp. 1–12.

[96] F. P. Tasse, J. Kosinka and N. Dodgson. 'Cluster-Based Point Set Saliency'. In: *2015 IEEE International Conference on Computer Vision (ICCV)*. 2015, pp. 163–171.

[97] X. Ding, W. Lin, Z. Chen and X. Zhang. 'Point Cloud Saliency Detection by Local and Global Feature Fusion'. In: *IEEE Transactions on Image Processing* 28.11 (2019), pp. 5379–5393.

[98] M. Limper, A. Kuijper and D. W. Fellner. 'Mesh Saliency Analysis via Local Curvature Entropy.' In: *Eurographics (Short Papers)*. 2016, pp. 13–16.

[99] D. Martin, A. Fandos, B. Masia and A. Serrano. 'SAL3D: a model for saliency prediction in 3D meshes'. In: *The Visual Computer* (2024), pp. 1–11.

[100] V. Sitzmann, A. Serrano, A. Pavel, M. Agrawala, D. Gutierrez, B. Masia and G. Wetzstein. 'Saliency in VR: How Do People Explore Virtual Environments?' In: *IEEE Transactions on Visualization and Computer Graphics* 24.4 (2018), pp. 1633–1642. DOI: 10.1109/TVCG.2018.2793599.

[101] A. Nguyen and Z. Yan. 'A saliency dataset for 360-degree videos'. In: *Proceedings of the 10th ACM Multimedia Systems Conference*. 2019, pp. 279–284.

[102] G. Lavoué, F. Cordier, H. Seo and M.-C. Larabi. 'Visual Attention for Rendered 3D Shapes'. In: *Computer Graphics Forum* 37.2 (2018), pp. 191–203. DOI: https://doi.org/10.1111/cgf.13353. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.13353.

[103] X. Ding and Z. Chen. 'Towards Mesh Saliency Detection in 6 Degrees of Freedom'. In: *arXiv preprint arXiv:2005.13127* (2020).

[104] M. Abid, M. P. Da Silva and P. Le Callet. 'Towards visual saliency computation on 3D graphical contents for interactive visualization'. In: *2020 IEEE International Conference on Image Processing (ICIP)*. IEEE. 2020, pp. 3448–3452.

[105] Y. Rai, J. Gutiérrez and P. Le Callet. 'A dataset of head and eye movements for 360 degree images'. In: *Proceedings of the 8th ACM on Multimedia Systems Conference*. 2017, pp. 205–210.

[106] H. Liu and I. Heynderickx. 'Visual Attention in Objective Image Quality Assessment: Based on Eye-Tracking Data'. In: *IEEE Transactions on Circuits and Systems for Video Technology* 21.7 (2011), pp. 971–982.

[107] W. Zhang and H. Liu. 'Toward a reliable collection of eye-tracking data for image quality research: Challenges, solutions, and applications'. In: *IEEE Transactions on Image Processing* 26.5 (2017), pp. 2424–2437.

[108] Y. Jin, M. Chen, T. Goodall, A. Patney and A. C. Bovik. 'Subjective and objective quality assessment of 2D and 3D foveated video compression in virtual reality'. In: *IEEE Transactions on Image Processing* 30 (2021), pp. 5905–5919.

[109] S. Bourbia, A. Karine, A. Chetouani, M. El Hassouni and M. Jridi. 'No-reference point clouds quality assessment using transformer and visual saliency'. In: *Proceedings of the 2nd Workshop on Quality of Experience in Visual Multimedia Applications*. 2022, pp. 57–62.

[110] W. Zhou, G. Yue, R. Zhang, Y. Qin and H. Liu. 'Reduced-reference quality assessment of point clouds via content-oriented saliency projection'. In: *IEEE Signal Processing Letters* 30 (2023), pp. 354–358.

[111] Z. Wang, Y. Zhang, Q. Yang, Y. Xu, J. Sun and S. Liu. 'Point cloud quality assessment using 3D saliency maps'. In: *arXiv preprint arXiv:2209.15475* (2022).

[112] O. Le Meur, A. Ninassi, P. Le Callet and D. Barba. 'Overt visual attention for free-viewing and quality assessment tasks: Impact of the regions of interest on a video quality metric'. In: *Signal Processing: Image Communication* 25.7 (2010), pp. 547–558.

[113] H. Alers, J. Redi, H. Liu and I. Heynderickx. 'Effects of task and image properties on visual-attention deployment in image-quality assessment'. In: *Journal of Electronic Imaging* 24.2 (2015), pp. 023030–023030.

[114] J. Hadnett-Hunter, G. Nicolaou, E. O'Neill and M. Proulx. 'The effect of task on visual attention in interactive virtual environments'. In: *ACM Transactions on Applied Perception (TAP)* 16.3 (2019), pp. 1–17.

[115] Z. Hu, A. Bulling, S. Li and G. Wang. 'Ehtask: Recognizing user tasks from eye and head movements in immersive virtual reality'. In: *IEEE Transactions on Visualization and Computer Graphics* 29.4 (2021), pp. 1992–2004.

[116] S. Malpica, D. Martin, A. Serrano, D. Gutierrez and B. Masia. 'Task-Dependent Visual Behavior in Immersive Environments: A Comparative Study of Free Exploration, Memory and Visual Search'. In: *IEEE transactions on visualization and computer graphics* (2023).

[117] E. Zerman, C. Ozcinar, P. Gao and A. Smolic. 'Textured Mesh vs Coloured Point Cloud: A Subjective Study for Volumetric Video Compression'. In: *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*. 2020, pp. 1–6. DOI: 10.1109/QoMEX48832.2020.9123137.

[118] Y. Xu, Y. Lu and Z. Wen. 'Owlii Dynamic Human Textured Mesh Sequence Dataset'. In: *ISO/IEC JTC1/SC29/WG1 1 input document m41658*. 2017.

[119] 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'. In: *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE. 2020, pp. 127–136.

[120] J. Munson and T. Khoshgoftaar. 'Regression modelling of software quality: empirical investigation'. In: *Information and Software Technology* 32.2 (1990), pp. 106–114.

[121] S. Pan, Z. Liu, Y. Han, D. Zhang, X. Zhao, J. Li and K. Wang. 'Using the Pearson ' s correlation coefficient as the sole metric to measure the accuracy of quantitative trait prediction: is it sufficient?' In: *Frontiers in Plant Science* 15 (2024), p. 1480463.

[122] J. Antkowiak, T. Jamal Baina, F. V. Baroncini, N. Chateau, F. FranceTelecom, A. C. F. Pessoa, F. Stephanie Colonnese, I. L. Contin, J. Caviedes and F. Philips. 'Final report from the video quality experts group on the validation of objective models of video quality assessment'. In: (March 2000).

[123] Y. Guo, H. Wang, Q. Hu, H. Liu, L. Liu and M. Bennamoun. 'Deep learning for 3d point clouds: A survey'. In: *IEEE transactions on pattern analysis and machine intelligence* 43.12 (2020), pp. 4338–4364.

[124] I. Sipiran, A. Mendoza, A. Apaza and C. Lopez. 'Data-driven restoration of digital archaeological pottery with point cloud analysis'. In: *International Journal of Computer Vision* 130.9 (2022), pp. 2149–2165.

[125] W. Cao, J. Wu, Y. Shi and D. Chen. 'Restoration of Individual Tree Missing Point Cloud Based on Local Features of Point Cloud'. In: *Remote Sensing* 14.6 (2022), p. 1346.

[126] 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'. In: *IEEE Transactions on Image Processing* 30 (2021), pp. 6623–6636.

[127] H. Su, Q. Liu, Y. Liu, H. Yuan, H. Yang, Z. Pan and Z. Wang. 'Bitstream-Based Perceptual Quality Assessment of Compressed 3D Point Clouds'. In: *IEEE Transactions on Image Processing* 32 (2023), pp. 1815–1828.

[128] Z. Liu, Q. Li, X. Chen, C. Wu, S. Ishihara, J. Li and Y. Ji. 'Point cloud video streaming: Challenges and solutions'. In: *IEEE Network* 35.5 (2021), pp. 202–209.

[129] J. Van Der Hooft, T. Wauters, F. De Turck, C. Timmerer and H. Hellwagner. 'Towards 6dof http adaptive streaming through point cloud compression'. In: *ACM MM*. 2019, pp. 2405–2413.

[130] M. Bassier, S. Vincke, H. De Winter and M. Vergauwen. 'Drift invariant metric quality control of construction sites using BIM and point cloud data'. In: *ISPRS International Journal of Geo-Information* 9.9 (2020), p. 545.

[131] P. Ye, J. Kumar, L. Kang and D. Doermann. 'Unsupervised feature learning framework for no-reference image quality assessment'. In: *CVPR*. 2012, pp. 1098–1105. DOI: 10.1109/CVPR.2012.6247789.

[132] L. Zhang, L. Zhang and A. C. Bovik. 'A Feature-Enriched Completely Blind Image Quality Evaluator'. In: *IEEE Transactions on Image Processing* 24.8 (2015), pp. 2579–2591. DOI: 10.1109/TIP.2015.2426416.

[133] J. Xu, P. Ye, Q. Li, H. Du, Y. Liu and D. Doermann. 'Blind Image Quality Assessment Based on High Order Statistics Aggregation'. In: *IEEE Transactions on Image Processing* 25.9 (2016), pp. 4444–4457. DOI: 10.1109/TIP.2016.2585880.

[134] W. Xue, L. Zhang and X. Mou. 'Learning without Human Scores for Blind Image Quality Assessment'. In: *CVPR*. June 2013.

[135] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, M. Manohara *et al.* 'Toward a practical perceptual video quality metric'. In: *The Netflix Tech Blog* 6.2 (2016), p. 2.

[136] X. Zhou, Y. Zhang, N. Li, X. Wang, Y. Zhou and Y.-S. Ho. 'Projection invariant feature and visual saliency-based stereoscopic omnidirectional image quality assessment'. In: *IEEE Transactions on Broadcasting* 67.2 (2021), pp. 512–523.

[137] Y. Zhang, S. Kwong and S. Wang. 'Machine learning based video coding optimizations: A survey'. In: *Information Sciences* 506 (2020), pp. 395–423.

[138] Y. Zhang, X. Gao, L. He, W. Lu and R. He. 'Objective video quality assessment combining transfer learning with CNN'. In: *IEEE transactions on neural networks and learning systems* 31.8 (2019), pp. 2716–2730.

[139] J. Kim, A.-D. Nguyen and S. Lee. 'Deep CNN-based blind image quality predictor'. In: *IEEE transactions on neural networks and learning systems* 30.1 (2018), pp. 11–24.

[140] X. Wang, J. Xiong, H. Gao and W. Lin. 'Regression-free Blind Image Quality Assessment'. In: *arXiv preprint arXiv:2307.09279* (2023).

[141] E. Alexiou, E. Upenik and T. Ebrahimi. 'Towards subjective quality assessment of point cloud imaging in augmented reality'. In: *2017 IEEE 19th International Workshop on Multimedia Signal Processing (MMSP)*. 2017, pp. 1–6. DOI: 10.1109/MMSP.2017.8122237.

[142] A. Javaheri, C. Brites, F. Pereira and J. Ascenso. 'Point cloud rendering after coding: Impacts on subjective and objective quality'. In: *IEEE Transactions on Multimedia* 23 (2020), pp. 4049–4064.

[143] D. G. A. Zaghetto and A. Tabatabai. 'Density-to-density (d3- psnr'. In: *ISO/IEC JTC1/SC29 WG7 input document M61195* (2022).

[144] D. Tian, H. Ochimizu, C. Feng, R. Cohen and A. Vetro. *Updates and Integration of Evaluation Metric Software for PCC*. Tech. rep. MPEG2017/M40522. Hobart, Australia: ISO/IEC JTC 1/SC 29/WG 11 (MPEG), Apr. 2017.

[145] E. Alexiou and T. Ebrahimi. 'Towards a Point Cloud Structural Similarity Metric'. In: *2020 IEEE International Conference on Multimedia Expo Workshops (ICMEW)*. 2020, pp. 1–6. DOI: 10.1109/ICMEW46912.2020.9106005.

[146] X. Zhou, E. Alexiou, I. Viola and P. Cesar. 'PointPCA+: Extending PointPCA objective quality assessment metric'. In: *2023 IEEE International Conference on Image Processing Challenges and Workshops (ICIPCW)*. IEEE. 2023, pp. 1–5.

[147] Y. Liu, Z. Shan, Y. Zhang and Y. Xu. 'MFT-PCQA: Multi-Modal Fusion Transformer for No-Reference Point Cloud Quality Assessment'. In: *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2024, pp. 7965–7969. DOI: 10.1109/ICASSP48485.2024.10445736.

[148] ITU-R BT.709-6. *Parameter values for the HDTV standards for production and international programme exchange*. International Telecommunication Unionn. June 2015.

[149] I. Guyon, J. Weston, S. Barnhill and V. Vapnik. 'Gene selection for cancer classification using support vector machines'. In: *Machine learning* 46 (2002), pp. 389–422.

[150] R. Diniz, P. G. Freitas and M. C. Q. Farias. 'Color and Geometry Texture Descriptors for Point-Cloud Quality Assessment'. In: *IEEE Signal Processing Letters* 28 (2021), pp. 1150–1154. DOI: 10.1109/LSP.2021.3088059.

[151] 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'. In: *2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)*. 2016, pp. 1–6. DOI: 10.1109/QoMEX.2016.7498936.

[152] A. Chetouani, G. Valenzise, A. Ak, E. Zerman, M. Quach, M. Tliba, M. A. Kerkouri and P. L. Callet. 'ICIP 2023 - Point cloud visual quality assessment grand challenge'. In: (2023).

[153] L. Breiman and R. Cutler. 'Random forests machine learning [J]'. In: *journal of clinical microbiology* 2 (2001), pp. 199–228.

[154] S. Wang, X. Wang, H. Gao and J. Xiong. 'Non-Local Geometry and Color Gradient Aggregation Graph Model for No-Reference Point Cloud Quality Assessment'. In: *ACM MM*. 2023, pp. 6803–6810.

[155] C. R. Qi, L. Yi, H. Su and L. J. Guibas. 'Pointnet++: Deep hierarchical feature learning on point sets in a metric space'. In: *NeurIPS*. Vol. 30. 2017.

[156] K. He, X. Zhang, S. Ren and J. Sun. 'Deep residual learning for image recognition'. In: *CVPR*. 2016, pp. 770–778.

[157] Y. Cao, Y. Ma, M. Zhou, C. Liu, H. Xie, T. Ge and Y. Jiang. 'Geometry aligned variational transformer for image-conditioned layout generation'. In: *ACM MM*. 2022, pp. 1561–1571.

[158] J. Ma, S. Yan, L. Zhang, G. Wang and Q. Zhang. 'ELMformer: Efficient Raw Image Restoration with a Locally Multiplicative Transformer'. In: *ACM MM*. 2022, pp. 5842–5852.

[159] B. Zhang, J. Yuan, B. Li, T. Chen, J. Fan and B. Shi. 'Learning cross-image object semantic relation in transformer for few-shot fine-grained image classification'. In: *ACM MM*. 2022, pp. 2135–2144.

[160] C. Zhang, H. Wan, X. Shen and Z. Wu. 'Patchformer: An efficient point transformer with patch attention'. In: *CVPR*. 2022, pp. 11799–11808.

[161] M.-H. Guo, J.-X. Cai, Z.-N. Liu, T.-J. Mu, R. R. Martin and S.-M. Hu. 'Pct: Point cloud transformer'. In: *Computational Visual Media* 7 (2021), pp. 187–199.

[162] Q. Yang, Y. Zhang, S. Chen, Y. Xu, J. Sun and Z. Ma. 'MPED: Quantifying point cloud distortion based on multiscale potential energy discrepancy'. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.5 (2022), pp. 6037–6054.

[163] I. Viola and P. Cesar. 'A Reduced Reference Metric for Visual Quality Evaluation of Point Cloud Contents'. In: *IEEE Signal Processing Letters* 27 (2020), pp. 1660–1664. DOI: 10.1109/LSP.2020.3024065.

[164] Z. Zhang, W. Sun, X. Min, T. Wang, W. Lu and G. Zhai. 'No-Reference Quality Assessment for 3D Colored Point Cloud and Mesh Models'. In: *IEEE Transactions on Circuits and Systems for Video Technology* 32.11 (2022), pp. 7618–7631. DOI: 10.1109/TCSVT.2022.3186894.

[165] Y. Fan, Z. Zhang, W. Sun, X. Min, N. Liu, Q. Zhou, J. He, Q. Wang and G. Zhai. 'A no-reference quality assessment metric for point cloud based on captured video sequences'. In: *2022 IEEE 24th International Workshop on Multimedia Signal Processing (MMSP)*. IEEE. 2022, pp. 1–5.

[166] Z. Zhang, W. Sun, H. Wu, Y. Zhou, C. Li, X. Min, G. Zhai and W. Lin. 'GMS-3DQA: Projection-based Grid Mini-patch Sampling for 3D Model Quality Assessment'. In: *arXiv preprint arXiv:2306.05658* (2023).

[167] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai and S. Chintala. 'PyTorch: An Imperative Style, High-Performance Deep Learning Library'. In: *NeurIPS*. Curran Associates, Inc., 2019, pp. 8024–8035. URL: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[168] I. Loshchilov and F. Hutter. *Decoupled Weight Decay Regularization*. arXiv preprint arXiv:1711.05101. 2017. DOI: 10.48550/arXiv.1711.05101. URL: https://arxiv.org/abs/1711.05101.

[169] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei. 'Imagenet: A large-scale hierarchical image database'. In: *CVPR*. Ieee. 2009, pp. 248–255.

[170] Z. Shan, Y. Zhang, Q. Yang, H. Yang, Y. Xu, J.-N. Hwang, X. Xu and S. Liu. 'Contrastive Pre-Training with Multi-View Fusion for No-Reference Point Cloud Quality Assessment'. In: *arXiv preprint arXiv:2403.10066* (2024).

[171] P. Yang, C. G. Snoek and Y. M. Asano. 'Self-Ordering Point Clouds'. In: *ICCV*. 2023, pp. 15813–15822.

[172] E. Alexiou, Y. Nehmé, E. Zerman, I. Viola, G. Lavoué, A. Ak, A. Smolic, P. Le Callet and P. Cesar. 'Subjective and objective quality assessment for volumetric video'. In: *Immersive Video Technologies*. Elsevier, 2023, pp. 501–552.

[173] Y. Zhang, L. Wang and Y. Dai. 'PLOT: a 3D point cloud object detection network for autonomous driving'. In: *Robotica* (2023), pp. 1–17.

[174] Q. Cheng, P. Sun, C. Yang, Y. Yang and P. X. Liu. 'A morphing-Based 3D point cloud reconstruction framework for medical image processing'. In: *Computer methods and programs in biomedicine* 193 (2020), p. 105495.

[175] S. F. Langa, M. Montagud, G. Cernigliaro and D. R. Rivera. 'Multiparty Holomeetings: Toward a New Era of Low-Cost Volumetric Holographic Meetings in Virtual Reality'. In: *Ieee Access* 10 (2022), pp. 81856–81876.

[176] I. Viola, J. Jansen, S. Subramanyam, I. Reimat and P. Cesar. 'VR2Gather: A collaborative social VR system for adaptive multi-party real-time communication'. In: *IEEE MultiMedia* (2023).

[177] H. Hadizadeh and I. V. Bajić. 'Saliency-aware video compression'. In: *IEEE Transactions on Image Processing* 23.1 (2013), pp. 19–33.

[178] H. Li, G. Chen, G. Li and Y. Yu. 'Motion guided attention for video salient object detection'. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 7274–7283.

[179] A. Borji and L. Itti. 'State-of-the-art in visual attention modeling'. In: *IEEE transactions on pattern analysis and machine intelligence* 35.1 (2012), pp. 185–207.

[180] T. Betz, T. C. Kietzmann, N. Wilming and P. Koenig. 'Investigating task-dependent top-down effects on overt visual attention'. In: *Journal of vision* 10.3 (2010), pp. 15–15.

[181] S. Kollmorgen, N. Nortmann, S. Schröder and P. König. 'Influence of low-level stimulus features, task dependent factors, and spatial biases on overt visual attention'. In: *PLoS computational biology* 6.5 (2010), e1000791.

[182] P. Polatsek, M. Waldner, I. Viola, P. Kapec and W. Benesova. 'Exploring visual attention and saliency modeling for task-based visual analysis'. In: *Computers & Graphics* 72 (2018), pp. 26–38.

[183] S. Rahman and N. Bruce. 'Visual saliency prediction and evaluation across different perceptual tasks'. In: *PloS one* 10.9 (2015), e0138053.

[184] C. Wloka and J. K. Tsotsos. 'Overt fixations reflect a natural central bias'. In: *Journal of Vision* 13.9 (2013), pp. 239–239.

[185] A. K. Sinha and K. Shukla. 'A study of distance metrics in histogram based image retrieval'. In: *Int. J. Comput. Technol* 4.3 (2013), pp. 821–830.

[186] Z. Hu. '[DC] Eye Fixation Forecasting in Task-Oriented Virtual Reality'. In: *2021 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. 2021, pp. 707–708.

[187] Z. Hu. 'Gaze Analysis and Prediction in Virtual Reality'. In: *2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. 2020, pp. 543–544.

[188] E. Psatha, D. Laskos, G. Arvanitis and K. Moustakas. 'Aggressive saliency-aware point cloud compression'. In: *arXiv preprint arXiv:2307.10741* (2023).

[189] Y. Zhang, K. Ding, N. Li, H. Wang, X. Huang and C.-C. J. Kuo. 'Perceptually weighted rate distortion optimization for video-based point cloud compression'. In: *IEEE Transactions on Image Processing* 32 (2023), pp. 5933–5947.

[190] P. Ruiu, L. Mascia and E. Grosso. 'Saliency-Guided Point Cloud Compression for 3D Live Reconstruction'. In: *Multimodal Technologies and Interaction* 8.5 (2024), p. 36.

[191] Y. Kim and A. Varshney. 'Saliency-guided enhancement for volume visualization'. In: *IEEE Transactions on Visualization and Computer Graphics* 12.5 (2006), pp. 925–932.

[192] R. Singh, M. Huzaifa, J. Liu, A. Patney, H. Sharif, Y. Zhao and S. Adve. 'Power, Performance, and Image Quality Tradeoffs in Foveated Rendering'. In: *2023 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*. 2023, pp. 205–214.

[193] Z. Wang, Y. Zhang, Q. Yang, Y. Xu, J. Sun and S. Liu. 'Point cloud quality assessment using 3D saliency maps'. In: *2023 IEEE International Conference on Visual Communications and Image Processing (VCIP)*. IEEE. 2023, pp. 1–5.

[194] E. d' Eon, B. Harrison, T. Myers and P. A. Chou. *JPEG Pleno Database: 8i Voxelized Full Bodies (8iVFB v2)–A Dynamic Voxelized Point Cloud Dataset*. 2019.

[195] 'ITU-T Rec. P.910 (04/2008) Subjective video quality assessment methods for multimedia applications'. In: 2009. URL: https://api.semanticscholar.org/CorpusID:30891831.

[196] S.Schwarz. *Common Test Conditions for PCC*. ISO/IEC JTC1/SC29/WG11 Doc. N18665. Gothenburg, Sweden, July 2019.

[197] I. Reimat, E. Alexiou, J. Jansen, I. Viola, S. Subramanyam and P. Cesar. 'CWIPC-SXR: Point Cloud dynamic human dataset for Social XR'. In: *Proceedings of the 12th ACM Multimedia Systems Conference*. 2021, pp. 300–306.

[198] J. Gutierrez, P. Perez, M. Orduna, A. Singla, C. Cortes, P. Mazumdar, I. Viola, K. Brunnström, F. Battisti, N. Cieplińska *et al.* 'Subjective Evaluation of Visual Quality and Simulator Sickness of Short 360° Videos: ITU-T Rec. P. 919'. In: *IEEE transactions on multimedia* 24 (2021), pp. 3087–3100.

[199] International Telecommunication Union. *Recommendation ITU-T P.910: Subjective video quality assessment methods for multimedia applications*. Recommendation ITU-T. Edition 6.0, approved October 29, 2023. 2023. URL: https://handle.itu.int/11.1002/1000/15697.

[200] F. L. Ferris III, A. Kassoff, G. H. Bresnick and I. Bailey. 'New visual acuity charts for clinical research'. In: *American journal of ophthalmology* 94.1 (1982), pp. 91–96.

[201] J. Clark. 'The Ishihara test for color blindness.' In: *American Journal of Physiological Optics* (1924).

[202] W. Zhang and H. Liu. 'Study of Saliency in Objective Video Quality Assessment'. In: *IEEE Transactions on Image Processing* 26.3 (2017), pp. 1275–1288. DOI: 10.1109/TIP.2017.2651410.

[203] A. van Kasteren, K. Brunnström, J. Hedlund and C. Snijders. 'Quality of experience of 360 video–subjective and eye-tracking assessment of encoding and freezing distortions'. In: *Multimedia tools and applications* 81.7 (2022), pp. 9771–9802.

[204] E. J. David, J. Gutiérrez, A. Coutrot, M. P. Da Silva and P. L. Callet. 'A dataset of head and eye movements for 360 videos'. In: *Proceedings of the 9th ACM Multimedia Systems Conference*. 2018, pp. 432–437.

[205] C. Ozcinar and A. Smolic. 'Visual attention in omnidirectional video for virtual reality applications'. In: *2018 Tenth international conference on quality of multimedia experience (QoMEX)*. IEEE. 2018, pp. 1–6.

[206] R. S. Kennedy, N. E. Lane, K. S. Berbaum and M. G. Lilienthal. 'Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness'. In: *The international journal of aviation psychology* 3.3 (1993), pp. 203–220.

[207] L. Sidenmark, M. N. Lystbæk and H. Gellersen. 'GE-Simulator: An Open-Source Tool for Simulating Real-Time Errors for HMD-Based Eye Trackers'. In: *Proceedings of the 2023 Symposium on Eye Tracking Research and Applications*. ETRA '23. Tubingen, Germany: Association for Computing Machinery, 2023. URL: https://doi.org/10.1145/3588015.3588417.

[208] I. B. Adhanom, S. C. Lee, E. Folmer and P. MacNeilage. 'Gazemetrics: An open-source tool for measuring the data quality of HMD-based eye trackers'. In: *ACM symposium on eye tracking research and applications*. 2020, pp. 1–5.

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

[210] H. Widdel. 'Operational problems in analysing eye movements'. In: *Advances in psychology*. Vol. 22. Elsevier, 1984, pp. 21–29.

[211] M. Ester, H.-P. Kriegel, J. Sander, X. Xu *et al.* 'A density-based algorithm for discovering clusters in large spatial databases with noise.' In: *kdd*. Vol. 96. 34. 1996, pp. 226–231.

[212] E. Schubert, J. Sander, M. Ester, H. P. Kriegel and X. Xu. 'DBSCAN revisited, revisited: why and how you should (still) use DBSCAN'. In: *ACM Transactions on Database Systems (TODS)* 42.3 (2017), pp. 1–21.

[213] M. Maguire and B. Delahunt. 'Doing a thematic analysis: A practical, step-by-step guide for learning and teaching scholars.' In: *All Ireland journal of higher education* 9.3 (2017).

[214] G. Gautier, A. Mercat, L. Fréneau, M. Pitkänen and J. Vanne. 'UVG-VPC: Voxelized Point Cloud Dataset for Visual Volumetric Video-based Coding'. In: *2023 15th International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE. 2023, pp. 244–247.

[215] Y. Xu, Y. Lu and Z. Wen. *Owlii Dynamic Human Mesh Sequence Dataset*. Tech. rep. M41658. Macau: ISO/IEC JTC 1/SC 29/WG 11 (MPEG), Oct. 2017.

[216] S. Rossi, I. Viola and P. Cesar. 'Behavioural analysis in a 6-DoF VR system: Influence of content, quality and user disposition'. In: *Proceedings of the 1st Workshop on Interactive eXtended Reality*. 2022, pp. 3–10.

[217] H. Alers, L. Bos and I. Heynderickx. 'How the task of evaluating image quality influences viewing behavior'. In: *2011 Third International Workshop on Quality of Multimedia Experience*. IEEE. 2011, pp. 167–172.

[218] I. Stein, H. Jossberger and H. Gruber. 'MAP3D: An explorative approach for automatic mapping of real-world eye-tracking data on a virtual 3D model'. In: *Journal of Eye Movement Research* 15.3 (2022).

[219] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba and F. Durand. 'What Do Different Evaluation Metrics Tell Us About Saliency Models?' In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41.3 (2019), pp. 740–757.

[220] N. Riche, M. Duvinage, M. Mancas, B. Gosselin and T. Dutoit. 'Saliency and human fixations: State-of-the-art and study of comparison metrics'. In: *Proceedings of the IEEE international conference on computer vision*. 2013, pp. 1153–1160.

[221] O. Le Meur, P. Le Callet and D. Barba. 'Predicting visual fixations on video based on low-level visual features'. In: *Vision research* 47.19 (2007), pp. 2483–2498.

[222] Y. Rubner, C. Tomasi and L. J. Guibas. 'The earth mover's distance as a metric for image retrieval'. In: *International journal of computer vision* 40 (2000), pp. 99–121.

[223] N. Wu, K. Liu, R. Cheng, B. Han and P. Zhou. 'Theia: Gaze-driven and Perception-aware Volumetric Content Delivery for Mixed Reality Headsets'. In: *Proceedings of the 22nd Annual International Conference on Mobile Systems, Applications and Services*. 2024, pp. 70–84.

[224] H. Kim, S. Lee and A. C. Bovik. 'Saliency Prediction on Stereoscopic Videos'. In: *IEEE Transactions on Image Processing* 23.4 (2014), pp. 1476–1490. DOI: 10.1109/TIP.2014.2303640.

[225] S. Cheng, J. Fan and Y. Hu. 'Visual saliency model based on crowdsourcing eye tracking data and its application in visual design'. In: *Personal and Ubiquitous Computing* 27.3 (2023), pp. 613–630.

[226] Y.-C. Sun, I.-C. Huang, Y. Shi, W. T. Ooi, C.-Y. Huang and C.-H. Hsu. 'A Dynamic 3D Point Cloud Dataset for Immersive Applications'. In: *Proceedings of the 14th Conference on ACM Multimedia Systems*. 2023, pp. 376–383.

[227] L. Xie, X. Mu, G. Li, W. Gao *et al.* 'PKU-DPCC: A New Dataset for Dynamic Point Cloud Compression'. In: *APSIPA Transactions on Signal and Information Processing* 13.6 (2024).

[228] U. Engelke, H. Liu, H.-J. Zepernick, I. Heynderickx and A. Maeder. 'Comparing two eye-tracking databases: The effect of experimental setup and image presentation time on the creation of saliency maps'. In: *28th Picture Coding Symposium*. 2010, pp. 282–285.

[229] O. Sidorov, J. S. Harvey, H. E. Smithson and J. Y. Hardeberg. 'Overt visual attention on rendered 3D objects'. In: *arXiv preprint arXiv:1905.10444* (2019).

[230] S. Kollmorgen, N. Nortmann, S. Schröder and P. König. 'Influence of Low-Level Stimulus Features, Task Dependent Factors, and Spatial Biases on Overt Visual Attention'. In: *PLOS Computational Biology* 6.5 (May 2010), pp. 1–20.

[231] A. Schmitz, A. MacQuarrie, S. Julier, N. Binetti and A. Steed. 'Directing versus Attracting Attention: Exploring the Effectiveness of Central and Peripheral Cues in Panoramic Videos'. In: *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. 2020, pp. 63–72.

[232] L. Paletta, K. Santner, G. Fritz, H. Mayer and J. Schrammel. '3D attention: measurement of visual saliency using eye tracking glasses'. In: *CHI'13 Extended Abstracts on Human Factors in Computing Systems*. 2013, pp. 199–204.

[233] M. Adamove, D. Padúch and P. Kapec. 'Evaluation of Visual Saliency Models in Immersive Analytics'. In: *Advances in Information and Communication*. Ed. by K. Arai. Cham: Springer Nature Switzerland, 2024, pp. 375–392. ISBN: 978-3-031-53963-3.

[234] M. S. Arefin, N. Phillips, A. Plopski, J. L. Gabbard and J. E. Swan. 'Impact of AR Display Context Switching and Focal Distance Switching on Human Performance: Replication on an AR Haploscope'. In: *2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. 2020, pp. 571–572.

[235] L. Itti, C. Koch and E. Niebur. 'A model of saliency-based visual attention for rapid scene analysis'. In: *IEEE Transactions on pattern analysis and machine intelligence* 20.11 (1998), pp. 1254–1259.

[236] Z. Lu, W. Lin, X. Yang, E. Ong and S. Yao. 'Modeling visual attention's modulatory aftereffects on visual sensitivity and quality evaluation'. In: *IEEE transactions on Image Processing* 14.11 (2005), pp. 1928–1942.

[237] X. Wang, A. Katsenou and D. Bull. 'UGC quality assessment: exploring the impact of saliency in deep feature-based quality assessment'. In: *Applications of Digital Image Processing XLVI*. Vol. 12674. SPIE. 2023, pp. 351–365.

[238] M. Tliba, X. Zhou, I. Viola, P. Cesar, A. Chetouani, G. Valenzise and F. Dufaux. 'Enhancing Immersive Experiences through 3D Point Cloud Analysis: A Novel Framework for Applying 2D Visual Saliency Models to 3D Point Clouds'. In: *2024 16th International Conference on Quality of Multimedia Experience (QoMEX)*. 2024, pp. 307–313.

[239] Q. Yang, Y. Liu, S. Chen, Y. Xu and J. Sun. 'No-Reference Point Cloud Quality Assessment via Domain Adaptation'. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022.

[240] E. Alexiou, P. Xu and T. Ebrahimi. 'Towards Modelling of Visual Saliency in Point Clouds for Immersive Applications'. In: *2019 IEEE International Conference on Image Processing (ICIP)*. 2019, pp. 4325–4329. DOI: 10.1109/ICIP.2019.8803479.

[241] A. Borji and L. Itti. 'State-of-the-Art in Visual Attention Modeling'. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.1 (2013), pp. 185–207. DOI: 10.1109/TPAMI.2012.89.

[242] J. Lou, H. Lin, D. Marshall, D. Saupe and H. Liu. 'TranSalNet: Towards perceptually relevant visual saliency prediction'. In: *Neurocomputing* 494 (2022), pp. 455–467.

[243] G. Lee, Y.-W. Tai and J. Kim. 'Deep saliency with encoded low level distance map and high level features'. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 660–668.

[244] S.-A. Rebuffi, R. Fong, X. Ji and A. Vedaldi. 'There and back again: Revisiting backpropagation saliency methods'. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 8839–8848.

[245] J. D. M.-W. C. Kenton and L. K. Toutanova. 'BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding'. In: *Proceedings of NAACL-HLT*. 2019, pp. 4171–4186.

[246] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.* 'An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale'. In: *International Conference on Learning Representations*. 2020.

[247] W. Sun, X. Min, W. Lu and G. Zhai. 'A deep learning based no-reference quality assessment model for ugc videos'. In: *ACM MM*. 2022, pp. 856–865.

[248] R. Mekuria, Z. Li, C. Tulvan and P. Chou. 'Evaluation criteria for point cloud compression'. In: *ISO/IEC MPEG* 16332 (2016).

[249] Video Quality Experts Group (VQEG). *Final Report from the Video Quality Experts Group on the Validation of Objective Models of Video Quality Assessment*. https://www.its.bldrdoc.gov/vqeg/vqeg-home.aspx. Accessed: 2026-02-04. 2003.

[250] Q.-Y. Zhou, J. Park and V. Koltun. 'Open3D: A modern library for 3D data processing'. In: *arXiv preprint arXiv:1801.09847* (2018).

[251] D. P. Kingma and J. Ba. 'Adam: A method for stochastic optimization'. In: *arXiv preprint arXiv:1412.6980* (2014).

[252] Y. Zhang, S. Song, E. Yumer, M. Savva, J.-Y. Lee, H. Jin and T. Funkhouser. 'Physically-based rendering for indoor scene understanding using convolutional neural networks'. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 5287–5295.

[253] M. Tliba, X. Zhou, I. Viola, P. Cesar, A. Chetouani, G. Valenzise and F. Dufaux. 'Enhancing Immersive Experiences through 3D Point Cloud Analysis: A Novel Framework for Applying 2D Visual Saliency Models to 3D Point Clouds'. In: *International Workshop on Quality of Multimedia Experience (QoMEX' 2024)*. 2024.

[254] G. Gautier, X. Zhou, T. Nguyen, J. Jansen, L. Fréneau, M. Viitanen, U. Phan, J. Käpylä, I. Viola, A. Mercat *et al.* 'UVG-CWI-DQPC: Dual-Quality Point Cloud Dataset for Volumetric Video Applications'. In: *Proceedings of the 33rd ACM International Conference on Multimedia*. 2025, pp. 13112–13118.

[255] MPEG 3DG and Requirements. *Call for proposals for point cloud compression v2*. ISO/IEC JTC1/SC29/WG11 Doc. N16763. Hobart, Australia, Apr. 2017.

[256] ITU-T J.149. *Method for specifying accuracy and cross-calibration of Video Quality Metrics (VQM)*. International Telecommunication Union. Mar. 2004.

[257] T. AB. *HTC Vive Pro Eye integration with Tobii eye tracking*. https://www.tobii.com/products/integration/xr-headsets/device-integrations/htc-vive-pro-eye. Accessed: 2025-05-26. 2024.

[258] X. Ding and Z. Chen. 'Towards mesh saliency in 6 degrees of freedom'. In: *Neurocomputing* 502 (2022), pp. 120–139.

Acknowledgements

I do not want to graduate, at least not at this moment. This PhD Odyssey has been unexpectedly comfortable, and I owe that comfort to the many who supported me along the way.

I would like to express my deepest gratitude to my promotor Pablo Cesar and my copromotor Irene Viola for their guidance, patience, and continuous support throughout my PhD. I thank Pablo for his close supervision and the trust he placed in me, offering both strong guidance early on and increasing freedom as I grew into an independent researcher. His mentorship extended beyond research, shaping my communication skills, research taste, career path, and more. I thank Irene for her daily supervision, rigorous guidance and patience with my many questions and evolving ideas. As a female researcher, I learned from her ambition, integrity, and openness, and our discussions on work–life balance and standing up in academia have been invaluable to me.

I am grateful to my thesis committee: Prof. Alan Hanjalic, Prof. Maria G. Martini, Prof. Patrick Le Callet, Dr. Maarten Wijntjes, Dr. Zerrin Yumak, and Prof. Alessandro Bozzon (reserve member), for their valuable feedback and suggestions that improved this manuscript.

I want to thank the senior researchers who generously shared their knowledge and expertise. Vaggelis introduced me to the world of point clouds; Elmar brought me to CGV, which became my second office at Delft. I am grateful to Elmar and Michael for their guidance on my survey paper, and to Guillaume, Alexandre, and Jarno for the internship in Tampere. I felt welcomed and supported at CGV & UVG and greatly enjoyed the dynamics in both.

My sincere thanks go to all my co-authors and collaborators inside and outside the lab! Minh, Shivi, Christian, Hermann, Louis, Alexandre, Guillaume, Marouane, Giuseppe, and Aladine...the full list is on my Google Scholar! The many insightful discussions and brainstorming challenged and motivated me. I also thank the supporting staff at CWI and TU Delft for the countless things they quietly take care of. How it works, I know not; that it works, I know well. Because of you, my life here was much easier.

I thank my DIS colleagues: Jack and Tom, whose ability to magically fix bugs gave me the courage to be the troublemaker. Steven, for checking my English in the thesis. My fellow PhDs: Shishir, for answering my endless questions and Tianyi, who picked me up from the airport and began my Amsterdam chapter. Thanks to Abdo, Alina, Moonisa, Pooja, Ashu, Karthikeya, Sueyoon, Silvia, Nacho... as well as visitors and office mates at M374 and other colleagues in CWI. To my colleagues at CGV in Delft–Mark, Benno, Lukas, Guowei and others... Even though I did not attend the group meetings regularly, I somehow never missed the group outings.

I thank the Chinese community at Lab 42 (UvA) and CWI for warmth and companionship: Yunlu, Pengwan, Ruihong, Tao, Yixian, Yijia, Xinyuan, Tianyuan, Gao, Yibin, Di,

and many more! I thank the “Chinese chefs” (Aozhu, Xiaoran, Longjiao, Tom, Youyou, and more!) who feed me well from time to time. I thank my master’s group, you fed me well in Shenzhen each time, listened as I cried and cursed the hard first year of my PhD —first soothing my emotions, and only later talking me through right and wrong. I thank Yiru, Gaole, and Ruomei for all kinds of help over the years, we have known each other long before coming to the Netherlands. I am grateful to Amir, Sueyoon, and Guichen for their friendship: we update regularly and talk about careers, politics, dreams, hobbies, gossip, impact, beauty and love. Special thanks to Yingkui Zhang for timely help with Unity, which kept my PhD on track.

爸爸妈妈，我爱你们！To my parents and elder sister: thank you for being there, no matter what. I hope I may still be your daughter/younger sister next life if there is one. To my brother-in-law and little nephews, you made me healthy (I now believe more in Chinese medicine!) and joyful. Xuemei, thank you for being real and free. And to all who crossed my path, you softened my days abroad and left quiet light behind. I thank you all!

Written on 25 Jan 2026, Amsterdam