APSIPA Transactions on Signal and Information Processing, 2022, 11, e28

Overview Paper

Visual Saliency and Quality Evaluation for 3D Point Clouds and Meshes: An Overview

Weisi Lin[1*] and Sanghoon Lee[2]

[1] *School of Computer Science and Engineering, Nanyang Technological University, Singapore*
[2] *Electrical and Electronic Engineering, Yonsei University, Seoul, Korea*

ABSTRACT

Three-dimensional (3D) point clouds (PCs) and meshes have increasingly become available and indispensable for diversified applications in work and life. In addition, 3D visual data contain information from any viewpoint when needed, introducing new challenges and opportunities. As in the cases of 2D images and videos, computationally modeling saliency and quality for 3D PCs and meshes are important for widespread, economical adaption and optimization. This paper aims to provide a comprehensive overview of the related signal presentation and existing saliency and quality models, with major perspectives from the ultimate users (i.e., humans or machines), modeling methodology (with handcrafted features or machine learning), and modeling scope (generic or utility-oriented models). Possible future research directions are also discussed.

Keywords: Point clouds, meshes, 3D visual data, saliency, quality, human uses, machine uses, keypoints, handcrafted features, learning-based modeling, utility-oriented evaluation, quality of experience (QoE), metaverse.

*Corresponding author: Weisi Lin, wslin@ntu.edu.sg.

Received 29 April 2022; Revised 26 June 2022
ISSN 2048-7703; DOI 10.1561/116.00000125

1 Introduction

There has been rapid technological development in photogrammetry, three-dimensional (3D) scanning, high-performance computing, transmission and internet, and machine learning. As a result, more 3D point clouds (PCs) and meshes [72, 204] have become available and are increasingly indispensable for representing real-world objects (as digital twins). The PCs and meshes have diversified scales (from a single object to a whole city, as illustrated in Figure 1) for different meaningful applications and even otherwise impossible missions. The relevant application scenarios include virtual objects and settings are expected to extend to the emerging metaverse [27, 117].

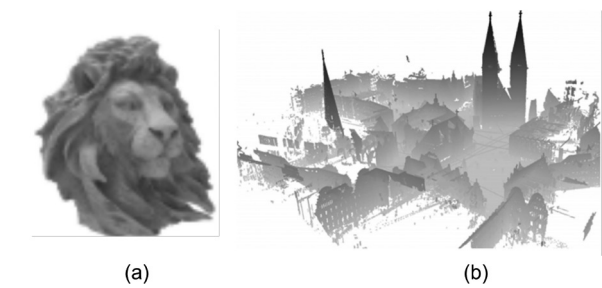

Figure 1: A PC can be with diversity for scale: (a) a single object and (b) a whole city.

A 3D PC is a set of points in 3D space with 3D geometry (with x, y, and z coordinates) and attribute information, such as color, texture, and derived/auxiliary information (e.g., surface normal and other more sophisticated feature descriptors). Ideally, infinite points and an unlimited depth range are required to express an object and scene exactly.

A PC can be acquired by a depth sensor or massive red, green and blue (RGB) images. Depth sensors can be divided into active stereo and time of flight (ToF) sensors [75]. The active stereo sensor [99] emits a pattern texture onto the scene surface and measures the depth based on the captured pattern differences in stereo images. The ToF sensor [75], including the light detection and ranging (LiDAR) sensor, measures the depth according to the time that passes from signal emission receipt. In addition, PCs can be obtained from massive RGB images [150, 220] by finding point correspondences between images; thus they can be considered as a substantial extension of traditional RGB images and video. The emergence of PCs has provided a bridge between computer graphics (CG) and computer vision (CV), which have long been two largely separate domains.

In practice with constrained resources, a PC has finite points and a bounded depth range; therefore, for an effective and efficient PC, points should be prop-

erly distributed [8, 16, 140, 142]. The depth information of a 3D PC for bounded scenes and objects can be easily obtained using laser scanners by maximizing the resolution of the depth information at a short distance by comparing the phase of the reflected wave using short-wavelength rays. In addition, it is also possible to acquire dynamic 3D PCs in real-time using multiple sensors to capture details (e.g., clothes wrinkles caused by movement) to satisfy the users' required quality [86, 94, 109, 231]. Substantial research has been conducted in related 3D segmentation [110, 175], object detection [34] and tracking, localization [36], classification [72], simplification [140], registration [81, 204], and compression [146].

As a PC may require numerous points (especially for realistic rendering), the PC is not always the best choice for the 3D data representation of an object or scene due to inefficiency in memory usage, transmission, and certain processing. Therefore, using a 3D mesh can be a good alternative for data representation [235]. A 3D mesh can be created by a graphics designer directly or automatically derived from a PC and has a connection relationship between vertices defined by a topology.

Application scenarios of 3D PCs and meshes include facial recognition access control [181], immersive telecommunication [148], virtual (VR) and augmented reality (AR) [13, 163], image-based localization [36], robot navigation [108, 216], autonomous driving or flying [31, 40], traffic management [61], criminal investigation [17], building information modeling (BIM) [143], cultural heritage preservation [39], smart manufacturing systems [239], restorative dentistry and orthodontics [134], and medical and biology science [242]. The list could go on and on.

The scarcity concept in economics [186] indicates that only a finite amount of resources are available to meet unlimited human demands at all times. The resources in the context of uses of PCs and meshes are computing power, bandwidth, storage space, energy or battery, device cost and size, and others. Visual saliency and quality can be formulated, similar to the cases of images and video to meet maximum demands with scarce resources and environment considerations [23, 24, 126, 236]. These enable economical and optimal resource allocation priority to deal with salient portions of signals and achieve the best possible quality within the resource limit.

Similar to their counterparts in images and video, PC/mesh saliency [51, 115, 240] and quality [3, 56, 171, 225] are important research topics, to facilitate the best decisions in different processes, as stated above. However, they are related and interact with each other as well. In general for visual signals, signal changes in salient regions affect quality more than in other regions [126, 236]. However, salient regions may also change when quality changes with different distortion types and levels [57, 237].

This paper aims to present an overview of the research on saliency and quality modeling for 3D PCs and meshes and presents views on future possibil-

ities. As mentioned, along with points (vertices), a mesh also establishes the associated connection and surfaces, and 3D PCs and meshes are closely related. A 3D mesh can be reconstructed from a 3D PC (as presented in Section 2). In addition, PCs have a more primitive form than meshes and are more versatile and flexible in 3D representation, especially for large-scale scenarios (buildings, terrains, and cities). In contrast, meshes play a key role in CG, especially with rendering and visualization-related tasks, particularly for a single object, and possibly shrink the data size and facilitate direct uses of existing algorithms developed in the CG domain.

The rest of this paper is organized as follows. Section 2 reviews issues related to the representation of PCs and meshes. Sections 3 and 4 review the existing research on saliency and quality models for 3D PCs and meshes, with human and machine uses, respectively. More specifically, we discuss the saliency problem for finding the most sensitive regions for processing 3D PCs and meshes. Afterward, we examine the quality evaluation methods of PCs and meshes. It is hoped that the review helps create awareness and interest in the important, related topics. Compared with the cases of 2D images and video, saliency and quality for 3D PCs and meshes are still less investigated, especially concerning aspects and factors unique to 3D visual signals. Section 5 presents and discusses future opportunities and possible new technical approaches in line with the interests in CV, CG, VR, AR and the emerging metaverse [27, 117]. Finally, a summary and concluding remarks are given in Section 6. Moreover, Table 1 lists the abbreviations used throughout this paper for easy reference.

2 Data Acquisition and Representation

This section describes the two major approaches to 3D PC acquisition in Section 2.1. Second, 3D visual data cannot be transmitted and stored in the raw format due to the massive data volume; hence, PC compression is presented in Section 2.2. Since 3D mesh compression has a relatively long history and has been well-reviewed, readers can refer to the related surveys [145, 170]. Third, the conversion of a PC into a 3D mesh is discussed in Section 2.3, because the mesh is a better form of data representation for rendering, visualization, direct uses of other existing CG algorithms, and other applications. In addition, implicit function based 3D representation is explored as the last topic of this section.

2.1 Visual Data Acquisition

As the first major data acquisition method, 3D PCs can be directly acquired via dedicated hardware RGB-depth (D) sensors (Section 2.1.1). The second major method is to compute the depth information of a scene from a single

Table 1: List of abbreviations.

Abbreviations	**Full term**
3DHoPD	3D histograms of point distributions
3D-SURF	3D SURF
AI	artificial intelligence
AR	augmented reality
BIM	building information modelling
B-SHOT	binary SHOT
CG	computer graphics
CV	computer vision
CVT	centroidal Voronoi tessellation
FPFH	fast point feature histogram
FR	full-reference
GGD	general Gaussian distribution
G-PCC	geometry-based PCC
GPS	global positioning system
GRNN	general regression neural network
HVS	human visual system
IMU	inertial measurement unit
IQA	image quality assessment
IR	infrared
JND	just noticeable difference
LBP	local binary pattern
LiDAR	light detection and ranging
LoD	level of details
MLS	moving least squares
MOS	mean opinion score
MSDM	mesh structural distortion measure
MSE	mean squared error
MRMS	maximum root mean square error

Table 1: Continued.

NeRF	neural radiance field
NR	no-reference
PSNR	peak signal-to-noise ratio
PC	point cloud
PCC	PC compression
PEFR	precise extreme feature region
PFH	point feature histogram
PSIM	pose similarity metric
QoE	quality of experience
RAHT	region adaptive hierarchical transform
RANSAC	random sample consensus
RGB	red, green and blue
RGB-D	red, green, blue and depth
RR	reduced-reference
RRE	relative rotation error
RTE	relative translation error
SDF	signed distance function
SE(3)	3-dimensional special Euclidean group
SIDE	single image depth estimation
SfM	structure from motion
SHOT	signature of histograms of orientations
SIFT	scale-invariant feature transform
SLAM	simultaneous localization and mapping
SSIM	structural similarity image metric
SURF	speeded-up robust feature
SVR	support vector regression
ToF	time of flight
TSDF	truncated signed distance function

Table 1: Continued.

UDF	unsigned distance field
UoQ	utility-oriented quality
UoS	utility-oriented saliency
V-PCC	video-based PCC
VR	virtual reality

2D image (enabled by machine learning) or multiple 2D images captured from different locations and viewing directions, as described in Section 2.1.2.

2.1.1 Direct Acquisition via RGB-D Hardware

Depth sensors typically obtain the depth map (reflecting the distance to the sensor) of a scene (i.e., the geometry information (x, y, and z) of a 3D PC from one viewing direction). The underlying technology can be structured using light, stereo/multi-camera vision, or ToF, including the LiDAR based on active remote sensing [75].

The RGB-D sensors refer to combinations of depth sensors and RGB cameras, to associate conventional images with 3D PCs on a per-pixel basis, for both geometry and attribute information (i.e., color in most cases, although auxiliary information, such as the derived surface normal and other feature descriptors, and may be included). Conveniently available RGB-D sensors include Microsoft Kinect, Intel RealSense, ASUS Xtion Pro, Structure Sensor Pro, and even iPhone12 Pro's cameras with a LiDAR sensor.

LiDAR is a type of ToF technology that measures round-trip time. An infrared (IR) laser released from the emitter hits the target object and returns to the receiver. The speed of light is constant; thus, the distance of the target object from the LiDAR sensor can be calculated from the round trip time. LiDAR is the most adopted approach with high accuracy and resolution, good detection range, and robustness to environmental conditions (temperature, dust, etc.). It can also include a global positioning system (GPS), and an inertial measurement unit (IMU) to provide the orientation. Color information (if not already captured) for each acquired point may be estimated via machine learning [78].

However, the LiDAR performance is highly dependent on the hardware specification of the IR sensors. The depth quality tends to be proportional to the device cost. In addition, LiDAR measures the distance by emitting and receiving an IR signal; thus, interference may occur among IR signals when multiple units are used simultaneously, resulting in performance degradation. When one LiDAR sensor is seen in a field of view of another LiDAR sensor,

serious performance degradation occurs, as shown in Figure 2. Therefore, LiDAR is more suitable for capturing a large-scale 3D scene than a small-scale 3D object (and it is better to use a multi-camera system, as described in the next paragraph).

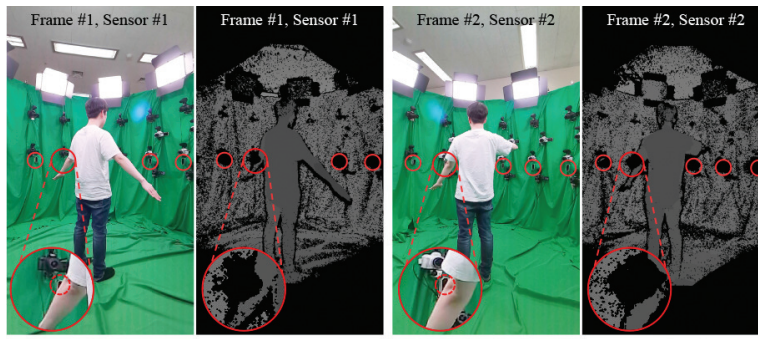

Figure 2: Depth quality degradation due to ToF interference between sensors demonstrated with a scene showing multi-sensors and a person. The location of the ToF unit in each sensor is marked with a red circle. A circled region is enlarged to view RGB and depth images better. Frames #1 and #2 were captured using two different sensors (namely, Sensors 1 and 2 respectively) which are not shown in the figures.

Stereo vision can measure depth from disparity using a geometrical formula [99] and has passive and active means. The passive stereo vision estimates the correspondence between stereo images without using IR patterns. Due to the matching ambiguity, it has lower accuracy in texture-less regions. However, the active stereo vision estimates the correspondence using IR patterns to enable higher accuracy even in texture-less regions. In such a system, IR sensors and patterns are generally used for depth estimation. Each stereo-vision RGB-D sensor has two calibrated IR cameras, one IR pattern emitter, and one RGB camera. The IR pattern and IR cameras capture the depth, and the RGB camera captures color. Moreover, Figure 3 depicts the depth quality comparison of passive and active stereo systems. In contrast to LiDAR, an active stereo system has diverse IR patterns, so multiple sensors can be simultaneously used to achieve accurate 3D reconstruction. A multi-camera system can perform a 360° reconstruction of small-scale objects.

2.1.2 Computational Depth Estimation

Depth information may be obtained without using dedicated hardware. Single image depth estimation (SIDE), as an ill-posed problem for a long time, has been made possible by deep learning to yield a dense depth map for a single

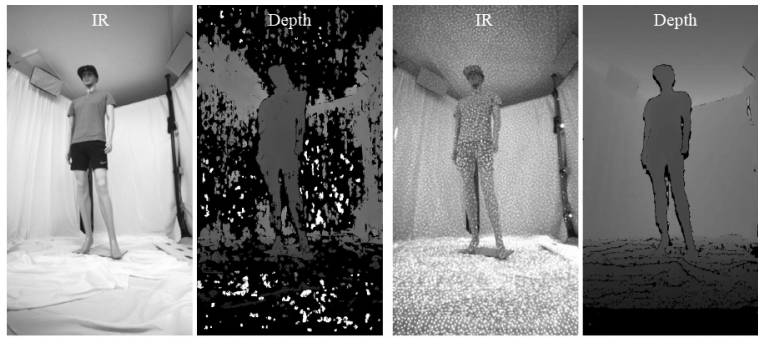

Figure 3: Depth estimation comparison between the passive and active stereo vision systems: A passive system does not use an IR pattern, whereas an active system uses an IR pattern (the snow-like pattern, emitted from the IR pattern projector).

image (i.e., a depth value can be determined for each pixel in a given image [21, 22, 150]). In addition, depth can be computed from multiple images of the same scene from different viewpoints. By matching 2D keypoints in images, a 3D model of a scene can be built, with the principles of stereo vision described in the previous part as a special case. The monocular-camera simultaneous localization and mapping (SLAM) is also based on the same concept, comparing the current frame with the previous frames for depth estimation.

Structure from motion (SfM) [220] is a well-known photogrammetric imaging process of estimating the 3D structure of a scene from a set of 2D images. The SfM reconstructs the 3D structure using the geometrical relations between image correspondences, which should be invariant under radiometric and geometric changes over multiple images. Traditionally, hand-crafted features, such as the scale-invariant feature transform (SIFT), are used to determine the image correspondences [136]. Recently, with the advances in deep learning, deep features have been used to determine correspondences [49].

The geometry between images is estimated using epipolar geometry [14], which describes geometry between 3D points and their projections on image planes. Image correspondences may be incorrectly matched because they are extracted with consideration of only the appearance in images. Therefore, it is necessary to remove outlier correspondences using epipolar geometry. Random sample consensus (RANSAC) [64] is the most popular outlier removal method. Based on the inlier image correspondences and camera projection matrics, the correspondence points of two images are mapped onto a 3D space to estimate the actual 3D coordinates of points. Finally, bundle adjustment [209], which jointly refines 3D point and camera matrics, is performed to reconstruct visually optimal 3D points. For a SfM-generated 3D PC, each 3D point stores the feature descriptors (e.g., SIFTs) of correspondences in the database images

employed to construct 3D points [30, 187, 189], apart from its geometry and attribute information.

To reconstruct a 3D structure that deforms over time, a nonrigid SfM was introduced [25]. The rigid SfM assumes that the target object is fixed, so it estimates only the camera parameters of a single frame. In contrast, since the nonrigid SfM assumes that the target object is moving, it estimates the camera parameters and object motion in all frames. Therefore, nonrigid SfM becomes an ill-posed problem because the number of unknowns grows with the number of frames. Due to the growing unknowns, non-rigid SfM was initially targeted on structured models [7, 166] to simplify the structure representation. As estimating the depth map and object masks has become feasible with the advent of deep learning, the authors of [211] proposed SfM-net which reconstructs the 3D structure and motion of a scene in an end-to-end manner. The SfM-net architecture is composed of structure and motion networks. The structure network reconstructs the depth map for a single frame, and the motion network estimates the object and camera motions. Given the depth and motion, the final optical flow map is generated and supervised using only the photometric loss without any other information.

2.1.3 General Data Structure of a PC

The j^{th} point in a PC, Ω, with K points, can be expressed as a vector:

$$\omega_j = [g_j, c_j, a_j], j = 1, 2, \ldots, K, \tag{1}$$

where $g_j = [x_j, y_j, z_j]$ and $c_j = [r_j, g_j, b_j]$ denote the geometry (point position) and attribute (RGB values) information, respectively, a_j represents the auxiliary descriptors, such as the derived surface normal, and in a SfM PC, it can represent the SIFTs and database image associated with the point:

$$\Omega = \{\omega_j, j = 1, 2, \ldots, K\}. \tag{2}$$

2.2 PC Compression

The standardization of PC compression technology promotes inter-operability and substantial cost reduction. In contrast, non-standard solutions may result in better coding performance (possibly allowing less computational requirements and reducing overhead in the bitstream), higher content controllability and protection, and more advanced research on the next-generation of standards.

2.2.1 Standards

The MPEG 3D graphics coding group has targeted efficient representation and compression of three major PC categories: static objects and scenes (Category 1), dynamic objects (Category 2), and dynamically-acquired LiDAR sequences (Category 3). Two distinct technological tracks have been identified for PC compression (PCC) standardization under two coordinated test models: TMC13 (i.e., geometry-based PCC, or G-PCC) [146] for Categories 1 and 3 and TMC2 (i.e., video-based PCC, or V-PCC) [233] for Category 2.

In TMC13 (G-PCC), the geometry and attribute information of a PC is coded separately. Geometry information should be decoded first to decode attribute information. For geometry compression, the codec represents a PC using an octree structure and assigns 1 bit to indicate the occupancy state of each octree node. It provides two attribute coding options: the region adaptive hierarchical transform (RAHT) based encoder [177], and the level of detail (LoD) based encoder [146].

Rather than directly encoding in 3D space, TMC2 (V-PCC) converts 3D points of a PC to 2D via projection from different viewpoints, a methodology that has inspired new research in related areas (e.g., see Section 4.3), and then takes advantage of the well-developed state-of-the-art video codec (high-efficiency video coding) to compress the resultant projected images/videos. The architecture of TMC2 consists of 4 major modules: patch segmentation, patch packing and occupancy map generation, image generation, and image padding and compression. The V-PCC is more suitable for dense PCs because a sparse PC results in holes (which must be padded) in image generation and inefficient compression. It may also be used for dense static objects and scenes (in Category 1) without a temporal dimension, although it was not originally designed for this. Fast coding mode decisions can be made for the V-PCC standard using occupancy maps [224]. Compared to TMC13, TMC2 is not suitable for large-scale, sparse, and noisy PC compression [131], because it has a higher time complexity and requires a large projection plane at the expense of increased coding overhead.

2.2.2 Non-standardized solutions

Compared with 2D images and videos, PCs have higher space dimensionality, exhibit characteristics of stronger structural irregularity, and pose challenges to exploiting point correlation to eliminate information redundancy. Enormous effort beyond the current standardization has been made in probing better solutions, which can be broadly classified into Class I for tailored codecs for PC data [15, 42, 107, 192], Class II for image/video-based compression [55, 233, 238], and Class III for emerging deep learning-based coding [80, 176, 214].

Studies of Class I have exploited the correlation between points based on octrees [42, 69, 107], the K-D tree [192] and subdivisional meshes [15], for geometry and attribute compression for PCs in Categories 1 and 2. For instance, the GSR codec [69] divides a voxelized PC into blocks of equal size and uses geometry-guided sparse representation to deal with structural irregularity. A block prediction scheme and entropy coding strategy were tailored to eliminate information redundancy within each block.

The research on Class II aims to take advantage of the well-developed image and video coding infrastructure for PC compression, in line with the strategy of TMC2. Work in this vein has focused on effective 3D to 2D transformations to map 3D points to 2D image pixels while maintaining the inherent spatial correlation between points, including rasterization [238], space-filling curves [55], and nonlinear dimensionality reduction [233].

The emerging trend (for Class III) is to apply deep learning to PC compression [80, 176, 214]. An octree-based deep entropy model, OctSqueeze, was introduced for LiDAR PC compression [80], and a conditional entropy model was designed to predict the occurrence of an 8-bit occupancy symbol for each octree node. A recent study [214] suggested a variational autoencoder based on an end-to-end stacked deep neural network for PC geometry compression to outperform the G-PCC codec by a large margin. Despite the progress, deep learning based PC compression is still in its infancy, and numerous open problems exist, such as objective distortion metrics and adaptivity to PCs with varied geometrical characteristics.

2.3 Surface construction from a PC

2.3.1 Mesh construction

Polygonal (triangulation) meshes have been applied in various use scenarios, such as 3D object animation and scene rendering. Mathematically, a mesh is a simplicial complex (simplex) structure to fit the geometric features of a 3D object, such as topology, surface and curvature (for rendering and other situations).

Three key requirements for mesh reconstruction from a PC [142] are (1) local region determination, (2) geometric feature maintenance, and (3) necessary point resampling toward an isotropic mesh with a specified number of points (a challenging requirement [50, 97]). Local regions can be determined in the local tangent space, such as the centroidal Voronoi tessellation (CVT) [33] and Gaussian kernel [241]. The neighbor relationship of points can be obtained and optimized in the local tangent space, but certain geometric features (e.g., external and internal edges) may be lost [142]. The external edges are outside a 3D object, whereas internal ones are inside (e.g., holes in the object).

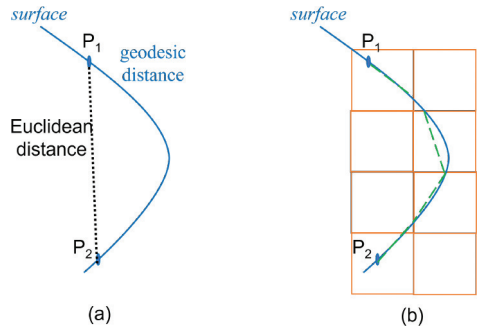

Figure 4: Illustration of the geodesic distance between two points (P_1 and P_2) on a surface: (a) geodesic distance against Euclidean distance (denoted by the dotted black line), (b) voxel-based approximation of the geodesic distance [140] (denoted by the dashed green lines), plotted in 2D space for illustration purpose.

More recently, mesh reconstruction has been performed using two steps: initial mesh reconstruction and mesh optimization [142]. For Step 1, based upon the intrinsic metric [67], the geodesic distance (not Euclidean) is used for calculating the distance between points, and the difference between Euclidean and geodesic distances is illustrated in Figure 4(a). This step leads to better determining local regions and neighbor relationships between points. For Step 2, the initial mesh is enhanced toward an isotropic one (Requirement 3). Remeshing can also be realized for isotropic surfaces by progressive eliminating obtuse triangles and improving small angles [217].

2.3.2 Function-based representation

Implicit function-based 3D representation is also possible, unlike direct PC (or mesh) representation which expresses 3D data only with points (and connections). For example, if a 3D sphere is expressed as a PC, it can be expressed as a set of points on the sphere surface. If it is expressed as a 3D mesh, it further defines the connection between points. However, when using a function, a 3D sphere with radius r can be expressed as $x^2 + y^2 + z^2 = r^2$, as an accurate and compact representation by an implicit function in an analytic form.

An implicit surface representation, such as truncated signed distance function (TSDF) [93], can be used to represent a 3D surface of an object from the depth image. A TSDF volumetrically represents the distance from the object surface to the voxel grid subdivided over world coordinates, as depicted in Figure 5. Contrary to a mesh, even if a TSDF does not represent the object surface, it effectively represents structural changes in which surfaces are combined and separated. The TSDF generated from a PC is dependent on the sensor viewpoint in cooperation with the perspective projection mechanism.

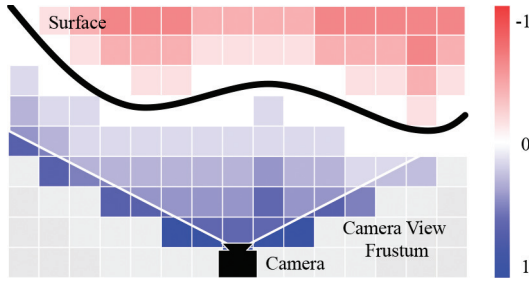

Figure 5: Illustration of TSDF representation. The outside of an object is represented as a positive value, and the inside is represented as a negative value. The object surface is represented as zero.

For a calibrated depth sensor, $\mathbf{X} = [X, Y, Z]^\top$ and $\mathbf{x} = [x, y, z]^\top$ are the discretized cubic voxels in the world and local sensor coordinates, respectively, and $\mathrm{T} : \mathbb{R}^3 \rightarrow \mathbb{R}^3$ denotes the transformation matrix from world to local coordinates. The TSDF, denoted as $\phi(\mathbf{X})$, is formed by computing the signed distances. A signed distance $\mathrm{dist}(\mathbf{X})$ is measured with a projection of the related points onto the depth map H [85, 158]:

$$\mathrm{dist}(\mathbf{X}) = H\left(\Pi\left(\mathrm{T}(\mathbf{X})\right)\right) - Z, \tag{3}$$

$$\phi(\mathbf{X}) = sgn\left(\mathrm{dist}(\mathbf{X})\right) \cdot \min\left(\left|\mathrm{dist}(\mathbf{X})\right|, \tau\right)/\tau, \tag{4}$$

$$\eta(\mathbf{X}) = \begin{cases} 1, & \text{if } \mathrm{dist}(\mathbf{X}) > -\tau \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

where $\Pi : \mathbb{R}^3 \rightarrow \mathbb{N}^2$, as the projection operator of $\mathbf{x}$ on to the depth map on the 2D coordinates, Z is the ground truth of depth, $sgn\left(\cdot\right)$ is the sign operator, $\eta(\mathbf{x})$ is the TSDF weight, and τ is the truncated margin to control the TSDF accuracy by ignoring regions far from the object surface and determining the thickness of the object surface. When TSDFs are yielded from consecutive depth frames, they can be combined by weighted averaging [46]. In addition, a TSDF can be conveniently changed to a mesh through the marching cube strategy [135].

KinectFusion [85, 158] is the first 3D reconstruction method with RGB-D sensing based on an explicit function. It obtains a 3D structure by fusing the depth sequences of a moving sensor to capture the entire surface of a static object. Thus, it only considers rigid motion when fusing depth sequences. In recent 3D reconstruction methods, handling the non-rigid motion of a moving object has been also proposed [58, 59, 82, 94, 157].

Recently, neural networks [41, 169] have been developed to obtain the signed distance function (SDF) rather than directly obtaining it from a depth sensor with camera parameters. The neural network estimates the 3D surface

from the PCs without using camera parameters. For example, DeepSDF [169] estimates the SDF from PCs using a neural network as the regression tool. Estimating the SDF using neural networks displays powerful performance in the dense surface completion from sparse PCs. Implicit surface representation using neural networks is also used to estimate surfaces from multi-view RGB images. Neural radiance field (NeRF) [154] computes a 3D surface and its color implicitly from multi-view RGB images.

3 Saliency Modeling

The 2D image (or video) saliency [23, 24, 84, 139, 227] can primarily be determined using contrast/distinctiveness of luminance, color, texture, pixel motion in 2D planes, as well as high-level semantics, center-bias phenomenon, and so on. A depth map or its derivatives can be used as supplements with 2D image saliency to determine the saliency of RGB-D images, as surveyed in [44].

Although these aforementioned attributes (color, texture, motion, semantics, and others) still play critical roles in PC/mesh saliency [200], PCs/meshes have higher dimensionality, and their modelling needs further incorporation of 3D object geometry and surfaces and also consideration of the viewpoint. Modeling for PCs can be even more difficult because points in a PC are more disordered and unstructured (i.e., lacking explicit information for connections). This section is devoted to a comprehensive review human and machine use, for handcrafted and learning-based approaches developed for computational PC/mesh saliency models, which can be applied in various utilities, including simplification, compression, point/shape registration, segmentation, localization, and viewpoint selection to name a few.

3.1 Saliency for human uses

Visual saliency for PCs and meshes can be defined according to the characteristics of the human visual system (HVS) evidenced psychologically [221], (i.e., to determine more perceptually important regions regarding their surrounding regions), for human observation, appreciation, and understanding.

3.1.1 Perceptual PC Saliency

Early research in [194] identifies and integrates the following three features for PC saliency maps based on a hierarchical HVS perception mechanism:

1. low-level (local) distinctness, evaluating if a point is dissimilar to its neighbor points;

2. point association, regarding regions near the focus of attention to be more interesting than distant regions; and

3. high-level (global) distinctness, comparing large neighborhoods.

With the work in [205] for PCs, a set of points are first decomposed into small clusters using fuzzy clustering, and cluster distinctness and spatial distribution of each cluster are combined as a cluster saliency function. Finally, a saliency score is assigned to each point using the probability of belonging to each cluster. Yun and Sim [232] detected saliency by first voxelizing a PC and grouping points into super-voxels. Then, the cluster hierarchy is constructed iteratively by calculating the cluster similarity, and the cluster saliency is computed based on the distinctness of geometric and color features. The method in [114] brought sparse coding into saliency detection, by extracting features in local neighborhoods and applying sparse coding to these features. Finally, the saliency is predicted based on the minimum description length principle. The work in [51] improved the ideas in previous work [194, 205] using the adaptive fusion of the local and global distinctness, including local features, and employing random walk ranking for global distinctness.

3.1.2 Perceptual Mesh Saliency

Early research on saliency detection on meshes aimed to project a 3D object to 2D planes [147, 229]. The work in [115] is among the first to measure mesh saliency directly in 3D space. The authors of [115] computed saliency based on the mean curvature, and the results can guide mesh simplification and viewpoint selection. Wu *et al.* [222] introduced the concept of global rarity and the method is based on local contrast. It can be used for mesh smoothing and simplification. Leifman *et al.* [120] introduced the region of interest for surfaces, by defining the vertex distinctness and shape extremities characterizing the local structure. The method can be applied for viewpoint selection. Limper *et al.* [124] estimated saliency based on Shannon entropy (i.e., the expected information value of visual content), using the mesh curvature. In [199], in addition to local saliency, the authors proposed a measure of global distinctness based on a statistical Laplacian-based algorithm that computes saliency at multiple scales. The recent work [200] investigated whether mesh saliency can be derived from 2D saliency, and a weakly supervised learning method was proposed to learn to predict 3D saliency.

3.2 Keypoint detection & 3D feature descriptors for machine uses

Keypoints (interest points) refer to stable (repeatable) points with well-defined positions that play critical roles in many CV and other machine-oriented

tasks. Keypoint detection has the general requirements of robustness to noise, compactness for description, repeatability under arbitrary 3D special Euclidean group (SE(3)) transformations, and computational efficiency [71, 194]. A 3D feature descriptor depicts the local geometric features of a point (including a keypoint) to facilitate many processes that follow (e.g., feature matching for recognition [92] and PC/mesh registration [137]).

Keypoint detection can be performed either independently [197, 202] or with a 3D feature descriptor [182, 183, 185]. The keypoints and related 3D feature descriptors can also be determined together [18, 138, 234]. If more keypoints are detected within a region, they form a salient region. Thus, detecting regions with a sufficient interest points is a reasonable first step for saliency detection models. For instance, a set of points with high curvature [174], instead of a few single points at higher curvature that might sometimes arise due to noise, can more reliably detect the keypoints and salient region in a PC.

In this section, two major, distinct approaches, handcrafted and learning-based, are reviewed for 3D keypoint detection and the associated 3D feature descriptors.

3.2.1 Handcrafted Approaches

For handcrafted 3D feature descriptors, most methods are based on histogram evaluation, as in the 2D cases. However, there have not been widely adopted descriptors unlike the great success of the SIFT [136] or speeded-up robust feature (SURF) [20] in 2D cases, although SIFT has been extended as 3D SIFT [191] and PointSIFT [90] for PCs, while 3D-SURF [106] is an extension of SURF [20]. This situation is primarily due to the lack or insufficiency of textual and semantic information.

The signature of histograms of orientations (SHOT) [184, 207] is a method based on the signature histogram that performs binning in local support to construct a repeatable local reference based on the covariance of the neighbors within a specific radius of each point. In each local region, points are binned based on the angles between their normals and those of the feature point. The histogram bins are defined by the radial, azimuth and elevation axes. The binary SHOT (B-SHOT) [173] has also been proposed with far less required memory and is much faster in descriptor matching. The point feature histogram (PFH) [183] describes the geometry of each point locally with a set of 16D features to exploit the neighborhood relationships of a point, and has been developed into the fast PFH (FPFH) technique [182]. Another histogram-based techniques is the 3D histograms of point distributions (3DHoPD) [172].

The spin image [92] is a shape descriptor that matches 3D surfaces represented as a PC or mesh, for efficient object recognition in cluttered 3D scenes. The method constructs a cylindrical system with two axes based on the

normal and tangent planes at each point. For every point under consideration, each neighboring point is mapped to this cylindrical system by computing the perpendicular distance to the line through the considered point parallel to its normal (normal line) and the signed distance to the tangent plane.

Keypoints (Schelling points) can also be found for 3D meshes via analysis [32]: symmetry, local curvature properties, and global properties (including segment centeredness and proximity to a symmetry axis).

3.2.2 Learning-based Approaches

Learning-based approaches usually achieve better performance in determining keypoints, but require more computational power than handcrafted methods. In [230], a weakly supervised keypoint detector, the first learning-based 3D keypoint detector, uses triplet loss and attention mechanisms to learn feature correspondences from PCs. In [121], an unsupervised keypoint detection method is proposed to produce highly repeatable and well-localized keypoints under arbitrary 3-dimensional special Euclidean group (SE(3)) transformations, with a feature proposal network to generate a set of keypoints and their respective saliency uncertainties from a 3D PC. In more recent work [137, 193] based upon PointNet++ (a PC segmentation network) [175], the point-wise normalized scores (to represent the probability that each point belongs to one of the k classes of segmented objects) are used to obtain keypoints [193], or the farthest point sampling from the PointNet++ output to choose a subset of points as the keypoints [137]. The log-Laplacian spectrum of a mesh [198] and angular deviation of normal vectors between neighboring faces [89] can be used as feature descriptors for 3D meshes.

Learning-based approaches [18, 138, 234] have been used to determine 3D feature descriptors or joint determination of keypoints and the related 3D feature descriptors. The 3DMatch [234] is a 3D ConvNet comprising the convolutional and pooling layers and the activation function, which takes in a 3D patch (30x30x30 voxel grid) around each keypoint. The output, a representation with 512 features, is the descriptor for the local region under consideration. The training minimizes the L2-distance between descriptors of corresponding points as matches, and maximizes the L2-distance between those from non-corresponding points as non-matches. The model was trained with 8 million correspondences from a collection of 62 3D (RGB-D) scenes, from multiple views. In [18], it is proposed to jointly learn a keypoint detector and a 3D feature descriptor, instead of training separate networks for keypoint detection and description.

The PC keypoint detectors and descriptors can be evaluated [121, 138, 230] using indoor settings (e.g., the 3DMatch [234] dataset) and outdoor settings (e.g., KITTI [65], Ford [165], and Oxford RobotCar [144] datasets), in

terms of repeatability, precision and point registration, under different SE(3) transformations. First, the repeatability (stability) of the detected keypoints is determined as the ratio of repeatable keypoints to all detected ones [121], for various transformations. Second, precision is used for jointly evaluating a keypoint detector and descriptor [230]. With a source keypoint p_s, the corresponding target keypoint p_t in a transformed PC is searched for based on the descriptor in the nearest neighbors. If p_s and p_t are within a distance threshold, the correspondence is found, and precision is the validation ratio. Finally, a point registration via RANSAC [63] is successful if the relative translation error (RTE) and the relative rotation error (RRE) are sufficiently small [121].

3.3 Utility-oriented Saliency

As introduced in Section 1, practical 3D PCs and meshes have diversified scales (from a single object to a large city, shown in Figure 1), utilities (recognition, localization, and so on), and ultimate users (i.e., humans or machines). Naturally, computational saliency models can be (or even better to be) utility-oriented.

A utility-oriented saliency (UoS) model can enable the optimized outcome of the focused utility and its related applications. In [240], 3D PC recognition performance has been considered in point-wise saliency evaluation. The resultant UoS (although this term is not used in the paper) map explains which points are salient for PC recognition. Dropping points with negative scores leads to better recognition performance. However, if the points with the highest saliency scores are dropped, recognition performance is significantly decreased, creating the potential to build an adversarial attack model. Most follow-up work has focused on adversarial attacks or defenses for PCs, such as [103, 130].

With city-scale PC simplification toward image-based localization as the utility [28, 37] using SfM PCs (presented in Sections 2.1.2 and 2.1.3), the saliency (relative visibility [37]) of a 3D point ω_j in Ω can be derived below, if the formulation in Equation (1) is followed:

$$\nu_j = \frac{O_j}{M}, \tag{6}$$

where O_j and M denote the number of database images observing ω_j and the total number of database images in Ω, respectively. In other words, a 3D point supported by more of the PC's constituent (database) images (from different locations and camera poses) is more salient for localization because it potentially provides more 2D-to-3D feature correspondences for the camera pose (SE(3) transformations) estimation. Local visible points have also been selected due to outlier filtering for the RANSAC stage [35], which are especially useful in urban scenes with usually strong visual occlusions.

Other utility information, domain knowledge and user requirements can influence saliency modeling (e.g., for PC pre-processing [140], shape registration [137, 230], compression (as introduced in Section 2.2), and so on). More, convincing research is called for UoS modeling.

4 Quality Evaluation

Similar to the exploration on saliency presented in the previous section, PC and mesh quality evaluation is essential for human and machine uses as well. There are several common sources of distortion and quality degradation. First of all, as sensors such as LiDAR/ToF rely on infrared reflection, they are vulnerable to various types of noise during an acquisition process. Furthermore, the related 3D expression is mainly for 3D objects at a relatively short distance because of the limited capability of depth finding. In addition, since 3D data usually is in very large volumes, various artifacts may occur during simplification and compression due to the limitations in practical transmission and storage. In the case of 3D reconstruction of photogrammetry, noise occurs when the correspondence between image feature points is not properly matched [14]. If PCs and meshes are delivered to human users, the user's sense of immersion and realism may be impaired, and sickness may occur when viewing 3D content [102, 161, 167].

Like the 2D counterpart of visual signals [126, 236], there are two basic types of PC and mesh quality assessment: subjective assessment and objective (computational) assessment, respectively. The first type conducts perceptual evaluation directly using human subjects (in Section 4.2), while the second type performs computational (machine-based) evaluation (in Sections 4.1 and 4.3). Furthermore, both basic types of 3D PC and mesh quality assessment can be with full-reference (FR) and no-reference (NR). Objective assessment can be of reduced-reference (RR) also. Most related research has been for FR so far. Quality evaluation can be done with geometry alone, or geometry and color (maybe normal and other contextual information as well) together.

4.1 Evaluation for signal fidelity

As in 2D image and video cases, the most straightforward objective PC quality metrics are for signal fidelity [60, 171] when the reference is available (i.e., for FR situations), based on certain geometric distance or/and color distortion measurement between points and/or local surfaces of a PC and its reference. Such FR evaluation can be with mean squared error (MSE), peak signal-to-noise ratio (PSNR) or one of their variations [62], to be used in situations such as a PC encoder for compression (Section 2.2), where a reference is available,

Correspondences can be identified (e.g., by a nearest neighbor algorithm) between a target PC and the reference one, if the registration is not already

available. Geometry-alone fidelity assessment can be divided into the following three types [68].

1. Point-to-Point [149]: For each point of a target PC, point error is computed based on the Euclidean distance, indicating the displacement of the distorted point from its reference position.
2. Point-to-Plane [206]: This is based on the projected error of a target point along the normal of the reference point.
3. Plane-to-Plane [12]: This is based on the angular similarity of tangent planes corresponding to the associated points between reference and target PCs. To be more specific, using the normal vectors for the two PCs, the similarity is computed with the two angles formed by the intersecting tangent planes.

The total error is then measured by evaluating all point/plane errors for each type above, with MSE, PSNR or a variation.

Similar methodology applies for meshes, and measures to provide fidelity analysis for meshes include the shortest distance map [74], Hausdorff distances [19], and moving least squares (MLS) error [123]. All these measures estimate the geometric consistency between a mesh and one of its variants.

4.2 Evaluation for human uses

4.2.1 Perceptual PC evaluation

With PCs for humans to observe and judge (usually for relatively small-scale objects or scenes), subjective viewing experiments can be conducted [171, 203], via visualization of 2D rendering and reconstructed 3D surfaces. Subjective assessment has been conducted for AR [13], PC denoising and compression [11], visualization strategies [10], and with voxelization and projection assistance [208].

In [29, 65, 73, 133, 196, 201, 203, 223], more subjective viewing data have been collected with different sources, scales, numbers of objects and points, geometric structures, distortion/distort levels, and represented scenarios (indoor/outdoor). The databases with mean opinion score (MOS) provide the ground truth of performance bench-marking, model training, or both (e.g., in [133, 171, 225]). However, subjective evaluation and the objective modeling associated with it are not meaningful for city-scale PCs and rapidly increasing applications with machines as users.

From the perspective of the HVS, due to foveation (like the 2D cases [118, 119]), it is expected that a region surrounding a keypoint contributes more to modeling quality of a PC than one that is far away. Besides, the HVS does not perceive the resolution of depth information as sensitively as that of textural information [100, 101, 104, 162]. Hence, it has been studied to provide a more realistic rendered view by representing texture more precisely than depth [95]. The human user's visual discomfort can be also predicted [102, 161, 167].

4.2.2 Perceptual mesh evaluation

There have been steady studies to obtain human-judged scores for mesh quality. When evaluating the subjective quality of meshes, it is difficult to grasp the overall evaluation of 3D content by visualizing one view. Therefore, multi-factors such as viewpoint, object rotation and translation, and effect of light and shading must be carefully considered [109]. It is necessary to consider the existence of self-occlusion and the back of the mesh that is not rendered from a specific viewpoint, and then to view the observed results from various viewpoints. Therefore, there have been various attempts to measure subjective mesh quality, depending on whether observers like to see it as static [219] or with rotation, zoom, and translation [45, 70]. Investigation has been also performed for the effect of the number of vertices and the resolution of texture [164].

Rogowitz *et al.* [180] showed that the distortion perceived by a human observer is greatly affected by the position and type of light. A subjective quality evaluation study [70] was conducted following the optimal lighting conditions [160] when evaluating the quality of a mesh. Since error is more visible in smooth regions of a mesh, an HVS bias is possible, and in [215], it was mimicked by using the visual masking effect and psychometric saturation as modulation for surface roughness values, which are first defined as the Laplacian of Gaussian curvature values obtained from the considered vertex and the surrounding vertices, and then modulated using a mapping function and a threshold.

Various conditions need to be considered to construct a database for mesh quality assessment. The general-purpose LIRIS/EPFL database [112] contains distorted meshes with smoothing and noise added to an entire mesh or a part of it, and the LIRIS Masking database [113] is composed of 6 distortions for the purpose of considering human visual masking effects. The UWB compression database [210] and the IEETA simplification database [195] were also published for implementing distortion from compression and simplification, respectively.

4.3 Quality evaluation with feature analysis for machine uses

4.3.1 PC evaluation with feature analysis

Geometric and contextual features can be analyzed for their respective implications for PC quality. Examples include the angular similarity of points [12], color statistics extracted from reference and target PCs [213], local luminance patterns [52], local binary pattern (LBP) at each point for texture [53], the precise extreme feature region (PEFR) [178], the generalization of the Hausdorff distance [87], and the influence of scale changes [88].

Following the concept of structural similarity image metric (SSIM) [203, 218], an FR geometry-alone metric, PC-mesh structural distortion measure (MSDM), has been proposed [152] for structural differences (captured via curvature statistics) in the corresponding local neighborhoods between a

reference and target PC. The method consists of three steps: (1) curvature computation to obtain the local structural information for each point, (2) establishment of correspondences between PC, and (3) calculation of the statistical curvature difference between the points in the spherical neighborhood around each target point and the corresponding reference points. The PC-MSDM has been extended to PCQM [153] for colored 3D PCs, using a weighted linear combination of geometry-based features (based on PC-MSDM) and color-based features. For the color-based features, it extends the image-difference measures [129] of lightness, chroma, and hue to PCs. Similar to PCQM, geometry and color descriptors [54] are proposed with the normal distance and CIEDE2000 color difference, respectively, to independently extract features from reference and target PCs. More recently, a method has also constructed relationships among points for a quality evaluation using graph signal processing [225].

The FR PC quality assessment can also be conducted by rendering PCs from different viewpoints (motivated by the projection in Section 2.2.1 for PC compression) using an image quality assessment (IQA) metric on the resultant projected 2D images. For instance, an image quality metric is applied to projected images from PCs [9]. In [226], a six-sided (front, back, left, right, upper, and bottom views) perpendicular projection is used for texture and depth maps, and then the features extracted from these maps are fused as a PC quality index. For each projection plane, the 2D texture and depth images represent the photometric and geometric information of a PC, respectively.

The RR and NR metrics have emerged in recent years. An RR PCQM has been proposed to extract 21 global features from geometry, color, and normal data [212] as the partial reference. Geometry-based features are variations of the statistical distribution of points along the three axes (X, Y, and Z). The color histogram represents the change in the general distribution of colors when artifacts are introduced. Normal-based features measure the similarity between the normal vector of a point and that of its neighbors. A quality score is obtained using a linear combination of multiple features. Another RR metric for PCs encoded with V-PCC was presented in [130], based on a linear model of geometry and color quantization parameters, determined using local and global color fluctuation features.

A NR PCQM has been presented in [132] by first projecting a PC into six 2D images as described above and then feeding the projected images along with their ground-truth distortion types and MOS, into the proposed deep-learning network, PQA-Net. In [38], three relevant low-level features (i.e., geometric distance, local curvature and luminance values) from local patches are used. Specifically, a set of N points is randomly selected in a PC. A region is formed around each selected point by finding its Γ nearest neighbors, characterized by these three features. Afterward, deep neural networks (AlexNet, VGG, or ResNet) learn the mapping from the features to the ground-truth MOS of the PCs.

4.3.2 Mesh evaluation with feature analysis

Early studies measured simple geometric distances for FR assessment. In [190], the authors measured the geometric distance (Figure 4(a)) between a reference mesh and its associated version generated by mesh simplification. To measure the distance caused by distortion, Hausdorff distance, MSE, and maximum root mean square error (MRMS) were used [43]. To evaluate the visual quality of the compressed meshes, Karni and Gotsman [96] measured the difference in geometric Laplacian values between a reference mesh and its distorted mehs. Each Laplacian value is obtained by the weighted sum of the Laplacian coordinate errors between one vertex and its neighborhood vertices to express the smoothness of each vertex. Lavoue *et al.* [112] defined a local window on the mesh and calculated the local curvature using the eigenvalue obtained from the curvature tensor. The MSDM uses the calculated local curvature features to define and compare the curvature, contrast, and structure terms in a method similar to the SSIM. However, the MSDM has the limitation that the reference and processed/distorted mesh must have the same connectivity, so MSDM2 [111] uses a multiscale window to allow different connectivity to be applied to the two meshes. The dihedral angle between the face normals of a mesh can be extracted as a feature [210].

The reference mesh for a given distorted mesh does not exist in many cases; thus the NR quality assessment has been studied alongside the development of FR metrics. Studies have been conducted on the regression of quality scores using limited features extracted from a processed/distorted mesh. With the available databases, it is possible to train a model. Abouelaziz *et al.* [5] used statistical distributions, such as Rayleigh, Gamma, and Weibull distributions, by fitting the dihedral angle feature obtained from a mesh using support vector regression (SVR). In [4], the quality score was predicted by training general regression neural network (GRNN) using a curvature feature similar to that used in MSDM as input. In [128], the shape and curvedness indices were obtained by fitting the histogram to general Gaussian distribution (GGD). The final score was predicted using a random forest method using the vertex

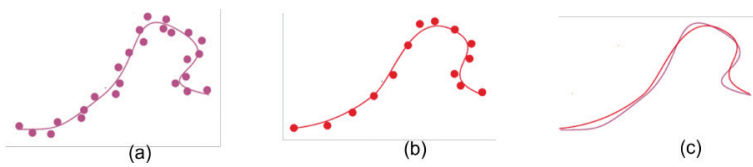

Figure 6: Illustration of the MLS surface fitting: (a) the purple curve represents the true (ideal) object surface, and purple points are an original PC (subject to noise and errors in data acquisition); (b) red points are the processed PC after preprocessing (denoising, resampling, simplification, etc.) [140–142], whereas the red curve is the MLS fitting; (c) the MLS result (red curve) is compared against the true object surface (purple curve).

scatter, structure, area features, shape index, and curvedness index. Other recent NR studies have used machine learning [1–3, 127].

Furthermore, quality related issues unique to 3D PCs and meshes include the isotropic property (distances between each point and its neighbor are approximately equal) [142], geodesic measurement and its approximation (Figure 4), and MLS surface fitting [8, 140] (illustrated in Figure 6). More careful research is needed for these aspects. In addition, the 3D feature descriptors discussed in Section 3 can help identify regions of high importance to develop quality metrics.

4.4 Utility-oriented quality evaluation

Similar to the cases of UoS (Section 3.3), a utility-oriented quality (UoQ) evaluation for PCs and meshes can be useful and might even be necessary for many situations due to the vast diversity of the scales and tasks. The UoQ can address a specific object or scene (e.g., a bridge [155] or an indoor setting [79]). The UoQ can be for a specific purpose or application (e.g., a PC should have more visible 3D points with an arbitrary view in the scene for image-based localization [37], as explained next), whereas PC-mesh construction has different requirements (see Section 2.3.1).

Explicit conceptualization of UoQ for machine uses has not been addressed in the literature yet; thus, an NR example (which is more realistic in practice) is provided, based on a closely related study, to present the concept and trigger discussion and further exploration.

For city-scale image-based localization with SfM PCs, UoQ (localization-oriented in this example) can be formulated based upon the findings in [37] with simplified assumptions (although the concept of utility-oriented PC quality was not presented in the cited paper). The density of the database images (denoted as D) in a PC is an important factor for UoQ because the quality of different, diversified is to be compared and calculated based on the overlap extent of co-visible 3D points from database images [37]. The relative visibility, ν_j, is derived in Equation (6) for the j^{th} point in a PC, ω_j, and the associated probability for visibility can be expressed as follows:

$$V_j = \nu_j f(D), \tag{7}$$

where $0 \leq f(.) \leq 1$, as an appropriate weighting function that increases with D, and a possible form of $f(.)$ is:

$$f(D) = 1 - \alpha e^{-\beta D} \tag{8}$$

with α and β as positive parameters.

For each point, the visibility probability, V_j, is assumed to be independent of the existence of other points (i.e., λ_j represents a related independent

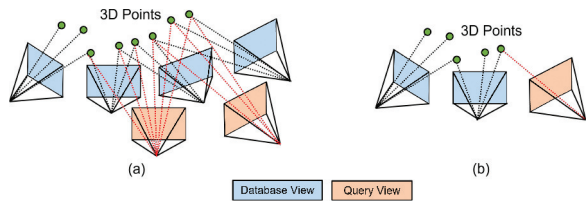

Figure 7: Illustration of two PCs with at least three visible points (as simplified examples because useful situations should have more visible points) from each database view: (a) a dense PC with $Q = 3$ and (b) a sparse PC with $Q = 1$ (although database views have three visible points).

Bernoulli trial, $Bernoulli(V_j)$. $\lambda_j = 1$ when ω_j is visible for a view (camera pose), and $\lambda_j = 0$ otherwise).

With an arbitrary view in the scene, whether it is a database image view or a query image view, the probability that ω_j is visible, $p(\lambda_j = 1)$, and the total number of visible points, Λ, can be determined as follows:

$$p(\lambda_j = 1) = V_j, \quad \text{and} \quad \Lambda = \Sigma_j \lambda_j. \tag{9}$$

The cumulative distribution for at least γ points visible with a view, $p(\lambda > \gamma)$, can be estimated via V_j in a given PC [37]. The UoQ of an SfM PC can be defined as follows:

$$Q = \gamma_{max} = argmax_{\gamma}(p(\Lambda > \gamma) = 1), \tag{10}$$

where the maximum γ value is sought. For a PC, the process expressed above determines the value of Q, and we know that at least Q visible 3D points (i.e., 2D-3D correspondences for RANSAC) exist from any view. A higher Q leads to higher localization accuracy by machines and higher UoQ for the PC under consideration. Figure 7 illustrates two cases with different Qs. An example of dense PCs (with a higher Q) is the Rome database [6] computed with considerable visual overlap between database images, whereas an example of sparse PCs (with a lower Q) is the Aachen database [188] from sparsely located images.

In [179], the influence of multiple echoes provided by LiDAR has been investigated for performance enhancement with PC classification; therefore, such additional information can be meaningful for UoQ evaluation. For the evaluation of mesh construction and remeshing, the intrinsic and isotropic properties [142, 217] are critical factors for building an associated UoQ metric.

5 Possibilities Ahead

In this section, the potential new research is explored and discussed for topics related to saliency and quality modeling for 3D PCs and meshes, apart from those mentioned in the previous sections whenever appropriate, based upon our R&D experience in the field, and extensive reading, thinking and reassessing during this work. This is organized into two sections: new important topics as an extension of existing research and new topics that are even more forwarding-looking.

5.1 Further research on important topics

Limited by scanning devices and usually complicated scenes under attack in the wild, most raw PCs cannot be directly used for effective and efficient applications due to the high data redundancy, unavailability of semantic information, object occlusion, scanning noise/interference, and environmental and weather (if outdoors) conditions. Further investigation is needed for a better methodology for research on various topics discussed in the previous sections, especially addressing issues unique to 3D PCs and meshes, such as geodesic measures in 3D space, and more global manipulations for keypoint detection to overcome the difficulty caused by no or insufficient textual and semantic information.

In particular, saliency and quality modeling for machine tasks required substantial attention from research communities. This need is anticipated because more PC- and mesh-related tasks are to be accomplished by machines rather than humans in the artificial intelligence (AI) era with waves of digital transformation in industries. There are possibilities for formulating an evaluation for human and machine uses. Humans and machines share a substantial commonality in saliency and quality requirements with 3D PCs and meshes, and the current machine learning architectures and algorithms were inspired by the HVS, and other related human brain and neural functioning.

Considerable PC saliency and quality research has adopted approaches similar to that of the 2D counterparts; thus, intrinsic geometric/topological analysis unique to 3D signals requires intensive effort to demonstrate the values and advantages in practice. Furthermore, related lightweight learning and model compression can also be explored for green computing. More work is also expected to be conducted on RR and NR PC quality evaluation. Another possible direction of exploration is PC compression, because PC representation is a fundamental issue, and like other forms of visual signals, a large PC in practice cannot be transmitted and stored in its raw format. Highly PC-dedicated codecs can be developed for effectiveness, and efficiency for all three PC categories presented in Section 2.2. Saliency and quality evaluation can certainly play active roles in new PC compression frameworks. The concept of a just-noticeable difference (JND) [91, 125] may be extended to PCs and

meshes for saliency and quality modeling, compression, and other related manipulations, for human and machine use. Work can be extended more for nonrigid 3D PCs and meshes [204] (e.g., human bodies are nonrigid by nature). In addition, more research is needed to explore the influence of saliency on quality and the interactions between quality and saliency.

It is impossible to express a surface analytically for arbitrary objects. However, with the advent of deep learning, functions that are non-linear and difficult to represent analytically can be expressed through deep learning, and we can reversely obtain a continuous and elaborate implicit function representation of 3D data. The 3D representations based on implicit function (as introduced in Section 2.3.2) have many different forms, including occupancy networks that express the area occupied by 3D data [151], the signed distance function (SDF) and the unsigned distance function (UDF) expressing the distance between an arbitrary point and a 3D surface [41, 169], and NeRF that enables volume rendering by expressing density and color [154]. Currently, implicit function-based 3D representation is actively studied, but no attempts have been made to quantify the saliency or quality evaluation using such representations for 3D PCs and meshes. We believe that the saliency and quality evaluation can benefit from by either extracting PC or mesh features from the implicit function or directly analyzing the implicit representation.

Utility-oriented models (i.e., for UoS and UoQ) are particularly meaningful for 3D PCs and meshes because of the vast diversity in scales (from a toy to an industrial part or from a building to a big city) and the nature of the utilities/tasks. As initial attempts, the research can start by referencing some principles applied to the saliency or quality metrics for PCs and meshes, as in Sections 3.3 and 4.4, for human and machine use. Careful UoQ modeling is useful in numerous practical PC/mesh processes mentioned in all sections above, including acquisition, preprocessing, compression, simplification, segmentation, detection, tracking, registration, classification, and so on.

Domain-specific perceptual analysis of the 3D human pose and face is one example of further research on 3D information, where it has been found that errors on some joints (e.g., head, neck, spine, shoulder, and hips) are more salient than other joints (such as knees, ankles, elbows and wrists) [116]. Temporal naturalness is also important because severe distortions of the pose degrade the overall quality of a pose sequence [105, 168]. With such prior knowledge, the pose similarity metric (PSIM) quantifies the spatio-temporal structural error of a 3D pose similar to the human visual perception [116]. By introducing the PSIM, the objective metric scores exhibited a much higher correlation with the subjective scores than the original Euclidean distance metric [83]. Such an approach is expected to significantly facilitate future research on 3D meshes and PCs in specific domains.

Hand-crafted models have been developed for compression, saliency detection, and quality evaluation. These models have several advantages, such as simplicity, generalizability, and interpretability, but with less accuracy.

Many deep-learning methods have been studied and can closely mimic human perception and fulfill machine tasks in a data-driven manner. The current limitation of learning-based models lies in the lack of interpretability, substantial required computation, and big data required to generalize in practice.

5.2 Exploring more advancement

To date, PC and mesh quality for human uses is primarily evaluated in terms of technical quality (similar to the relevant concept in 2D cases [76]), which accounts for the major factors in the signal life cycle (from acquisition to processing to the final utility). These factors usually are low-level defects/changes (e.g., noise, compression distortion, transmission error, processing artifacts, etc.) as discussed in the main body of the preceding sections. For human uses, aesthetic quality [48] also matters apart from the technical quality, and concerns more abstract and higher-level judgment of beauty (e.g., object composition, lighting, color harmony, and even personality [77, 122]). The quality of experience (QoE) [26] is a holistic concept of the delight or annoyance of users' entire experience with a PC or mesh, which is worthy of careful exploration. It should consider technical quality, aesthetic quality, visual discomfort (as mentioned in Section 4.2.1), and so on.

Furthermore, as already highlighted in Section 1, the recent rapid development and availability of 3D sensing/computing have enabled the integration of CV and CG when this is needed or preferred. The massive and economical depth information with PCs, used alone or fused with RGB data, provides significant solutions to solve or simplifies many challenging CV and CG tasks. Examples include object detection, scene parsing, pose estimation, visual tracking, semantic segmentation, shape analysis, image-based rendering, and 3D model reconstruction, which were ill-posed, prone to errors/mistakes, expensive, or even otherwise impossible. Therefore, more research is called for regarding an integrated CV and CG framework and exploiting new territories of applications.

In the emerging metaverse [27, 117], humans expect to behave and interact with others through a 3D virtual world, similar to the real, physical world. Although the nature and working of the metaverse must still be evolved and defined, 3D visual content plays a crucial role because it is an enabler for the experience of realism, immersion, flexibility, and more importantly, user interaction. Consequently, the saliency and quality evaluation of 3D visual content, captured (for physical people) or generated (for avatars), is expected to be a prerequisite for realizing the needed realism, immersion, and QoE. It will facilitate seamless interactions, via fast, low-latency processing enabled by allocating more system resources to handle more significant data to achieve better quality or experience. Considerable latency may lead to an action performed by an avatar representing a real person lagging behind the intended consequence or a virtual object failing to move to a position as expected. A saliency model makes it possible to reduce the latency in the

metaverse by avoiding the transmission of unnecessary data and non-essential computation.

The introduction of full multimedia [66, 125], with visual, hearing, olfactory, haptic and gustatory signals would also differentiate the metaverse from conventional VR and AR. Thus far, visual and speech/audio have been better explored. Emphasis can be placed on olfactory, haptic, and even gustatory (the most difficult) signals. We take olfaction [98] as an example to add to the PC representation for an arbitrary point:

$$\widetilde{\omega}_j = [\omega_j, o_j], j = 1, 2, \ldots, K, \tag{11}$$

where ω_j is defined in Equation (1), and $o_j = [e_j^1, e_j^2, \ldots, e_j^P, f_j, d_j, i_j]$ denotes the olfactory descriptor. Moreover, e_j^p (for $p = 1, 2, \ldots, P$; assuming P olfactory sensors) denotes the normalized value from the p^{th} olfactory sensor:

$$\sum_{p=1}^{P} e_j^p = 1, \tag{12}$$

and f_j, d_j and i_j are the frequency, duration and concentration of odor releasing, respectively.

Apart from the visual and speech/audio, the olfactory, haptic and gustatory signals, and environmental settings (such as temperature and wind) greatly influence the evaluation of PC saliency and quality/QoE. The modeling must formulate interaction in real-time within virtual scenarios. Cross-modal effects must also be considered (e.g., between visual and olfaction [47] and between visual, olfaction, and taste [156]). We take saliency as an example: the overall saliency, $\mathbf{S}$, from n modalities may be fused as follows [159]:

$$\boldsymbol{S} = \sum_{\iota=1}^{n} S(\iota) - \sum_{\iota=1}^{n} \sum_{\rho=\iota+1}^{n} \varepsilon(\iota, \rho)\Psi(S(\iota), S(\rho)), \tag{13}$$

where $S(\iota)$ $(\iota = 1, 2, \ldots, n)$ represents the saliency effect for a modality (i.e., visual, hearing, olfactory, haptic, or gustatory); $\Psi(.,.)$ is a function (usually nonlinear) for evaluating the overlapping effect of the two modalities under consideration; $\varepsilon(.,.)$ is the gain-controlling parameter for overlapping. Equation (13) was previously tested for visual JND fusion [228].

While full multimedia is used and QoE is improved, an effort is required to avoid the possible Uncanny Valley, because this is expected to be a more likely problem for the metaverse.

6 Summary and Concluding Remarks

More PCs and meshes have become available and facilitate rapidly increasing practical applications in physical and virtual worlds due to powerful technology

for 3D data acquisition and big visual data computing. They cover a wide spectrum, have high scalability in terms of the number, size, and complexity of objects, and further enable an ever-expanding scope of utilities. Using the signal saliency and quality to guide the allocation of various limited system resources for the optimal quality of performance for human and machine uses is desirable to meet the high demand in a green and cost-effective manner.

"If you cannot measure it, you cannot improve it" (William Thomson, 1824–1907). Substantial exploration has been internationally conducted in the related research communities, and this paper is devoted to a comprehensive overview of computational models for the saliency and quality of PCs and meshes using handcrafted and learning-based approaches and generic and utility-oriented modeling. Relevant new exploration possibilities in the existing topics and emerging research areas have also been discussed.

Acknowledgement

We are grateful for the research collaboration and constant discussion in the related areas with Wentao Cheng, Baoquan Zhao, Chenlei Lv, Xiaoying Ding, Jingwen Hou and Zhi Hu, and would like to thank Anh-Duc Nguyen, Seongmin Lee, Jeonghaeng Lee, and Jungwoo Huh for some literature survey, discussion, and proofreading. We would also like to thank the three anonymous reviewers for their recognition and helpful comments to our work documented in this paper.

Financial Support

This work was supported by the Ministry of Education, Singapore (W.L., grant number Tier-1 Fund MOE2021, RG14/21) and National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (S.L., grant number 2020R1A2C3011697).

Biographies

Weisi Lin is a Professor and the Associate Chair (Research) in the School of Computer Science and Engineering, Nanyang Technological University, Singapore. His research interests include intelligent image processing, perceptual signal modeling, video compression, and multimedia communication. He is a Chartered Engineer and a fellow of IEEE and IET. He was the Technical Program Chair of the IEEE ICME 2013, PCM 2012, QoMEX 2014, and the IEEE VCIP 2017. He has been a Keynote/Invited/Panelist/Tutorial Speaker at 40+ international conferences and was a Distinguished Lecturer of the IEEE

Circuits and Systems Society from 2016 to 2017 and the AsiaPacific Signal and Information Processing Association (APSIPA) from 2012 to 2013. He has been an Associate Editor of the IEEE Trans. Image Process., the IEEE Trans. Circuits Syst. Video Technol., the IEEE Trans. Multimedia, and the IEEE Signal Process. Lett. He is a Highly Cited Researcher 2019, 2020, 2021 (awarded by Clarivate Analytics)

Sanghoon Lee is a Professor at the EE Department, Yonsei University, Korea. His current research interests include image processing, computer vision, and graphics. He was an Associate Editor of the IEEE Trans. on Image Processing from 2010 to 2014. He served as a Guest Editor for the IEEE Trans. on Image Processing in 2013. He was the General Chair of the 2013 IEEE IVMSP Workshop. He has been serving as the Chair of the IEEE P3333.1 Working Group since 2011. He served as an Associate Editor for the IEEE SPL from 2014 to 2018. He was the IEEE IVMSP/MMSP TC (2014–2019)/(2016–2021) and the IVM TC Chair of APSIPA from 2018 to 2019. He has been serving as a Senior Area Editor of the IEEE SPL and an Associate Editor of IEEE Trans. on Multimedia. He is a Board of Governors member of APSIPA, and also an Editor in Chief of APSIPA News Letters.

References

[1] I. Abouelaziz, A. Chetouani, M. El Hassouni, H. Cherifi, and L. J. Latecki, "Learning Graph Convolutional Network for Blind Mesh Visual Quality Assessment", *IEEE Access*, 9, 2021, 108200–11.

[2] I. Abouelaziz, A. Chetouani, M. El Hassouni, L. J. Latecki, and H. Cherifi, "3D Visual Saliency and Convolutional Neural Network for Blind Mesh Quality Assessment", *Neural Computing and Applications*, 2019, 1–15.

[3] I. Abouelaziz, A. Chetouani, M. El Hassouni, L. J. Latecki, and H. Cherifi, "No-reference Mesh Visual Quality Assessment via Ensemble of Convolutional Neural Networks and Compact Multi-linear Pooling", *Pattern Recognition*, 100, 2020, 107174.

[4] I. Abouelaziz, M. El Hassouni, and H. Cherifi, "A Curvature Based Method for Blind Mesh Visual Quality Assessment Using a General Regression Neural Network", in *2016 12th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*, IEEE, 2016, 793–7.

[5] I. Abouelaziz, M. El Hassouni, and H. Cherifi, "Blind 3d Mesh Visual Quality Assessment Using Support Vector Regression", *Multimedia Tools and Applications*, 77(18), 2018, 24365–86.

[6] S. Agarwal, Y. Furukawa, N. Snavely, I. Simon, B. Curless, S. S. M., and R. Szeliski, "Building Rome in a Day", *Commun. ACM*, 54(10), 2011, 105–12.

[7] R. Akhter, N. M. M. Hassan, J. Aida, S. Takinami, and M. Morita, "Relationship between Betel Quid Additives and Established Periodontitis Among Bangladeshi Subjects", in, Vol. 35, No. 1, 2008, 9–15.

[8] M. Alexa, J. Behr, D. Cohen-Or, S. Fleishman, D. Levin, and C. T. Silva, "Point Set Surfaces", *IEEE Visualization*, 2001, 21–9.

[9] E. Alexiou, "Exploiting User Interactivity in Quality Assessment of Point Cloud Imaging", in *Int'l Conf. Quality of Multimedia Experience (QoMEX)*, 2019.

[10] E. Alexiou and T. Ebrahimi, "Impact of Visualisation Strategy for Subjective Quality Assessment of Point Clouds", in *IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, 2018.

[11] E. Alexiou and T. Ebrahimi, "On the Performance of Metrics to Predict Quality in Point Cloud Representations", in *Applications of Digital Image Processing XL, SPIE Optical Engineering + Applications*, Vol. 10396, 2017.

[12] E. Alexiou and T. Ebrahimi, "Point Cloud Quality Assessment Metric Based on Angular Similarity", in *IEEE Int'l Conf. Multimedia and Expo*, 2018.

[13] E. Alexiou, E. Upenik, and T. Ebrahimi, "Towards Subjective Quality Assessment of Point Cloud Imaging in Augmented Reality", in *IEEE Int'l Workshop on Multimedia Signal Proc. (MMSP)*, Poisson reconstruction, 2017.

[14] A. M. Andrew, "Multiple View Geometry in Computer Vision", *Kybernetes*, 2001.

[15] A. Anis, P. A. Chou, and A. Ortega, "Compression of Dynamic 3D Point Clouds Using Subdivisional Meshes and Graph Wavelet Transforms", in *IEEE Int'l Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, 6360–4.

[16] U. Assarsson and T. Moller, "Optimized View Frustum Culling Algorithms for Bounding Boxes", *Journal of Graphics Tools*, 5(1), 2000, 9–22.

[17] K. Bahirat and B. Prabhakaran, "A Study on LiDAR Data Forensics", in *IEEE Int'l Conf. on Multimedia and Expo.* 2017, 679–84.

[18] X. Bai, Z. Luo, L. Zhou, H. Fu, L. Quan, and C.-L. Tai, "D3Feat: Joint Learning of Dense Detection and Description of 3D Local Features", in *IEEE/CVF Int'l Conf. Computer Vision and Pattern Recognition (CVPR)*, 2020, 6358–66.

[19] M. Bartoň, I. Hanniel, G. Elber, and others., "Precise Hausdorff Distance Computation Between Polygonal Meshes", *Computer Aided Geometric Design*, 27(8), 2010, 580–91.

[20] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded Up Robust Features", *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 3951 LNCS, 2006, 404–17.

[21] J. Bechtold, M. Tatarchenko, V. Fischer, and T. Brox, "Fostering Generalizationin Single-view 3D Reconstruction by Learning a Hierarchy of Local and Globalshape Priors", in *IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2021, 15880–9.

[22] A. Bhattad, A. Dundar, G. Liu, A. Tao, and B. Catanzaro, "View Generalization for Single Image Textured 3D Models", in *IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2021.

[23] A. Borji, M. Cheng, Q. Hou, H. Jiang, and J. Li, "Salient Object Detection: A Survey", *Comp. Visual Media*, 5, 2019, 117–50.

[24] A. Borji, L. Itti, and M. Bennamou, "State-of-the-Art in Visual Attention Modeling", *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(1), 2013, 185–207.

[25] C. Bregler, A. Hertzmann, and H. Biermann, "Recovering Non-rigid 3D Shape from Image Streams", 2000, 690–6.

[26] K. Brunnström, S. A. Beker, K. d. Moor, A. Dooms, S. Egger, and others., "Qualinet White Paper on Definitions of Quality of Experience", ffhal-00977812, 2013.

[27] N. C., "Mark Zuckerberg is Betting Facebook's Future on the Metaverse", The Verge, Archived from the original on Oct. 2021.

[28] S. Cao and N. Snavely, "Minimal Scene Descriptions from Structure from Motion Models", in *IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2014, 461–8.

[29] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, J. Xiao, L. Yi, and F. Yu, "ShapeNet: An Information-rich 3D Model Repository", 2015, arXiv preprint arXiv:1512.03012.

[30] D. M. Chen, G. Baatz, K. Koser, S. Tsai, R. Vedantham, T. Pylvanainen, K. Roimela, X. Chen, J. Bach, M. Pollefeys, B. Girod, and R. Grzeszczuk, "City-scale Landmark Identification on Mobile Devices", in *IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2011, 737–44.

[31] S. Chen, B. Liu, C. Feng, C. Vallespi-Gonzalez, and C. Wellington, "3D Point Cloud Processing and Learning for Autonomous Driving: Impacting Map Creation, Localization, and Perception", *IEEE Signal Processing Magazine*, 38(1), 2021, 68–86.

[32] X. Chen, A. Saparov, B. Pang, and T. A. Funkhouser, "Schelling Points on 3D Surface Meshes", *ACM Trans. Graph.*, 31(4), 2012, 1–29.

[33] Z. Chen, T. Zhang, J. Cao, Y. J. Zhang, and C. Wang, "Point Cloud Resampling Using Centroidal Voronoi Tessellation Methods", *Computer-Aided Design*, 103, 2018, 12–21.

[34] B. Cheng, L. Sheng, S. Shi, M. Yang, and D. Xu, "Back-tracing Representative Points for Voting-Based 3D Object Detection in Point Clouds", in *IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2021, 8963–72.

[35] W. Cheng, K. Chen, W. Lin, M. Goesele, X. Zhang, and Y. Zhang, "A Two-stage Outlier Filtering Framework for City-Scale Localization using 3D SfM Point Clouds", *IEEE Trans. Image Process.*, 28(10), 2019, 4857–69.

[36] W. Cheng, W. Lin, K. Chen, and X. Zhang, "Cascaded Parallel Filtering for Memory Efficient Image-based Localization", *Int'l Conf. Computer Vision (ICCV)*, 2019.

[37] W. Cheng, W. Lin, X. Zhang, M. Goesele, and M.-T. Sun, "A Data-driven Point Cloud Simplification Framework for City-scale Image-based Localization", *IEEE Trans. Image Proc.*, 26(1), 2017, 262–75.

[38] A. Chetouani, M. Quach, G. Valenzise, and F. Dufaux, "Deep Learning-Based Quality Assessment of 3D Point Clouds Without Reference", in *IEEE Int'l Conf. Multimedia and Expo (ICME) Workshops*, 2021.

[39] F. Chiabrando, M. Lo Turco, and F. Rinaudo, "Modeling the Decay in AN Hbim Starting from 3d Point Clouds. A Followed Approach for Cultural Heritage Knowledge, ISPRS - International Archives of the Photogrammetry", *Remote Sensing and Spatial Information Sciences*, 62(5), 2017, 605–12.

[40] K.-W. Chiang, G.-J. Tsai, Y.-H. Li, and N. El-Sheimy, "Development of LiDAR-based UAV System for Environment Reconstruction", *IEEE Geoscience and Remote Sensing Lett.*, 14(10), 2017, 1790–4.

[41] J. Chibane, A. Mir, and G. Pons-Moll, "Neural Unsigned Distance Fields for Implicit Function Learning", 2020, arXiv preprint arXiv:2010.13938.

[42] P. A. Chou, M. Koroteev, and M. Krivokuća, "A Volumetric Approach to Point Cloud Compression—Part I: Attribute Compression", *IEEE Trans. Image Proc.*, 29, 2019, 2203–16.

[43] P. Cignoni, C. Rocchini, and R. Scopigno, "Metro: Measuring Error on Simplified Surfaces", in *Computer Graphics Forum*, Vol. 17, Wiley Online Library, 1998, 167–74.

[44] R. Cong, J. Lei, H. Fu, M. Cheng, W. Lin, and Q. Huang, "Review of Visual Saliency Detection with Comprehensive Information", *IEEE Trans. on Circuits and Syst. Video Technol.*, 29(10), 2019, 2941–59.

[45] M. Corsini, E. D. Gelasca, T. Ebrahimi, and M. Barni, "Watermarked 3-D Mesh Quality Assessment", *IEEE Transactions on Multimedia*, 9(2), 2007, 247–56.

[46] B. Curless and M. Levoy, "A Volumetric Method for Building Complex Models from Range Images", in *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques*, 1996, 303–12.

[47] M. L. Dematte, D. Sanabria, and C. Spence, "Olfactory Discrimination: When Vision Matters?", *Chem. Senses*, 34(2), 2009, 103–9.

[48] Y. Deng, C. C. Loy, and X. Tang, "Image Aesthetic Assessment: An Experimental Survey", *IEEE Signal Processing Magazine*, 34(4), 2017, 80–106.

[49] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superpoint: Self-Supervised Interest Point Detection and Description", in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018, 224–36.

[50] J. Digne, J. M. Morel, and C. M. Souzani, "Scale Space Meshing of Raw Data Point Sets", *Computer Graphics Forum*, 30(6), 2013, 1630–42.

[51] X. Ding, W. Lin, and Z. Chen, "Point Cloud Saliency Detection by Local and Global Feature Fusion", *IEEE Transactions on Image Proc.*, 28(11), 2019, 5379–93.

[52] R. Diniz, P. G. Freitas, and M. Farias, "Local Luminance Patterns for Point Cloud Quality Assessment", in *Int'l Workshop Multimedia Signal Processing*, 2020.

[53] R. Diniz, P. G. Freitas, and M. Farias, "Towards a Point Cloud Quality Assessment Model using Local Binary Patterns", in *Int'l Conf. Quality of Multimedia Experience*, 2020.

[54] R. Diniz, P. G. Freitas, and M. C. Q. Farias, "Color and Geometry Texture Descriptors for Point-Cloud Quality Assessment", *IEEE Signal Process. Lett.*, 28, 2021, 1150–4.

[55] D. Dolonius, E. Sintorn, V. Kämpe, and U. Assarsson, "Compressing Color Data for Voxelized Surface Geometry", *IEEE Trans. Visualization and Computer Graphics*, 25(2), 2017, 1270–82.

[56] L. Dong, Y. Fang, W. Lin, and H. S. Seah, "Perceptual Quality Assessment for 3D Triangular Mesh based on Curvature", *IEEE Trans. Multimedia*, 17(12), 2015, 2174–84.

[57] L. Dong, W. Lin, C. Zhu, and H. S. Seah, "Selective Rendering with Graphical Saliency Model", *10-th IEEE IVMSP Workshop on Perception and Visual Signal Analysis*, 2011.

[58] M. Dou, P. Davidson, S. R. Fanello, S. Khamis, A. Kowdle, C. Rhemann, and others., "Motion2fusion: Real-time Volumetric Performance Capture", *ACM Transactions on Graphics (TOG)*, 36(6), 2017, 1–16.

[59] M. Dou, S. Khamis, Y. Degtyarev, P. Davidson, S. R. Fanello, A. Kowdle, and others., "Fusion4d: Real-time Performance Capture of Challenging Scenes", *ACM Transactions on Graphics (ToG)*, 35(4), 2016, 1–13.

[60] E. Dumic, C. R. Duarte, and d. S. C. A. Luis, "Subjective Evaluation and Objective Measures for Point Clouds – State of the Art", in *First Int'l Colloquium on Smart Grid Metrology (SmaGriMet), Split, Croatia*, 2018.

[61] S. Eckelmann, T. Trautmann, H. Ußler, B. Reichelt, and O. Michler, "V2V-communication, LiDAR System and Positioning Sensors for Future Fusion Algorithms in Connected Vehicles", *Transportation research procedia*, 27, 2017, 69–76.

[62] P. Eskicioglu A.M.and Fisher, "Image Quality Measures and Their Performance", *IEEE Trans. Communications*, 43(12), 1995, 2959–65.

[63] M. A. Fischler, "Random Sampling Concensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography", *Commun. ACM*, 24(6), 1981, 381–95.

[64] M. A. Fischler and R. C. Bolles, "Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography", *Communications of the ACM*, 24(6), 1981, 381–95.

[65] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision Meets Robotics: The KITTI Dataset", *Int'l J. Robotics Research*, 32(11), 2013, 1231–7.

[66] G. Ghinea, C. Timmerer, W. Lin, and S. Gulliver, "Mulsemedia: State-of-the- Art, Perspectives and Challenges", *ACM Transactions on Multimedia Computing Communications and Applications*, 11(1s), 2014, Article 17.

[67] R. Grossmann, N. Kiryati, and R. Kimmel, "Computational Surface Flattening: A Voxel-based Approach", *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(4), 2002, 433–41.

[68] 3. Group., "Common Test Conditions for Point Cloud Compression", in *ISO/IEC JTC1/SC29/WG11, N18474, Geneva, Switzerland*, Mar. 2019.

[69] S. Gu, J. Hou, H. Zeng, H. Yuan, and K. Ma, "3D Point Cloud Attribute Compression Using Geometry-Guided Sparse Representation", *IEEE Trans. Image Proc.*, 29, 2019, 796–808.

[70] J. Guo, V. Vidal, I. Cheng, A. Basu, A. Baskurt, and G. Lavoue, "Subjective and Objective Visual Quality Assessment of Textured 3D Meshes", *ACM Transactions on Applied Perception (TAP)*, 14(2), 2016, 1–20.

[71] Y. Guo, M. Bennamoun, F. Sohel, M. Lu, J. Wan, and N. Kwok, "A Comprehensive Performance Evaluation of 3D Local Feature Descriptors", *Int'l J. of Computer Vision*, 116, 2016, 66–89.

[72] Y. Guo, H. Wang, Q. Hu, H. Liu, L. Liu, and M. Bennamou, "Deep Learning for 3D Point Clouds: A Survey", *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(12), 2021, 4338–64.

[73] P. Henry, D. Fox, A. Bhowmik, and R. Mongia, "Patch Volumes: Segmentation-based Consistent Mapping with RGB-D Cameras", in *Int'l Conf. 3D Vision-3DV*, 2013, 398–405.

[74] K. Hildebrandt, K. Polthier, and M. Wardetzky, "On the Convergence of Metric and Geometric Properties of Polyhedral Surface", *Geometriae Dedicata*, 123(1), 2006, 89–112.

[75] R. Horaud, M. Hansard, G. Evangelidis, and C. Menier, "An Overview of Depth Cameras and Range Scanners Based on Time-of-Flight Technologies", *Machine Vision and Applications*, 27, 2016, 1005–20.

[76] V. Hosu, D. Saupe, B. Goldluecke, W. Lin, W.-H. Cheng, J. See, and L.-K. Wong, "From Technical to Aesthetics Quality Assessment and Beyond: Challenges and Potential, Joint Workshop on Aesthetic and Technical Quality Assessment of Multimedia and Media Analytics for Societal Trends", *ACM Multimedia*, 202.

[77] J. Hou, S. Yang, and W. Lin, "Object-level Attention for Aesthetic Rating Distribution Prediction", *ACM Multimedia*, 2020.

[78] J. Hou, B. Zhao, N. Ansari, and W. Lin, "Range Image Based Point Cloud Colorization Using Conditional Generative Model", in *IEEE Int'l Conf. Image Proc. (ICIP)*, 2019.

[79] F. Huang, C. Wen, H. Luo, M. Cheng, C. Wang, and J. Li, "Local Quality Assessment of Point Clouds for Indoor Mobile Mapping", *Neurocomputing*, 196, 2016, 59–69.

[80] L. Huang, S. Wang, K. Wong, J. Liu, and R. Urtasun, "Octsqueeze: Octree-Structured Entropy Model for LiDAR Compression", in *IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 2020, 1313–23.

[81] X. Huang, L. Fan, Q. Wu, J. Zhang, and C. Yuan, "Fast Registration for Cross-source Point Clouds by Using Weak Regional Affinity and Pixel-wise Refinement", in *IEEE Int'l Conf. Multimedia and Expo (ICME)*, 2019, 1552–7.

[82] M. Innmann M.; Zollhöfer, M. Nießner, C. Theobalt, and M. Stamminger, "Volumedeform: Real-time Volumetric Non-rigid Reconstruction", in *European Conference on Computer Vision*, 2016, 362–79.

[83] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3. 6m: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7), 2013, 1325–39.

[84] L. Itti, C. Koch, and E. Niebur, "A Model of Saliency-based Visual Attention for Rapid Scene Analysis", *IEEE Trans. Pattern Anal. and Mach. Intell.*, 20(11), 1998, 1254–9.

[85] S. Izadi, D. Kim, O. Hilliges, D. Molyneaux, R. Newcombe, P. Kohli, and others., "KinectFusion: Real-time 3D Reconstruction and Interaction Using a Moving Depth Camera", in *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology*, 2011, 559–68.

[86] M. Jang, S. Lee, J. Kang, and S. Lee, "Active Stereo Matching Benchmark for 3D Reconstruction Using Multi-view Depths", in *2021 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*, IEEE, 2021, 215–20.

[87] A. Javaheri, C. Brites, F. Pereira, and J. Ascenso, "A Generalized Hausdorff Distance-based Quality Metric for Point Cloud Geometry", in *Int'l Conf. Quality of Multimedia Experience*, 2020.

[88] A. Javaheri, C. Brites, F. Pereira, and J. Ascenso, "Mahalanobis Based Point to Distribution Metric for Point Cloud Geometry Quality Evaluation", *IEEE Signal Processing Letters*, 2020, 1350–4.

[89] S. Jeong and J.-Y. Sim, "Saliency Detection for 3D Surface Geometryusing Semi-regular Meshes", *IEEE Trans. Multimedia*, 19(12), 2017, 2692–705.

[90] M. Jiang, Y. Wu, T. Zhao, Z. Zhao, and C. Lu, "PointSIFT: A SIFT-like Network Module for 3D Point Cloud Semantic Segmentation", 2018, arXiv:1807.00652v2 [cs.CV].

[91] J. Jin, X. Zhang, X. Fu, H. Zhang, W. Lin, J. Lou, and Y. Zhao, "Just Noticeable Difference for Deep Machine Vision", *IEEE Trans. Circuits and Systems for Video Technology*, 2021, in press.

[92] A. E. Johnson and M. Hebert, "Using Spin Images for Efficient Object Recognition in Cluttered 3D Scene", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(5), 1999, 433–49.

[93] J. Kang, S. Lee, M. Jang, and S. Lee, "Gradient Flow Evolution for 3D Fusion from a Single Depth Sensor", *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

[94] J. Kang, S. Lee, M. Jang, H. Yoon, and S. Lee, "WarpingFusion: Accurate Multi-view TSDF Fusion with Local Perspective Warp", in *2021 IEEE International Conference on Image Processing (ICIP)*, 2021, 1564–8.

[95] J. Kang, S. Lee, and S. Lee, "Competitive Learning of Facial Fitting and Synthesis Using UV Energy", *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2021.

[96] Z. Karni and C. Gotsman, "Spectral Compression of Mesh Geometry", in *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques*, 2000, 279–86.

[97] M. Kazhdan and H. Hoppe, "Screened Poisson Surface Reconstruction", *ACM Transactions on Graphics*, 32(3), 2013, 1–13.

[98] A. Keller, R. Gerkin, Y. Guan, A. Dhurandhar, G. Turu, B. Szalai, J. D. Mainland, Y. Ihara, C. W. Yu, R. Wolfinger, and others., "Predicting Human Olfactory Perception from Chemical Features of Odor Molecules", *Science*, 355(6327), 2017, 820–6.

[99] L. Keselman, A. Iselin Woodfill J. Grunnet-Jepsen, and . Bhowmik, "Intel Realsense Stereoscopic Depth Cameras", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017.

[100] H. Kim, S. Ahn, W. Kim, and S. Lee, "Visual Preference Assessment on Ultra-high-definition Images", *IEEE Transactions on Broadcasting*, 62(4), 2016, 757–69.

[101] H. Kim, J. Kim, T. Oh, and S. Lee, "Blind Sharpness Prediction for Ultrahigh-definition Video Based on Human Visual Resolution", *IEEE Transactions on Circuits and Systems for Video Technology*, 27(5), 2016, 951–64.

[102] H. Kim and S. Lee, "Transition of Visual Attention Assessment in Stereoscopic Images with Evaluation of Subjective Visual Quality and Discomfort", *IEEE Transactions on Multimedia*, 17(12), 2015, 2198–209.

[103] J. Kim, B.-S. Hua, D. T. Nguyen, and S.-K. Yeung, "Minimal Adversarial Examples for Deep Learning on 3D Point Clouds", in *Int'l Conf. Computer Vision (ICCV)*, 2021.

[104] W. Kim, S. Ahn, A.-D. Nguyen, J. Kim, J. Kim, H. Oh, and S. Lee, "Modern Trends on Quality of Experience Assessment and Future Work", in *APSIPA Transactions on Signal and Information Processing*, Vol. 8, 2019.

[105] W. Kim, J. Kim, S. Ahn, J. Kim, and S. Lee, "Deep Video Quality Assessor: From Spatio-temporal Visual Sensitivity to a Convolutional Neural Aggregation Network", in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, 219–34.

[106] J. Knopp, M. Prasad, G. Willems, R. Timofte, and L. Van Gool, "Hough Transform and 3D SURF for Robust Three Dimensional Classification", *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 6316 LNCS(PART 6), 2010, 589–602.

[107] M. Krivokuća, P. A. Chou, and M. Koroteev, "A Volumetric Approach to Point Cloud Compression–Part II: Geometry Compression", *IEEE Trans. Image Proc.*, 29, 2019, 2217–29.

[108] P. Krusi, P. Furgale, M. Bosse, and R. Siegwart, "Driving on Point Clouds: Motion Planning, Trajectory Optimization, and Terrain Assessment in Generic Nonplanar Environments", *J. of Field Robotics*, 34(5), 2017, 940–84.

[109] B. Kwon, J. Huh, K. Lee, and S. Lee, "Optimal Camera Point Selection Toward the Most Preferable View of 3-D Human Pose", *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 52(1), 2022, 533–53.

[110] L. Landrieu and M. Simonovsky, "Large-Scale Point Cloud Semantic Segmentation With Superpoint Graphs", in *IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2018, 4558–67.

[111] G. Lavoué, "A Multiscale Metric for 3D Mesh Visual Quality Assessment", *Computer Graphics Forum*, 30, 2011, 1427–37, Wiley Online Library.

[112] G. Lavoué, E. D. Gelasca, F. Dupont, A. Baskurt, and T. Ebrahimi, "Perceptually Driven 3D Distance Metrics with Application to Watermarking", in *Applications of Digital Image Processing XXIX*, Vol. 6312, International Society for Optics and Photonics, 2006, 63120L.

[113] G. Lavoué, M. C. Larabi, and L. Váša, "On the Efficiency of Image Metrics for Evaluating the Visual Quality of 3D Models", *IEEE Transactions on Visualization and Computer Graphics*, 22(8), 2015, 1987–99.

[114] E. A. Leal Narvaez, G. Sanchez Torres, and J. W. Branch, "Bedoya: Point Cloud Saliency Detection via Local Sparse Coding", *DYNA*, 86(209), 2019, 238–47.

[115] C. H. Lee, A. Varshney, and D. W. Jacobs, "Mesh Saliency", *ACM Transactions on Graphics*, 24(3), 2005, 659–66.

[116] K. Lee, W. Kim, and S. Lee, "From Human Pose Similarity Metric to 3D Human Pose Estimator: Temporal Propagating LSTM Networks", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[117] L. Lee, T. Braud, P. Zhou, and others., "All One Needs to Know About Metaverse: A Complete Survey on Technological Singularity, Virtual Ecosystem, and Research Agenda", 2021, arXiv preprint arXiv:2110.05352.

[118] S. Lee, M. S. Pattichis, and A. C. Bovik, "Foveated Video Compression with Optimal Rate Control", *IEEE Transactions on Image Processing*, 10(7), 2001, 977–92.

[119] S. Lee, M. S. Pattichis, and A. C. Bovik, "Foveated Video Quality Assessment", *IEEE Transactions on Multimedia*, 4(1), 2002, 129–32.

[120] G. Leifman, E. Shtrom, and A. Tal, "Surface Regions of Interest for View-point Selection", in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2012, 414–21.

[121] J. Li and G. H. Lee, "USIP: Unsupervised Stable Interest Point Detection From 3D Point Clouds", in *IEEE Int'l Conf. Computer Vision (ICCV)*, 2019.

[122] L. Li, H. Zhu, S. Zhao, G. Ding, and W. Lin, "Personality-assisted Multi-task Learning for Generic and Personalized Image Aesthetics Assessment", *IEEE Transactions on Image Processing*, 29(2), 2020, 3898–910.

[123] J. H. Lim, S. Im, and Y. S. Cho, "MLS-based Finite Elements for Three-dimensional Nonmatching Meshes and Adaptive Mesh Refinement", *Computer Methods in Applied Mechanics and Engineering*, 196(17–20), 2007, 2216–28.

[124] M. Limper, A. Kuijper, and D. W. Fellner, "Mesh Saliency Analysis via Local Curvature Entropy", in *European Association for Computer Graphics - 37th Annual Conference, EUROGRAPHICS 2016 - Short Papers*, 2016, 13–6.

[125] W. Lin and G. Ghinea, "Progress and Opportunities in Modelling Just-Noticeable Difference (JND) for Multimedia", *IEEE Trans. Multimedia*, 2021, in press.

[126] W. Lin and C.-C. J. Kuo, "Perceptual Visual Quality Metrics: A Survey", *J. Visual Commun. Image Represent.*, 22(4), 2011, 297–312.

[127] Y. Lin, M. Yu, K. Chen, G. Jiang, F. Chen, and Z. Peng, "Blind Mesh Assessment Based on Graph Spectral Entropy and Spatial Features", *Entropy*, 22(2), 2020, 190.

[128] Y. Lin, M. Yu, K. Chen, G. Jiang, Z. Peng, and F. Chen, "Blind Mesh Quality Assessment Method Based on Concave, Convex and Structural Features Analyses", in *2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, IEEE, 2019, 282–7.

[129] I. Lissner, J. Preiss, P. Urban, M. S. Lichtenauer, and P. Zolliker, "Image Difference Prediction: From Grayscale to Color", *IEEE Trans. on Image Process*, 22(2), 2013, 435–46.

[130] H. Liu, J. Jia, and N. Z. Gong, "PointGuard: Provably Robust 3D Point Cloud Classification", in *IEEE/CVF Int'l Conf. Computer Vision and Pattern Recognition (CVPR)*, 2021, 6186–95.

[131] H. Liu, H. Yuan, Q. Liu, J. Hou, and J. Liu, "A Comprehensive Study and Comparison of Core Technologies for MPEG 3-D Point Cloud Compression", *IEEE Trans. Broadcasting*, 66(3), 2019, 701–17.

[132] 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", *IEEE Trans. Circuits Syst. Video Technol.*, 31(12), 2021, 4645–60.

[133] Y. Liu, Q. Yang, Y. Xu, and L. Yang, "Point Cloud Quality Assessment: Dataset Construction and Learning-based No-Reference Approach", 2020, arXiv:2012.11895 [eess.IV].

[134] S. Logozzo, G. Franceschini, A. Kilpela, M. Caponi, L. Governiy, and L. Blois, "A Comparative Analysis of Intraoral 3D Digital Scanners for Restorative Dentistry", *The Internet Journal of Medical Technology*, 5(1), 2011, 1–18.

[135] W. E. Lorensen and H. E. Cline, "Marching Cubes: A High Resolution 3D Surface Construction Algorithm", *ACM siggraph computer graphics*, 21(4), 1987, 163–9.

[136] D. G. Lowe, "Object Recognition from Local Scale-Invariant Features", in *Int'l Conf. Computer Vision (ICCV)*, 1999, 1150–7.

[137] F. Lu, G. Chen, Y. Liu, L. Zhang, S. Qu, S. Liu, and R. Gu, "HRegNet: A Hierarchical Network for Large-scale Outdoor LiDAR Point Cloud Registration", in *Int'l Conf. Computer Vision (ICCV)*, 2021.

[138] P. Lu, G. Chen, Y. Liu, and A. C. Qu Z. anbd Knoll, "RSKDD-Net: Random Sample-based Keypoint Detector and Descriptor", in *Conf. and Workshop Neural Information Processing Systems (NeurIPS)*, 2020.

[139] Z. Lu, W. Lin, X. Yang, E. Ong, and S. Yao, "Modeling Visual Attention's Modulatory Aftereffects on Visual Sensitivity and Quality Evaluation", *IEEE Trans. Image Processing*, 14(11), 2005, 1928–42.

[140] C. Lv, W. Lin, and B. Zhao, "Approximate Intrinsic Voxel Structure for Point Cloud Simplification", *IEEE Trans. Image Proc.*, 30(9), 2021, 7241–55.

[141] C. Lv, W. Lin, and B. Zhao, "Intrinsic and Isotropic Resampling for 3D Point Clouds", *IEEE Trans on Pattern Analysis and Machine Intelligence*, 2022.

[142] C. Lv, W. Lin, and B. Zhao, "Voxel Structure-based Mesh Reconstruction from a 3D Point Cloud", *IEEE Trans. Multimedia*, 2021.

[143] H. Macher, T. Landes, and P. Grussenmeyer, "From Point Clouds to Building Information Models: 3D Semi-Automatic Reconstruction of Indoors of Existing Buildings", *Applied Sciences*, 7(10), 2017, 1030.

[144] W. Maddern, G. Pascoe, C. Linegar, and P. Newman, "1 Year, 1000km: The Oxford RobotCar Dataset", *Int'l J. Robotics Research (IJRR)*, 36(1), 2017, 3–15.

[145] A. Maglo, G. Lavoué, F. Dupont, and C. Hudelot, "Technologies for 3D Mesh Compression: A Survey", *ACM Computing Surveys*, 47(3), 2015, Article No.: 44.

[146] K. Mammou, P. A. Chou, D. Flynn, M. Krivokuca, O. Nakagami, and T. Sugio, "G-PCC Codec Description v2", in *ISO/IEC JTC1/SC29/WG11, MPEG, N18189, Marrakech*, Jan. 2019.

[147] R. Mantiuk, K. Myszkowski, and S. Pattanaik, "Attention Guided MPEG Compression for Computer Animations", *Proceedings of the 18th spring conference on Computer graphics - SCCG '03*, 2003.

[148] R. Mekuria, K. Blom, and P. Cesar, "Design, Implementation, and Evaluation of a Point Cloud Codec for Tele-Immersive Video", *IEEE Trans. on Circuits and Syst. Video Technol.*, 27(4), 2017, 828–42.

[149] R. N. Mekuria, Z. Li, C. Tulvan, and P. Chou, "Evaluation Criteria for PCC (Point Cloud Compression)", in *ISO/IEC JTC1/SC29/WG11/ N16332, Geneva, Switzerland*, Jun. 2016.

[150] A. Mertan, D. Duff, and G. Unal, "Single Image Depth Estimation: An Overview", 2021, arXiv:2104.06456 [cs.CV].

[151] L. Mescheder, M. Oechsle, M. Niemeyer, S. Nowozin, and A. Geiger, "Occupancy Networks: Learning 3D Reconstruction in Function Space", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, 4460–70.

[152] G. Meynet, J. Digne, and G. Lavoué, "PC-MSDM: A Quality Metric for 3D Point Clouds", in *Int'l Conf. Quality of Multimedia Experience (QoMEX)*, 2019.

[153] G. Meynet, Y. Nehmé, J. Digne, and G. Lavoué, "PCQM: A Full Reference Quality Metric for Colored 3D Point Clouds", in, 2020.

[154] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing Scenes as Neural Radiance Fields for View Synthesis", in *European Conference on Computer Vision*, Springer, 2020, 405–21.

[155] M. Mohammadi, M. Rashidi, V. Mousavi, A. Karami, Y. Yu, and B. Samali, "Quality Evaluation of Digital Twins Generated Based on UAV Photogrammetry and TLS: Bridge Case Study", *Remote Sensing*, 13(17), 2021, 3499.

[156] T. Narumi, T. Kajinami, S. Nishizaka, T. Tanikawa, and M. Hirose, "Pseudo-Gustatory Display System Based on Cross-Modal Integration of Vision, Olfaction and Gustation", in *IEEE Virtual Reality Conf.* 2011, 127–30.

[157] R. A. Newcombe, D. Fox, and S. M. Seitz, "Dynamicfusion: Reconstruction and Tracking of Non-rigid Scenes in Real-time", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, 343–52.

[158] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, and others., "Kinectfusion: Real-Time Dense Surface Mapping and Tracking", in *2011 10th IEEE International Symposium on Mixed and Augmented Reality*, 2011, 127–36.

[159] H. Nothdurft, "Salience from Feature Contrast: Additivity Across Dimensions", *Vision Research*, 40, 2000, 1183–201.

[160] J. P. O'Shea, M. S. Banks, and M. Agrawala, "The Assumed Light Direction for Perceiving Shape from Shading", in *Proceedings of the 5th Symposium on Applied Perception in Graphics and Visualization*, 2008, 135–42.

[161] H. Oh, S. Ahn, S. Lee, and A. C. Bovik, "Deep Visual Discomfort Predictor for Stereoscopic 3D Images", *IEEE Transactions on Image Processing*, 27(11), 2018, 5420–32.

[162] H. Oh and S. Lee, "Visual Presence: Viewing Geometry Visual Information of UHD S3D Entertainment", *IEEE Transactions on image processing*, 25(7), 2016, 3358–71.

[163] R. Pagés, K. Amplianitis, D. Monaghan, J. Ondřej, and A. Smolić, "Affordable Content Creation for Free-Viewpoint Video and VR/AR Applications", *Journal of VCIP*, 53, 2018, 192–201.

[164] Y. Pan, I. Cheng, and A. Basu, "Quality Metric for Approximating Subjective Evaluation of 3-D Objects", *IEEE Transactions on Multimedia*, 7(2), 2005, 269–79.

[165] G. Pandey, J. R. McBride, and R. M. Eustice, "Ford Campus Vision and LiDAR Data Set", *Int'l Journal Robotics Research*, 30(13), 2011, 1543–52.

[166] H. S. Park, T. Shiratori, I. Matthews, and Y. Sheikh, "3D Reconstruction of a Moving Point from a Series of 2D Projections", in *European Conference on Computer Vision*, Springer, 2010, 158–71.

[167] J. Park, H. Oh, S. Lee, and A. C. Bovik, "3D Visual Discomfort Predictor: Analysis of Disparity and Neural Activity Statistics", *IEEE Transactions on Image Processing*, 24(3), 2014, 1101–14.

[168] J. Park, K. Seshadrinathan, S. Lee, and A. C. Bovik, "Video Quality Pooling Adaptive to Perceptual Distortion Severity", *IEEE Transactions on Image Processing*, 22(2), 2012, 610–20.

[169] J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove, "Deepsdf: Learning Continuous Signed Distance Functions for Shape Representation", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, 165–74.

[170] J. Peng, C.-S. Kim, and C.-C. J. Kuo, "Technologies for 3D Mesh Compression: A Survey", *J. Visual Commun. Image Represent.*, 16, 2005, 688–733.

[171] S. Perry, A. Pinheiro, E. Dumic, and L. A. da Silva Cruz, "Study of Subjective and Objective Quality Evaluation of 3D Point Cloud Data by the JPEG Committee", in *Electronic Imaging, Image Quality and System Performance XVI*, 2019, 312-1–312-7.

[172] S. M. Prakhya, J. Lin, V. Chandrasekhar, W. Lin, and B. Liu, "3DHoPD: A Fast Low Dimensional 3D Descriptor", *IEEE Robotics and Automation Letters*, 2(3), 2017, 1472–9.

[173] S. M. Prakhya, B. Liu, and W. Lin, "B-SHOT: A Binary Feature Descriptor for Fast and Efficient Keypoint Matching on 3D Point Clouds", in *IEEE/RSJ Int'l Conf. Intelligent Robots and Systems (IROS)*, 2015.

[174] S. M. Prakhya, B. Liu, and W. Lin, "Detecting Keypoint Sets on 3D Point Clouds via Histogram of Normal Orientations", *Pattern Recognition Letters*, 83(Part 1), 2016, 42–8.

[175] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space", in *Conf. and Workshop Neural Information Processing Systems (NIPS)*, 2017.

[176] Z. Que, L. Guo, and D. Xu, "VoxelContext-Net: An Octree Based Framework for Point Cloud Compression", in *IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 2021, 6042–51.

[177] R. L. de Queiroz and P. A. Chou, "Compression of 3D Point Clouds Using a Region-Adaptive Hierarchical Transform", *IEEE Trans. Image Proc.*, 25(8), 2016, 3497–956.

[178] Y. Rao, B. Fan, Q. Wang, J. Pu, X. Luo, and R. Jin, "Extreme Feature Regions Detection and Accurate Quality Assessment for Point-cloud 3D Reconstruction", *IEEE Access*, 2019, 37757–69.

[179] C. Reymann and S. Lacroix, "Improving LiDAR Point Cloud Classification Using Intensities and Multiple Echoes", in *IEEE/RSJ Int'l Conf. Intelligent Robots and Systems*, 2015, 5122–8.

[180] B. E. Rogowitz and H. E. Rushmeier, "Are Image Quality Metrics Adequate to Evaluate the Quality of Geometric Objects?", in *Human Vision and Electronic Imaging VI*, 4299.

[181] J. Ruiz-Sarmiento, C. Galindo, and J. Gonzalez, "Improving Human Face Detection Through ToF Cameras for Ambient Intelligence Applications", *Ambient Intelligence-Software and Applications*, 2011, 125–32.

[182] R. Rusu, N. Blodow, and M. Beetz, "Fast Point Feature Histograms (FPFH) for 3D Registration", in *IEEE Int'l Conf. Robotics and Automation (ICRA)*, 2009, 3212–7.

[183] R. B. Rusu, N. Blodow, Z. C. Marton, and M. Beetz, "Aligning Point Cloud Views Using Persistent Feature Histograms", in *IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS*, 2008, 3384–91.

[184] S. Salti, F. Tombari, and L. Di Stefano, "SHOT: Unique Signaturesof Histograms for Surface and Texture Description", *Computer Vision and Image Understanding*, 125, 2014, 251–64.

[185] S. Salti, F. Tombari, R. Spezialetti, and L. Di Stefano, "Learning a Descriptor-Specific 3D Keypoint Detector", in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, 2318–26.

[186] P. A. Samuelson and W. F. Samuelson, *Economics*, New York: McGraw-Hill, 1980.

[187] T. Sattler, W. Maddern, C. Toft, A. Torii, L. Hammarstrand, E. Stenborg, D. Safari, M. Okutomi, M. Pollefeys, J. Sivic, F. Kahl, and T. Pajdla, "Benchmarking 6DOF Outdoor Visual Localization in Changing Conditions", in *IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018.

[188] T. Sattler, T. Weyand, B. Leibe, and L. Kobbelt, "Image Retrieval Forimage-based Localization Revisited", in *British Machine Vision Conference (BMVC)*, 2012.

[189] G. Schindler, M. Brown, and R. Szeliski, "City-Scale Location Recognition", in *IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2007, 1–7.

[190] W. J. Schroeder, J. A. Zarge, and W. E. Lorensen, "Decimation of Triangle Meshes", in *Proceedings of the 19th Annual Conference on Computer Graphics and Interactive Techniques*, 1992, 65–70.

[191] P. Scovanner, S. Ali, and M. Shah, "A 3-Dimensional Sift Descriptor and its Application to Action Recognition", in *ACM Int'l Conf. Multimedia (ACM-MM)*, 2007, 357–60.

[192] Y. Shao, Q. Zhang, G. Li, Z. Li, and L. Li, "Hybrid Point Cloud Attribute Compression Using Slice-based Layered Structure and Block-based Intra Prediction", in *ACM Int'l Conf. on Multimedia*, 2018, 1199–207.

[193] R. Shi, Z. Xue, Y. You, and C. Lu, "Skeleton Merger: An Unsupervised Aligned Keypoint Detector", in *IEEE/CVF Int'l Conf. Computer Vision and Pattern Recognition (CVPR)*, 2021, 43–52.

[194] E. Shtrom, G. Leifman, and A. Tal, "Saliency Detection in Large Point Sets", in *Int'l Conf. Computer Vision (ICCV)*, 2013, 3591–8.

[195] S. Silva, B. S. Santos, C. Ferreira, and J. Madeira, "A Perceptual Data Repository for Polygonal Meshes", in *2009 Second International Conference in Visualisation*, IEEE, 2009, 207–12.

[196] L. A. da Silva Cruz, E. Dumić, E. Alexiou, J. Prazeres, R. Duarte, M. Pereira, A. Pinheir, and T. Ebrahimi, "Point Cloud Quality Evaluation: Towards a Definition for Test Conditions", in *Int'l Conf. Quality of Multimedia Experience (QoMEX)*, 2019.

[197] I. Sipiran and B. Bustos, "Harris 3D: A Robust Extension of the Harris Operator for Interest Point Detection on 3D Meshes", *The Visual Computer*, 27(11), 2011, 963.

[198] R. Song, Y. Liu, R. Martin, and P. Rosin, "Mesh Saliency via Spectral Processing", *ACM Trans. Graph.*, 33(1), 2014.

[199] R. Song, Y. Liu, R. R. Martin, and K. R. Echavarria, "Local-to-Global Mesh Saliency", *Visual Computer*, 34(3), 2018, 323–36.

[200] R. Song, W. Zhang, Y. Zhao, Y. Liu, and P. L. Rosin, "Mesh Saliency: An Independent Perceptual Measure or A Derivative of Image Saliency?", in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, 8849–58.

[201] S. Song, S. P. Lichtenberg, and J. Xiao, "Sun RGB-D: A RGB-D Scene Understanding Benchmark Suite", in *IEEE Conf. Computer Vision and Pattern Recognition*, 2015.

[202] B. Steder, R. Rusu, K. Konolige, and W. Burgard, "Point Feature Extraction on 3D Range Scans Taking into Account Object Boundaries", in *IEEE International Conference on Robotics and Automation (ICRA)*, 2011, 2601–8.

[203] H. Su, Z. Duanmu, W. Liu, Q. Liu, and Z. Wang, "Perceptual Quality Assessment of 3d Point Clouds", in *IEEE Int'l Conf. Image Proc, (ICIP)*, 2019.

[204] G. K. L. Tam, Z.-Q. Cheng, Y.-K. Lai, F. C. Langbein, Y. Liu, D. Marshall, R. R. Martin, X.-F. Sun, and P. L. Rosin, "Registration of 3D Point Clouds and Meshes: A Survey from Rigid to Nonrigid", *IEEE Trans. Visualization and Computer Graphics*, 19(7), 2013, 1199–217.

[205] F. P. Tasse and J. Kosinka, "Dodgson, Cluster-Based Point Set Saliency", in *Int'l Conf. Computer Vision (ICCV)*, 2015, 163–71.

[206] D. Tian, H. Ochimizu, C. Feng, R. Cohen, and A. Vetro, "Geometric Distortion Metrics for Point Cloud Compression", in *IEEE Int'l Conf. Image Proc. (ICIP)*, 2017, 3460–4.

[207] F. Tombari, S. Salti, and L. Di Stefano, "Unique Signatures of Histograms for Local Surface Description", *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 6313 LNCS(PART 3), 2010, 356–69.

[208] 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, SPIE Optical Engineering + Applications*, Vol. 10752, 2018.

[209] B. Triggs, P. F. McLauchlan, R. I. Hartley, and A. W. Fitzgibbon, "Bundle Adjustment—A Modern Synthesis", in *International workshop on vision algorithms*, Springer, Berlin, Heidelberg, 1999, 298–372.

[210] L. Váša and J. Rus, "Dihedral Angle Mesh Error: A Fast Perception Correlated Distortion Measure for Fixed Connectivity Triangle Meshes", *Computer Graphics Forum*, 31, 2012, 1715–24, Wiley Online Library.

[211] S. Vijayanarasimhan, S. Ricco, C. Schmid, R. Sukthankar, and K. Fragkiadaki, "Sfm-net: Learning of Structure and Motion from Video", 2017, arXiv preprint arXiv:1704.07804.

[212] I. Viola and P. Cesar, "A Reduced Reference Metric for Visual Quality Evaluation of Point Cloud Contents", *IEEE Signal Processing Letters*, 27, 2020, 1660–4.

[213] I. Viola, S. Subramanyam, and P. Cesar, "A Color-based Objective Quality Metric for Point Cloud Contents", in *Int'l Conf. Quality of Multimedia Experience*, 2020.

[214] J. Wang, H. Zhu, H. Liu, and Z. Ma, "Lossy Point Cloud Geometry Compression Via End-to-End Learning", *IEEE Trans. on Circuits and Syst. Video Technol.*, 31(12), 2021, 4909–23.

[215] K. Wang, F. Torkhani, and A. Montanvert, "A Fast Roughness-based Approach to the Assessment of 3D Mesh Visual Quality", *Computers & Graphics*, 36(7), 2012, 808–18.

[216] X. Wang, Y. Mizukami, M. Tada, and F. Matsuno, "Navigation of a Mobile Robot in a Dynamic Environment Using a Point Cloud Map", *Artificial Life and Robotics*, 26, 2021, 10–20.

[217] Y. Wang, D.-M. Yan, X. Liu, C. Tang, J. Guo, X. Zhang, and P. Won, "Isotropic Surface Remeshing without Large and Small Angles", *IEEE Trans. Visualization and Computer Graphics*, 25(7), 2019, 2430–42.

[218] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image Quality Assessment: From Error Visibility to Structural Similarity", *IEEE Trans. Image Proc.*, 13(4), 2004, 600–12.

[219] B. Watson, A. Friedman, and A. McGaffey, "Measuring and Predicting Visual Fidelity", in *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*, 2001, 213–20.

[220] M. Westoby, J. Brasington, N. Glasser, M. Hambrey, and J. Reynolds, "Structure-from-Motion Photogrammetry: A Low-cost, Effective Tool for Geoscience Applications", *Geomorphology*, 179, 2012, 300–14.

[221] J. M. Wolfe, "Guided Search 2.0 A Revised Model of Visual Search", *Psychonomic Bulletin & Review*, 1(2), 1994, 202–38.

[222] J. Wu, X. Shen, W. Zhu, and L. Liu, "Mesh Saliency with Global Rarity", *Graphical Models*, 75(5), 2013, 255–64.

[223] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao, "3D ShapeNet: A Deep Representation for Volumetric Shapes", in *IEEE Conf. Computer Vision and Pattern Recognition*, 2015, 1912–20.

[224] J. Xiong, H. Gao, M. Wang, H. Li, and W. Lin, "Occupancy Map Guided Fast Video based Dynamic Point Cloud Coding", *IEEE Trans. on Circuits and Syst. Video Technol.*, 2021.

[225] Q. Yang, Z. Ma, Y. Xu, Z. Li, and J. Sun, "Inferring Point Cloud Quality via Graph Similarity", in *IEEE Trans. Pattern Anal. Mach. Intell.* 2021.

[226] Q. Yang, 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 Trans. Multimedia*, 2021.

[227] S. Yang, G. Lin, Q. Jiang, and W. Lin, "A Dilated Inception Network for Visual Saliency Prediction", *IEEE Trans. on Multimedia*, 22(8), 2020, 2163–76.

[228] X. Yang, W. Lin, Z. Lu, E. Ong, and S. Yao, "Just Noticeable Distortion Model and Its Applications in Video Coding", *Signal Processing: Image Communication*, 20(7), 2005, 662–80.

[229] H. Yee, S. Pattanaik, and D. P. Greenberg, "Spatiotemporal Sensistivity and Visual Attention for Efficient Rendering of Dynamic Environments", *CM Trans. Graph.*, 20(1), 2001, 3965.

[230] Z. J. Yew and G. H. Lee, "3DFeat-Net: Weakly Supervised Local 3D Features for Point Cloud Registration", in *European Conf. Computer Vision (ECCV)*, 2018, 630–46.

[231] H. Yoon, M. Jang, J. Huh, J. Kang, and S. Lee, "Multiple Sensor Synchronization with Therealsense RGB-D Camera", *Sensors*, 21(18), 2021, 6276.

[232] J. S. Yun and J. Y. Sim, "Supervoxel-based Saliency Detection for Large-scale Colored 3D Point Clouds", in *Proceedings - International Conference on Image Processing, ICIP*, Vol. 2016-Augus, Aug. 2016, 4062–6.

[233] V. Zakharchenko, "V-PCC Codec Description, ISO/IEC TC1/SC29/WG11 MPEG, N18190, Marrakech", Jan. 2019.

[234] A. Zeng, S. Song, M. Niessner, M. Fisher, J. Xiao, and T. Funkhouser, "3DMatch: Learning Local Geometric Descriptors From RGB-D Reconstructions", in *IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2017, 1802–11.

[235] 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)*, IEEE, 2020, 1–6.

[236] G. Zhai and X. Min, "Perceptual Image Quality Assessment: A Survey", *Science China Information Sciences*, 63, 2020.

[237] W. Zhang, W. Zou, and F. Yang, "Linking Visual Saliency Deviation to Image Quality Degradation: A Saliency Deviation-based Image Quality Index", *Signal Processing: Image Communication*, 75, 2019, 168–77.

[238] B. Zhao, W. Lin, and C. Lv, "Fine-Grained Patch Segmentation and Rasterization for 3-D Point Cloud Attribute Compression", *IEEE Trans. on Circuits and Syst. Video Technol.*, 31(12), 2021, 4590–602.

[239] P. Zheng, H. Wang, Z. Sang, and others., "Smart Manufacturing Systems for Industry 4.0: Conceptual Framework, Scenarios, and Future Perspectives", *Front. Mech. Eng.*, 13, 2018, 137–50.

[240] T. Zheng, C. Chen, J. Yuan, B. Li, and K. Ren, "PointCloud Saliency Maps", in *Int'l Conf. Computer Vision (ICCV)*, 2019, 1598–606.

[241] S. Zhong, Z. Z., and J. Hua, "Surface Reconstruction by Parallel and Unified Particle-based Resampling from Point Clouds", *Computer Aided Geometric Design*, 71, 2019, 43–62.

[242] Y. Zou, P. X. Liu, Q. Cheng, P. Lai, and C. Li, "A New Deformation Model of Biological Tissue for Surgery Simulation", *IEEE Trans. Cybernetics*, 47(11), 2016, 3494–503.