Multimodal Technologies and Interaction

MDPI

Article

Saliency-Guided Point Cloud Compression for 3D Live Reconstruction

Pietro Ruiu *,† **, Lorenzo Mascia** † **and Enrico Grosso** †

Department of Biomedical Sciences, University of Sassari, 07100 Sassari, Italy; l.mascia2@studenti.uniss.it (L.M.); grosso@uniss.it (E.G.)

* Correspondence: pruiu@uniss.it

† These authors contributed equally to this work.

Abstract: 3D modeling and reconstruction are critical to creating immersive XR experiences, providing realistic virtual environments, objects, and interactions that increase user engagement and enable new forms of content manipulation. Today, 3D data can be easily captured using off-the-shelf, specialized headsets; very often, these tools provide real-time, albeit low-resolution, integration of continuously captured depth maps. This approach is generally suitable for basic AR and MR applications, where users can easily direct their attention to points of interest and benefit from a fully user-centric perspective. However, it proves to be less effective in more complex scenarios such as multi-user telepresence or telerobotics, where real-time transmission of local surroundings to remote users is essential. Two primary questions emerge: (i) what strategies are available for achieving real-time 3D reconstruction in such systems? and (ii) how can the effectiveness of real-time 3D reconstruction methods be assessed? This paper explores various approaches to the challenge of live 3D reconstruction from typical point cloud data. It first introduces some common data flow patterns that characterize virtual reality applications and shows that achieving high-speed data transmission and efficient data compression is critical to maintaining visual continuity and ensuring a satisfactory user experience. The paper thus introduces the concept of saliency-driven compression/reconstruction and compares it with alternative state-of-the-art approaches.

Keywords: point cloud; compression; 3D live reconstruction; saliency; visual attention

Citation: Ruiu, P.; Mascia, L.; Grosso, E. Saliency-Guided Point Cloud Compression for 3D Live Reconstruction. *Multimodal Technol. Interact.* **2024**, *8*, 36. https://doi.org/10.3390/mti8050036

Academic Editor: Cristina Portales

Received: 29 March 2024
Revised: 22 April 2024
Accepted: 26 April 2024
Published: 3 May 2024

1. Introduction

Nowadays, the 3D representation of objects and environments is increasingly utilized in common applications like telepresence and remote collaboration [1,2], autonomous driving [3], remote operations in hazardous environments [4,5], telemedicine [6,7]. Specifically, thanks to the use of special sensors such as depth (RGB-D) cameras or LiDAR scanners, it is possible to faithfully reconstruct reality and reproduce it in virtual environments, enabling new forms of interaction. The typical representation used in this kind of application is the point cloud; it involves the storage and visualization of large sets of points in 3D space, where each point typically corresponds to a single spatial coordinate along with optional attributes such as color, intensity, or surface normal. One of the main advantages of point cloud representation is its ability to capture detailed geometric information of complex surfaces and environments, making it a valuable tool in various contexts for tasks such as object recognition, scene reconstruction, and measurement. However, managing and processing large point cloud datasets can be computationally intensive and challenging due to their high dimensionality and potentially massive size. Moreover, in the case of network transmission, several factors come into play to ensure efficient and reliable transfer: for example, data compression turns to be a critical aspect when dealing with large dataset, and latency can cause delays in data transmission, affecting the responsiveness of the applications. Overall, efficient network transmission of point cloud data requires a balance between data size, transmission speed, reliability, and latency to meet the requirements of

specific applications [8]. For this reason, the quest for methods to compress or reduce 3D live data is gaining considerable interest within the scientific community.

The objective of this study is twofold. First, a novel approach for the compression and reconstruction of point cloud representations is proposed; this approach is based on the detection of saliency points within a scene and is inspired by the way human vision operates. Just as the human eye does not attempt to capture every detail in the visual field but instead focuses on specific areas or points, this approach optimizes the amount of information captured and consequently processed. More precisely, two novel encoding methods are introduced: the Alternate Depth Compression (ADC) and the Log-Polar (LP) method, detailed in Section 3.1 and Section 3.2, respectively. Both methods utilize depth maps and RGB data as inputs and share the same basic idea, although they employ different compression filters.

Second, in order to assess the effectiveness of the proposed solutions, the developed methods are compared with state-of-the-art compression techniques. The comparison takes into account both the compression efficiency and the quality of reconstructed surfaces. Two distinct scenarios are explored in detail: telepresence in video conferencing, which emphasizes the faces of participants as the focal point of attention, and telerobotics, where a dynamically updated 3D environment captured by a mobile remote camera is shared with telepresence users. In both scenarios, the system autonomously extracts and utilizes a subset of available points, aiming to minimize data size while ensuring high-quality 3D information.

2. Background

2.1. 3D Live Reconstruction for Mixed Reality

Although both 3D live reconstruction and traditional 3D reconstruction aim to generate 3D models from acquired data, they differ significantly in terms of real-time processing and scope. In live reconstruction, the goal is to reconstruct a 3D model of a scene or object at the moment it is captured, allowing for immediate visualization and interaction in a virtual environment. Reconstruction algorithms can, therefore, prioritize speed over accuracy to ensure that 3D models can be generated quickly enough for real-time interaction. Differences also exist in relation to acquisition systems. Real-time capture devices such as depth cameras, LiDAR sensors, or RGB-D cameras, can provide depth information along with color imagery. These devices are optimized to acquire data quickly and efficiently, and this enables real-time reconstruction, but often at the expense of resolution and accuracy.

A point-based representation relative to the observer (in the form of depth) is the most widely used of those provided by real-time acquisition devices. Depth data can either undergo pre-processing (for example, background cleaning or generation of a point cloud independent from the observer) or be directly utilized as input data to compute a full 3D representation of the target scene, employing approaches such as surface meshes or volumetric fusion [9,10].

With reference to Figure 1, it is worth noting that a full 3D reconstruction performed immediately after acquisition (server side) greatly affects the transmission phase. In fact, in this case, the amount of data to be transmitted is very large (fine-scale volumetric representations are usually memory-greedy) and requires specific manipulation in order to limit the transmission time. For example, an online system for large and fine-scale volumetric reconstruction has been proposed in [11]. This approach uses a memory-efficient spatial hashing scheme, enabling real-time access and updates of implicit surface data without relying on hierarchical grid structures. Surface data are stored densely only where measurements are observed, and efficient data streaming ensures scalability, especially during sensor motion. SLAM-CAST [12] is a novel scene representation and transmission protocol based on Marching Cubes (MC) indices, enabling operation in low-bandwidth remote connection scenarios. Instead of reconstructing geometry on the server side or performing server-side rendering, the scene is encoded as a compressed sequence of voxel block indices and values. The final geometry reconstruction is left to the exploration client.

This approach significantly reduces bandwidth requirements compared to previous voxel-based methods. In [13], a method for producing a single full 3D reconstruction surface of a moving user in real-time is introduced by using multiple depth sensors and a marching square approach to produce a single full 3D reconstruction surface. This method has been specifically designed for Mixed Reality (MR) telepresence.

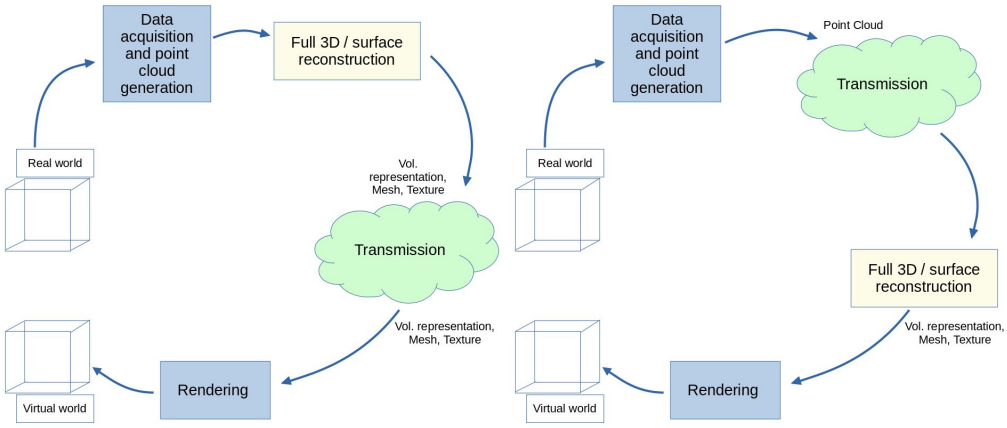


Figure 1. Data flow for 3D live reconstruction before (**left**) and after (**right**) transmission.

Figure 1 also shows an alternative approach to 3D live reconstruction. In this case, direct use of point cloud is favored due to its ability to reduce the computational burden on the server side and limit the overall memory requirements associated with volumetric approaches [14]. Obviously, 3D reconstruction, in this case, occurs after transmission (client or exploration side). Point cloud compression and transmission can be achieved either by considering the raw 3D data and attempting to regularize the non-uniform and sparse structure of point clouds or by converting (or keeping) the point cloud into the original depth format [15,16]. In this case, data can be compressed and transmitted over a network in a way quite similar to what is done for common 2D images [17,18]. The main 2D-based techniques leverage traditional image or video compression methods, such as JPEG, MPEG, or dictionary-based compression [19].

2.2. Depth Compression

Although depth image compression has been studied for many years, it remains an open discussion topic in the scientific community. Various lossy and lossless methods have been proposed [15,19]; however, a standard reference method has not yet been established.

Several methods adopt standard RGB compression techniques to compress depth. For instance, in [20], the original 16-bit depth map is encoded into an 8-bit, three-channel image, which is then processed by a video encoder and transferred over the network.

Other methods rely on the segmentation of depth or RGB data. In [21], a compression method based on segmenting the contour of the image is proposed. Depth information for each segment within the depth image is represented by a piecewise-linear function, enabling the representation of surfaces not parallel to the image plane and characterized by a linear gradient in the depth image. In [22,23], segmentation of depth based on planar information is exploited.

In [24], Microsoft introduces a 2D lossless method named RVL, which attains comparable compression rates to commonly used lossless techniques but operates notably faster. The proposed method is a blend of Run Length Encoding and Variable Length Encoding schemes, which consider the number of zeros, the number of non-zeros, and the differences (deltas) of successive non-zero pixel values.

In [25], Intel introduces a colorization method for depth images utilizing the Hue color space. This enables the treated image to be regarded as a standard RGB image, thereby facilitating compression using widely available standard compression methods. In 2022, a

novel approach was proposed by Chen and colleagues [26], leveraging shared structural information within RGB-D data to reduce cross-modal redundancies in the depth map. This method utilizes machine learning techniques, employing convolutional layers and activation layers to extract structural information from the latent features of RGB data.

FitDepth [27] is one of the most recent methods proposed, relying on multiple curve fittings for encoding noisy depth images. The compression mechanism processes each row of the depth image, describing it as a set of polynomial functions. This method claims superior speed and compression performance compared to JPEG2000 and PNG formats.

In contrast to the above methods, attention-based compression methods have been proposed in [28,29]. Both methods are grounded in the concept of region-of-interest (ROI), defined through segmentation approaches that extract edges or main objects.

In [30], a similar approach is applied to 3D medical images, introducing the concept of volume-of-interest (VOI). This compression scheme relies on object-based coding, employing segmentation approaches to separate objects from the background and compressing object and background regions separately.

2.3. Saliency

In the field of computer vision, saliency detection has long been a tool for the identification of the most visually significant regions within an image. Traditional approaches to saliency detection are based on low-level features such as color, intensity, texture, and orientation. Examples include the Itti–Koch model [31], an implementation of earlier psychological theories of bottom-up attention, and the Koch and Ullman Computational Attention Architecture [32]. Although effective for simple images with few featured elements, these methods struggle in complex scenes with cluttered backgrounds. To overcome this problem, Liu and colleagues [33] reformulated saliency detection as an image segmentation problem, focusing on the separation of a "salient object" from the image background. For this purpose, they proposed a set of novel features, including contrast analysis over multiple spatial scales and color spatial distribution, and suggested the use of Random Fields to effectively combine these features. Deep Learning-Based Techniques have also been recently adopted; leveraging neural networks, deep learning models have revolutionized the field [34] showing how to capture hierarchical saliency information from deep, coarse layers (global saliency response) to shallow, fine layers (local saliency response). The CNNs have proven to be highly effective in identifying areas of greatest interest, although the relationship between CNN results and the prediction of human visual attention remains largely obscure and subject to study [35,36].

3. Methodology

In this paper, we present and compare different innovative techniques for optimizing remote 3D data reconstruction. Two novel compression methods are proposed: the *Alternate Depth Compression (ADC)* and the *Log-Polar (LP)* method, detailed in Section 3.1 and Section 3.2, respectively. Furthermore, two distinct application scenarios characterized by different attention mechanisms are taken into consideration: the *single-attention* scenario described in Section 3.3 and the *multi-attention* scenario presented in Section 3.4.

3.1. ADC Compression Method

The ADC uses quantization to reduce the overall amount of information. Essentially, it preserves maximum resolution for the portion of the image that is most interesting to the viewer (i.e., the focus of attention) while under-sampling pixels in other areas. ADC was designed specifically for telepresence applications, where the face of the interlocutor is considered the focal point for the observer. For this purpose, ADC exploits a basic face detection tool [37] and selects through this the face bounding box (FBB) of interest. As depicted in Algorithm 1, the ADC code accepts as inputs the image file (RGB or depth), the quantization step, and a data structure FBB defining the face bounding box. The code iterates through each row of the matrix derived from the input image, examining

individual pixels. If a pixel falls within one of the rows defined by the quantization steps or resides within the region delineated by the FBB, it is appended to the encoded list. This list constitutes the ultimate output of the encoding procedure.

Algorithm 1 ADC compression

```
procedure ADC(image, step, FBB)
    w = width of the depth file
    h = height of the depth file
    encoded = newList()
    jumpCounter ← 0
    for y ← 1, h do
        if jumpCounter = step then
            jumpCounter ← 0
        end if
        for x ← 1, w do
            if (x, y) ∈ FBB OR jumpCounter = 0 then
                encoded.Append(depth[x, y])
            end if
        end for
        jumpCounter ← jumpCounter + 1
    end for
    return encoded
end procedure
```

3.2. Log-Polar Compression Method

Log-polar compression is a well-known type of 2D compression based on the log-polar transformation and inspired by the mapping of retinal receptive fields. Images or depth images are transformed from their usual Cartesian coordinates to log-polar coordinates (ρ, θ), where ρ represents the eccentricity (logarithm of the distance from the origin) and θ represents the angle. Log-polar compression is particularly useful in computer vision tasks where variations in scale and rotation need to be accommodated. It can also be applied in tasks like feature extraction and pattern matching, where the inherent properties of log-polar coordinates facilitate robust and efficient processing. Since the late eighties, the log-polar has been successfully applied to shape analysis [38], video compression [39] and robotics [40]. The log-polar transformation applied to the case study is the one described in [41] but with a number of receptive fields (24 radial fields along 32 angular directions) suitable to ensure the quality of the reconstruction. Figure 2 shows the application of LP compression to a single frame; in analogy with ADC, the FBB is placed on the subject's face.

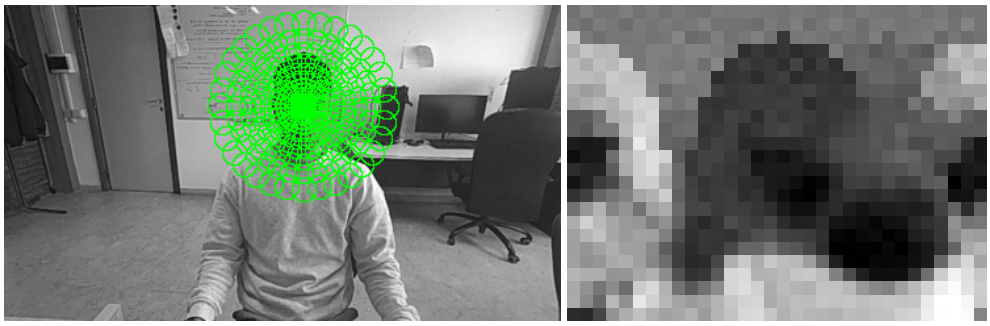


Figure 2. (**left**) Example of the LP compression scheme applied to a single frame (424 × 240 pixels), the center of the filter is the subject's face. (**right**) Resulting in LP image (24 × 32 pixels, enlarged for visualization purposes).

Algorithm 2 shows the pseudo-code of the log-polar compression implementation. The image serves as the input file (RGB or depth), while cx and cy denote the center of

the filter. Parameters overlayAngles and overlayRad govern the degree of oversampling of the log-polar filter, while numAngles and numRads determine the dimensions of the log-polar filter. As a preliminary step, the method computes the positions of all filtering areas (receptive fields) using the functions CalculateAngles() and CalculateRads(). Thus, for each radial distance, a kernel is computed with the CreateMeanKernel() function, which is then iteratively applied along the circumference. The final output is the compressed pixels stored in the encoded list.

Algorithm 2 Log-polar compression

```
procedure LOGPOLAR(image, cx, cy, overlayAngles, overlayRad, numAngles, numRads)
    angles ← CalculateAngles()
    rads ← CalculateRads()
    encoded = newList()
    for i ← 1, numRads do
        kernel ← CreateMeanKernel()
        for j ← 1, numAngles do
            x ← cx + cos(angles[j]) ∗ rads[i])
            y ← cy + sin(angles[j]) ∗ rads[i])
            encoded.Append(ApplyKernel(image, x, y))
        end for
    end for
    return encoded
end procedure
```

3.3. Single-Attention Scenario

The single-attention scenario pertains to a scene where the focus of attention is clear: the observer is looking at a specific point, typically in the foreground, and directs his attention to it. An example of this is a telepresence application where the observer's attention is primarily focused on a person, specifically the face, represented in the virtual environment. In Figure 3, the schema of the enconding-decoding mechanisms for the single attention scenario is shown.

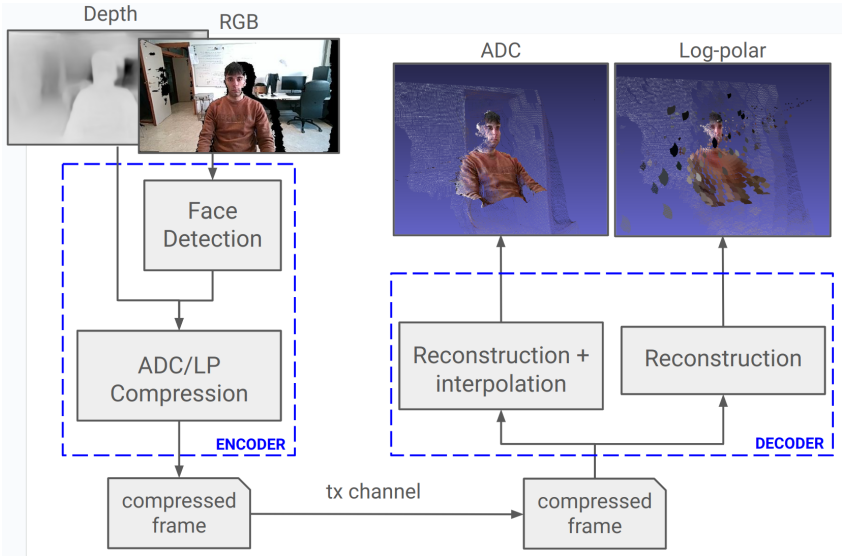


Figure 3. Single-attention compression schema for ADC and the LP method.

The RGB image is processed with the face detector, which generates a bounding box (FBB) around the subject's face in the scene. This FBB can then be utilized to apply an ADC or log-polar compression method to either the depth or RGB data. The resulting

compressed data can be transmitted over the network. At the receiver, the decoder generates the 3D data in the form of a point cloud. If the ADC method has been employed, missing data resulting from lossy compression are estimated using bilinear interpolation. If the log-polar method has been utilized, the final value assigned to each point of the cloud (either depth or color) is computed as the average of all affected (overlapping) receptive fields.

3.4. Multiple-Attention Scenario

The multiple-attention scenario assumes that the observer does not have a specific subject to focus on but is instead actively observing a scene, such as an environment that needs to be reconstructed remotely. This applies, for example, to applications like telerobotics or remote operations. In this case, simple saliency detection techniques, based on the Laplacian of Gaussian (LoG) filter, are applied [42], and the most prominent points are used to guide the log-polar or ADC compression (Figure 4).

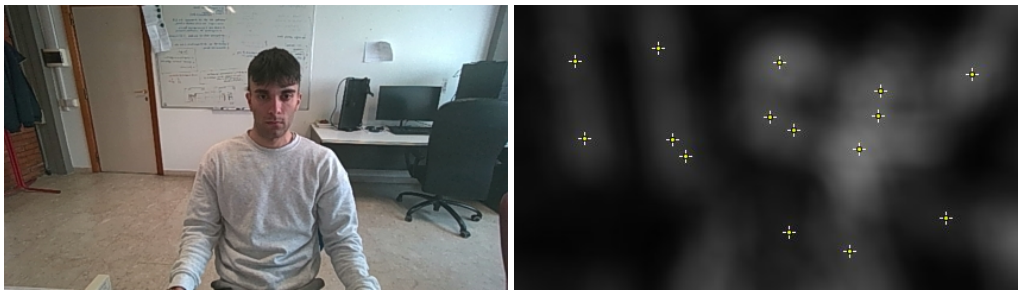


Figure 4. Example of a frame (**left**) processed with the LoG filter in order to extract local presence (density) of image contrast; prominent points (excluding the image borders) are indicated (**right**).

In Figure 5, the encoding-decoding blocks of the multiple-attention scenario are depicted. In this scenario, a set of frames is supplied as input. The RGB image of the initial frame is employed to extract the saliency points using the LoG filter. These points are then fed into the log-polar or ADC compressor, which, for each frame (to both depth and RGB), applies a log-polar or ADC filter, respectively, with a new saliency point as the center. Once transferred, the compressed data are used for the reconstruction. A 3D reconstruction refinement process generates the point cloud by merging the newly compressed frame with the previously received frames. The final outcome is a cumulative point cloud with a progressively increased resolution.

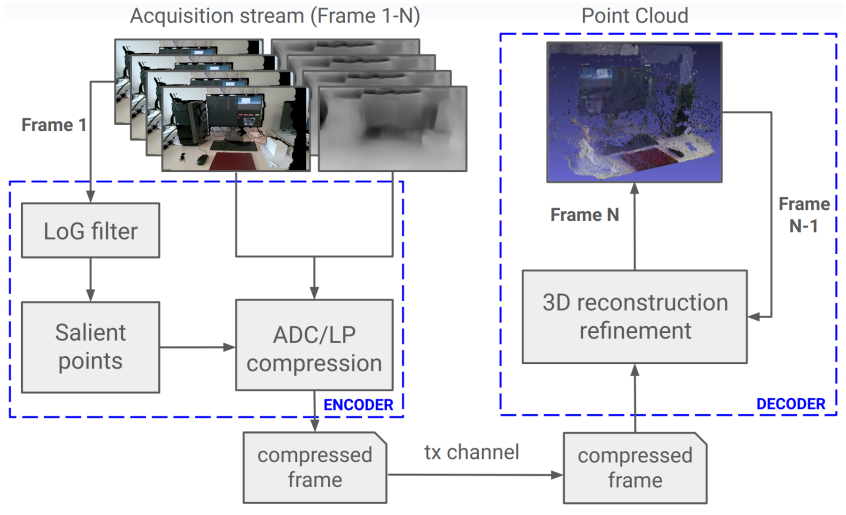


Figure 5. Multiple-attention compression schema using ADC and LP methods.

3.5. Comparison Metrics

To ensure a fair comparison, two classes of metrics have been selected: one for evaluating compression performance and the other for assessing the final reconstruction quality. Performance metrics are based on 2D data generated from the encoder; quality metrics, instead, evaluate the decoder's output, i.e., the point cloud. For the point cloud quality assessment, both Full-Reference (FR) and No-Reference (NR) metrics are used; while FR metrics compare a reconstructed point cloud against its original reference, NR metrics can estimate the subjective quality score of point clouds without relying on a reference. The metrics chosen are listed and briefly described below.

3.5.1. Performance Metrics

- Bit per pixel (BPP): size in bit used to represent a single pixel;
- Bit-rate (BR): total size of the computed frames divided by the frame transmission rate;
- Compression ratio (CR): the ratio between the BPP raw image and the BPP of encoded image;
- Encoding and decoding time (ET, DT): time to process the data and generate the corresponding output.

3.5.2. Quality Metrics

- Peak Signal-to-Noise Ratio (PSNR)—FR metric: is the ratio between the peak signal and the MSE error. The peak is calculated as the diagonal distance of the bounding box containing all the points of the point cloud [43]. Both the PNSR point-to-point (PSNR-D1), which calculates the intra-point MSE error and the PSNR point-to-plane (PSNR-D2), which calculates the distances with respect to a plane, are used;
- PointSSIM (PSSIM)—FR metric: is a family of statistical dispersion measurements for the prediction of perceptual degradations of point cloud data. It encompasses four types of attributes: colors, curvatures, normals, and geometry [44]. In this work, only colors (luminance values) and geometry (Euclidean distances) have been considered;
- Video Quality Assessment Point Cloud (VQA-PC)—NR metric: spatial and temporal features extracted from a point cloud are used to estimate the quality level through an ML model [45];
- Multi-Modal Point Cloud Quality Assessment (MM-PCQA)—NR metric: the point cloud is split into various 3D sub-models and rendered into 2D image projections used for feature extraction. Both sub-models and projections are then encoded with two neural networks, and symmetric cross-modal attention is performed. Finally, the quality level is estimated through a quality regression block [46].

4. Comparison Experiments

The frames used for the comparison were captured using the Intel® RealSense™ Depth Camera D455 (Intel Corporation, 2200 Mission College Blvd. Santa Clara, CA 95052 USA), an off-the-shelf device equipped with a depth sensor, an RGB camera, and an IMU unit, ensuring a comprehensive capture of the scene. The depth field of view (FOV) of the device spans 87 × 58 degrees, with a depth range between 0.6 and 60 m. RGB data have been acquired and aligned with the depth map for each frame to obtain the RGB-D data used for the point cloud generation. Both the depth and RGB images acquired by the RealSense camera have a resolution of 424 × 240 pixels. Note that the data acquired with this equipment constitute a partial scan of a scene (persons or objects, background), capturing only the visible portion thereof. To achieve a comprehensive scan, various frames need to be stitched together, considering the acquisition position of the device. Illustrative samples of acquisition data (RGB and depth) for both scenarios are reported in Figure 6.

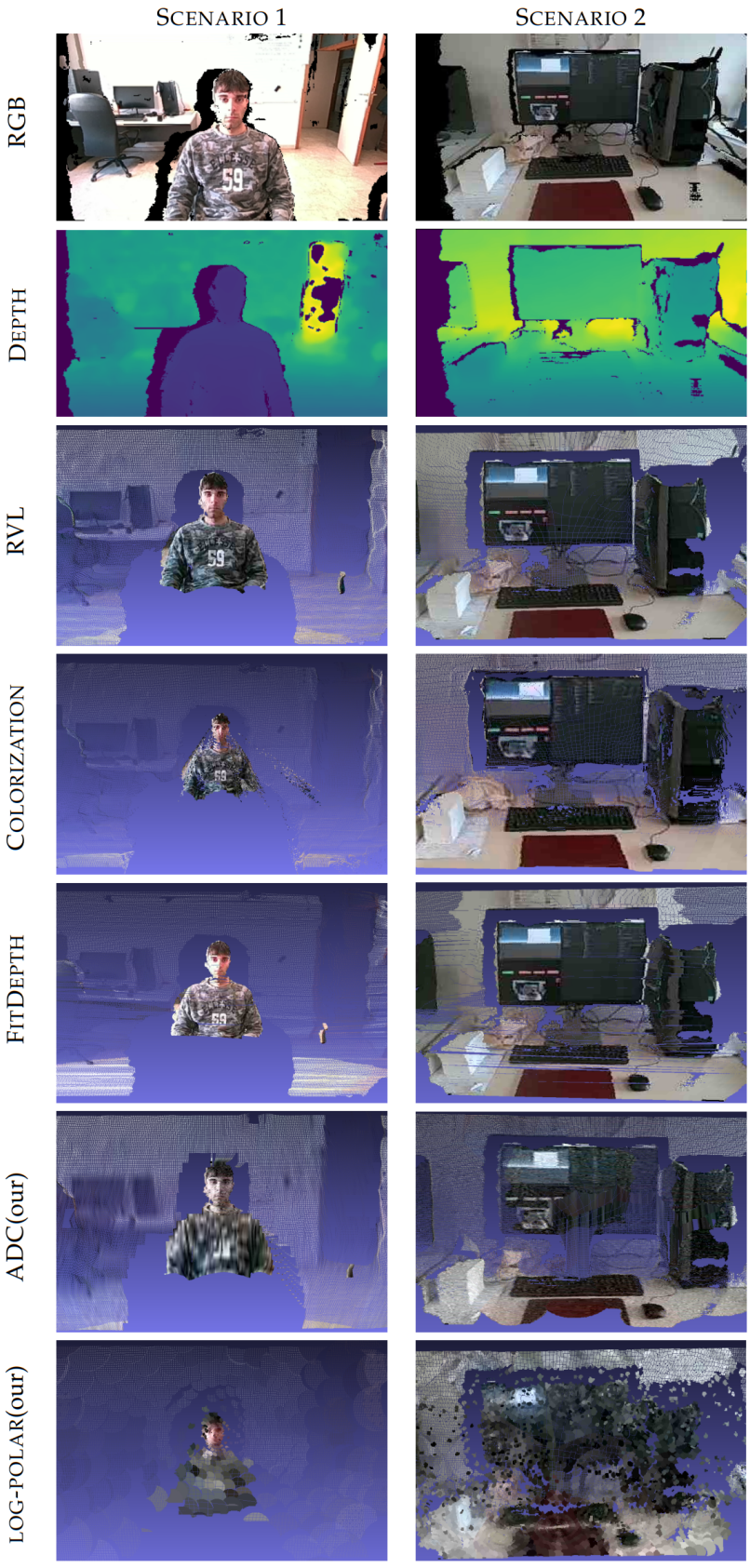


Figure 6. Visual comparisons of the reconstructed point clouds for all compression methods employed in the experiments are provided for both the single attention (1) and multiple attention scenarios (2).

Two experiments have been conducted based on the scenarios described in Section 3. In Table 1, the parameters used for the encoding/decoding filters in both scenarios are reported. The computing hardware utilized for the experiments comprised a Dell Alienware workstation equipped with an Intel i7 processor, 32 GB of RAM, and an RTX1080 GPU with 8 GB of memory.

Table 1. Specifications of the filters used in the two scenarios.

Scenario	ADC (Quantization Step)	Log-Polar (Receptive Fields)
Single-attention	10 px	30 × 32
Multiple-attention	40 px	24 × 32

In the absence of an actual standard, RVL lossless compression method [24], has been chosen as a reference for the evaluation of the results. Furthermore, an extensive comparison with two recent methods retrieved from the literature has been conducted: the FitDepth compression method [27] and the Realsense Depth Colorization (Colorization) [25]. Both these methods stand out as particularly noteworthy among the recent advancements in in-depth image compression techniques.

Regarding FitDepth, we opted for linear polynomial encoding without residual encoding, as it demonstrated superior performance compared to other encoding modalities characterizing the method. For Realsense Depth Colorization, we used JPEG compression with a quality level set to 25. This decision was made to attain a high compression ratio while preserving the quality of the colorized depth frame.

When employing the log-polar or ADC method for RGB data compression, the same compression approach is applied to the corresponding RGB images, along with JPEG compression. Conversely, for the RVL, FitDepth, and Colorization methods, only the standard JPEG compression is employed.

5. Results

Table 2 shows the results concerning performance metrics. Note that, for both scenarios, coding of the original raw frame requires 40 BPP (Bits Per Pixel), with a BR (Bit Rate) at 30 fps of about 117.6 Mbps. The results show a superior performance of the log-polar method in both scenarios, considering BPP, BR, and CR metrics. ADC instead outperforms other methods only in the multiple-attention scenario. However, both proposed methods entail longer encoding and decoding times (ET and DT metrics), primarily due to the non-optimized code utilized in the experiments.

Table 2. Performance metrics calculated for all methods considered across both single-attention and multiple-attention scenarios. Arrows in the columns heading indicate whether a higher or lower value is preferred. In bold, the values with the best performance.

Scenario	Method	BPP [bit] ↓	BR@30fps [Mbps] ↓	CR ↑	ET [ms] ↓	DT [ms] ↓
Single	RVL	6.24	18.38	6.40	4.41	6.4
	Colorization	1.35	3.99	29.62	**4.0**	**4.0**
	FitDepth	2.31	6.79	17.31	53.0	15.0
	ADC	2.60	7.67	15.33	11.93	57.42
	Log-polar	**0.21**	**0.62**	**187.1**	10.6	4.5
Multiple	RVL	4.65	13.70	8.58	3.87	3.17
	Colorization	1.59	4.67	25.15	**4.0**	**4.0**
	FitDepth	1.75	5.17	22.85	52.0	16.0
	ADC	1.25	3.67	32.0	11.20	54.94
	Log-polar	**0.17**	**0.50**	**231.3**	8.45	4.46

In Table 3, the results of the quality analysis are reported. For RVL compression, PSNR and PSSIM metrics are not reported due to the lossless nature of the method.

Table 3. Quality metrics calculated for all methods considered across both single-attention and multiple-attention scenarios. Higher values are better. In bold, the values with the best quality.

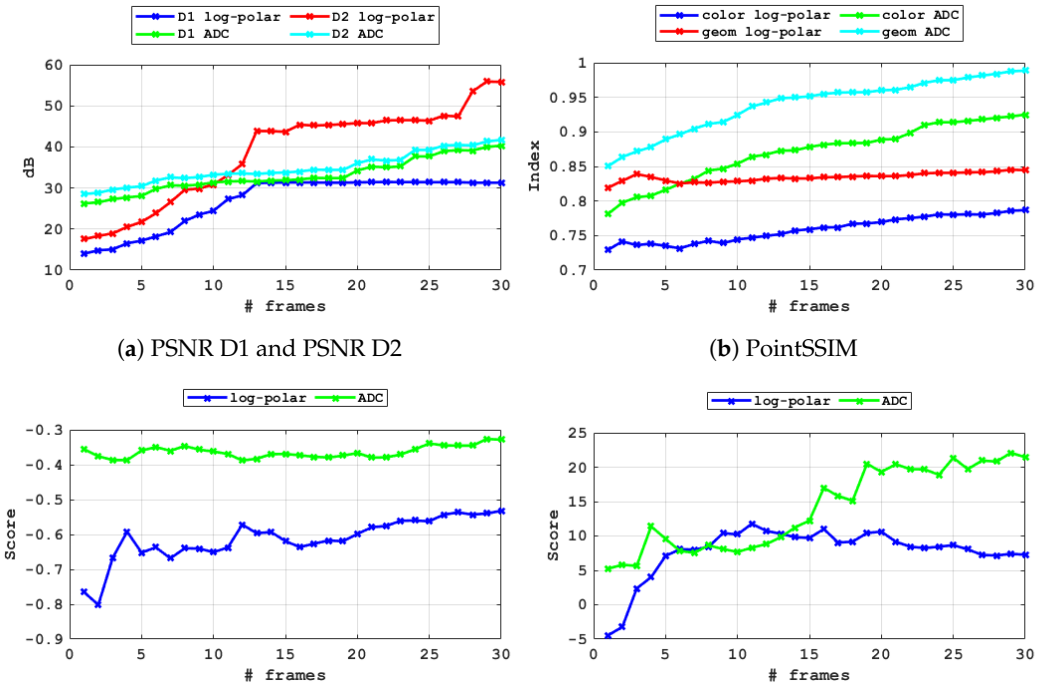
Scenario	Method	PSNR D1 [dB]	PSNR D2 [dB]	PSSIM Color	PSSIM Geom	VQA-PC	MM-PCQA
Single	RVL	–	–	0.97	–	−0.53	31.74
	Colorization	35.05	40.56	0.85	0.78	−1.02	7.35
	FitDepth	39.79	**44.43**	0.84	0.82	**−0.55**	**22.86**
	ADC	**40.46**	41.85	**0.89**	**0.90**	−0.60	1.51
	Log-polar	29.41	38.87	0.73	0.73	−0.85	2.81
Multiple	RVL	–	–	0.97	–	−0.325	27.38
	Colorization	27.61	31.08	0.91	0.81	−0.89	9.92
	FitDepth	**41.09**	44.55	0.80	0.86	−0.58	5.12
	ADC	39.97	41.70	**0.92**	**0.98**	**−0.328**	**21.45**
	Log-polar	31.28	**55.79**	0.78	0.84	−0.53	7.26

Upon examining the quality results presented in the table for the single-attention scenario for the proposed methods, it is evident that ADC surpasses log-polar in most of the FR and NR metrics. Notably, the ADC PSSIM Geom score approaches the peak value of 1, indicating a close resemblance to the geometric quality of the original frame. Similarly, the VQA-PC score for ADC closely aligns with the score achieved by the RVL compression method. Extending the comparison to other methods, again, in the single attention scenario, ADC outperforms others considering FR metrics (except for PSNR D2), whereas FitDepth emerges as the top performer for NR metrics.

In the multiple-attention scenario, ADC maintains superior performance across all metrics except for PSNR, where FitDepth and Log-polar achieve higher values. In the multiple-attention scenario, the log-polar and ADC quality value refers to the fully reconstructed point cloud after applying all filters to every salient point. Compared to the single-attention scenario, where only one log-polar or ADC center is processed, we observe higher values, approaching the quality achieved by lossless methods like RVL.

The improvement is even more evident when we look at Figure 7, which shows the trend over time of the quality metrics for the multiple-attention scenario obtained by the log-polar and ADC methods. Note that both VQA-PC (Figure 7c) and MM-PCQA (Figure 7d) output a score value indicating a subjective quality level of the point cloud. These metrics are learned by training the system on human subjective scores, referred to as Mean Opinion Scores (MOS), which usually range from 0 to 5 points. As reference values for these NR metrics, we can look at the scores achieved by the RVL method (−0.325 and 27.38), which is the only lossless method considered; RVL values can thus be regarded as the upper bound achievable by this type of data. The graph shows the gradual increase in quality as the point cloud is reconstructed frame by frame for both metrics. Initially, the quality curve shows some instability, especially for NR metrics. However, after about fifteen frames, quality increases steadily until it stabilizes at higher values. It is worth noting that the initial instability most likely depends on the order in which the salient points are processed. However, this aspect has not been extensively tested.

Finally, it should be noted that the VQA-PC and MM-PCQA NR models were trained on datasets containing humans and small objects [47], thus very different from the large environments that characterize the experiments performed. This inconsistency does not compromise the assessment of the quality of the point cloud; in fact, one can still appreciate an increasing trend for both metrics as the number of frames computed in the multiple attention scenario increases.



(**a**) PSNR D1 and PSNR D2 (**b**) PointSSIM

(**c**) VQA-PC (**d**) MM-PCQA

Figure 7. Comparison of the quality of reconstructed point clouds post-compression using the two proposed techniques (ADC and LP) for the multiple-attention scenario. (**a**) PSNR expressed as a value in dB indicating the robustness of the signal. (**b**) SSIM expressed as an index capturing perceptual quality. (**c**) VQA-PC expressed as a score capturing the video quality. (**d**) MM-PCQA expressed as a score indicating the quality level of the point cloud.

In Figure 6, a visual comparison of all the methods employed for the experiments along with the input files are shown: images confirm the results presented in Table 3. The methods selected from the literature (i.e., RVL, Color, and FitDepth) exhibit good quality, albeit at the expense of compression performance. The ADC method demonstrates a reconstruction quality comparable to that of methods from the literature for both scenarios. The log-polar method, despite achieving superior compression performance, exhibits more noise that compromises the quality of the reconstruction. However, in the single-attention experiment, the focus of the observer's attention (such as the face of a person) is well reconstructed, whereas the background appears to be more confused. This is also evident in Figure 8, which depicts the visual comparison of a detail from the single-attention scenario for ADC, LP, and FitDepth methods. It is worth noting that while the FitDepth method achieves very good quality performance for the global frame (as reported in Table 3 for VQA-PC and MM-PCQA metrics, which score -0.55 and 22.86, respectively), upon closer inspection of the detail of the face, the superior quality of the proposed methods becomes evident (see Figure 8b,c).

(**a**) FitDepth (**b**) ADC (**c**) LogPolar

Figure 8. Visual comparison of a detail of the face for the single-attention scenario among the FitDepth, ADC, and Log-polar methods.

6. Conclusions

This paper introduces a new approach for compression and live reconstruction of point clouds suitable for augmented and mixed reality applications. The approach exploits the concept of observer attention and outlines two different application scenarios: single attention, where attention is focused on a single object or point in the scene, and multiple attention, where attention is distributed over multiple points in the scene, resulting in an incremental reconstruction that stabilizes after a few frames. The proposed method involves two different types of compression: ADC, which quantizes RGB and depth data and uses linear interpolation for point cloud reconstruction, and log-polar filtering centered on one or more salient points.

Experimental evaluations, utilizing performance and quality metrics, were conducted across both scenarios, comparing the proposed techniques with a benchmark lossless compression method (RVL) and two recently released state-of-the-art methods (Intel Colorization and FitDepth). Results reveal the superior performance of the Log-polar method for the performance and ADC for the quality, albeit with slightly longer execution times. Specifically, in the single-attention scenario, ADC surpasses log-polar in quality, while in multiple-attention, log-polar achieves commendable quality levels after approximately fifteen frames, significantly reducing transmitted data due to its robust compression.

Despite being exploratory in nature, both approaches demonstrate promising potential and applicability to live 3D reconstruction applications, including telepresence, remote operations in hazardous environments, and telemedicine. Future endeavors encompass the exploration of alternative attention mechanisms, such as Vision Transformers or CNNs, for salient point identification, integration of segmentation models to distinguish foreground from background, and use of eye-tracking systems embedded in standard headsets to actively guide live 3D reconstruction.

Author Contributions: Conceptualization P.R. and E.G., methodology P.R., L.M. and E.G., software L.M., validation P.R. and L.M., formal analysis P.R. and E.G., investigation P.R., L.M. and E.G., writing—original draft preparation P.R., L.M. and E.G., writing—review and editing P.R., L.M. and E.G. All authors have read and agreed to the published version of the manuscript.

Funding: This work has been developed within the framework of the project e.INS—Ecosystem of Innovation for Next Generation Sardinia (cod. ECS 00000038) funded by the Italian Ministry for Research and Education (MUR) under the National Recovery and Resilience Plan (NRRP)—MISSION 4 COMPONENT 2, "From research to business" INVESTMENT 1.5, "Creation and strengthening of Ecosystems of innovation" and construction of "Territorial R&D Leaders". This work has received financial support under the National Recovery and Resilience Plan (NRRP), Mission 4, Component 2, Investment 1.1, Call for tender No. 1409 published on 14 September 2022 by the Italian Ministry of University and Research (MUR), funded by the European Union—NextGenerationEU—Project Title "METATwin—Metaverse & Human Digital Twin: digital identity, Biometrics and Privacy in the future virtual worlds", CUP J53D23015030001—Grant As-

signment Decree No. 0001382 adopted on 1 September 2023 by the Italian Ministry of Ministry of University and Research (MUR).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study, in the collection, analyses, or interpretation of data, in the writing of the manuscript, or in the decision to publish the results.

References

1. Orts-Escolano, S.; Rhemann, C.; Fanello, S.; Chang, W.; Kowdle, A.; Degtyarev, Y.; Kim, D.; Davidson, P.L.; Khamis, S.; Dou, M.; et al. Holoportation: Virtual 3d teleportation in real-time. In Proceedings of the 29th Annual Symposium on User Interface Software and Technology, Tokyo, Japan, 16–19 October 2016; pp. 741–754.
2. Fernandez, S.; Montagud, M.; Rincón, D.; Moragues, J.; Cernigliaro, G. Addressing Scalability for Real-time Multiuser Holoportation: Introducing and Assessing a Multipoint Control Unit (MCU) for Volumetric Video. In Proceedings of the 31st ACM International Conference on Multimedia, Ottawa, ON, Canada, 29 October–3 November 2023; pp. 9243–9251.
3. Geiger, A.; Lenz, P.; Urtasun, R. Are we ready for autonomous driving? the kitti vision benchmark suite. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Kolkata, India, 30 November–1 December 2012; pp. 3354–3361.
4. Szczurek, K.A.; Prades, R.M.; Matheson, E.; Rodriguez-Nogueira, J.; Castro, M.D. Multimodal Multi-User Mixed Reality Human–Robot Interface for Remote Operations in Hazardous Environments. *IEEE Access* **2023**, *11*, 17305–17333. [CrossRef]
5. Fairchild, A.J.; Campion, S.P.; García, A.S.; Wolff, R.; Fernando, T.; Roberts, D.J. A mixed reality telepresence system for collaborative space operation. *IEEE Trans. Circuits Syst. Video Technol.* **2016**, *27*, 814–827. [CrossRef]
6. Lo, S.; Rose, A.; Fowers, S.; Darko, K.; Britto, A.; Spina, T.; Ankrah, L.; Godonu, A.; Ntreh, D.; Lalwani, R.; et al. Ghana 3D Telemedicine International MDT: A proof-of-concept study. *J. Plast. Reconstr. Aesthetic Surg.* **2024**, *88*, 425–435. [CrossRef] [PubMed]
7. Lo, S.; Fowers, S.; Darko, K.; Spina, T.; Graham, C.; Britto, A.; Rose, A.; Tittsworth, D.; McIntyre, A.; O'Dowd, C.; et al. Participatory development of a 3D telemedicine system during COVID: The future of remote consultations. *J. Plast. Reconstr. Aesthetic Surg.* **2023**, *87*, 479–490. [CrossRef] [PubMed]
8. Murroni, M.; Anedda, M.; Fadda, M.; Ruiu, P.; Popescu, V.; Zaharia, C.; Giusto, D. 6G—Enabling the New Smart City: A Survey. *Sensors* **2023**, *23*, 7528. [CrossRef]
9. Hauswiesner, S.; Straka, M.; Reitmayr, G. Coherent image-based rendering of real-world objects. In Proceedings of the Symposium on Interactive 3D Graphics and Games, San Francisco, CA, USA, 18–20 February 2011; pp. 183–190.
10. Alexiadis, D.S.; Zarpalas, D.; Daras, P. Real-time, realistic full-body 3D reconstruction and texture mapping from multiple Kinects. In Proceedings of the IVMSP 2013, Seoul, Republic of Korea, 10–12 June 2013; pp. 1–4.
11. Nießner, M.; Zollhöfer, M.; Izadi, S.; Stamminger, M. Real-time 3D reconstruction at scale using voxel hashing. *ACM Trans. Graph.* **2013**, *32*, 1–11. [CrossRef]
12. Stotko, P.; Krumpen, S.; Hullin, M.B.; Weinmann, M.; Klein, R. SLAMCast: Large-scale, real-time 3D reconstruction and streaming for immersive multi-client live telepresence. *IEEE Trans. Vis. Comput. Graph.* **2019**, *25*, 2102–2112. [CrossRef]
13. Ishigaki, S.A.K.; Ismail, A.W. Real-time 3D reconstruction for mixed reality telepresence using multiple depth sensors. In Proceedings of the International Conference on Advanced Communication and Intelligent Systems, Virtual, 20–21 October 2022; pp. 67–80.
14. Fadzli, F.E.; Ismail, A.W.; Abd Karim Ishigaki, S. A systematic literature review: Real-time 3D reconstruction method for telepresence system. *PLoS ONE* **2023**, *18*, e0287155. [CrossRef] [PubMed]
15. Cao, C.; Preda, M.; Zaharia, T. 3D point cloud compression: A survey. In Proceedings of the 24th International Conference on 3D Web Technology, Los Angeles, CA, USA, 26–28 July 2019; pp. 1–9.
16. Liu, H.; Yuan, H.; Liu, Q.; Hou, J.; Liu, J. A comprehensive study and comparison of core technologies for MPEG 3-D point cloud compression. *IEEE Trans. Broadcast.* **2019**, *66*, 701–717. [CrossRef]
17. Nardo, F.; Peressoni, D.; Testolina, P.; Giordani, M.; Zanella, A. Point cloud compression for efficient data broadcasting: A performance comparison. In Proceedings of the 2022 IEEE Wireless Communications and Networking Conference (WCNC), Austin, TX, USA, 10–13 April 2022; pp. 2732–2737.
18. Bletterer, A.; Payan, F.; Antonini, M.; Meftah, A. Point Cloud Compression using Depth Maps. *Electron. Imaging* **2016**, *28*, art00005. [CrossRef]

19. Quach, M.; Pang, J.; Tian, D.; Valenzise, G.; Dufaux, F. Survey on deep learning-based point cloud compression. *Front. Signal Process.* **2022**, *2*, 846972. [CrossRef]
20. Pece, F.; Kautz, J.; Weyrich, T. Adapting standard video codecs for depth streaming. In Proceedings of the EGVE/EuroVR, Nottingham, UK, 20–21 September 2011; pp. 59–66.
21. Jäger, F. Contour-based segmentation and coding for depth map compression. In Proceedings of the 2011 Visual Communications and Image Processing (VCIP), Tainan, Taiwan, 6–9 November 2011; pp. 1–4.
22. Kumar, S.H.; Ramakrishnan, K. Depth compression via planar segmentation. *Multimed. Tools Appl.* **2019**, *78*, 6529–6558. [CrossRef]
23. Duch, M.M.; Morros, J.R.; Ruiz-Hidalgo, J. Depth map compression via 3D region-based representation. *Multimed. Tools Appl.* **2017**, *76*, 13761–13784. [CrossRef]
24. Wilson, A.D. Fast lossless depth image compression. In Proceedings of the 2017 ACM International Conference on Interactive Surfaces and Spaces, Brighton, UK, 17–20 October 2017; pp. 100–105.
25. Sonoda, T.; Grunnet-Jepsen, A. Depth Image Compression by Colorization for Intel RealSense Depth Cameras. Intel Rev. 1.0. 2021. Available online: https://dev.intelrealsense.com/docs/depth-image-compression-by-colorization-for-intel-realsense-depth-cameras (accessed on 26 April 2024).
26. Chen, M.; Zhang, P.; Chen, Z.; Zhang, Y.; Wang, X.; Kwong, S. End-to-end depth map compression framework via rgb-to-depth structure priors learning. In Proceedings of the 2022 IEEE International Conference on Image Processing (ICIP), Bordeaux, France, 16–19 October 2022; pp. 3206–3210.
27. D'Amato, J.P. FitDepth: Fast and lite 16-bit depth image compression algorithm. *EURASIP J. Image Video Process.* **2023**, *2023*, 5. [CrossRef]
28. Zanuttigh, P.; Cortelazzo, G.M. Compression of depth information for 3D rendering. In Proceedings of the 2009 3DTV Conference: The True Vision-Capture, Transmission and Display of 3D Video, Potsdam, Germany, 4–6 May 2009; pp. 1–4.
29. Krishnamurthy, R.; Chai, B.B.; Tao, H.; Sethuraman, S. Compression and transmission of depth maps for image-based rendering. In Proceedings of the 2001 International Conference on Image Processing (Cat. No. 01CH37205), Thessaloniki, Greece, 7–10 October 2001; Volume 3, pp. 828–831.
30. Boopathiraja, S.; Punitha, V.; Kalavathi, P.; Prasath, V.S. Computational 2D and 3D medical image data compression models. *Arch. Comput. Methods Eng.* **2022**, *29*, 975–1007. [CrossRef] [PubMed]
31. Itti, L.; Koch, C.; Niebur, E. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **1998**, *20*, 1254–1259. [CrossRef]
32. Koch, C.; Ullman, S. Shifts in selective visual attention: Towards the underlying neural circuitry. *Hum. Neurobiol.* **1985**, *4*, 219–227. [PubMed]
33. Liu, T.; Yuan, Z.; Sun, J.; Wang, J.; Zheng, N.; Tang, X.; Shum, H.Y. Learning to detect a salient object. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *33*, 353–367.
34. Wang, W.; Shen, J. Deep visual attention prediction. *IEEE Trans. Image Process.* **2017**, *27*, 2368–2378. [CrossRef]
35. Cadoni, M.; Lagorio, A.; Khellat-Kihel, S.; Grosso, E. On the correlation between human fixations, handcrafted and CNN features. *Neural Comput. Appl.* **2021**, *33*, 11905–11922. [CrossRef]
36. Cadoni, M.; Lagorio, A.; Grosso, E. Face detection based on a human attention guided multi-scale model. *Biol. Cybern.* **2023**, *117*, 453–466. [CrossRef] [PubMed]
37. Viola, P.; Jones, M.J. Robust real-time face detection. *Int. J. Comput. Vis.* **2004**, *57*, 137–154. [CrossRef]
38. Massone, L.; Sandini, G.; Tagliasco, V. "Form-invariant" topological mapping strategy for 2D shape recognition. *Comput. Vis. Graph. Image Process.* **1985**, *30*, 169–188. [CrossRef]
39. Weiman, C.F. Video compression via log polar mapping. In Proceedings of the Real-Time Image Processing II, Orlando, FL, USA, 16–20 April 1990; SPIE: Bellingham, WA, USA, 1990; Volume 1295, pp. 266–277.
40. Traver, V.J.; Bernardino, A. A review of log-polar imaging for visual perception in robotics. *Robot. Auton. Syst.* **2010**, *58*, 378–398. [CrossRef]
41. Bicego, M.; Grosso, E.; Lagorio, A.; Brelstaff, G.; Brodo, L.; Tistarelli, M. Distinctiveness of faces: A computational approach. *ACM Trans. Appl. Percept.* **2008**, *5*, 1–18. [CrossRef]
42. Bruce, N.D. Features that draw visual attention: An information theoretic perspective. *Neurocomputing* **2005**, *65–66*, 125–133. [CrossRef]
43. Tian, D.; Ochimizu, H.; Feng, C.; Cohen, R.; Vetro, A. Geometric distortion metrics for point cloud compression. In Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP), Beijing, China, 17–20 September 2017; pp. 3460–3464.
44. Alexiou, E.; Ebrahimi, T. Towards a point cloud structural similarity metric. In Proceedings of the 2020 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), London, UK, 6–10 July 2020; pp. 1–6.
45. Zhang, Z.; Sun, W.; Zhu, Y.; Min, X.; Wu, W.; Chen, Y.; Zhai, G. Evaluating point cloud from moving camera videos: A no-reference metric. *IEEE Trans. Multimed.* 2023, *early access*. [CrossRef]

46. Zhang, Z.; Sun, W.; Min, X.; Zhou, Q.; He, J.; Wang, Q.; Zhai, G. MM-PCQA: Multi-modal learning for no-reference point cloud quality assessment. *arXiv* **2022**, arXiv:2209.00244.
47. Yang, Q.; Chen, H.; Ma, Z.; Xu, Y.; Tang, R.; Sun, J. Predicting the perceptual quality of point cloud: A 3d-to-2d projection-based exploration. *IEEE Trans. Multimed.* **2020**, *23*, 3877–3891. [CrossRef]