Occlusion-Aware Volumetric Video Streaming

for Bandwidth-Efficient 3D Viewing

Group 12

D13949003 Mohammadreza Kamrani, M11207327 吳忠翰

1. Introduction

Volumetric video, often represented as point clouds, enables immersive 6-Degrees-of-Freedom (6DoF) experiences in Mixed Reality (MR) headsets. Unlike traditional 2D videos, volumetric content allows users to view scenes from any angle, providing a highly realistic sense of presence. However, streaming high-fidelity volumetric content presents significant challenges. It requires extremely high bandwidth—often exceeding 3.6 Gbps for 1 million points per frame—and substantial computational power.

Current mobile MR headsets, such as the HoloLens 2, have limited compute resources and battery life, making real-time decoding and rendering of raw point clouds difficult. Without optimization, users suffer from low visual quality, high latency, and poor Quality of Experience (QoE).

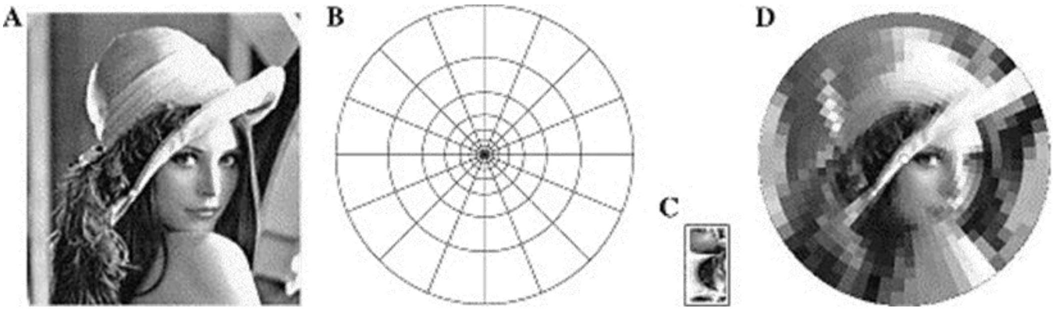


Fig. 1. Log-Polar Transformation simulates human foveated vision [1]

To address these issues, we propose a bandwidth-efficient streaming system inspired by "Theia". Our goal is to enable high-quality, low-bandwidth volumetric streaming on mobile headsets. We adopt a "Foveated Streaming" strategy, which leverages the human visual system's characteristics by streaming high-quality content only to the foveal area [1] (where the user is looking) while reducing quality in the peripheral vision. Additionally, we implement an occlusion-aware mechanism to ignore hidden points, further reducing unnecessary data transmission.

2. Methodology

Our methodology focuses on mapping 3D point clouds to a 2D space for efficient filtering and then back to 3D for rendering. Before detailing the transformation pipeline, we first introduce the dataset used to benchmark our system and define the bandwidth challenge.

2.1. Dataset

To evaluate the performance of our system on high-fidelity volumetric content, we utilized the industry-standard 8i Voxelated Full Bodies (8i VFB) dataset [2]. Specifically, we selected two sequences representing different geometric characteristics:

- LongDress: Characterized by complex geometry and self-occlusions due to the flowing dress.
- Loot: Characterized by high surface detail and texture.

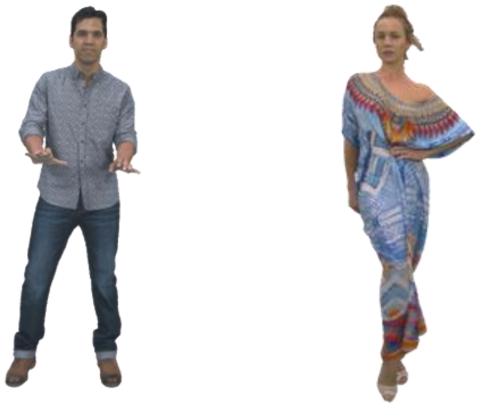


Fig. 2. The Loot and LongDress Dataset [2]

These datasets represent a significant challenge for mobile streaming due to their high density. Each frame contains approximately 800,000 points (834K for LongDress, 794K for Loot). Without compression, the raw bitrate for transmission exceeds 2.8 Gbps (3,002 Mbps for LongDress, 2,858 Mbps for Loot). This massive bandwidth requirement serves as the baseline for evaluating our compression efficiency.

2.2. Method

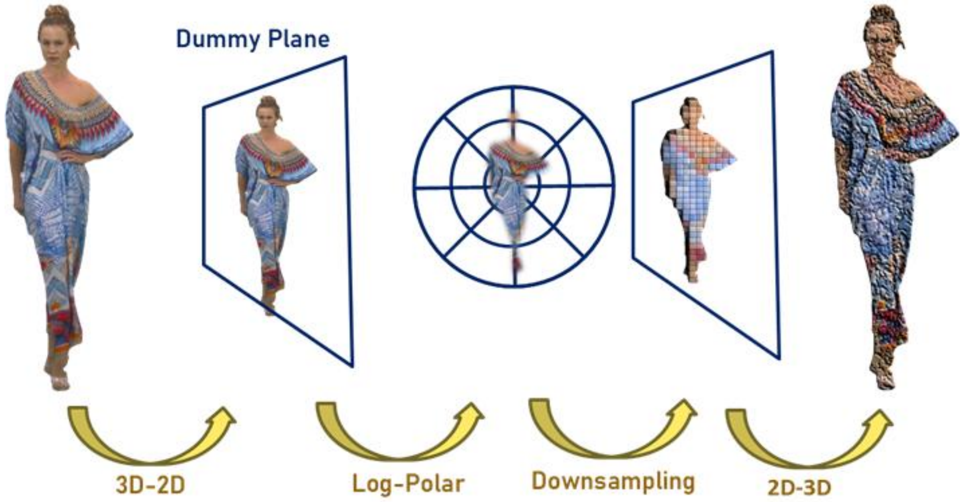


Fig. 3. Overview of the proposed coordinate transformation pipeline

2.2.1. Log-Polar Transformation

To simulate human vision and reduce data size, we utilize a Log-Polar transformation. This transformation maps the 3D points onto a 2D "Dummy Plane" relative to the user's gaze direction. The transformation allows us to maintain high sampling density in the center (fovea) and exponentially decrease density towards the periphery. Mathematically, for a point P_{Point} and a gaze origin P_{Gaze}, we calculate the relative position $P_{Relative}$:

$$P_{Relative} = P_{Point} - P_{Gaze}$$

We then project this onto a 2D plane based on the gaze direction D_{gaze}. The Log-Polar coordinates (u, v) are computed as follows:

$$u = \log_b \frac{\rho}{\tan(MAR_0)}$$

$$v = \frac{\theta}{2\pi W}$$

Where ρ is the radial distance $\sqrt{x^2 + y^2}$, θ is the angular coordinate arctan $(\frac{y}{x})$, and MAR0 (Minimum Angle of Resolution) is set to 1 arcminute, and b we set as 1.002. This mathematical mapping ensures that points in the peripheral vision are aggressively downsampled, while central points are preserved.

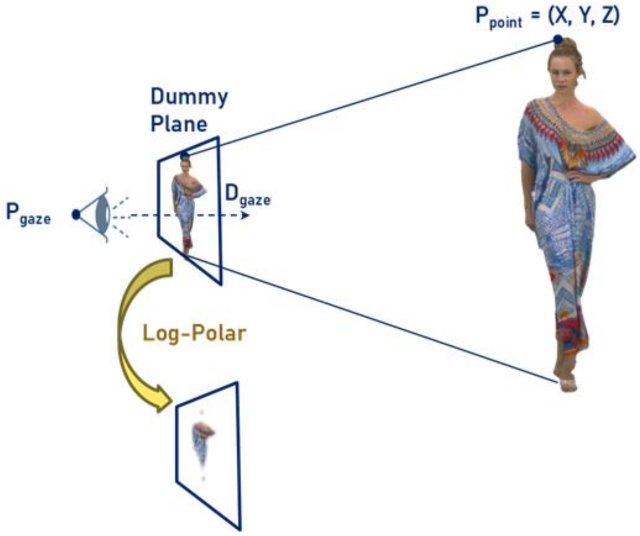


Fig. 4. Geometric illustration of the gaze-driven projection model

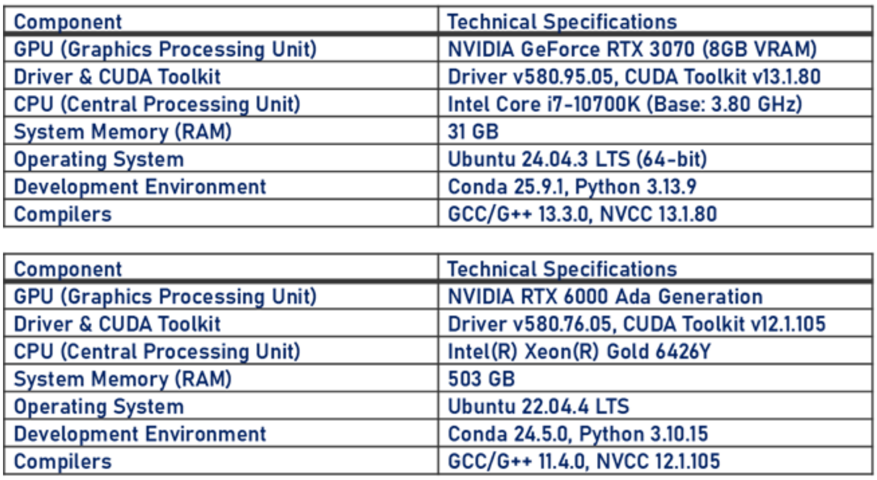
2.2.2. Occlusion Awareness (Implicit HPR)

A major inefficiency in traditional streaming is transmitting points that are occluded (blocked by other points) from the user's perspective. Our system implements Hidden Point Removal (HPR) implicitly through the 3D-2D mapping process. When multiple 3D points map to the same 2D pixel in the Log-Polar grid, only the point closest to the camera is retained. This effectively filters out hidden points without complex geometric calculations.

3. Result

We evaluated our system using the 8i Voxelized Full Bodies dataset. To demonstrate the adaptability of our Log-Polar sampling strategy, we conducted experiments under two distinct parameter configurations: High-Efficiency Mode (maximizing compression) and High-Fidelity Mode (maximizing visual detail).

3.1. System Specifications for Implementation



Component	Technical Specifications
GPU (Graphics Processing Unit)	NVIDIA GeForce RTX 3070 (8GB VRAM)
Driver & CUDA Toolkit	Driver v580.95.05, CUDA Toolkit v13.1.80
CPU (Central Processing Unit)	Intel Core i7-10700K (Base: 3.80 GHz)
System Memory (RAM)	31 GB
Operating System	Ubuntu 24.04.3 LTS (64-bit)
Development Environment	Conda 25.9.1, Python 3.13.9
Compilers	GCC/G++ 13.3.0, NVCC 13.1.80

Component	Technical Specifications
GPU (Graphics Processing Unit)	NVIDIA RTX 6000 Ada Generation
Driver & CUDA Toolkit	Driver v580.76.05, CUDA Toolkit v12.1.105
CPU (Central Processing Unit)	Intel(R) Xeon(R) Gold 6426Y
System Memory (RAM)	503 GB
Operating System	Ubuntu 22.04.4 LTS
Development Environment	Conda 24.5.0, Python 3.10.15
Compilers	GCC/G++ 11.4.0, NVCC 12.1.105

Fig. 5. System Specifications for Implementation

3.2. Extreme Compression (High-Efficiency Mode)

we applied aggressive foveation parameters (higher rate_adapt and lower r_bins) on the "LongDress" sequence.

- Data Reduction: The input frame contained 765,821 points, while the output was reduced to 24,704 points.
- Compression Ratio: This achieved a massive 31:1 compression ratio.
- Performance: The processing time was 11 ms, enabling extremely low-latency transmission (< 90 Mbps) suitable for congested mobile networks (ran on 3070).

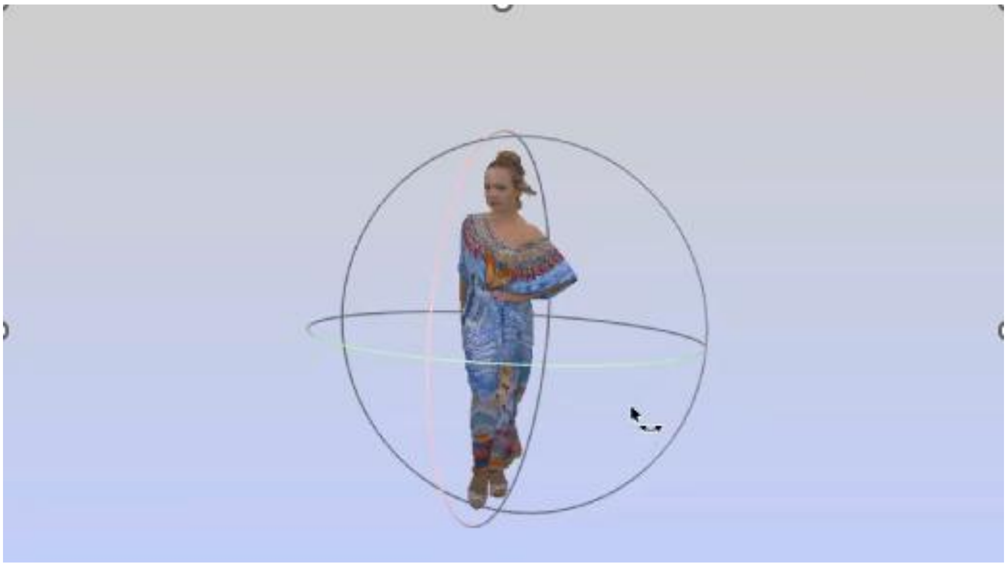


Fig. 6. Original LongDress Point Cloud Dataset

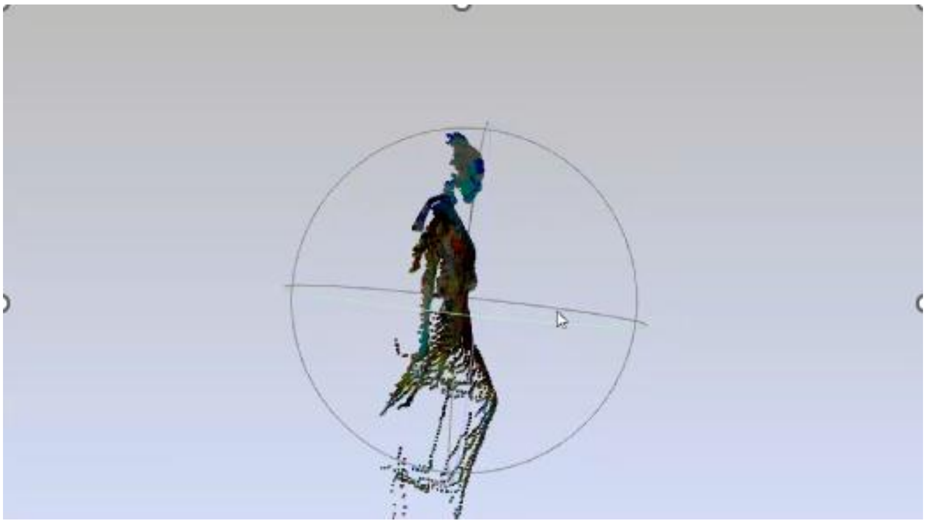


Fig. 7. LongDress Dataset after Hidden Point Removal

3.3. High Fidelity (Quality-First Mode)

In contrast, the "Loot" sequence processed with parameters tuned for quality.

- Detail Preservation: From an input of 784,142 points, the system retained 170,959 points.
- Compression Ratio: The ratio is 4.6:1, retaining significantly more visual information in the para-foveal and peripheral regions.
- Performance: The processing time remained extremely fast at 5.65 ms (ran on 6000 ada).

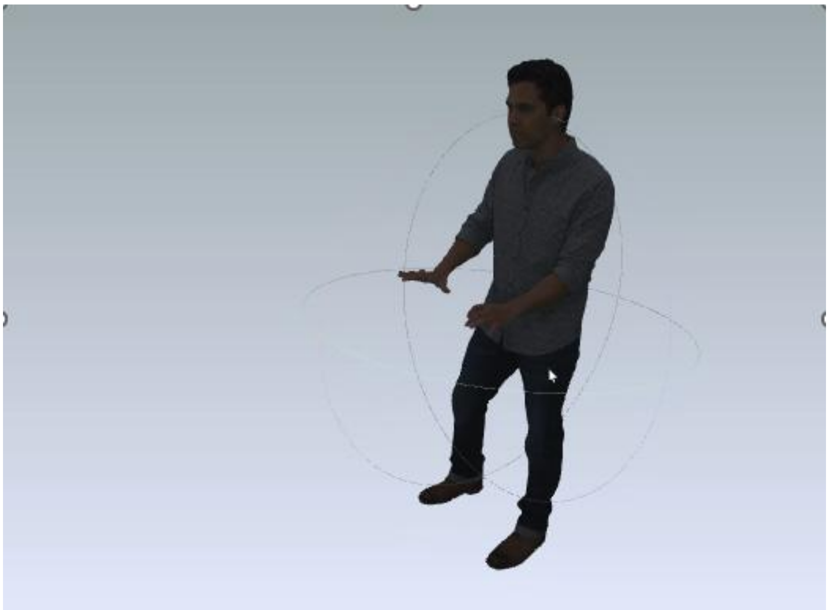


Fig. 8. Original Loot Point Cloud Dataset

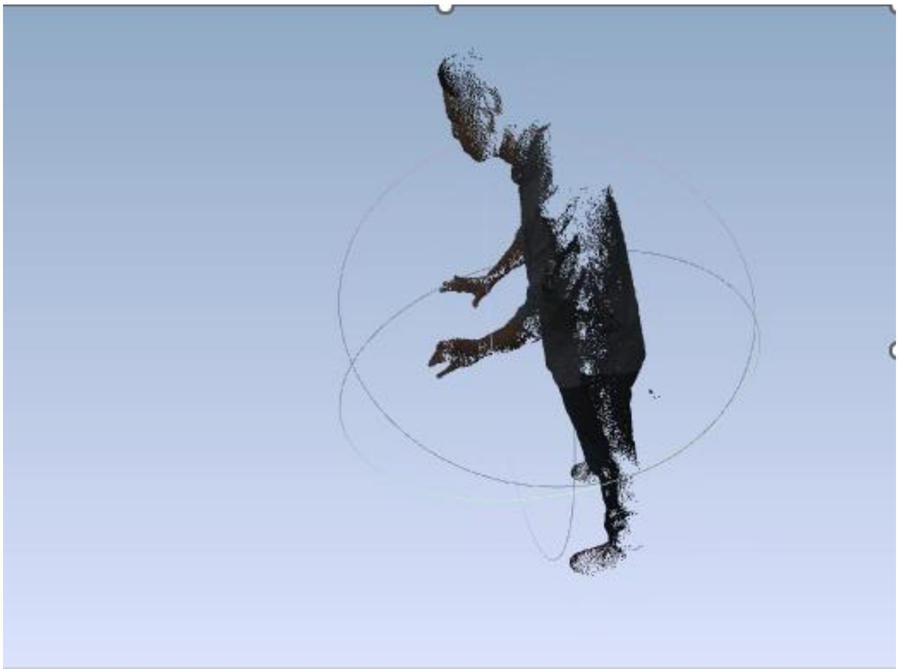


Fig. 9. Loot Dataset after Hidden Point Removal

These two experiments illustrate the system's flexibility. By adjusting the Log-Polar grid density (r_bins) and the decay rate (rate_adapt), users or applications can dynamically trade-off between bandwidth savings and visual resolution.

3.4. Quality Perception and Resource Allocation Strategy

To balance visual fidelity with bandwidth constraints, we implemented a multi-zone quality strategy based on the human visual system's acuity fall-off. We partitioned the user's Field of View (FoV) into three distinct concentric regions: Foveal, Para-foveal, and Peripheral.

Distinct parameters were assigned to each region:

- Foveal Region (Central): This covers the immediate gaze center 0.75 degrees. Here, we prioritize quality over compression, aiming for a Target PSNR of 45-50 dB (Near lossless). Although we apply a moderate point reduction (50-60%), this region is allocated a significant portion of the bitrate (30-40%) relative to its small size to ensure critical details are preserved.

- Para-foveal Region (Middle): Spanning 2.5 degrees to 3.75 degrees, this transition zone maintains structural context. We apply a high point reduction rate (85-90%) to achieve a "Good" quality rating (38-42 dB).
- Peripheral Region (Outer): Extending up to 15 degrees, this region occupies the largest visual area. To prevent bandwidth saturation, we apply extreme compression, removing 95-98% of the points3. Despite this reduction, due to the vast volumetric space it covers, it still utilizes 35-45% of the total bitrate to maintain acceptable spatial awareness (30-35 dB).

3.5. Temporal Optimization: Differential Frame Streaming

Beyond spatial compression using Log-Polar transformation, we further reduced bandwidth consumption by exploiting temporal redundancy between consecutive frames. In a typical volumetric video, a significant portion of the geometry remains static or exhibits only minor displacements due to sensor noise. To address this, we implemented a Distance Thresholding mechanism.

We calculate the Euclidean distance of corresponding points between the current frame (t) and the previous frame (t-1). If the displacement is below a specific threshold (∈), the point is considered static and is not transmitted.

- Baseline (Threshold = 0): All 24,723 spatially compressed points are transmitted.
- Noise Filtering (Threshold = 0.001): By filtering out minor sensor jitter, we achieve a 10.6% reduction in data size without affecting visual motion.
- Aggressive Optimization (Threshold = 0.005): Increasing the threshold drastically reduces the point count to 3,765, achieving an 84.7% savings rate. This setting is ideal for extremely low-bandwidth scenarios where transmitting only major movements is prioritized.

4. References

[1] Bo Han, Yu Liu, and Feng Qian. ViVo: Visibility-Aware Mobile Volumetric Video Streaming. In The 26th Annual International Conference on Mobile Computing and Networking (MobiCom '20). Association for Computing Machinery, NY, USA, Article 16, 1–14.

[2] Yu Liu, Bo Han, Feng Qian, Arvind Narayanan, and Zhi-Li Zhang. Vues: Practical Mobile Volumetric Video Streaming Through Multiview Transcoding. In The 28th Annual International Conference on Mobile Computing and Networking (MobiCom '22). Association for Computing Machinery, NY, USA, 367–381.

[3] Yongjie Guan, Xueyu Hou, Nan Wu, Bo Han, and Tao Han. MetaStream: Live Volumetric Content Capture, Creation, Delivery, and Rendering in Real Time. In The 29th Annual International Conference on Mobile Computing and Networking (ACM MobiCom '23). Association for Computing Machinery, NY, USA, 1–15

[4] Zhongzheng Yuan, Yifei Zhu, Yuye Zhang, Zhigi Peng, Wei Tsang Ooi, and Yunxin Liu. Theia: Gaze-driven and Perception-aware Volumetric Content Delivery for Mixed Reality Headsets. In Proceedings of the 22nd Annual International Conference on Mobile Systems, Applications and Services (MobiSys '24). Association for Computing Machinery, NY, USA, 302–315