Evaluating the Impact of Tiled User-Adaptive Real-Time Point Cloud Streaming on VR Remote Communication

Shishir Subramanyam
Irene Viola
Jack Jansen
Evangelos Alexiou
s.subramanyam@cwi.nl
Centrum voor Wiskunde en
Informatica
Amsterdam, Netherlands

Alan Hanjalic
a.hanjalic@tudelft.nl
TU Delft
Delft, Netherlands

Pablo Cesar
p.s.cesar@cwi.nl
Centrum voor Wiskunde en
Informatica
Amsterdam, Netherlands
TU Delft
Delft, Netherlands

ABSTRACT

Remote communication has rapidly become a part of everyday life in both professional and personal contexts. However, popular video conferencing applications present limitations in terms of quality of communication, immersion and social meaning. VR remote communication applications offer a greater sense of co-presence and mutual sensing of emotions between remote users. Previous research on these applications has shown that realistic point cloud user reconstructions offer better immersion and communication as compared to synthetic user avatars. However, photorealistic point clouds require a large volume of data per frame and are challenging to transmit over bandwidth-limited networks. Recent research has demonstrated significant improvements to perceived quality by optimizing the usage of bandwidth based on the position and orientation of the user's viewport with user-adaptive streaming. In this work, we developed a real-time VR communication application with an adaptation engine that features tiled user-adaptive streaming based on user behaviour. The application also supports traditional network adaptive streaming. The contribution of this work is to evaluate the impact of tiled user-adaptive streaming on quality of communication, visual quality, system performance and task completion in a functional live VR remote communication system. We performed a subjective evaluation with 33 users to compare the different streaming conditions with a neck exercise training task. As a baseline, we use uncompressed streaming requiring approximately 300 megabits per second and our solution achieves similar visual quality with tiled adaptive streaming at 14 megabits per second. We also demonstrate statistically significant gains in the quality of interaction and improvements to system performance and CPU consumption with tiled adaptive streaming as compared to the more traditional network adaptive streaming.

CCS CONCEPTS

• **Information systems** → **Multimedia streaming**; • **Human-centered computing** → **Virtual reality**.

KEYWORDS

Virtual Reality, Remote Communication, 3D Point Clouds, Adaptive Streaming

1 INTRODUCTION

Remote communication and collaboration have rapidly become a necessity in a globalized and connected world. Video conferencing applications have become ubiquitous in everyday life in both professional and personal environments. Notwithstanding their popularity, it is estimated that travel for the purpose of in-person communication is responsible for roughly eight percent of US energy consumption [35]. With recent events like the Covid-19 global pandemic, there is an increased need for applications that can deliver a greater sense of co-presence and mutual sensing of emotions in remote communication. Current video conferencing solutions have clear limitations in this regard [7, 26, 31, 55].

Immersive Virtual Reality (VR) applications offer an increased sense of presence and immersion. These applications have emerged as a promising alternative for remote communication and telepresence [6, 9, 18, 24, 32, 36, 52, 54]. They allow users to employ both verbal and non-verbal communication in a shared virtual space. In such applications, users can be embodied in the virtual space either using avatars or real-time photorealistic 3D reconstructions typically using depth sensors. Previous work in the field has demonstrated that realistic user reconstructions improve immersion and communication [23, 33] as compared to avatars.

Among the different 3D formats, point clouds have emerged as a popular representation for user reconstructions as they are relatively easy to acquire in real-time using consumer depth sensors [9, 18]. This format represents the object's geometry as an unorganized collection of surface point coordinates with additional attributes such as color provided at each point location. They are generally resilient to noise and do not incur an additional computational overhead to triangulate mesh faces making them suitable for real-time applications. Owing to their unorganized nature they

are also easy to partition into non-overlapping segments. However, photorealistic point clouds present a large volume of data per frame that requires real-time compression to transmit over bandwidth-limited networks [29, 50]. In order to alleviate this requirement, recent research has looked into adapting the point cloud stream to the user's viewport location and orientation in order to optimize how the available bandwidth is spent. This is done by prioritizing objects or surfaces facing the viewer and reducing the wastage of bandwidth on surfaces or objects that are occluded or outside the viewport [8, 14, 37, 47, 48, 50, 51]. However, this approach to adaptive streaming has not yet been evaluated for live communication with real-time point cloud reconstructions.

In this work, we present the first implementation and subjective evaluation of a two-user social VR system with real-time adaptive streaming of user reconstructions shown in Figure 1. We set out to assess the impact of tiled adaptive streaming on the quality of communication, visual quality and subjective task related performance. We constructed a social VR pipeline extending the system proposed in [18] with network adaptive and tiled adaptive streaming. Two confederate users were recruited and trained to play the role of a trainer in every experiment session while 33 users (16 females, 17 males) were recruited to play the role of trainee. The participants were asked to learn and perform three neck exercises during the session. The contribution of this work is to evaluate and compare tiled adaptive streaming (TA) with traditional network adaptive streaming (NA) and baseline uncompressed streaming in a functional live VR remote communication system. We propose and employ a novel evaluation methodology using a training task to perform the assessment. We address the following four research questions for VR remote communication:

- R1: How does tiled user-adaptive point cloud streaming impact quality of interaction/ quality of communication *(QoI)*?
- R2: How does tiled user-adaptive point cloud streaming impact the experience of performing a training task?
- R3: How does tiled user-adaptive point cloud streaming impact the perceived quality of remote user reconstruction?
- R4: What is the computational overhead of using tiled adaptive point cloud streaming?

From our results, we observe statistically significant improvements to QoI (R1). We observe no statistically significant change to task experience (R2). We observe significant improvements to visual quality (R3) and at 14Mbps we observe similar visual quality to uncompressed point clouds streamed at ca. 300Mbps. We also see a reduction in CPU utilization and an improvement to playback performance (R4). We validated the communication system and checked that the training task provides coherent results in the evaluation.

2 RELATED WORK

2.1 VR Remote Communication using Point Clouds

Advances in low-latency streaming and volumetric point cloud delivery mechanisms have led to the emergence of novel teleimmersion systems that allow distributed remote users to communicate as themselves in a shared environment with realistic user reconstructions. Microsoft released the RoomAlive Toolkit for creating interactive Augmented Reality (AR) experiences [19, 52]. Mekuria *et al.* proposed a teleimmersive system that blends avatar representations and photo-realistic reconstructions of users in a shared virtual environment [33]. Gunkel *et al.* introduced VRComm [9], a web-based social VR communication system using photo-realistic user reconstructions that was evaluated using both simulations and subjective studies. They present a video-based transmission approach as an initial step toward delivering volumetric content. Cernigliaro *et al.* propose a point cloud multi-point control unit for optimizing holo-conferencing systems [2]. Jansen *et al.* [18] proposed a pipeline for volumetric videoconferencing using low latency DASH with photo-realistic point cloud user reconstructions. None of these systems implements real-time tiled adaptive streaming of live-captured point clouds. **In this work**, we present a system design with network adaptive *(NA)* and tiled user-adaptive *(TA)* streaming. We transmit tiled point cloud user reconstructions at fixed target bitrates to assess the experience without the influence of a volatile network.

2.2 Point Cloud Delivery

2.2.1 Compression Standards. Point cloud compression has received significant research attention in recent years with the launch of two new MPEG compression standards [44]. The V-PCC standard codec for dynamic point clouds that projects point clouds geometry and attributes onto separate 2D patches that are them packed into video tracks along with the occupancy. These video tracks are then encoded using legacy video codecs making this approach suitable for relatively dense and uniform distribution of points. The G-PCC standard codec uses an octree space partitioning structure to code geometry and can be optionally combined with an additional surface reconstruction step using the TriSoup approach [38]. G-PCC also includes several modules for attribute coding, the lowest complexity coder uses the Region Adaptive Hierarchical Transform (RAHT) [4]. This codec is targeted at irregular sparse distributions of points making it suitable for live captured point clouds. However, both codecs introduce high complexity encoding making them unsuitable for real-time communication. At the start of the MPEG standardization activity an anchor codec proposed by Mekuria *et al.* [32] was introduced. This codec utilizes octree occupancy to code geometry and scans attributes to map them to a 2D grid and encodes them with legacy JPEG image compression. This approach offers low encode and low decode complexity making it suitable for real-time framerate sensitive applications such as VR remote communication. **In this work**, we use the anchor codec to encode point cloud tiles at multiple quality levels in real-time before streaming.

2.2.2 Adaptive Streaming. Initial works on adaptive streaming of point clouds utilized entire point cloud objects as the basic unit of bandwidth allocation in scenes containing multiple point cloud objects. Hosseini *et al.* [13, 14] propose DASH-PC for dynamic adaptive view aware point cloud streaming. They propose three spatial subsampling techniques to create multiple representations of point cloud objects in a scene. The density of each object representation is used by the client for bitrate allocation based on human visual acuity. Hooft *et al.* [50] propose PCC-DASH, a standards-compliant

Table 1: System Setup

Hardware	HMD	HTC Vive Pro Eye		
	CPU	Intel(R) Core(TM) i7-7700K CPU @ 4.20GHz (8 CPUs), ~4.2GHz		
	GPU	NVIDIA GeForce GTX 1080 Ti		
	Memory	32768MB RAM		
	Depth Sensors	3 x Azure Kinect DK		
Display Parameters	Resolution	1440 x 1600 pixels per eye		
	Application Target Framerate	90 Hz		
	Point Cloud Playback Target Framerate	15 Hz		
Fixed System Parameters	Audio Codec	Ogg Speex 48 KHz		
	Point Cloud Codec Configurations	Octree Depth 6	Octree Depth 7	Octree Depth 9
	Kinect Depth Configuration	NFOV unbinned 640x576		
	Kinect Color Configurations	1280x720		
	Point Cloud Capture	ca. 130k points per frame at 15 fps		
Conditions	Target Bitrates	7 Mbps	14 Mbps	
	Streaming Conditions	Network Adaptive	Tiled Adaptive	Uncompressed

means for HTTP adaptive streaming. They present three heuristics based on the users viewport and distance to the object to allocate bitrate to different objects in the scene. Different ranking metrics and bitrate allocation heuristics had to be selected for different scenes and user navigation paths.

Another approach used in previous work is to split each point cloud object into tiles that are then used as the unit of bandwidth allocation. Park *et al.* [37] define a utility per tile based on the user's proximity, point cloud surface quality and display device resolution. To account for interactions, they propose a window-based design for the Client Buffer Manager with greedy utility maximization. This type of rasterization and pixel occupancy based approach is not suitable as computing this at every frame is computationally expensive. He *et al.* [11] propose view-dependent streaming over hybrid networks. Each point cloud frame is projected onto the six faces of a bounding cube, with a color and a depth video created per face. The user can request videos that correspond to particular faces of the cube in high quality from the edge node of a bidirectional broadband network, reconstructing the point cloud from the downloaded depth and color videos at the receiver end. This approach requires a redundant extra reconstruction step at each receiver. We instead perform the reconstruction on the sender side in order to generate a self view to embody the user and transmit the reconstruction to all receivers. Li *et al.* [27] propose a joint communication and computational resource allocation framework to stream and decode pre-recorded point clouds. They also propose a QoE metric to guide tile selection based on the users viewport, distance to tile and available quality levels. Lee *et al.* [25] propose GROOT a real-time streaming system to reduce decoding overhead by dividing the point cloud into cells defined by the leaf nodes of an octree represented in a parallel decodable tree. Han *et al.* [10] propose ViVo using a similar approach and employ machine learning models to predict viewport movement. Liu *et al.* [28] follow a similar approach, they include an uncompressed base layer and use fuzzy logic based quality selection. This type of approach using the leaf nodes of the octree as an enhancement layer is currently not suitable for real-time systems as it adds an extra surface orientation estimation step that introduces additional delays in the

pipeline. Subramanyam *et al.* [48] build on the ideas presented in PCC-DASH [50] to tile point cloud content using low complexity surface estimation suitable for frame rate sensitive real-time applications. They performed an objective quality evaluation using prerecorded navigation paths and image distortion metrics. **In this work**, we build on their approach and create tiles based on surfaces visible to multiple depth sensors and estimate their orientation using the transformation matrix associated with each sensor.

2.3 VR Communication Evaluation

Quality assessment for remote communication is usually conducted using subjective user studies that are either passive or active. Passive tests involve asking users to rate prerecorded clips of content. This approach to evaluation is more suited for standardized testing of codecs with offline content and has limited ecological validity in remote communication [15, 43]. Active tests involve multiple remote participants being placed in an interactive live communication system. The International Telecommunication Union published recommendations to define evaluation methods for quantifying the impact of terminal and communication link performance on point-to-point audiovisual communication [16]. The recommendation contains sample tasks such as name guessing, story comparison, picture comparison, object description and building blocks. Schmitt *et al.* [43] utilize the building blocks task to develop and evaluate personalized quality of experiment metrics for multiparty video conferencing at varying bitrates. Smith *et al.* [46] compare face-to-face communication with embodied and unembodied remote VR communication. They propose a task involving negotiating apartment layout and furniture placement based on blueprints. Li *et al.* [26] compare face-to-face, videoconferencing and Facebook spaces VR communication in the context of a photosharing task. They found that Facebook spaces is able to closely approximate face-to-face photosharing. In general, these methods have been used to compare VR remote communication with other technologies and with face to face communication. The tasks proposed either focus on the audio quality or rely on external objects for the evaluation. **In this work**, we focus on evaluating adaptive streaming within VR remote communication. We define a new visually focused training
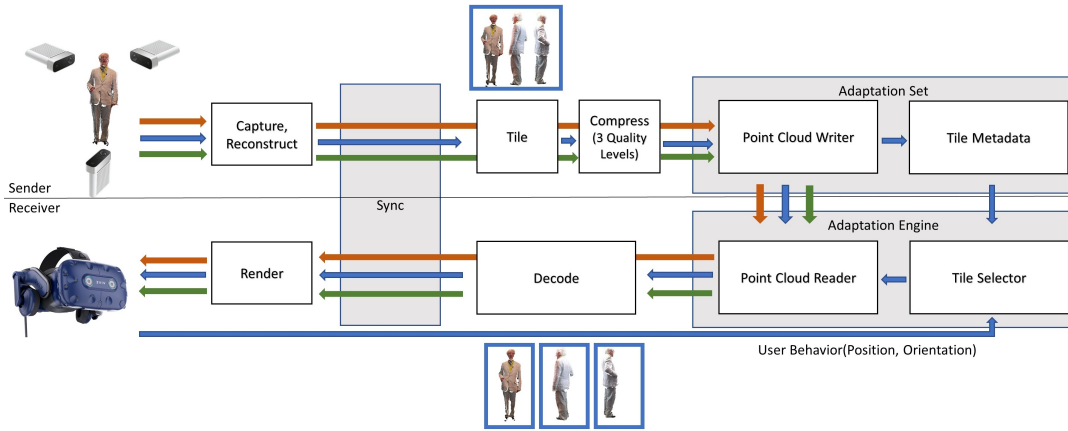
Figure 1: Architecture and Data Flow for Baseline (Orange), NA (Green) and TA (Blue) Streaming

task where participants are taught a neck exercise from a trainer and are asked to perform the exercise in order to complete the task.

In order to evaluate communication, several questionnaires have been proposed in the literature. Toet *et al.* [49] propose the holistic mediated social communication (H-MSC) framework and associated questionnaire to evaluate the experience of spatial presence as well as social presence. The framework is general enough to be used for any mediated social communication system. Slater *et al.* [45] and Witmer *et al.* [53] have proposed two popular questionnaires aimed at measuring presence in virtual environments. Kangas *et al.* [20] present a pragmatic task related questionnaire that they use to evaluate VR interaction techniques in a rigid object manipulation task. Li *et al.* [26] propose a social VR questionnaire that evaluates Quality of Interaction/Communication *(QoI)*, Presence/Immersion and Social Meaning. **In this work**, we combine the QoI part of this questionnaire along with ITU visual quality questions [16], and task related questions from the pragmatic task related questionnaire [20].

3 STUDY DESIGN

3.1 Experiment Task

The goal was to define a task that could be used to evaluate QoI, visual quality and task experience in VR remote communication. We needed to facilitate a conversation with a fixed general outline and with a focus on the visual aspect of communication. We considered the tasks presented in ITU-T P.920 [16], the building blocks task presented in [43] and the photosharing task presented in [26]. We found that these tasks are either heavily dependent on the audio quality (story comparison, name-guessing and object description task) or required external objects that had to either be live-captured along with the user or digitally represented in the virtual world with appropriate interaction tools (picture comparison, building blocks, photo-sharing). These objects could occlude and distract from the visual quality of the user reconstruction. In order to assess the impact of tiled adaptive point cloud streaming, we chose to keep the audio quality consistent across sessions and manipulated the quality of the point cloud representation of the participants. Based on pre-trials with seven colleagues, we selected a training

task where participants were asked to first learn and then perform a neck exercise. This allowed for a fixed general outline of the conversation and focused on the visual representation of the remote participant. The neck exercises were found not to induce severe motion sickness as validated in section 4 and provided coherent results. In our experiment design, we use a confederate user as the trainer. In this approach, all participants (trainees) in the same condition are always paired with the same trainer. This allows us to focus on the individual as the unit of study and we mitigate social context as a confounding variable. In addition, we isolate the basic elements of communication by attempting to hold constant the behavior of one conversation partner. In order to adhere to the recommendations on using confederate users [22], we used an asymmetric communication channel. The trainers are always shown the same quality point clouds of the trainee (octree depth 9) in order to ensure that the confederate users are as naive as possible to the current experimental condition. In addition, the trainers were not briefed on the hypothesis associated with each experimental condition. Confederate behavior could be scripted as these users were always the initiators and addressers as they provided instructions on how to perform the neck exercise. We then evaluate the quality ratings only from the point of view of the trainees in line with the recommendations in ITU-T P.1301 [15].

3.2 Adaptive VR Remote Communication System

3.2.1 System. The overall architecture and dataflow of the real-time VR remote communication system are shown in Figure 1 with baseline, TA and NA streaming. This is an extension of the system presented in [18] with novel components for NA and TA streaming. Apart from the point cloud delivery pipeline described here the actual implementation also contains an audio delivery pipeline and a module for session management. The point cloud capture, and codec modules are implemented in C++, the other modules in C# with overall control in the Unity game engine.

The capture module interfaces with three Azure Kinect Depth sensors as shown in Figure 1. The sensors are calibrated in advance to generate the transformation matrix with intrinsics to combine
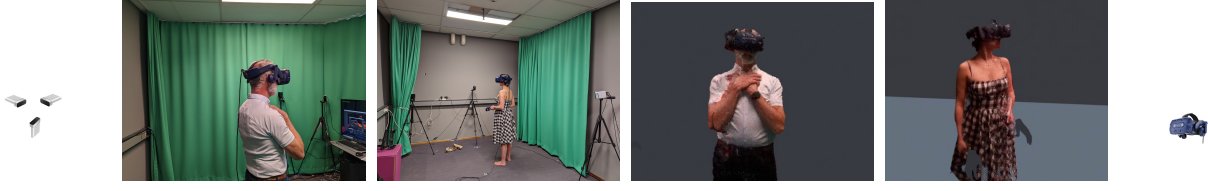
Figure 2: Capture Nodes and Real-Time Point Cloud Reconstructions

color and depth as well as extrinsics to bring all sensors to a common coordinate system. The color and depth images from the sensors are then transformed and fused in order to reconstruct the point cloud. We set a target capture framerate of 15 fps based on pre-tests with colleagues as this was the maximum achievable framerate at an acceptable baseline quality. The point cloud generated is first sent directly to the renderer in order to generate a self view to embody the participant. The layout of the capture node is shown in Figure 2.

In baseline uncompressed, the point cloud is serialized and sent directly to the writer to forward to the receiver. For NA, the point cloud is sent to the encoder where it is compressed to three quality levels and sent to the writer to prepare the adaptation set with the associated encoded size.

For TA, the point cloud is split into tiles based on the contributing sensor. Each tile contains an orientation vector that is derived from the transformation matrix of the sensor and the centroid of it's bounding box. The tiles are then fed to the compression module that launches encoders in parallel for each tile and quality level. In this way we create an adaptation set with multiple representations for each tile. In addition, we prepare a tile meta data structure that contains information on the number of tiles, meta data for each tile, available quality levels and the associated encoded size.

At the receiver, for baseline uncompressed, the point cloud is sent directly to the renderer. For NA, the adaptation engine selects the highest possible quality within the available bandwidth budget. This is then decoded and sent to the renderer. We apply the bitrate budget per frame based on the target capture framerate of 15 frames per second (fps).

For TA, the adaptation engine utilizes both the tile metadata from the sender and the receiver's interactions in the system in terms of viewport position and orientation to select an appropriate representation for each tile within the available bandwidth budget. The tiles are then decoded in parallel and sent to the synchronizer. The synchronizer module was developed to playback tiled point cloud sequences with tiles of varying sizes and quality in a consistent manner. The primary goal of the synchronizer is to playback tiles of the same frame together with a secondary goal of playing back frames at the right time to match the received frame rate. The point clouds are then sent to the renderer.

Finally, the renderer stores the point locations and colors on a vertex buffer and draws procedural geometry on the GPU. Points are rendered as camera-facing quads with a fixed offset based on the selected quality level.

3.2.2 Tiling and Tile Selection. We use a modified version of the approach proposed by Subramanyam *et al.* [48] in order to create

tiles. The point cloud is split into tiles based on the contributing sensor. We use the forward vector of the contributing sensor to estimate the orientation of the tile surface. We also compute the centroid of the bounding box of the tile. Each tile T_i has an orientation $\vec{T}_i$ and a bounding box centroid $T_i^{(bc)}$. The current viewport V of the user is defined by a position $V^{(pos)}$ and an orientation $\vec{V}$. The utility of each tile is calculated based on the following formula:

$$u(V, T_i) = \begin{cases} |\vec{T}_i \cdot \vec{V}|, & \text{if } d(V^{(c)}, T_i^{(bc)}) < d_{max}(V^{(bc)}, T_j^{(c)}) \\ -|\vec{T}_i \cdot \vec{V}|, & \text{otherwise.} \end{cases} \quad (1)$$

We use the absolute value of the dot product to identify surfaces directly facing the user. In addition, to account for the position of the user we ensure that the two tiles closest to the user always have a positive utility.

In the next step, the calculated utility is used to allocate the bandwidth budget to each tile. In this work, we re-use the allocation strategies presented in previous research [48, 50]. Based on pre-tests with colleagues, we found that uniform bit rate allocation achieved a higher median score and a lower spread of scores. We use the utility to rank the available tiles. The quality of each tile is then increased one step at a time in order of this ranking.

3.3 Experiment Conditions

The primary components determining the point cloud quality are the capture sensors, the codec configurations or adaptation set, target bitrate, streaming condition and the display device used. In order to evaluate adaptive streaming, we vary the streaming condition and target bitrate. The other factors do not change dynamically over a session and are uninteresting for streaming optimization. They are held constant across all participants. In addition, the audio quality is also maintained at the same level across all experimental sessions using the Ogg container format and the Speex codec with an ultra wide band sampling rate. In order to conduct a pre-test and set the experimental conditions, we used the dataset published by Reimat *et al.* [42] sub-sampled to three cameras (1,5,6) as this most closely resembles our capture setup. Based on pre-tests on encode time and captured point count we use three codec configurations on the MPEG anchor codec shown in Table 1. All codec configurations were encoded at JPEG QP 75. We also set two target bitrates exclusively for remote user point cloud reconstructions at 7Mbps and 14Mbps based on the approximate bandwidth requirement for full point clouds using the two highest quality levels from the pre-tests. We selected three neck exercises that were taught to all participants. In the experiment, we evaluate three streaming conditions: baseline uncompressed, NA and TA.

3.4 Experiment Design and Protocol

With three neck exercises and three streaming conditions, we used a Greco-Latin square design to randomize and counter balance the different levels of each variable. In order to avoid fatigue, we separated the target bitrates so each participant only took part in three sessions at a fixed target bitrate. We recruited two confederate users to play the role of trainer for each target bitrate. We recruited 16 and 17 participants for the two target bitrates of 7Mbps and 14Mbps.

Upon arrival, participants were led to the experiment room and were briefed about the purpose of the study, after which they were asked for written consent for data gathering. Participants were asked to provide some background information on themselves and to take a Ishihara test [3] for color perception. Participants were then asked to fill the simulator sickness questionnaire before starting and again after each experiment session. We then conducted a brief training session where participants were shown the highest and lowest available quality of the remote user point cloud to serve as an anchor. Participants were then taught how to use the HMD controllers to teleport and navigate the virtual space. Participants were informed that during the experiment they are free to move about the virtual space. Participants then entered the first session, in each session there was a brief introduction by the trainer, a training stage where the trainer demonstrated the exercise technique and finally a performance where participants were asked to repeat the exercise three times in order to complete the session. Each session took 2 to 3 minutes to complete. Participants were asked to fill in a questionnaire to report their experience after each session. Participants completed the experiment in ca. 30 minutes.

3.5 Experiment Setup

Participants took part in the study in a separate room from the trainer. Each room had a workstation, an HTC Vive Pro Eye HMD (with controllers and base stations) and three azure kinect depth sensors. The configuration of the setup used by participants is shown in Table 1. Figure 2 shows the setup used to capture each user and the resulting point cloud. The two labs were connected using a dedicated gigabit ethernet connection in order to control the connection quality for the duration of the study. Each of the azure kinect cameras was set to use a depth mode of NFOV unbinned with a resolution of 640x576 and the lowest supported color resolution of 1280x720. This was done as the color image is later mapped to the depth image in order to reconstruct the point cloud similar to the method proposed by Reimat *et al.* [42].

During each experiment session the system resource consumption was measured using the Resources Consumption Metrics (RCM) tool [34]. The conversation audio was recorded with audio only capture using the Open Broadcaster Software tool. The VR remote communication application recorded log files that contained performance information on each component including latency and framerate.

3.6 Research Questions and Data Collected

After each condition participants filled in a questionnaire about the experience they just had. The first six questions were related to QoI [26] with a 5-point Likert scale to address R1. The next
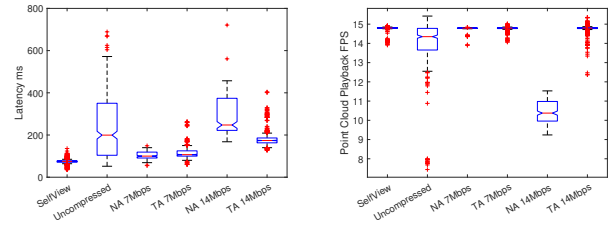

Figure 3: System Performance in terms of latency (ms) and framerate (fps)

four questions were about the visual quality of the point cloud representation [16] with a 5-point ACR scale to address R3. The last five questions were about task related experience [20] with a 7-point Likert scale to address R2. In addition, on the trainee node we record system resource consumption and log playback performance to address R4.

4 RESULTS

4.1 Performance Results

We evaluate system performance based on resource consumption, framerate and latency. The machine used for the evaluation is described in Table 1. Resource consumption was logged at the trainee node for each experiment session using the RCM measurement tool [34]. This tool is a windows application that allows for the capture of CPU, GPU and memory usage for a given process in a 1-second interval. The results are shown in Table 2. As expected we see higher memory consumption for large uncompressed point clouds. At 7Mbps we observe ca. 10% reduction in CPU usage on account of additional parallelization in encoding and decoding in TA as compared to NA with similar GPU and memory utilization. At 14Mbps we observe similar results with a 9% reduction in CPU usage.

In addition, our VR application logs the capture to render latency and framerate while accounting for clock sync between the trainee and trainer node. The latency results are shown in Figure 3. The selfview latency describes the time needed for reconstructing and rendering only. The selfview is rendered with a median latency of 75ms across all experiment sessions. We observe the largest range of latency for baseline uncompressed streaming with the largest point clouds requiring ca. 300Mbps to transmit and render. At 7Mbps we observe similar latency across the two streaming conditions. However, at 14Mbps we observe a 74ms increase in median latency for NA. This is caused by larger encode and decode times required for the highest quality point clouds in our adaptation set. In case of TA we have some performance gains due to parallel encoding and decoding of tiles and generally smaller point clouds decoded at the receiver.

The application runs at a near consistent 90 fps with motion reprojection. The point cloud target capture framerate is set to 15 fps based on the capability of the system. On the receiver end, we observe a drop in median framerate to 10.4 fps for NA at 14Mbps caused by the encode and decode times required for the highest quality full point clouds in our adaptation set. For the remaining streaming conditions we generally observe similar performance of ca. 15 fps point cloud playback with a larger range for uncompressed

Table 2: System Resource Consumption (1000+ samples each, mean values)

	Streaming Condition	Target Bitrate	CPU (%)	GPU(%)	Memory (MB)
Uncompressed	Baseline	-	31.1% (SD 7.8)	46.2% (SD 6.8)	3146 (SD 1423)
Compressed	Network Adaptive	7Mbps	49.7% (SD 5.9)	45.2% (SD 6.9)	1025 (SD 112)
	Tiled Adaptive	7Mbps	39.2% (SD 10.7)	46.9% (SD 5.4)	1088 (SD 94)
	Network Adaptive	14Mbps	57.4% (SD 4.6)	45.7% (SD 4.8)	1071 (SD 68)
	Tiled Adaptive	14Mbps	48.4% (SD 6.4)	46.5% (SD 4.7)	1096 (SD 89)

point clouds as shown in Figure 3. To summarize, we observe significant gains in playback performance (framerate and latency) by employing TA with lower CPU usage as compared to NA to address R4.

4.2 Subjective Results

4.2.1 Quality of Communication. In this section of the questionnaire we included the QoI questions from [26]. These questions are meant to assess four types of experience: (1) feeling understood (2) engaging in conversations (3) sensing other's emotion and (4) feeling comfortable in the environment. The overall QoI scores are obtained by adding up the scores for each of the six questions. We split the analysis for each target bitrate of 7Mbps and 14Mbps.

Table 3: Pairwise post-hoc test QoI and streaming condition at 7Mbps

Streaming Condition	Z	p	r
NA – TA	-2.2340	0.0025	0.3950
NA – Baseline	-2.6410	0.0083	0.4670
TA – Baseline	-1.7230	0.0849	0.3050

For the 7Mbps case a Shapiro-Wilk normality test issued on the entirety of the scores indicates that they do not follow a normal distribution ($W = 0.9163, p = 0.003$). We use non-parametric statistical tools to perform an exploratory data analysis and check if statistical differences could be found amongst the different streaming conditions. To compare the QoI across the different streaming conditions, we first conduct a Friedman's test to check if any groups exist with statistically significant differences ($\chi^2 = 12.04, p = 0.0024$). We then conduct a Wilcoxon signed-rank test with Bonferroni correction. The results are shown in table 3

We observe statistically significant differences in two of the comparisons. TA outperforms NA with a medium effect size ($r = 0.39505$). As expected, baseline uncompressed streaming outperforms NA with a large effect size ($r = 0.467$). On the other hand, we do not observe statistically significant differences between TA and baseline uncompressed indicating similar performance in terms of QoI.

For the 14Mbps case, a Shapiro-Wilk normality test indicates that the scores are not normally distributed ($W = 0.9484, p = 0.002$). In order to check if any of the groups exhibit statistically significant differences we run Friedman's test ($\chi^2 = 24.96, p < 0.001$). We then conduct a Wilcoxon signed-rank test with Bonferroni correction. The results are shown in table 4.

This time we observe statistically significant differences in all comparisons. TA outperforms NA with a medium effect size ($r =$

Table 4: Pairwise post-hoc test QoI and streaming condition at 14Mbps

Streaming Condition	Z	p	r
NA – TA	-3.3720	<.001	0.3780
NA – Baseline	-3.3310	<.001	0.5710
TA – Baseline	2.1650	0.0304	0.3710

0.3780). Baseline uncompressed streaming outperforms NA with a large effect size ($r = 0.5710$) and outperforms TA with a medium effect size ($r = 0.3710$). In general, we observe that TA leads to statistically significant gains in terms of QoI with respect to NA across both bitrates to address R1.

In order to validate the three exercises we used, we checked if they led to different QoI scores and we found no statistically significant differences using the Friedman test ($\chi^2 = 3.16, p = 0.206$) at 7Mbps and ($\chi^2 = 2.28, p = 0.3198$) at 14Mbps.

4.2.2 Visual Quality. In order to assess the visual quality we include a question about the visual quality of the trainer's point cloud representation. Participants were asked to indicate the quality on a scale from 1 to 5 (*1-Bad*, *2-Poor*, *3-Fair*, *4-Good*, and *5-Excellent*). We analyzed these scores separately for each target bitrate.

Table 5: Pairwise post-hoc test visual quality and streaming condition at 7Mbps

Streaming Condition	Z	p	r
NA – TA	-2.5840	0.0098	0.4570
NA – Baseline	-2.5550	0.0106	0.4520
TA – Baseline	-1.2650	0.2059	0.2240

For the 7Mbps case, a Shapiro-Wilk normality test issued on the scores indicates that they do not follow a normal distribution ($W = 0.8106, p < 0.001$). To compare the remote user visual quality across the different streaming conditions, we first conduct a Friedman's test to check if any groups exist with statistically significant differences ($\chi^2 = 9.8, p = 0.0074$). We then conduct a Wilcoxon signed-rank test with Bonferroni correction. The results are shown in table 5. We observe statistically significant differences in two pairwise comparisons. TA outperforms NA with a medium effect size ($r = 0.4570$). Baseline uncompressed streaming outperforms NA with a medium effect size ($r = 0.4520$). We do not observe statistically significant differences between TA and baseline indicating similar performance in terms of visual quality.

Table 6: Pairwise post-hoc test visual quality and streaming condition at 14Mbps

Streaming Condition	Z	p	r
NA – TA	-2.8140	0.0049	0.4830
NA – Baseline	-3.1400	0.0017	0.5390
TA – Baseline	-2.1210	0.0339	0.3640
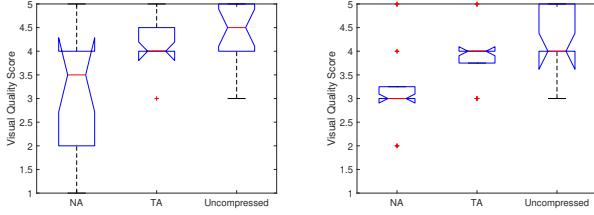
Figure 4: Point cloud quality at 7Mbps and 14Mbps

For the 14Mbps case, a Shapiro-Wilk normality test indicates the scores do not follow a normal distribution ($W = 0.8622$, $p < 0.001$). Friedman's test reveals that there are groups with statistically significant differences ($\chi^2 = 19$, $p < 0.001$). The results of the Wilcoxon signed-rank test with bonferroni corrections are shown in Table 6. This time we observe statistically significant differences in two of the pairwise comparisons. TA outperforms NA with a medium effect size ($r = 0.4830$). Baseline uncompressed outperforms NA with a large effect size ($r = 0.5390$). We do not observe statistically significant differences between TA and baseline uncompressed streaming indicating similar performance in terms of visual quality at 14Mbps. The distribution of the scores is shown in Figure 4. To address R3, we observe significant gains in perceived visual quality by employing TA over NA. At 14Mbps we even observe the same median score as baseline uncompressed streaming that required ca. 300Mbps.

4.2.3 Task Related Experience. For the task related experience questions, we used the questionnaire presented in [20]. The questions are meant to assess the participant's confidence in using the system and if the system was natural and easy to use while performing the task. We compute an overall score by adding up the five questions from this section. We split the analyses by target bit rate. A Shapiro Wilk normality test issued on the scores revealed that they are not normally distributed; ($W = 0.9019$, $p = 0.0013$) for 7Mbps and ($W = 0.9653$, $p = 0.014$) for 14Mbps. We then checked if there are any groups with statistically significant differences using Friedman's test ($\chi^2 = 2.1$, $p = 0.3504$) at 7Mbps and ($\chi^2 = 2.56$, $p = 0.278$) at 14Mbps. We found no statistically significant differences amongst the different streaming conditions at both target bitrates. This indicates participants were able to adapt their behavior to compensate for changes in the point cloud quality and were able to complete the task within the same time regardless. We observe no gains in task experience by employing TA (R2). This can be explained by the relative simplicity of completing our training task as compared to tasks used in other works [16, 26, 43].

4.2.4 Simulator Sickness Questionnaire. We calculate the total severity of cybersickness based on the SSQ questionnaire for all 33 participants across the three experiment sessions. We observe low

post-exposure total severity scores for cybersickness, as defined in [1, 21] ([mean, median, std] baseline: [10.54, 3.74, 18.22], after session 1: [9.63, 3.74, 14.88], after session 2: [9.41, 3.74, 16.14], after session 3:[11.22, 3.74, 16.95]). In general, no users reported severe symptoms after participating in any of the experiment sessions.

5 DISCUSSION

In this paper, we presented a VR remote communication system with tiled adaptive real-time point cloud streaming using commodity hardware. We present an evaluation framework and a training task to evaluate the impact of adaptive streaming on QoI, visual quality and task-related experience. Our system at 14Mbps was able to achieve similar visual quality as compared to uncompressed streaming at ca. 300Mbps. We also demonstrate statistically significant improvements to QoI as compared to traditional network adaptive streaming as well as improvements to playback performance with up to 10% reduction in CPU usage.

In our pre-trials, we observed a higher median score and a lower spread of scores with uniform tile allocation. This is different from the results with a prerecorded dataset [48], where hybrid tile allocation was shown to yield higher perceived quality. The point clouds used in that study were captured offline. They were dense and voxelized with ca. 1 million points per frame and the adaptation set comprised of 30 quality levels. In our study, we used real-time live captured point clouds with ca. 130K points with an adaptation set of 3 quality levels defined by octree depths.

6 CONCLUSION AND FUTURE WORK

In general, tiled adaptive streaming techniques have received significant research attention for omnidirectional videos [5, 40, 56] and point clouds [12, 30, 39, 41, 51]. However, further study is required for live real-time human point cloud reconstructions. Although the tiles are independently decodable, their quality cannot be optimized in isolation based on available bandwidth and viewport. Some participants reported that seeing artifacts in body extremities and in the face of the reconstruction was unpleasant. Further study into body part segmentation and quality perception is required to optimize tiling and tile selection strategies for humans engaging in conversation as compared to prerecorded content, objects and scenes.

There is a need for new standardized tasks that can be used to evaluate VR remote communication. The existing ITU recommendations are insufficient to handle novel interaction techniques and immersive content inherent to VR communication. In this work, we utilized a neck exercise training task as it was more visually focused and the interaction was repeatable with a confederate trainer. Further study into other use cases and scenarios are required to evaluate emerging VR communication systems. To this end, ITU-T has recently launched a new activity [17] to develop assessment methods for extended reality meetings.

REFERENCES

[1] Pauline Bimberg, Tim Weissker, and Alexander Kulik. 2020. On the Usage of the Simulator Sickness Questionnaire for Virtual Reality Research. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. 464–467. https://doi.org/10.1109/VRW50115.2020.00098

[2] Gianluca Cernigliaro, Marc Martos, Mario Montagud, Amir Ansari, and Sergi Fernandez. 2020. PC-MCU: Point Cloud Multipoint Control Unit for Multi-User Holoconferencing Systems. In *Proceedings of the 30th ACM Workshop on Network and Operating Systems Support for Digital Audio and Video*. Association for Computing Machinery, New York, NY, USA, 47–53.

[3] J. H. Clark. 1924. The Ishihara Test for Color Blindness.. In *American Journal of Physiological Optics*. 269–276.

[4] Ricardo L. de Queiroz and Philip A. Chou. 2016. Compression of 3D Point Clouds Using a Region-Adaptive Hierarchical Transform. *IEEE Transactions on Image Processing* 25, 8 (2016), 3947–3956. https://doi.org/10.1109/TIP.2016.2575005

[5] Ching-Ling Fan, Wen-Chih Lo, Yu-Tung Pai, and Cheng-Hsin Hsu. 2019. A Survey on 360° Video Streaming: Acquisition, Transmission, and Display. *ACM Comput. Surv.* 52, 4, Article 71 (aug 2019), 36 pages. https://doi.org/10.1145/3329119

[6] H. Fuchs, A. State, and J. Bazin. 2014. Immersive 3D Telepresence. *Computer* 47, 07 (jul 2014), 46–52. https://doi.org/10.1109/MC.2014.185

[7] Edd Gent. 2020. Forget Video Conferencing—Host Your Next Meeting in VR. https://spectrum.ieee.org/forget-video-conferencinghost-your-next-meeting-in-vr

[8] Serhan Gül, Dimitri Podborski, Anna Hilsmann, Wieland Morgenstern, Peter Eisert, Oliver Schreer, Thomas Buchholz, Thomas Schierl, and Cornelius Hellge. 2020. INTERACTIVE VOLUMETRIC VIDEO FROM THE CLOUD.

[9] Simon N. B. Gunkel, Rick Hindriks, Karim M. El Assal, Hans M. Stokking, Sylvie Dijkstra-Soudarissanane, Frank ter Haar, and Omar Niamut. 2021. VRComm: An End-to-End Web System for Real-Time Photorealistic Social VR Communication. In *Proceedings of the 12th ACM Multimedia Systems Conference* (Istanbul, Turkey) *(MMSys '21)*. Association for Computing Machinery, New York, NY, USA, 65–79. https://doi.org/10.1145/3458305.3459595

[10] Bo Han, Yu Liu, and Feng Qian. 2020. *ViVo: Visibility-Aware Mobile Volumetric Video Streaming*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3372224.3380888

[11] Lanyi He, Wenjie Zhu, Ke Zhang, and Yiling Xu. 2018. View-Dependent Streaming of Dynamic Point Cloud over Hybrid Networks. In *Advances in Multimedia Information Processing – PCM 2018*. Springer International Publishing, Cham, 50–58.

[12] Jeroen van der Hooft, Maria Torres Vega, Tim Wauters, Christian Timmerer, Ali C. Begen, Filip De Turck, and Raimund Schatz. 2020. From Capturing to Rendering: Volumetric Media Delivery with Six Degrees of Freedom. *IEEE Communications Magazine* 58, 10 (2020), 49–55. https://doi.org/10.1109/MCOM.001.2000242

[13] Mohammad Hosseini. 2017. Adaptive Rate Allocation for View-Aware Point-Cloud Streaming. (2017). https://doi.org/10.13140/RG.2.2.23436.26244

[14] Mohammad Hosseini and Christian Timmerer. 2018. Dynamic Adaptive Point Cloud Streaming. In *Proceedings of the 23rd Packet Video Workshop* (Amsterdam, Netherlands) *(PV '18)*. ACM, New York, NY, USA, 25–30. https://doi.org/10.1145/3210424.3210429

[15] ITU-T P.1301. 2017. P.1301 : Subjective quality evaluation of audio and audiovisual multiparty telemeetings. International Telecommunication Union.

[16] ITU-T P.920. 2000. P.920 : Interactive test methods for audiovisual communications. International Telecommunication Union.

[17] ITU-T P.QXM. 2022. QoE assessment of eXtended Reality (XR) meetings. International Telecommunication Union. https://www.itu.int/ITU-T/workprog/wp_item.aspx?isn=17791

[18] Jack Jansen, Shishir Subramanyam, Romain Bouqueau, Gianluca Cernigliaro, Marc Martos Cabré, Fernando Pérez, and Pablo Cesar. 2020. A Pipeline for Multiparty Volumetric Video Conferencing: Transmission of Point Clouds over Low Latency DASH. In *Proceedings of the 11th ACM Multimedia Systems Conference* (Istanbul, Turkey) *(MMSys '20)*. ACM, New York, NY, USA.

[19] Brett Jones, Rajinder Sodhi, Michael Murdock, Ravish Mehra, Hrvoje Benko, Andrew Wilson, Eyal Ofek, Blair MacIntyre, Nikunj Raghuvanshi, and Lior Shapira. 2014. RoomAlive: Magical Experiences Enabled by Scalable, Adaptive Projector-Camera Units. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology* (Honolulu, Hawaii, USA) *(UIST '14)*. Association for Computing Machinery, New York, NY, USA, 637–644. https://doi.org/10.1145/2642918.2647383

[20] Jari Kangas, Sriram Kishore Kumar, Helena Mehtonen, Jorma Järnstedt, and Roope Raisamo. 2022. Trade-Off between Task Accuracy, Task Completion Time and Naturalness for Direct Object Manipulation in Virtual Reality. *Multimodal Technologies and Interaction* 6, 1 (2022). https://doi.org/10.3390/mti6010006

[21] Robert S. Kennedy, Norman E. Lane, Kevin S. Berbaum, and Michael G. Lilienthal. 1993. Simulator Sickness Questionnaire: An Enhanced Method for Quantifying Simulator Sickness. *The International Journal of Aviation Psychology* 3, 3 (1993), 203–220. https://doi.org/10.1207/s15327108ijap0303_3

[22] AK Kuhlen and Brennan SE. 2013. Language in dialogue: when confederates might be hazardous to your data. *Psychon Bull Rev.* 20, 1 (2013), 54–72. https://doi.org/10.3758/s13423-012-0341-8

[23] Marc Erich Latoschik, Daniel Roth, Dominik Gall, Jascha Achenbach, Thomas Waltemate, and Mario Botsch. 2017. The Effect of Avatar Realism in Immersive Social Virtual Realities. In *Proceedings of the 23rd ACM Symposium on Virtual Reality Software and Technology* (Gothenburg, Sweden) *(VRST '17)*. Association for Computing Machinery, New York, NY, USA, Article 39, 10 pages. https://doi.org/10.1145/3139131.3139156

[24] Jason Lawrence, Danb Goldman, Supreeth Achar, Gregory Major Blascovich, Joseph G. Desloge, Tommy Fortes, Eric M. Gomez, Sascha Häberling, Hugues Hoppe, Andy Huibers, Claude Knaus, Brian Kuschak, Ricardo Martin-Brualla, Harris Nover, Andrew Ian Russell, Steven M. Seitz, and Kevin Tong. 2021. Project Starline: A High-Fidelity Telepresence System. *ACM Trans. Graph.* 40, 6, Article 242 (dec 2021), 16 pages. https://doi.org/10.1145/3478513.3480490

[25] Kyungjin Lee, Juheon Yi, Youngki Lee, Sunghyun Choi, and Young Min Kim. 2020. *GROOT: A Real-Time Streaming System of High-Fidelity Volumetric Videos*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3372224.3419214

[26] Jie Li, Yiping Kong, Thomas Röggla, Francesca De Simone, Swamy Ananthanarayan, Huib de Ridder, Abdallah El Ali, and Pablo Cesar. 2019. Measuring and Understanding Photo Sharing Experiences in Social Virtual Reality. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3290605.3300897

[27] Jie Li, Cong Zhang, Zhi Liu, Wei Sun, and Qiyue Li. 2020. Joint Communication and Computational Resource Allocation for QoE-driven Point Cloud Video Streaming. https://doi.org/10.48550/ARXIV.2001.01403

[28] Zhi Liu, Jie Li, Xianfu Chen, Celimuge Wu, Susumu Ishihara, Yusheng Ji, and Jie Li. 2020. Fuzzy Logic-Based Adaptive Point Cloud Video Streaming. *IEEE Open Journal of the Computer Society* 1 (2020), 121–130. https://doi.org/10.1109/OJCS.2020.3006205

[29] Zhi Liu, Qiyue Li, Xianfu Chen, Celimuge Wu, Susumu Ishihara, Jie Li, and Yusheng Ji. 2021. Point Cloud Video Streaming: Challenges and Solutions. *IEEE Network* 35, 5 (2021), 202–209. https://doi.org/10.1109/MNET.101.2000364

[30] Zhi Liu, Qiyue Li, Xianfu Chen, Celimuge Wu, Susumu Ishihara, Jie Li, and Yusheng Ji. 2021. Point Cloud Video Streaming: Challenges and Solutions. *IEEE Network* 35, 5 (2021), 202–209. https://doi.org/10.1109/MNET.101.2000364

[31] Danielle S. McNamara and Jeremy N. Bailenson. 2021. Nonverbal Overload: A Theoretical Argument for the Causes of Zoom Fatigue. *Technology, Mind, and Behavior* 2, 1 (23 2 2021). https://doi.org/10.1037/tmb0000030 https://tmb.apaopen.org/pub/nonverbal-overload.

[32] Rufael Mekuria, Kees Blom, and Pablo Cesar. 2016. Design, Implementation and Evaluation of a Point Cloud Codec for Tele-Immersive Video. *IEEE Transactions on Circuits and Systems for Video Technology* (January 2016).

[33] Rufael Mekuria, Pablo Cesar, Ioannis Doumanis, and Antonella Frisiello. 2015. Objective and subjective quality assessment of geometry compression of reconstructed 3D humans in a 3D virtual room. *Proc. SPIE 9599, Applications of Digital Image Processing XXXVIII, 95991M* (September 2015).

[34] Mario Montagud, Juan Antonio De Rus, Rafael Fayos-Jordan, Miguel Garcia-Pineda, and Jaume Segura-Garcia. 2020. Open-Source Software Tools for Measuring Resources Consumption and DASH Metrics. In *Proceedings of the 11th ACM Multimedia Systems Conference* (Istanbul, Turkey) *(MMSys '20)*. Association for Computing Machinery, New York, NY, USA, 261–266. https://doi.org/10.1145/3339825.3394931

[35] U.S. Department of Energy Advanced Research Projects Agency Energy (ARPA-E). 2017. Facsimile Appearance to Create Energy Savings (FACES). https://arpa-e-foa.energy.gov/Default.aspx?foaId=b437e046-1235-4620-bb13-e87bdecdd537

[36] Sergio Orts-Escolano, Christoph Rhemann, Sean Fanello, Wayne Chang, Adarsh Kowdle, Yury Degtyarev, David Kim, Philip L. Davidson, Sameh Khamis, Mingsong Dou, Vladimir Tankovich, Charles Loop, Qin Cai, Philip A. Chou, Sarah Mennicken, Julien Valentin, Vivek Pradeep, Shenlong Wang, Sing Bing Kang, Pushmeet Kohli, Yuliya Lutchyn, Cem Keskin, and Shahram Izadi. 2016. Holoportation: Virtual 3D Teleportation in Real-Time. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology* (Tokyo, Japan) *(UIST '16)*. Association for Computing Machinery, New York, NY, USA, 741–754. https://doi.org/10.1145/2984511.2984517

[37] Jounsup Park, Philip A. Chou, and Jenq-Neng Hwang. 2019. Rate-Utility Optimized Streaming of Volumetric Media for Augmented Reality. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems* 9, 1 (2019), 149–162.

[38] Eduardo Pavez, Philip A Chou, Ricardo L de Queiroz, and Antonio Ortega. 2016. Dynamic Polygon Cloud Compression. *Microsoft Research Technical Report* (October 2016).

[39] Stefano Petrangeli, Gwendal Simon, Haoliang Wang, and Vishy Swaminathan. 2019. Dynamic Adaptive Streaming for Augmented Reality Applications. In *2019 IEEE International Symposium on Multimedia (ISM)*. 56–567. https://doi.org/10.1109/ISM46123.2019.00017

[40] Stefano Petrangeli, Viswanathan Swaminathan, Mohammad Hosseini, and Filip De Turck. 2017. An HTTP/2-Based Adaptive Streaming Framework for 360° Virtual Reality Videos. In *Proceedings of the 25th ACM International Conference on Multimedia* (Mountain View, California, USA) *(MM '17)*. Association for Computing Machinery, New York, NY, USA, 306–314. https://doi.org/10.1145/3123266.3123453

[41] Feng Qian, Bo Han, Jarrell Pair, and Vijay Gopalakrishnan. 2019. Toward Practical Volumetric Video Streaming on Commodity Smartphones. In *Proceedings of the 20th International Workshop on Mobile Computing Systems and Applications* (Santa Cruz, CA, USA) *(HotMobile '19)*. Association for Computing Machinery, New York, NY, USA, 135–140. https://doi.org/10.1145/3301293.3302358

[42] Ignacio Reimat, Evangelos Alexiou, Jack Jansen, Irene Viola, Shishir Subramanyam, and Pablo Cesar. 2021. CWIPC-SXR: Point Cloud Dynamic Human Dataset for Social XR. In *Proceedings of the 12th ACM Multimedia Systems Conference* (Istanbul, Turkey) *(MMSys '21)*. Association for Computing Machinery, New York, NY, USA, 300–306. https://doi.org/10.1145/3458305.3478452

[43] Marwin Schmitt, Judith Redi, Dick Bulterman, and Pablo S. Cesar. 2018. Towards Individual QoE for Multiparty Videoconferencing. *IEEE Transactions on Multimedia* 20, 7 (2018), 1781–1795. https://doi.org/10.1109/TMM.2017.2777466

[44] Sebastian Schwarz, Marius Preda, Vittorio Baroncini, Madhukar Budagavi, Pablo Cesar, Philip A Chou, Robert A Cohen, Maja Krivokuća, Sebastien Lasserre, Zhu Li, et al. 2019. Emerging MPEG Standards for Point Cloud Compression. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems* 9, 1 (March 2019), 133–148. https://doi.org/10.1109/JETCAS.2018.2885981

[45] Mel Slater, Martin Usoh, and Anthony Steed. 1994. Depth of Presence in Virtual Environments. *Presence: Teleoper. Virtual Environ.* 3, 2 (jan 1994), 130–144. https://doi.org/10.1162/pres.1994.3.2.130

[46] Harrison Jesse Smith and Michael Neff. 2018. Communication Behavior in Embodied Virtual Reality. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3173574.3173863

[47] Jangwoo Son, Serhan Gül, Gurdeep Singh Bhullar, Gabriel Hege, Wieland Morgenstern, Anna Hilsmann, Thomas Ebner, Sven Bliedung, Peter Eisert, Thomas Schierl, Thomas Buchholz, and Cornelius Hellge. 2020. Split Rendering for Mixed Reality: Interactive Volumetric Video in Action. In *SIGGRAPH Asia 2020 XR* (Virtual Event, Republic of Korea) *(SA '20)*. Association for Computing Machinery,

New York, NY, USA, Article 8, 3 pages. https://doi.org/10.1145/3415256.3421491

[48] Shishir Subramanyam, Irene Viola, Alan Hanjalic, and Pablo Cesar. 2020. User Centered Adaptive Streaming of Dynamic Point Clouds with Low Complexity Tiling. In *Proceedings of the 28th ACM International Conference on Multimedia* (Seattle, WA, USA) *(MM '20)*. Association for Computing Machinery, New York, NY, USA, 3669–3677. https://doi.org/10.1145/3394171.3413535

[49] Alexander Toet, Tina Mioch, Simon N Gunkel, Omar Niamut, and Jan B van Erp. 2021. Holistic Framework for Quality Assessment of Mediated Social Communication. osf.io/4twyh

[50] Jeroen van der Hooft, Tim Wauters, Filip De Turck, Christian Timmerer, and Hermann Hellwagner. 2019. Towards 6DoF HTTP Adaptive Streaming Through Point Cloud Compression. In *Proceedings of the 27th ACM International Conference on Multimedia* (Nice, France) *(MM '19)*. Association for Computing Machinery, New York, NY, USA, 2405–2413. https://doi.org/10.1145/3343031.3350917

[51] Lisha Wang, Chenglin Li, Wenrui Dai, Shaohui Li, Junni Zou, and Hongkai Xiong. 2022. QoE-Driven Adaptive Streaming for Point Clouds. *IEEE Transactions on Multimedia* (2022), 1–1. https://doi.org/10.1109/TMM.2022.3148585

[52] Andrew D. Wilson and Hrvoje Benko. 2016. Projected Augmented Reality with the RoomAlive Toolkit. In *Proceedings of the 2016 ACM International Conference on Interactive Surfaces and Spaces* (Niagara Falls, Ontario, Canada) *(ISS '16)*. Association for Computing Machinery, New York, NY, USA, 517–520. https://doi.org/10.1145/2992154.2996362

[53] Bob G. Witmer and Michael J. Singer. 1998. Measuring Presence in Virtual Environments: A Presence Questionnaire. *Presence: Teleoper. Virtual Environ.* 7, 3 (jun 1998), 225–240. https://doi.org/10.1162/105474698565686

[54] Zhenyu Yang, Wanmin Wu, Klara Nahrstedt, Gregorij Kurillo, and Ruzena Bajcsy. 2010. Enabling Multi-Party 3D Tele-Immersive Environments with <i>ViewCast</i>. *ACM Trans. Multimedia Comput. Commun. Appl.* 6, 2, Article 7 (mar 2010), 30 pages. https://doi.org/10.1145/1671962.1671963

[55] Amal Yassien, Passant ElAgroudy, Elhassan Makled, and Slim Abdennadher. 2020. *A Design Space for Social Presence in VR.* Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3419249.3420112

[56] Michael Zink, Ramesh Sitaraman, and Klara Nahrstedt. 2019. Scalable 360° Video Stream Delivery: Challenges, Solutions, and Opportunities. *Proc. IEEE* 107, 4 (2019), 639–650. https://doi.org/10.1109/JPROC.2019.2894817