Task-oriented and Semantics-aware Communication Framework for Augmented Reality

Zhe Wang, Yansha Deng, and A. Hamid Aghvami

Abstract

Upon the advent of the emerging metaverse and its related applications in Augmented Reality (AR), the current bit-oriented network struggles to support real-time changes for the vast amount of associated information, hindering its development. Thus, a critical revolution in the Sixth Generation (6G) networks is envisioned through the joint exploitation of information context and its importance to the task, leading to a communication paradigm shift towards semantic and effectiveness levels. However, current research has not yet proposed any explicit and systematic communication framework for AR applications that incorporate these two levels. To fill this research gap, this paper presents a task-oriented and semantics-aware communication framework for augmented reality (TSAR) to enhance communication efficiency and effectiveness in 6G. Specifically, we first analyse the traditional wireless AR point cloud communication framework and then summarize our proposed semantic information along with the end-to-end wireless communication. We then detail the design blocks of the TSAR framework, covering both semantic and effectiveness levels. Finally, numerous experiments have been conducted to demonstrate that, compared to the traditional point cloud communication framework, our proposed TSAR significantly reduces wireless AR application transmission latency by 95.6%, while improving communication effectiveness in geometry and color aspects by up to 82.4% and 20.4%, respectively.

Index Terms

Metaverse, augmented reality, semantic communication, end-to-end communication.

Z. Wang, Y. Deng, and A. Hamid Aghvami (Emeritus Professor) are with the Department of Engineering, King’s College London, Strand, London WC2R 2LS, U.K. (e-mail: tylor.wang@kcl.ac.uk; yansha.deng@kcl.ac.uk; hamid.aghvami@kcl.ac.uk) (Corresponding author: Yansha Deng). Youtube: https://youtu.be/n9YPF979m_0

I. Introduction

The metaverse, as an expansion of the digital universe, has the potential to significantly influence people's lives, affecting their entertainment experiences and social behaviors. Specific applications such as Augmented Reality (AR), Virtual Reality (VR), and other immersive technologies within the metaverse have demonstrated remarkable potential in various areas, including virtual conferences, online education, and real-time interactive games, capturing the attention of both industry and academia [1]. These applications, also referred as eXtended Reality (XR), need to process rich and complex data, such as animated avatars, point cloud, and model mesh, to create immersive experiences for clients [2]. However, the extensive transmission of information and high bandwidth requirements within the XR pose significant challenges for its wider applications, particularly in avatar-related applications that necessitate real-time client communication and interaction. The existing communication networks fails to achieve such high bandwidth requirement and thus can not adequately support XR applications, necessitating the development of 6G technology to enhance its applications for further advancement [3, 4]. Specifically, to ensure a good Quality of Experience (QoE) in AR applications, a transmission latency of less than 20 ms is required, which is 20 times less than the transmission latency tolerated in video communication applications [5]. Due to the nature of numerous sensing data in AR applications, more packets need to be transmitted in such a short time, which consequently increases the demand for bandwidth. This growing concern about the transmission latency and bandwidth in AR application services highlights the need for further research in communication technology to ensure a real-time immersive experiences for clients in AR-related applications.

To address the high bandwidth requirements diploma in wireless communication in AR applications, the concept of semantic communication has been proposed [6]. This approach aims to facilitate communication at the semantic level by exploring only the content of traditional text and speech data or the information freshness. Initial research on semantic communication for text [7], speech [8], and image data [9] mainly focused on identifying the semantic content of traditional data. Other semantic communication research on sensor and control data emphasize on using information freshness, such as Age of Information (AoI) [10], as a semantic metric to estimate timeliness and evaluate the importance of the information. Note that these AoI-related semantic communication is unable to adequately capture the importance of specific information with inherent importance in the emerging AR dataset. This underscores the need to develop new strategies and techniques that effectively incorporate semantic communication in AR, considering not only information timeliness but also the relevance and sufficiency of the data content for a given application. In [11], a generic task-oriented and semantics-aware communication framework taking into account the designs at the semantic and effectiveness levels is envisioned for various tasks with diverse data types. However, an explicit and concrete task-oriented and semantics-aware communication framework has not been designed for AR application so far.

Current XR-related application research typically requires users to utilize Head-Mounted Displays (HMD) [12]. These applications generally focus on avatar-centric services, where the use of avatar animation in replacement of real human figures can decrease HMD computing requirements, reduce transmission data, and protect user privacy [13]. This avatar representation method has been implemented in social media platforms, such as TikTok and Instagram, where avatar characters is used for augmented reality video effects. Interestingly, using avatars instead of human has shown no significant differences in social behavior transmission and can even attract users to complete tasks more quickly in gaming situations [14]. For instance, fitness coaches can employ virtual avatars for AR conferencing to guide training. Games, like Pokémon Go, use avatars in mixed reality to encourage gamer interaction [15]. Avatar-based communication has been considered in [16], where the point cloud of avatars, structures, and models are transmitted between transmitter and receiver. Task-related effectiveness level performance metrics, including point-to-point [17], peak signal-to-noise ratio for the luminance component [18], mean per joint position error [19] have been considered to assess the telepresence task [20], point cloud video displaying task [21], and avatar pose recovery task [22], respectively. However, these AR-related applications have not fully addressed the issue of the effectiveness of avatar transmission, and bandwidth requirements for such applications still remain high. Users continue to experience suboptimal and lagging AR experiences in areas with moderate signal strength, indicating that the current AR communication framework has limitations, particularly in identifying a better avatar representation method for more effective communication, which need to be addressed.

Several studies have recently begun to explore the representation of avatars in wired communication. Different data types have been designed to represent avatars, which results in diverse avatar reconstruction required at the client side and limited transmission effectiveness evaluation capabilities for AR. For instance, skeleton elements have been proposed as a means to represent avatars, where motion capture devices are used to record skeletal positions. The recorded avatar movements are then replayed in wired Head-Mounted Displays (HMDs), and the differences in skeleton position between transmitter and receiver are measured to evaluate wired AR communication [23]. However, how to best extract semantic information that reflects the importance and context of information related to the avatar-centric display task is still unclear in a wireless communication AR application. The presence of redundant messaging can lead to an increase in transmission packets, resulting in decreased efficiency of wireless communication and ultimately impacting the user's viewing experience.

Inspired by the 3D keypoints extraction method presented in [24], we propose a task-oriented and semantics-aware communication framework in AR (TSAR) for avatar-centric end-to-end AR communication. In contrast to traditional point cloud AR communication frameworks that rely solely on point cloud input, our proposed TSAR extracts and transmits only essential semantic information. To the best of our knowledge, our contributions can be summarized as follows:

1) We propose a task-oriented and semantics-aware communication framework in augmented reality (TSAR) for interactive avatar-centric displaying applications with an integration of the semantic and effectiveness levels design, which includes semantic information extraction, task-oriented semantics-aware wireless communication, avatar pose recovery and rendering.
2) We apply an avatar-based semantic ranking (AbSR) algorithm to abstract features from the avatar skeleton graph using shared base knowledge and measure the importance of different semantic information. By utilizing Channel State Information (CSI) feedback, the AbSR can improve the avatar transmission quality in the wireless AR communication.
3) We have conducted a series of experiments comparing our proposed TSAR framework with the traditional point cloud communication framework. Our results indicate that our proposed TSAR framework outperforms the traditional point cloud communication framework in terms of color quality, geometry quality, and transmission latency for avatar-centric displaying task, with improvements of up to 20.4%, 82.4% and 95.6% respectively.

The rest of the paper is organized as follows: In section II, we present the system model and problem formation, covering both the traditional point cloud and the TSAR frameworks. Section

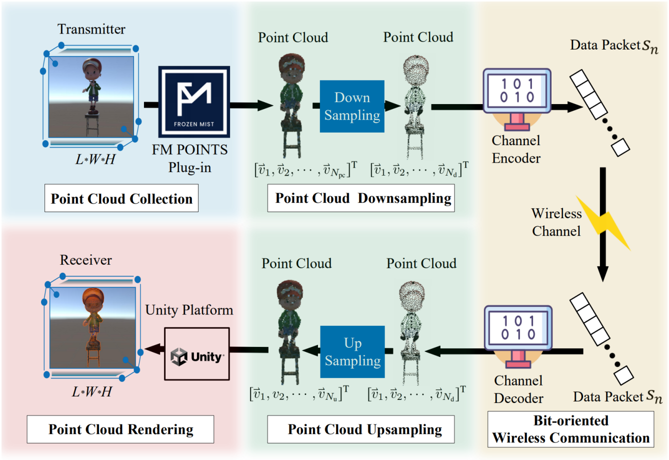


Fig. 1: Traditional point cloud communication framework

III details the design principles for semantic level. Section IV details the design principles for effectiveness level. Section V demonstrates the avatar movement and experimental performance evaluation. Finally, Section VI concludes this paper.

II. System Model and Problem Formation

In this section, we first describe the existing traditional point cloud communication framework for AR applications. Then, we present our wireless communication channel model implemented in both the point cloud communication framework and the TSAR. We further introduce our proposed TSAR in detail, which considers not only the bit-level but also the semantic and effectiveness levels. Finally, we present the problem formation and the objective function.

A. Traditional Point Cloud Communication Framework

As shown in Fig. 1, the procedures for traditional point cloud communication in AR applications typically consist of point cloud collection, downsampling, upsampling, and rendering.

1) Point Cloud Collection: We focus on interactive avatar-centric displaying and gaming AR applications, which are promising applications in the metaverse [13]. These AR applications require transmitting avatar animations and other stationary background models to the client side for displaying on an HMD in the area with dimensions length L, height H, and width W. To guarantee a smooth viewing experience of the AR scenery at the client side, high-resolution point cloud of both the moving avatar and stationary background models need to be captured and transmitted to the client side. Current Unity3D platform have numerous plugins for generating sensor data in real time, such as FM POINTS, which is a comprehensive point cloud visualization plugin that can transform the whole AR scenery or any 3D models into real-time point cloud. The information for each point $\vec{v}_i$ can be represented as

$$\vec{v}_i = (\vec{l}_i, \vec{c}_i) = (l_{\mathrm{x}}, l_{\mathrm{y}}, l_{\mathrm{z}}, c_{\mathrm{r}}, c_{\mathrm{g}}, c_{\mathrm{b}}), \tag{1}$$

where the $\vec{l}_i$ and $\vec{c}_i$ represent the three-dimensional location and RGB color of point, respectively. The generated point cloud $\mathbf{P}_{\mathrm{pc}}$ of the whole AR scenery consist of thousands of points v_i, which can be represented as

$$\mathbf{P}_{\mathrm{pc}} = [\vec{v}_1, \vec{v}_2, \cdots, \vec{v}_{N_{\mathrm{pc}}}]^{\mathrm{T}}, \tag{2}$$

where N_{pc} denotes the total number of generated point cloud of AR scenery. Typically, each 3D object needs to be represented by over 1,500 thousand point cloud in each frame to achieve a satisfactory viewing experience for clients [25].

2) Point Cloud Downsampling and Upsampling: In the traditional point cloud wireless communication framework, the transmission of a large number of point cloud can lead to data congestion at the wireless channel, causing intolerable delays and thus hinders AR application development [26]. To minimize transmission delays, current research explores the use of compression algorithms in point cloud transmission [27]. By introducing an downsample algorithm at the transmitter and an upsample algorithm at the receiver, the transmission latency can be reduced through transmitting only the compressed point cloud. The farthest point sampling algorithm [28] is ultilized as the downsample method, which enables the selection of representative points from the original point cloud while maintaining the overall features of the 3D objects. This algorithm reduces the number of points to be transmitted, thus improving the efficiency of the communication system. The process of farthest point downsampling $\mathcal{D}(\cdot)$, can be expressed as

$$\mathbf{P}_{\mathrm{dpc}} = [\vec{v}_1, \vec{v}_2, \cdots, \vec{v}_{N_{\mathrm{d}}}]^{\mathrm{T}} = \mathcal{D}(\mathbf{P}_{\mathrm{pc}}), \tag{3}$$

where $\mathbf{P}_{\mathrm{dpc}}$ represents the downsampled point cloud data awaiting transmission, and N_{d} is the total number of downsampled point cloud data. Then, the client's view experience can be enhanced by employing an upsampling algorithm for high-resolution point cloud recovery. Due to the instability of the wireless channel, the receiver faces the challenge of converting a sparse, irregular, and non-uniform point cloud into a dense, complete, and uniform one. To address this challenging issue [29], the linear interpolation algorithm [30] is introduced for the point cloud upsampling process. This algorithm involves estimating the positions of the missing points based on the positions of their neighbors, effectively generating a denser point cloud that closely resembles the original point cloud structure. The point cloud upsampling process, denoted as $\mathcal{U}(\cdot)$, can be expressed as

$$\mathbf{P}_{\text{upc}} = [\vec{v}_1, \vec{v}_2, \cdots, \vec{v}_{N_{\text{u}}}]^{\text{T}} = \mathcal{U}(\mathbf{P}'_{\text{dpc}}), \tag{4}$$

where $\mathbf{P}_{\text{upc}}$ is the reconstructed point cloud after upsampling, N_{u} represents the total number of upsampled point cloud, and $\mathbf{P}'_{\text{dpc}}$ is the received point cloud data after transmitting $\mathbf{P}_{\text{dpc}}$ over wireless channels. The upsampling process aims to accurately reconstruct the original point cloud, ensuring that the client-side viewing experience is maintained at a high quality despite the data compression and transmission through an unstable wireless channel.

3) Point Cloud Rendering: The point cloud rendering process begins when all the N_{u} point clouds for the AR scenery are received and upsampled. This process prepares the point cloud data for the Unity3D platform and facilitates high-resolution rendering. The rendering process needs to create a comprehensive 360° view of the avatar, along with immersive background scenery, which involves point cloud preparation and procedures:

(1) Point cloud preparation: Point cloud preparation involves formatting points from the received point cloud data. Each point contains information such as three-dimensional location and RGB color value, which determines the point's position and visual depiction within the virtual environment.

(2) Point cloud processing: The procedure of point cloud processing includes mesh reconstruction along with positioning. It commences with the transformation of these discrete points into a compatible mesh format for the Unity3D platform. Subsequently, the Shader, a uniquely designed program, is employed during the rendering process to regulate the gradients of illumination, obscurity, and chromaticity within the virtual environment. The final step of this process involves implementing the positioning phase to optimize the visualization, encompassing translation, rotation, and scaling elements. Concurrently, the Level of Detail (LoD) strategy is invoked in the whole processing process, which dynamically modulates the complexity of a 3D model representation contingent upon its spatial relation

to the clients. It renders fewer points when clients are distant and, conversely, more points as they step closer, thereby providing a better viewing experience.

B. Wireless Channel Model

The wireless communication model is characterized by a Rayleigh fading channel, impacted by additive white Gaussian noise and utilizing an Orthogonal Frequency-division Multiplexing (OFDM) scheme. The OFDM approach divides the wireless channel into multiple parallel subchannels. Each subchannel experiences varying levels of noise, leading to different Signal-to-Noise Ratios (SNRs).

The wireless communication process begins with source encoding, transforming the awaiting transmit data into the bitstream. Following this, a standard channel encoding is implemented to inject redundancy into the data to be transmitted, safeguarding data integrity and enabling the correction of potential errors during transmission. Traditional communication coding methods, such as turbo coding and low-density parity-check coding, can be utilized in the channel coding process [31]. The encoded bits generated by channel encoding are then carried forward as b_n. Following channel encoding, we implement Binary Phase-shift Keying (BPSK), a widely used modulation technique. BPSK alters the phase of a carrier signal based on the encoded bits b_n, resulting in modulated signals denoted as s_n. Finally, we take into account the multi-path channel within the OFDM, represented as $\overrightharpoon{H}_{\mathrm{c}}$. In the wireless channel, each modulated bit s_n is allocated to a subchannel, denoted as h_n, and is then ready for transmission over that subchannel. This approach allows for the simultaneous transmission of multiple modulated bits over different subchannels, the channel gains in wireless multi-path channel is represented as

$$\overrightharpoon{H}_{\mathrm{c}} = [h_1, h_2, \cdots, h_{N_{\mathrm{c}}}]^{\mathrm{T}}, \tag{5}$$

where N_{c} stands for the total number of subchannels in $\mathbf{H}_c$, and h_n signifies the channel gain of the nth subchannel.

Considering the characteristics of each subchannel, the cumulative SNR of the communication process within channel $\overrightharpoon{H}_{\mathrm{c}}$ is expressed as

$$\mathrm{SNR} = \frac{\sum_{n=1}^{N_{\mathrm{c}}} \left\| h_n \cdot s_n \right\|^2}{\sum_{n=1}^{N_{\mathrm{c}}} \sigma_n^2}, \tag{6}$$

where, σ_n^2 represents the noise within the nth subchannel. The received bits after the wireless channel, marked as $s_n^{'}$, are articulated by the subsequent equation, which can be expressed as

$$s_n^{'} = s_n \otimes h_n + \sigma_n^2, \tag{7}$$

where the $\otimes$ refers to circular convolution, an operation that correlates the input signal with a finite impulse response. Subsequently, the received data $s_n^{'}$ undergoes traditional channel decoder and source decoder at the receiver to recover the original data.

C. Novel Task-oriented and Semantics-aware Framework

In this section, we provide a detailed description of our proposed TSAR framework, that not only compare with the traditional point cloud communication framework but also incorporates several task-oriented strategies, including effectiveness level optimization methodology. The TSAR framework leverages shared base knowledge and utilizes a task-oriented context at the semantic level, to exploit more efficient and effective communication for AR application. As illustrated in Fig. 2 in the next page, the modules in TSAR include semantic information extraction, task-oriented semantics-aware wireless communication, avatar pose recovery and rendering.

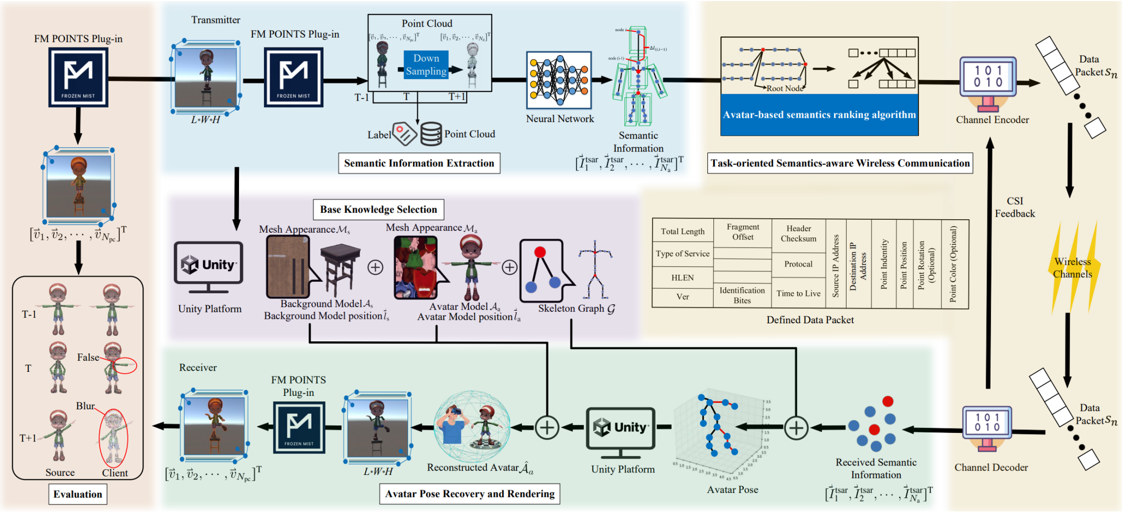

FM POINTS Plug-in
Transmitter
FM POINTS Plug-in
Point Cloud
$[\vec{v}_1, \vec{v}_2, \cdots, \vec{v}_{N_{pc}}]^T$
$[\vec{v}_1, \vec{v}_2, \cdots, \vec{v}_{N_d}]^T$
Down Sampling
T-1
T
T+1
$L*W*H$
Label
Point Cloud
Semantic Information Extraction
Neural Network
node i
node (i-1)
$\Delta l_{(i,i-1)}$
Semantic Information
$[\vec{I}_1^{tsar}, \vec{I}_2^{tsar}, \cdots, \vec{I}_{N_a}^{tsar}]^T$
Root Node
Avatar-based semantics ranking algorithm
Task-oriented Semantics-aware Wireless Communication
Channel Encoder
Data Packet s_n
$[\vec{v}_1, \vec{v}_2, \cdots, \vec{v}_{N_{pc}}]^T$
Base Knowledge Selection
Unity
Unity Platform
Mesh Appearance $\mathcal{M}_s$
Background Model $\mathcal{A}_s$
Background Model position $\vec{l}_s$
Mesh Appearance $\mathcal{M}_a$
Avatar Model $\mathcal{A}_a$
Avatar Model position $\vec{l}_a$
Skeleton Graph $\mathcal{G}$
Total Length
Type of Service
HLEN
Ver
Fragment Offset
Identification Bites
Header Checksum
Protocal
Time to Live
Source IP Address
Destination IP Address
Point Indentity
Point Position
Point Rotation (Optional)
Point Color (Optional)
Defined Data Packet
CSI Feedback
Wireless Channels
T-1
T
False
T+1
Blur
Source
Client
Evaluation
Receiver
FM POINTS Plug-in
$[\vec{v}_1, \vec{v}_2, \cdots, \vec{v}_{N_{pc}}]^T$
$L*W*H$
Reconstructed Avatar $\hat{\mathcal{A}}_a$
Avatar Pose Recovery and Rendering
Unity
Unity Platform
Avatar Pose
Received Semantic Information
$[\vec{I}_1^{tsar}, \vec{I}_2^{tsar}, \cdots, \vec{I}_{N_a}^{tsar}]^T$
Channel Decoder
Data Packet s_n

Fig. 2: Task-oriented and semantics-aware communication framework

1) Semantic Information Extraction: Unlike traditional point cloud communication framework, which primarily relies on raw point cloud data for AR scenery representation and transmission, our proposed TSAR framework provides a more sophisticated approach to extract a rich depth of semantic and effectiveness levels data from the raw point cloud. The process begins with the downsampled point cloud sensing data, $\mathbf{P}_{\text{dpc}}$, as the input. This point cloud data encapsulates all the AR scenery, which are broadly divided into two categories: the moving avatar model $\mathcal{A}_{\text{a}}$ and the stationary model $\mathcal{A}_{\text{s}}$. Only the avatar's moving position is considered essential information and needs to be refreshed at every frame. Thus, the output of this semantic information extraction process is the skeletons information of the moving avatar, $\vec{I}_i^{\text{tsar}}$, which can be represented as

$$\vec{I}_i^{\text{tsar}} = (\vec{l}_i, \vec{r}_i) = (l_{\text{x}}, l_{\text{y}}, l_{\text{z}}, r_{\text{x}}, r_{\text{y}}, r_{\text{z}}, r_{\text{w}}),\ i \in [0, N_{\text{a}}], \tag{8}$$

where N_{a} represents the total number of skeletons in the avatar, $\vec{l}_i$ represents the three-dimensional location and $\vec{r}_i$ represents the quaternion rotation of the ith skeleton in the avatar model.

Apart from quaternion rotation, current research also employs euler angles to represent rotations in AR scenery. In comparison to quaternion, euler angles offer a simpler and more information-efficient method to represent rotation and calculate root node position when a fixed root node point is available. This approach needs less information to reconstruct the avatar's pose compared to quaternion, resulting in less data packets and potentially more efficient communication [32]. The transformation from rotation to euler angles can be expressed as

$$\begin{bmatrix} e_{\text{p}} \\ e_{\text{r}} \\ e_{\text{y}} \end{bmatrix} = \begin{bmatrix} \arctan \frac{2(r_{\text{y}} r_{\text{z}} + r_{\text{w}} r_{\text{x}})}{1-2\left(r_{\text{x}}^2 + r_{\text{y}}^2\right)} \\ \arcsin\left(2\left(r_{\text{w}} r_{\text{y}} - r_{\text{x}} r_{\text{z}}\right)\right) \\ \arctan \frac{2(r_{\text{x}} r_{\text{y}} + r_{\text{w}} r_{\text{z}})}{1-2\left(r_{\text{y}}^2 + r_{\text{z}}^2\right)} \end{bmatrix} * \frac{180}{\pi}. \tag{9}$$

where the e_{p}, e_{y}, and e_{r} are defined as the pitch, roll, and yaw in euler angles to represent rotations around the three primary axes with an associated root point. The semantic information of the AR application, denoted as $\mathbf{D}_{\text{tsar}}$, represents all the skeleton information $\vec{I}_i^{\text{tsar}}$ of the avatar model generated through a semantic information extraction process from the downsampled point cloud $\mathbf{P}_{\text{dpc}}$, which can be expressed as

$$\mathbf{D}_{\text{tsar}} = [\vec{I}_1^{\text{tsar}}, \vec{I}_2^{\text{tsar}}, \cdots, \vec{I}_{N_{\text{a}}}^{\text{tsar}}]^{\text{T}} = \mathcal{S}(\mathbf{P}_{\text{dpc}}, \theta_{\text{s}}), \tag{10}$$

where $\mathcal{S}(\cdot)$ represents the semantic information extraction process, and θ_{s} encompasses all the experimental and neural network parameters. This equation represents the entire semantic information extraction process, which maps the downsampled point cloud data $\mathbf{P}_{\text{dpc}}$ to a more meaningful semantic representation $\mathbf{D}_{\text{tsar}}$ for further transmitting over wireless channels.

2) Task-oriented Semantics-aware Wireless Communication: Building upon the extracted semantic information, we develop an avatar-based semantic ranking algorithm to integrate task-oriented semantic information ranking into end-to-end wireless communication to exploit the importance of semantic information to an avatar-based AR displaying task. The algorithm correlates the importance evaluation of semantic information and task relevance with channel state information feedback, thereby prioritizing more important semantic information for optimal transmission over more reliable subchannels. More specifically, each skeleton is represented as a node in the avatar skeleton graph $\mathcal{G}$ as shown in the Fig. 4, and the skeleton ranking is determined by a calculated weight in the skeleton graph, which indicates the level of importance in the later avatar pose recovery. The weights of all semantic information $\mathbf{D}_{\text{tsar}}$ are denoted as $\overrightarrow{W}_{\text{tsar}}$ and can be formulated as

$$\overrightarrow{W}_{\text{tsar}} = [\omega_{I_1}, \omega_{I_2} ..., \omega_{I_{N_{\text{a}}}}]^{\text{T}} = \mathcal{W}(\mathbf{D}_{\text{tsar}}, \mathcal{G}), \tag{11}$$

where w_{I_i} represents the weight of the semantic information of the ith skeleton in avatar skeleton graph, these node weights essentially represent the importance of the semantic information to the avatar representation, with higher weights indicating greater importance of the skeleton information for avatar pose recovery. By correlating these weights representing the importance of semantic information with Channel State Information (CSI) feedback during wireless communication, the effectiveness of the avatar transmission in AR application could be optimized. Specifically, the semantically important information is mapped and transmitted over more reliable subchannels. Current research in the OFDM has demonstrated that CSI can be accurately estimated at the transmitter side using suitable algorithms and feedback mechanisms [33]. Consequently, the subchannel gains h_n at the receiver side are assumed to be added in the CSI feedback, enabling the transmitter to be aware of the accurate all the subchannel state in the OFDM. According to Eq. (6), the subchannel with a higher SNR will have a better subchannel state and thus achieve a more reliable transmission for semantic information. Therefore, an ascending sorting is employed to establish a mapping function $\mathcal{M}(\cdot)$ between the semantic information and various subchannels. This mapping relies on the weights calculated for the semantic information and the CSI. Higher weights, indicating greater importance of the semantic information in the avatar pose recovery, are assigned to more reliable subchannels. The mapping function is expressed as

$$\mathcal{M}(\overrightarrow{W}_{\text{tsar}}, \mathcal{G}, \vec{H}_{\text{c}}) = \{\vec{I}_i^{\text{tsar}}, h_j\}, i \in [1, N_{\text{a}}], j \in [1, N_{\text{c}}], \tag{12}$$

where the map $\{\vec{I}_i^{\text{tsar}}, h_j\}$ refers to transmit the semantic information $\vec{I}_i^{\text{tsar}}$ at the subchannel h_j. Based on the channel mapping results, each semantic information is transmitted through different subchannels in the OFDM subchannels.

3) Avatar Pose Recovery and rendering: In contrast to traditional point cloud wireless communication framework, the TSAR framework approaches avatar pose recovery differently with the transmission of the base knowledge at the beginning of AR application. As illustrated in Fig. 2, the data could be used for base knowledge $\boldsymbol{B}_*$ encompasses different types of information, which include avatar skeleton graph $\mathcal{G}$, avatar initial position l_o, avatar model $\mathcal{A}_{\text{a}}$, stationary background model $\mathcal{A}_{\text{s}}$, stationary initial position l_s, and their respective appearance meshes, $\mathcal{M}_{\text{a}}$ and $\mathcal{M}_{\text{s}}$. Whenever a new 3D object appears in the AR scenery, the base knowledge at both transmitter and receiver need to be updated synchronously.

In this way, the TSAR framework considers the avatar as a whole entity and recover the avatar's pose using a limited set of skeleton points instead of treating individual points as the smallest recovery unit. The avatar pose recovery process $\mathcal{R}(\cdot)$ can be expressed as

$$\hat{\mathcal{A}}_a = \mathcal{R}(\mathbf{D}_{\text{tsar}}^{'}, \boldsymbol{B}_{\text{tsar}}), \tag{13}$$

where $\boldsymbol{B}_{\text{tsar}}$ represents the base knowledge of TSAR, and $\hat{\mathcal{A}}_a$ denotes the avatar model $\mathcal{A}_{\text{a}}$ with appearance $\mathcal{M}_{\text{a}}$ after pose recovery with semantic information $\mathbf{D}_{\text{tsar}}^{'}$.

The AR displaying process is quite straightforward by presenting the reconstructed avatar $\hat{\mathcal{A}}_{\text{a}}$ and the stationary background model S_o in the AR scenery. The process of avatar pose recovery in the TSAR framework is intricately designed and hinges on associating each piece of skeleton information $\vec{I}_i^{\text{tsar}}$ with the avatar model $\mathcal{A}_a$ on the Unity3D platform. In traditional point cloud communication frameworks, the entire point cloud data must be refreshed for each frame, which can be a computationally expensive and time-consuming process. In contrast, the TSAR framework only requires the updating of the skeleton information associated with the avatar's movements, and update the avatar's pose based on these information.

D. Problem Formation

In summary, the overall framework aims to achieve task-oriented semantics-aware communication with efficient data transmission for better avatar representation in wireless AR applications. The primary objective of the framework is to maximize the client-side AR viewing experience based on the transmitted semantic information. The objective function can be represented as

$$\begin{aligned} \mathcal{P}: \quad & \min_{\{\theta_\mathrm{s}, \left(\vec{I}_i, h_j\right)\}} \lim_{T \to +\infty} \frac{1}{T} \sum_{t=0}^{T} \sum_{i=0}^{N_a} \left(\vec{I}_{i,t}^{\mathrm{tsar}} - \vec{I}_{i,t}^{\mathrm{tsar}'} \right) \cdot \omega_{I_i}, \\ \text{s.t.} \quad & i \in [1, N_\mathrm{a}], \quad j \in [1, N_\mathrm{c}], \end{aligned} \tag{14}$$

where $\vec{I}_{i,t}^{\mathrm{tsar}}$ represents the semantic information of the ith skeleton at time t, and $\vec{I}_{i,t}^{\mathrm{tsar}'}$ is the received semantic information after the wireless channel. The weights ω_{I_i} reflect the importance of each skeleton node i in representing the avatar graph. This equation formulates the problem of minimizing the error in avatar representation during transmission.

III. Semantic Level Design

In this section, we will discuss the semantic extraction and recovery blocks, including semantic information extraction with deep learning, base knowledge selection, avatar pose recovery, and evaluation metric.

A. Semantic Extraction with Deep Learning

Inspired by the KeypointNet proposed in [24], we propose a semantics-aware network called SANet to extract the skeleton keypoint information of a moving avatar from the whole point cloud of AR scenery. The extraction is an integral step towards creating a more interactive and

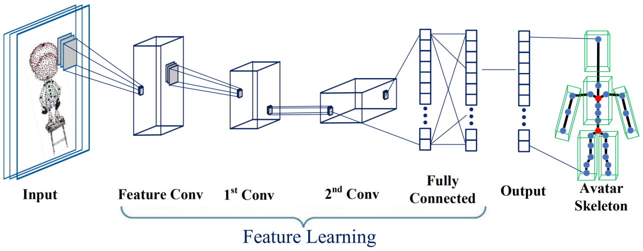


Fig. 3: Semantic Information Extraction Network

TABLE I: SANet parameters and training setup

Parameter	Value
Cell	
Semantic network	In (2048,3), out (25,1)
Feature conv	(In feature=2048, out feature=1440)
1^{st} Conv2d	(In feature=256, out feature=256)
2^{nd} Conv2d	(In feature=256, out feature=128)
Output layer	(In feature=128, out feature=25)
Simulation	
Learning rate	10^{-4}
Optimizer	Adam
Episode	900
Activation function	ReLU

immersive augmented reality experience. The SANet operates by using downsampled point cloud data $\mathbf{P}_{\text{dpc}}$ as input, which represents the 3D coordinates of both the stationary models and the moving avatar. This data is then processed by the SANet to extract accurate avatar skeleton information, crucial for reproducing the avatar's movements in the virtual environment. The design objective of the SANet is to minimize the Euclidean distance ($\mathcal{L}_2$) between the predicted semantic information, denoted as $\mathcal{S}(\mathbf{D}_{\text{dpc}})$, and the labeled semantic information of the skeleton location, represented as $\mathbf{D}^{l}_{\text{tsar}}$. The interplay between these variables is captured as

$$\text{Loss} = \arg\min_{(\theta_{\text{s}})} \mathcal{L}_2 \left(\mathcal{S}(\mathbf{P}_{\text{dpc}}), \mathbf{D}^{l}_{\text{tsar}} \right). \tag{15}$$

where θ_{s} represents all the neural networks and experiment parameters in the SANet, which is defined in Table I and Fig. 3. Training the SANet involves optimizing these parameters to minimize the loss, thus enhancing the accuracy of semantic information extraction.

To determine the most suitable backbone for the designed SANet, we train the SANet with various backbone networks, including ResNet, RsCNN, PointNet, SpiderCNN, PointConv, and DGCNN [34]. Similar to [24], we use the mean Average Precision (mAP) as the performance evaluation metric to assess the semantic information extraction accuracy of the predicted keypoint probabilities in relation to the ground truth semantic information labels.

B. Base Knowledge Selection

To better explore the most suitable base knowledge, we propose basic TSAR framework (TSAR) and euler angle based TSAR framework (E-TSAR) that considers different shared base knowledge and semantic information definition[1].

TSAR: For the basic TSAR framework, semantic information for each skeleton is defined as the data pertaining to position and quaternion rotation as in Eq. (8). The shared base knowledge, denoted as $\boldsymbol{B}_{\text{tsar}}$, comprises the stationary background model, stationary model initial position moving avatar model, and their corresponding appearance meshes, which is denoted as

$$\boldsymbol{B}_{\text{tsar}} = \{\mathcal{A}_{\text{o}}, \mathcal{A}_{\text{s}}, \mathcal{M}_{\text{o}}, \mathcal{M}_{\text{s}}, \vec{l}_{\text{s}}\}. \tag{16}$$

E-TSAR: As an extension of TSAR, the semantic information in each skeleton I_i is defined as the euler angle rotation in E-TSAR, according to Eq. (9), which could be defined as

$$\vec{I}_i^{\text{etsar}} = (\vec{e}_i) = (e_{\text{r}}, e_{\text{y}}, e_{\text{p}}),\ i \in [0, N_{\text{a}}], \tag{17}$$

where the shared base knowledge $\boldsymbol{B}_{\text{etsar}}$ encompasses the avatar skeleton graph, avatar initial position, stationary background model, stationary model initial position, moving avatar model, and their appearance meshes, defined as

$$\boldsymbol{B}_{\text{etsar}} = \{\mathcal{M}_{\text{a}}, \mathcal{M}_{\text{s}}, \mathcal{A}_{\text{a}}, \mathcal{A}_{\text{s}}, \vec{l}_{\text{a}}, \vec{l}_{\text{s}}, \mathcal{G}\}. \tag{18}$$

C. Avatar Pose Recovery

The avatar pose recovery involves using the skeleton graph $\mathcal{G}$ in the base knowledge and the received semantic information to reconstruct the avatar pose. The entire avatar pose recovery process is shown in **Algorithm 1**. Specifically, a recursive algorithm is employed to traverse and assign all skeleton information to the avatar model $\mathcal{A}_a$ with initialized parameters. However, due to differences in the definition of the semantic information and the shared base knowledge, the avatar poses recovery process has variations between the TSAR and E-TSAR framework.

On the one hand, the basic TSAR framework employs a simple avatar pose recovery method, assigning the avatar model with value based on the skeleton point identity using the received

[1] Semantic information, as presented in Fig. 2, consists of the skeleton information that need to be transmitted in every frame. Conversely, base knowledge encompasses information used primarily in the first frame.

position vector and quaternion rotation. On the other hand, the E-TSAR framework, which only transmits the euler angle of each skeleton point as semantic information, requires calculating each skeleton position with respect to its root point in the skeleton graph before assigning the skeleton information to the avatar model. The E-TSAR framework reconstructs the avatar pose by first determining the relationships between the skeleton points in the avatar skeleton graph $\mathcal{G}$. It then computes the position of each skeleton point by considering its euler angle and the position of its root point within the $\mathcal{G}$, the relative distance vector $\Delta\vec{l}_{(i,i-1)}$ between the ith skeleton node and the previous $(i-1)$th node can be represented as

$$\Delta\vec{l}_{(i,i-1)} = (\Delta\mathrm{x}, \Delta\mathrm{y}, \Delta\mathrm{z}) = \vec{e}_i \times \vec{l}_{i-1}, \tag{19}$$

where e_i represents the eular angle of the ith skeleton node, $(\Delta\mathrm{x}, \Delta\mathrm{y}, \Delta\mathrm{z})$ represents the distance between two skeleton node towards the x, y, and z coordinates, and the actual position of the ith skeleton node will be calculated by combining $\Delta\vec{l}_{(i,i-1)}$ and $\vec{l}_{i-1}$, which can be expressed as

$$\vec{l}_i = \vec{l}_{i-1} + \Delta\vec{l}_{(i,i-1)}, \tag{20}$$

where the root node position $\vec{l}_0$ is equal to the avatar initial position $\vec{l}_{\mathrm{a}}$ in the base knowledge, and $\vec{l}_i$ represents the position of the ith skeleton node in the avatar, with its three components representing the x, y, and z coordinates respectively.

D. Evaluation Metric

The semantic level of our proposed TSAR aims to enhance the communication effectiveness to achieve accurate avatar moving of the AR application, specifically, the skeleton information accuracy between the transmitter and the receiver. The optimization seeks to minimize the Euclidean distance of the semantic information transmitted at the transmitter and received at receiver. Thus, the MPJPE is used to estimate and evaluate the avatar pose error in geometry aspect between the transmitter and receiver, including the x-axis, y-axis, and z-axis values, which can be expressed as

$$\text{MPJPE } = \frac{1}{N_{\mathrm{a}}} \sum_{i=1}^{N_{\mathrm{a}}} \sqrt{|\vec{l}_i - \vec{l}'_i|^2}, \tag{21}$$

where the $\vec{l}_i$ and $\vec{l}'_i$ represent the three dimensional position value of skeleton at the transmitter and the receiver respectively.

Algorithm 1 Avatar Pose Recovery

```
Initialization: Received base knowledge $\boldsymbol{B}_*$, received data $\mathbf{D}'_{\text{tsar}}$
Get skeleton graph $\mathcal{G}$, avatar initial position $\vec{l}_a$ avatar model $\mathcal{M}_a$, and avatar appearance mesh $\mathcal{A}_a$ from $\boldsymbol{B}_*$
Count the skeleton number $N_{\text{a}} = \mathbf{C}_{\text{s}}(\mathcal{G})$
Count the received semantic information $N_{\text{r}} = \mathbf{C}_{\text{r}}(\mathbf{D}'_{\text{tsar}})$
if ($\mathcal{G} \notin \boldsymbol{B}_*$ & $l_i \in \mathbf{D}'_{\text{tsar}}$) then
    for each $i$ in $N_{\text{r}}$ do
        Attach $\vec{l}_i^{\text{tsar}}$ to model $\mathcal{A}_a$ (Avatar pose recovery for the TSAR)
    end for
else
    for each $i$ in $N_a$ do
        update $\vec{l}_i$ according to Eq. (20) and Eq. (19)
        Attach $\vec{l}_i^{\text{etsar}}$ to model $\mathcal{A}_a$ (Avatar pose recovery for the E-TSAR)
    end for
end if
Generate avatar $\hat{\mathcal{A}}_a$ with appearance mesh $\mathcal{M}_a$ and model initial position $l_a$ according to Eq. (13).
Output: Avatar $\hat{\mathcal{A}}_a$ with reconstructed pose
```

IV. Effectiveness Level Design

In this section, we will demonstrate the design principles of TSAR optimization at the effectiveness level based on the above defined semantic information. In the following, we present task-oriented semantics-aware wireless communication and its evaluation metric.

A. Task-oriented Semantics-aware Wireless Communication

To further enhance the effectiveness of avatar communication in AR applications, we propose an avatar-based semantic ranking algorithm to calculate an importance weight value among all the extracted semantic information, which plays a more advantageous role in avatar representation. More specifically, we calculate the importance of the skeleton nodes in the skeleton graph $\mathcal{G}$

using a ranking method based on the PageRank algorithm proposed by Google [35], the detailed process of AbSR algorithm is proposed in **Algorithm 2**, and the weight is calculated as

$$\omega_{I_i} = \frac{N_J}{(1-\alpha)} + \sum_{j=0}^{N_J} \left(|\Delta \vec{l}_{(i,j)}| \times \omega_{J_j} \right). \tag{22}$$

where ω_{I_i} represents the weight of the semantic information $\vec{I}_i$ in the ith skeleton node of skeleton graph, and $|\Delta \vec{l}_{(i,j)}|$ denotes the Euclidean distance between the ith and jth skeleton. J_j denotes the node index which are connected to the ith node, ω_{J_j} is the weight value of the J_jth skeleton, N_{j} represents the total number of nodes J_j in the skeleton graph, and α is a discount factor ranging from 0 to 1. As suggested in [36], we set the discount factor to 0.7 in this paper. A detailed diagram is shown in Fig. 4, which illustrates that skeletons with more connections and longer distances from other connected skeletons are more critical. The underlying rationale is that a node with more connections will have a greater impact on connected skeleton nodes if it have bit error in wireless communication. Furthermore, nodes that are more isolated, indicated by their greater distance from other skeletons, are likely to have a more substantial impact on the avatar representation due to their distinctive appearance contributions, highlighting the importance of these skeletons.

After calculating the critical node weight of skeleton graph, a descending sort algorithm is applied to arrange the skeleton nodes in descending order of rank. Leveraging our proposed AbSR algorithm, we consider the effectiveness level optimization during the wireless communication, focusing on avatar semantic preservation. This shift advancing the semantic level design in Section III, thus ensuring that crucial avatar semantic information is prioritized in our task-based wireless communication approach. As shown in Eq. (12), this approach maps higher weight semantic information to transmit in OFDM subchannels with better CSI. This is the so called euler angle and channel-based TSAR framework (EC-TSAR), with details below.

EC-TSAR: Based on the E-TSAR, the CSI information is considered to implement the AbSR and channel mapping in **Algorithm 2** to improve communication effectiveness in AR application. The semantic information is defined as the vector position and euler angle rotation of all skeletons in the moving avatar as shown in Eq. (17), while the base knowledge encompasses the avatar skeleton graph, shared background model, moving avatar model, and their appearance meshes, as shown in Eq. (18).

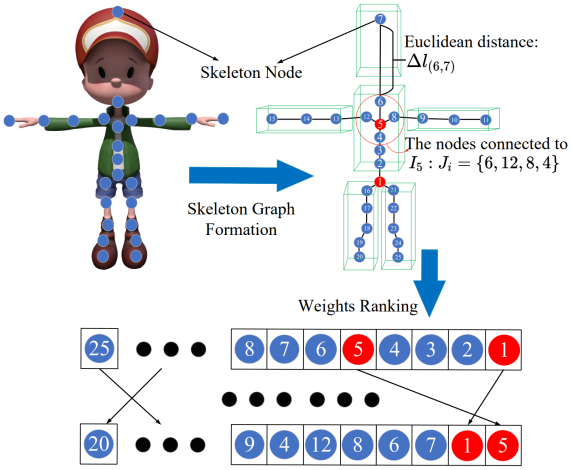


Fig. 4: Skeleton Graph Formation and Ranking

Algorithm 2 Avatar-based Semantic Ranking Algorithm

```
Initialization: Base Knowledge $\boldsymbol{B}_*$, Semantic
information $\mathbf{D}_{\text{tsar}}$
Get $\mathcal{G}, \mathcal{A}_a$ from $\boldsymbol{B}_*$,
Get $\Delta\vec{l}_{(i,i-1)}$ from $\mathcal{A}_a$
Count skeleton number $N_{\text{a}} = \mathbf{C}_{\text{s}}(\mathcal{G})$
repeat
    $k = k + 1$
    for each $i$ in $N_{\text{a}}$ do
        Update $\omega_{I_i}^k$ with $\Delta\vec{l}_{(i,i-1)}$ based on Eq. (22)
        $\delta = ||\omega_{I_i}^k - \omega_{I_i}^{k-1}||$
    end for
until $\delta < \varepsilon$
Update $\{\vec{I}_i^{\text{tsar}}, h_j\}$ according to Eq. (12)
Output: Channel Mapping $\{\vec{I}_i^{\text{tsar}}, h_j\}$
```

B. Evaluation Metric

Building upon semantic level optimization, the overall goal of the task in AR application is to recover the avatar for better clients viewing experience. To achieve this, we use point cloud to evaluate the entire virtual scenery, which includes Point-to-Point (P2Point), Peak Signal-to-Noise Ratio for the luminance component (PSNR$_\text{y}$), and transmission latency:

P2Point: To evaluate the viewing experience of clients in AR applications, the P2Point metric is employed to assess the AR scenery from a 360° viewing angles, comparing the geometry difference between the point cloud data at transmitter $\mathbf{P}_\text{t}$ and the point cloud data at receiver $\mathbf{P}_\text{r}$. The P2Point error calculation can be expressed as

$$\text{P2Point} = \max\left(d_\text{rms}^{(\mathbf{P}_\text{t},\mathbf{P}_\text{r})}, d_\text{rms}^{(\mathbf{P}_\text{r},\mathbf{P}_\text{t})}\right), \tag{23}$$

where the function d_rms is the root mean square error between two point cloud.

PSNR$_\text{y}$: The color difference plays a crucial role in avatar displaying task of AR applications, as it can significantly impact the user viewing experience if there are discrepancies in the colors transmitted. The PSNR$_\text{y}$ is used to evaluate the luminance component of the AR scenery difference between the receiver and transmitter . The PSNR$_\text{y}$ is then calculated as

$$\text{PSNR}_\text{y} = 10\log_{10}\left(\frac{255^2}{\frac{1}{N_\text{t}}\sum_{\vec{v}_i\in\mathbf{P}_\text{t}}\left[y_{\vec{v}_i} - y_{\vec{v}_\text{near}^{\mathbf{P}_\text{r}}}\right]^2}\right), \tag{24}$$

where $\vec{v}_\text{near}^{\mathbf{P}_\text{r}}$ represents the nearest point to $\vec{v}_i$ from point cloud $\mathbf{P}_\text{r}$, N_t represents the total number of point cloud in the $\mathbf{P}_\text{t}$, and $y_{\vec{v}_i}$ represents the luminance elements of point $\vec{v}_i$.

Transmission Latency: Transmission Latency is a critical metric in AR applications and plays a crucial role in evaluating client QoE. The transmission latency of the AR application can be divided into different components, including semantic information extraction time T_s, wireless communication time T_w, avatar pose recovery and rendering time T_r. The combination of all these times results in the transmission delay of the AR application, which can be expressed as

$$\text{Transmission Latency} = T_\text{s} + T_\text{w} + T_\text{r}, \tag{25}$$

by analyzing and optimizing each component of the transmission latency, we can justify and indicate the efficiency of our proposed framework.

V. Simulation Results

In this section, we evaluate the performance of our proposed TSAR framework and compare it with the traditional point cloud communication framework as well as the enhanced frameworks such as E-TSAR and EC-TSAR, as described in section III and IV. For assessing the performance

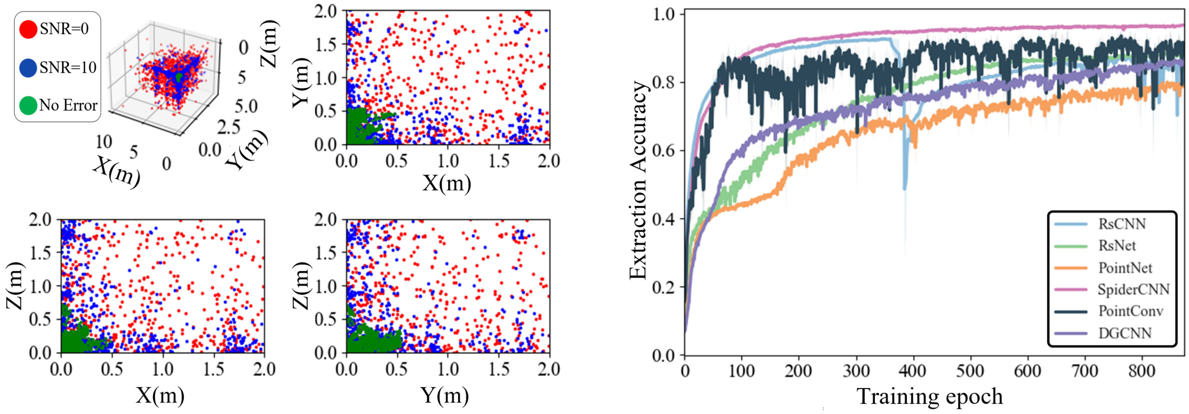
TABLE II: Experiment Setup

Dance type	***Last time***
Upper body dance	2min 10s
Slight shaking	50s
Full body dance	2min 5s
Simulation	***Value***
Data type	Point cloud
FPS	60
Avatar skeleton number	25
Stationary model skeleton number	15
Point cloud number	2,048
Attribute information 1	Point index
Attribute information 2	Position
Attribute information 3	Rotation (optional)
Attribute information 4	Color (optional)

of the semantic information extraction, we use several different avatar dance types as specified in Table I, and we configure the hyperparameters for the SANet as listed in Table II. The SANet initially undergoes a learning phase where it is trained until it converges to an optimal state. Once the training phase is complete, the resulting trained neural network is implemented across TSAR, E-TSAR, and EC-TSAR. The subsequent sections present the results of our proposed frameworks. Section V-A offers insights into the avatar movement distribution and Section V-B first provides data on the semantic information extraction accuracy achieved by the SANet, and following that, we present experimental results examining various metrics to evaluate the XR application and avatar transmission. These metrics include the MPJPE, the adjacent frame MPJPE, transmission latency, P2Point, and PSNR_y.

A. Avatar Skeleton Distribution

To obtain a comprehensive understanding of avatar movement in the AR environment, several avatar dance types were conducted upon the Unity3D and Mixamo platform. Mixamo is a robust 3D character creation and animation tool offering a wide array of diverse and dynamic 3D character animations suitable for a broad spectrum of movement analysis. Three distinct dance



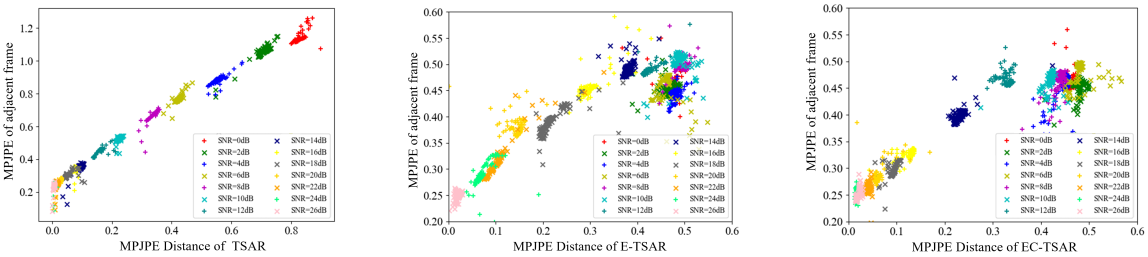
(a) Avatar movement range of adjacent frame.

(b) Semantic information extraction accuracy.

Fig. 5: Avatar movement distribution and semantic information extraction accuracy

types from Mixamo were selected for our experiments: an upper-body dance, a slight shaking dance, and a full-body dance. These dances cover a wide range of avatar movements, from localized to full-body motions, and each dance has a specific duration, as detailed in Table II. The transmitter used for these experiments operates at 60 Frames Per Second (FPS), ensuring a smooth and continuous displaying of the avatar's movements at the transmitter. The moving avatar, with 25 skeletons, is placed on a stationary background stage model.

Fig. 5 (a) plots the data analysis of the experiments, which is carried out based on the skeleton difference between the adjacent frames across the X, Y, and Z axes under different SNR sceneries. Green points correspond to adjacent frame skeleton position differences under optimal wireless channels, which reveals that the shifts in position from one frame to the next were typically minimal. The adjacent difference ranges for the three axes are (0, 0.46), (0, 0.48), and (0, 0.48) meters, respectively, suggesting that the maximum movement of the avatar's skeleton usually remains less than 0.5 meters per frame in the Unity3D platform. Furthermore, with the SNR increases, the adjacent skeleton difference indicates that the received data might be distorted under highly noisy conditions and the Rayleigh fading channel. This can result in significant positional differences between adjacent frames, potentially surpassing the realistic movement capabilities of the avatar and subsequently causing disjointed in the virtual environment.



(a) Adjacent MPJPE of TSAR. (b) Adjacent MPJPE of E-TSAR. (c) Adjacent MPJPE of EC-TSAR.

Fig. 6: Adjacent MPJPE difference among TSAR, E-TSAR, and EC-TSAR

B. Performance Evaluation

1) Semantic information Extraction Performance: Figure 5 (b) plots the semantic extraction precision of the SANet, anchored on a variety of backbone networks over equivalent training epochs. Each network exhibits commendable proficiency, corroborating the viability of employing such a deep learning mechanism to extract semantic information from point cloud data. The degree of accuracy serves as a benchmark for the effectiveness of semantic extraction capabilities, the accuracy of which is delineated as follows: SpiderCNN >PointConv >RsNet >RsCNN >DGCNN. This pecking order underscores the pronounced superiority of the SpiderCNN-based SANet, achieving an impressive accuracy surpassing 96% within the same epoch duration. As outlined in Table II, the SpiderCNN boasts a unique structural design that performs better in point cloud structure feature extraction. This advantage may become particularly obvious in handling complex, high-dimensional data such as avatars and 3D model structures. This could also illuminate the other backbone networks' less efficient processing and learning capacities. It is likely that other backbones struggle with adequately extracting and learning from the structure of point cloud structure, which could consequently impact semantic information extraction accuracy. These findings highlight the importance of not just the SANet, but also the backbone choice while performing semantic information extraction over point cloud data.

2) Avatar Transmission Performance: Fig. 6 (a) plots the MPJPE of adjacent frames, alongside the MPJPE error between the receiver and transmitter, under different wireless channel conditions for the proposed TSAR. With the diminishing SNR, a visible degradation in AR displaying fluency with uncontinued avatar movement of adjacent frames, marked by an increase in both the adjacent MPJPE and the MPJPE. This result reemphasize the insights drawn from Fig. 5 (a), signifying that a lower SNR channel generates noise and blur in the received packets, thereby increasing the MPJPE. Furthermore, with the SNR decrease below 10 dB, the MPJPE of adjacent frames amplifies with the decreasing SNR and transcends the general avatar movement range under optimal wireless channels explicated in section V-A. This demonstrates that concerning the adjacent MPJPE, with the SNR decrease, it alludes to precipitous movements of the avatar's constituent parts, potentially inducing stutters when substantial positional discrepancies arise between successive frames. Simultaneously, if the MPJPE escalates excessively, it could engender distortions in the avatar, with skeletal elements manifesting in aberrant positions, such as a foot emerging at the head. Both the uninfluenced and distortion of the avatar in the AR application could damage the viewing experience on the client side [37].

Fig. 6 (b) plots the MPJPE of adjacent frames, alongside the MPJPE error between the receiver and transmitter, under different wireless channel conditions for the proposed E-TSAR. In contrast to the outcomes of our proposed TSAR shown in Fig. 6 (a), E-TSAR profoundly decreased the MPJPE between the transmitter and the receiver with the SNR increase and achieved a 40% decrease in MPJPE within the 0dB SNR scenery. Such observations denote a smoother and more fluent avatar movement of the E-TSAR compared to the TSAR, given the E-TSAR a reduced likelihood of confronting disconcerting avatar distortions compared to TSAR. Additionally, unlike the basic TSAR results, where the MPJPE continues to increases as the SNR decreases, the E-TSAR MPJPE does not increase after the SNR drops below 10 dB. This indicates that using the avatar model as base knowledge in semantic communication helps the avatar maintain its undistorted appearance in the poor wireless channel scenarios. This improvement in avatar representation can lead to an enhanced user experience and a higher QoE for clients, thereby underscoring the effectiveness of employing the avatar model as a shared base knowledge in the domain of wireless AR implementations.

Fig. 6 (c) plots the MPJPE of adjacent frames, alongside the MPJPE error between the receiver and transmitter, under different wireless channel conditions for the proposed EC-TSAR. With a result generally similar to E-TSAR's shown in 6 (b), EC-TSAR achieves a significant decrease when the SNR increase above 10 dB, generating a more fluent video with lower adjacent frames MPJPE. This illustrates that with the assistance of the AbSR algorithm and adaptive channel mapping, more important semantic information is effectively transmitted through wire-

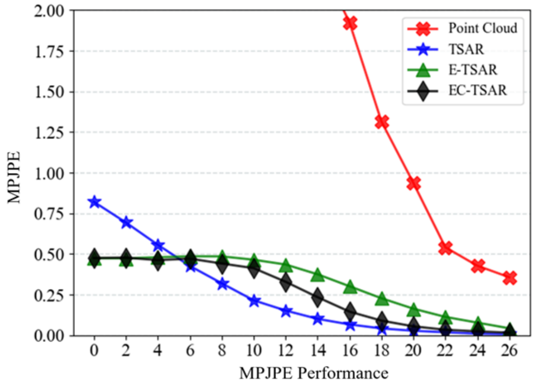
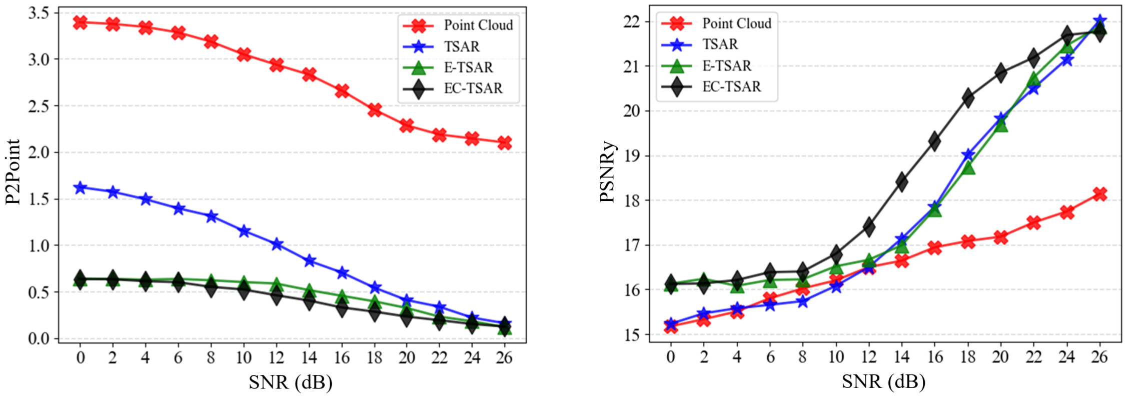


Fig. 7: Mean Per Joint Position Error.

less communication, ultimately aiding in avatar recovery on the client side. This highlights the effectiveness of the AbSR algorithm and adaptive channel mapping in improving the efficacy of avatar transmission, especially in higher SNR scenarios. Besides, similar to the E-TSAR, the MPJPE does not continue to increase as the SNR decreases below 10 dB, which reemphasizes the advantages of employing the avatar model as a shared base knowledge.

Fig. 7 plots the MPJPE performance results, which reveal the differences in the avatar skeleton's position between the receiver and transmitter. A lower MPJPE indicates a better avatar pose recovery ability in wireless communication, and the overall results of The MPJPE results are ranked as TSAR $<$ EC-TSAR $<$ E-TSAR $<$ Point Cloud. Specifically, the TSAR framework achieves the lowest MPJPE with the SNR increase above 6 dB, achieving about an 83% decrease compared to the point cloud framework at 26 dB scenery. In contrast, the EC-TSAR framework achieves lower MPJPE than the TSAR framework when the SNR continues to decrease below 6 dB. Besides, the point cloud framework struggles to generate key points within the 3D scenery with the SNR decrease below 16 dB. This observation indicates that in the cloud point communication framework, the avatars are displayed with distorted proportions, such as an arm's length longer than the avatar's entire body, which can cause the SANet to fail in distinguishing the skeleton key points accurately. Meanwhile, in the EC-TSAR, the avatar model used in the shared base knowledge functions not to allow movements exceeding the avatar's capabilities, resulting in a better and undistorted AR avatar displayed on the client side compared with other frameworks with the SNR continue to decrease below 6 dB.

Fig. 8 (a) plots the P2Point error, revealing the geometry differences of the AR scene between



(a) Point to point.

(b) Peak signal-to-noise ratio in the luminance (Y).

Fig. 8: Point to point and peak signal-to-noise ratio in the luminance(Y).

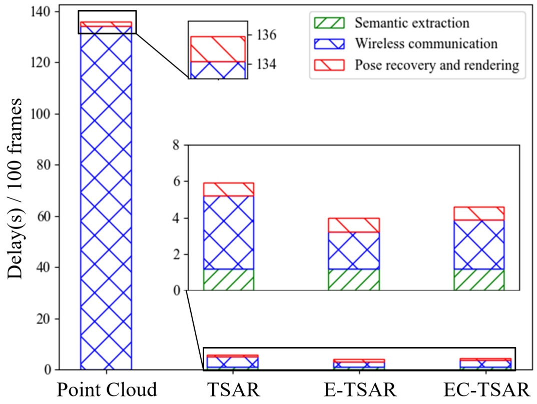


Fig. 9: Transmission Latency.

the transmitter and receiver. A lower P2Point value indicates a better viewing experience of the geometry aspect on the client side, and the overall P2Point value is ranked as EC-TSAR $<$ E-TSAR $<$ TSAR $<$ Point Cloud. With the SNR increases, the P2Point of all the frameworks witnessed an increase, indicating all the frameworks are affected by the worse wireless channel conditions. Besides, The EC-TSAR and E-TSAR frameworks both achieve a flat P2Point value increase with the SNR decrease below 8 dB compared with TSAR and Point Cloud, indicating that the avatar model transmitted in the base knowledge works to prevent the avatar displaying distortion, and make avatar only generates some odd positions in both frameworks, while the avatar displaying in the point cloud framework and TSAR already shows distortion.

Fig. 8 (b) plots the PSNR_y results, which reveal the color differences of the AR displaying scenery between the transmitter and receiver. A higher PSNR_y value represents a better viewing experience on the client side, and the PSNR_y results are ranked as EC-TSAR $>$ E-TSAR $>$ TSAR $>$ Point Cloud. All the frameworks shown an increase with the SNR increase, indicating the viewing experience is affected by the wireless channel conditions. Besides, all the TSAR, E-TSAR, and EC-TSAR achieve a significant increase when the SNR increase above 14 dB, while the point cloud communication framework has a relatively flat increase. This indicates the avatar model used in the shared base knowledge makes the avatar transmitted as a whole model, which helps to more effectively transmit the exact color of the avatar model in wireless communication, whereas the color value in the traditional point cloud framework totally up to the channel conditions and will exhibit distortions through wireless communication.

Fig. 9 plots the transmission latency of all frameworks as defined in Eq. (25). A lower latency could contribute to a better QoE on the client side, which is ranked as E-TSAR $<$ EC-TSAR $<$ TSAR $<$ Point Cloud. Compared to the point cloud communication framework, the TSAR, E-TSAR, and EC-TSAR save a substantial amount of transmission time due to significantly fewer packets transmitted. Although these frameworks introduce an additional semantic information extraction step with the DL-based semantic information extractor, it only takes about one second per 100 frames, constituting only a tiny portion of the total transmission time. Concerning pose recovery and rendering, which are inherently linked to the data packets, the point cloud requires rendering all the upsampled point cloud data based on 2,048 points. Conversely, the TSAR, E-TSAR, and EC-TSAR merely require 25 skeletal points to update the pose of an already rendered avatar, thereby significantly reducing time consumption on the client side. Moreover, although both E-TSAR and EC-TSAR necessitate calculating the skeletal position according to Eq. (19) and Eq. (20) before avatar pose recovery, while the TSAR can directly update the avatar pose. The limited calculation time of 25 cycles renders the time consumption of this pose recovery and rendering process relatively uniform among TSAR, E-TSAR, and EC-TSAR. This substantial reduction in data transmission volume concurrently minimizes bandwidth usage spent on wireless communication compared with the traditional point cloud framework.

VI. Conclusion

This paper has presented a novel task-oriented and semantics-aware communication framework designed to enhance the effectiveness and efficiency of avatar-based communication in wireless

AR applications. By introducing new semantic information in AR and representing relationships between different types of semantic information using a graph, our proposed task-oriented and semantics-aware communication framework extracted and transmitted only essential semantic information in wireless AR communication, substantially reducing communication bandwidth requirements. This selective transmission of important semantic information provided a more effective approach to semantic information extraction compared to traditional communication frameworks, ensuring minimal errors and lower bandwidth usage. Furthermore, we have extracted effectiveness level features from the complete avatar skeleton graph using shared base knowledge based on end-to-end wireless communication, distinguishing it from and enhancing general semantic communication frameworks. This pioneering work opened research for further advancements in wireless AR communication frameworks. Our future work will focus on improvements by integrating other semantic features, such as model recognition and interaction, to further improve effectiveness and efficiency in the avatar-centric wireless AR application.

References

[1] H. Ning, H. Wang, Y. Lin, W. Wang, S. Dhelim, F. Farha, J. Ding, and M. Daneshmand, "A survey on metaverse: the state-of-the-art, technologies, applications, and challenges," *arXiv preprint arXiv:2111.09673*, Nov. 2021.

[2] Y. Wang, Z. Su, N. Zhang, R. Xing, D. Liu, T. H. Luan, and X. Shen, "A survey on metaverse: Fundamentals, security, and privacy," *IEEE Commun. Surveys Tuts.*, vol. 25, no. 1, pp. 319–352, Sept, 2022.

[3] F. Hu, Y. Deng, W. Saad, M. Bennis, and A. Aghvami, "Cellular-connected wireless virtual reality: Requirements, challenges, and solutions," *IEEE Commun. Mag.*, vol. 58, no. 5, pp. 105–111, 2020.

[4] F. Hu, Y. Deng, H. Zhou, T. Jung, C.-B. Chae, and A. Aghvami, "A vision of an xr-aided teleoperation system toward 5g/b5g," *IEEE Commun. Mag.*, vol. 59, no. 1, pp. 34–40, 2021.

[5] S. Van Damme, M. T. Vega, and F. De Turck, "Human-centric quality management of immersive multimedia applications," in *Proc. IEEE Conf. Netw. Softwarization (NetSoft)*, June 2020, pp. 57–64.

[6] M. Kountouris and N. Pappas, "Semantics-empowered communication for networked intelligent systems," *IEEE Commun. Mag.*, vol. 59, no. 6, pp. 96–102, June 2021.

[7] L. Yan, Z. Qin, R. Zhang, Y. Li, and G. Y. Li, "Resource allocation for text semantic communications," *IEEE Wireless Commun. Lett.*, vol. 11, no. 7, pp. 1394–1398, Apr. 2022.

[8] Z. Weng, Z. Qin, and G. Y. Li, "Semantic communications for speech signals," in *Proc. IEEE Int. Conf. Commun. (ICC)*. IEEE, June 2021, pp. 1–6.

[9] P. Jiang, C.-K. Wen, S. Jin, and G. Y. Li, "Wireless semantic communications for video conferencing," " *IEEE J. Sel. Areas Commun.*, vol. 41, no. 1, pp. 230–244, Nov. 2022.

[10] A. Maatouk, M. Assaad, and A. Ephremides, "The age of incorrect information: An enabler of semantics-empowered communication," *IEEE Trans. Commun.*, Oct. 2022.

[11] H. Zhou, X. Liu, Y. Deng, N. Pappas, and A. Nallanathan, "Task-oriented and semantics-aware 6G networks," *arXiv preprint arXiv:2210.09372*, Oct. 2022.

[12] H. Du, D. Niyato, C. Miao, J. Kang, and D. I. Kim, "Optimal targeted advertising strategy for secure wireless edge metaverse," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*. IEEE, 2022, pp. 4346–4351.

[13] C. B. Fernandez and P. Hui, "Life, the metaverse and everything: An overview of privacy, ethics, and governance in metaverse," in *Proc. IEEE Int. Conf. Distrib. Comput. Syst. Workshops (ICDCSW)*. IEEE, July 2022, pp. 272–277.

[14] L. S. Pauw, D. A. Sauter, G. A. van Kleef, G. M. Lucas, J. Gratch, and A. H. Fischer, "The avatar will see you now: Support from a virtual human provides socio-emotional benefits," *Comput. Human Behav.*, vol. 136, p. 107368, May 2022.

[15] J. S. Lemmens and I. A. Weergang, "Caught them all: Gaming disorder, motivations for playing and spending among core pok'emon go players," *Entertain. Comput.*, p. 100548, March 2023.

[16] L. A. da Silva Cruz, E. Dumi'c, E. Alexiou, J. Prazeres, R. Duarte, M. Pereira, A. Pinheiro, and T. Ebrahimi, "Point cloud quality evaluation: Towards a definition for test conditions," in *Proc. IEEE Int. Conf. Quality of Multimedia Experience (QoMEX)*. IEEE, June 2019, pp. 1–6.

[17] Q. Yang, Y. Liu, S. Chen, Y. Xu, and J. Sun, "No-reference point cloud quality assessment via domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit.*, 2022, pp. 21 179–21 188.

[18] D. Lazzarotto, M. Testolina, and T. Ebrahimi, "Influence of spatial rendering on the performance of point cloud objective quality metrics," in *Proc. 10th European Workshop on Visual Information Processing (EUVIP)*. IEEE, 2022, pp. 1–6.

[19] D. Pavllo, C. Feichtenhofer, D. Grangier, and M. Auli, "3d human pose estimation in video with temporal convolutions and semi-supervised training," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019, pp. 7753–7762.

[20] K. Yu, G. Gorbachev, U. Eck, F. Pankratz, N. Navab, and D. Roth, "Avatars for teleconsultation: Effects of avatar embodiment techniques on user perception in 3d asymmetric telepresence," *IEEE Trans. Vis. Comput. Graph.*, vol. 27, no. 11, pp. 4129–4139, 2021.

[21] Y. Xu, Q. Yang, L. Yang, and J.-N. Hwang, "Epes: Point cloud quality modeling using elastic potential energy similarity," *IEEE Trans. Broadcasting*, vol. 68, no. 1, pp. 33–42, 2021.

[22] J. Liu, N. Akhtar, and A. Mian, "Deep reconstruction of 3d human poses from video," *IEEE Trans. Artif. Intell.*, pp. 1–1, March 2022.

[23] Y. Wu, Y. Wang, S. Jung, S. Hoermann, and R. W. Lindeman, "Towards an articulated avatar in vr: Improving body and hand tracking using only depth cameras," *Entertain. Comput.*, vol. 31, p. 100303, 2019.

[24] Y. You, Y. Lou, C. Li, Z. Cheng, L. Li, L. Ma, C. Lu, and W. Wang, "Keypointnet: A large-scale 3d keypoint

dataset aggregated from numerous human annotations," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, June 2020, pp. 13 647–13 656.

[25] Z.-L. Zhang, U. K. Dayalan, E. Ramadan, and T. J. Salo, "Towards a software-defined, fine-grained QoS framework for 5g and beyond networks," in *Proc. ACM SIGCOMM Workshop Netw.-Appl. Integr. (NAI)*, Aug. 2021, pp. 7–13.

[26] Y. Huang, B. Bai, Y. Zhu, X. Qiao, X. Su, and P. Zhang, "Iscom: Interest-aware semantic communication scheme for point cloud video streaming," *arXiv preprint arXiv:2210.06808*, Oct. 2022.

[27] F. Nardo, D. Peressoni, P. Testolina, M. Giordani, and A. Zanella, "Point cloud compression for efficient data broadcasting: A performance comparison," in *Proc. IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE, March 2022, pp. 2732–2737.

[28] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," *Adv. Neural Inf. Process. Syst.*, vol. 30, Dec. 2017.

[29] A. Akhtar, Z. Li, G. Van der Auwera, L. Li, and J. Chen, "Pu-Dense: Sparse tensor-based point cloud geometry upsampling," *IEEE Trans. Image Process.*, vol. 31, pp. 4133–4148, July 2022.

[30] Y. Chen, V. T. Hu, E. Gavves, T. Mensink, P. Mettes, P. Yang, and C. G. Snoek, "Pointmixup: Augmentation for point clouds," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Springer, June 2020, pp. 330–345.

[31] Z. B. K. Egilmez, L. Xiang, R. G. Maunder, and L. Hanzo, "Development, operation, and performance of 5G polar codes," *IEEE Commun. Surv. Tutor.*, vol. 22, no. 1, pp. 96–122, 2019.

[32] L. Quintero, P. Papapetrou, J. E. Mu noz, J. De Mooij, and M. Gaebler, "Excite-o-meter: an open-source unity plugin to analyze heart activity and movement trajectories in custom vr environments," in *2022 IEEE Conf. Virtual Reality 3D User Interfaces Abstracts Workshops (VRW)*. IEEE, 2022, pp. 46–47.

[33] S. S. Thoota and C. R. Murthy, "Massive MIMO-OFDM systems with low resolution adcs: Cram'er-rao bound, sparse channel estimation, and soft symbol decoding," *IEEE Trans. Signal Process.*, vol. 70, pp. 4835–4850, 2022.

[34] S. Qiu, S. Anwar, and N. Barnes, "Dense-resolution network for point cloud classification and segmentation," in *Proc. IEEE/CVF Winter Conf. on Applications of Computer Vision (WACV)*. IEEE, 2021, pp. 3813–3822.

[35] M. A. Joshi and P. Patel, "Google page rank algorithm and it's updates," in *Proc. Int. Conf. Emerg. Trends Sci. Eng. Manage.(ICETSEM)*, 2018.

[36] A. K. Srivastava, R. Garg, and P. Mishra, "Discussion on damping factor value in pagerank computation," *Int. J. Intell. Syst. Appl.*, vol. 9, no. 9, p. 19, Sept. 2017.

[37] M. Fiedler, T. Hossfeld, and P. Tran-Gia, "A generic quantitative relationship between quality of experience and quality of service," *IEEE Netw.*, vol. 24, no. 2, pp. 36–41, March 2010.