Enriching Telepresence with Semantic-driven Holographic Communication

Ruizhi Cheng
George Mason University
rcheng4@gmu.edu

Kaiyan Liu
George Mason University
kliu23@gmu.edu

Nan Wu
George Mason University
nwu5@gmu.edu

Bo Han
George Mason University
bohan@gmu.edu

Abstract

Achieving the optimal balance of minimizing bandwidth consumption and end-to-end latency while preserving a satisfactory level of visual quality becomes the ultimate goal of live, interactive holographic communication, a fundamental building block of immersive telepresence envisioned for 6G. Nevertheless, achieving this ambitious goal poses significant challenges for mobile devices with limited computing power, considering the substantial amount of 3D data to stream, the demanding latency requirements, and the high computation workload involved. Instead of distributing immersive content *bit by bit*, in this position paper, we propose to deliver *semantic information* extracted from telepresence participants to drastically reduce Internet bandwidth usage for task-oriented applications such as remote collaboration. We contribute a taxonomy by categorizing related semantics into three different types (*i.e.,* keypoints, 2D images, and text), pinpoint the open research challenges associated with developing a practical system for each category in our comprehensive research agenda, and delve into the potential solutions for overcoming these challenges. The preliminary results from our proof-of-concept implementation that harnesses keypoint-based semantics (partially) validate the feasibility of our research agenda.

CCS Concepts

• **Information systems** → **Multimedia streaming**; • **Computing methodologies** → **Mixed / augmented reality**.

Keywords

Semantic Communication, Immersive Telepresence

HotNets '23, November 28–29, 2023, Cambridge, MA, USA

ACM ISBN 979-8-4007-0415-4/23/11.
https://doi.org/10.1145/3626111.3628184

ACM Reference Format:
Ruizhi Cheng, Kaiyan Liu, Nan Wu, and Bo Han. 2023. Enriching Telepresence with Semantic-driven Holographic Communication. In *The 22nd ACM Workshop on Hot Topics in Networks (HotNets '23), November 28–29, 2023, Cambridge, MA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3626111.3628184

1 Introduction

The way humans communicate remotely has evolved from post service (first invented around 2400 before the common era), telegraph (in 1844), and telephone (in 1876), to videoconferencing (in the 1950s). While the pandemic highlighted the importance of and significantly improved the quality of experience (QoE) for remote communication, we still largely prefer to attend conferences and hold business meetings in person, due to the inefficiency of current videoconferencing techniques (*e.g.,* lack of social signal interference such as eye contact and body language) [11, 49, 69]. Thus, it has been widely believed that hologram-based telepresence, which has been envisioned as a top use case of 6G [23, 87, 89], bears the potential to revolutionize remote communication by offering truly immersive and interactive experiences.

Holographic communication [21] benefits from the delivery of 3D content. A hologram, which can be generated with volumetric content to capture 3D objects/scenes, is typically represented by a point cloud or mesh [5, 16]. Furthermore, recent advancements in implicit neural representations, such as neural radiance fields (NeRF) [65], have gained popularity as a viable alternative for representing volumetric content [64, 77]. Nevertheless, NeRF is primarily designed for static scenes and requires prior knowledge for training, making its direct application to live, interactive holographic communication challenging (§3.2). One distinctive aspect of volumetric content is its ability to grant viewers the freedom to not only alter their viewing direction but also freely move in 3D space, known as six degrees of freedom (6DoF) motion.

While recent years have observed a growing endeavor to optimize volumetric content delivery and boost its QoE [32, 39, 47, 48, 56, 57, 102, 103, 105, 106], existing work falls short

in the following aspects. First, prior accomplishments [34, 35, 39, 48, 57, 80, 102, 103, 105, 106] have mainly focused on *video on demand* (VOD) that streams pre-recorded content. Different from VOD, live streaming facilitates more exciting use cases of holographic communication, such as telesurgery [20] and remote collaboration [90]. Second, even with medium-quality volumetric content, state-of-the-art methods still demand considerable bandwidth requirements, for example, ∼100 Mbps in ViVo [39]. Third, previous efforts [39, 48, 57] primarily target *smartphones* that display volumetric content on a 2D screen, which, compared to mixed reality (MR) headsets, leads to a barely satisfactory user experience.

To achieve a truly immersive and engaging experience for telepresence, holographic communication should *facilitate interactive and live streaming of high-quality volumetric content for MR headsets*. However, achieving this ambitious goal is challenging due to the following reasons.

• Given its 3D nature, streaming high-quality volumetric content leads to a tremendous amount of data to deliver (*e.g.,* >1 Gbps throughput [73]), even with compression [48].

• Due to heat dissipation issues [63], head-mounted displays, which offer an intuitive way to consume and interact with volumetric content, usually operate with limited computing capabilities, compared to smartphones.

• Interactive live streaming exacerbates the complexity of the problem by mandating extremely low end-to-end latency, typically less than 100ms, to ensure a desirable QoE [9, 15].

In this position paper, we argue that in order to realize the envisaged holographic communication that facilitates participants of telepresence around the world, it is imperative to remarkably reduce the bandwidth demands of delivering volumetric content while preserving high visual quality and minimizing end-to-end latency. To this end, we propose SemHolo, a first-of-its-kind semantic-driven holographic communication framework. Semantic communication is an emerging paradigm that transmits only the crucial, relevant, and useful information extracted from a vast quantity of data [36, 58, 95], instead of leveraging bit-by-bit communication.

The motivation for incorporating semantic communication into immersive telepresence stems from its task-driven nature. To accomplish a task, an exact duplication of the 3D volumetric content of remote peers is often unnecessary. The key, instead, is on the delivery of core interactions or significant events in real time. Such critical elements could include a speaker's prominent gestures and facial expressions in the online meeting or the critical maneuvers in remote surgery.

Figure 1 depicts the end-to-end pipeline of holographic communication for telepresence with traditional and our proposed semantic-based communications. For simplicity, we show only two sites with participants wearing an MR headset, and there are multiple RGB(-D) sensors capturing them.

Semantics	Comp. Overhead		Data Size	Visual Quality	Output Format
	Extract	Recon.			
Keypoint	L	H	L	M	Mesh
Image	-	H	M	H	Image
Text	H	H	L	M	PtCl/Img

Table 1: Comparison of three holographic communication semantics. L: low; M: medium; H: high. PtCl: point cloud.

Due to the resource restrictions of mobile headsets, users are served by an edge server that is capable of executing computation-intensive tasks, for example, executing deep learning (DL) models. After rendering, remote participants are displayed in real time on the MR headsets of other users.

Compared with traditional communication methods that transmit 3D data in point cloud or mesh formats, semantic-based approaches deliver mainly abstracted data, offering a promising solution for network-friendly content distribution. Nonetheless, the effective design of semantic-driven holographic communication remains largely uncharted territory, especially in networking and systems research. To shed light on it, we undertake the following efforts in designing SemHolo, representing our key contributions.

• We first establish a taxonomy by categorizing related semantic information into three distinct types: keypoints, 2D images, and text, based on the data transmitted over the Internet (§2). Our investigation reveals that while the computer vision and graphics communities have provided potential foundations of various semantics for holographic communication, none of them could satisfy all criteria for low data size, high visual quality, and real-time extraction and reconstruction, as depicted in Table 1.

• We then delve into the research challenges associated with each category and propose potential solutions (§3). For example, although image-based semantics that leverage NeRF [65] hold promise for photorealistic reconstruction, employing it in holographic communication necessitates the capability for real-time and continuous learning of the NeRF model, given that the future frames are unknown in live streaming. To this end, we propose a novel acceleration approach encompassing an initial phase of offline pre-training for the preliminary scene, subsequently complemented by frame-specific fine-tuning. On the other hand, while keypoint- and text-based semantics can significantly reduce bandwidth consumption, they bring numerous challenges for real-time reconstruction. Thus, we propose to exploit the unique features of the human visual system (*e.g.,* foveated vision [33]) and inter-frame similarities to mitigate the reconstruction overhead.

• Finally, we implement a proof-of-concept for SemHolo, leveraging keypoint-based semantics for holographic communication (§4). Our preliminary results indicate that it can

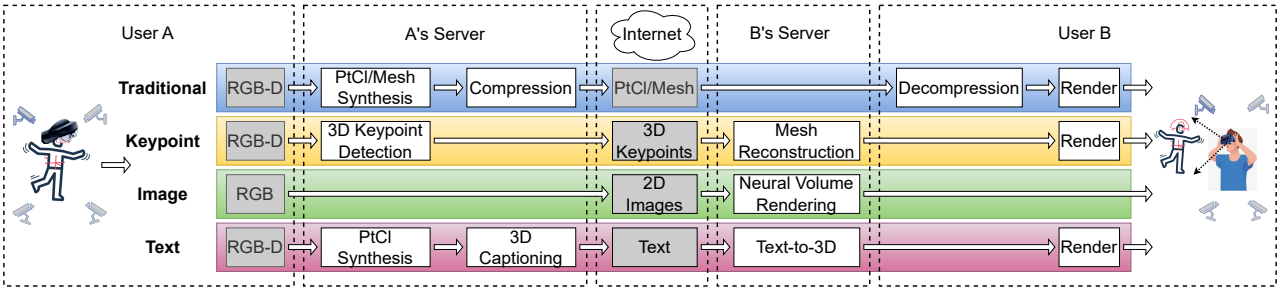

Figure 1: The end-to-end pipeline of traditional and three semantic-based approaches for holographic communication. For simplicity, we show only the communication process from A to B. The process from B to A mirrors this structure. PtCl: point cloud.

deliver bandwidth savings of up to ~207×, with the required bandwidth being only 0.30 Mbps at 30 FPS. However, such compact data representation yields subpar performance in terms of visual quality. Moreover, the intricate process of generating 3D human models from keypoints with limited information incurs considerable reconstruction overhead, leading to an extremely low frame rate (*e.g.,* <1).

2 Background

2.1 Holographic Communication

Holographic communication involves capturing, creating, delivering, and rendering volumetric content in real time. Volumetric content is usually captured with multiple RGB-D cameras positioned to cover different viewing angles [22, 25, 32, 47, 73, 99]. By merging RGB-D images from multiple cameras via synchronization, calibration, and filtering, we can obtain free-view 3D models, typically rendered by textured meshes or point clouds [60]. Recently, implicit neural representations such as NeRF [65], which employ multilayer perceptron (MLP) networks to predict scene properties for any point in 3D space, have emerged as another popular and powerful representation of volumetric content [64, 77].

To maintain satisfactory QoE, delivering traditional volumetric content in point clouds or meshes requires substantially higher network bandwidth than regular 2D video streaming. Relying solely on compressing volumetric data may not yield an optimal experience due to the on-device decoding overhead and potential network constraints [39]. Thus, recent research has pivoted toward optimization strategies for communication and computation overhead [39, 48]. However, the bandwidth requirement, even with medium-quality volumetric content (*e.g.,* ~100 Mbps [39]), still exceeds the standard broadband service in the U.S. (*i.e.,* 25 Mbps [59]).

2.2 Semantic Communication

Semantic communication embodies task-oriented processes where only pertinent, valuable, and beneficial information is extracted from original data and conveyed to receivers [36, 58, 84, 95], instead of traditional bit-by-bit transmission. This approach strategically optimizes communication overhead by underscoring the intrinsic utility of transmitted data. Early work on semantic communication centers around its literal interpretation, focusing on text data delivery [94]. Recent advancements have expanded this paradigm to encompass other modalities, such as image information [95], broadening the scope of semantic communication to other domains, for example, the emerging metaverse [55, 104].

In the context of semantic-based holographic communication for immersive telepresence, this process involves the extraction of semantic information at the sender side, transmission over the Internet, and subsequent reconstruction of volumetric content of the sender at the receiver side. Based on our investigation of existing work, we identify three types of semantics in holographic communication, which is mainly determined by what is transmitted over the Internet (§2.3).

Note that a recent work considers vectors as a form of semantics for point-cloud-based volumetric content [107]. This approach typically leverages encoder-decoder neural network architectures [4]. It maps input data to a low-dimension vector via the encoder (at the sender side) and then recovers the original data with the decoder (at the receiver side). However, it is similar to traditional point cloud compression techniques [31, 41]. Moreover, it offers a limited compression ratio and yields poor visual quality.

2.3 Taxonomy of Semantics

Keypoint-based Semantics. Keypoints represent specific and unique features of an object. For human beings, keypoints are primarily located in areas such as the body, hands, and face [14, 44]. Human keypoint detection (*a.k.a.* human pose estimation) aims to predict the positions (*i.e.,* coordinates) of body parts or joints of a person from an image or video. It has been extensively studied, typically leveraging DL models

for high accuracy [14, 44, 88, 91]. Tracking the temporal motion of keypoints can offer valuable insight into overall body posture and movements [85]. Thus, they serve as a valuable repository of semantics for holographic communication.

Most 3D keypoint detection schemes employ a two-step process: initially predicting 2D keypoints and then lifting them into 3D space (*e.g.,* by utilizing DL models) [13, 75]. Despite requiring only RGB images without depth information, this approach brings additional computational overhead for learning and inference. On the other hand, by leveraging depth information captured from off-the-shelf RGB-D cameras such as Microsoft Kinect [2], we can directly extract 3D poses, which potentially offers faster processing and higher accuracy than inferring from 2D poses [92].

Recovering the human model from keypoints, however, is non-trivial. Existing methods heavily rely on prior knowledge of the complete 3D data [54, 81], which considerably increases communication overhead. Note that given keypoints cannot encode texture information, solely relying on them for reconstruction will result in a non-clothed body structure [19, 66, 67].

Image-based Semantics capitalizes NeRF [65], an innovative technique that employs implicit neural representations for 3D scene reconstruction. NeRF allows the delivery of 2D images as semantic information for holographic communication. It utilizes a (pre-trained) MLP neural network to learn a static 3D scene from 2D images. The MLP network processes 3D spatial coordinates and viewing directions to output color and volume density at any location. By integrating the predicted color and volume density along each camera ray, which is the line that originates from the camera and passes through each pixel in the 2D image, NeRF leverages volume rendering [26] to generate a 3D scene that can be viewed from any angle. NeRF has demonstrated the potential to reconstruct high-fidelity 3D scenes [62, 64], including human models [77, 93].

Text-based Semantics lies in the intersection between 3D visual understanding [37], natural language processing, and generative AI [51]. It aims to translate 3D representations into textual descriptions and vice versa. On the sender side, the conversion of 3D data into textual information can benefit from 3D dense captioning [17, 18, 101], which primarily focuses on generating detailed text for point clouds. It involves feature extraction from the point cloud, for example, with PointNet++ [79], followed by a caption decoder to transform these features into text. On the receiver side, the emerging text-to-3D generative techniques [45, 51, 70, 78] have set the stage for reconstructing point clouds from text. Existing approaches typically utilize pre-trained text-to-2D diffusion models [71] as an initial step to convert text into 2D images. Subsequently, they leverage NeRF [65] to generate free-view 3D scenes [45, 78] or design additional diffusion models to create point clouds [70].

3 Research Agenda

In this section, we outline the research challenges and their potential solutions for each semantics in SemHolo.

3.1 Keypoint-based Semantics

The major advantage of keypoint-based semantics is its small data size, as keypoints are denoted as 2D/3D coordinates, and a modest number of keypoints (*e.g.,* ∼100 [10]) can represent the human model. While the state-of-the-art [66, 83, 96] tends to encode keypoints into parametric body models, such as SMPL-X [74], before reconstruction, the size of transferred data remains low (*e.g.,* ∼1.91 KB per frame, as demonstrated in §4). However, such low data size poses significant challenges for time-efficient and visually satisfactory content reconstruction.

Real-time Reconstruction. The sparse keypoints preclude the recovery of a detailed mesh, necessitating further processing such as additional training [19, 83] to produce a fine-grained mesh. However, doing this may exceed latency requirements, as demonstrated in §4. One potential solution is to exploit the unique feature of the human visual system that only content near foveal areas requires high resolution [33]. Therefore, we could opt to directly transmit the compressed 3D mesh for the foveal region to maintain high visual quality while delivering keypoints for only peripheral regions to reconstruct the mesh with limited refinement.

This approach poses the following challenges. First, accurately predicting the future foveal area of users is difficult due to the high-speed movements of eye gaze [7, 40]. To address this issue, one can classify gaze movements into three patterns: fixation, smooth pursuit, and saccades, determined by their speeds ranging from low to high [52]. Saccades, in particular, often account for errors in gaze prediction due to their high velocity [40, 68]. However, by leveraging saccadic omission [24], we can predict mainly the landing positions of saccades [6, 7, 68] to improve QoE. Second, there exists a trade-off between the communication overhead for delivering the 3D mesh for the foveal area and the reconstruction overhead for peripheral regions. A larger foveal area implies a higher bandwidth consumption. However, this could alleviate the burden of refining the mesh generated from keypoints in the peripheral. On the other hand, a smaller foveal area can save bandwidth usage. Nonetheless, it may require refinement for the mesh generated from keypoints to maintain a satisfactory QoE. Third, given that the reconstructed human model consists partly of the original mesh and partly of the mesh reconstructed from keypoints, seamlessly integrating these components necessitates further exploration.

High-quality Texture Alignment. Since texture could not be encoded in keypoints, the recovered mesh derived from keypoints contains only geometry, resulting in a non-clothed body structure [19, 66, 100]. A feasible solution could be to directly deliver the compressed 2D texture, given its high compression ratio and thus relatively small data size [72], to the receiver, who can then align it with the reconstructed geometry. However, considering that geometry reconstruction may result in information loss and thus the reconstructed geometry may be inconsistent with the original one, achieving high-quality texture mapping presents a significant challenge. A promising solution is to initially project the 2D texture onto the recovered 3D geometry by leveraging projection mapping [27, 28], followed by applying deformation techniques [12], which adjust the texture according to the alterations in geometry to improve visual quality.

Trade-off between Size of Transferred Data, Computation Overhead, and Visual Quality. To improve the visual quality of the reconstructed mesh, an intuitive strategy is to extract more keypoints by utilizing intricate models. While doing this may not significantly increase bandwidth requirements, given that only hundreds of keypoints should be sufficient to represent the human body [14], it inevitably heightens computational overhead. Moreover, state-of-the-art efforts [66, 83, 96] may not entirely capitalize on the additional information offered by these extra keypoints. This is because they choose to encode keypoints into parametric human models, such as SMPL-X [74], before reconstruction to enable smooth streaming. Given that these models operate with fixed parameters, the enhancements in visual quality achievable through the extraction of additional keypoints may be limited. Adding more parameters to these models may disrupt their inherent 3D rotation representations, leading to discontinuities [46].

The model-free method [19] that directly maps keypoints to 3D mesh is a potential solution to exploit the benefits of additional keypoints. However, it functions on a single-frame basis and neglects the temporal dynamics inherent in video frames. Consequently, applying such a technique to video streaming may yield unsatisfactory visual quality due to temporal discontinuity and visual artifacts [75]. In light of these limitations, we plan to develop a non-parametric, temporal-aware framework, followed by exploring the trade-offs between the number of extracted keypoints, computation overhead, and visual quality within this framework.

3.2 Image-based Semantics

By leveraging NeRF [65], image-based semantics offers two distinct advantages. First, it requires only RGB images as input, which makes it suitable for outdoor use cases for which depth sensors usually do not work well [62]. Second, with high-resolution images for training and inference, NeRF can reconstruct high-fidelity, photorealistic 3D scenes. However, in the context of live, interactive holographic communication, it poses significant challenges associated with dynamic scene reconstruction and rate adaptation.

Dynamic Scene Reconstruction. The original NeRF is designed for mainly static scenes, which is not suitable for streaming. While recent advances introduce the temporal dimension into NeRF to make it streamable, they depend on a pre-trained MLP model [8, 76, 86]. Hence, they are capable of mainly VOD services that stream pre-recorded content and are not compatible with live, interactive holographic communication where the content of future frames is unknown.

To incorporate NeRF into live, interactive holographic communication, we face the complex challenge of real-time, continuous training of NeRF models for 3D scene reconstruction. In light of this, we propose a solution hinged on the observation that changes in a user's profile over time are likely to be limited. For instance, during a meeting, the major change in the user's appearance may be only facial expressions. Thus, once a user-specific NeRF model has been trained, there is no need to retrain the model from scratch.

Based on this observation, we can include a cold start session. Before a user's inaugural engagement, we train a dedicated NeRF model. Recent research indicates that it can be completed within a few minutes [30]. Subsequently, during the user's ongoing engagement, we fine-tune the pre-trained model by feeding features extracted from the changed pixels [98]. This adjustment is designed to equip the model with the ability to reconstruct the 3D scene for the current frame, potentially expediting continuous training.

Reducing Latency with Rate Adaption. Given that image-based semantic communication involves the delivery of 2D images over the Internet, it bears similarities to traditional 2D video streaming. Since delivering multiple high-resolution 2D videos may still require substantial bandwidth [42], it is necessary to design a rate adaption scheme, for example, by adjusting the resolution of images based on the predicted bandwidth available for the receiver [43, 61].

An ideal design for accommodating inputs with diverse resolutions involves adjusting the model size in accordance with the resolution. This approach can utilize a portion of the model to handle smaller input resolutions, leading to accelerated fine-tuning and inference processes and thus diminishing end-to-end latency. This is because smaller models generally demonstrate expedited training (fine-tuning) and inference times compared to their larger counterparts [53]. However, the weight parameters within the NeRF model are intricately interconnected, implying that the omission of even a small portion of them could lead to reconstruction failure [29]. Consequently, the original NeRF architecture is ill-suited for segmentation and cannot be readily adapted to accommodate

varying input resolutions. A naive solution could involve training models of different sizes, achieved by increasing their depth and width to suit different input resolutions. Nevertheless, this would inevitably lead to significant increases in memory footprint and storage overhead.

Thus, we are investigating scalable neural networks [97], such as slimmable networks [50] and progressive networks [82]. They are designed to train a single model that can be divided into multiple executable sub-networks with different widths and layers. To enable rate adaptation, each sub-network could be trained to accommodate a particular input resolution. For example, a narrower sub-network handles low-resolution input, whereas a wider sub-network, which incorporates narrower ones, manages high-resolution input. By progressively adjusting the network parameters between narrower and wider sub-networks, we are able to dynamically adapt the model size in correlation with the input resolution.

3.3 Text-based Semantics

The principal advantage of text-based semantics is its compact data representation. Its major challenges lie in real-time semantic extraction and content reconstruction, as well as in improving the visual quality of reconstructed content.

Real-time Extraction and Reconstruction. The real-time requirements of holographic communication pose significant challenges for building blocks of text-based semantics, which heavily rely on complex DL models for extracting semantics [18, 101] and reconstructing content [45, 51, 70]. To address this problem, we propose to capitalize on the continuity of human motion, where inter-frame differences may be small. Specifically, for the first frame, we encode the information of the entire point cloud into text-based semantics. For subsequent frames, we can encode only the differences from the preceding frame, reducing the computation overhead of extraction and reconstruction. However, both existing extraction and reconstruction models for text-based semantics operate on a frame-by-frame basis, falling short of utilizing inter-frame similarities. Therefore, it is non-trivial to incorporate temporal features into them.

High-quality Reconstruction. Existing accomplishments demonstrate the capability to reconstruct only simple objects, such as cartoon avatars [78] and tableware [45, 70]. They might be inadequate for reconstructing photorealistic human models. We propose to partition the human model into cells and utilize multiple text channels to describe each cell. Doing this will add only limited communication overhead, given that the text size is usually small. On the receiver side, each channel could be reconstructed at different quality levels by leveraging content reduction techniques [33, 39]. By doing this, we can not only reduce reconstruction overhead but also maintain a satisfactory QoE.

Semantic-based		Traditional	
w/o compre.	w/ compre.	w/o compre.	w/ compre.
0.46	0.30	95.4	10.1

Table 2: Comparison of required bandwidth (Mbps) at 30 FPS for keypoint-based semantic and traditional communication approaches before and after data compression.

One associated challenge arises from the potential loss of global information, such as the overall body pose, caused by the segmentation of human models. This could lead to inaccurate reconstruction [74]. Thus, we will conduct a two-step encoding. First, we encode global features with a dedicated text channel. Following this, we design fine-grained local feature channels with reference to the global one to ensure their correctness and coherent relationship with global features during reconstruction.

4 Preliminary Results and Discussion

To better understand the challenges faced by SemHolo, we build a proof-of-concept that partially implements keypoint-based semantics for holographic communication (§3.1).

4.1 Experiment Setup

We utilize X-Avatar [83], a state-of-the-art model for generating human meshes from keypoints. It encompasses two networks. The first one takes 3D keypoints aligned with SMPL-X parameters as input and outputs geometry. The second one leverages created geometry and original RGB-D data to learn texture, which is needed because X-Avatar aims to apply the model to unseen humans. However, as proposed in §3.1, given that we possess the ground-truth texture for telepresence, we could deliver the compressed textures instead of learning them. Therefore, we retrain the X-Avatar model without the texture part. X-Avatar can tune the output resolution, indicated by the number of voxels along each dimension. Specifically, we generate mesh at resolutions of 128, 256, 512, and 1024, enabling us to explore the trade-off between visual quality and computation overhead. We employ the RGB-D image datasets released by X-Avatar and its provided 3D poses for our experiments.

4.2 Experimental Results

Data Size. We first compare the size of transferred data over the Internet between traditional and semantic communication methods. Table 2 shows the required bandwidth at 30 frames per second (FPS) before and after compression for two communication approaches. For semantic communication, the transmitted data is the 3D pose aligned with SMPL-X. The data size per frame before compression is 1.91 KB (*i.e.,*

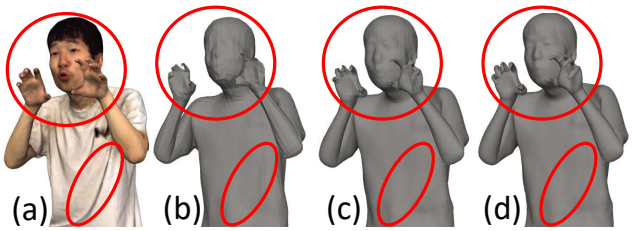

Figure 2: (a): Textured mesh generated from RGB-D data; (b) – (d): Mesh (without texture) generated from keypoints with output resolutions of 128, 256, and 1024, respectively. The visual quality of reconstructed mesh at 512 resolution is similar to that of 1024 resolution.

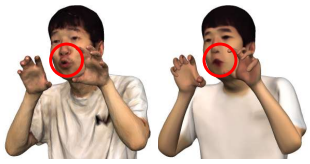

Figure 3: Textured mesh from RGB-D data (left) and from learning with resolution of 1024 (right).

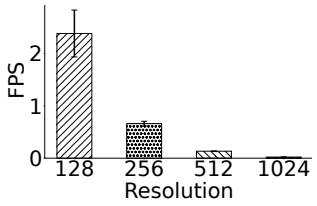

Figure 4: Reconstruction FPS of different mesh resolutions on an NVIDIA A100 GPU.

0.46 Mbps at 30 FPS). After compressing it with the Lempel–Ziv–Markov chain algorithm (LZMA) [38], the data size per frame is 1.23 KB (*i.e.,* 0.30 Mbps at 30 FPS). For the traditional approach, the transmitted data is the untextured 3D mesh of the sender generated from SMPL-X parameters.

In this case, the data size per frame is 397.7 KB (*i.e.,* 95.4 Mbps at 30 FPS) before compression and 42.1 KB (*i.e.,* 10.1 Mbps at 30 FPS) after compression with Draco [1]. Thus, keypoint-based semantic communication can potentially offer ~207× and ~34× bandwidth savings before and after compression, respectively. Note that while the required bandwidth at 30 FPS for compressed mesh in our experiment is only ~10 Mbps, it does not involve textures, and the human model used for experiments is still not photorealistic. The envisioned telepresence that renders photorealistic 3D human models may significantly increase bandwidth requirements [73].

Visual Quality. Figure 2 shows the 3D mesh directly generated from the raw RGB-D data (with texture) as the baseline and reconstructed from keypoints (without texture). The preliminary results indicate that in general, as the resolution increases, the detail in the mesh generated from keypoints augments. Specifically, at the resolution of 1024, the generated mesh is capable of revealing intricate details such as hand joints and facial contours. However, it still cannot recover the details of the clothes, such as folds.

Figure 3 presents the mesh with texture from the raw RGB data and that generated with the learning approach of X-Avatar at a resolution of 1024. The mesh learned by X-Avatar fails to accurately mirror detailed expressions. For instance, in the mesh created from raw RGB-D data, the person displays an open mouth with a pout. However, the learned mesh only reflects the open-mouth action, missing out on capturing the pouting expression.

Reconstruction Time. Figure 4 illustrates the FPS of mesh reconstruction for different resolutions on an NVIDIA A100 GPU, which is lower than 1 for most resolutions. Even with a resolution of 128, the FPS is <3, far below the required 30 FPS for real-time telepresence. Note that A100 [3] is one of the most powerful workstation GPUs currently available. When executed on an NVIDIA RTX 3080 GPU for laptops, it cannot handle the mesh reconstruction at resolutions of 512 and 1024, further exacerbating the reconstruction overhead.

Discussion. Our preliminary results demonstrate the inherent trade-offs between data size, computation overhead, and visual quality in semantic-based holographic communication. For keypoint-based semantics, despite its small data size, the high compression ratio introduces significant challenges for real-time reconstruction and maintaining the high visual quality of reconstructed content. Thus, we should judiciously consider the trade-offs when designing the full-fledged SemHolo.

5 Conclusion

In this position paper, we presented a holistic research agenda for semantic-driven, live, interactive holographic communication, a cornerstone of emerging immersive telepresence. To mitigate the huge bandwidth consumption stemming from the 3D nature of volumetric content, we proposed a pioneering approach that transmits semantic information in lieu of traditional bit-by-bit communication. Aiming to minimize both bandwidth consumption and end-to-end latency while maintaining a satisfactory level of visual quality, we delved into each semantic category to elucidate the open research challenges and propose potential solutions. Our preliminary results from a proof-of-concept implementation for keypoint-based semantics demonstrated that while it can significantly reduce bandwidth consumption, it presents considerable challenges in achieving high FPS and satisfactory visual quality. We hope that our study will inspire further research in this domain, propelling toward the realization of live, interactive holographic communication.

Acknowledgment

We thank the anonymous reviewers for their insightful comments. This work was supported in part by the U.S. NSF under Grants 2212296 and 2235049 and a Google Research Scholar Award.

References

[1] Draco 3D Data Compression. https://google.github.io/draco/. [accessed on 24-October-2023].

[2] Kinect for Windows. https://developer.microsoft.com/en-us/windows/kinect.

[3] NVIDIA A100 Tensor Core GPU. https://www.nvidia.com/en-us/data-center/a100/. [accessed on 24-October-2023].

[4] P. Achlioptas, O. Diamanti, I. Mitliagkas, and L. Guibas. Learning Representations and Generative Models for 3d Point Clouds. In *Proceedings of International Conference on Machine Learning (ICML)*, 2018.

[5] L. Ahrenberg, P. Benzie, M. Magnor, and J. Watson. Computer Generated Holograms from Three Dimensional Meshes Using an Analytic Light Transport Model. *Applied Optics*, 47(10):1567–1574, 2008.

[6] E. Arabadzhiyska, C. Tursun, H.-P. Seidel, and P. Didyk. Practical Saccade Prediction for Head-Mounted Displays: Towards a Comprehensive Model. *ACM Transactions on Applied Perceptions*, 20(1):1–23, 2023.

[7] E. Arabadzhiyska, O. T. Tursun, K. Myszkowski, H.-P. Seidel, and P. Didyk. Saccade Landing Position Prediction for Gaze-contingent Rendering. *ACM Transactions on Graphics*, 36(4):1–12, 2017.

[8] B. Attal, J.-B. Huang, C. Richardt, M. Zollhoefer, J. Kopf, M. O'Toole, and C. Kim. HyperReel: High-Fidelity 6-DoF Video with Ray-Conditioned Sampling. In *Proceedings of IEEE/CVF CVPR*, 2023.

[9] M. Baldi and Y. Ofek. End-to-end Delay Analysis of Videoconferencing Over Packet-switched Networks. *IEEE/ACM Transactions On Networking*, 8(4):479–492, 2000.

[10] R. Bashirov, A. Ianina, K. Iskakov, Y. Kononenko, V. Strizhkova, V. Lempitsky, and A. Vakhitov. Real-time Rgbd-based Extended Body Pose Estimation. In *Proceedings of IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2021.

[11] R. Bergmann, S. Rintel, N. Baym, A. Sarkar, D. Borowiec, P. Wong, and A. Sellen. Meeting (the) Pandemic: Videoconferencing Fatigue and Evolving Tensions of Sociality in Enterprise Video Meetings During COVID-19. *Computer Supported Cooperative Work*, pages 1–37, 2022.

[12] A. Burov, M. Nießner, and J. Thies. Dynamic Surface Function Networks for Clothed Human Bodies. In *Proceedings of IEEE/CVF ICCV*, 2021.

[13] Y. Cai, L. Ge, J. Liu, J. Cai, T.-J. Cham, J. Yuan, and N. M. Thalmann. ploiting Spatial-temporal Relationships for 3D Pose Estimation via Graph Convolutional Networks. In *Proceedings of IEEE/CVF CVPR*, 2019.

[14] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. In *Proceedings of IEEE/CVF CVPR*, 2017.

[15] K. Chen, T. Li, H.-S. Kim, D. E. Culler, and R. H. Katz. MARVEL: Enabling Mobile Augmented Reality with Low Energy and Low Latency. In *Proceedings of ACM SenSys*, 2018.

[16] R. H.-Y. Chen and T. D. Wilkinson. Computer Generated Hologram from Point Cloud Using Graphics Processor. *Applied Optics*, 48(6):6841–6850, 2009.

[17] S. Chen, H. Zhu, X. Chen, Y. Lei, G. Yu, and T. Chen. End-to-end 3D Dense Captioning with Vote2cap-detr. In *Proceedings of IEEE/CVF CVPR*, 2023.

[18] Z. Chen, A. Gholami, M. Nießner, and A. X. Chang. Scan2Cap: Context-aware Dense Captioning in RGB-D Scans. In *Proceedings of IEEE/CVF CVPR*, 2021.

[19] H. Choi, G. Moon, and K. M. Lee. Pose2Mesh: Graph Convolutional Network for 3D Human Pose and Mesh Recovery from a 2D Human Pose. In *Proceedings of Springer ECCV*, 2020.

[20] P. J. Choi, R. J. Oskouian, and R. S. Tubbs. Telesurgery: Past, Present, and Future. *Cureus*, 10(5):e2716, 2018.

[21] A. Clemm, M. T. Vega, H. K. Ravuri, T. Wauters, and F. D. Turck. Toward Truly Immersive Holographic-Type Communication: Challenges and Solutions. *IEEE Communications Magazine*, 58(1):93–99, 2020.

[22] A. Collet, M. Chuang, P. Sweeney, D. Gillett, D. Evseev, D. Calabrese, H. Hoppe, A. Kirk, and S. Sullivan. High-quality Streamable Free-viewpoint Video. *ACM Transactions on Graphics*, 34(4):1–13, 2015.

[23] C. De Alwis, A. Kalla, Q.-V. Pham, P. Kumar, K. Dev, W.-J. Hwang, and M. Liyanage. Survey on 6G Frontiers: Trends, Applications, Requirements, Technologies and Future Research. *IEEE Open Journal of the Communications Society*, 2:836–886, 2021.

[24] M. R. Diamond, J. Ross, and M. C. Morrone. Extraretinal Control of Saccadic Suppression. *Journal of Neuroscience*, 20(9):3449–3455, 2000.

[25] M. Dou, S. Khamis, Y. Degtyarev, P. Davidson, S. R. Fanello, A. Kowdle, S. O. Escolano, C. Rhemann, D. Kim, J. Taylor, et al. Fusion4D: Real-time Performance Capture of Challenging Scenes. *ACM Transactions on Graphics*, 35(4):1–13, 2016.

[26] R. A. Drebin, L. Carpenter, and P. Hanrahan. Volume Rendering. *ACM Siggraph Computer Graphics*, 22(4):65–74, 1988.

[27] R. Du, S. Bista, and A. Varshney. Video Fields: Fusing Multiple Surveillance Videos into a Dynamic Virtual Environment. In *Proceedings of International Conference on Web3D Technology*, 2016.

[28] R. Du, M. Chuang, W. Chang, H. Hoppe, and A. Varshney. Montage4D: Real-time Seamless Fusion and Stylization of Multiview Video Textures. *Journal of Computer Graphics Techniques*, 1(15):1–34, 2019.

[29] E. Dupont, H. Kim, S. A. Eslami, D. J. Rezende, and D. Rosenbaum. From Data to Functa: Your Data Point is a Function and You Can Treat it Like One. In *Proceedings of International Conference on Machine Learning (ICML)*, 2022.

[30] C. Geng, S. Peng, Z. Xu, H. Bao, and X. Zhou. Learning Neural Volumetric Representations of Dynamic Humans in Minutes. In *Proceedings of IEEE/CVF CVPR*, 2023.

[31] T. Golla and R. Klein. Real-time Point Cloud Compression. In *Proceedings of International Conference on Intelligent Robots and Systems*, 2015.

[32] Y. Guan, X. Hou, N. Wu, B. Han, and T. Han. MetaStream: Live Volumetric Content Capture, Creation, Delivery, and Rendering in Real Time. In *Proceedings of ACM MobiCom*, 2023.

[33] B. Guenter, M. Finch, S. Drucker, D. Tan, and J. Snyder. Foveated 3D graphics. *ACM Transactions on Graphics*, 31(6):1–10, 2012.

[34] S. Gül, D. Podborski, T. Buchholz, T. Schierl, and C. Hellge. Low-latency Cloud-based Volumetric Video Streaming Using Head Motion Prediction. In *Proceedings of ACM Workshop on Network and Operating Systems Support for Digital Audio and Video*, 2020.

[35] S. Gül, D. Podborski, J. Son, G. S. Bhullar, T. Buchholz, T. Schierl, and C. Hellge. Cloud Rendering-based Volumetric Video Streaming System for Mixed Reality Services. In *Proceedings of ACM MMSys*, 2020.

[36] D. Gündüz, Z. Qin, I. E. Aguerri, H. S. Dhillon, Z. Yang, A. Yener, K. K. Wong, and C.-B. Chae. Beyond Transmitting Bits: Context, Semantics, and Task-Oriented Communications. *IEEE Journal on Selected Areas in Communications*, 41(1):5–41, 2022.

[37] Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu, and M. S. Lew. Deep Learning for Visual Understanding: A Review. *Neurocomputing*, 187:27–48, 2016.

[38] A. Gupta, A. Bansal, and V. Khanduja. Modern Lossless Compression Techniques: Review, Comparison and Analysis. In *Proceedings of IEEE International Conference on Electrical, Computer and Communication Technologies*, 2017.

[39] B. Han, Y. Liu, and F. Qian. ViVo: Visibility-Aware Mobile Volumetric Video Streaming. In *Proceedings of ACM MobiCom*, 2020.

[40] Z. Hu, S. Li, C. Zhang, K. Yi, G. Wang, and D. Manocha. DGaze: CNN-Based Gaze Prediction in Dynamic Scenes. *IEEE Transactions on Visualization and Computer Graphics*, 26(5):1902–1911, 2020.

[41] E. Hubo, T. Mertens, T. Haber, and P. Bekaert. The Quantized kd-Tree: Efficient Ray Tracing of Compressed Point Clouds. In *Proceedings of IEEE Symposium on Interactive Ray Tracing*, 2006.

[42] H. Iqbal, A. Khalid, and M. Shahzad. Dissecting Cloud Gaming Performance with DECAF. In *Proceedings of ACM on Measurement and Analysis of Computing Systems (SIGMETRICS)*, 2021.

[43] J. Jiang, V. Sekar, and H. Zhang. Improving Fairness, Efficiency, and Stability in HTTP-based Adaptive Video Streaming with FESTIVE. In *Proceedings of ACM CoNEXT*, 2012.

[44] H. Joo, T. Simon, and Y. Sheikh. Total Capture: A 3D Deformation Model for Tracking Faces, Hands, and Bodies. In *Proceedings of IEEE/CVF CVPR*, 2018.

[45] H. Jun and A. Nichol. Shap-E: Generating Conditional 3D Implicit Functions. https://arxiv.org/abs/2305.02463, 2023. [accessed on 24-October-2023].

[46] N. Kolotouros, G. Pavlakos, and K. Daniilidis. Convolutional Mesh Regression for Single-image Human Shape Reconstruction. In *Proceedings of IEEE/CVF CVPR*, 2019.

[47] K. Lee, J. Yi, and Y. Lee. FarfetchFusion: Towards Fully Mobile Live 3D Telepresence Platform. In *Proceedings of ACM MobiCom*, 2023.

[48] K. Lee, J. Yi, Y. Lee, S. Choi, and Y. M. Kim. GROOT: A Real-Time Streaming System of High-Fidelity Volumetric Videos. In *Proceedings of ACM MobiCom*, 2020.

[49] M. Lee, W. Park, S. Lee, and S. Lee. Distracting Moments in Videoconferencing: A Look Back at the Pandemic Period. In *Proceedings of ACM Conference on Human Factors in Computing Systems (CHI)*, 2022.

[50] C. Li, G. Wang, B. Wang, X. Liang, Z. Li, and X. Chang. Dynamic Slimmable Network. In *In Proceedings of IEEE/CVF CVPR*, 2021.

[51] C. Li, C. Zhang, A. Waghwase, L.-H. Lee, F. Rameau, Y. Yang, S.-H. Bae, and C. S. Hong. Generative AI meets 3D: A Survey on Text-to-3D in AIGC Era. https://arxiv.org/abs/2305.06131, 2023. [accessed on 24-October-2023].

[52] T. Li and X. Zhou. Battery-Free Eye Tracker on Glasses. In *Proceedings of ACM MobiCom*, 2018.

[53] Z. Li, E. Wallace, S. Shen, K. Lin, K. Keutzer, D. Klein, and J. Gonzalez. Train Large, Then Compress: Rethinking Model Size for Efficient Training and Inference of Transformers. In *Proceedings of International Conference on Machine Learning (ICML)*, 2020.

[54] J.-M. Lien, G. Kurillo, and R. Bajcsy. Multi-camera Tele-immersion System with Real-time Model Driven Data Compression. *The Visual Computer*, 26(3):3–15, 2010.

[55] Y. Lin, Z. Gao, H. Du, D. Niyato, J. Kang, A. Jamalipour, and X. S. Shen. A Unified Framework for Integrating Semantic Communication and AI-Generated Content in Metaverse. https://arxiv.org/abs/2305.11911, 2023. [accessed on 24-October-2023].

[56] K. Liu, R. Cheng, N. Wu, and B. Han. Toward Next-generation Volumetric Video Streaming with Neural-based Content Representations. In *Proceedings of ACM Workshop on Mobile Immersive Computing, Networking, and Systems (ImmerCom 2023)*, 2023.

[57] Y. Liu, B. Han, F. Qian, A. Narayanan, and Z.-L. Zhang. Vues: Practical Volumetric Video Streaming through Multiview Transcoding. In *Proceedings of ACM MobiCom*, 2022.

[58] X. Luo, H.-H. Chen, and Q. Guo. Semantic Communications: Overview, Open Issues, and Future Research Directions. *IEEE Wireless Communications*, 29(1):210–219, 2022.

[59] K. MacMillan, T. Mangla, J. Saxon, and N. Feamster. Measuring the Performance and Network Utilization of Popular Video Conferencing Applications. In *Proceedings of ACM IMC*, 2021.

[60] A. Maglo, G. Lavoué, F. Dupont, and C. Hudelot. 3D Mesh Compression: Survey, Comparisons, and Emerging Trends. *ACM Computing Surveys*, 47(3), 2015.

[61] H. Mao, R. Netravali, and M. Alizadeh. Neural Adaptive Video Streaming with Pensieve. In *Proceedings of ACM SIGCOMM*, 2017.

[62] R. Martin-Brualla, N. Radwan, M. S. Sajjadi, J. T. Barron, A. Dosovitskiy, and D. Duckworth. NeRF in the Wild: Neural Radiance Fields for Unconstrained Photo Collections. In *Proceedings of IEEE/CVF CVPR*, 2021.

[63] R. McAfee, C. Haxton, M. Harrison, and J. Gess. Thermal Characterization of a Virtual Reality Headset during Transient and Resting Operation. In *Proceedings of Semiconductor Thermal Measurement, Modeling & Management Symposium (SEMI-THERM)*, 2020.

[64] G. Metzer, E. Richardson, O. Patashnik, R. Giryes, and D. Cohen-Or. Latent-NeRF for Shape-guided Generation of 3D Shapes and Textures. In *Proceedings of IEEE/CVF CVPR*, 2023.

[65] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

[66] G. Moon, H. Choi, and K. M. Lee. Accurate 3D Hand Pose Estimation for Whole-Body 3D Human Mesh Estimation. In *Proceedings of IEEE/CVF CVPR*, 2022.

[67] G. Moon and K. M. Lee. I2L-MeshNet: Image-to-Lixel Prediction Network for Accurate 3D Human Pose and Mesh Estimation from a Single RGB Image. In *Proceedings of Springer ECCV*, 2020.

[68] A. Morales, F. M. Costela, and R. L. Woods. Saccade Landing Point Prediction Based on Fine-Grained Learning Method. *IEEE Access*, 9:52474–52484, 2021.

[69] T. Neate, V. Kladouchou, S. Wilson, and S. Shams. "Just Not Together": The Experience of Videoconferencing for People with Aphasia during the Covid-19 Pandemic. In *Proceedings of ACM Conference on Human Factors in Computing Systems (CHI)*, 2022.

[70] A. Nichol, H. Jun, P. Dhariwal, P. Mishkin, and M. Chen. Point-E: A System for Generating 3D Point Clouds from Complex Prompts. https://arxiv.org/abs/2212.08751, 2022. [accessed on 24-October-2023].

[71] A. Q. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. Mcgrew, I. Sutskever, and M. Chen. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In *Proceddings of International Conference on Machine Learning (ICML)*, 2022.

[72] J. Nystad, A. Lassen, A. Pomianowski, S. Ellis, and T. Olson. Adaptive Scalable Texture Compression. In *Proceedings of ACM SIGGRAPH/Eurographics Conference on High-Performance Graphics*, 2012.

[73] S. Orts-Escolano, C. Rhemann, S. Fanello, W. Chang, A. Kowdle, Y. Degtyarev, D. Kim, P. Davidson, S. Khamis, M. Dou, V. Tankovich, C. Loop, Q. Cai, P. Chou, S. Mennicken, J. Valentin, V. Pradeep, S. Wang, S. B. Kang, P. Kohli, Y. Lutchyn, C. Keskin, and S. Izadi. Holoportation: Virtual 3D Teleportation in Real-time. In *Proceedings of ACM Symposium on User Interface Software and Technology (UIST)*, 2016.

[74] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. Osman, D. Tzionas, and M. J. Black. Expressive Body Capture: 3D Hands, Face, and Body from a Single Image. In *Proceedings of IEEE/CVF CVPR*, 2019.

[75] D. Pavllo, C. Feichtenhofer, D. Grangier, and M. Auli. 3D Human Pose Estimation in Video with Temporal Convolutions and Semi-supervised

Training. In *Proceedings of IEEE/CVF CVPR*, 2019.

[76] S. Peng, Y. Yan, Q. Shuai, H. Bao, and X. Zhou. Representing Volumetric Videos as Dynamic MLP Maps. In *Proceedings of IEEE/CVF CVPR*, 2023.

[77] S. Peng, Y. Zhang, Y. Xu, Q. Wang, Q. Shuai, H. Bao, and X. Zhou. Neural Body: Implicit Neural Representations with Structured Latent Codes for Novel View Synthesis of Dynamic Humans. In *Proceedings of IEEE/CVF CVPR*, 2021.

[78] B. Poole, A. Jain, J. T. Barron, and B. Mildenhall. Dreamfusion: Text-to-3d Using 2d Diffusion. https://arxiv.org/abs/2304.12932, 2022. [accessed on 24-October-2023].

[79] C. R. Qi, L. Yi, H. Su, and L. J. Guibas. PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space. In *Proceedings of Conference on Neural Information Processing Systems*, 2017.

[80] F. Qian, B. Han, J. Pair, and V. Gopalakrishnan. Toward Practical Volumetric Video Streaming On Commodity Smartphones. In *Proceedings of ACM HotMobile*, 2019.

[81] S. Raghuraman, K. Venkatraman, Z. Wang, B. Prabhakaran, and X. Guo. A 3D Tele-immersion Streaming Approach Using Skeleton-based prediction. In *Proceedings of ACM International Conference on Multimedia*, 2013.

[82] A. A. Rusu, N. C. Rabinowitz, G. Desjardins, H. Soyer, J. Kirkpatrick, K. Kavukcuoglu, R. Pascanu, and R. Hadsell. Progressive Neural Networks. https://arxiv.org/abs/1606.04671, 2016. [accessed on 24-October-2023].

[83] K. Shen, C. Guo, M. Kaufmann, J. J. Zarate, J. Valentin, J. Song, and O. Hilliges. X-avatar: Expressive Human Avatars. In *Proceedings of IEEE/CVF CVPR*, 2023.

[84] G. Shi, Y. Xiao, Y. Li, and X. Xie. From Semantic Communication to Semantic-Aware Networking: Model, Architecture, and Open Problems. *IEEE Communications Magazine*, 59(8):44–50, 2021.

[85] L. Sigal. Human Pose Estimation. In *Computer Vision: A Reference Guide*, pages 573–592. 2021.

[86] L. Song, A. Chen, Z. Li, Z. Chen, L. Chen, J. Yuan, Y. Xu, and A. Geiger. NeRFPlayer: A Streamable Dynamic Scene Representation with Decomposed Neural Radiance Fields. *IEEE Transactions on Visualization and Computer Graphics*, 29(5):2732–2742, 2023.

[87] E. C. Strinati, S. Barbarossa, J. L. Gonzalez-Jimenez, D. Ktenas, N. Cassiau, L. Maret, and C. Dehos. 6G: The Next Frontier: From Holographic Messaging to Artificial Intelligence Using Subterahertz and Visible Light Communication. *IEEE Vehicular Technology Magazine*, 14(3):42–50, 2019.

[88] K. Sun, B. Xiao, D. Liu, and J. Wang. Deep High-resolution Representation Learning for Human Pose Estimation. In *Proceedings of IEEE/CVF CVPR*, 2019.

[89] F. Tariq, M. R. A. Khandaker, K.-K. Wong, M. A. Imran, M. Bennis, and M. Debbah. A Speculative Study on 6G. *IEEE Wireless Communications*, 27(4):118–125, 2020.

[90] B. Thoravi Kumaravel, F. Anderson, G. Fitzmaurice, B. Hartmann, and T. Grossman. Loki: Facilitating Remote Instruction of Physical Tasks Using Bi-Directional Mixed-Reality Telepresence . In *Proceedings of ACM Symposium on User Interface Software and Technology (UIST)*, 2019.

[91] A. Toshev and C. Szegedy. Deeppose: Human Pose Estimation via Deep Neural Networks. In *Proceedings of IEEE/CVF CVPR*, 2014.

[92] M. Wang and W. Deng. Deep Face Recognition: A Survey. *Neurocomputing*, 429:215–244, 2021.

[93] C.-Y. Weng, B. Curless, P. P. Srinivasan, J. T. Barron, and I. Kemelmacher-Shlizerman. HumanNeRF: Free-viewpoint Rendering of Moving People from Monocular Video. In *Proceedings of IEEE/CVF CVPR*, 2022.

[94] H. Xie, Z. Qin, G. Y. Li, and B.-H. Juang. Deep Learning Enabled Semantic Communication Systems. *IEEE Transactions on Signal Processing*, 69:2663–2675, 2021.

[95] H. Xie, Z. Qin, X. Tao, and K. B. Letaief. Task-Oriented Multi-user Semantic Communications. *IEEE Journal on Selected Areas in Communications*, 40(9):2584–2597, 2022.

[96] H. Xue, Y. Ju, C. Miao, Y. Wang, S. Wang, A. Zhang, and L. Su. mmMesh: Towards 3D Real-Time Dynamic Human Mesh Construction Using Millimeter-Wave. In *Proceedings of ACM MobiSys*, 2021.

[97] G.-W. Yang, W.-Y. Zhou, H.-Y. Peng, D. Liang, T.-J. Mu, and S.-M. Hu. Recursive-NeRF: An Efficient and Dynamically Growing NeRF. *IEEE Transactions on Visualization and Computer Graphics*, 2022.

[98] A. Yu, V. Ye, M. Tancik, and A. Kanazawa. pixelNeRF: Neural Radiance Fields from One or Few Images. In *Proceedings of IEEE/CVF CVPR*, 2021.

[99] T. Yu, Z. Zheng, K. Guo, P. Liu, Q. Dai, and Y. Liu. Function4D: Real-time Human Volumetric Capture from Very Sparse Consumer RGBD Sensors. In *Proceedings of IEEE/CVF CVPR*, 2021.

[100] Z. Yu, J. Wang, J. Xu, B. Ni, C. Zhao, M. Wang, and W. Zhang. Skeleton2Mesh: Kinematics Prior Injected Unsupervised Human Mesh Recovery. In *Proceedings of IEEE/CVF CVPR*, 2021.

[101] Z. Yuan, X. Yan, Y. Liao, Y. Guo, G. Li, S. Cui, and Z. Li. X-trans2cap: Cross-modal Knowledge Transfer Using Transformer for 3D Dense Captioning. In *Proceedings of IEEE/CVF CVPR*, 2022.

[102] A. Zhang, C. Wang, B. Han, and F. Qian. Efficient Volumetric Video Streaming Through Super Resolution. In *Proceedings of ACM HotMobile*, 2021.

[103] A. Zhang, C. Wang, B. Han, and F. Qian. YuZu: Neural-enhanced Volumetric Video Streaming. In *Proceedings of USENIX NSDI*, 2022.

[104] B. Zhang, Z. Qin, Y. Guo, and G. Y. Li. Semantic Sensing and Communications for Ultimate Extended Reality. https://arxiv.org/abs/2212.08533, 2022. [accessed on 24-October-2023].

[105] D. Zhang, B. Han, P. Pathak, and H. Wang. Innovating Multi-user Volumetric Video Streaming through Cross-layer Design. In *Proceedings of ACM HotNets*, 2021.

[106] D. Zhang, P. Zhou, B. Han, and P. Pathak. M5: Facilitating Multi-User Volumetric Content Delivery with Multi-Lobe Multicast over mmWave. In *Proceedings of ACM SenSys*, 2022.

[107] Y. Zhu, Y. Huang, X. Qiao, Z. Tan, B. Bai, H. Ma, and S. Dustdar. A Semantic-aware Transmission with Adaptive Control Scheme for Volumetric Video Service. *IEEE Transactions on Multimedia*, 2022.