Demo Abstract: Semantic Communications for Immersive Multi-view Media Delivery

Jingxuan Men[1], Carl Udora[2], Ning Wang[3], Mahdi Boloursaz Mashhadi[4], Yi Ma[5], and Mike Nilsson[6]

[1,2,3,4,5]University of Surrey , E-mail: {j.men, cu00029, n.wang, m.boloursazmashhadi, y.ma}@surrey.ac.uk
[6]British Telecom , E-mail: {mike.nilsson}@bt.com

***Abstract*—Our demonstration highlights the use of semantic communication to transmit stereo frame streams and its application in immersive media. Considering the limitations of traditional stereo compression with digital transmission under poor channel conditions, such as the cliff effect and high latency in computation and transmission, we employ the Deep Joint Source and Channel Coding (Deep JSCC) framework to transmit semantic information of stereo streams between the sender and receiver. To address channel instability, we propose a dynamic rate adjustment method that adapts to channel conditions while maintaining transmission efficiency and reconstruction quality. Furthermore, we extend this work to stereo stream applications, enabling the real-time synthesis of multiple novel view streams of the streamer. The overall design of this demo enables an immersive multi-view experience by transmitting rate-controlled semantic features from only two perspectives.**

I. Introduction

The increasing demand for high-quality stereo camera systems in applications like immersive streaming and autonomous driving has significantly heightened interest in the compression and transmission of stereo image streams. Stereo camera arrays capture vast amounts of data, and simplistic image compression methods fail to account for overlapping information between cameras. This highlights the need for advanced image processing algorithms that can extract deeper data features to reduce transmission rates effectively.

Semantic communication has recently garnered significant attention, leveraging deep learning models to extract semantic features from data sources and generate similar content at the receiver end for more efficient transmission. Some studies have proposed using Deep Joint Source-Channel Coding (DeepJSCC) for semantic communication. Unlike traditional communication methods that separate source and channel coding, JSCC integrates these processes, focusing on transmitting task-relevant semantic information. Through end-to-end optimization, JSCC effectively mitigates the "cliff effect" caused by fixed source and channel coding rates, making it an ideal approach for semantic communication [1].

In recent years, immersive live streaming has garnered increasing attention and adoption. However, it typically requires substantial transmission resources. Currently, the limitations of immersive streaming using stereo transmission methods are as follows:

1) Existing stereo codecs that rely on digital transmission are complex and dependent on entropy encoding, resulting in high latency. Additionally, no methods have been proposed for stereo transmission using semantic communication.
2) Unstable channel conditions lead to transmission delays and a decline in image reconstruction quality. For stereo transmission, no adaptive rate control methods based on channel conditions have been proposed.

To address these challenges and enhance the user experience of immersive live streaming, we propose a semantic communication method for stereo media along with an adaptive rate control architecture. This method ensures stable transmission quality and minimal bandwidth consumption, while integrating a multi-view rendering mechanism that allows users to experience high-resolution live streaming from multiple angles in specific stereo media applications.

II. Proposed Method

Our proposed architecture, illustrated in Fig.1, employs multiple cameras placed around the streamer to capture real-time imagery. The captured video streams are sent to an edge server, which provides computational resources and performs semantic encoding of the stereo frames. To further mitigate the cliff effect associated with fixed source-channel coding, we incorporate a Deep JSCC mechanism. This approach enhances transmission robustness and quality under varying channel conditions. When a user accesses the stream, their edge server provides computational resources to perform semantic decoding and real-time rendering. For stereo decoding, we adopt three Stereo Attention Modules (SAM) [2], which leverage the correlation between transmitted features from the two views for high-quality reconstruction. After that, the reconstructed information are rendered using Gaussian splatting [3], which is state-of-the-art rendering method, enabling users to dynamically switch to a new viewing angle in real time. Our demo can support dynamic multi-view (1-10) live streaming at the same time, providing an immersive experience with minimal device complexity for both senders and receivers.

In the semantic encoding process, we utilize an adaptive semantic masking method to dynamically extract important features and adjust the mask based on the current channel conditions. This controls the amount of transmitted data to match the available bandwidth. Specifically, after CNN-based feature extraction for the left and right views, correlation scores between the features are calculated to determine their relative importance. These scores, combined with the current SNR, are input into the decision module to generate the final

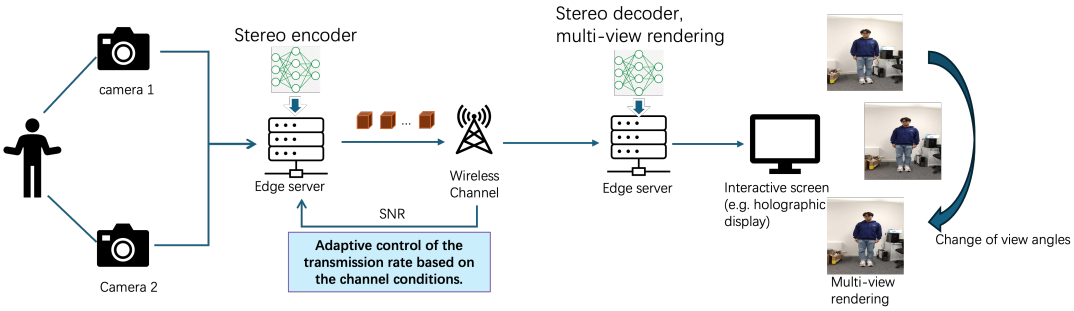

Fig. 1: The architecture of semantic communication for immersive multi-view streaming transmission.

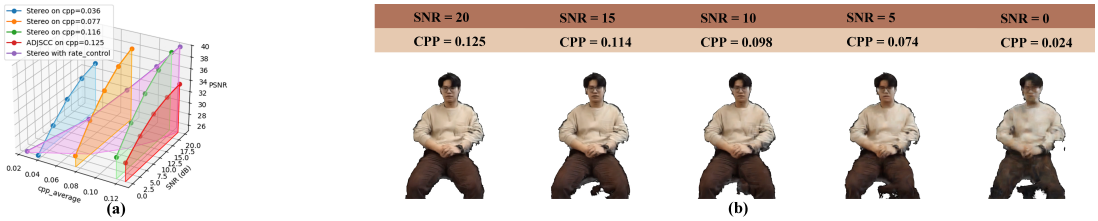

Fig. 2: PSNR value of reconstruction after transmission (a) and visual rendering quality (b) in different channel conditions.

masking matrix. Before transmission, this matrix is applied to the CNN-encoded features of both the left and right views, ensuring that the most critical features have a lower probability of being masked. This process enhances the integration of inter-view correlations, thereby improving the quality of the final reconstruction.

III. Demonstration

In this demo abstract, we have introduced the application of semantic communication and rate control method for transmitting stereo media streams. Building upon this, we integrated Gaussian splatting as a rendering method to form a comprehensive demo. This demo showcases the effectiveness of semantic transmission, the impact of rate control, and the practical application of stereo video stream rendering.

The detailed process is as follows: the streamer places two cameras on tripods in front of them, with an angle of 10–30 degrees between the devices. The streamer stands approximately 2 meters away from each camera and activates the cameras. The captured video streams are uploaded to an edge server, where joint source-channel encoding is streamed. These encoded streams are then transmitted over a simulated AWGN channel to the user's edge server, where joint source-channel decoding is executed. Stereo stream rendering is subsequently achieved using Gaussian splatting. To simplify the demo, a single server is utilized to simulate the encoding, transmission, decoding, and rendering processes.

To demonstrate the performance in our demo, channel quality is represented by the Signal-to-Noise Ratio (SNR), while the transmission rate is measured using wireless channel utilization per pixel (CPP). CPP is defined as the ratio of the total number of semantic symbols transmitted to the total number of pixels in the original image.

In Figure 2(a), we evaluate the PSNR values of the single-view semantic communication method ADJSCC [4] and our proposed fixed-rate and rate-controlled stereo semantic communication methods after transmitting the captured data. Compared to the single-view method, the stereo method improves reconstruction quality while reducing the transmission CPP. Furthermore, for the rate control method, although the CPP varies with SNR, the reconstructed PSNR value remains largely unaffected compared to the fixed-rate methods. Additionally, figure 2(b) displays the live rendered human in the intermediate view between two cameras, along with the corresponding channel conditions and transmission rates. It is evident that as the channel quality deteriorates, the adaptive rate control method effectively reduces the data transmission rate without causing significant degradation in visual quality. This is due to our framework's ability to account for the correlation between viewpoints, enabling the extraction and transmission of key semantic information. Moreover, the stereo reconstruction mechanism helps improve reconstruction quality, ensuring visually consistent output even under challenging channel conditions.

Acknowledgment

This work is supported in part by the UK Department for Science, Innovation and Technology under the Future Open Networks Research Challenge project TUDOR (Towards Ubiquitous 3D Open Resilient Network) and Horizon Europe SPIRIT Project (101070672). The views expressed are those of the authors and do not necessarily represent the project.

References

[1] Y. Shi *et al.*, "Task-oriented communications for 6G: Vision, principles, and technologies," *IEEE Wirel. Commun.*, 2023.

[2] X. Ying *et al.*, "A Stereo Attention Module for Stereo Image Super-Resolution," *IEEE Signal Process. Lett.*, 2020.

[3] S. Zheng *et al.*, "GPS-Gaussian: Generalizable pixel-wise 3D gaussian splatting for real-time human novel view synthesis," *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024.

[4] J. Xu *et al.*, "Wireless image transmission using deep source channel coding with attention modules," *IEEE Trans. Circuits Syst. Video Technol.*, 2021.