IBC2024

NERF BASED 3D GENERATIVE VIDEO CONFERENCING SYSTEM

Jianglong Li, Jun Xu, Yuelin Hu, Zhiyu Zhang, Li Song

Institute of Image Communication and Network Engineering,
Shanghai Jiao Tong University, China

ABSTRACT

Video conferencing, is the most demanding form of video communication in terms of real-time requirements and it remains a challenge to maintain quality in weak network conditions. Traditional block-based encoding video conferencing systems can experience freezing and significant degradation when bandwidth is extremely low or network conditions deteriorate suddenly. Recent advancements in 3D facial representation offer novel promising solutions for video conferencing under weak network conditions. This paper introduces a generative 3D video conferencing system using pre-trained Neural Radiance Field (NeRF) models for high-fidelity 3D head reconstruction and real-time rendering. Clients extract and encode facial parameters for transmission, while simultaneously receiving and decoding parameters from peers to generate visuals. Our system maintains good video quality at bit-rates under 5kbps, with objective and subjective quality comparable to HEVC encoders at 18kbps and 50kbps, respectively. By integrating real-time face tracking of facial parameters, Real-Time Communication (RTC), and real-time volumetric video rendering, our system enhances the potential for 3D video conferencing collaboration. A live demonstration showcases the significant innovation of the system, promising to forge a new paradigm for video conferencing in the context of future spatial computing.

INTRODUCTION

According to the latest report from Cisco, video traffic accounted for 82% of internet traffic in 2022 [32]. As a crucial form of video communication, video conferencing demands high real-time performance, necessitating a lightweight overall system and strong network adaptability. Meeting these requirements continues to pose a challenge to both the academic and industrial communities.

Currently, mainstream video conferencing systems such as Zoom, Microsoft Teams and Tencent Meeting typically rely on traditional video codec frameworks. Traditional video codecs like High Efficiency Video Coding (HEVC/H.265) [1], Versatile Video Coding (VVC/H.266) [2], and AV1 [3] have several advantages: (1) they aim for pixel-level fidelity, staying true to the original images; (2) they are universal and stable, providing good encoding performance across a wide range of scenarios; (3) they have low hardware performance requirements, facilitating large-scale deployment. However, traditional video encoding also has significant drawbacks, particularly its inability to handle extremely low-bandwidth conditions. This often results in freezing and a dramatic decrease in quality, leading to a poor user experience. These challenges have prompted the exploration of new encoding solutions suitable for video conferencing systems.

In the academic community, numerous studies have examined the characteristics of video within video conferencing, leading to the development of various solutions. Specifically, it has been noted that in video conferencing scenarios, the background image behind a participant is typically static, focusing attendees' attention primarily on the facial region. This implies that only the facial video needs to be encoded and transmitted. Facial images usually have similar structures and semantic meanings (such as eyes, mouths, etc.), which suggests that we can learn *a priori* on a facial dataset and then use minimal semantic information to reconstruct facial images. Recent deep learning methods have demonstrated potential in generating facial imagery from limited information, making them promising for facial semantic communication. Feng *et al* [4] proposed a generative video compression framework based on FSGAN [9], achieving a low bit-rate of around 1 kB/s, but it struggles with significant facial movements. FOMM [10] uses keypoints and Jacobians to represent sparse motion, which is then used to animate a talking face. Building on FOMM, Konuko *et al* [8] utilize one raw frame as a reference frame and add generated frames to the reference frame pool, which may lead to error accumulation. Xu *et al* [11], building on [6], proposed a hybrid encoding framework based on facial keypoints. However, it needs keyframes on both sides from which to reconstruct intermediate frames. This introduces additional latency, which is less acceptable in real-time communication (RTC) scenarios. All these methods can only reconstruct 2D faces.

The recent surge in Artificial Intelligence Generated Content (AIGC) has fostered innovative 3D representation methods, such as Neural Radiance Fields (NeRF) [12] and 3D Gaussian Splatting (3D GS) [13], which promise to shift video communication towards generative approaches. These generative video communication technologies not only aim to address the issues associated with low-bandwidth networks as mentioned earlier, but also offer users novel experiences. NeRF is a technique for creating realistic 3D models from 2D images, utilizing neural networks to predict the color and density of 3D points in space, based on their coordinates and camera viewpoints, and achieving the rendering through volumetric rendering. 3D GS uses 3D Gaussians to model three-dimensional scenes explicitly and optimizes parameters using the capabilities of neural networks. There have already been applications using NeRF or 3D GS to represent parameterized 3D human heads, capable of real-time rendering [14][19]. Building on these advances within the 3D community, this paper proposes and implements an ultra-low bit-rate generative 3D video conferencing system. Our approach requires the receiver to have a personalized NeRF model [14] of the other participants. Compared to conventional system such as WebRTC, only facial parameters are extracted and transmitted to the other end, instead of image data. The receiver end then reconstructs the 3D head based on the decoded expression parameters and the personalized NeRF model of the sender’s head. In particular, the receiver can use custom pose parameters for rendering from any chosen viewpoint. This system not only addresses issues related to weak network environments but also provides an enhanced 3D viewing experience.

The main contributions of this paper are:

1) We innovatively propose a 3D video conferencing system, which is based on open-source components, with all modules being real-time and practical. The overall end-to-end latency of the system is below 90ms. To the best of our knowledge, this is the first practical real-time system that integrates a 3D representation model.

2) The parameter encoding module and pose control module allow the system to achieve ultra-low bit-rates while supporting free-viewpoint watching.

3) Experiments demonstrate that our approach outperforms traditional encoding methods. At bit-rates under 5kbps, our video quality achieves levels comparable to those of the x265 encoder at 18kbps and 50kbps for objective and subjective metrics, respectively.

RELATED WORK

Video Conference System

As industries accelerate their digital transformation, the demand for Real-Time Communication (RTC) applications has surged dramatically. Video conferencing has become a cornerstone of professional collaboration, remote education, and personal connections. This evolution has driven widespread adoption of video conferencing across various platforms, including mobile devices, personal computers, and dedicated conference systems. With the expansion of application scopes and the increasing demand for advanced features, many systems have transitioned from traditional on-premises setups to more scalable cloud-based solutions to enhance computational efficiency and manageability. As spatial computing has advanced, facial representation has also evolved from traditional 2D generation to 3D reconstruction. For example, Google Project Starline [34] and Apple Vision Pro [35] have introduced new possibilities for immersive experience video conferencing. Video conferencing systems have stringent requirements for latency and bandwidth efficiency to maintain optimal Quality of Experience (QoE) across diverse network conditions. This necessitates the support of efficient transport protocols, among which the QUIC protocol is considered more promising than traditional UDP-based transmission protocols due to its enhanced efficiency and reliability.

Facial Presentation

In traditional video conferencing systems, a block-based hybrid encoding architecture, such as HEVC and AV1, is used. This approach is suboptimal when encoding video specifically for video conferencing scenarios. With the development of deep learning technologies, some approaches have adopted generative methods for 2D facial generation. These methods, such as those outlined in [5] [6] [11], primarily rely on the extraction of 2D keypoints. At the sending end, facial images are divided into keyframes and non-keyframes, with keypoints extracted from non-keyframes. At the receiving end, deep generative models use these keypoints and keyframes to reconstruct the non-keyframes. Due to their dependency between frames, these methods lack robustness in handling large movements.

As spatial computing continues to advance, research into 3D facial representation has expanded significantly, exploring a variety of innovative approaches. These methods enable the utilization of minimal semantic parameters to drive models, facilitating the accurate reconstruction of 3D faces. This technological progress has made 3D video conferencing a practical reality. Some studies focus on 3D keypoints; for example, in [7] the authors extract 3D keypoints from facial images to simulate facial movements. Although this technique allows for rendering from arbitrary viewpoints at the receiver end, it lacks robustness for large movements. Other research has adopted the parameterized

NeRF model for heads. For example, NeRFace [16] inputs facial parameters to achieve dynamic 3D modeling, but the training of this model is slow and fidelity is relatively low. Research presented in [14] introduces a method for reconstructing a facial semantic NeRF model from monocular videos, facilitating rapid training and high-fidelity reconstruction. With a pre-trained NeRF model, this approach can generate 3D faces using a set of expression parameters and allows for specifying the angle of 2D rendering using pose parameters. Additionally, some research has utilized explicit expressions of 3D GS to represent 3D human heads. In [18], the explicit representation of 3D GS is blended with a set of learnable latent features, enabling the driving of a parametrized head model with low-dimensional linear parameters. [19] employs a 3D Gaussian field to represent a parametrized facial model, capturing facial details using geometric priors, and achieves high-fidelity rendering at 300 fps on consumer-grade GPUs.

Transport Protocol

The choice of transport protocol is crucial to video conferencing systems. The most commonly used transport layer protocols are Transmission Control Protocol (TCP) and User Datagram Protocol (UDP).

While TCP provides ordered and reliable data delivery, its susceptibility to high latency due to head-of-line blocking and inefficient retransmission tactics renders it less ideal for RTC applications. Consequently, most advanced video conferencing systems leverage UDP-based protocols that prioritize timeliness over reliability. These systems commonly utilize the Real-Time Transport Protocol (RTP) [20] and its secure variant, the Secure Real-time Transport Protocol (SRTP) [21], with industry leaders like Zoom enhancing these protocols with custom extensions [22].

The Quick UDP Internet Connections (QUIC) [23] protocol has recently attracted considerable attention for its impressive performance and flexibility. QUIC dramatically reduces the time required to establish a reliable and secure connection to just one Round-Trip Time (RTT), a significant improvement over TCP’s cumbersome three-way handshake. It also facilitates stream-multiplexing and connection migration, which enhance performance under fluctuating network conditions and simplify congestion control upgrades through its modular approach.

This growing interest in QUIC indicates its potential to revolutionize real-time video streaming. Ongoing research is focused on integrating QUIC with traditional video streaming protocols, such as adapting RTP to operate over QUIC [24] and extending it to support unreliable transmissions [25]. These developments position QUIC as a transformative element in the video conferencing domain.

SYSTEM ARCHITECTURE

Overview

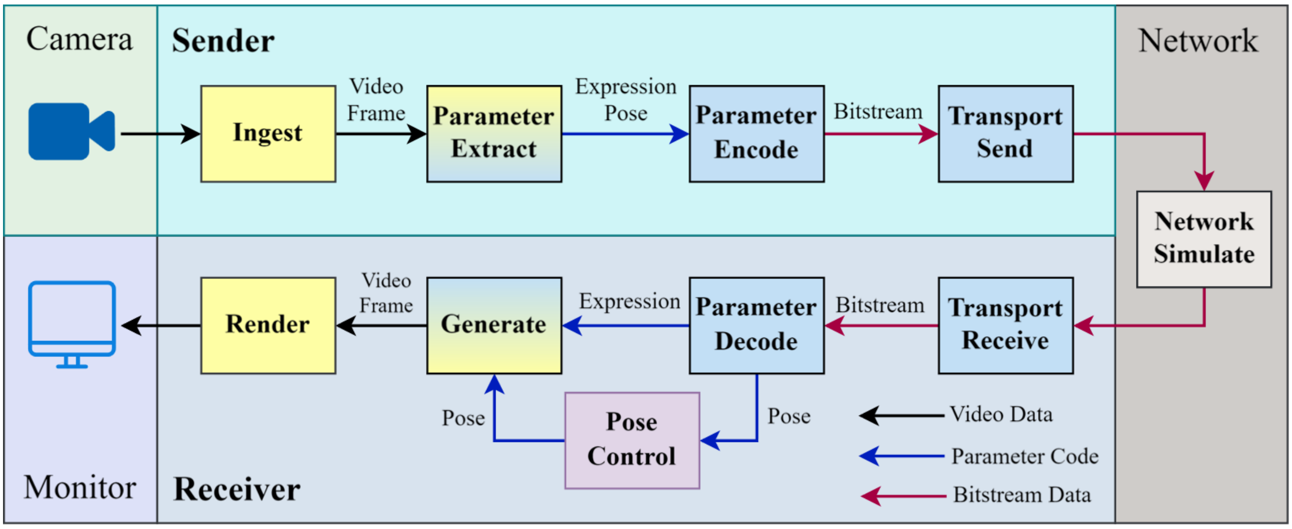

Figure 1– The proposed framework for the 3D generative video conferencing system. The data format in the yellow modules is images, while in the blue modules, the data format consists of expression and pose parameters. Black arrows represent video data, blue arrows indicate parameter codes, and red arrows represent bitstream data.

The proposed 3D generative video conferencing architecture is depicted in Figure 1, consisting of a sender, network simulation, and receiver. The workflow of our video conferencing system is as follows: At the sender side, video is first captured from the camera. Then facial expression parameters and pose parameters are extracted from each frame. These parameters are encoded into bitstreams for network transmission. At the receiver side, bitstreams are received from the network, decoded to retrieve the expression parameters and pose parameters, and used to drive a NeRF-based 3D head representation model to generate and display facial images. In our system, different modules communicate through FIFO (a First-In First-Out buffer), which enable asynchronous data transfer and buffering between modules, thus facilitating real-time operation.

Our 3D video conferencing system not only supports viewing from a realistic perspective at the receiver side but also from any arbitrary angle. Moreover, since only the expression parameters and pose parameters are transmitted for each frame, the bit-rate of our system is significantly lower, while still providing video quality comparable to traditional video conferencing systems at a much higher bit-rate.

In the following subsections, we provide a detailed description of the functionalities and implementations of each module within the 3D generative video conferencing system.

Ingestion and Rendering

At the sender side, the Ingest module captures video data from a camera. This video data is subsequently segmented by frames and fed into a FIFO queue, which reliably forwards it to subsequent processing modules.

On the receiver side, the Render module is responsible for displaying the received video frames. To ensures that the video is rendered smoothly. Specifically, we have established a playback buffer for K frames and set the playback time for the first frame as the base time. Each subsequent frame is assigned a target playback time according to the frame rate.

Parameter Extract Module

This module extracts facial expression parameters and head pose parameters from video frames, employing the open-source CPEM model [15] equipped with pre-trained weights.

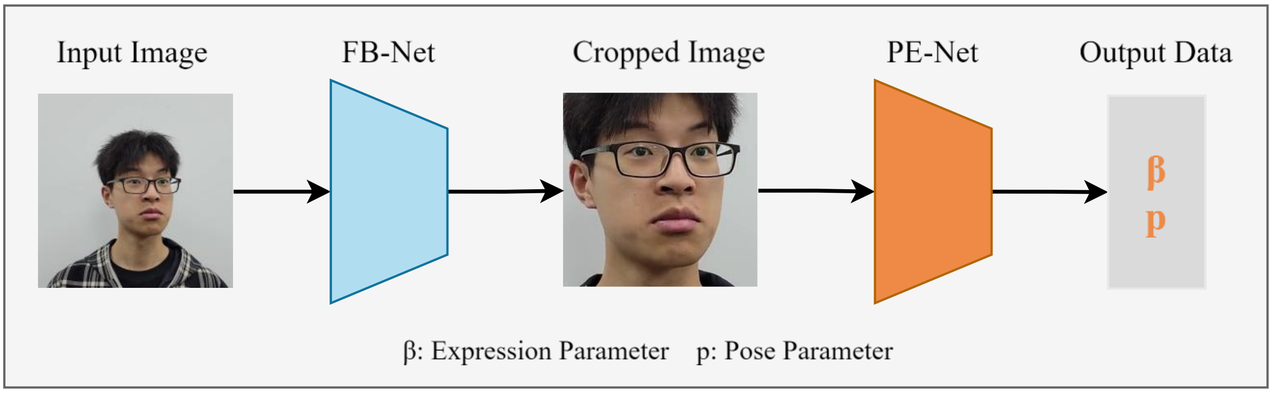

Figure 2– Parameter extraction module workflow. An input frame image is processed using a Face Bounding Network (FB-Net) to obtain a cropped image, which is then input into a Parameter Estimation Network (PE-Net) to extract expression and pose parameters

CPEM utilizes a linear 3D Morphable Model (3DMM) [26] as its 3D facial model, which comprises both shape and texture components. The shape component is subdivided into a facial base and an expression base. The expression parameters extracted by CPEM represent the coefficients of the expression base, guided by the FaceWarehouse [27] database, which distinctly annotates each expression base with specific semantics (e.g., Eye Close Left, Eye Squint Left). The pose parameters generated by CPEM include both rotational and translational elements. In our system, to ensure a consistent communication experience, the face is fixed at the centre, therefore only the rotational element is utilized.

The operational workflow of the Extract module is depicted in Figure 2. It starts with the Face Bounding Network, which segments out the facial region from the image. Subsequently, the segmented facial image is input into the Parameter Estimation Network, based on a ResNet50 architecture, to estimate the expression coefficients and pose parameters. The extracted parameter information is sent to a FIFO and then passed to the encoding module for encoding.

Encoding and Decoding

This process encodes the extracted expression and pose parameters (floating point numbers), including quantization, prediction, and zero-order exponential Golomb coding.

In the quantization step, each floating-point number is converted to an 8-bit integer. To improve the quantization accuracy, we record the maximum and minimum values of each dimension from our model training dataset. These bounds serve as the upper and lower limits for the expression parameters and pose parameters.

The prediction module utilizes inter-frame prediction to compute differences between successive frames, encoding only the resultant residuals via zero-order exponential Golomb coding, thus markedly reducing the volume of data to encode.

Zero-order exponential Golomb coding, an effective lossless compression technique, is particularly suited for data sequences predominantly comprising small values. This method, which builds on Golomb coding, utilizes straightforward rules for encoding non-negative integers, thereby efficiently compressing the data.

The decoding module performs inverse operations to recover the expression parameters and pose parameters.

Generate Module

The generation module at the receiver end utilizes expression and pose parameters to drive a NeRF model for generating facial images from specific viewpoints. During generation, the expression parameters dictate the facial expressions, while the pose parameters determine the viewing angle of the generated face. The receiver can either use the pose parameters transmitted from the sender for true-to-perspective facial generation or generate from arbitrary viewpoints through a pose control module. Our system supports free-viewpoint generation along the x, y, and z axes, within a range of ±45°.

This functionality is implemented using the open-source NerfBlenderShape [14][17]. NerfBlenderShape is a semantic facial model based on NeRF that can be trained in 10-20 minutes using short monocular RGB video inputs and can render facial images in tens of milliseconds based on expression and pose parameters. It represents facial semantics as an MLP-based implicit function and links expression bases with multi-level hash tables. Each hash table corresponds to specific facial semantics, and expression coefficients can combine encodings from multiple hash tables. Camera parameters used during rendering include intrinsic parameters and extrinsic parameters represented by the head pose.

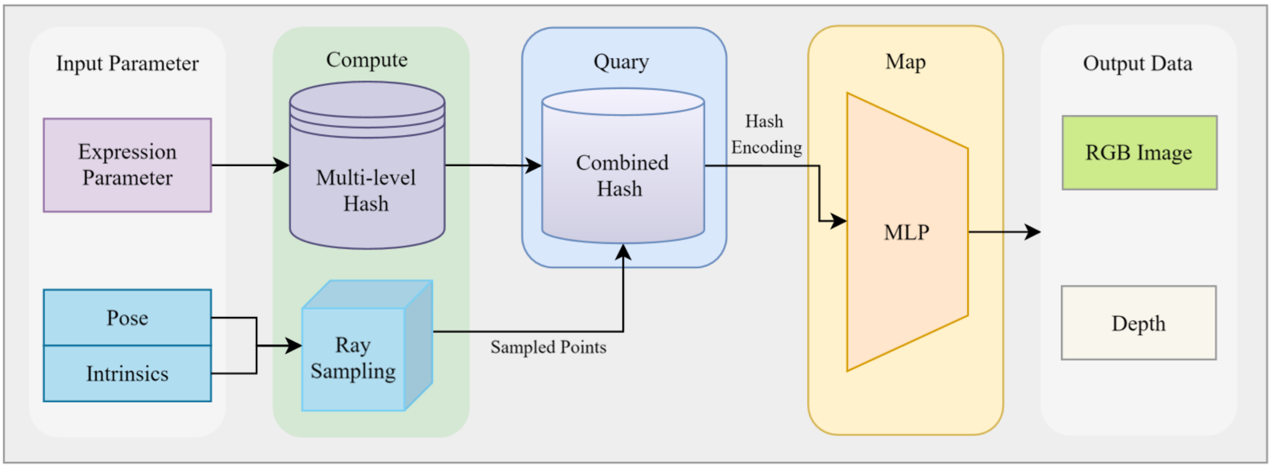

Figure 3– Generate module workflow

The detailed rendering process of NerfBlenderShape is illustrated in Figure 3. For each image to be rendered, the input parameters include expression parameters and camera parameters. The expression parameters are used to linearly combine multi-level hash tables in linear space to form a combined hash table. Camera parameters are fed into the Ray Sampling module to cast rays and obtain sample points. These sample points are then queried in the combined hash table to obtain hash encodings. Finally, these hash encodings are input into a lightweight MLP to produce the RGB image and depth information.

Building on the open-source methodology provided by NerfBlenderShape, we retrained the NeRF model using our proprietary dataset to derive the weights employed in the Generate module. Detailed descriptions of the training process are presented in Implement section.

Sending and Receiving

The sending and receiving modules use QUIC as the underlying transport protocol to enhance the efficiency and reliability of data transmission. The sending module first encapsulates the encoded facial expression and pose parameters into QUIC packets. These packets are then sent over the network, where they quickly and reliably reach the peer via the QUIC protocol. Leveraging QUIC's sequential characteristics and reliable transmission, the system ensures the continuity and integrity of expression and pose data. Additionally, QUIC's one Round-Trip Time (1RTT) handshake and robust congestion control mechanisms effectively reduce transmission latency.

We also provide a network simulation module that utilizes the TC tool to simulate network traces, which facilitates the verification of our proposed system's adaptability to weak network conditions. To better simulate various and extremely weak network conditions, we have opted to deploy our system on a single host. Network transmission occurs over the local loopback, controlled using the TC tool.

IMPLEMENTATION

We implement our 3D generative facial video conferencing system on an Ubuntu 20.04 64-bit operating system using Python. The entire system operates on two Intel® Xeon® Gold 6240 CPUs at 2.60GHz with 256 GB of RAM, and utilizes two GeForce RTX 4090 GPUs for neural network computations. To ensure real-time performance, the extraction module on the sender side and the generation module on the receiver side are each run on separate GPUs.

In our system, the dimension of the expression parameters is set to 46, and the dimension of the pose parameters (rotational elements only) is set to 3. This section details the specific implementation of key system modules.

Nerf Model Training

Following the methodology provided by RAD-NeRF [28], we constructed our own dataset. The dataset construction involves the following steps:

1. Video Recording: A 100-second video capturing a variety of facial expressions and head poses was filmed.
2. Semantic Segmentation: Each frame was semantically segmented to identify the background, head, neck, and torso.
3. Background Extraction: A stable background image was intelligently extracted by analysing the foreground and background within the image sequence.
4. Torso Image Extraction: Using the semantic segmentation images from step 2 and the original images, torso images were extracted, and the backgrounds in the original images were replaced with the stable background from step 3 to create ground truth images.
5. Facial Tracking: Pose parameters for each image were obtained using facial tracking, including 3-dimensional Euler angles and 3-dimension translations, which were then converted into standard 4x4 position matrices.
6. Expression Parameter Extraction: 46-dimensional expression parameters were extracted from each image using the CPEM model.
7. Training File Preparation: Expression and pose parameters were compiled into a training file.

Using the training methodology provided by NerfBlenderShape, we retrained our NeRF head model to suit our conference system. For optimization, we employed the Adam optimizer with an initial learning rate of 0.001, and momentum betas configured to (0.9, 0.99). To facilitate effective learning rate management throughout the training, we incorporated a MultiStepLR scheduler that dynamically adjusted the learning rate at predetermined epochs. The entire training process was designed to run 200 epochs.

Network Transmission Based on QUIC

Quiche [29] is an implementation of the QUIC transport protocol and HTTP/3 as specified by the IETF. It provides a QUIC kernel implemented in Rust and offers C/C++ APIs. Utilizing the Quiche library, we designed send and receive modules using C++. Communication between the Python-based system and the C++-based network transmission modules is achieved through a FIFO named pipe.

Specifically, at the sender side, a 2-byte delimiter is appended to the encoded data of each frame to distinguish between data of different frames. This bitstream, including the delimiter, is then written into the sender's FIFO. The C++-based send module reads the bitstream from the FIFO and transmits it using the QUIC protocol. On the receiving end, the receive module reads data from the QUIC stream and writes it into the receiver's FIFO. Subsequently, the system reads the received bitstream from the FIFO and separates it into individual frames based on the delimiter.

Accelerating Model Inference Speed

In order to facilitate real-time functionality in our 3D video conferencing system, we have adopted various methods to accelerate inference. Specifically, at the sender side, the facial segmentation and CPEM models are exported and inferred in the ONNX format, rather than utilizing direct PyTorch implementations. On the receiving end, the rendering of facial images is conducted using fp16 precision. Furthermore, both modules undergo a pre-warming process prior to system activation to improve the efficiency of model inference.

EXPERIMENT

In this section, we present the experimental results of the proposed system. Bit-rate and latency performance are the most critical indicators for video conferencing systems, directly impacting the user experience. We conducted experiments on these two metrics separately.

Bit-rate

The test videos include ten video sequences, each with a length of 1000 frames. Each frame is cropped to 512x512 and encoded with an 8-bit quantization depth.

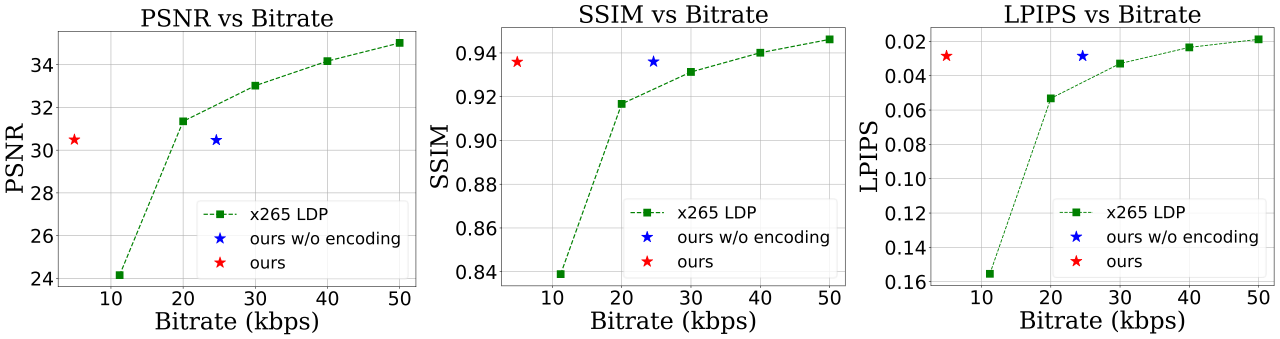

Figure 4–RD performance comparison.

We conduct a comparative analysis with the traditional HEVC encoder, specifically x265, using quantitative metrics such as PSNR, SSIM, and LPIPS [33]. Given that our system generates images containing only the head, while the x265 encodes image blocks, we adopt the following experimental setup to ensure comparability and fairness. We use semantic segmentation to replace the background of the test video with a plain white background, preserving only the head portion as our reference source video. We then encode the source video into a specific bit-rate H.265 stream using the FFmpeg tool. An example command is: *ffmpeg -i input -c:v libx265 -x265-params bframes=0 -b:v bitrate -r 20 output.* The background used at the reception side for generating facial images is also plain white. Figure 4 compares the performance of our NeRF-based approach and x265, and illustrates the impact of quantization on generation quality during the codec process.

As shown in Figure 4, with encoding process at 20fps and 512x512 resolution, our system achieved an average bit-rate of only 4.94kbps (red star), compared to 25.6kbps without encoding process (blue star). Despite a fourfold difference in bit-rate, the video quality achieved by both methods is nearly identical. This is attributed to our facial transmission being semantics-based, which is robust against minor semantic parameter errors, having negligible impact on pixel-based objective metrics. Therefore, the information loss due to quantization does not affect our generation quality, allowing us to save a significant amount of bit-rate.

The curves for PSNR, SSIM, and LPIPS for x265 are shown with green line. Due to bit-rate compression constraints, x265 can only encode test video at a minimum bit-rate of 11.2kbps, at which the PSNR is 24.15dB. It is noteworthy that our encoding scheme at 4.94kbps achieves a PSNR of 30.49 dB, equivalent to x265 at 18kbps. Our SSIM and LPIPS metrics are 0.9360 and 0.0285, respectively, comparable to x265 at 35kbps. Our video transmission is generation-based and does not aim for pixel-perfect recovery, yet we still outperform x265 significantly in objective assessment metrics.

Figure 5 displays a human perception comparison of the test images. It is evident that our advantages are more pronounced during fast head movements. For rapidly moving images, our system achieves higher subjective quality using only one-tenth the bit-rate of x265. This is because x265 relies on prediction between adjacent frames, whereas our semantics-based method independently processes expression parameters and pose parameters for each frame.

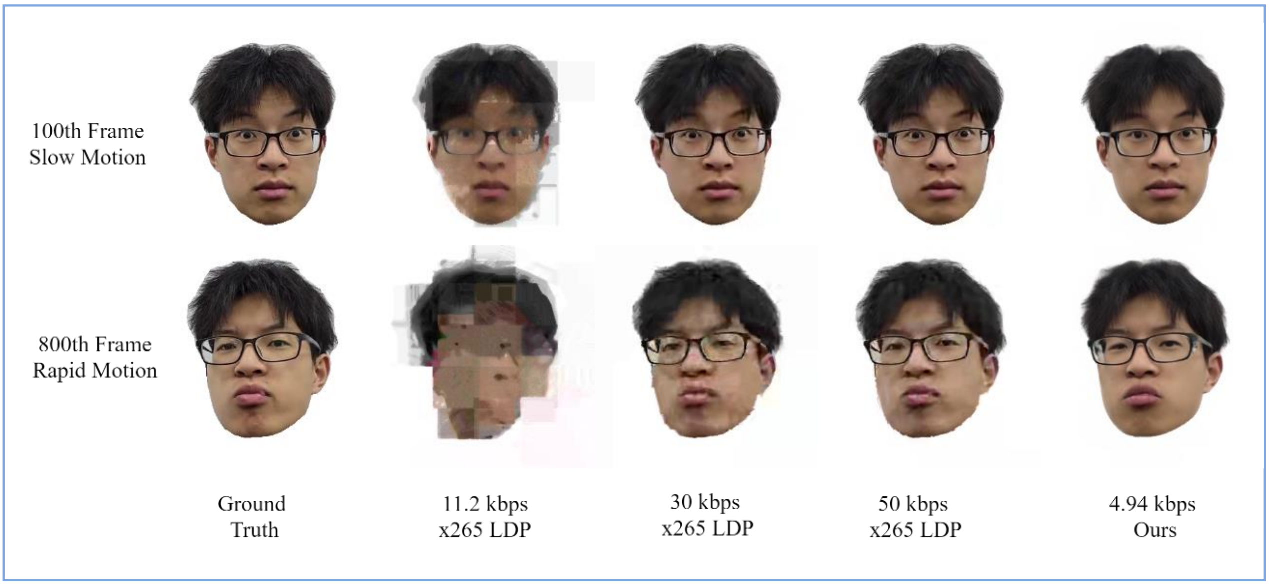

Figure 5– Subjective performance comparison

Latency

We set the Round-Trip Time (RTT) of the link to 20ms through our network simulation module. We have measured the computational latencies of individual modules when operating independently and as part of the integrated system, as shown in Table 1.

Delay test	Extract	Encode	Transport	Decode	Generate	End to End
Individual(ms)	14.7	0.5	12.46	0.15	27.6	---
Overall(ms)	25.6	0.8	13.69	0.21	49.2	89.5

Table 1– Latency of each system module and end-to-end latency

During full system operation, the latency of each module was less than 50ms, indicating that our system can operate in real-time at a frame rate of 20 fps. Moreover, the system's end-to-end latency of 89.5ms supports a smooth real-time communication (RTC) experience. The predominant contributor to the overall system latency is the generate module at the receiver end, which imposes a computational delay of 49.2ms, thus limiting our frame rate to 20 fps. Currently, our generate module operates using the PyTorch framework. Transitioning to ONNX for inference acceleration could yield higher frame-rates and further reduce the end-to-end latency.

DISCUSSION

This section discusses the limitations of our system and outlines future work:

- Due to current limitations in 3D reconstruction technology, our system only generates images of the head. In the future, we plan to include the upper body to enhance realism.
- Our system transmits only facial information, and the generated images lack backgrounds. Future efforts will focus on integrating 2D virtual backgrounds or placing participants' 3D head avatars within the same 3D virtual environment.
- We employ residual-based predictive coding for encoding and decoding. Network packet loss can impact the decoding of adjacent frames. To combat weak network conditions, we will introduce a group of pictures (GOP) strategy to limit decoding dependencies within a GOP. In the event of packet loss, we will retransmit parameters from I-frames, using them as references for subsequent frames to ensure reliable transmission.
- Our conferencing system requires participants to have pre-trained NeRF models of each other, and these NeRF-based head models are substantial in size (exceeding 500MB). For future practical deployments, we plan to incorporate online training. For new participants without models, we will initially conduct traditional video conferences and train individual NeRF head models based on received facial images within 20 minutes. When network conditions deteriorate, these pre-trained models can be utilized for conferencing. Additionally, there has been some work on compressing these models to less than 20MB [30][31], which we aim to integrate into our system.

CONCLUSIONS

In this paper, we propose and implement an ultra-low bit-rate generative 3D video conferencing system. Utilizing the latest advancements in 3D facial reconstruction technology, we achieved the goal of conducting 3D video conferences at an ultra-low bit-rate while providing acceptable video quality. Our system supports viewing from any angle, offering participants an immersive experience. In our future work, we will further refine our system, providing a new paradigm for video conferencing systems.

REFERENCES

1. G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, “Overview of the High Efficiency Video Coding (HEVC) Standard,” IEEE Trans. Circuits Syst. Video Technol., vol. 22, no. 12, pp. 1649–1668, Dec. 2012.

2. B. Bross et al., “Overview of the Versatile Video Coding (VVC) Standard and its Applications,” IEEE Trans. Circuits Syst. Video Technol., vol. 31, no. 10, pp. 3736– 3764, 2021.

3. J. Han et al., “A Technical Overview of AV1,” Proc. IEEE, pp. 1–28, 2021.

4. D. Feng, Y. Huang, Y. Zhang, J. Ling, A. Tang, and L. Song, “A Generative Compression Framework For Low Bandwidth Video Conference,” in 2021 IEEE International Conference on Multimedia Expo Workshops (ICMEW), Jul. 2021, pp. 1–6.

5. M. Oquab et al., “Low Bandwidth Video-Chat Compression using Deep Generative Models,” in 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Nashville, TN, USA, Jun. 2021, pp. 2388–2397.

6. Tang, Anni, et al. "Generative compression for face video: A hybrid scheme." 2022 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2022.

7. T.-C. Wang, A. Mallya, and M.-Y. Liu, “One-Shot Free-View Neural Talking-Head Synthesis for Video Conferencing,” in 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, Jun. 2021, pp. 10034–10044.

8. Konuko, Goluck, Giuseppe Valenzise, and Stéphane Lathuilière. "Ultra-low bitrate video conferencing using deep image animation." ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021.

9. Nirkin, Yuval, Yosi Keller, and Tal Hassner. "Fsgan: Subject agnostic face swapping and reenactment." Proceedings of the IEEE/CVF international conference on computer vision. 2019.

10. Siarohin, Aliaksandr, et al. "First order motion model for image animation." Advances in neural information processing systems 32 (2019).

11. J. Xu et al., “An ultra-low bitrate video conferencing system with flexible virtual access patterns” in IBC 2022.

12. 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, vol. 65, no. 1, pp. 99-106, 2021.

13. Kerbl, Bernhard, et al. "3d gaussian splatting for real-time radiance field rendering." ACM Transactions on Graphics 42.4 (2023): 1-14.

14. Gao, Xuan, et al. "Reconstructing personalized semantic facial nerf models from monocular video." ACM Transactions on Graphics (TOG) 41.6 (2022): 1-12.

15. Mo, Langyuan, et al. "Towards accurate facial motion retargeting with identity-consistent and expression-exclusive constraints." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 36. No. 2. 2022.

16. G. Gafni, J. Thies, M. Zollhöfer and M. Nießner, "Dynamic neural radiance fields for monocular 4d facial avatar reconstruction", Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8649-8658, June 2021.

17. Xuan, USTC-3DV/NeRFBlendShape-code. 2024. Accessed: May. 12, 2024. [Online]. Available: https://github.com/USTC3DV/NeRFBlendShape-code

18. Dhamo, Helisa, et al. "Headgas: Real-time animatable head avatars via 3d gaussian splatting." arXiv preprint arXiv:2312.02902 (2023).

19. Xiang, Jun, et al. "FlashAvatar: High-Fidelity Digital Avatar Rendering at 300FPS." arXiv preprint arXiv:2312.02214 (2023).

20. H. Schulzrinne, S. L. Casner, R. Frederick, and V. Jacobson, “RTP: A Transport Protocol for Real-Time Applications,” Internet Engineering Task Force, Request for Comments RFC 3550, 2003. doi: 10.17487/RFC3550.

21. K. Norrman, D. McGrew, M. Naslund, E. Carrara, and M. Baugher, “The Secure Realtime Transport Protocol (SRTP),” Internet Engineering Task Force, Request for Comments RFC 3711, 2004.

22. B. Marczak and J. Scott-Railton, “Move Fast and Roll Your Own Crypto: A Quick Look at the Confidentiality of Zoom Meetings,” University of Toronto, Citizen Lab Research Report No. 126, Apr. 2020.

23. A. Langley et al., “The QUIC Transport Protocol: Design and Internet-Scale Deployment,” in Proceedings of the Conference of the ACM Special Interest Group on Data Communication, Los Angeles CA USA, Aug. 2017, pp. 183–196.

24. C. Perkins and J. Ott, “Real-time Audio-Visual Media Transport over QUIC,” in Proceedings of the Workshop on the Evolution, Performance, and Interoperability of QUIC, Heraklion Greece, Dec. 2018, pp. 36–42.

25. M. Palmer, T. Krüger, B. Chandrasekaran, and A. Feldmann, “The QUIC Fix for Optimal Video Streaming,” in Proceedings of the Workshop on the Evolution, Performance, and Interoperability of QUIC, Heraklion Greece, Dec. 2018, pp. 43–49.

26. Blanz, Volker, and Thomas Vetter. "A morphable model for the synthesis of 3D faces." Seminal Graphics Papers: Pushing the Boundaries, Volume 2. 2023. 157-164.

27. Cao, Chen, et al. "Facewarehouse: A 3d facial expression database for visual computing." IEEE Transactions on Visualization and Computer Graphics 20.3 (2013): 413-425.

28. Tang, Jiaxiang, et al. "Real-time neural radiance talking portrait synthesis via audio-spatial decomposition." arXiv preprint arXiv:2211.12368 (2022).

29. Cloudflare, cloudflare/quiche. 2024. Accessed: May.12, 2024. [Online]. Available: https://github.com/cloudflare/quiche

30. Z. Zhang, A. Tang, C. Zhu, G. Lu, R. Xie and L. Song, "High-Fidelity Free-View Talking Head Synthesis for Low-Bandwidth Video Conference," 2023 IEEE International Conference on Visual Communications and Image Processing (VCIP), Jeju, Korea, Republic of, 2023, pp. 1-5

31. Zhang, Zhiyu, et al. "Efficient Dynamic-NeRF Based Volumetric Video Coding with Rate Distortion Optimization." arXiv preprint arXiv:2402.01380 (2024).

32. Cisco, “Cisco Visual Networking Index: Forecast and Trends, 2017–2022”,2018

33. Zhang, Richard, et al. "The unreasonable effectiveness of deep features as a perceptual metric." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

34. Lawrence, Jason, et al. "Project starline: A high-fidelity telepresence system." ACM Transactions on Graphics (TOG) 40.6 (2021): 1-16.

35. Apple Vision Pro. Apple. Accessed May 12, 2024. [Online]. Available: https://www.apple.com/apple-vision-pro/