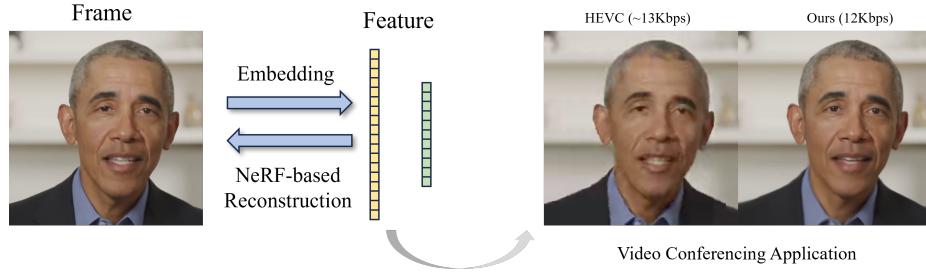


# Resolution-Agnostic Neural Compression for High-Fidelity Portrait Video Conferencing via Implicit Radiance Fields

Yifei Li<sup>1</sup>, Xiaohong Liu<sup>1\*</sup>, Yicong Peng<sup>1</sup>, Guangtao Zhai<sup>1</sup>, and Jun Zhou<sup>1\*</sup>

<sup>1</sup>Shanghai Jiao Tong University, Shanghai, China



**Fig. 1.** Illustration of our NeRF-based video compression. The core idea of our framework is frame-feature substitution for extremely low bandwidth. With NeRF-based face reconstruction model ensuring high-fidelity portrait generation, our framework shows significant compression performance for video conferencing application.

**Abstract.** Video conferencing has caught much more attention recently. High fidelity and low bandwidth are two major objectives of video compression for video conferencing applications. Most pioneering methods rely on classic video compression codec without high-level feature embedding and thus can not reach the extremely low bandwidth. Recent works instead employ model-based neural compression to acquire ultra-low bitrates using sparse representations of each frame such as facial landmark information, while these approaches can not maintain high fidelity due to 2D image-based warping. In this paper, we propose a novel low bandwidth neural compression approach for high-fidelity portrait video conferencing using implicit radiance fields to achieve both major objectives. We leverage dynamic neural radiance fields to reconstruct high-fidelity talking head with expression features, which are represented as frame substitution for transmission. The overall system employs deep model to encode expression features at the sender and reconstruct portrait at the receiver with volume rendering as decoder for ultra-low bandwidth. In particular, with the characteristic of neural radiance fields based model,

---

\* Corresponding author

our compression approach is resolution-agnostic, which means that the low bandwidth achieved by our approach is independent of video resolution, while maintaining fidelity for higher resolution reconstruction. Experimental results demonstrate that our novel framework can (1) construct ultra-low bandwidth video conferencing, (2) maintain high fidelity portrait and (3) have better performance on high-resolution video compression than previous works.

**Keywords:** Video conferencing · Neural radiance fields · Neural compression.

## 1 Introduction

Video conferencing enables individuals or groups to participate in a virtual meeting by using video, which has caught much more attention since the online lifestyle becomes prevalent. Nowadays, the demand for video conferencing with the large amount of simultaneous users also determines its extremely low bandwidth limitations in application, which relies more heavily on efficient video compression technologies. Video compression aims to reduce video bandwidth while maintaining high fidelity. Over the past several decades, the dominant video compression methods are based on classic video compression frameworks, such as H.262, AVS [4], H.264, HEVC [2], and VVC [3], which have achieved significant results. However, most classic methods reducing redundancy fully based on images and pixels without high level feature coding, thus can not reach the extremely limited low bandwidth while maintaining acceptable results in present video conferencing scenarios.

With the development of computer science and deep learning, video, as the main medium for simultaneous auditory and visual outputs, has received extensive attention for its applications and research in related fields [15,40,50,23,41,57,5,42,22]. At the same time, with the development of computer vision and graphics [13,45,25,52], there are more and more neural network based methods targeting the resolution and quality of videos [6,14,19,30,34,39,49,54,55,56]. Among them, the field of neural video compression has attracted much attention, where some neural compression methods [7,9,8,10,11] leverage face image generative models to deliver extreme compression by reconstructing video frame from a high-level feature, such as motion keypoints [8,10,12]. Specifically, most previous works use 2D warping based synthesis models to reconstruct portrait images. These warping methods deliver good reconstructions when the difference between the reference and target images is small, but they fail (possibly catastrophically) when there is large head pose movement or occlusion. As a result, lacking of 3D representation, these warping based compression frameworks are not robust in maintaining high fidelity for some cases. Furthermore, most of these generative approaches have restrictions on input resolution (e.g., usually  $256 \times 256$ ), which means when it comes to high resolution applications (e.g., typical video conferencing are designed for HD videos), corresponding neural compression will not work. Meanwhile, with the superior capability in multi-view image synthesis of

Neural Radiance Fields (NeRF) [16], several feature-conditioned dynamic neural radiance fields [1,17,18,24] have been proposed for talking head and dynamic face reconstruction. Rather than 2D warping, these models propose to use neural radiance fields to reconstruct portrait scene and represent the dynamics (e.g., expressions and head motion) as high-level features. Thanks to implicit 3D representation and volume rendering, these works are capable of producing natural portraits with high fidelity and more specifics (e.g., illumination and reflection) even in large movements. Nevertheless, to the best of our knowledge, the applications of such NeRF-based reconstruction model have not been delivered to neural video compression or video conferencing.

To address the defects of classic video codec and previous neural model-based compression and preserve both high fidelity and ultra-low bandwidth, we propose to leverage Neural Radiance Fields (NeRF) [16] to reconstruct portrait in implicit 3D space for model-based neural compression and video conferencing. Specifically, we propose a novel neural compression framework using implicit neural radiance fields. At the sender, instead of using warping keypoints, we leverage 3D Morphable Face Models (3DMMs) [20] to extract facial expression feature and head pose from portrait frame. Due to its disentanglement of face attributes as a 3D representation, 3DMMs can gain control of face synthesis better. Besides, to obtain higher-level information representation and better compression performance, we propose to employ an attention-based model [21] as encoder for feature embedding, which is called *fine-tuning embedding*. Before the features substituting frames to be transmitted, entropy coding as a lossless coding strategy is employed to compress the features further. Once the features have been received at the receiver, we leverage the feature-conditioned dynamic neural radiance fields to reconstruct the portrait video. It's worth noting that we refer to [1], which has desirable performance in both face and torso rendering, and replace the audio feature with expression feature to build the face reconstruction model employed in our approach. We conduct extensive experiments in both quantitative and qualitative aspects with comparisons to classic video codec and previous model-based video compression. We demonstrate that our approach can reach extremely low bandwidth while maintaining high fidelity for video conferencing application. Furthermore, thanks to the characteristic of NeRF rendering with unlimited resolution [16], our neural compression approach is *resolution-agnostic*.

To summarize, the contributions of our approach are:

- Firstly, we leverage neural radiance fields for extremely low-bandwidth video compression and high-fidelity video conferencing, which is resolution-agnostic. To the best of our knowledge, our approach is the first NeRF-based video compression method.
- Secondly, we holistically construct the framework for NeRF-based video compression and design fine-tuning embedding model to obtain fine-tuned feature as frame substitution to be transmitted for better and adaptive compression performance.

- Lastly, extensive experiments demonstrate that our proposed approach can achieve resolution-agnostic and ultra-low bandwidth with high fidelity preserving for applications in video conferencing, which significantly outperforms classic video codec (HEVC) and previous model-based compression methods.

## 2 Related Work

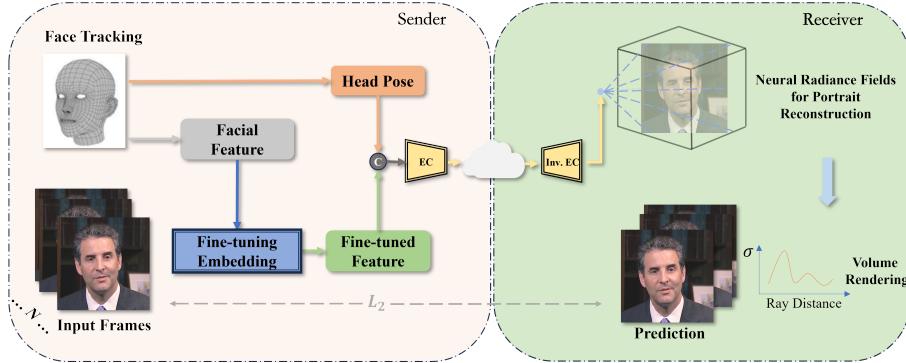
**Classic Video Codec** Many video applications utilize standard video compression modules, commonly known as codecs, including AVS, H.264/H.265 [27,2], VP8 [26], and AV1 [28]. These codecs employ a technique that divides video frames into key frames (I-frames), capitalizing on spatial redundancies within a frame, and predicted frames (P-/B-frames), leveraging both temporal and spatial redundancies across frames. Over time, these standards have undergone enhancements, incorporating concepts like variable block sizes and low-resolution encoding [28] to optimize performance at lower bitrates.

These codecs demonstrate notable efficiency in their slow modes, if ample time and computational resources to compress videos at high quality are available. Nevertheless, for real-time applications like video conferencing, they still demand a few hundred Kbps, even at moderate resolutions such as 720p. In situations with limited bandwidth, these codecs face challenges and may only transmit at lower quality, experiencing issues like packet loss and frame corruption [29].

**Face Animation Synthesis** Historically, face animation synthesis methods can be categorized into warping-based, mesh-based, and NeRF-based approaches. Among these, warping-based methods [31,32,33,9] are particularly popular within 2D generation techniques. In these methods, source features are warped using estimated motion fields to align the driving pose and expression with the source face. For example, Monkey-Net [35] constructs a 2D motion field from sparse keypoints detected by an unsupervised trained detector. Da-GAN [36] integrates depth estimation to enhance the 2D motion field by supplementing missing 3D geometry information. OSFV [8] attempts to extract 3D appearance features and predict a 3D motion field for free-view synthesis.

Certain traditional approaches [38,43] make use of 3D Morphable Models (3DMM) [20,37], enabling a broad range of animations through disentangled shape, expression, and rigid motions. Models like StyleRig [44] and PIE [46] leverage semantic information in the latent space of StyleGAN [47] to modulate expressions using 3DMM. PIRender [48] employs 3DMM to predict flow and warp the source image.

NeRF [16], a more recent method, represents implicit 3D scenes by rendering static scenes with points along different view directions, which initially gained prominence in audio-driven approaches [1,51,18] due to its compatibility with latent codes learned from audio.



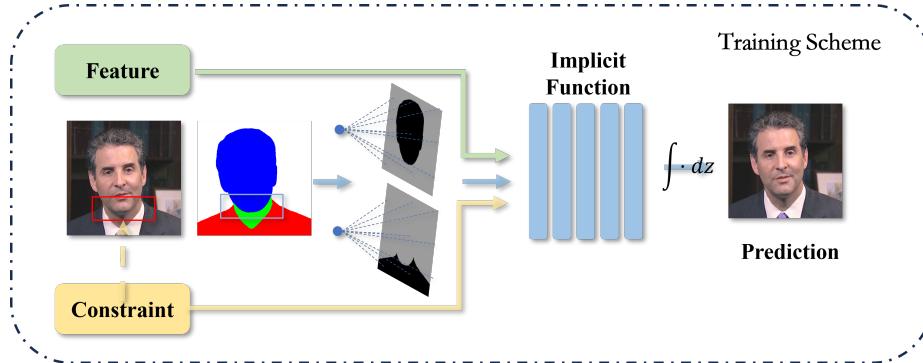
**Fig. 2.** The overall framework of our proposed method. Face feature is extracted at the sender and substitutes frame to be transmitted with ultra-low bandwidth. At the receiver, NeRF-based model takes the received feature as input to reconstruct portrait frame.

**Neural Compression for Video Conferencing** The limitations of classic codecs in achieving extremely low bitrates for high-resolution videos have prompted researchers to explore neural approaches for reconstructing videos from highly compact representations. Neural codecs have been specifically tailored for applications such as video streaming, live video, and video conferencing.

However, video conferencing presents distinct challenges compared to other video applications. Firstly, the unavailability of the video ahead of time hinders optimization for the best compression-quality trade-off. Additionally, video conferencing content predominantly consists of facial data, allowing for a more targeted model design for generating facial videos. Several models [7,9,8,10,11] have been proposed over the years, typically utilizing keypoints or facial landmarks as a compact intermediary representation of a specific pose. These representations are then used to compute the movement between two poses before generating the reconstruction. The models may incorporate 3D keypoints [8], off-the-shelf keypoint detectors [9], or a variety of reference frames [11] to enhance prediction.

**Neural Radiance Field and Dynamic Rendering** Our approach aligns with recent advancements in neural rendering and novel view synthesis, particularly drawing inspiration from Neural Radiance Fields (NeRF) [16]. NeRF employs a Multi-Layer Perception (MLP), denoted as  $F$ , to acquire a volumetric representation of a scene.  $F$ , for each 3D point and viewing direction, predicts color and volume density. Through hierarchical volume sampling,  $F$  is densely evaluated throughout the scene for a given camera pose, followed by volume rendering to generate the final image. The training process involves minimizing the error between the predicted color and the ground truth value of a pixel.

While NeRF is originally designed for static scenes, several efforts have been made to extend its applicability to dynamic objects or scenes. Some approaches [1,17,18] introduce a time component as input and impose temporal constraints



**Fig. 3.** Training scheme of the NeRF-based reconstruction model. We leverage consistency constraint code to get better generative results.

by utilizing scene flow or a canonical frame for talking head and face animation synthesis. For example, AD-NeRF [1] proposes to use an audio feature as additional input with head-torso separate modeling to reconstruct natural and photo-realistic face animation. Nevertheless, NeRF-based face reconstruction model has not been proposed for video compression and video conferencing.

### 3 Methodology

#### 3.1 NeRF-based Compression Framework

Our objective is to leverage neural radiance fields to design a video compression framework for extremely low-bandwidth video conferencing with high fidelity. Therefore, the overall framework of our proposed approach can be regarded as a communication system which is composed of the sender, the receiver and transmission. The key insight of the proposed approach is substituting face image with feature which can be represented as low-dimensional vector for transmission. Face tracking model and entropy coding are employed for facial feature extraction and further compression at the sender before transmission. At the receiver, the face animation model based on NeRF is used to reconstruct high-fidelity and photo-realistic portrait frames from the received features. Furthermore, the overall system is end-to-end which is illustrated in Fig. 2.

**Face Feature Extraction** To reconstruct high-fidelity portrait frame with low-bandwidth limitation, an appropriate representation of face is essential. Rather than extracting motion keypoints in self-supervised manner described in [7,12], we propose to employ 3DMM [20,37] as face tracking model to extract facial expression feature and head pose for face reconstruction. 3DMMs (3D Morphable Models) utilize a PCA (Principal Component Analysis)-based linear subspace to independently control face shape, facial expressions, and appearance. This

approach allows for a disentangled representation of these facial features, enabling more flexible and intuitive manipulation of individual components. This disentanglement is particularly valuable in applications such as face modeling and synthesis, which delivers precise control over specific aspects of the face. Therefore, we employ 3DMM as an intermediate 3D representation model to extract facial expression feature  $\delta$  and head pose  $p$ . Following [37], the primitive facial expression feature can be represented as a 79-dimensional vector. In terms of head pose, a 12-dimensional vector is employed: 9 numbers for the rotation and 3 numbers for the translation.

**Fine-tuning Embedding** However, in fact, the primitive face feature extracted using the pre-trained face tracking model is still redundant. To obtain lower bandwidth transmission and better performance in compression, we leverage an attention-based encoder network [21] to construct a fine-tuning embedding of primitive feature into lower-dimensional and higher-level representation. Specifically, the fine-tuned feature used in our experiment is a 30-dimensional vector.

**Further Compression** As for face feature, it's actually represented as vector with floating point values in 16 bits precision, and some classic compression schemes can be employed for further compression. Due to the characteristic of employed reconstruction model, the accuracy of the input features has a significant impact on generative performance. Therefore, the lossless compression scheme is recommended. In our approach, we compress fine-tuned face features further using *Entropy Coding*. Then the coded fine-tuned face feature, together with the coded pose, are transmitted to receiver.

**Portrait Frame Reconstruction** At the receiver, portrait frames are reconstructed from received face features using NeRF-based face reconstruction model, which hold the common facial expression and head poses as in *source* input images. Following the recent work of Guo Y *et al.* [1], we employ two individual neural radiance fields to represent head part and torso part separately which demonstrates significant performance in talking-head synthesis. Nevertheless, rather than the audio feature used in [1], we build the reconstruction model with facial expression feature from 3DMM as animation driving in order to maintain consistency in source and reconstructed facial expressions. Furthermore, we propose a learnable constraint to optimize the degree of fit between head and torso for better performance. More details of the reconstruction model are described in Sec. 3.2.

### 3.2 Neural Radiance Fields for Face Reconstruction

Inspired by audio driven neural radiance fields for talking-head synthesis introduced by Guo Y *et al.* [1], we utilize facial expression feature driven reconstruction model for neural compression. In addition to the view directions  $(\theta, \phi)$

and 3D locations  $(x, y, z)$ , the facial expression feature  $\delta$  is introduced as an additional input to the neural radiance field which is represented as an implicit function  $\mathcal{N}_\Theta$ . With the concatenated input vectors  $(\delta, \theta, \phi, x, y, z)$ , the network estimates color values  $\mathbf{c}$  accompanied by volume densities  $\sigma$  along the dispatched rays:

$$\mathcal{N}_\Theta(\delta, \theta, \phi, x, y, z) = (\mathbf{c}, \sigma). \quad (1)$$

**Consistency Constraint** In addition, apart from the different selection of driving feature, we observe that there will be a gap between head and torso in reconstruction following the individual optimization strategy introduced in [1], and thus we propose a *learnable constraint code* to improve the consistency between head and torso part, which is substantiated in the ablation study of our experiments. The overall training scheme of the reconstruction model is illustrated in Fig. 3.

**Volumetric Rendering of Face Radiance Fields** To generate images from this implicit geometry and appearance representation, we employ volumetric rendering. The process involves casting rays through each individual pixel of a frame, accumulating the sampled density and RGB values along the rays to calculate the final output color. Leveraging head pose tracking with 3DMM, we transform the ray sample points to the canonical space of the head model and then evaluate the dynamic neural radiance field at these locations. It's important to note that the pose  $P$ , obtained from head pose tracking, provides us with control over the head pose during test time. This control over head pose allows for dynamic adjustments and customization when rendering the images.

Once the color  $\mathbf{c}$  and volume density  $\sigma$  have been predicted by the implicit function  $\mathcal{N}_\Theta$ , the expected color  $\mathcal{C}$  of a camera ray  $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$  with camera center  $\mathbf{o}$  and viewing direction  $\mathbf{d} = (\theta, \phi)$  is accumulated as:

$$\mathcal{C}(\mathbf{r}; \Theta, P, \delta) = \int_{b_{near}}^{b_{far}} \sigma_\Theta(\mathbf{r}(t)) \cdot \mathbf{c}_\Theta(\mathbf{r}(t), \mathbf{d}) \cdot T(t) dt, \quad (2)$$

where  $b_{near}$  and  $b_{far}$  are near bounds and far bounds of sampling along the ray.  $T(t)$  is the accumulated transmittance along the ray from  $b_{near}$  to  $t$ :

$$T(t) = \exp\left(-\int_{b_{near}}^t \sigma(\mathbf{r}(x)) dx\right). \quad (3)$$

Besides, it's worth noting that we use a similar two-stage volumetric integration approach to Mildenhall *et al.* [16].

### 3.3 Optimization Details

**Dataset** We employ HDTF [53] as the main dataset for face animation reconstruction in the applications of video conferencing. We select videos of different identities from HDTF dataset [53]. There are several input resolutions for training:  $128 \times 128$ ,  $256 \times 256$ ,  $512 \times 512$  and  $1024 \times 1024$ .



**Fig. 4.** Qualitative results of the proposed framework compared with previous model-based compression (FOMM [7] and Bi-layer [9]) and classic video codec (HEVC [2]). Our approach, which employs NeRF-based model for high-fidelity reconstruction and feature-frame substitution for ultra-low bandwidth, outperforms other methods in image quality significantly. *f.t.* represents fine-tuning embedding employed in the framework.

**Training Loss** As the overall system is end-to-end, we leverage a photo-metric reconstruction error metric over the training images  $I_i$  to optimize both the coarse network and fine network:

$$L = \sum_{i=1}^M L_i(\Theta_c) + L_i(\Theta_f), \quad (4)$$

where  $\Theta_c$  and  $\Theta_f$  are parameters of coarse and fine networks and  $L_i$  is:

$$L_i = \sum_{j \in pixels} \|\mathcal{C}(\mathbf{r}_j; \Theta, P_i, \delta_i) - I_i[j]\|^2. \quad (5)$$

## 4 Experiments

### 4.1 Overview

The goal of the proposed framework is to construct resolution-agnostic NeRF-based compression for high-fidelity portrait video conferencing with extremely

low bandwidth. To demonstrate the significant performance of our approach for applications in video conferencing, we conduct both quantitative and qualitative evaluation compared with state-of-the-art model-based video compression approach and classic video codec and discuss ablation studies of our method.

**Metrics & Setting** We measure the performance of reconstruction-based models and classic codec using both quality metrics (**SSIM**, **PSNR**, **LPIPS**) and fidelity metrics. Specifically, following [32], we employ **CSIM**, **AUCON** and **PRMSE** to evaluate the fidelity. Cosine similarity (**CSIM**) is used to evaluate the quality of identity preservation. **PRMSE**, the root mean square error of the head pose angles is leveraged to inspect the capability of the model to properly reenact the pose and the expression of the driver. And **AUCON** represents the ratio of identical facial action unit values between generated images and driving images. As for qualitative evaluation, we design the similar bandwidth of classic codec as other methods and evaluate the quality of images. In terms of quantitative evaluation, we first compare both quality metrics and bitrate-quality trade-off, which is represented as **SSIM/b.r.**, **PSNR/b.r.** and **LPIPS×b.r.**, where **b.r.** represents bitrate. And then we compare the fidelity metrics and bitrate-fidelity trade-off, which is represented as **CSIM/b.r.**, **AUCON/b.r.** and **PRMSE×b.r.**. Furthermore, we also demonstrate the compression performance using rate-distortion curve.

## 4.2 Qualitative Evaluation

In terms of qualitative evaluation, we compare our method with the SOTA model-based compression Bi-layer [9] and FOMM [7] together with the most available and efficient classic video codec, HEVC. Specifically, we preserve the regular keypoint/landmark settings proposed in Bi-layer and FOMM and employ *Entropy Coding* for further compression as well to make comparisons. For HEVC, we choose the appropriate *Constant Rate Factor* to obtain similar bandwidth as our proposed method and compare the compression performance.

As illustrated in Fig. 4, our method can generate more realistic and high-fidelity results under extremely low bandwidth. Neither Bi-layer nor FOMM can reconstruct high-fidelity portrait due to their 2D warping based method. Classic codec HEVC has little implication on fidelity, while in similar condition (compared to ultra-low bandwidth in our method) there is much distortion that degrades the image quality. Consequently, our proposed framework delivers more appealing results for applications in video conferencing.

## 4.3 Quantitative Evaluation

With regards to quantitative evaluation, we first compare both the quantitative metrics (quality and fidelity metrics) and trade-off between the bitrate and quality/fidelity represented as **SSIM/b.r.**, **PSNR/b.r.**, **LPIPS×b.r.**, **CSIM/b.r.**,

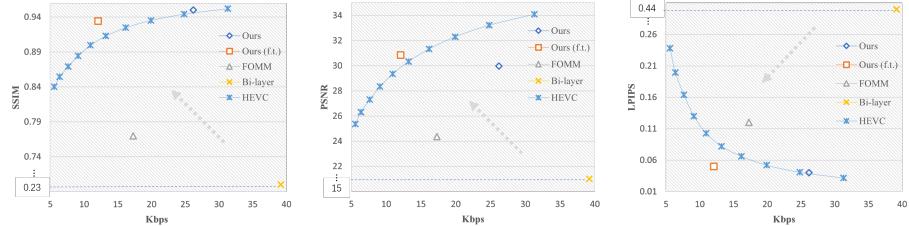
**Table 1.** Quantitative results over quality metrics and bitrate trade-off.

Methods	Quality				Quality-bitrate Tradeoff		
	L1↓	SSIM↑	PSNR↑	LPIPS↓	SSIM/b.r.↑	PSNR/b.r.↑	LPIPS×b.r.↓
FOMM [7]	0.038	0.77	24.37	0.12	0.04	1.41	2.07
Bi-layer [9]	0.23	0.55	15.88	0.44	0.014	0.41	17.23
HEVC [2]	0.019	0.89	28.83	0.091	0.068	2.22	1.21
Ours	<b>0.014</b>	<b>0.95</b>	<b>29.97</b>	<b>0.048</b>	0.036	1.144	<b>1.19</b>
Ours(f.t.)	<u>0.015</u>	<u>0.934</u>	<b>30.85</b>	<u>0.05</u>	<b>0.077</b>	<b>2.55</b>	<b>0.6</b>

**Table 2.** Quantitative results over fidelity metrics and bitrate trade-off.

Methods	Fidelity			Fidelity-bitrate Tradeoff		
	CSIM↑	AUCON↑	PRMSE↓	CSIM/b.r.↑	AUCON/b.r.↑	PRMSE×b.r.↓
FOMM [7]	0.829	0.856	2.79	0.048	0.0495	48.2
Bi-layer [9]	0.518	0.626	4.86	0.013	0.016	190
Ours	<b>0.956</b>	<b>0.989</b>	<b>1.21</b>	0.036	0.0377	<b>31.7</b>
Ours(f.t.)	<u>0.945</u>	<u>0.967</u>	<u>1.29</u>	<b>0.078</b>	<b>0.08</b>	<b>15.609</b>

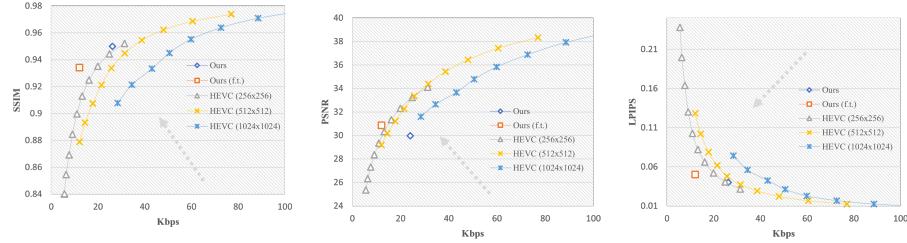
**AUCON/b.r.** and **PRMSE×b.r.**, as shown in Table 1 and Table 2. Furthermore, to demonstrate compression performance more clearly, we employ rate-distortion curve analysis in Fig. 5.



**Fig. 5.** Rate-distortion curve for our proposed framework compared with existing model-based compression method and classic codec HEVC. The resolution for HEVC codec is 256 × 256.

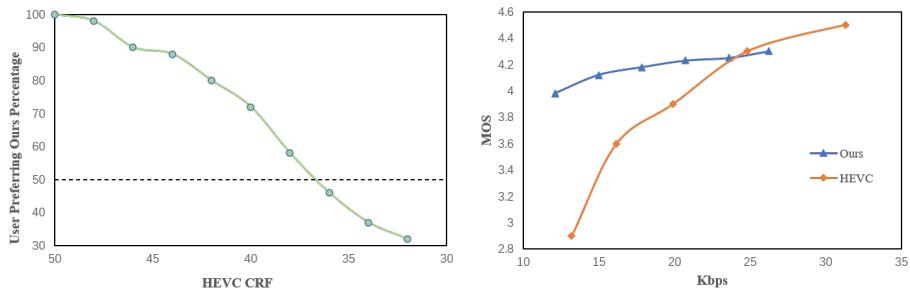
**Resolution-agnostic Analysis** It's worth noting that besides the significant bitrate-quality trade-off, our NeRF-based compression framework is resolution-agnostic due to the characteristic of neural radiance fields. That is the extremely low bandwidth achieved by our approach is independent of video resolution, while maintaining fidelity for higher resolution reconstruction and compression.

As illustrated in Fig. 6 rate-distortion curve, Bi-layer [9] and FOMM [7] have no support for variation in resolution, while higher resolution has significant affect on performance of HEVC.



**Fig. 6.** Rate-distortion curve of resolution-agnostic analysis for our proposed framework compared with classic codec HEVC in several different resolution settings.

**Subjective Evaluation** Following [8] and [12], we conduct subjective evaluation as well. We compress several clips using our framework and HEVC separately and show the compressed clips to users in video conferencing application. With various bitrate settings, we ask the users to choose the preference and compute the percentage as shown in the left side of Fig 7, and to rate the clips by *Mean Opinion Score* (MOS) as shown in the right side of Fig 7. And in extremely low bandwidth setting, our framework shows significant performance compared to HEVC.



**Fig. 7.** Subjective evaluation on the proposed framework and HEVC.

#### 4.4 Ablation Study

We also benchmark our performance gain upon our modules. Specifically, we conduct ablations about our proposed *fine-tuning embedding* model and head-torso *consistency constraint* code. As for fine-tuning embedding, we have demonstrated its significant effects on video compression performance from previous experimental results, where fine-tune embedding has little affect on image quality with similar fidelity (PSNR is even higher), and reduces bandwidth significantly. The ablation study of consistency constraint is described in Table 3.

**Table 3.** Ablation study on head-torso consistency constraint.

Setting	L1	SSIM	PSNR	LPIPS	CSIM	AUCON	PRMSE
w/o. constraint	0.019	0.91	28.68	0.06	0.94	0.958	1.31
w. constraint	<b>0.015</b>	<b>0.934</b>	<b>30.85</b>	<b>0.05</b>	<b>0.945</b>	<b>0.967</b>	<b>1.29</b>

## 5 Conclusion

In this work, we propose to leverage neural radiance fields face reconstruction model for neural video compression. Based on our NeRF-based reconstruction model, we substitute frames with features to be transmitted for video conferencing. With extensive experiments in both qualitative and quantitative aspects, we demonstrate that our novel framework implements resolution-agnostic neural compression with high-fidelity portraits in extremely low bandwidth for video conferencing, which outperforms the existing methods. As for future work, there are more further compression methods for the extracted facial feature besides lossless Entropy Coding, and we plan to leverage deep compression scheme for further feature compression to obtain better performance.

**Acknowledgement.** This work was supported in part by the Shanghai Pujiang Program under Grant 22PJ1406800.

## References

1. Guo Y, Chen K, Liang S, et al. Ad-nerf: Audio driven neural radiance fields for talking head synthesis[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 5784-5794.
2. Sullivan G J, Ohm J R, Han W J, et al. Overview of the high efficiency video coding (HEVC) standard[J]. IEEE Transactions on circuits and systems for video technology, 2012, 22(12): 1649-1668.
3. Bross B, Chen J, Liu S, et al. Versatile video coding (draft 5)[J]. Joint Video Experts Team (JVET) of Itu-T Sg, 2019, 16: 3-12.

4. Ma S, Zhang L, Wang S, et al. Evolution of AVS video coding standards: twenty years of innovation and development[J]. *Science China Information Sciences*, 2022, 65(9): 192101.
5. Nie X, Hu Y, Shen X, et al. Reconstructing and editing fluids using the adaptive multilayer external force guiding model[J]. *Science China Information Sciences*, 2022, 65(11): 212102.
6. Shi Z, Xu X, Liu X, et al. Video frame interpolation transformer[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 17482-17491.
7. Siarohin A, Lathuilière S, Tulyakov S, et al. First order motion model for image animation[J]. *Advances in neural information processing systems*, 2019, 32.
8. Wang T C, Mallya A, Liu M Y. One-shot free-view neural talking-head synthesis for video conferencing[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 10039-10049.
9. Zakharov E, Ivakhnenko A, Shysheya A, et al. Fast bi-layer neural synthesis of one-shot realistic head avatars[C]//Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16. Springer International Publishing, 2020: 524-540.
10. Oquab M, Stock P, Haziza D, et al. Low bandwidth video-chat compression using deep generative models[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 2388-2397.
11. Volokitin A, Brugger S, Benlalah A, et al. Neural Face Video Compression using Multiple Views[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 1738-1742.
12. Konuko G, Valenzise G, Lathuilière S. Ultra-low bitrate video conferencing using deep image animation[C]//ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021: 4210-4214.
13. Liu M, Wei Y, Wu X, et al. Survey on leveraging pre-trained generative adversarial networks for image editing and restoration[J]. *Science China Information Sciences*, 2023, 66(5): 1-28.
14. Shi Z, Liu X, Li C, et al. Learning for unconstrained space-time video super-resolution[J]. *IEEE Transactions on Broadcasting*, 2021, 68(2): 345-358.
15. Chen Y, Hao C, Yang Z X, et al. Fast target-aware learning for few-shot video object segmentation[J]. *Science China Information Sciences*, 2022, 65(8): 182104.
16. Mildenhall B, Srinivasan P P, Tancik M, et al. Nerf: Representing scenes as neural radiance fields for view synthesis[J]. *Communications of the ACM*, 2021, 65(1): 99-106.
17. Gafni G, Thies J, Zollhofer M, et al. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 8649-8658.
18. Yao S, Zhong R Z, Yan Y, et al. DFA-NeRF: Personalized talking head generation via disentangled face attributes neural rendering[J]. arXiv preprint arXiv:2201.00791, 2022.
19. Shi Z, Liu X, Shi K, et al. Video frame interpolation via generalized deformable convolution[J]. *IEEE transactions on multimedia*, 2021, 24: 426-439.
20. Blanz V, Vetter T. A morphable model for the synthesis of 3D faces[M]//Seminal Graphics Papers: Pushing the Boundaries, Volume 2. 2023: 157-164.
21. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. *Advances in neural information processing systems*, 2017, 30.
22. Wang H, Wu Y, Li M, et al. Survey on rain removal from videos or a single image[J]. *Science China Information Sciences*, 2022, 65(1): 111101.

23. Tian Y, Fu H, Wang H, et al. RGB oralscan video-based orthodontic treatment monitoring[J]. Science China Information Sciences, 2024, 67(1): 112107.
24. Lombardi S, Simon T, Saragih J, et al. Neural volumes: Learning dynamic renderable volumes from images[J]. arXiv preprint arXiv:1906.07751, 2019.
25. Ma S, Gao J, Wang R, et al. Overview of intelligent video coding: from model-based to learning-based approaches[J]. Visual Intelligence, 2023, 1(1): 15.
26. Bankoski J, Wilkins P, Xu Y. Technical overview of VP8, an open source video codec for the web[C]//2011 IEEE International Conference on Multimedia and Expo. IEEE, 2011: 1-6.
27. Schwarz H, Marpe D, Wiegand T. Overview of the scalable video coding extension of the H. 264/AVC standard[J]. IEEE Transactions on circuits and systems for video technology, 2007, 17(9): 1103-1120.
28. Chen Y, Murherjee D, Han J, et al. An overview of core coding tools in the AV1 video codec[C]//2018 picture coding symposium (PCS). IEEE, 2018: 41-45.
29. Fouladi S, Emmons J, Orbay E, et al. Salsify:Low-Latency network video through tighter integration between a video codec and a transport protocol[C]//15th USENIX Symposium on Networked Systems Design and Implementation (NSDI 18). 2018: 267-282.
30. Liu X, Shi K, Wang Z, et al. Exploit camera raw data for video super-resolution via hidden Markov model inference[J]. IEEE Transactions on Image Processing, 2021, 30: 2127-2140.
31. Dong H, Liang X, Gong K, et al. Soft-gated warping-gan for pose-guided person image synthesis[J]. Advances in neural information processing systems, 2018, 31.
32. Ha S, Kersner M, Kim B, et al. Marionette: Few-shot face reenactment preserving identity of unseen targets[C]//Proceedings of the AAAI conference on artificial intelligence. 2020, 34(07): 10893-10900.
33. Liu W, Piao Z, Min J, et al. Liquid warping gan: A unified framework for human motion imitation, appearance transfer and novel view synthesis[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 5904-5913.
34. Liu X, Kong L, Zhou Y, et al. End-to-end trainable video super-resolution based on a new mechanism for implicit motion estimation and compensation[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2020: 2416-2425.
35. Siarohin A, Lathuilière S, Tulyakov S, et al. Animating arbitrary objects via deep motion transfer[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 2377-2386.
36. Hong F T, Zhang L, Shen L, et al. Depth-aware generative adversarial network for talking head video generation[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 3397-3406.
37. Paysan P, Knothe R, Amberg B, et al. A 3D face model for pose and illumination invariant face recognition[C]//2009 sixth IEEE international conference on advanced video and signal based surveillance. Ieee, 2009: 296-301.
38. Suwajanakorn S, Seitz S M, Kemelmacher-Shlizerman I. Synthesizing obama: learning lip sync from audio[J]. ACM Transactions on Graphics (ToG), 2017, 36(4): 1-13.
39. Liu X, Chen L, Wang W, et al. Robust multi-frame super-resolution based on spatially weighted half-quadratic estimation and adaptive BTV regularization[J]. IEEE Transactions on Image Processing, 2018, 27(10): 4971-4986.
40. Huang Y, Yang C, Chen Z. 3DPF-FBN: video inpainting by jointly 3D-patch filling and neural network refinement[J]. SCIENCE CHINA-INFORMATION SCIENCES, 2022, 65(7).

41. Yi Z, Song W, Li S, et al. Automatic image matting and fusing for portrait synthesis[J]. Science China Information Sciences, 2022, 65(2): 124101.
42. Qian R, Lin W, See J, et al. Controllable augmentations for video representation learning[J]. Visual Intelligence, 2024, 2(1): 1-15.
43. Thies J, Zollhöfer M, Nießner M. Deferred neural rendering: Image synthesis using neural textures[J]. Acm Transactions on Graphics (TOG), 2019, 38(4): 1-12.
44. Tewari A, Elgharib M, Bharaj G, et al. Stylerig: Rigging stylegan for 3d control over portrait images[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 6142-6151.
45. Fan D P, Ji G P, Xu P, et al. Advances in deep concealed scene understanding[J]. Visual Intelligence, 2023, 1(1): 16.
46. Tewari A, Elgharib M, Bernard F, et al. Pie: Portrait image embedding for semantic control[J]. ACM Transactions on Graphics (TOG), 2020, 39(6): 1-14.
47. Karras T, Laine S, Aila T. A style-based generator architecture for generative adversarial networks[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 4401-4410.
48. Ren Y, Li G, Chen Y, et al. Pirenderer: Controllable portrait image generation via semantic neural rendering[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 13759-13768.
49. Huang S, Liu X, Tan T, et al. TransMRSR: Transformer-based Self-Distilled Generative Prior for Brain MRI Super-Resolution[J]. arXiv preprint arXiv:2306.06669, 2023.
50. Li K, Guo D, Wang M. ViGT: proposal-free video grounding with a learnable token in the transformer[J]. Science China Information Sciences, 2023, 66(10): 202102.
51. Shen S, Li W, Zhu Z, et al. Learning dynamic facial radiance fields for few-shot talking head synthesis[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022: 666-682.
52. Li W, Wang Z, Mai R, et al. Modular design automation of the morphologies, controllers, and vision systems for intelligent robots: a survey[J]. Visual Intelligence, 2023, 1(1): 2.
53. Zhang Z, Li L, Ding Y, et al. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 3661-3670.
54. Yin G, Jiang X, Jiang S, et al. Online Video Streaming Super-Resolution with Adaptive Look-Up Table Fusion[J]. arXiv preprint arXiv:2303.00334, 2023.
55. Wu G, Liu X, Luo K, et al. Accflow: Backward accumulation for long-range optical flow[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 12119-12128.
56. Wei B, Wen Y, Liu X, et al. SOFNet: Optical-flow based large-scale slice augmentation of brain MRI[J]. Displays, 2023, 80: 102536.
57. Zhou Z, Meng M, Zhou Y, et al. Model-guided 3D stitching for augmented virtual environment[J]. Science China Information Sciences, 2023, 66(1): 112106.