FAIVCONF: FACE ENHANCEMENT FOR AI-BASED VIDEO CONFERENCE WITH LOW BIT-RATE

Zhengang Li\textsuperscript{1,2}, Sheng Lin\textsuperscript{1}, Shan Liu\textsuperscript{1}, Songnan Li\textsuperscript{1}, Xue Lin\textsuperscript{2}, Wei Wang\textsuperscript{1} and Wei Jiang\textsuperscript{1}

\textsuperscript{1}Tencent Media Lab, \textsuperscript{2}Northeastern University

\{barneylin, shanl, sunnysnli, rickwang101, vwjiang\}@tencent.com, \{li.zhen, xue.lin\}@northeastern.edu.

ABSTRACT

Recently, high-quality video conferencing with fewer transmission bits becomes a very hot and challenging problem. We propose FAIVConf, a specially designed video compression framework for video conferencing, based on the effective neural human face generation techniques. FAIVConf brings together several designs to improve the system robustness in real video conference scenarios: face swapping to avoid artifacts in background animation; facial blurring to decrease transmission bit-rate and maintain quality of extracted facial landmarks; and dynamic source update for face view interpolation to accommodate a large range of head poses. Our method achieves significant bit-rate reduction in video conference and gives much better visual quality under the same bit-rate compared with H.264 and H.265 coding schemes.

Index Terms— AI Video Conference, Face Generation, Deep Learning

1. INTRODUCTION

There has been a growing demand for video conferencing solutions to provide stable and high-quality video communication with low latency. Due to the limited bandwidth resources, it is critical to effectively reduce the video bit-rate to ensure a smooth user experience with little transmission lag or artifacts. Traditional video coding tools such as H.264/H.265 aim at compressing general video content, which, albeit may be improved through continuous study, are suboptimal for the video conferencing scenario with its unique characteristics.

Human faces are especially important in video conferences, which usually occupy a large portion of the frames and also are the main focus of the frames. Therefore, it is particularly important to improve the compression quality of the human faces in this scenario. With the great success of the generative adversarial network (GAN), significant progress has been made in human face generation. These methods \cite{1,2} usually generate face images based upon face-related semantic information such as facial landmarks \cite{3}, estimated poses \cite{4}, and segmentation masks \cite{5}. Such face-related data have a much smaller scale compared to the original frames, which makes it possible to build an extremely low bit-rate and high-quality coding framework for enhancing visual quality in video conferencing. For example, NVIDIA’s video conferencing solution \cite{6} transfers only a keypoint representation of faces, which is used to reconstruct the source frames in decoder. Although theoretically appealing, such methods suffer from several challenges in real applications, such as (complex) background generation, large differences for pose transfer, large illumination mismatch, facial occlusions, etc.

Targeting at robust performance in real applications, we propose FAIVConf, a video compression framework specially designed for practical video conferences. Our contribution can be summarized as follows:

First, we propose a simple yet effective facial blurring mechanism to decrease the bit-rate of transmitting the facial area while keeping the essence of facial features, since blurring can effectively reduce the blocky artifacts caused by heavy compression, which affects accuracy of facial landmark extraction. To ensure high-quality face generation with a large range of head poses in real applications, we propose a dynamic updating mechanism for the view interpolation method in the face reenactment module. Based on the above improvements, we propose the FAIVConf to reduce bit-rate and optimize the video transmission quality in the video conference scenario. Our proposed framework achieves significant bit-rate reduction, \textit{i.e.}, only 0.001875 bits per pixel for transmitting a conference video with $800 \times 800$ pixels, which is about 0.8\% of streaming the original video. Compared with the commercial H.264 and H.265, our method gives much better visual quality under the same bit-rate.

2. MOTIVATION

GAN-based guided face generation enables highly efficient compression in the scenario of video conference. Guided by only facial landmarks, one can achieve optimal compression rate since only one source image and successive keypoints need to be transferred to the decoder for face generation. Such
methods assumes simple clean background and limited pose difference between the source and the drive image, without significant occlusions over the face area. Unfortunately, these assumptions are often violated in practical video conference sessions, and these methods suffer from severe artifacts.

To avoid modeling unconstrained backgrounds, we need to confine the face generation only to the true face area, and compress facial and non-facial areas separately. To avoid the difficulty of face reenactment with large pose and illumination differences, we also transfer a low-quality (e.g., highly blurred) version of the original frame to guide the face generation. Fig. 1 gives an example of face generation with or without using the low-quality face area in the original frame. The face generation result with Gaussian blurred face (bottom right) can simulate the lighting condition much better than the result using only background (top right). In addition, we propose dynamic updating face view interpolation mechanism using multiple source images to further improve the robustness of the generated face with regard to flexible head pose in real video conference. One caveat here is the increased transmission overhead. The background is compressed and transferred by itself to maintain the fidelity of the video background. Since face is the main changing part in video conferences, which usually occupy the majority of bits. The background can be compressed efficiently in general. Regarding the detected face area, to reduce the transmission overhead as much as possible, we perform facial landmark extraction and face segmentation all on the decoder side, and only transfer a blurred version of an extended face area. We conduct Gaussian blur before encoding the face area to achieve a good balance between reduced bit-rate and quality of extracted landmarks. We observe that the accuracy of facial landmark extraction is quite sensitive to the blocky artifacts from compression. Gaussian blur can not only perform low-pass filtering to reduce the blocky artifacts for high compression ratio, but also lead to better bit-rate due to the blurred content.

3. THE FAIVConf FRAMEWORK

As shown in Fig. 2, the proposed FAIVConf framework mainly consists of two parts. Encoder and Decoder.

**Encoding Process.** Let \( \{D_1, D_2, D_3, ..., D_n\} \) be the input conference video frames (driving video) to be transferred, where \( D_i \) is the \( i \)-th driving frame in a total number of \( n \) frames. Let \( \{S_1, S_2, S_3, ..., S_m\} \) be a set of source frames which has been collected from encoder side and stored in decoder side. The goal of encoding process is to extract the feature of each frame which can be transferred with small bandwidth to decoder. The transferred feature should be enough for the decoder to generate the output video \( \{O_1, O_2, O_3, ..., O_n\} \), which is highly similar to driving video.

Specifically, we first apply face detection on the driving frame to get the bounding box of the face position \( B_i \). Then we compute the head pose \( P_i \) and perform face segmentation in parallel. The head pose information is compared with that of each source frame to see if similar poses exist. If the answer is yes, Gaussian blur will be applied on the segmented face area to decrease the video compression load. Then, the blurred frame will be encoded by the H.265 codec, and transferred to decoder side with the face bounding box and head pose. If the answer is no, the whole frame \( D_i \) will be transferred to the decoder side directly, and the face area in \( B_i \) will be stored in decoder as a new source face.

**Decoding Process.** The decoding process is executed at the user terminal. After receiving the transferred data of the \( i \)-th driving frame, i.e., the recovered frame \( D_i' \) (after H.265 decoding), the head pose \( P_i \) (expressed as Euler angles), and the face bounding box \( B_i \), the decoder extracts both the facial landmarks \( L_i \) and the segmented face of \( D_i' \) in parallel. Then one or several source frames will be chosen to perform the face reenactment according to the head pose similarity as shown in Sec. 3. The landmarks \( L_i \) and reenacted output from the \( k \)-th source frame is represented as \( O_i(k) (1 \leq k \leq m) \):

\[
L_i = G_l(B_i; D_i'); \\
O_i(k) = G_r(S_k; L_i)
\]

(1)

where \( G_l \) and \( G_r \) are the facial landmark detector and reenactment generator, respectively. Then we compute the interpolated reenacted output \( O_i \) through face view interpolation, which will be discussed in detail in Sec. 3. Then, inpainting [7] and blending [8] will be performed to swap \( O_i \) back into the transferred driving frame \( D_i' \) within the bounding box area defined by \( B_i \) to generate the final output frame.

**Dynamic source pool Update for Face View Interpolation.** To overcome the difficulty of image animation posed by a large range of facial orientations in real applications, we incorporate the face view interpolation technique [1] in our framework. This method allows using multiple source frames rather than a single image to generate the face reenactment result and makes it possible to synthesize faces with different pose with little distortion. We further improve this method.
by dynamically updating the source frames through the video conference session to ensure small pose differences for quality reenactment.

For the source frames set (SFS) \( \{ S_1, S_2, ... S_m \} \), the corresponding Euler angles are represented as \( \{ e_1, e_2, ... e_m \} \). Each angle \( e_j \) consists of yaw \( y_j \), pitch \( p_j \) and roll \( r_j \). We first project the Euler angle into a plane by dropping roll, i.e., \( \{ e_1(y, p), e_2(y, p), ... e_m(y, p) \} \), and then remove points in the angular domain that are too close to each other with a distance threshold \( L \). A mesh can be built based upon the remaining points in the angular domain by Delaunay Triangulation.

For the \( i \)-th driving frame with head pose \( P_i \), its corresponding projected point in the angular domain is \( e^d_i(y, p) \). We find the most related three source frames \( S_{k_1}, S_{k_2}, S_{k_3} \) at the vertices of the triangle which contains \( e^d_i(y, p) \). With the barycentric coordinates \( (\lambda_{k_1}, \lambda_{k_2}, \lambda_{k_3}) \) of \( e^d_i(y, p) \) in the triangle, the face view interpolation result \( O_i \) can be calculated as:

\[
O_i = \sum_{r=1}^{3} \lambda_k O_i(k_r),
\]

where \( O_i(k_r) \) is the reenacted output from the source frame \( S_{k_r} \) corresponding to the \( k_r \)-th vertex.

In addition, we update the set of source frames dynamically. When the projected point of a driving frame in the angular domain is not in any of the existing triangles, we only use the source frame of the closest existing vertex, and set \( \lambda \) as 1. If a driving frame is far from all source frames in the angular domain, i.e., distances between the projected points of the driving frame and the source frames are all above a threshold \( L \), this driving frame will be added into the set of source frames and transferred to the decoder. The Delaunay Triangulation mesh will be rebuilt to include the new source frame to enable a more comprehensive range of head pose.

### 4. EXPERIMENTS

**Experiment Setup.** We used the mixed video sequences of the IJB-C [9] and Human-centric video matting [10] dataset to train our reenactment model, and used the WFLW dataset [11] to train the facial landmark detector. Hopenet [4] was used as our pose estimator in our framework.

For the reenactment model training, we used two VGG-19 models [12] pretrained on VGGFace2 [13] and CelebA [14] datasets to compute the perceptual loss [15] towards face recognition and face attribute classification, respectively. This perceptual loss has been wildly used recently [1, 6]. Besides that, we also used the multi-scale GAN loss [16] and the pixel-wise loss in the training process to enhance the generator’s performance.

**Landmark Detection of Blurred Face.** We first preprocessed the WFLW dataset with Gaussian blur within the face area to simulate the face blurring of frames transmitted between the encoder and decoder. Then, we trained HRNetV2 to extract facial landmark with the same setting in the work of [3]. As shown in Table 1, we compare the normalized mean error (NME) of HRNetV2 (Trained on original WFLW) and HRNetV2-blur (trained on blurred WFLW data). We can see that training landmark detector with blurred data can significantly decrease the NME of all the subsets in the Gaussian blur situation, and the loss is just a little higher than that of the original model on non-blurred dataset. This means that HRNetV2-blur on blurred image has the same accuracy level compared with the original model on the same image without blurring. Therefore, HRNetV2-blur is good enough to handle transmitted frames in decoder.

**Performance on Conference Video.** To evaluate our performance in video conference, we propose a High-resolution Talking Head (HSTH) dataset, which contains a scene where a single person is talking with high resolution up to 800×800 collected from YouTube. Our method can re-create a talking-
When the head rotation angle is larger than 45°, our approach shows better performance especially when it is greater than 45 degrees, the face generated by OSFV gradually deforms and becomes flattened, while FAIVConf can still generate high quality images which maintain the identity and fidelity similar to the original frames.

5. CONCLUSION

We introduce FAIVConf, a framework dedicated to highly efficient video compression for the video conference scenario. Compared with the commercial H.264/H.265 coding schemes, our approach achieves much better visual video quality with less transmission data. We achieve significant bit-rate reduction under similar visual quality via only 0.8% of streaming the original video in a resolution of 800×800 pixels.

6. REFERENCES

[1] Nirkin et al., “Fsgan: Subject agnostic face swapping and reenactment,” in ICCV, 2019.
[2] Siarohin et al., “First order motion model for image animation,” NeurIPS, 2019.
[3] Wang et al., “Deep high-resolution representation learning for visual recognition,” TPAMI, 2020.
[4] Ruiz et al., “Fine-grained pose estimation without keypoints,” in CVPR workshops, 2018.
[5] Ronneberger et al., “U-net: Convolutional networks for biomedical image segmentation,” in MICCAI, 2015.
[6] Wang et al., “One-shot free-view neural talking-head synthesis for video conferencing,” in CVPR, 2021.
[7] Yeh et al., “Semantic image inpainting with deep generative models,” in CVPR, 2017.
[8] Wu et al., “Gp-gan: Towards realistic high-resolution image blending,” in ACM/MM, 2019.
[9] Maze et al., “Larpa janus benchmark-c: Face dataset and protocol,” in ICB, 2018.
[10] CVPR, “Human-centric video matting challenge,” https://maadaa.ai/cvpr2021-human-centric-video-matting-challenge/, 2021.
[11] Wu et al., “Look at boundary: A boundary-aware face alignment algorithm,” in CVPR, 2018.
[12] Simonyan et al., “Very deep convolutional networks for large-scale image recognition,” arXiv, 2014.
[13] Cao et al., “Vggface2: A dataset for recognising faces across pose and age,” in FG, 2018.
[14] Liu et al., “Large-scale CelebFaces attributes (celeba) dataset,” Retrieved August, 2018.
[15] Johnson et al., “Perceptual losses for real-time style transfer and super-resolution,” in ECCV, 2016.
[16] Wang et al., “Few-shot video-to-video synthesis,” arXiv, 2019.