Wireless Semantic Communications for Video Conferencing

Peiwen Jiang, *Graduate Student Member, IEEE*, Chao-Kai Wen, *Senior Member, IEEE*, Shi Jin, *Senior Member, IEEE*, and Geoffrey Ye Li, *Fellow, IEEE*

***Abstract*— Video conferencing has become a popular mode of meeting despite consuming considerable communication resources. Conventional video compression causes resolution reduction under a limited bandwidth. Semantic video conferencing (SVC) maintains a high resolution by transmitting some keypoints to represent the motions because the background is almost static, and the speakers do not change often. However, the study on the influence of transmission errors on keypoints is limited. In this paper, an SVC network based on keypoint transmission is established, which dramatically reduces transmission resources while only losing detailed expressions. Transmission errors in SVC only lead to a changed expression, whereas those in the conventional methods directly destroy pixels. However, the conventional error detector, such as cyclic redundancy check, cannot reflect the degree of expression changes. To overcome this issue, an incremental redundancy hybrid automatic repeat-request framework for varying channels (SVC-HARQ) incorporating a novel semantic error detector is developed. SVC-HARQ has flexibility in bit consumption and achieves a good performance. In addition, SVC-channel state information (CSI) is designed for CSI feedback to allocate the keypoint transmission and enhance the performance dramatically. Simulation shows that the proposed wireless semantic communication system can remarkably improve transmission efficiency.**

***Index Terms*— IR-HARQ, CSI feedback, semantic communication, video conferencing, facial keypoints.**

I. Introduction

IN SEMANTIC communications [1], the shared, local knowledge, extracted from the set of transmission contents, helps compress the transmission information and correct the transmission errors according to the semantic correlation. However, the design of a practical semantic transceiver is difficult, especially the extraction of semantic knowledge using traditional methods due to the lack of appropriate mathematical models. Deep learning (DL) enables the implementation of semantic communication, as evidenced by many related works [2], [3], [4].

DL can address many challenging issues [5], [6], [7], [8], [9] in communication systems, such as nonlinear interference, insufficient pilots, channel estimation, channel coding, channel state information (CSI) feedback, and autoencoder-based communication systems. These studies focus on the technical level and exploit DL to extract the features of channel environments and outperform the traditional design. In semantic communications, DL is used to extract the semantic information from the contents, and shared, local knowledge at the semantic level is implicitly contained in the trained parameters. DL-based semantic communication systems are usually designed using joint source and channel coding methods, and trained for a specific transmission content, including image [10], [11], [12], [13], video [14], speech [15], and text [16], [17], [18]. Furthermore, for some specific tasks, such as object recognition [13] and scene classification [19], semantic communications can substantially reduce transmission overhead.

Video conferencing has become an essential part of our work at present, especially during the pandemic of COVID-19. A high-resolution video transmission requires a substantial amount of transmission resources. Therefore, it is very challenging, especially for conferencing over mobile phones. In [20], DL is exploited to enhance the conventional video compression algorithm through extracting semantic areas. In [21], a text transcript is used to represent the voice and lip motions, which considerably reduces the bandwidth. Apart from driving the voice and lip with a text transcript, driving a video with a few keypoints is widely applied in face swapping in [22]. The keypoints of the frames from a driving video are extracted to represent the motion of facial expressions, and a generator is exploited to enable a source image to move similarly to the driving video. In [23], the facial driving video is considered the transmitted video, and the receiver can restore the driving video from the transmitted keypoints and the photo of the speaker. However, the existing studies only relate to the source coding module, and the effect of errors from the physical channel transmission is unclear. In addition, a semantic-based video conferencing transmission should protect key information facing physical transmission

Manuscript received 8 April 2022; revised 16 September 2022; accepted 2 October 2022. Date of publication 21 November 2022; date of current version 19 December 2022. The work of Chao-Kai Wen was supported in part by the National Science and Technology Council of Taiwan under Grant MOST 111-2218-E-110-003. This work was supported in part by the National Key Research and Development Program under Grant 2018YFA0701602 and in part by the National Natural Science Foundation of China (NSFC) under Grant 61941104 and Grant 61921004. *(Corresponding author: Shi Jin.)*

Peiwen Jiang and Shi Jin are with the National Mobile Communications Research Laboratory, Southeast University, Nanjing 210096, China (e-mail: PeiwenJiang@seu.edu.cn; jinshi@seu.edu.cn).

Chao-Kai Wen is with the Institute of Communications Engineering, National Sun Yat-sen University, Kaohsiung 80424, Taiwan (e-mail: chaokai.wen@mail.nsysu.edu.tw).

Geoffrey Ye Li is with the Department of Electrical and Electronic Engineering, Imperial College London, SW7 2AZ London, U.K. (e-mail: geoffrey.li@imperial.ac.uk).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/JSAC.2022.3221968.

Digital Object Identifier 10.1109/JSAC.2022.3221968

errors, and the received videos should be acceptable under varying channels.

This paper concentrates on keypoint transmission and the changes it brings to the design of the communication system. A basal network for semantic video conferencing (SVC) is established into a three-level framework. First, the entire transmission is investigated, and the difference between semantic and conventional errors is analyzed. Then, acknowledgement (ACK) feedback is added to the SVC; it is a widely used technique in the conventional wireless communications to ensure a successful semantic transmission. An incremental redundancy hybrid automatic repeat request (IR-HARQ) framework for SVC, called SVC-HARQ, is proposed to guarantee the quality of the received frames when facing wicked channels. Then, the transmitter learns to allocate information with varying importance according to signal-to-noise ratios (SNR) at different subchannels with the help of CSI, which is called SVC-CSI. The major contributions of this work are summarized as follows:

1) Establishing SVC Framework: The state-of-art technology to restore the image from several keypoints [23] has achieved a considerable compression ratio. Thus, the technology is exploited to cope with channel distortion. Different from the substantial compression ratio (only few number of keypoints) that causes the transmitted image to lose the detailed expressions, the simulation results show that transmission errors in physical channel transmission may change the locations of the keypoints and lead to inaccurate expressions. Nevertheless, these errors may be visually acceptable through semantic processing. On the contrary, the errors usually directly destroy the pixels for the conventional methods.

2) Combining With HARQ Scheme: To guarantee the feasibility of SVC under varying channels, the IR-HARQ feedback framework for SVC, called SVC-HARQ, is developed. Compared with the conventional bit error detection using cyclic redundancy check, a semantic error detector is used to decide whether the received frame requires an incremental transmission. The semantic error detector exploits the fluency of the video to check the received frame. The simulation results demonstrate that inaccurate keypoints usually reduce the fluency. The proposed SVC-HARQ can adapt different bit error rates (BER) and require transmitting fewer bits than the competing methods.

3) Exploiting CSI: CSI is exploited so that optimal transmitted information with different importance can be allocated automatically on different subchannels, which is called SVC-CSI. SVC-CSI learns to allocate more information at the subchannels with high SNRs than at those with low SNRs. An extra incremental transmission is trained without employing CSI because the performance of SVC-CSI worsens when the testing channel environment is different from the training environment. Thus, it is robust to varying channels and is called SVC-CSI-HARQ.

The rest of this paper is organized as follows. Section II introduces the system model and the related methods, including conventional IR-HARQ and adaptive modulation. Section III describes the proposed networks. Section IV introduces module designs and training strategies. Section V demonstrates the superiority of the proposed networks in terms of semantic metrics and required bits. Section VI concludes this paper.

II. System Model and Related Works

In this section, the existing frameworks on semantic networks are first described, and some important techniques in wireless communications that can help semantic transmission are introduced. Finally, the challenges when applying semantic transmission over wireless communication systems are discussed.

A. Semantic Frameworks

To transmit source information, such as a picture $\mathbf{p}$, the semantic transmitter first extracts its meaning. The semantic extractor plays a role similar to that of the source encoder in the conventional communication systems and is denoted as $S(\mathbf{p})$. Then, the channel encoder, $C(\cdot)$, can be designed separately or jointly with the semantic extractor, and the encoded symbols are generated for channel transmission. The whole encoder process can be expressed as

$$\mathbf{s} = C(S(\mathbf{p})). \tag{1}$$

For a conventional orthogonal frequency division multiplexing (OFDM) system, the encoded symbols, $\mathbf{s}$, are modulated into OFDM symbols with K subcarriers, $[s_1, \ldots, s_K]$.

After passing through a wireless channel, the demodulated OFDM signal can be expressed as

$$\mathbf{y} = \mathbf{h} \cdot \mathbf{s} + \mathbf{z}, \tag{2}$$

where $\mathbf{h}$ is the channel in the frequency domain, $\cdot$ is the Hadamard product, and $\mathbf{z}$ is the Gaussian noise vector with each entry being mean 0 and variance σ^2.

For a frequency-selective channel, $\mathbf{h} = [h_1, \cdots, h_K]$ has different channel gains at the subchannels. With channel gains, the received symbols can be estimated by

$$\hat{\mathbf{s}} = \left[\frac{y_1}{\hat{h}_1}, \ldots, \frac{y_K}{\hat{h}_K}\right]. \tag{3}$$

Then, the overall SNR can be expressed as

$$SNR = \frac{\sum_{k=1}^{K} \|h_k \cdot s_k\|^2}{K\sigma^2}. \tag{4}$$

The transmitted picture can be recovered at the receiver by

$$\hat{\mathbf{p}} = S^{-1}(C^{-1}(\hat{\mathbf{s}})), \tag{5}$$

where $S^{-1}(\cdot)$ and $C^{-1}(\cdot)$ represent the semantic source decoder and channel decoder, respectively. As indicated in [1], the semantic processing and transmission in semantic communications are remarkably different from the conventional ones. The local, shared knowledge in the semantic systems plays a major role. Semantic knowledge can be exploited implicitly or explicitly, as summarized in the following:

1) Implicit Semantic Knowledge: In these designs [14], [15], [16], [17], the local, shared knowledge is implicitly contained in the trainable parameters and the transceivers are usually trained in an end-to-end manner. The effect of the

physical channels is also learned implicitly. These methods automatically extract the semantic features and cope with the distortion and interference in the physical channels. However, the trained parameters are difficult to adjust under changing transmit sources or physical environments.

2) Explicit Semantic Knowledge: In some specific tasks, the semantic knowledge is shared explicitly. For example, the semantic network in [24] shares the most important features in the image so that the received image can be classified better than the conventional methods. In [23], the photo of the speaker is shared because the appearance of the speaker does not change much during a speech. The explicit shared knowledge can be adjusted according to the change in the source information, such as replacing the photo for the next speaker.

Apart from the semantic knowledge, the existing methods have not considered adjusting the settings under different channel environments. Therefore, the semantic methods cannot adapt to physical channel variation.

B. Link Settings in Conventional Methods

In this section, two key techniques in wireless communications to cope with changing environments, which can be exploited in semantic system design, are introduced. In modern communication systems with HARQ, corrupted packets are retransmitted. IR-HARQ can balance the requirements of transmission resources and accuracy, and is a popular option. To establish an IR-HARQ system, a channel encoder and an error detector are needed. If the semantic symbols, $\mathbf{k}$, are protected by a conventional channel encoder, $C(\cdot)$, then the coded symbol can be expressed as

$$\mathbf{s} = C(\mathbf{k}). \tag{6}$$

In [25], the coded symbol vector can be divided into $\mathbf{s}_1$ and $\mathbf{s}_2$ with $\mathbf{s} = [\mathbf{s}_1, \mathbf{s}_2]$, where $\mathbf{s}_1$ corresponds to the coded semantic symbols with a high coding rate, and $\mathbf{s}_2$ represents the incremental symbols. The high code rate symbols, $\mathbf{s}_1$, are transmitted first. Denote $\hat{\mathbf{s}}_1$ as the received symbols corresponding to $\mathbf{s}_1$. The recovered semantic symbols can be expressed as

$$\hat{\mathbf{k}} = C_1^{-1}(\hat{\mathbf{s}}_1). \tag{7}$$

The conventional CRC error detector is widely used for HARQ systems and extra parity bits are coded from $\mathbf{p}$ and transmitted at the very beginning. With extra parity bits at the receiver, $\mathbf{b}_{\mathrm{CRC}}$, ACK information can be generated by

$$\mathrm{ACK} = Det_{\mathrm{CRC}}(\hat{\mathbf{k}}, \mathbf{b}_{\mathrm{CRC}}), \tag{8}$$

where $Det_{\mathrm{CRC}}(\cdot)$ denotes the error detection. When no error is found, feedback signal ACK is assigned a value of 1, else 0.

The incremental symbols, $\mathbf{s}_2$, need to be transmitted to decrease the code rate if some errors are founded (ACK=0). The received coded symbols are combined and decoded again, yielding

$$\hat{\mathbf{k}} = C^{-1}([\hat{\mathbf{s}}_1, \hat{\mathbf{s}}_2]). \tag{9}$$

If the decoded result still has errors, then retransmission starts, and the above process is repeated. This IR-HARQ method can deal with varying channels in different time slots.

If OFDM is used, then the overall channel bandwidth can be divided into K parallel flat fading subchannels with different SNRs [26]. Then, the diversity on channel conditions at different frequencies of the same time slot can be exploited. For OFDM systems, the subchannel gains are different, whereas the noise powers at different subchannels are the same, that is, once σ^2 and $[h_1, \ldots, h_K]$ are available at the transmitter, the modulation of s_k can be adaptive to cope with the changing gains of the subchannels.

Although the combination of the semantic networks and conventional link adaptive methods is naturally considered, the novel mechanism on semantic transmission brings challenges in the design at the technical level. As a deep and inexplicable network, the performance of the semantic-based transceiver must be guaranteed under varying physical environments.

III. Novel Frameworks for SVC

In this section, novel architectures for SVC, which exploit conventional strategies in wireless communications, are introduced. Starting with a basic network as a semantic source encoder, a novel error detector is proposed to generate an ACK feedback. The basic network is expanded into the HARQ mode to cope with varying channels at different time slots. Finally, the CSI for each subchannel is fed back to the semantic transmitter for adaptive modulation.

A. Basic Semantic Network for Video Conferencing

Restoring a specific face in an image from few keypoints has been studied in [23] and [22]. In these methods, the keypoints contain the changing information of facial expression and manner. Other information, such as appearance features, does not change during a speech and can be shared with the receiver in advance. Moreover, [23] presents that the keypoints can be compressed and encoded to improve transmission efficiency. The above methods dramatically reduce the requirement of the transmit resource. However, the existing methods only focus on the framework of source coding and ignore the influence of varying wireless channels.

A complete SVC framework is shown in Fig. 1, where simple dense layers are introduced as a channel coding module. The whole framework consists of three levels similar to [1], namely, effectiveness, semantic, and technical levels. The effectiveness level delivers the motion and expression of the speaker. The conventional goal is to minimize the difference of the transmitted and recovered frames. At the semantic level, the photo of the speaker is shared in advance given that the speaker has no remarkable change during the speech. Usually, the first frame of the video is shared with the receiver for convenience, whereas a photo with a distinct face is beneficial to generating a good image at the receiver. The keypoint detector extracts the movement of the face in the current frame, and these keypoints are transmitted at the technical level. Based on the received keypoints and the shared photo, the semantic part of the receiver reconstructs the

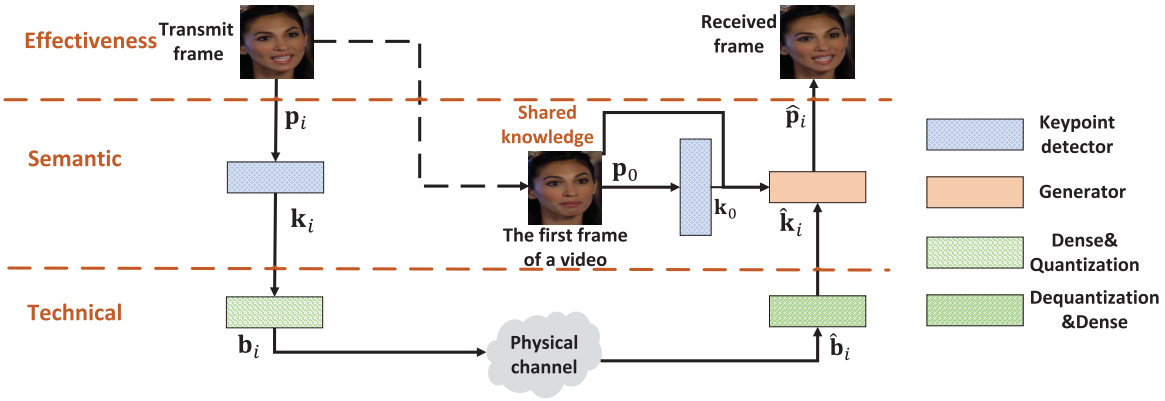


Fig. 1. Three-level framework of semantic video conferencing.

frame. The networks at the technical level are trained to cope with distortion and interference from physical channels. Based on the above description, SVC has three subnets, namely, a keypoint detector and a generator at the semantic level, and an encoder-decoder at the technical level.

The keypoint detector extracts n coordinates of the keypoints, $\mathbf{k}_i \in \mathbb{R}^{2\times n}$, from the i-th frame, $\mathbf{p}_i \in \mathbb{R}^{256\times256\times3}$, yielding

$$\mathbf{k}_i = KD(\mathbf{p}_i; \mathbf{W}_{\rm KD}), \tag{10}$$

where $\mathbf{W}_{\rm KD}$ denotes the set of trainable parameters of the keypoint detector. Especially, the first frame, $\mathbf{p}_0$, with its keypoints $\mathbf{k}_0$, is shared with the receiver.

The encoder-decoder consists of dense and quantization/dequantization layers. The whole process is expressed as

$$\mathbf{b}_i = Q(f_{\rm en}(\mathbf{k}_i; \mathbf{W}_{\rm en})), \tag{11}$$

where $\mathbf{W}_{\rm en}$ is the set of trainable parameters in the dense layers.

The dequantizer, $Q^{-1}(\cdot)$, at the receiver at the technical level is the inverse process of $Q(\cdot)$ to recover m real numbers from the received bits, $\hat{\mathbf{b}}_i$. This process can be expressed as

$$\hat{\mathbf{k}}_i = f_{\rm de}(Q^{-1}(\hat{\mathbf{b}}_i); \mathbf{W}_{\rm de}), \tag{12}$$

where $\mathbf{W}_{\rm de}$ is the set of trainable parameters in the dense layers.

The generator reconstructs the current frame from the shared image, $\mathbf{p}_0$, with its keypoints, $\mathbf{k}_0$, and the received keypoints of the i-th frame, $\hat{\mathbf{k}}_i$. This process is denoted as $G(\cdot; \mathbf{W}_{\rm G})$, where $\mathbf{W}_{\rm G}$ denotes the set of trainable parameters. Therefore, the i-th frame can be recovered by

$$\hat{\mathbf{p}}_i = G(\mathbf{p}_0, \mathbf{k}_0, \hat{\mathbf{k}}_i; \mathbf{W}_{\rm G}). \tag{13}$$

The proposed SVC is a combination of video synthesis and theoretic three-level semantic transmission in wireless communications. This basic framework is established and trained to study the effect of replacing video transmission with semantic keypoint transmission. The performance of semantic transmission can be improved further by introducing ACK feedback in wireless networks, as shown in the following section.

B. Semantic HARQ With ACK Feedback for Video Conferencing

HARQ can cope with time-varying channels in wireless communications. Retransmission and transmitting incremental symbols are flexible under changing channels with ACK feedback. Thus, a novel SVC framework with HARQ, called SVC-HARQ, is developed to improve semantic transmission.

Fig. 2 shows the receiver feeds an ACK signal back to the transmitter after the first transmission. The first transmission is the same as in Fig. 1, and the trained parameters can be used directly. The first transmitted bit vector, $\mathbf{b}_{1,i}$, can be expressed as

$$\mathbf{b}_{1,i} = Q(f_{\rm en}(KD(\mathbf{p}_i; \mathbf{W}_{1,\rm KD}); \mathbf{W}_{1,\rm en})), \tag{14}$$

where $\mathbf{W}_{1,\rm KD}$ is the set of trainable parameters in the keypoint detector, and $\mathbf{W}_{1,\rm en}$ is the set of trainable parameters in the encoder. Then, the recovered frame is

$$\hat{\mathbf{p}}_{1,i} = G(\mathbf{p}_0, \mathbf{k}_0, f_{\rm de}(Q^{-1}(\hat{\mathbf{b}}_{1,i}); \mathbf{W}_{1,\rm de}); \mathbf{W}_{1,\rm G}), \tag{15}$$

where $\mathbf{W}_{1,\rm G}$ is the set of parameters in the generator of the first transmission, and $\hat{\mathbf{b}}_{1,i}$ represents the received bit sequence at the first transmission. Then, the reconstructed frame, $\hat{\mathbf{p}}_{1,i}$, is evaluated by a semantic detector. If the detector finds $\hat{\mathbf{p}}_{1,i}$ unacceptable, then ACK=0 is fed back to the transmitter, and an incremental transmission is triggered.

The incremental bit sequence is transmitted to correct the errors. Different from the first transmission, the incremental transmission only concentrates on fallible keypoints under wicked channel conditions. Thus, the incremental transmission also needs to be trained and has different trainable parameters, namely, $\mathbf{W}_{2,\rm KD}$, $\mathbf{W}_{2,\rm en}$, and $\mathbf{W}_{2,\rm G}$, for the keypoint detector, decoder, and generator, respectively. The incremental transmitted bit sequence is

$$\mathbf{b}_{2,i} = Q(f_{\rm en}(KD(\mathbf{p}_i; \mathbf{W}_{2,\rm KD}); \mathbf{W}_{2,\rm en})), \tag{16}$$

and the recovered frame is

$$\hat{\mathbf{p}}_{2,i} = G(\mathbf{p}_0, \mathbf{k}_0, f_{\rm de}(Q^{-1}([\hat{\mathbf{b}}_{1,i}, \hat{\mathbf{b}}_{2,i}]); \mathbf{W}_{2,\rm de}); \mathbf{W}_{2,\rm G}), \tag{17}$$

where $[\hat{\mathbf{b}}_{1,i}, \hat{\mathbf{b}}_{2,i}]$ is the received symbol vector that includes symbols corresponding to the first and incremental transmission. This framework only shows one incremental

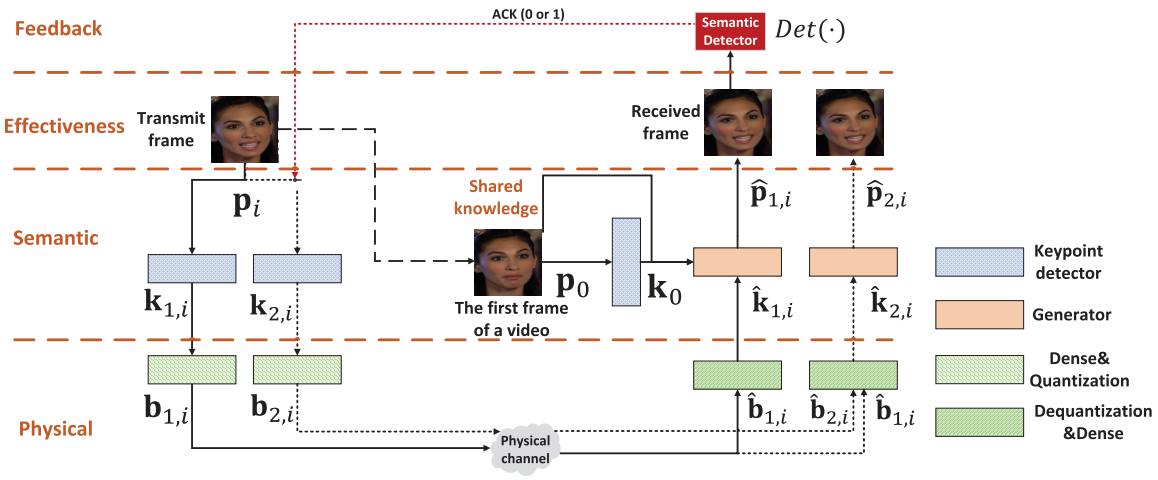


Fig. 2. Structure of SVC-HARQ with ACK feedback.

transmission. But it can be extended to more incremental transmissions if the semantic detector finds errors.

The above description indicates that the semantic detector is the key module of SVC-HARQ because the detector directly decides whether incremental transmission or retransmission is required. The conventional error detector, CRC, in the HARQ system is unsuitable for SVC-HARQ because the difference between some subtle errors in the received frames is acceptable for the conferee. An image quality assessment method [27] is used to evaluate whether the received frame is acceptable.

The errors in the received keypoints do not directly decrease the image quality, but they change the facial expressions. The generator can reconstruct an acceptable face image even if the keypoints have some errors because the general appearance is obtained from the shared image. The error keypoints only change the current expression and cause the video to be not fluent. To detect these changes, a novel fluency detector is proposed and its detailed architecture is introduced in Section IV.

The proposed quality and fluency detectors are compared for HARQ systems under semantic-based video conferencing. Overall, an extra incremental transmission can make SVC more adaptive under changing channels if the semantic detector is effective. Moreover, retransmission is started if the incremental symbols cannot correct the errors to reach the criterion of the detector.

C. Adaptive Encoding With CSI Feedback

The above SVC methods do not exploit CSI further. However, the noise power of the subchannels can be obtained by the receiver. For example, frequency-selective channels can be divided into different subchannels with different SNRs. The CSI of all subchannels is estimated by the receiver and shared with the transmitter. These channel conditions are exploited by the encoder-decoder at the technical level, which helps protect the most important keypoints. The accurate CSI of each subchannel cannot be obtained in practice, and the feedback of the entire CSI values requires resources. Thus, the receiver sorts the subchannels by their channel conditions and feeds this sequence back to the transmitter. This method simplifies the design of the encoder-decoder at the technical level and reduces the feedback cost.

Compared with the original SVC, that with CSI feedback (SVC-CSI) only needs to add a sort module $SN(\cdot)$, as shown in Fig. 3(a). The output of the sort module is denoted as $\mathbf{b}_i^{\mathrm{CSI}} = SN(\mathbf{b}_i)$, where its elements, representing subchannel gains, are in decreasing order. This process is expressed as

$$\mathbf{b}_i^{\mathrm{CSI}} = SN(\mathbf{b}_i) = [\ldots, \mathbf{b}_{i,K}, \ldots], \tag{18}$$

where $[\mathbf{b}_{i,1}, \ldots, \mathbf{b}_{i,K}]$ is rearranged according to $SN(\cdot)$ to make $\mathbf{b}_{i,1}$ at the subchannel with the highest gain, $\mathbf{b}_{i,2}$ at the subchannel with the second highest gain, ..., and $\mathbf{b}_{i,K}$ at the subchannel with the lowest gain. Then, $\mathbf{b}_i^{\mathrm{CSI}}$ is sent to the receiver. At the receiver, the received $\hat{\mathbf{b}}_i^{\mathrm{CSI}}$ is restored by $SN^{-1}(\cdot)$, yielding

$$\hat{\mathbf{b}}_i = SN^{-1}(\hat{\mathbf{b}}_i^{\mathrm{CSI}}) = [\hat{\mathbf{b}}_{i,1}, \ldots, \hat{\mathbf{b}}_{i,K}]. \tag{19}$$

The above methods encode the keypoints into a bit sequence, which can be easily applied in the conventional wireless communication systems, such as the OFDM system with quadrature amplitude modulation (QAM). Furthermore, joint design with modulation module to encode keypoints into constellation points directly can further improve performance. Thus, the benefit of CSI feedback on the encoding keypoints into constellation points is also investigated.

On the left structure of Fig. 3(b), the quantization in Fig. 3(a) is directly replaced with a dense layer and Tanh activation function. Its output has m real symbols, which denote $m/2$ constellation points. These points are also rearranged according to CSI feedback. This method is called full-resolution constellation because the learned constellation points can appear anywhere in the constellation.

However, the full-resolution constellation is extremely complex for practical systems due to finite precision. Thus, the constellation points need to be limited. Two bits are combined into a real symbol and every two real symbols are combined into a complex constellation point similar to 16-QAM. Each 2-bit vector is coded by the two shared trainable parameters,

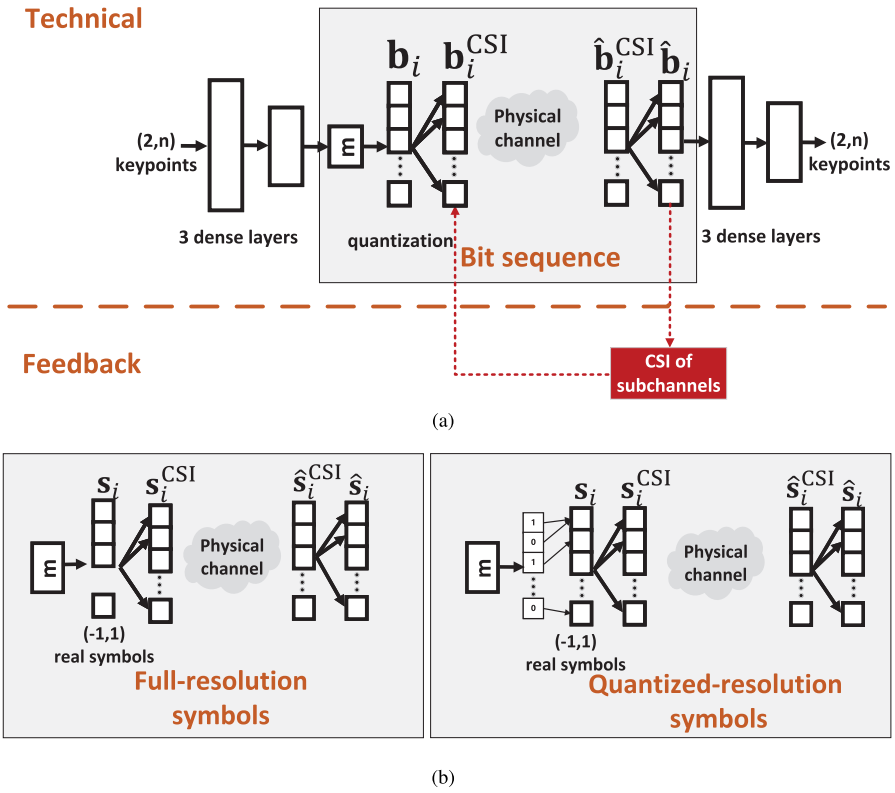


Fig. 3. (a) Structure of SVC with CSI feedback. (b) Two different modulation methods for SVC-CSI.

α and β, yielding

$$s_{i,j} = \alpha b_{i,2j-1} + \beta b_{i,2j}, j = 1, \ldots, m, \tag{20}$$

where the $2m$ bits in $\mathbf{b}$ are first modulated into the real symbols, $s_{i,j}$, which only have four possible values, that is, the constellation points only appear in 16 locations. The m-symbol vector $\mathbf{s}_i$ is also divided into K subchannels and multiplied with different transmit powers, $\boldsymbol{\rho} = [\rho_1, \ldots, \rho_K]$. Then, $\mathbf{s}_i$ is rearranged to $\mathbf{s}_i^{\text{CSI}}$ according to CSI feedback and sent to the receiver. The training of these two methods with constellation points is still the same as that of SVC. Especially, this method is called quantized-resolution constellation and only introduces $K + 2$ parameters, α, β, and $\boldsymbol{\rho}$.

In general, some bits/symbols always transmit under better channel conditions than others with CSI feedback. Thus, the networks can learn to transmit important keypoints at subchannels with high SNRs.

IV. Proposed Modules and Training Strategy

In this section, the experimental details of the proposed frameworks are discussed. At the beginning, the specific settings of the modules in this paper are provided, including the keypoint detector and the generator at the semantic level, and the encoder-decoder at the technical level. Then, the training strategies are described.

A. Module Details

The design of the modules is shown in Fig. 4 and their parameters are listed as follows:

1) Keypoint Detector: The keypoint detector consists of convolution neural networks (CNNs) similar to [28]. The inputted image matrix with sizes (256, 256, 3) is first downsampled to (64, 64, 3) by anti-alias interpolation to reduce complexity of the keypoint detector. Then, the image is processed by an hourglass network [29] with three blocks. Each block has a 3×3 convolution operation with a ReLU activation function, a batch normalization, and a 2×2 average pooling. The network has 1024 maximum channels and 32 output channels. After the hourglass network, a 7×7 convolution converts the output of CNN blocks from (64, 64, 32) into (64, 64, n), thereby dividing the image into n 64×64 grids. SoftMax activation function is applied in an 64×64 grid to generate an 64×64 matrix with the sum of 1, yielding

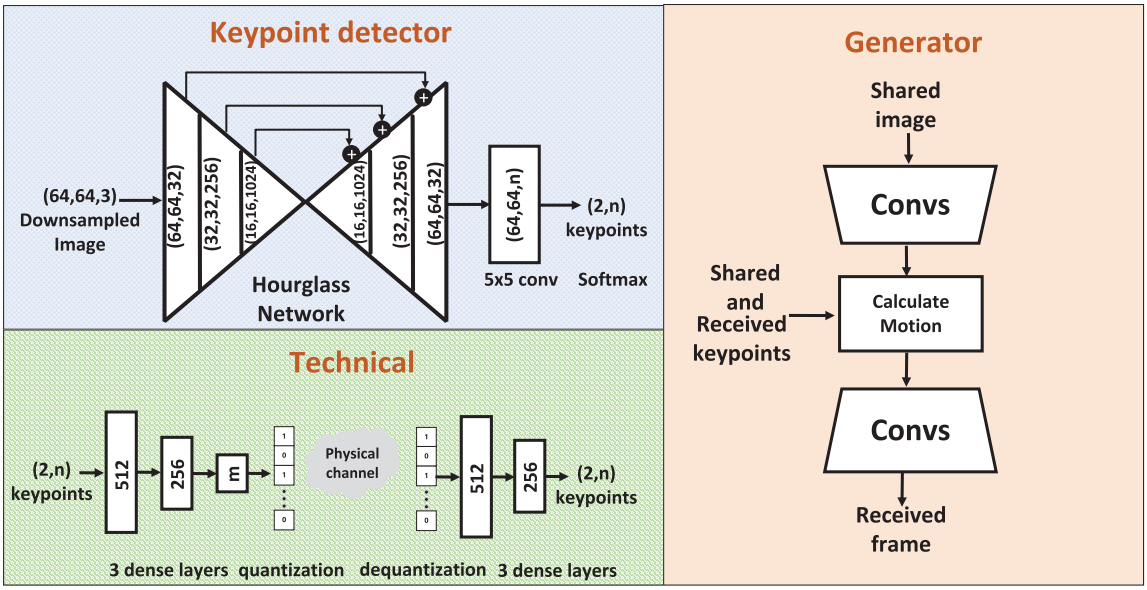
$$\mathbf{C} = \begin{pmatrix} c_{0,0} & \cdots & c_{0,63} \\ \vdots & \ddots & \vdots \\ c_{63,0} & \cdots & c_{63,63} \end{pmatrix}, \quad \sum_{j=0}^{63}\sum_{i=0}^{63} c_{i,j} = 1. \tag{21}$$

The keypoint coordinate generated by this 64×64 grid is

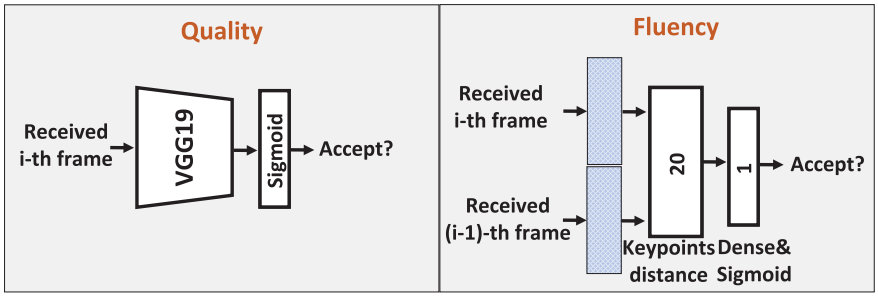
$$\sum_{i=0}^{63}\sum_{j=0}^{63} c_{i,j} \cdot (i, j). \tag{22}$$

Thus, n 64×64 grids generate n coordinates. Then, the values of the n coordinates in [0, 63] are normalized to $[-1, 1]$.

2) Encoder-Decoder: For the encoder, the n keypoints of the i-th frame, $\mathbf{k}_i$ (expressed in n coordinates), are considered $2n$ real numbers and processed by three dense layers, $f_{\text{en}}(\cdot)$, with 512, 256, and m neurons, where m is the number of the transmit symbols. The first two layers use ReLU activation



(a)



(b)

Fig. 4. (a) Structure of SVC-HARQ with ACK feedback. (b) Two potential methods for semantic error detection $Det(\cdot)$.

function, and the last one uses Sigmoid activation function. Then, a two-bit quantization $Q(\cdot)$ is applied to generate $2m$ transmitted bits, $\mathbf{b}_i$. All values before quantization are converted into $[0, 1]$ with Sigmoid activation function. Then, the 2-bit quantization uniformly divides $[0, 1]$ into four value spaces and encodes them into 00, 01, 11, and 10. The choice of a proper quantization method can also improve performance [30]. In this manuscript, the quantization choice is not concentrated on because it has no effect on expressing the advantages of the SVC methods. The three dense layers, $f_{\rm de}(\cdot; \mathbf{W}_{\rm de})$, have 512, 256, and $2n$ neurons, where the first two layers use ReLU activation function and the last one uses Tanh to restore n keypoints, $\hat{\mathbf{k}}_i$. Tanh activation converts the restored keypoints into $[-1, 1]$, which has the same value range as the output of the keypoint detector. The derivative of the quantization layer is replaced by that of the expectation in the backward pass because the gradient is truncated by the quantization [31].

3) Generator: The architecture of the generator is similar to that in [22] but without the Jacobian matrix.

4) Quality Detector: This quality assessment network can be obtained by transfer-learning a VGG-19 based classifier, as shown on the left of Fig. 4(b). The VGG-19 based quality detector consists of VGG-19 and one dense layer with Sigmoid activation function to output the frame quality indicator.

5) Fluency Detector: On the right of Fig. 4(b), the detector needs to distinguish inappropriate expressions. We use a keypoint detector to capture the keypoints after $\hat{\mathbf{p}}_{1,i}$. Then, we calculate the distance between the keypoints of $\hat{\mathbf{p}}_{1,i}$ and $\hat{\mathbf{p}}_{1,i-1}$. A large distance means that the expression has a sudden change and the transmitted keypoints have some errors. The whole detection process can be expressed as

$$Det_{\rm KD}(\hat{\mathbf{p}}_{1,i-1}, \hat{\mathbf{p}}_{1,i}) = f_{\rm Det}((KD(\hat{\mathbf{p}}_{1,i-1}) - KD(\hat{\mathbf{p}}_{1,i}))^2; \mathbf{W}_{\rm Det}), \quad (23)$$

where $f_{\rm Det}(\cdot)$ is a dense layer with one neuron output and Sigmoid activation function.

B. Losses and Metrics

The loss function consists of perceptual loss (Ploss) [32] based on a pretrained CNN, called VGG-19 [33], a patch-level discriminator loss [34], and an equivariance loss [22], denoted as $L_{\rm P}(\cdot)$, $L_{\rm D}(\cdot)$, and $L_{\rm E}(\cdot)$, respectively. These losses are briefly discussed below.

1) Ploss: The five outputs of convolution layers before the maxpooling in the VGG-19 are exploited. The process

from the input of VGG-19 to the j-th output is denoted as $f_{\mathrm{VGG},j}(\cdot)$. Thus, Ploss can be calculated by the mean-squared-error (MSE), yielding

$$L_{\mathrm{P}}(\mathbf{p}_i, \hat{\mathbf{p}}_i) = \sum_{j=1}^{5} \|f_{\mathrm{VGG},j}(\mathbf{p}_i) - f_{\mathrm{VGG},j}(\hat{\mathbf{p}}_i)\|^2. \quad (24)$$

2) Patch-Level Discriminator Loss: The patch-level discriminator in [34] is used and denoted as $D(\cdot)$. This loss function is calculated as

$$L_{\mathrm{D}}(\mathbf{p}_i, \hat{\mathbf{p}}_i) = \|1 - D(\hat{\mathbf{p}}_i)\|^2. \quad (25)$$

After a step to train the proposed network, the discriminator should update itself with a training step and the loss function for the discriminator can be expressed as

$$L_{\mathrm{D,update}}(\mathbf{p}_i, \hat{\mathbf{p}}_i) = \|1 - D(\mathbf{p}_i)\|^2 + \|D(\hat{\mathbf{p}}_i)\|^2. \quad (26)$$

3) Equivariance Loss: The MSE between the keypoints is extracted from the affine transformation of the true frame and the affine transformation of the received keypoints, yielding

$$L_{\mathrm{E}}(\mathbf{p}_i, \hat{\mathbf{k}}_i) = \|KD(\tau(\mathbf{p}_i)) - \tau(\hat{\mathbf{k}}_i)\|^2, \quad (27)$$

where $\tau()$ is an affine transformation introduced in [35].

As a result, the overall loss function is

$$L(\mathbf{p}_i, \hat{\mathbf{p}}_i) = L_{\mathrm{P}}(\mathbf{p}_i, \hat{\mathbf{p}}_i) + L_{\mathrm{D}}(\mathbf{p}_i, \hat{\mathbf{p}}_i) + L_{\mathrm{E}}(\mathbf{p}_i, \hat{\mathbf{k}}_i). \quad (28)$$

Given that the trainable parameters at the technical level are much fewer than other parts but still important, an MSE loss function is added to train the technical level, yielding

$$L_{\mathrm{MSE}} = \frac{\|\mathbf{k}_i - \hat{\mathbf{k}}_i\|^2}{2n}. \quad (29)$$

Metrics: Three metrics are used to evaluate the results:

1) Average Keypoint Distance (AKD): An pretrained facial landmark detector [36] is used to evaluate the errors in our transmission. This pretrained detector extracts keypoints from the received and transmitted frame, and their average distance is computed. The AKD metric represents the motion and the changing expression of the face.

2) Structural Similarity Index Measure (SSIM): SSIM evaluates the structural similarity among patches of the input images [37]. Therefore, SSIM, which is more robust than PSNR, is widely used as the metric of images.

3) Ploss: Ploss is commonly used as a regularization method when training a network in computer vision. Through calculating the sum of MSEs between the estimated and the true image at different layers of a pretrained network, such as VGG, the similarity of the features is represented by Ploss. Here, the Ploss metric proposed in [38] is selected.

C. Training Process

1) SVC: The training processes are divided into three steps. At the beginning, the technical level is ignored, and the parameters in the keypoint detector and the generator are trained by $L(\mathbf{p}_i, \hat{\mathbf{p}}_i)$, yielding

$$(\hat{\mathbf{W}}_{\mathrm{KD}}, \hat{\mathbf{W}}_{\mathrm{G}}) = \underset{\mathbf{W}_{\mathrm{KD}}, \mathbf{W}_{\mathrm{G}}}{\arg\min}\, L\left(\mathbf{p}_i, G(\mathbf{p}_0, \mathbf{k}_0, KD(\mathbf{p}_i; \mathbf{W}_{\mathrm{KD}}); \mathbf{W}_{\mathrm{G}})\right). \quad (30)$$

Then, the parameters at the technical level are trained by L_{MSE} to restore the $\mathbf{k}_i$ with some symmetric bits due to the effect of channels, yielding

$$(\hat{\mathbf{W}}_{\mathrm{en}}, \hat{\mathbf{W}}_{\mathrm{de}}) = \underset{\mathbf{W}_{\mathrm{en}}, \mathbf{W}_{\mathrm{de}}}{\arg\min}\, L_{\mathrm{MSE}}\left(\mathbf{k}_i, f_{\mathrm{de}}(Q^{-1}(Q(f_{\mathrm{en}}(KD(\mathbf{p}_i; \hat{\mathbf{W}}_{\mathrm{KD}}); \mathbf{W}_{\mathrm{en}})); \mathbf{W}_{\mathrm{de}})\right). \quad (31)$$

Finally, all trainable parameters of SVC are fine-tuned in the end-to-end manner as

$$(\hat{\mathbf{W}}_{\mathrm{KD}}, \hat{\mathbf{W}}_{\mathrm{en}}, \hat{\mathbf{W}}_{\mathrm{de}}, \hat{\mathbf{W}}_{\mathrm{G}}) = \underset{\mathbf{W}_{\mathrm{KD}}, \mathbf{W}_{\mathrm{en}}, \mathbf{W}_{\mathrm{de}}, \mathbf{W}_{\mathrm{G}}}{\arg\min}\, L\left(\mathbf{p}_i, \hat{\mathbf{p}}_i\right). \quad (32)$$

2) SVC-HARQ: The training of the first transmission is the same as SVC and its BER is set as 0 to focus on video compression. Then, all trained parameters are used as the initial values when training the parameters at the incremental transmission. Moreover, the trained parameters, $\hat{\mathbf{W}}_{1,\mathrm{KD}}$ and $\hat{\mathbf{W}}_{1,\mathrm{en}}$, in the first transmission are fixed, and the process can be written as

$$(\hat{\mathbf{W}}_{2,\mathrm{KD}}, \hat{\mathbf{W}}_{2,\mathrm{en}}, \hat{\mathbf{W}}_{2,\mathrm{de}}, \hat{\mathbf{W}}_{2,\mathrm{G}}) = \underset{\mathbf{W}_{2,\mathrm{KD}}, \mathbf{W}_{2,\mathrm{en}}, \mathbf{W}_{2,\mathrm{de}}, \mathbf{W}_{2,\mathrm{G}}}{\arg\min}\, L\left(\mathbf{p}_i, G(\mathbf{p}_0, \mathbf{k}_0, f_{\mathrm{de}}(Q^{-1}([\hat{\mathbf{b}}_{1,i}, \hat{\mathbf{b}}_{2,i}]); \mathbf{W}_{2,\mathrm{de}}); \mathbf{W}_{2,\mathrm{G}})\right), \quad (33)$$

where

$$\hat{\mathbf{b}}_{2,i} = h(Q(f_{\mathrm{en}}(KD(\mathbf{p}_i; \mathbf{W}_{2,\mathrm{KD}}); \mathbf{W}_{2,\mathrm{en}}))), \quad (34)$$

$$\hat{\mathbf{b}}_{1,i} = h(Q(f_{\mathrm{en}}(KD(\mathbf{p}_i; \hat{\mathbf{W}}_{1,\mathrm{KD}}); \hat{\mathbf{W}}_{1,\mathrm{en}}))), \quad (35)$$

and $h(\cdot)$ is a binary symmetric channel (BSC), and 5% transmitted bits are randomly errors. Thus, incremental transmission can learn to repair errors. More incremental transmissions can be added if necessary. For example, one more incremental transmission is trained as

$$(\hat{\mathbf{W}}_{3,\mathrm{KD}}, \hat{\mathbf{W}}_{3,\mathrm{en}}, \hat{\mathbf{W}}_{3,\mathrm{de}}, \hat{\mathbf{W}}_{3,\mathrm{G}}) = \underset{\mathbf{W}_{3,\mathrm{KD}}, \mathbf{W}_{3,\mathrm{en}}, \mathbf{W}_{3,\mathrm{de}}, \mathbf{W}_{3,\mathrm{G}}}{\arg\min}\, L\left(\mathbf{p}_i, G(\mathbf{p}_0, \mathbf{k}_0, f_{\mathrm{de}}(Q^{-1}([\hat{\mathbf{b}}_{1,i}, \hat{\mathbf{b}}_{2,i}, \hat{\mathbf{b}}_{3,i}]); \mathbf{W}_{3,\mathrm{de}}); \mathbf{W}_{3,\mathrm{G}})\right) \quad (36)$$

where the error bits of the BSC can be increased to 10% for a stronger correction capability.

3) Error Detector for SVC-HARQ: A total of 10,000 images received by SVC under different BERs, i.e., BER $\in [0, 0.2]$ are collected. The received image is labeled as 1 (acceptable) when its corresponding Ploss$<$0.2 and AKD$<$5. Otherwise, it is labeled as 0. These two performance thresholds are set because the received and the true frames cannot be distinguished subjectively if Ploss$<$0.2 and AKD$<$5.

With the collected training data and the labels, the VGG based quality estimator is trained directly with the cross-entropy loss function. With a trained detector, $Det_{\mathrm{VGG}}(\cdot)$, the ACK feedback of the j-th transmission can be expressed as

$$\mathrm{ACK} = \begin{cases} 1, & Det_{\mathrm{VGG}}(\hat{\mathbf{p}}_{j,i}) > 0.5, \\ 0, & Det_{\mathrm{VGG}}(\hat{\mathbf{p}}_{j,i}) \le 0.5. \end{cases} \quad (37)$$

However, the transmission errors may not be found by $Det_{\mathrm{VGG}}(\cdot)$ directly because the errors in keypoint transmission usually relate to expression rather than image quality. Thus, the fluency detector is trained for comparison. Training the keypoint based fluency detector for the i-th received frame also needs the $(i-1)$-th received frame with the "acceptable" quality. Thus, the $(i-1)$-th frame is transmitted under varying BERs and the received $(i-1)$-th frame with Ploss$<$0.2 and AKD$<$5 is used. The loss function is cross-entropy. After training, fluency detector $Det_{\mathrm{KD}}(\cdot)$ is used as

$$\mathrm{ACK} = \begin{cases} 1, & Det_{\mathrm{KD}}(\hat{\mathbf{p}}_{j,i-1}, \hat{\mathbf{p}}_{j,i}) > 0.5, \\ 0, & Det_{\mathrm{KD}}(\hat{\mathbf{p}}_{j,i-1}, \hat{\mathbf{p}}_{j,i}) \leq 0.5. \end{cases} . \tag{38}$$

The keypoint detector trained in $Det_{KD}(\cdot)$ only concentrates 10 keypoints, whereas the AKD network generates 63 keypoints for a face. We train $Det_{KD}(\cdot)$ by AKD rather than using AKD network directly because some keypoints for detailed expression changes cannot be found by users and can be ignored.

4) With CSI: The SVC-CSI has the same training strategy as the SVC because the other parts are the same as the SVC and the sort module of the SVC-CSI has no impact on the gradient.

V. Numerical Results

In this section, the numerical results of different frameworks are presented, and the pros and cons of SVC are discussed. Their bit consumption (required number of bits) are also compared with competing ones.

A. Configurations of Simulation System

1) Configurations: The VoxCeleb dataset [39] has considerable face videos of speakers. These videos are pre-processed into the size of 256×256, and those without a distinct face are removed. All the videos have only one speaker. After pre-processing, the training dataset has 2000 videos, and the testing dataset has 100 videos, which have about 500 different speakers. The number of extracted keypoints is set as $n = 10$. The encoded dimension is set as $m = 80$. After 2-bit quantization, the SVC encodes a video frame into 160 bits and these bits are converted into 40 subcarriers. Each OFDM symbol has 16 subcarriers. Thus, the video frame uses three OFDM symbols. The first OFDM symbol uses eight pilots inserted uniformly for channel estimation, and the eight remaining subcarriers are used for data transmissions. The two subsequent OFDM symbols have 32 data subcarriers. All networks are trained with Adam optimizer [40], and their initial learning rate is 0.0002.

2) Baseline: H264 is widely used as a commercial standard and usually occupies a 1/10 bandwidth of the original video. The constant rate factor (CRF) values of H264 can be set between 0 and 51, where the lower values result in better quality at the expense of larger file sizes. AV1 is a state-of-art video codec with higher complexity but higher compression ratio than H264. With the development of hardware acceleration, AV1 becomes more popular to reduce network traffic requirements. AV1 can also be adjusted by CRF. Reed Solomon (RS) code is commonly applied in storage and channel coding in the wicked environment, such as deep space communications. In the following, RS(p, q) means encoding p information symbols into q symbols. The redundancy $(q-p)$ symbols can correct $(q-p)/2$ error symbols. In this paper, the IR-HARQ strategy [41] encodes 64 information symbols into 255 symbols and initially transmits 127 symbols. The 128 remaining symbols are transmitted as incremental redundancy while the CRC detects errors after the first transmission. Apart from RS code, low-density parity check (LDPC) code is used as a state-of-art method.

TABLE I
The Performance Versus Bpp of the SVC and Conventional Methods

	SVC	H264D	H264	AV1D	AV1
Bpp	**0.0024**	0.0025	**0.0157**	0.0028	**0.0092**
SSIM↑	0.671	0.6091	0.803	0.750	0.867
Ploss↓	**0.186**	0.5738	**0.187**	0.340	**0.182**
AKD↓	3.12	45.4	2.50	5.14	2.42

B. Differences Between Conventional and Semantic-Based Methods

In TABLE I, SVC only transmits 160 bits per frame and is trained under no transmission errors. The shared photos have no need to be transmitted in real time given the fixed speakers at the beginning of the conference. Thus, the bits per pixel (bpp) of the shared photos is not considered when bpp is used to represent the requirement of the transmission resources in this paper. The CRF of H264 is set as 46. Thus, H264 with 0.0157 bpp (bold in the table) has a similar Ploss as SVC. Given that H264 cannot reach the bpp of SVC when its CRF is set to the maximum value, that is, 51, H264 with downsampling (H264D) is applied to demonstrate the superiority of SVC under the similar bpp. The downsampling rate (DR) of H264 is four, which means the 256×256 image is downsampled into 64×64. The CRF of AV1 is 57, and AV1D is the downsamples of AV1 with DR=2 and CRF=63.

The conventional AV1 always have a better SSIM performance than SVC. Then, the competing methods are compared in the metrics of Ploss and AKD, where facial features are more important than structural similarity. SVC has a tremendous superiority in bit consumption. The required bpp of H264 is about six times that of SVC when their Ploss performance is similar and more than four times those of SVC when their AKD performance is similar. The required bits of AV1 are less than H264 but still more than those of SVC. In general, these metrics only represent different perspectives in the evaluation of semantic transmissions and the transmitted results should be acceptable to humans. In the following section, the semantic errors are analyzed through examples.

Although H264 has the same Ploss performance as SVC, the content loss is different, as shown in the examples in Figs. 5(a)-(c). H264 loses the pixel information in all the areas of this frame. Thus, the frame shows a lower resolution than the original one. The lost information of SVC usually

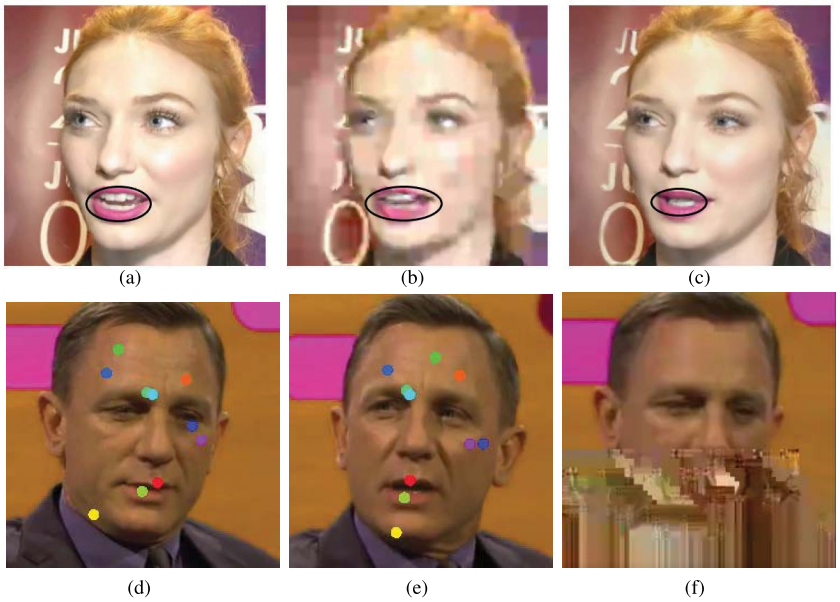


Fig. 5. Different content losses after conventional H264 and the SVC coding: (a) Original frame, (b) Frame coded by the H264 (Ploss=0.187), (c) Frame coded by the SVC. (Ploss=0.186); Error examples of the conventional and semantic methods under BER=0.05: (d) Transmit keypoints. (e) Received keypoints and restored frame. The AKD of the semantic method is about 9. (f) Received frame with H264+RS(64, 127). The keypoints cannot be detected to calculate AKD because the face is blurred.

cannot be distinguished as an independent image. Only the detailed expressions in the frame coded by SVC, such as the mouth in the circles, are different from the original because semantic information is ignored when coding. This phenomenon is demonstrated in the three metrics in TABLE I. Considered a lower-resolution image of the original frame, the structural information of the H264 frame is still reserved and the locations of the detected keypoints are unchanged. Thus, H264 has better SSIM and AKD metrics than SVC. However, the quality of the SVC frame seems higher than that of H264 to the human if the detailed expressions cause no ambiguities. Considering that SVC only requires 1/6 of bits of the H264, the semantic transmission is a better option for video conferencing, especially when some conferees are using mobile phones in the crowd.

The difference between the semantic and bit errors is also considered. Figs. 5(d)-(f) show the bit errors in H264+RS (64, 127) directly blur the frame and even cause the speaker to become unrecognizable, where no semantic errors are found in the same channel condition independently. These keypoints of the received frame in Fig. 5(e) are out of position. Apart from the loss of detailed expressions in Figs. 5(a)-(c), the transmission errors lead to the change in the expressions. Thus, an effective error detector should be proposed to guarantee the quality of SVC-HARQ. The frames coded by H264 are divided into two categories, namely, I and P frames. I frames transmit the entire pixel information, and P frames only transmit the changed part with the previous I frame as a reference. This frame is chosen from the I frames of H264. Thus, the same errors in the I frame happen in the following P frames if the P frames can be decoded normally. Moreover, the errors in the P frames make the frame become more blurred, indistinct, and even black. Figs. 5(d) and (e) also show the correlation of the coordinates. For example, the three points around the mouth and neck are always at the bottom of the other points.

Errors in SVC are difficult to find independently. Thus, the VGG-based quality detector always achieves an accepted ratio higher than 96%. Therefore, SVC can preserve visual quality even under a wicked environment because the main appearance features are shared in advance. Thus, the VGG-based quality detector is insufficiently effective as a semantic detector. Apart from the quality of the received frame, the received video should be fluent. The performance of the current frame cannot be obtained because the true current frame is unknown to the receiver in practice. Thus, the keypoints are detected again at the receiver by the trained keypoint detector, and the average distances of the detected keypoints between the adjacent frames, called detected AKD, are related to video fluency. Fig. 6(b) shows the detected AKDs of most frames are lower than 0.05 when no bit error exists and increase with BER. Compared with the 32-bit CRC code used in the conventional HARQ systems, the fluency detector helps guarantee the quality of the video without any extra parity code.

In general, SVC can save transmit resources for a high resolution video conferencing because it only transmits keypoints and has no need to compress pixel information such as H264. Moreover, SVC is superior under an extremely high BER. However, the training BER affects the performance of the SVC, similar to selecting the code rate of channel coding. Thus, the IR-HARQ frame of SVC is proposed and tested in the following section.

C. Performance of SVC-HARQ

The HARQ frameworks are tested under the AWGN channel shown in Figs. 7(a) and (b). The comparison uses BER rather

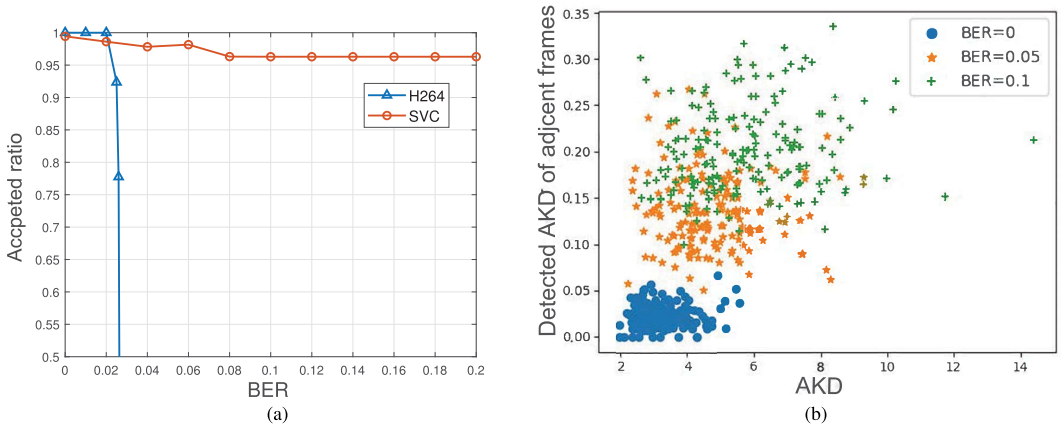


Fig. 6. (a) Accepted ratio of received frames using VGG-based detector under different BERs. (b) Detected AKD of adjacent frames and their AKD performances under different BERs.

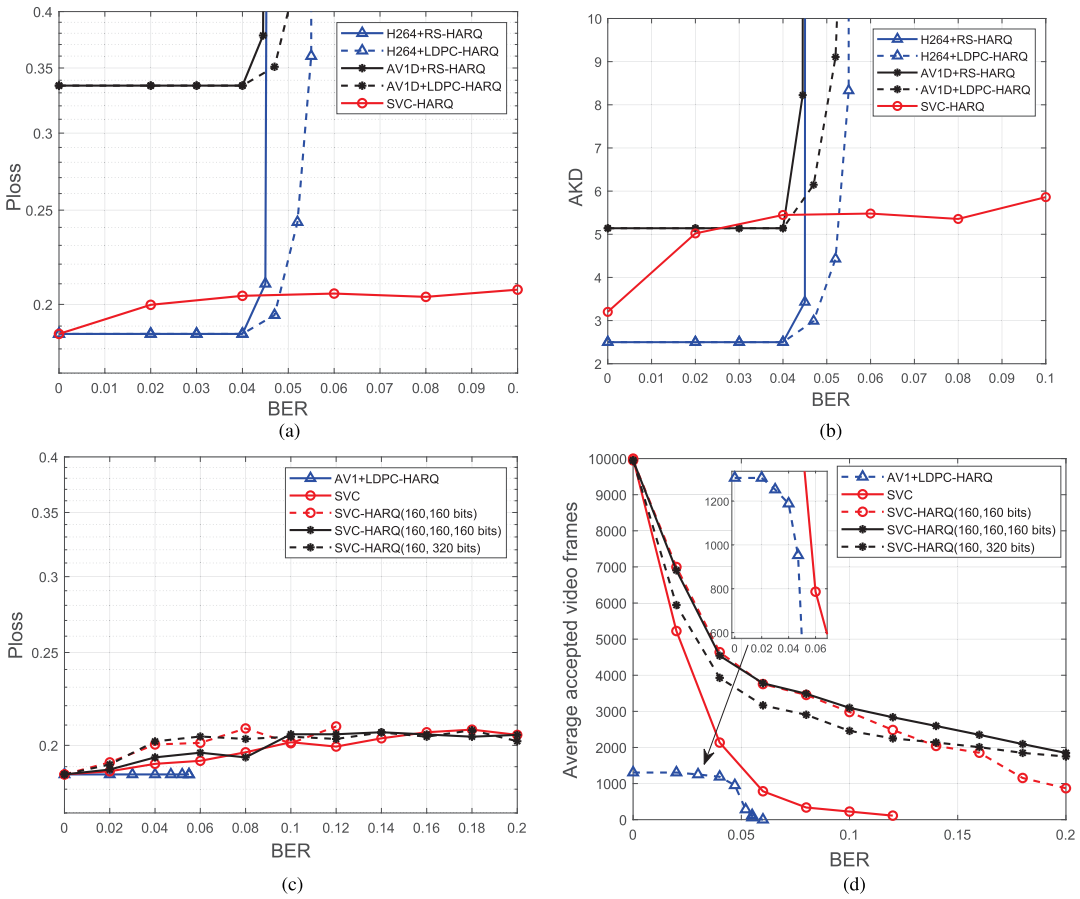


Fig. 7. (a) Ploss performance of SVC-HARQ and competing methods. (b) AKD performance of SVC-HARQ and competing methods. (c) (d) Ploss performance and throughput of conventional method and different settings of SVC-based methods. Throughput is the number of received video frames with acceptable quality under 1.526M transmit bits, where 1.526M bits can transmit 10,000 video frames encoded by SVC when BER=0.

than SNR because the noiseless performances demonstrate the capability of these source coding methods. The average performance of all received frames and the error correction ability of these methods are exhibited. The competing method, H264+RS-HARQ, first transmits 127 bits per 64 information bits. Then, the 128 incremental bits are transmitted if the CRC detector finds errors. However, the incremental symbols cannot repair all error bits when BER>0.04. Compared with H264+RS-HARQ, H264+LDPC-HARQ can correct more errors when BER>0.04. The tendency of AV1D-based methods is similar to H264 based methods, whereas AV1D has worse Ploss and AKD performances than H264. Overall, H264+RS-HARQ requires about 12 times bpp but performs worse than SVC-HARQ when BER is high. AV1D has the same compression ratio as SVC, but its performance is far worse than SVC.

The received frames are rejected by the error detector if their quality remains unacceptable after all the incremental transmissions, which means this transmission is failed and the entire transmission should be restarted. Fig. 7(d) shows the

average number of accepted frames per bit is calculated after a transmission including the first transmission and the following incremental transmissions. The accepted frames reflect the throughput of the competing methods. AV1 has a similar Ploss performance similar to SVC but requires about four times bpp, which brings the throughput gap between AV1+LDPC-HARQ and SVC-based methods when BER is close to 0. The CRC error detector can ensure that all accepted frames have no transmission error. In Fig. 7(c), the Ploss performance of AV1+LDPC-HARQ is unchanged until no received frames can be accepted by CRC (throughput=0). The proposed fluency detector can also reject all the received frames of SVC when BER>0.12, which demonstrates the effectiveness of the fluency detector. For the accepted frames, the fluency also ensures that these frames are not worse than the training threshold (Ploss<0.2). However, some frames whose Ploss performance is slightly higher than 0.2, cannot be found by the fluency detector. Sometimes, their Ploss performance may reach 0.21.

Three different settings of SVC-HARQ are also tested in Figs. 7(c) and (d), which provides an insight into designing an adaptive SVC-HARQ. SVC-HARQ (160, 160 bits) means SVC-HARQ has the first transmission with 160 bits and one incremental transmission with 160 bits, which is the default SVC-HARQ mentioned above. Then, an extra incremental transmission is added for comparison called SVC-HARQ (160, 160 bits) or the incremental transmission has more bits called SVC-HARQ (160, 320 bits). All three methods require 160 bits for the first transmission and their throughputs are equal when BER is close to 0. However, the throughput of SVC-HARQ (160, 320 bits) is slightly worse than that of SVC-HARQ (160, 160, 160 bits) when BER$\in [0, 0.2]$ because this method directly transmits extra 320 bits when the first transmission has errors, whereas SVC-HARQ (160, 160, 160 bits) has two 160-bit incremental transmissions. SVC-HARQ (160, 160, 160 bits) and SVC-HARQ (160, 320 bits) have similar throughput when BER=0.2 because they have the same maximum code length and all incremental bits are transmitted when BER=0.2. SVC-HARQ (160, 160 bits) has fewer incremental bits and less throughput than the other methods when BER is high. However, SVC-HARQ (160, 160 bits) and SVC-HARQ (160, 160, 160 bits) have the same throughputs when BER<0.08. Therefore, the last incremental transmission of SVC-HARQ (160, 160 bits) has no effect when BER is low because only one 160-bit incremental transmission is enough.

The absolute value changes of Ploss are small because it represents the whole content and expression errors only occupy a small part of the image. For example, the Ploss of SVC-HARQ is about 0.18 when BER is close to 0, and about 0.23 when BER=0.2. AKD only concentrates on the facial expression, and its change is large. Especially, the tendencies of Ploss and AKD metrics are similar. Therefore, inaccurate facial expression is the major factor for lower Ploss due to higher BER.

According to the above discussion, SVC-HARQ shows its flexibility in bit consumption with the change in BER. SVC-HARQ can reach the best performance of the SVCs trained under different BERs as an adaptive method. However, the semantic detector can only protect the fluency of the video. A detector to find expression errors should be proposed for an important conference.

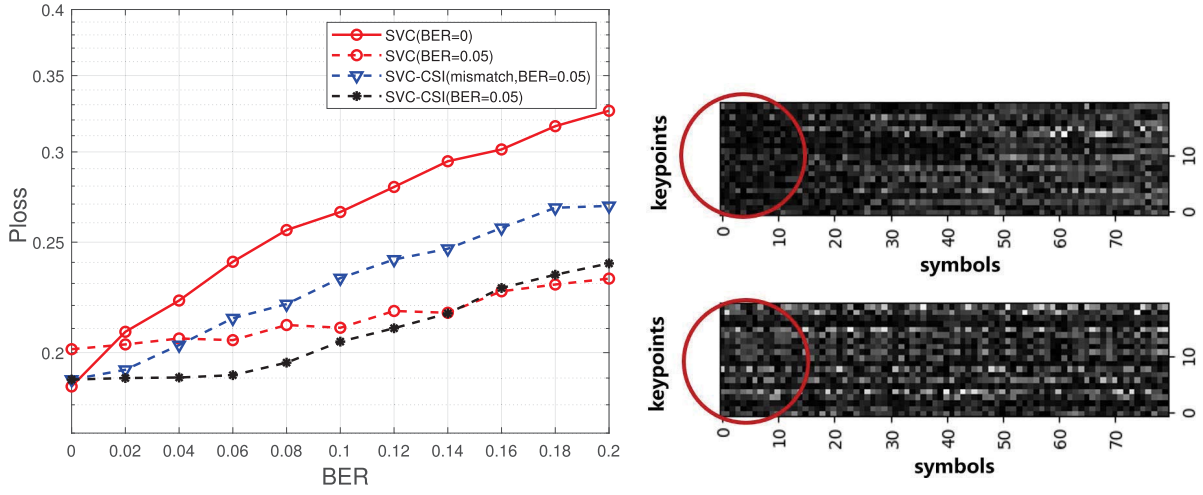
D. Performance of SVC-CSI

Frequency-selective channels are considered to train and test the effectiveness of CSI feedback. All the SVCs with CSI feedback are trained under channels with exponential power delay profile of three paths. Each path obeys complex Gaussian distribution. Given that channel model is introduced when training, untrained channels with five paths are also simulated test the robustness of the SVC. The testing environment is called mismatched channel environments because their statistical parameters, such as delay spread, are different from the trained channels. The number of subchannels in the frequency domain is 16. The transmit bits are modulated to 16-QAM.

Fig. 8(a) shows SVC-CSI (BER=0.05) means the average BER of the training channels is 0.05. SVC-CSI (BER=0.05) has a Ploss performance similar to SVC (BER=0.05) when BER>0.14 and has a better performance than SVC (BER=0.05) when BER<0.14. It reaches the performance of SVC (BER=0) when BER=0 because SVC-CSI learns to protect the information according to the qualities of the channels. However, the performance of SVC-CSI decreases more sharply under mismatched channel environments. Thus, SVC-CSI (mismatch) performs worse than SVC (BER=0.05) when BER>0.04. Overall, CSI feedback enhances the performance when BER is low under matched channel environments but loses its robustness under mismatched environments.

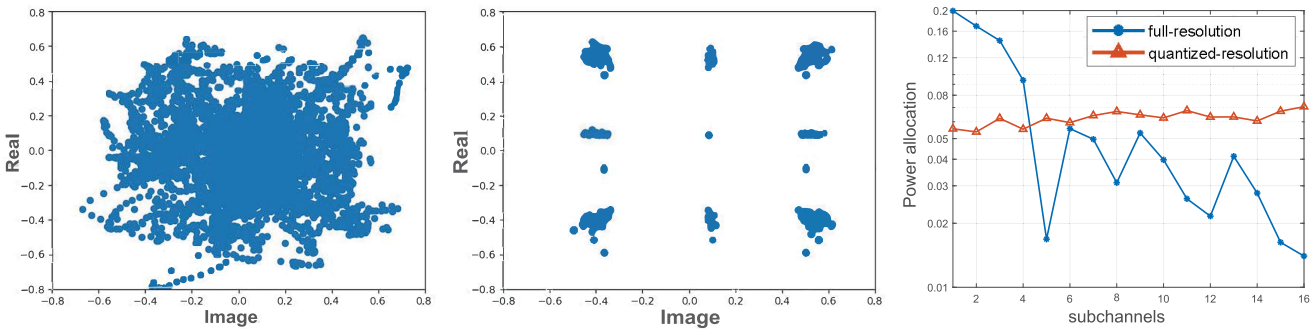
To visualize the influence of CSI feedback, the three dense layers at the transmitter at the technical level are replaced with one dense layer, whose input includes keypoints (20 real numbers) and output includes 80 symbols (160 bits with 2-bit quantization). In Fig. 8(b), the absolute values of trained multiplicative weights in this dense layer are shown as a gray picture, and only the weights on the right picture are trained with CSI feedback. Thus, the $0-10$ symbols on the right picture are usually transmitted at better channels than those on the left one due to CSI feedback. The absolute values in the circle of the right picture are larger than those of the left picture. This finding means that the transmitter learns to place more information at better channel conditions. SVC-CSI performs better than SVC when BER is higher because most information is transmitted at the first several channels with lower noise power.

The keypoints are also transmitted as symbols directly under Rayleigh fading channels in Fig. 9. SVC-CSI (full-resolution) learns to transmit 160 real symbols. Owing to CSI feedback with 16 subchannels, the first five symbols are always transmitted under the highest SNR and the last five symbols are under the lowest SNR. The constellation points of SVC-CSI (full-resolution) are spread in Fig. 9(a). This method learns to transmit more information at better subchannels. Thus, transmit power decreases when channel condition worsens, as shown in Fig. 9(c). SVC-CSI (quantized-resolution) has the same modulation method at all subchannels, and its constellation



(a) (b)

Fig. 8. (a) Ploss performance of the SVC-CSI methods. (b) Gray image of the trained weights to visualize the effects of CSI feedback. The left image is trained without CSI information, whereas the right one is trained under CSI feedback.



(a) (b) (c)

Fig. 9. Different constellations of the two SVC methods: (a) Constellation points of SVC-CSI (full-resolution). (b) Constellation points of SVC-CSI (quantized-resolution). (c) Different transmit powers allocated at different subchannels. The channel condition worsens with the increase of the channel number.

TABLE II

PLOSS PERFORMANCE OF SVC-CSI WITH DIFFERENT MODULATION METHODS AND SVC-CSI-HARQ UNDER MATCHED AND MISMATCHED CHANNELS

SNR (dB)	0	4	8	12	16	20
SVC-CSI (16-QAM)	0.2356	0.2273	0.1994	0.1901	0.1895	0.1894
SVC-CSI (quantized-resolution)	0.2100	0.2039	0.2000	0.1990	0.1982	0.1980
SVC-CSI (full-resolution)	0.1949	0.1899	0.1876	0.1862	0.1861	0.1860
SVC-HARQ	0.2300	0.2189	0.2054	0.2014	0.1881	0.1865
SVC-CSI-HARQ	0.2293	0.2188	0.2002	0.1901	0.1895	0.1894
SVC-CSI-HARQ(mismatch)	0.2297	0.2189	0.2050	0.1968	0.1899	0.1894

points in Fig. 9(b) are similar to 16-QAM. To cope with noise, the transmit power of SVC-CSI (quantized-resolution) increases as channel condition worsens.

The Ploss performance of SVC-CSI under fading channels is compared. SVC-CSI (16-QAM) is trained to transmit 320 bits under SNR=10 dB and modulated to 80 16-QAM symbols. The three methods in TABLE II have the same transmit resources. SVC-CSI (full-resolution) always has best performance, but its complexity is impractical. SVC-CSI (quantized-resolution) learns a different modulation method from 16-QAM. This method performs worse than 16-QAM when SNR$\geq$8 dB but better than 16-QAM when SNR is low. Therefore, the trained modulation of SVC-CSI (quantized-resolution) is suitable for wicked environments but cannot perfectly reconstruct the frame when SNR is high.

The introduction of HARQ can improve the performance of SVC-CSI under mismatched channels, and this method is called SVC-CSI-HARQ. These methods are tested under fading channels with 16-QAM modulation. In this method, the second transmission is trained without CSI feedback and

TABLE III
COMPLEXITY OF CONVENTIONAL AND SVC BASED METHODS

BER	Parameters	Memory	Fps
SVC	91 M	751Mbytes	51.9
SVC-CSI-HARQ	205 M	1432 Mbytes	31.2
H264	/	/	564
AV1	/	/	93

* The above experiments are based on two cores of E5-269 (CPU) and one Tesla V100 (GPU).
* The runtime of the keypoint detector is 6.3 ms and the generator is 9.9 ms.

is robust to varying environments. This strategy can guarantee that SVC-CSI-HARQ under mismatched environment is not worse than SVC-HARQ. SVC-CSI-HARQ performs better than SVC-HARQ under mismatched channels when SNR is between 8 and 12 dB because mismatched power correlation at subchannels is slight when subchannel gain is large. In addition, SVC-CSI-HARQ shows its superiority under matched channels. Especially, SVC-HARQ is slightly better than the methods with CSI feedback when SNR$\geq$16 because CSI feedback is ineffective under approximately noiseless channels. By contrast, CSI feedback may mislead SVC to transmit less information at the last subchannels.

E. Complexity Analysis

A brief complexity analysis is given in TABLE III. The frame rates of the proposed methods are sufficient for video conferencing without any acceleration design. SVC-CSI-HARQ needs more than twice of parameters and memory of SVC because SVC-CSI-HARQ has one more incremental transmission link than SVC and a fluency detector. The runtimes of the two most important modules, the keypoint detector and the generator, play a major role from encoding a new frame to receiving by users. The higher resolution does not increase the complexity of the transmitter because the transmitter only deals with keypoint from a downsampled 64×64 frame. The user can achieve a high-resolution video with the equipment of a powerful terminal.

The frame rates of H264 and AV1 are shown for reference under the default settings of Python+ffmpeg. With the development of hardware acceleration techniques, AV1 is becoming more competitive than H264 for high-resolution videos. Similarly, the DL-based methods can be compressed and accelerated. In [18], model compression is studied and the performance loss is small after 90% of the parameters are reduced. Thus, the application of the proposed methods will be feasible after model compression and hardware acceleration.

VI. CONCLUSION

Semantic transmission framework for video conferencing is investigated. The knowledge of the speaker's photos can be shared explicitly because the faces play an essential role at a conference. The three-level framework, called SVC, is established by utilizing only keypoint transmission to represent the motion of the facial expressions. Compression by SVC only loses detailed expressions, whereas the conventional methods reduce the resolution. Furthermore, transmission errors in the conventional methods directly destroy pixels, whereas those in SVC leads to a changed expression.

The effect of feedback in SVC is also considered, and an IR-HARQ framework called SVC-HARQ with ACK feedback is designed. The changed expression of error keypoints obtains nonsmooth adjacent frames. To detect semantic errors, a semantic detectoris developed. SVC-HARQ is flexible, and it can combine the performance of networks trained under different BERs, and always achieves a good performance.

CSI feedback can also further enhance the performance. Transmitted symbols or bits are sorted by SNRs at different subchannels, called SVC-CSI. SVC-CSI learns to allocate more information at subchannels with higher gains and performs better than SVC without CSI feedback. However, the robustness of SVC-CSI decreases because the channel model is exploited when training. The combination of CSI and ACK feedback can balance performance, bit consumption, and robustness.

REFERENCES

[1] J. Bao et al., "Towards a theory of semantic communication," in *Proc. IEEE Netw. Sci. Workshop*, Jun. 2011, pp. 110–117.
[2] Z. Qin, X. Tao, J. Lu, W. Tong, and G. Y. Li, "Semantic communications: Principles and challenges," 2021, *arXiv:2201.01389*.
[3] K. Lu et al., "Rethinking modern communication from semantic coding to semantic communication," 2021, *arXiv:2110.08496*.
[4] M. Sana and E. C. Strinati, "Learning semantics: An opportunity for effective 6G communications," in *Proc. IEEE 19th Annu. Consum. Commun. Netw. Conf. (CCNC)*, Jan. 2022, pp. 631–636.
[5] H. Ye, G. Y. Li, and B.-H. Juang, "Power of deep learning for channel estimation and signal detection in OFDM systems," *IEEE Wireless Commun. Lett.*, vol. 7, no. 1, pp. 114–117, Feb. 2018.
[6] H. Ye, L. Liang, G. Y. Li, and B.-H. F. Juang, "Deep learning-based end-to-end wireless communication systems with conditional GANs as unknown channels," *IEEE Trans. Wireless Commun.*, vol. 19, no. 5, pp. 3133–3143, May 2020.
[7] T. O'Shea and J. Hoydis, "An introduction to deep learning for the physical layer," *IEEE Trans. Cogn. Commun. Netw.*, vol. 3, no. 4, pp. 563–575, Dec. 2017.
[8] Z. Qin, H. Ye, G. Y. Li, and B.-H.-F. Juang, "Deep learning in physical layer communications," *IEEE Wireless Commun.*, vol. 26, no. 2, pp. 93–99, Apr. 2019.
[9] H. He, S. Jin, C. Wen, F. Gao, G. Y. Li, and Z. Xu, "Model-driven deep learning for physical layer communications," *IEEE Wireless Commun.*, vol. 26, no. 5, pp. 77–83, Oct. 2019.
[10] O. Y. Bursalioglu, G. Caire, and D. Divsalar, "Joint source-channel coding for deep-space image transmission using rateless codes," *IEEE Trans. Commun.*, vol. 61, no. 8, pp. 3448–3461, Aug. 2013.
[11] D. B. Kurka and D. Gunduz, "DeepJSCC-f: Deep joint source-channel coding of images with feedback," *IEEE J. Sel. Areas Inf. Theory*, vol. 1, no. 1, pp. 178–193, Dec. 2020.
[12] E. Bourtsoulatze, D. B. Kurka, and D. Gunduz, "Deep joint source-channel coding for wireless image transmission," *IEEE Trans. Cognit. Commun. Netw.*, vol. 5, no. 3, pp. 567–579, Sep. 2019.
[13] C. Lee, J. Lin, P. Chen, and Y. Chang, "Deep learning-constructed joint transmission-recognition for Internet of Things," *IEEE Access*, vol. 7, pp. 76547–76561, 2019.
[14] F. Zhai, Y. Eisenberg, and A. K. Katsaggelos, "Joint source-channel coding for video communications," in *Handbook of Image and Video Processing*, A. Bovik, Ed., 2nd ed. Burlington, MA, USA: Academic, 2005.
[15] Z. Weng and Z. Qin, "Semantic communication systems for speech transmission," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 8, pp. 2434–2444, Aug. 2021.
[16] N. Farsad, M. Rao, and A. Goldsmith, "Deep learning for joint source-channel coding of text," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 2326–2330.

[17] H. Xie, Z. Qin, G. Y. Li, and B.-H. Juang, "Deep learning enabled semantic communication systems," *IEEE Trans. Signal Process.*, vol. 69, pp. 2663–2675, 2021.
[18] H. Xie and Z. Qin, "A lite distributed semantic communication system for Internet of Things," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 1, pp. 142–153, Jan. 2021.
[19] X. Kang, B. Song, J. Guo, Z. Qin, and F. Richard Yu, "Task-oriented image transmission for scene classification in unmanned aerial systems," 2021, *arXiv:2112.10948*.
[20] D. Ding, Z. Ma, D. Chen, Q. Chen, Z. Liu, and F. Zhu, "Advances in video compression system using deep neural network: A review and case studies," *Proc. IEEE*, vol. 109, no. 9, pp. 1494–1520, Sep. 2021.
[21] P. Tandon et al., "Txt2Vid: Ultra-low bitrate compression of talking-head videos via text," 2021, *arXiv:2106.14014*.
[22] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe, "First order motion model for image animation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 7137–7147.
[23] T.-C. Wang, A. Mallya, and M.-Y. Liu, "One-shot free-view neural talking-head synthesis for video conferencing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 10039–10049.
[24] Y. Yang et al., "Semantic communications with AI tasks," 2021, *arXiv:2109.14170*.
[25] M. L. B. Riediger and P. K. M. Ho, "Application of Reed–Solomon codes with erasure decoding to type-II hybrid ARQ transmission," in *Proc. IEEE Global Telecommun. Conf.*, vol. 1, Dec. 2003, pp. 55–59.
[26] J. Abot, C. Olivier, C. Perrine, and Y. Pousset, "A link adaptation scheme optimized for wireless JPEG 2000 transmission over realistic MIMO systems," *Signal Process., Image Commun.*, vol. 27, no. 10, pp. 1066–1078, Nov. 2012.
[27] W. Zhang, K. Ma, J. Yan, D. Deng, and Z. Wang, "Blind image quality assessment using a deep bilinear convolutional neural network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 1, pp. 36–47, Jan. 2020.
[28] A. Siarohin, S. Lathuiliere, S. Tulyakov, E. Ricci, and N. Sebe, "Animating arbitrary objects via deep motion transfer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2377–2386.
[29] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 483–499.
[30] B. Sun, H. Feng, K. Chen, and X. Zhu, "A deep learning framework of quantized compressed sensing for wireless neural recording," *IEEE Access*, vol. 4, pp. 5169–5178, 2016.
[31] L. Theis, W. Shi, A. Cunningham, and F. Huszár, "Lossy image compression with compressive autoencoders," 2017, *arXiv:1703.00395*.
[32] J. Johnson, A. Alahi, and F.-F. Li, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 694–711.
[33] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[34] T.-C. Wang, M.-Y. Liu, A. Tao, G. Liu, J. Kautz, and B. Catanzaro, "Few-shot video-to-video synthesis," 2019, *arXiv:1910.12713*.
[35] Y. Zhang, Y. Guo, Y. Jin, Y. Luo, Z. He, and H. Lee, "Unsupervised discovery of object landmarks as structural representations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2694–2703.
[36] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2D & 3D face alignment problem? (And a dataset of 230,000 3D facial landmarks)," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1–10.
[37] A. Hore and D. Ziou, "Image quality metrics: PSNR vs. SSIM," in *Proc. 20th Int. Conf. Pattern Recognit.*, Aug. 2010, pp. 2366–2369.
[38] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 586–595.
[39] A. Nagrani, J. Son Chung, and A. Zisserman, "VoxCeleb: A large-scale speaker identification dataset," 2017, *arXiv:1706.08612*.
[40] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[41] S. B. Wicker and M. J. Bartz, "Type-II hybrid-ARQ protocols using punctured MDS codes," *IEEE Trans. Commun.*, vol. 42, no. 234, pp. 1431–1440, Feb. 1994.

Peiwen Jiang (Graduate Student Member, IEEE) received the B.S. degree from Southeast University, Nanjing, China, in 2019, where he is currently pursuing the Ph.D. degree in information and communications engineering. His research interests include deep learning-based channel estimation, signal detection, and semantic transmission in communications.

Chao-Kai Wen (Senior Member, IEEE) received the Ph.D. degree from the Institute of Communications Engineering, National Tsing Hua University, Taiwan, in 2004. He was with the Industrial Technology Research Institute, Hsinchu, Taiwan, and MediaTek Inc., Hsinchu, from 2004 to 2009, where he was engaged in broadband digital transceiver design. In 2009, he joined the Institute of Communications Engineering, National Sun Yat-sen University, Kaohsiung, Taiwan, where he is currently a Professor. His research interests include optimization of wireless multimedia networks.

Shi Jin (Senior Member, IEEE) received the B.S. degree in communications engineering from the Guilin University of Electronic Technology, Guilin, China, in 1996, the M.S. degree from the Nanjing University of Posts and Telecommunications, Nanjing, China, in 2003, and the Ph.D. degree in information and communications engineering from Southeast University, Nanjing, in 2007. From June 2007 to October 2009, he was a Research Fellow with the Adastral Park Research Campus, University College London, London, U.K. He is currently with the Faculty of the National Mobile Communications Research Laboratory, Southeast University. His research interests include space time wireless communications, random matrix theory, and information theory. He and his coauthors have been awarded the 2011 IEEE Communications Society Stephen O. Rice Prize Paper Award in the field of communication theory and the 2010 Young Author Best Paper Award by the IEEE Signal Processing Society. He served as an Associate Editor for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, IEEE COMMUNICATIONS LETTERS, and *IET Communications*.

Geoffrey Ye Li (Fellow, IEEE) is currently a Chair Professor at Imperial College London, U.K. Before moving to Imperial in 2020, he was a Professor at the Georgia Institute of Technology, USA, for 20 years and a Principal Technical Staff Member with AT&T Labs-Research, Middletown, NJ, USA, for five years. His research interests include statistical signal processing and machine learning for wireless communications. In the related areas, he has published over 600 journals and conference papers in addition to over 40 granted patents and several books. His publications have been cited over 55,000 times with an H-index over 110 and he has been recognized as a Highly Cited Researcher, by Thomson Reuters, almost every year. He was awarded IEEE Fellow and IET Fellow for his contributions to signal processing for wireless communications. He won several prestigious awards from IEEE Signal Processing, Vehicular Technology, and Communications Societies, including IEEE ComSoc Edwin Howard Armstrong Achievement Award in 2019. He also received the 2015 Distinguished ECE Faculty Achievement Award from Georgia Tech. He has organized and chaired many international conferences, including the Technical Program Vice-Chair of the IEEE ICC 2003, the General Co-Chair of the IEEE GlobalSIP 2014, the IEEE VTC 2019 Fall, the IEEE SPAWC 2020, and IEEE VTC 2022 Fall. He has been involved in editorial activities for over 20 technical journals, including the Founding Editor-in-Chief of the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS Special Series on ML in Communications and Networking.