Traffic Reduction in Video Call and Chat using DNN-based Image Reconstruction

Shota Watanabe*, Takuya Fujihashi†, Shunsuke Saruwatari* and Takashi Watanabe*
*Graduate School of Information Science and Technology, Osaka University, Japan
†Graduate School of Science and Engineering, Ehime University, Japan

Abstract—**In this paper, a traffic reduction scheme for video-based call and chat applications that uses deep neural network (DNN) based super resolution is proposed. Specifically, a sender transmits low-quality, low-resolution video frames containing face information in order to reduce the amount of video traffic. The receiver uses DNN-based super resolution to reconstruct high-quality, high-resolution video frames from the low-quality video frames. The proposed scheme makes two contributions. First, face features are adopted for parameter optimization of DNN-based super resolution for high-quality image reconstruction. Second, the scheme includes a newly designed loss function that considers face features that allows high-quality face-containing video frames to be reconstructed at the receiver. According to our evaluation results using real video frames of video calls, the proposed scheme reduces the amount of video traffic by more than 90% as compared with the conventional schemes that implement the standard video encoder. In this case, the proposed scheme achieves a reconstructed image quality up to 0.85 in terms of structural similarity (SSIM).**

Index Terms—**Deep Neural Network, Video Call, Video Chat, Super-Resolution**

I. Introduction

As the number of mobile users increases, the number of users of social networking service (SNSs), e.g., Facebook and Instagram, communicating with on-line/off-line friends is also increasing. In such services, the user frequently captures his/her face using a mobile device's camera and shares the captured video with users so that they can enjoy video calls and chat with friends [1]. Face-containing video contents are also applied in many applications for teleconference and customer support to allow smooth communications with remote users [2].

A major issue pertaining to applications that use face-containing video contents is the large amount of video traffic that they generate as compared with audio- and text-based applications. A large amount of video traffic causes low video quality in band-limited links/networks. To reduce video traffic, video compression techniques, such as Advanced Video Coding (H.264/AVC) and High Efficiency Video Coding (H.265/HEVC), have been proposed for general video applications [3], [4]. These techniques use motion estimation, quantization, and entropy coding to attain better compression gains; however, the problems caused by the large amounts of video traffic in video-based call and chat applications remain. To achieve significant traffic reduction in such applications, an application-specific method is required that considers the video features in video call and chat applications: many video frames captured in the applications may contain the sender's face.

For this purpose, a novel transmission scheme is proposed for video call and chat applications, motivated by deep neural network (DNN)-based super resolution, i.e., deep convolutional generative adversarial networks (DCGANs) [5]. DCGANs reconstruct high-quality and high-resolution face-containing video frames from low-quality and low-resolution face-containing video frames by using a pre-trained generation model. The key idea of our scheme is that a sender transmits only low resolution and quality face-containing video frames over networks and the receiver uses a DCGAN to reconstruct high-resolution, high-quality face-containing video frames. Since the quality of the reconstructed video frames depends on the parameters of the generation model, the parameters must be optimized using the sender's face-containing video frames and a loss function before video transmissions. The optimized parameters are then shared with the receiver. The evaluation results show that our scheme achieves an approximately 90% traffic reduction as compared with the existing scheme, which uses the conventional video compression. In this case, the proposed scheme provides a clean face image after the image reconstruction with a structural similarity (SSIM) of 0.85.

The contributions of the proposed scheme are three fold: 1) it uses the face features in addition to the face-containing video frames for parameter optimization of the generation model to achieve high-quality reconstruction, 2) it includes a new loss function designed such that the face features are considered in order to enhance the image reconstruction quality at the receiver, and 3) it both realizes significant traffic reduction and maintains visual quality for applications using face-containing video contents, such as video call and chat.

II. Related Research

Our study is related to studies on super resolution-based image reconstruction, DNN/generative adversarial network (GAN) based image reconstruction, and DNN-based video delivery.

A. Super Resolution-based Image Reconstruction

In super resolution techniques, low-resolution images are restored to high-resolution images. There are two types of super resolution techniques: single-image and multiple-image. For example, A+ [6] and RAISR [7] have been proposed for single-image super resolution, while [8], [9] proposed super resolution for multiple images. In addition, the scheme in [10], [11] realizes high-quality super resolution for multiple images by using recurrent convolutional networks.

In our study, we used single-image-oriented super resolution for face-containing video frame reconstruction in video call and chat applications. For high-quality image reconstruction, the proposed scheme considers DNN- and GAN-based super resolution at the receiver. The super resolution uses generation and discriminant models to reconstruct a high-quality image from the corresponding noisy image to realize low-traffic video delivery. In this case, the parameters of both models are trained using the face features in addition to the face-containing video frames to allow high-quality reconstruction.

B. Deep Neural Networks/Generative Adversarial Network-based Image Reconstruction

In recent years, some studies were aimed at DNN/GAN-based image reconstruction, such as super resolution using convolutional neural networks (SRCNN) [12], DCGAN [5], Laplacian pyramid of GAN (LPGAN), and super resolution using a GAN (SRGAN) [13] to further improve the reconstruction quality. Specifically, DCGAN combines a CNN with a GAN to generate a high-quality image from a noisy image. LPGAN and SRGAN extend DCGAN for super resolution. For example, LPGAN stacks multiple layers of DCGAN to gradually improve the reconstructed image quality. The study most relevant to our proposed scheme is that on srez [14]. In this study, the authors used DCGAN to reconstruct high-resolution face-containing images from low-resolution images.

The super resolution technique in the proposed scheme is based on srez. We use super resolution to reconstruct face-containing images to achieve traffic reduction for video call and chat applications. To further improve the reconstruction quality, we focused on the face features and designed a new loss function for super resolution that considers them.

C. Deep Neural Network-based Video Delivery

In a few studies, DNNs were applied to video delivery over the Internet [15], [16]. The main objective of these studies was to obtain better streaming quality by using DNN-based predictions.

In [15], a rate control method for real-time video streaming using DNN-based future bit-rates and delay predictions was proposed. In [16], the authors presented a video delivery framework to improve video quality in order to enhance users' quality of experience (QoE) when viewing video-on-demand contents, such as films and TV shows.

The approach most closely related to our study was included in [16], that is, the utilization of DNNs for video quality improvement. This approach is focused on the overall system structure and partially uses the conventional DNN for video-on-demand streaming. However, in our approach a new DNN is designed that includes a loss function that uses face features, for real-time video call and chat applications.

III. Proposed Scheme

In this paper, a transmission scheme for face-containing video contents is proposed that realizes a significant traffic reduction for video-based call and chat applications. Fig. 1 shows an overview of the proposed scheme. The scheme first extracts the face features from the captured face-containing video frames using the face detector in dlib. The captured video frames are then converted to low resolution and encoded by the H.264/AVC encoder prior to transmission. After the encoding, the sender transmits the low-resolution video frames and the corresponding face features to the receiver. The receiver first uses the H.264/AVC decoder to decode the received video frames. Since the resolution of the decoded video frames is lower than that of the original video frames, they are then reconstructed to the original resolution by using DCGAN and the face features.

A. Generation and Discriminant Model

DCGAN uses both generation and discriminant models for image reconstruction. For high-quality image reconstruction, the optimized parameters of the generation and discriminant models are determined by using numerous face-containing video frames in the training phase.

We consider an original video frame containing the sender's face I_{org} with a resolution of $W \times H \times C$ pixels and the low-resolution video frame I_{down} with a resolution of $rW \times rH \times C$ pixels. Here, W and H represent the width and height of the original video frame, respectively, and C is the number of color channels. In addition, r is a downscaling factor. The generation model $G (= G_{\theta_G}(I_{\mathrm{down}}))$ is one that reconstructs a video frame I_{rec} with a resolution of $W \times H \times C$ pixels from the low-resolution video frame I_{down} using the parameters of the generation model θ_G. In this case, the parameters need to be optimized to improve the reconstruction video quality using the generation model. This optimization is expressed by the following equation using N original video frames

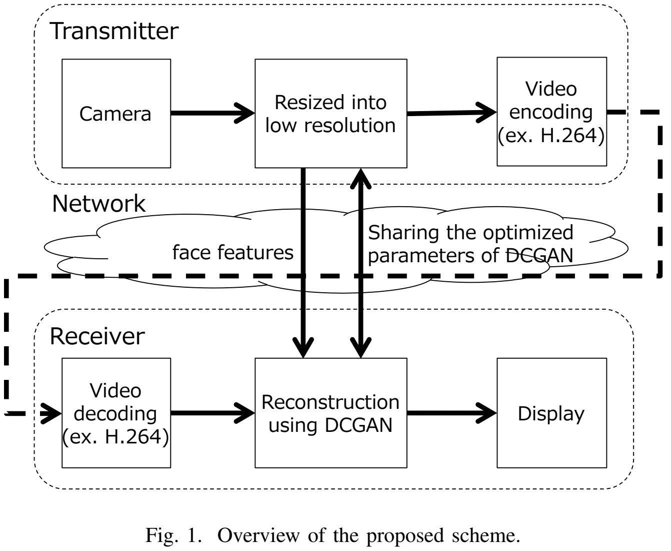

Transmitter
Camera
Resized into low resolution
Video encoding (ex. H.264)
Network
face features
Sharing the optimized parameters of DCGAN
Receiver
Video decoding (ex. H.264)
Reconstruction using DCGAN
Display

Fig. 1. Overview of the proposed scheme.

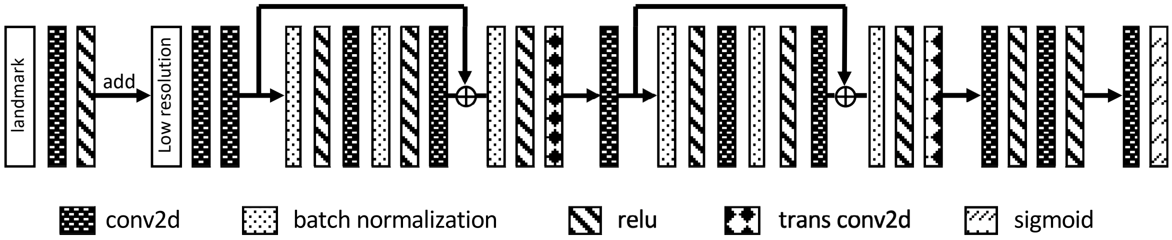


Fig. 2. Generation model of the proposed scheme.

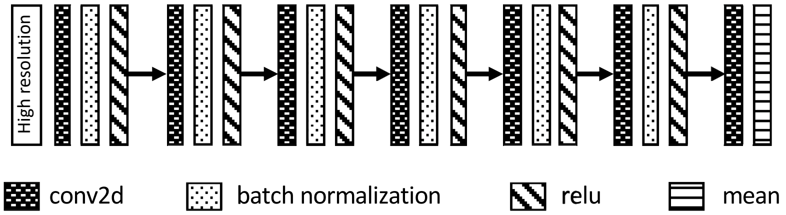


Fig. 3. Discriminant model of the proposed scheme.

$I_{\text{org}}^{(i)}, i = 1, ..., N$ and the same number of low-resolution images $I_{\text{down}}^{(i)}, i = 1, ..., N$:

$$\hat{\theta}_G = \arg\min_{\theta_G} \frac{1}{N} \sum_{n=1}^{N} l\left(G_{\theta_G}\left(I_{\text{down}}^{(i)}\right), I_{\text{org}}^{(i)}\right), \tag{1}$$

where $l\big(G_{\theta_G}(I_{\text{down}}^{(i)}), I_{\text{org}}^{(i)}\big)$ represents the loss function of the generation model with the inputs of the original and reconstructed video frames for parameter optimization of the generation model. The details of the loss function implemented in our scheme are described in Section III-B.

The discriminant model D is a binary classifier and used for the parameter optimization of the generation model G. The discriminant model attempts to distinguish between the reconstructed (false) data and the true data of the dataset. To improve the reconstruction accuracy, the proposed scheme solves the following minimax optimization by using both generation and discriminant models based on the method in [17]:

$$\begin{aligned}\min_{\theta_G} \max_{\theta_D} \mathbb{E}_{I_{\text{org}} \sim P_{\text{data}}(I)} \left[\log\left(D_{\theta_D}(I_{\text{org}})\right)\right] \\ + \mathbb{E}_{I_{\text{down}} \sim P_{\text{mosaic}}(I)} \left[\log\left(1 - D_{\theta_D}\left(G_{\theta_G}(I_{\text{down}})\right)\right)\right],\end{aligned} \tag{2}$$

where $D_{\theta_D}(I_{\text{org}})$ is the probability that the original video frame I_{org} belongs to the true dataset, $1 - D_{\theta_D}(G_{\theta_G}(I_{\text{down}}))$ is the probability that the reconstructed video frame generated by the generation model G belongs to the false dataset, and $\mathbb{E}[\cdot]$ is the expected value. Since the discriminant model D needs to correctly distinguish between the original and the reconstructed video frame, the parameters of the discriminant model θ_D should be optimized to maximize both probabilities. However, in view of the generation model, the reconstructed image should be classified into the correct dataset. This means that the generation model should find the optimized parameters that minimize the probability of $1 - D_{\theta_D}(G_{\theta_G}(I_{\text{down}}))$.

Figs. 2 and 3 show our generation model and discriminant model, respectively, based on the srez scheme [14]. In the generation model, the proposed scheme uses the convolution layer at the first layer. The convolution extracts the independent features of each image using certain convolution kernels. Note that the image features are shown in the weights between the layers. In the second layer, we use a rectified linear unit (ReLU) layer. Before the third layer, the proposed scheme adds the second layer output and the low-resolution frame generated by bilinear interpolation. This is the third layer input. At both the third and fourth layers, the proposed scheme uses convolution layers. In the fifth to seventh layers, the proposed scheme uses batch normalization, ReLU, and convolutional layers, respectively. The batch normalization standardizes the input values to balance the values in order to reduce the effect of outliers. The proposed scheme repeats these 3 layers one time to construct up to the 10-th layer, and then adds the 10-th layer output and 4-th layer output. At the 11-th to 13-th layers, we use batch normalization, ReLU, and the transconvolution layer, respectively, and then repeat the 4-th to 13-th layers one time to obtain the 14-th to 23-rd layers. Furthermore, the proposed scheme uses the convolution and the ReLU layers at the 24-th and 25-th layers and repeats the same layers for the 26-th and 27-th layers, and uses the convolution layer in the 28-th layer. Finally, the sigmoid layer is allocated to the final layer.

In the discriminant model, the first layer is convolution, the second layer is batch normalization, and the third layer is ReLU. The proposed scheme repeats these 3 layers up to the 18-th layer. In the 19-th and 20-th layers, the model uses the convolution and finally average operations.

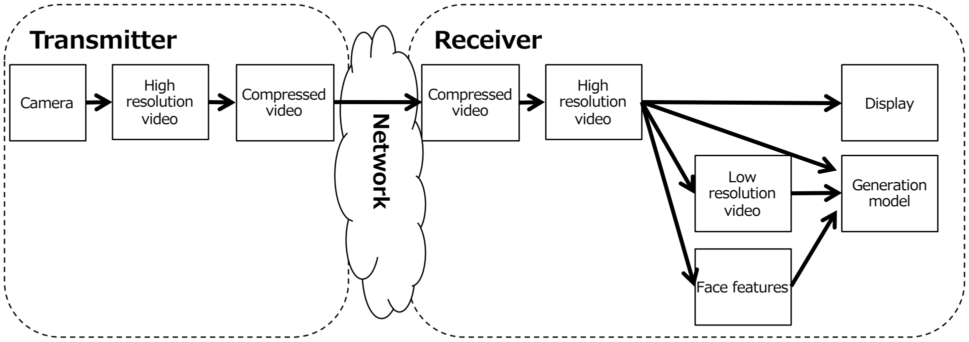


Fig. 4. Learning phase in sequential sharing.

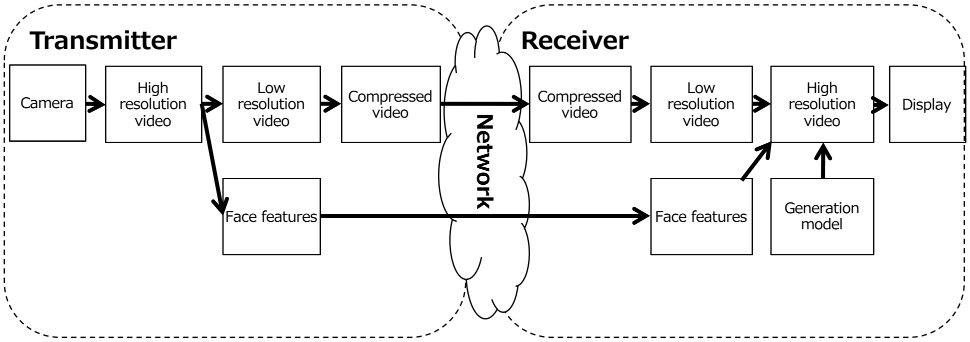


Fig. 5. Reconstruction phase in sequential sharing.

B. Loss Function

In our scheme, the loss function consists of three factors: generation loss, pixel-domain loss, and face feature loss. Specifically, the loss function is expressed as

$$l\Big(G_{\theta_G}\Big(I^{(i)}_{\text{down}}\Big), I^{(i)}_{\text{org}}\Big) = \alpha_1 \cdot l^{(i)}_{\text{adversarial}} + \alpha_2 \cdot l^{(i)}_{\text{pixel}} + \alpha_3 \cdot l^{(i)}_{\text{face}} \tag{3}$$

Here, α_1, α_2, α_3 are weights of each factor, the value of which is from 0 to 1. $l^{(i)}_{\text{adversarial}}$ represents the generation loss:

$$l^{(i)}_{\text{adversarial}} = -\log\Big(D_{\theta_D}\Big(G_{\theta_G}\Big(I^{(i)}_{\text{down}}\Big)\Big)\Big) \tag{4}$$

The generation loss increases as the probability that the reconstructed video frame generated by the generation model G belongs to the false dataset increases. $l^{(i)}_{\text{pixel}}$ represents the pixel-domain loss:

$$l^{(i)}_{\text{pixel}} = \frac{1}{r^2WH}\sum_{x=1}^{rW}\sum_{y=1}^{rH}\left|I^{(i)}_{\text{down}}(x,y) - R\Big(G_{\theta_G}\Big(I^{(i)}_{\text{down}}(x,y)\Big)\Big)\right|, \tag{5}$$

where $R(\cdot)$ represents a resize function and $I^{(i)}_{\text{down}}(x,y)$ represents the (x,y)-th pixel value of the i-th low-resolution video frame. This loss factor represents the average difference between the pixel values of the low-resolution original and the reconstructed video frames.

$l^{(i)}_{\text{face}}$ represents the difference between the face features in the original and reconstructed video frames:

$$l^{(i)}_{\text{face}} = \frac{1}{68}\sum_{k=1}^{68}\left|\left|\Psi\Big(I^{(i)}_{\text{org}}\Big) - \Psi\Big(G_{\theta_G}\Big(I^{(i)}_{\text{down}}\Big)\Big)\right|\right|_2 \tag{6}$$

where $\Psi(\cdot)$ is an extraction function of face features from face-containing video frames. To extract the face features from the face-containing video frames, the proposed scheme uses histogram of oriented gradient (HOG) features [18] and dlib detection. Here, we used the existing dataset http://dlib.net/files/shape_predictor_68_face_landmarks.dat.bz2 published by iBUG 300-W to obtain the HOG features in each video frame. The dlib detection algorithm extracts 68 face features, including jaw, eyes, eyebrows, nose, and mouth, from each face-containing video frame based on the HOG features. Note that the proposed scheme calculates $l^{(i)}_{\text{face}}$ only when the sender's face is detected by dlib in the i-th face-containing video frame. If dlib does not detect the sender's face in the video frames, the proposed scheme uses the following loss function instead of Eq. (3):

$$l\Big(G_{\theta_G}\Big(I^{(i)}_{\text{down}}\Big), I^{(i)}_{\text{org}}\Big) = (1-\alpha) \cdot l^{(i)}_{\text{adversarial}} + \alpha \cdot l^{(i)}_{\text{pixel}}. \tag{7}$$

Here, α is a parameter that takes a real value between 0 and 1.

C. Parameter Sharing of Generation Model

The proposed scheme finally shares the parameters of the generation model between the transmitter and the receiver before video transmissions. It considers two methods for parameter sharing: sequential sharing and pre-sharing. The sequential sharing method periodically updates the parameters of the generation model based on the new sender's face-containing video frames while the parameter optimization overhead becomes larger. The pre-sharing method always sends the pre-trained parameters to the receiver to decrease the overhead. The reconstruction quality is degraded if the captured video frames contain objects and scenes that differ from those of the training phase.

1) Sequential Sharing: The sequential sharing method consists of two phases: learning and reconstruction. Fig. 4 shows an overview of the learning phase. In this phase, the receiver simultaneously displays the high-resolution video frames and trains the parameters of the generation model using these frames. Specifically, a sender first encodes the high-resolution video frames using the standard video encoder H.264/AVC and transmits the encoded video frames to the receiver. The receiver first obtains the high-resolution frames by decoding the received video frames and displays the decoded video frames. At the same time, the receiver resizes the decoded video frames into low-resolution frames and extracts the face features from the high-resolution video frames. The receiver then trains the parameters of the generation model using both the high- and low-resolution video frames and the face features to reconstruct the high-resolution video frames from the low-resolution video frames and the corresponding face features. When the training in the learning phase has been completed, the receiver sends a signal to the transmitter that the reconstruction phase should be initiated by the transmitter and the receiver.

Fig. 5 shows an overview of the reconstruction phase. The transmitter resizes the high-resolution video frames into low-resolution frames and extracts the face features from the high-resolution video frames. The low-resolution video frames are encoded by H.264/AVC, and then, the encoded video frames and the corresponding face features are transmitted to the receiver. The receiver reconstructs high-resolution video frames by using the generation model obtained in the learning phase. In this case, the receiver uses the low-resolution face-containing video frames and the corresponding face features as inputs of the generation model. Finally, the reconstructed video frames are displayed on the receiving device.

2) Pre-Sharing Method: The pre-sharing method sends the pre-trained parameters of the generation model before initializing video-based call and chat applications. After sharing the pre-trained parameters with the receiver, the transmitter sends low-resolution face-containing video frames and the extracted face features. Finally, the receiver reconstructs high-resolution video frames using the generation model and the face features; the operations are the same as in the reconstruction phase in the sequential sharing method.

IV. Performance Evaluation

A. Evaluation Settings

In the performance evaluation of the proposed scheme for video-based call and chat applications, we used real video sequences of video call applications. The video frames were captured by a PENTAX KS-2. The frame rate was 30 frames per second (fps) and the duration of the video sequence was 10 m. Each video frame was compressed by using Joint Photographic Experts Group (JPEG). The compressed video frames were resized to two resolutions, i.e., 80×80 pixels and 160×160 pixels, and regarded as the input high-resolution video frames. We prepared 18,000 input video frames for training and testing. The batch size of both training and testing was 16.

The parameters of α, α_1, α_2, and α_3 were 0.90, 0.90, 0.09, and 0.40, respectively. The initial value of the learning coefficient was 0.0002. In addition, the initial values of the weight and bias were obtained from the normal distribution and 0, respectively. In our scheme, the parameter optimization algorithm and the initial values were based on those of Adam [19].

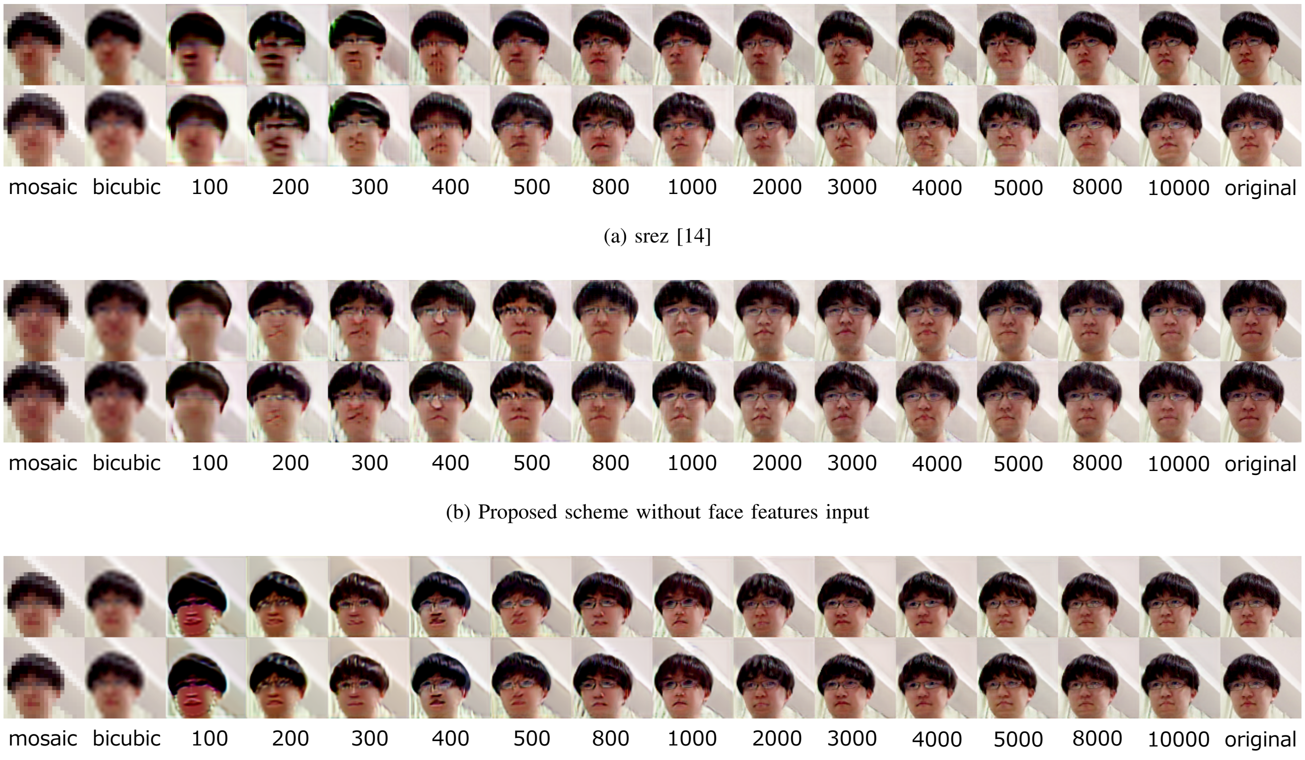
B. Effect of Batch Counts

We first discuss the visual quality of the reconstructed video frames as a function of batch counts. Figs. 6(a)–(c) show the visual quality of the reconstructed video frames for different numbers of batch counts. In Fig. 6(a), the reconstruction algorithm is based on srez [14]. Figs. 6(b) and (c) show the visual quality of the results of the proposed scheme using both loss functions, Eqs. (3) and (7). Figs. 6(c) and (b) respectively show the results when the proposed scheme does and does not use the face features for image reconstruction. From left to right, each image is the mosaic image, the reconstructed image using bicubic interpolation, the reconstructed image after 100, 200, 300, 400, 500, 800, 1,000, 2,000, 3,000, 4,000, 5,000, 8,000, and 10,000 batch counts, and the original image. To create the mosaic image, we resize the input image to 20×20 pixels and adopt nearest neighbor interpolation to reconstruct the image. Here, two images are selected from the testing images to compare the visual quality.

The evaluation results show that the proposed scheme reconstructs a genuine face in a smaller number of batch counts as compared with the existing srez scheme. This means the proposed loss function reconstructs high-resolution images with a low overhead requirement. In addition, we can see that the reconstructed images in the proposed scheme that includes the face features show a clear face at a smaller batch count as compared with those in the proposed scheme that does not include the face features. This means that the face features lead to a better reconstruction quality even in a small number of batch counts.

C. Data Size Reduction

This section evaluates the data size of the proposed scheme to demonstrate the effect on traffic reduction. Fig. 7 shows the JPEG-coded (intra-coded), low-resolution, and reconstructed



(a) srez [14]

(b) Proposed scheme without face features input

(c) Proposed scheme with face features input

Fig. 6. Visual quality of reconstructed and original video frames as a function of batch counts.

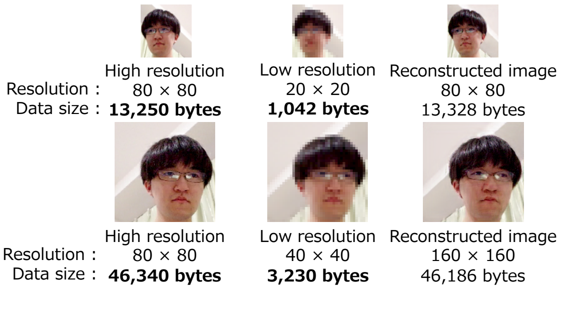


Fig. 7. Data size of intra-coded and low-resolution images.

images and their data size at a batch count of 1,250. Here, the resolution of the upper and lower images is 80×80 pixels and 160×160 pixels, respectively.

It is demonstrated that the data size of the intra-coded and low-resolution images is 13,250 Bytes and 1,042 Bytes, respectively, at a resolution of 80×80 pixels, while at a resolution of 160×160 pixels the data size is 46,340 Bytes and 3,230 Bytes, respectively. Thus, it is demonstrated that the proposed scheme reduces the amount of video traffic by more than 90% as compared with the intra coding in H.264/AVC, irrespective of image resolutions.

D. Image Quality

In this section, we evaluate the quality of the reconstructed image in each reference scheme in terms of the weighted peak signal-to-noise ratio (WPSNR) and SSIM [20]. The WPSNR is obtained by performing a weighted average of peak signal-to-noise ratios (PSNR) in the YCbCr color components, since the Y component contains considerable information as compared with the other color components. Here, the weight in each color component is $\mathrm{Y} : \mathrm{Cb} : \mathrm{Cr} = 8 : 1 : 1$ [21]. SSIM predicts the perceived video streaming quality. Larger values of SSIM close to 1 indicate a higher perceptual similarity between the original and reconstructed images. In contrast to WPSNR, SSIM considers the correlation between each pixel and its surroundings based on the luminance, contrast, and structure. It has been confirmed that SSIM reflects human visual characteristics as compared with PSNR [22].

Fig. 8 shows the quality of the mosaic image, a reconstructed image using bicubic interpolation, srez, and the proposed scheme in terms of WPSNR. The evaluation results show that the performance of srez and the proposed scheme are almost the same, even at a large number of batch counts. Since WPSNR considers pixel-wise distortion and suffers from low quality even in a 1-bit pixel shift, its use as an image reconstruction metric is not suitable.

Fig. 9 shows the quality of the mosaic image, the reconstructed image using bicubic interpolation, srez, and the proposed scheme in terms of SSIM. As seen in this figure, the proposed scheme with face features achieves the highest video quality. This means the proposed scheme should use face features to improve the reconstructed image quality. In

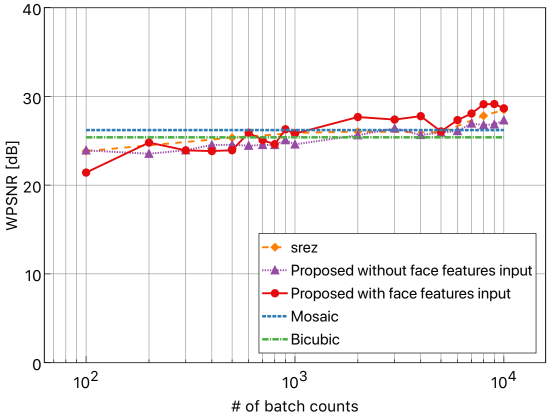


Fig. 8. Video quality of reference schemes as a function of batch counts in terms of weighted peak signal-to-noise ratio.

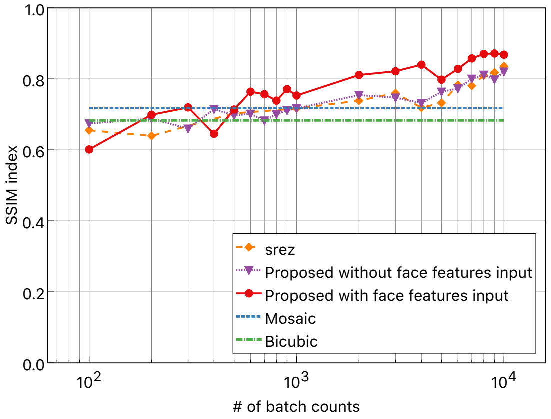


Fig. 9. Video quality of reference schemes as a function of batch counts in terms of structural similarity.

addition, the reconstruction quality of the proposed scheme without face features is almost the same as that of srez. It is noted that the proposed scheme obtains a clear face with a small number of batch counts as compared with srez, as shown in Fig. 6.

V. Conclusion

In this paper, a traffic reduction scheme for video call and chat applications that uses super resolution based on DNNs was proposed. Specifically, the proposed scheme uses face features for loss functions in the parameter optimization and high-quality image reconstruction. According to the evaluation results, the proposed scheme significantly reduces the amount of video traffic as compared with the conventional video encoding and simultaneously maintains better reconstruction quality.

References

[1] E. McClure and R. Barr, “Building family relationships from a distance: Supporting connections with babies and toddlers using video and video chat,” in *Media Exposure During Infancy and Early Childhood*. Springer, 2017, pp. 227–248.

[2] R. Janghorban, R. L. Roudsari, and A. Taghipour, “Skype interviewing: The new generation of online synchronous interview in qualitative research,” *International Journal of Qualitative Studies on Health and Well-being*, vol. 9, no. 1, p. 24152, 2014.

[3] T. Wiegand, G. J. Sullivan, G. Bjontegaard, and A. Luthra, “Overview of the h. 264/avc video coding standard,” *IEEE Transactions on circuits and systems for video technology*, vol. 13, no. 7, pp. 560–576, 2003.

[4] D. Grois, D. Marpe, A. Mulayoff, B. Itzhaky, and O. Hadar, “Performance comparison of h. 265/mpeg-hevc, vp9, and h. 264/mpeg-avc encoders,” in *Proceedings of Picture Coding Symposium 2013 (PCS’13)*. IEEE, 2013, pp. 394–397.

[5] A. Radford, L. Metz, and S. Chintala, “Unsupervised representation learning with deep convolutional generative adversarial networks,” in *Proceedings of International Conference on Learning Representations 2016 (ICLR’16)*, 2016.

[6] R. Timofte, V. De Smet, and L. Van Gool, “A+: Adjusted anchored neighborhood regression for fast super-resolution,” in *Proceedings of The 12th Asian Conference on Computer Vision (ACCV’14)*. Springer, 2014, pp. 111–126.

[7] R. Yaniv, I. John, and M. Peyman, “Raisr: Rapid and accurate image super resolution,” *IEEE Transactions on Computational Imaging*, vol. 3, no. 1, pp. 110–125, 2017.

[8] S. Borman and R. L. Stevenson, “Super-resolution from image sequences-a review,” in *Proceedings of The IEEE Midwest Symposium on Circuits and Systems 1998*. IEEE, 1998, pp. 374–378.

[9] S. Farsiu, M. D. Robinson, M. Elad, and P. Milanfar, “Fast and robust multiframe super resolution,” *IEEE transactions on image processing*, vol. 13, no. 10, pp. 1327–1344, 2004.

[10] M. S. Sajjadi, R. Vemulapalli, and M. Brown, “Frame-recurrent video super-resolution,” in *Proceedings of The IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2018 (CVPR’18)*. IEEE Computer Society, 2018.

[11] Y. Huang, W. Wang, and L. Wang, “Video super-resolution via bidirectional recurrent convolutional networks,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 1015–1028, 2018.

[12] C. Dong, C. C. Loy, K. He, and X. Tang, “Image super-resolution using deep convolutional networks,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 2, pp. 295–307, 2016.

[13] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, “Photo-realistic single image super-resolution using a generative adversarial network,” in *Proceedings of The IEEE Conference on Computer Vision and Pattern Recognition 2017 (CVPR’17)*, July 2017.

[14] D. Garcia, “Image super-resolution through deep learning (access: 2017/12/21),” https://github.com/david-gpu/srez, 2016.

[15] T. Huang, R.-X. Zhang, C. Zhou, and L. Sun, “Delay-constrained rate control for real-time video streaming with bounded neural network,” in *Proceedings of the 28th ACM SIGMM Workshop on Network and Operating Systems Support for Digital Audio and Video*, ser. NOSSDAV ’18. ACM, 2018, pp. 13–18.

[16] H. Yeo, Y. Jung, J. Kim, J. Shin, and D. Han, “Neural adaptive content-aware internet video delivery,” in *Proceedings of 13th USENIX Symposium on Operating Systems Design and Implementation (OSDI’18)*. USENIX Association, 2018, pp. 645–661.

[17] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in *Proceedings of Advances in Neural Information Processing Systems 27 (NIPS’14)*. Curran Associates, Inc., 2014, pp. 2672–2680.

[18] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *Proceedings of The IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2005 (CVPR’05)*, vol. 1. IEEE, 2005, pp. 886–893.

[19] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *Proceedings of The International Conference on Learning Representations 2015 (ICLR’15)*, 2015.

[20] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: From error visibility to structural similarity,” *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[21] Z. Wang, L. Lu, and A. C. Bovik, “Video quality assessment based on structural distortion measurement,” *Signal Processing: Image Communication*, vol. 19, no. 2, pp. 121–132, 2004.

[22] Z. Wang and A. C. Bovik, “Mean squared error: Love it or leave it? a new look at signal fidelity measures,” *IEEE Signal Processing Magazine*, vol. 26, no. 1, pp. 98–117, 2009.