PROCEEDINGS OF SPIE

SPIEDigitalLibrary.org/conference-proceedings-of-spie

A generative adversarial network for video compression

Pengli Du, Ying Liu, Nam Ling, Lingzhi Liu, Yongxiong Ren, et al.

Pengli Du, Ying Liu, Nam Ling, Lingzhi Liu, Yongxiong Ren, Ming Kai Hsu, "A generative adversarial network for video compression," Proc. SPIE 12097, Big Data IV: Learning, Analytics, and Applications, 120970E (31 May 2022); doi: 10.1117/12.2618714

SPIE.

Event: SPIE Defense + Commercial Sensing, 2022, Orlando, Florida, United States

A Generative Adversarial Network for Video Compression

Pengli Du[a], Ying Liu[a], Nam Ling[a], Lingzhi Liu[b], Yongxiong Ren[b], and Ming Kai Hsu[b]

[a]Santa Clara University, Santa Clara, CA 95053, USA
[b]Kwai Inc., Palo Alto, CA 94306, USA

ABSTRACT

Video data has occupied people's daily professional and entertainment activities. It imposes a big pressure on the Internet bandwidth. Hence, it is important to develop effective video coding techniques to compress video data as much as possible and save the transmission bandwidth, while still providing visually pleasing decoded videos. In conventional video coding such as the high efficiency video coding (HEVC) and the versatile video coding (VVC), signal processing and information theory-based techniques are mainstream. In recent years, thanks to the advances in deep learning, a lot of deep learning-based approaches have emerged for image and video compression. In particular, the generative adversarial networks (GAN) have shown superior performance for image compression. The decoded images are usually sharper and present more details than pure convolutional neural network (CNN)-based image compression and are more consistent with human visual system (HVS). Nevertheless, most existing GAN-based methods are for still image compression, and truly little research investigates the potential of GAN for video compression. In this work, we propose a novel inter-frame video coding scheme that compresses both reference frames and target (residue) frames by GAN. Since residue signals contain less energy, the proposed method effectively reduces the bit rates. Meanwhile, since we adopt adversarial learning, the perceptual quality of decoded target frames is well-preserved. The effectiveness of our proposed algorithm is demonstrated by experimental studies on common test video sequences.

Keywords: Deep learning, generative adversarial network, human visual system, image compression, motion compensation, perceptual quality, residue-frame coding, video coding

1. INTRODUCTION

Video data has occupied people's daily professional and entertainment activities, such as video conferences, live commerce, online video games, live concerts and shows. When a popular live show has a high demand among audiences, it imposes a big pressure on the transmission bandwidth. Hence, it is important to develop effective video coding techniques to compress the video signal as much as possible and save the bandwidth, while still providing visually pleasing decoded videos to let the audiences enjoy the live show.

In conventional video coding such as the high efficiency video coding (HEVC) [1] and the versatile video coding (VVC) [2], [3], signal processing and information theory based techniques are mainstream. In recent years, thanks to the advances in deep learning [4–6], a lot of deep learning-based video coding techniques have emerged. In particular, generative adversarial network (GAN)-based image compression schemes have shown superior performance in offering high perceptual-quality decoded images [7–9]. GAN [10] was originally proposed to generate photo-realistic images from random noise, by training two competing networks: the generator and the discriminator. Recently, studies showed that the decoded images of GAN-based compression systems are usually sharper and present more details than pure convolutional neural network (CNN)-based image compression that merely adopts the mean-squared-error (MSE) loss to train the network [11], [12].

Nevertheless, most existing works utilize GAN only for image compression [8], [13], [14], image style translation [15], [16], artifact removal [17], or for video frame prediction and generation [18–25]. Truly little research directly applied GAN to residue-frame coding. In this work, we propose a novel video coding system in which both reference frames and target (residue) frames are compressed by GAN. Since residue signals contain less energy,

Further author information: (Send correspondence to Ying Liu)
Pengli Du: pdu@scu.edu, Ying Liu: yliu15@scu.edu, Nam Ling: nling@scu.edu,
Lingzhi Liu: l.liu@kwai.com, Yongxiong Ren: yongxiongren@kwai.com, Ming Kai Hsu: mingkaihsu@kwai.com.

Big Data IV: Learning, Analytics, and Applications, edited by Fauzia Ahmad, Panos P. Markopoulos, Bing Ouyang, Proc. of SPIE Vol. 12097, 120970E
 · 0277-786X · doi: 10.1117/12.2618714

the proposed method effectively reduces the bit rates. Meanwhile, since we adopt adversarial learning, the perceptual quality of decoded target frames are well-preserved.

The rest of the paper is organized as follows: Section 2 introduces the related work in GAN-based image and video compression and processing, Section 3 elaborates the proposed GAN-based video coding system, and Section 4 presents the experimental studies and analysis. Finally, Section 5 concludes the paper.

2. RELATED WORKS

GAN-based image compression was first proposed in [7]. It trained a regular GAN network for image generation, and then used the trained generator as the image decoder, followed by training an encoder to work with the previously learned decoder. In this way, the decompressed image is expected to look more like a natural image that has higher perceptual quality. In the same work, a GAN-based video compression system was also proposed, with frame interpolation in the latent space (encoded feature space). However, the experiments in this work were only conducted on an old gray-scale action recognition data set [26] instead of HEVC test sequences.

The work in [8] proposed a GAN to compress images and to synthesize unimportant regions in the decoded image from a semantic label map. The method in [13] proposed a GAN-based face image compression scheme that not only considers the pixel-domain distortion loss and the adversarial loss, but also incorporates a semantic loss that preserves features for face recognition. Nevertheless, the major problem in these works is they do not consider videos and are limited to only image compression.

In [9], a GAN-based video coding scheme is proposed. It has a deterministic encoder that encodes the edges of the video frames with high compression rates, and only the decoder is trained as a generator in a GAN setup. However, this work does not explore motion and does not perform residue coding.

In addition, GAN has been used for artifact removal to enhance the quality of traditional intra-frame coding [17]. However, in this work the encoder and decoder that perform the compression do not involve any idea of GAN.

GAN has been more often investigated as a means for video prediction and generation [18–25], such as frame extrapolation [18], [24], slow-motion [20] and multi-view [25] video generation. However, in all of these works, GAN is only utilized as a prediction tool. It does not participate in the actual compression module.

3. THE PROPOSED GAN-BASED INTER-FRAME CODING FRAMEWORK

Our proposed video coding architecture starts with a reference-frame coding module, which encodes and decodes the reference frame $\mathbf{X}_{t-1} \in \mathbb{R}^{H \times W \times 3}$ at time-slot $t-1$, using the GAN-based image compression scheme in [8]. It generates the decoded reference frame $\widehat{\mathbf{X}}_{t-1}$. Here H, W, and 3 represent the height, width, and channel of the frame.

Subsequently, as shown in Fig. 1, our proposed GAN-based inter-frame coding scheme compresses the target frame $\mathbf{X}_t$ with two modules: (a) a motion-compensated target-frame prediction module; and (b) a GAN-based residue-frame coding module.

3.1 Motion-Compensated Target-Frame Prediction

As shown in Fig. 1 (a), the decoded reference frame $\widehat{\mathbf{X}}_{t-1}$ and the ground-truth target frame $\mathbf{X}_t$ are fed into a motion estimation and compensation module, to generate a motion-compensated prediction $\mathbf{X}_t^p$ of the target frame. In particular, a motion estimation module first estimates the optical flow $\mathbf{F}_{t-1 \to t}$ between $\widehat{\mathbf{X}}_{t-1}$ and $\mathbf{X}_t$, then an optical-flow compression block encodes $\mathbf{F}_{t-1 \to t}$ into a bit stream. Then, warping is performed on the decoded reference frame $\widehat{\mathbf{X}}_{t-1}$ and the decoded flow $\widehat{\mathbf{F}}_{t-1 \to t}$, to generate a warped target frame $\mathbf{X}_t^w$. Finally, a motion compensation block processes $\mathbf{X}_t^w$, $\widehat{\mathbf{X}}_{t-1}$, and $\widehat{\mathbf{F}}_{t-1 \to t}$ to generate the final prediction $\mathbf{X}_t^p$ of the target frame.

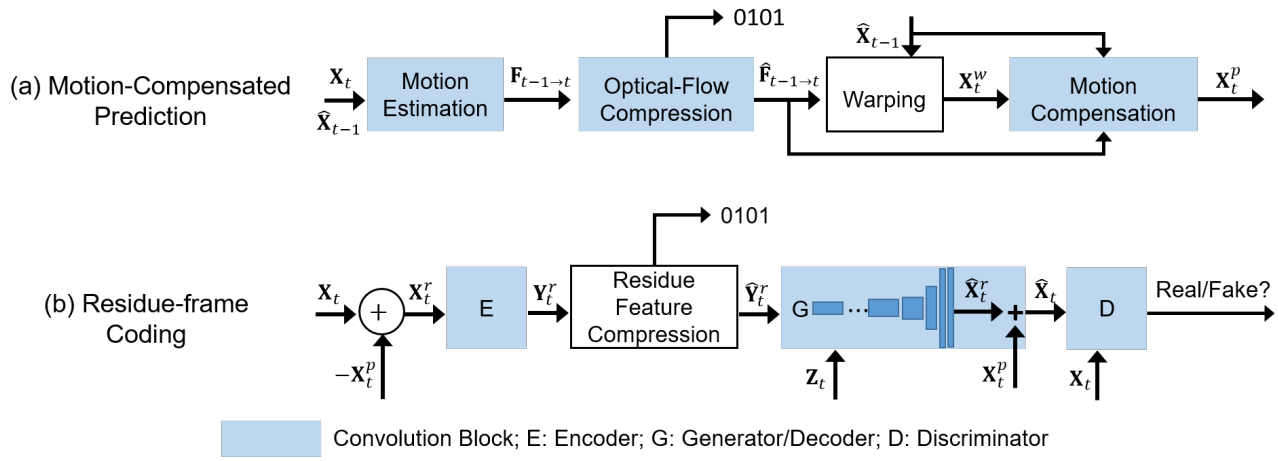

Figure 1. The proposed GAN-based inter-frame video coding architecture.

3.2 Residue-Frame Coding

Fig. 1 (b) shows the residue-frame coding module. The motion-compensated prediction $\mathbf{X}_t^p$ is subtracted from the target frame $\mathbf{X}_t$ to obtain the residue frame $\mathbf{X}_t^r$, which is the input of the residue-frame encoder. In particular, the encoder E is consisted of several convolutional layers and outputs the encoded residue feature $\mathbf{Y}_t^r \in \mathbb{R}^{h \times w \times c}$, where h, w, and c represent the height, width, and channel of the feature maps. The output layer of the encoder adopts the ReLU activation function. Then, a residue feature compression block performs quantization, arithmetic coding, and de-quantization.

The input of the generator/decoder G is the channel concatenation of the de-quantized residue feature $\widehat{\mathbf{Y}}_t^r \in \mathbb{R}^{h \times w \times c}$ and a random noise $\mathbf{Z}_t \in \mathbb{R}^{h \times w \times c}$. The generator is consisted of a convolutional layer with 960 feature maps, 9 residual blocks, each with 960 feature maps, several up-sampling layers that convert the height and width of the feature maps to the original frame size, followed by a convolutional layer that generates the decoded residue frame $\widehat{\mathbf{X}}_t^r \in \mathbb{R}^{H \times W \times 3}$. It is then added to the prediction $\mathbf{X}_t^p$ to form the final output of the generator, that is, the decoded target frame $\widehat{\mathbf{X}}_t \in \mathbb{R}^{H \times W \times 3}$.

3.3 Multi-scale discriminator

For the discriminator D, we use the multi-scale architecture of [8], which was originally proposed in [27]. The inputs of the discriminator are $\mathbf{X}_t^s$ and $\widehat{\mathbf{X}}_t^s$, representing the ground-truth and decoded t-th frame at scale s. Scale $s = 1$ refers to the original resolution, scales $s = \frac{1}{2}$ and $s = \frac{1}{4}$ refer to the frame down-sampled by a factor of 2 and 4, respectively.

For each scale s, the discriminator processes the ground-truth $\mathbf{X}_t^s$ to extract features $\mathbf{F}_{t,l}^s$ of five convolutional layers $l = 1, 2, \cdots, 5$. The same operation is conducted on the decoded target frame $\widehat{\mathbf{X}}_t^s$ to extract features $\widehat{\mathbf{F}}_{t,l}^s$ of five layers $l = 1, 2, \cdots, 5$.

3.4 Loss functions

The loss function adopted to train the encoder and generator $L_{encoder\text{-}generator}(E, G)$ is consisted of three components. The first component is the distortion loss defined as the MSE between the ground-truth target frame $\mathbf{X}_t$ and the decoded target frame $\widehat{\mathbf{X}}_t$

$$L_{distortion}(E, G) = \text{MSE}(\mathbf{X}_t, \widehat{\mathbf{X}}_t). \tag{1}$$

The second component is the generator loss defined as

$$L_{generator}(E, G) = \|\widehat{\mathbf{F}}_{t,5}^1 - \mathbf{1}^1\|_F^2, \tag{2}$$

where $\widehat{\mathbf{F}}^1_{t,5}$ is the discriminator layer-5 feature extracted from the full-resolution decoded target frame $\widehat{\mathbf{X}}^1_t$, $\mathbf{1}^1$ is an all-one tensor of the same size as $\widehat{\mathbf{F}}^1_{t,5}$, and $\|\cdot\|^2_F$ represents the squared Frobenius-norm. Minimizing such generator loss enforces each element in $\widehat{\mathbf{F}}^1_{t,5}$ to approach 1, which trains an encoder-generator pair that fools the discriminator. The third component is the feature matching loss defined as

$$L_{\text{feature}}(E,G) = \sum_{s=1,\frac{1}{2},\frac{1}{4}} \sum_{l=1}^{4} \text{MSE}(\mathbf{F}^s_{t,l}, \widehat{\mathbf{F}}^s_{t,l}). \tag{3}$$

That is, the element-wise MSE between the discriminator features extracted from the ground-truth target frame and those extracted from the decoded target frame, summed over all scales and all intermediate discriminator layers $l = 1, 2, 3, 4$.

Hence, the overall encoder-generator loss is defined as

$$L_{encoder\text{-}generator}(E,G) = L_{generator}(E,G) + \lambda_x L_{distortion}(E,G) + \lambda_f L_{feature}(E,G), \tag{4}$$

where λ_x and λ_f are the weights that trade off the three components.

The loss function adopted to train the multi-scale discriminator is defined as

$$L_{\text{discriminator}}(D) = \sum_{s=1,\frac{1}{2},\frac{1}{4}} \left(\|\mathbf{F}^s_{t,5} - \mathbf{1}^s\|^2_F + \|\widehat{\mathbf{F}}^s_{t,5}\|^2_F \right), \tag{5}$$

where $\mathbf{1}^s$ is an all-one tensor of the same size as $\mathbf{F}^s_{t,5}$, $s = 1, \frac{1}{2}, \frac{1}{4}$. Minimizing $L_{\text{discriminator}}(D)$ means that each element in $\mathbf{F}^s_{t,5}$ should approach 1, and each element of $\widehat{\mathbf{F}}^s_{t,5}$ should approach 0, which can learn a discriminator that distinguishes the fake (decoded) target frames from the ground-truth target frames.

4. EXPERIMENTAL STUDIES

4.1 Datasets

We conduct experimental studies on three HEVC test sequences, *BlowingBubbles*, *BQSquare*, and *Johnny*. The resolution of *BlowingBubbles* and *BQSquare* is 240×416. The original resolution of *Johnny* is 720×1280. To save the training time, we resizes *Johnny* such that it has the same resolution 240×416 as the other two videos.

We group the frames in each video sequence as pairs of odd and even frames $(\mathbf{X}_{t-1}, \mathbf{X}_t)$, $t = 2, 4, 6, \cdots$, in which the odd frames are the reference frames, and the even frames are the target frames. The *Johnny* sequence has 600 frames in total, so there are 300 pairs of odd and even frames. We randomize these 300 pairs and split them into 3 groups, each with 100 pairs. We first train a GAN-based intra-frame compression network [8] with the even frames in group 1. Afterwards, we use the trained intra-frame compression model to encode and decode the odd frames of group 2.

Next, we applied the motion-compensated prediction model in Fig. 1 (a) to predict the ground-truth even frames of group 2 $(\mathbf{X}_t, t = 2, 4, 6, \cdots)$ from the decoded odd frames of group 2 $(\widehat{\mathbf{X}}_{t-1}, t = 2, 4, 6, \cdots)$. The model was pre-trained by [28] using the Vimeo-90K [29] data set. The predicted frames are denoted as $\mathbf{X}^p_t, t = 2, 4, 6, \cdots$. Then, we use the predicted and ground-truth even frames of group 2 to train the proposed residue-frame compression network in Fig. 1 (b). Finally, the 100 pairs in group 3 are used to test the trained GAN-based video coding model with three steps: reference-frame coding, motion-compensated prediction, and residue-frame coding. A similar approach is used for the training and testing of the other two video sequences.

4.2 Evaluation metrics

We evaluate the performance of the proposed GAN-based video coding system subjectively by the visual quality of the decoded frames and quantitatively by the rate-distortion metrics. The distortion is measured by the peak signal-to-noise ratio (PSNR) and the multi-scale structural similarity (MS-SSIM) [30] between the decoded and the ground-truth target frames, and the bit rates are measured by the average bits per pixel (bpp) of the encoded target frames.

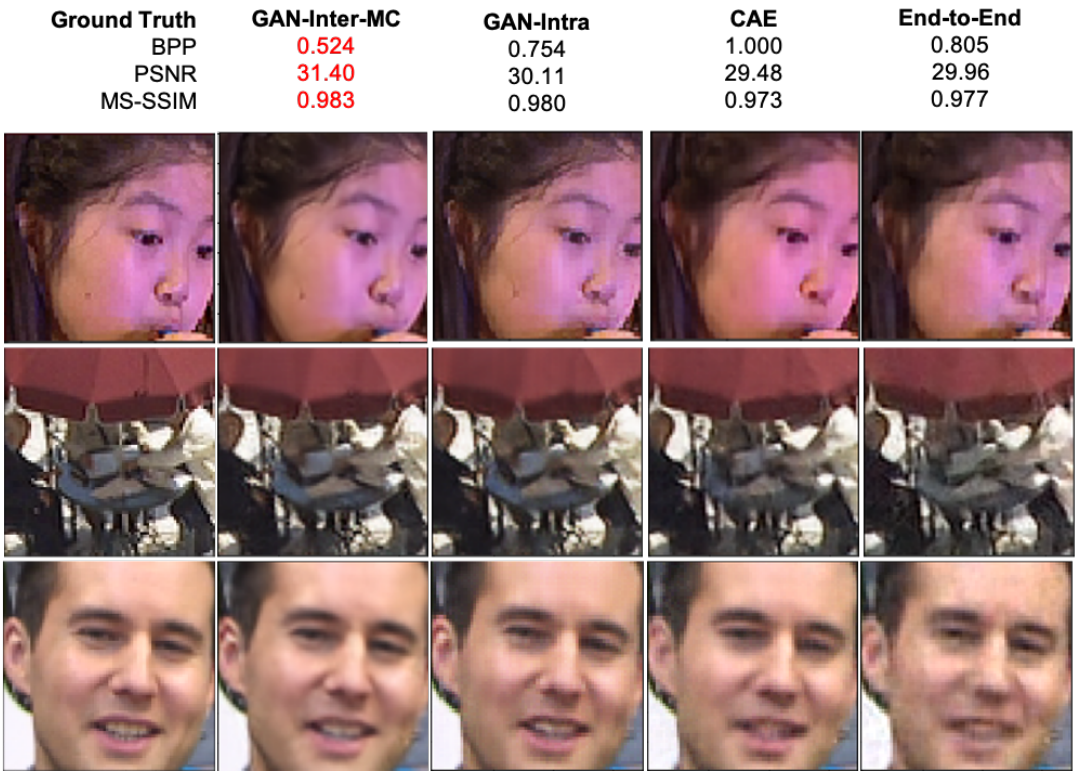

Figure 2. Sample decoding of *BlowingBubbles* (first-row), *BQSquare* (second-row) and *Johnny* (third-row). The average bit rates (bpp), PSNR (dB), and MS-SSIM of the three frames are labeled above the figure.

4.3 Comparison with other methods

We compare the performance of our proposed GAN-based motion-compensated inter-frame video coding system (GAN-Inter-MC) to the original GAN-based image compression network (GAN-Intra) [8] and another two state-of-the-art deep learning-based image compression schemes: the end-to-end optimized image compression network [11] (End-to-End) and the compressive auto-encoder [12] (CAE). These two models are CNN-based approaches with different network structures, but both of them adopt the MSE between the ground-truth frame and the decoded frame as the loss function.

Fig. 2 shows an enlarged 70×70 region of a ground-truth target frame for each of the three video sequences, and the corresponding decoding by our proposed GAN-Inter-MC, GAN-Intra [8], CAE [12] and End-to-End [11]. The average bit rates (bpp), PSNR (dB) and MS-SSIM values are labeled above the figure. Our proposed GAN-Inter-MC scheme reduces the average bpp by 30.58%, 47.63%, and 34.98% while increasing the average PSNR by 1.29 dB, 1.92 dB, and 1.44 dB compared to GAN-Intra, CAE and End-to-End, respectively. The average MS-SSIM of GAN-Inter-MC is higher than that of the other three schemes. Besides, the perceptual qualities of the proposed GAN-Inter-MC and GAN-Intra are significantly better than those of CAE and End-to-End. They both recover texture details in the decoded frames, such as the girl's hair and ear areas for *BlowingBubbles* (Fig. 2 first-row), the things on the table for *BQSquares* (Fig. 2 second-row), and the man's mouth for *Johnny* (Fig. 2 third-row). In contrast, the decoded frames of CAE and End-to-End are much more blurry. The reason is CAE and End-to-End merely adopt MSE as their loss functions, and MSE only enforces consistency in pixel intensity values. Although the perceptual quality of the proposed GAN-Inter-MC is similar to that of GAN-Intra, it requires 30.58% less bit rates.

To provide quantitative performance evaluation in detail, we plot the PSNR and MS-SSIM versus bit rate curves in Fig. 3, and compare our proposed GAN-Inter-MC to the other three methods. We observe that for all three videos, GAN-Inter-MC is able to achieve the same or higher PSNR and MS-SSIM values at much lower bit rates, compared to all other schemes.

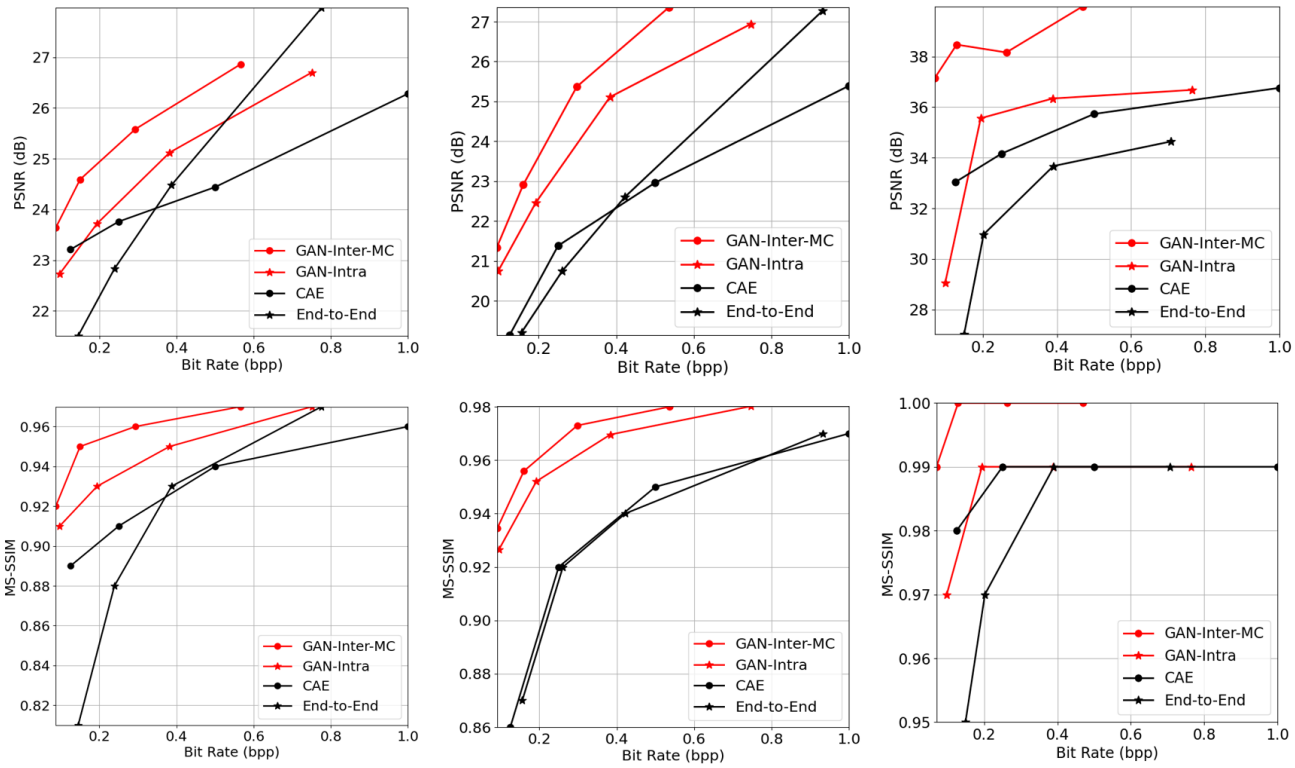

Figure 3. The peak signal-to-noise ratio (PSNR) and the multi-scale structural similarity index (MS-SSIM) versus the bit rates (bpp) of *BlowingBubbles* (left-column), *BQSquare* (middle-column), and *Johnny* (right-column).

5. CONCLUSIONS

In this paper, we proposed a novel GAN-based video coding system that uses GAN to encode and decode both the reference frames and the residue between the target and predicted frames. The method effectively reduces the bit rates compared to GAN-based intra-frame compression. Compared to CNN-based image compression, the proposed method has leveraged the adversarial learning of GAN to generate decoded frames that contain more texture details and are more consistent with HVS. Overall, our scheme simultaneously achieved high perceptual quality and reduced bit rates. In terms of future studies, we will test our method on videos of different resolutions, and will incorporate hierarchical motion prediction in our GAN-based video coding system to further enhance the coding efficiency.

REFERENCES

[1] Sullivan, G. J., Ohm, J.-R., Han, W.-J., and Wiegand, T., "Overview of the high efficiency video coding (hevc) standard," *IEEE Trans. Circuits Syst. Video Technol.* **22**, 1649–1668 (Dec. 2012).

[2] Segall, A., Baroncini, V., Boyce, J., Chen, J., and Suzuki, T., "Joint call for proposals on video compression with capability beyond hevc," *JVET-H1002* , 18–24 (Oct. 2017).

[3] Bross, B., Andersson, K., Bläser, M., Drugeon, V., Kim, S.-H., Lainema, J., Li, J., Liu, S., Ohm, J.-R., Sullivan, G. J., et al., "General video coding technology in responses to the joint call for proposals on video compression with capability beyond hevc," *IEEE Trans. Circuits Syst. Video Technol.* **30**, 1226–1240 (Oct. 2019).

[4] Krizhevsky, A., Sutskever, I., and Hinton, G. E., "Imagenet classification with deep convolutional neural networks," in [*Proc. Adv. Neural Inf. Process. Syst.*], 1097–1105 (Dec. 2012).

[5] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A., "Going deeper with convolutions," in [*Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*], 1–9 (June 2015).

[6] He, K., Zhang, X., Ren, S., and Sun, J., "Deep residual learning for image recognition," in [*Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*], 770–778 (June–Jul. 2016).
[7] Santurkar, S., Budden, D., and Shavit, N., "Generative compression," in [*Proc. IEEE Picture Coding Symp.*], 258–262 (June 2018).
[8] Agustsson, E., Tschannen, M., Mentzer, F., Timofte, R., and Van Gool, L., "Extreme learned image compression with gans," in [*Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit. Workshops*], 2587–2590 (June 2018).
[9] Kim, S., Park, J. S., Bampis, C. G., Lee, J., Markey, M. K., Dimakis, A. G., and Bovik, A. C., "Adversarial video compression guided by soft edge detection," in [*Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*], 2193–2197 (May 2020).
[10] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y., "Generative adversarial nets," in [*Proc. Adv. Neural Inf. Process. Syst.*], 2672–2680 (Dec. 2014).
[11] Ballé, J., Laparra, V., and Simoncelli, E. P., "End-to-end optimized image compression," in [*Proc. Int. Conf. on Learning Representations*], 1–27 (Apr. 2017).
[12] Theis, L., Shi, W., Cunningham, A., and Huszár, F., "Lossy image compression with compressive autoencoders," *arXiv:1703.00395* (2017).
[13] Chen, Z. and He, T., "Learning based facial image compression with semantic fidelity metric," *Neurocomputing* **338**, 16–25 (Apr. 2019).
[14] Wang, R., Sun, Z., and Kamata, S.-i., "Adaptive image compression using gan based semantic-perceptual residual compensation," in [*2020 25th International Conference on Pattern Recognition (ICPR)*], 9030–9037, IEEE (2021).
[15] Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A., "Image-to-image translation with conditional adversarial networks," in [*Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*], 1125–1134 (Jul. 2017).
[16] Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A., "Unpaired image-to-image translation using cycle-consistent adversarial networks," in [*Proc. Int. Conf. Comput. Vision*], 2223–2232 (Oct. 2017).
[17] Jin, Z., An, P., Yang, C., and Shen, L., "Quality enhancement for intra frame coding via cnns: An adversarial approach," in [*Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*], 1368–1372 (Apr. 2018).
[18] Mathieu, M., Couprie, C., and LeCun, Y., "Deep multi-scale video prediction beyond mean square error," in [*Proc. Int. Conf. on Learning Representations*], 1–14 (May 2016).
[19] Vondrick, C., Pirsiavash, H., and Torralba, A., "Generating videos with scene dynamics," in [*Proc. Adv. Neural Inf. Process. Syst.*], 613–621 (Dec. 2016).
[20] Zhou, Y. and Berg, T. L., "Learning temporal transformations from time-lapse videos," in [*Proc. Eur. Conf. Comput. Vision*], 262–277 (Oct. 2016).
[21] Saito, M., Matsumoto, E., and Saito, S., "Temporal generative adversarial nets with singular value clipping," in [*Proc. IEEE Int. Conf. Comput. Vision*], 2830–2839 (Oct. 2017).
[22] Liang, X., Lee, L., Dai, W., and Xing, E. P., "Dual motion gan for future-flow embedded video prediction," in [*Proc. IEEE Int. Conf. Computer Vision*], 1744–1752 (Oct. 2017).
[23] Tulyakov, S., Liu, M.-Y., Yang, X., and Kautz, J., "Mocogan: Decomposing motion and content for video generation," in [*Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*], 1526–1535 (June 2018).
[24] Lin, J., Liu, D., Li, H., and Wu, F., "Generative adversarial network-based frame extrapolation for video coding," in [*Proc. IEEE Vis. Commun. and Image Process.*], 1–4 (Dec. 2018).
[25] Mahmud, T., Billah, M., and Roy-Chowdhury, A. K., "Multi-view frame reconstruction with conditional gan," in [*2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*], 1164–1168, IEEE (2018).
[26] Schuldt, C., Laptev, I., and Caputo, B., "Recognizing human actions: a local svm approach," in [*Proc. Int. Conf. Pattern Recognit.*], 32–36 (Aug. 2004).
[27] Wang, T.-C., Liu, M.-Y., Zhu, J.-Y., Tao, A., Kautz, J., and Catanzaro, B., "High-resolution image synthesis and semantic manipulation with conditional gans," in [*Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*], 8798–8807 (June 2018).

[28] Yang, R., Mentzer, F., Van Gool, L., and Timofte, R., "Learning for video compression with hierarchical quality and recurrent enhancement," in [*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*], (2020).

[29] Xue, T., Chen, B., Wu, J., Wei, D., and Freeman, W. T., "Video enhancement with task-oriented flow," *International Journal of Computer Vision (IJCV)* **127**(8), 1106–1125 (2019).

[30] Wang, Z., Simoncelli, E. P., and Bovik, A. C., "Multiscale structural similarity for image quality assessment," in [*The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*], **2**, 1398–1402, Ieee (2003).