Generative Video Compression with a Transformer-Based Discriminator[‡]

Pengli Du[*], Ying Liu[*], Nam Ling[*], Yongxiong Ren[†], Lingzhi Liu[†]
[*]Department of Computer Science and Engineering, Santa Clara University, Santa Clara, CA 95053 USA
[†]Kwai Inc., Palo Alto, CA 94306 USA
Emails: {pdu, yliu15, nling}@scu.edu, {yongxiongren, l.liu}@kwai.com

***Abstract*—Deep learning has been successfully applied to image and video compression. Specifically, generative adversarial network (GAN) can compress images at low bit rates with sharp details and high perceptual quality. In this work, we propose a novel generative video compression (GVC) model with a transformer-based discriminator (TD), which learns non-local correlations within video frames to improve adversarial training. Besides, our GVC model incorporates a new loss to train the generator, which combines a base loss, a discriminator-dependent feature loss, and a perceptual loss. Experiments on HEVC test sequences demonstrate that the proposed GVC model provides superior performance at extremely low bit rates, compared to existing learned and traditional video coding schemes.**

***Index Terms*—Generative adversarial network (GAN), learned video compression, perceptual quality, transformer.**

I. Introduction

With the growing popularity of video streaming, the requirement for effective video coding (VC) schemes has risen exponentially. In the past decades, various VC standards were developed, such as H.265 [1] and versatile video coding (VVC) [2]. Nevertheless, traditional video codecs are handcrafted and are not able to be end-to-end optimized.

Recently, deep learning schemes are applied to VC. In [3], it's proposed to predict target frames using CNN-based auto-encoders. Habibian *et al.* proposed to learn context interaction for VC through 3D convolutions [4]. Lu *et al.* developed a CNN-based deep video compression (DVC) approach [5] that jointly optimizes motion and residual compression modules. Afterwards, the learned VC methods in hierarchical ways [6], [7] and recurrent learned video compression (RLVC) [8] are put forward to compress frames with a larger group-of-pictures (GOP) size and have achieved state-of-the-art rate-distortion performances. Although the aforementioned VC methods have demonstrated effectiveness, their decoded frames often suffer from blur at low bit rates, due to the use of mean-squared error (MSE) as the loss. Recently, GAN draws much attention in the field of image coding [9], since it can preserve sharper and more detailed textures compared to non-adversarial learning methods, especially at low bit rates. A natural next-step development is to extend it to GAN-based VC [10], [11]. Lately, perceptual learned video compression (PLVC) [12] integrates adversarial learning into a learned VC system. It has achieved state-of-the-art perceptual quality in learned VC and outperforms pure CNN- and RNN-based schemes. The discriminator in [12] is a CNN- and RNN-based classifier. Though effective in catching local features, it doesn't model long-distance dependencies and extract non-local features thoroughly. The transformer [13] was first proposed in natural language processing (NLP) to explore the non-local correlations among input sequences. Vision transformers also shows success in image classification and object detection.

Inspired by the ability of transformers in exploring non-local correlations among sequences and the potential of GAN to compress frames at extreme low bit rates, we propose a novel generative video compression (GVC) approach. Our contribution is twofold: 1) For the first time in the literature, a transformer is used in a GAN-based VC system; 2) We propose a new generator (G) loss function that not only constrains the collective pixel distortion and entropy of multiple compressed frames, but also employs a discriminator-dependent feature loss and a perceptual loss [14] to improve the perceptual quality. Experiments on HEVC test sequences reveal that GVC outperforms several state-of-the-art learned video compression approaches and the Low-Delay P (LDP) very fast and default configurations of the H.265 codec both quantitatively and qualitatively, especially at low bit rates. The paper is organized as follows: section II elaborates the proposed GVC method, section III presents the experiments and performance analysis. Section IV concludes the paper.

II. Methodology

We consider T successive video frames $\mathbf{X}_t$, $t = 1, 2, ..., T$. The first I frame $\mathbf{X}_1$ is intra-encoded and decoded as $\widehat{\mathbf{X}}_1$ using the traditional image compression approach BPG [15]. The remaining frames, $\mathbf{X}_2, ..., \mathbf{X}_T$, are P frames that are inter-encoded with our proposed GVC. Fig. 1 (a) illustrates the proposed GVC framework within three successive time slots.

A. Generator

Fig. 1 (b) shows the structure of the generator G, which includes motion estimation, recurrent motion auto-encoder (RMAE), motion compensation (MC) and recurrent residual auto-encoder (RRAE) modules. The RMAE and RRAE modules explore the temporal corrections between adjacent video frames. At time slot t, $\mathbf{X}_t \in \mathbb{R}^{H \times W \times 3}$ and $\widehat{\mathbf{X}}_{t-1} \in \mathbb{R}^{H \times W \times 3}$ are the target P frame and the decoded previous frame that serves as the reference frame. Firstly, the spatial pyramid

[‡]This work was supported in part by Kwai, Inc. under the Kwai, Inc. grant and in part by the National Science Foundation under Grant ECCS-2138635.

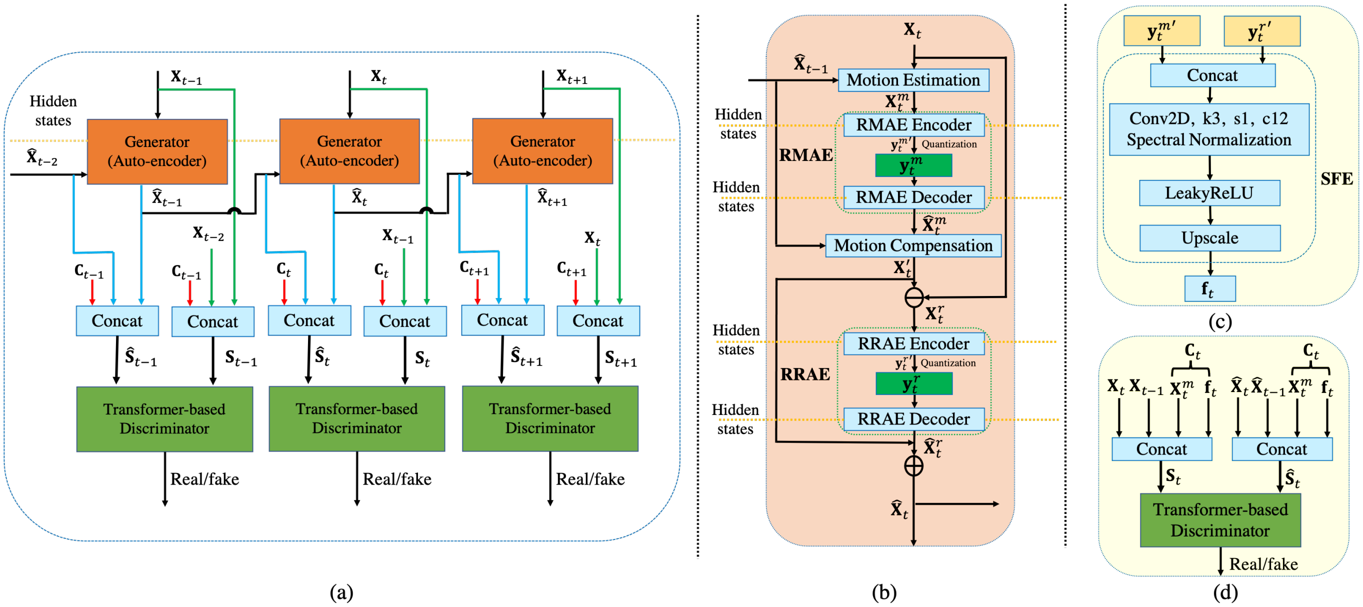

Fig. 1. (a) An illustration of the GVC framework at time slots $t-1$, t and $t+1$; (b) the structure of the G; (c) the structure of SFE; (d) the process of generating the condition $\mathbf{C}_t$, real $\mathbf{S}_t$ and fake sample $\widehat{\mathbf{S}}_t$ of the TD. k3: $kernelsize = 3$, s1: $stride = 1$, c12: $channel = 12$.

network (SPN) [16] is applied on $\mathbf{X}_t$ and $\widehat{\mathbf{X}}_{t-1}$ to estimate the motion $\mathbf{X}_t^m \in \mathbb{R}^{H\times W\times 2}$. Then it is encoded, quantized and decoded by RMAE, denoted as $\mathbf{y}_t^{m\prime}$, $\mathbf{y}_t^m$ and $\widehat{\mathbf{X}}_t^m$. The MC module takes $\widehat{\mathbf{X}}_t^m$ and $\widehat{\mathbf{X}}_{t-1}$ as the input to perform warping and convolutions to refine the output, which is the predicted target frame $\mathbf{X}_t'$. Subsequently, RRAE is applied on the residual $\mathbf{X}_t^r = \mathbf{X}_t - \mathbf{X}_t'$. It is encoded, quantized and decoded as $\mathbf{y}_t^{r\prime}$, $\mathbf{y}_t^r$ and $\widehat{\mathbf{X}}_t^r$. $\widehat{\mathbf{X}}_t^r$ is added to the motion-compensated frame $\mathbf{X}_t'$ to get the decoded P frame $\widehat{\mathbf{X}}_t$. The quantized motion and residual are encoded into bit streams by range coding.

B. *Transformer-based discriminator*

It is the first time that a transformer-based discriminator (TD) is adopted in a GAN-based VC framework. Besides, in TD, we integrate spatial-temporal side information as conditions which are essential and effective in video coding systems [12]. It is expected that the conditional GAN have the potential to generate frames with temporal consistency and rich texture, which motivates us to adopt conditions in GVC. Fig. 1 (d) depicts the real and fake inputs of the TD. The real pair $(\mathbf{X}_t, \mathbf{X}_{t-1})$ and fake pair $(\widehat{\mathbf{X}}_t, \widehat{\mathbf{X}}_{t-1})$ are each concatenated with the condition $\mathbf{C}_t$ to form the real $\mathbf{S}_t$ and fake input sample $\widehat{\mathbf{S}}_t$. The condition is consisted of the estimated temporal motion information $\mathbf{X}_t^m$, and the spatial feature $\mathbf{f}_t$ that is extracted by spacial feature extractor (SFE) as shown in Fig. 1 (c).

The structure of the TD is depicted in Fig. 2. To leverage both local and non-local features, TD starts with a convolution layer to extract features $\mathbf{Z}_t \in \mathbb{R}^{\frac{H}{16}\times\frac{W}{16}\times c}$, where c is the embedded dimension. Then $\mathbf{Z}_t$ is pre-processed to get the tensor $\mathbf{Z}_t^{f\prime}$. After going through a dropout layer, it is then processed by the 12-layer transformer blocks to get $\mathbf{Z}_t^{Trans}$. Each layer includes a multi-head self-attention (MSA) block with $h = 8$ heads and a multi-layer perceptron (MLP) block. Layer normalization (LayerNorm) is applied before each block, dropout layer and residual connection are applied after each block. After that, class token is extracted and a MLP-head (Fig. 2 bottom left) and a Sigmoid function are performed to output the class probability. The aforementioned workflow applies to both the real and fake input samples.

C. *Loss functions*

We propose to train G and TD alternately. The loss function adopted to train the G is consisted of 5 terms. The first component is the adversarial loss [17] defined as eq. (1). Minimizing it enforces the classification label of the decoded pair $(\widehat{\mathbf{X}}_t, \widehat{\mathbf{X}}_{t-1})$ to approach 1 to fool the TD. The second distortion loss eq. (2) is the MSE between the raw $\mathbf{X}_t$ and the decoded target P frame $\widehat{\mathbf{X}}_t$. The third entropy term controls the bit rates, where $\phi(\mathbf{y}_t^m)$ and $\phi(\mathbf{y}_t^r)$ are the estimated entropy of the compressed motion $\mathbf{y}_t^m$ and residual $\mathbf{y}_t^r$, as shown in eq. (3). These three terms make up the base loss as eq. (4), where λ_g, λ_d and λ_f are the hyper-parameters.

The fourth component is the feature matching loss defined as eq. (5). It's the mean absolute error (MAE) between the transformer's output $\mathbf{Z}_t^{Trans}$ and $\widehat{\mathbf{Z}}_t^{Trans}$ which are extracted from the real and fake input respectively. The last component is the perceptual loss [14] as defined in eq. (7), which measures the perceptual and semantic differences between the raw and decoded target frame. The 19-layer VGG network is adopted to extract features from 5 layers: *relu1-1*, *relu2-1*, *relu3-1*, *relu4-1* and *relu5-1*. $\mathbf{F}_{t,l}$ and $\widehat{\mathbf{F}}_{t,l}$ denote the l-th layer feature with N_l elements of the VGG-net, extracted from $\mathbf{X}_t$ and $\widehat{\mathbf{X}}_t$.

$$L_{adv}(G) = \sum_{t=2}^{T} \ln\Big(1 - D(\widehat{\mathbf{X}}_t, \widehat{\mathbf{X}}_{t-1}|\mathbf{C}_t)\Big) \tag{1}$$

$$L_d(G) = \sum_{t=2}^{T} \mathrm{MSE}(\mathbf{X}_t, \widehat{\mathbf{X}}_t) \tag{2}$$

$$L_e(G) = \sum_{t=2}^{T} \Big(\phi(\mathbf{y}_t^m) + \phi(\mathbf{y}_t^r)\Big). \tag{3}$$

$$L_{base}(G) = \lambda_g L_{adv}(G) + \lambda_d L_d(G) + \lambda_e L_e(G), \tag{4}$$

$$L_f(G) = \sum_{t=2}^{T} \mathrm{MAE}(\mathbf{Z}_t^{Trans}, \widehat{\mathbf{Z}}_t^{Trans}). \tag{5}$$

$$L_{bf}(G) = L_{base}(G) + \lambda_f L_f(G), \tag{6}$$

$$L_{VGG}(G) = \sum_{t=2}^{T} \sum_{l=1}^{5} \frac{||\mathbf{F}_{t,l} - \widehat{\mathbf{F}}_{t,l}||^2}{N_l} \tag{7}$$

The overall generator loss is defined as eq. (8), where λ_v and λ_e trade-off the three components. The loss function adopted to train the TD is defined as eq. (9). Minimizing $L_{\mathrm{D}}(D)$ means that $D(\widehat{\mathbf{X}}_t, \widehat{\mathbf{X}}_{t-1}|\mathbf{C}_t)$ should approach 0 (fake label), and $D(\mathbf{X}_t, \mathbf{X}_{t-1}|\mathbf{C}_t)$ should approach 1 (real label), which can learn a discriminator that distinguishes the decoded target frames from the raw target frames.

$$L_G(G) = L_{base}(G) + \lambda_f L_f(G) + \lambda_v L_{VGG}(G), \tag{8}$$

$$L_D(D) = \sum_{t=2}^{T} \Big(-\ln\big(1 - D(\widehat{\mathbf{X}}_t, \widehat{\mathbf{X}}_{t-1}|\mathbf{C}_t)\big) - \ln\big(D(\mathbf{X}_t, \mathbf{X}_{t-1}|\mathbf{C}_t)\big)\Big). \tag{9}$$

III. Experimental studies

A. Datasets and experiment settings

Our proposed GVC model is trained on the Vimeo-90k [18] dataset that contains 91k video sequences, each having 7 consecutive and the frame resolution is 448×256. During the training, we crop the frames to 256×256. The first I frame is compressed by BPG [15] and others are 6 P frames. Hence, in training, we set $T = 7$. The hyper-parameters of the generator loss (8) are set as $\lambda_g = 0.1$, $\lambda_d = 100$, $\lambda_f = 1$ and $\lambda_v = 1$. For the entropy term, we follow [12] to adjust λ_e by using α_1 and α_2 to approach the target bit rate R_t. As shown in Table I, we set 5 levels of target bit rate R_t. During training, if the bit rate is higher than R_t, λ_e is set as α_1, otherwise, it is set to be α_2 and $\alpha_2 \ll \alpha_1$. We conduct experimental studies on the HEVC [19] test sequences (Class B, C, D and E). To prevent error propagation, we adopt the bi-directional IPPP (bi-IPPP) structure [8] in testing. A GOP has 13 frames: $\mathbf{X}_t$, $t = 1, 2, ..., 13$. The first frame $\mathbf{X}_1$ is an I frame, followed by 6 P frames $\mathbf{X}_t$, $t = 2, ..., 7$ which are compressed by forward predictive coding as proposed in Section II. Afterwards, the first I frame $\mathbf{X}_{14}$ in the next GOP, is used to conduct predictive coding for $\mathbf{X}_t$, $t = 13, 12, ..., 8$ in a backward direction. We compare our GVC model with five existing VC methods, including three state-of-the-art learned video compression schemes: PLVC [12], RLVC (PSNR) model [8], RLVC (MS-SSIM) model [8], and two configurations of the traditional H.265 video codec: the LDP very fast and the LDP default setting of x265. For fair comparison, we adopt the probability distribution model introduced in appendix 6.1 and 6.2 of [20] to estimate the entropy for GVC, PLVC [12], RLVC (PSNR) and RLVC (MS-SSIM) model [8].

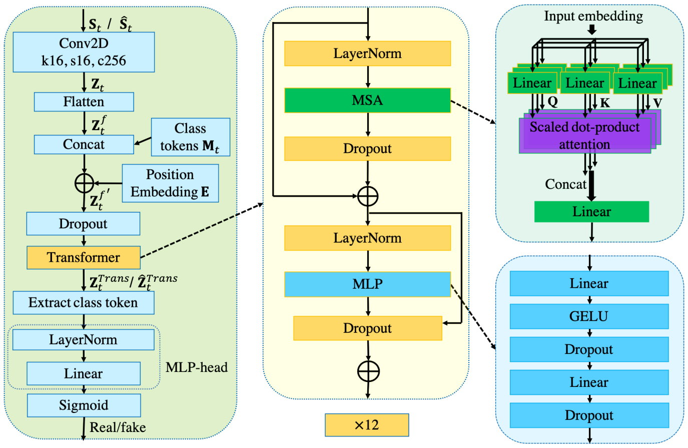

Fig. 2. The transformer-based discriminator (TD). k16: $kernelsize = 16$, s16: $stride = 16$, c256: $channel = 256$.

B. Evaluation metrics

The performance is evaluated subjectively by the visual quality of the decoded frames and quantitatively by perceptual quality metrics: Fréchet Inception Distance (FID) [21] and Kernel Inception Distance (KID) [22]. They evaluate the similarity between the distributions of the raw frames and the decoded frames, and have been validated to be effective for evaluating perceptual quality [23].

C. Performance analysis

The quantitative results, FID and KID curves against various bit rates, are shown in Fig. 3 (a) (1) and (2). We observe that GVC achieves the best FID and KID scores at almost all bit rates. At low bit rate range (<0.1 bpp), it significantly outperforms all other methods. The GAN-based video coding method PLVC also outperforms two RLVC models which do not unitize adversarial learning, however, PLVC is still worse than the x265 (LDP very fast) and x265 (LDP default).

Fig. 3 (b), (c) and (d) show the enlarged spatial textures of the decoded frames. Compared to the other five schemes, GVC exhibits excellent performance at extreme low bit rates (0.036 bpp to 0.067 bpp), by decoding richer photo-realistic textures.

x265 (LDP very fast), x265 (LDP default), and RLVC (MS-SSIM) model require much higher bit rates (1.25× to 2.43×) than GVC. However, their decoded frames are quite blurry and noisy, such as the leaves and walls in Fig. 3 (b), the girl's face in Fig. 3 (c) and the eye area of the basketball player in Fig. 3 (d). In contrast, the decodings of GVC are much clearer. It preserves fidelity to the ground-truth frames. The visual quality of the decoded frames produced by the RLVC (PSNR) model and the PLVC model is similar to that of the proposed GVC, but they require higher bit rates (1.03× to 1.90×). For example, in Fig. 3 (c), the RLVC (PSNR) model requires 1.89× bit rates, while PLVC requires 1.14× bit rates of GVC. Additionally, the HEVC test sequences include fast-motion videos (Fig. 3 (d)) and slow-motion sequences (Fig. 3 (b)) which indicate the generalization ability of GVC model.

TABLE I
HYPER-PARAMETERS OF THE ENTROPY LOSS.

level	R_t	α_1	α_2
1	0.0025	60	0.01
2	0.0125	20	0.01
3	0.025	3	0.01
4	0.05	1	0.01
5	0.1	0.3	0.001

D. Ablation study

The ablation study is to validate the effectiveness of the proposed feature and perceptual loss in eq. (5) and (7). The GVC baseline model is trained with the base loss in eq. (4). Feature loss is added in eq. (6) to train the GVC baseline-feature model. Then we add the the perceptual loss to form our proposed GVC model. In Fig. 3 (a) (3) and (4), with the additional feature loss and perceptual loss, the FID and KID scores on HEVC test sequences are improved (GVC $<$ GVC baseline-feature $<$ GVC baseline). Each component contributes to the perceptual quality, especially the VGG-based perceptual loss in eq. (7), which demonstrates the effectiveness of our proposed generator loss in eq. (8).

IV. CONCLUSIONS

In this paper, we proposed GVC for low bit-rate video compression. For the first time in the literature, transformer is used as the discriminator to guide the encoding and decoding of the target frames. Besides, we propose a new generator loss to facilitate generating decoded frames that contain more texture details which are more consistent with human vision system (HVS). Compared to existing learned video compression approaches as well as the LDP very fast mode and LDP default mode of the H.265 codec, our scheme achieves significantly higher perceptual quality at low bit rates. In terms of future studies, we will investigate advanced transformer structures to further improve the quality of decoded frames.

REFERENCES

[1] G. J. Sullivan, J. R. Ohm, W. J. Han, and T. Wiegand, "Overview of the high efficiency video coding (hevc) standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1649–1668, Dec. 2012.

[2] B. Bross, Y. K. Wang, Y. Ye, S. Liu, J. Chen, G. J. Sul-livan, and J. R. Ohm, "Overview of the versatile video coding (vvc) standard and its applications," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 10, pp. 3736–3764, Aug. 2021.

[3] C. Y. Wu, N. Singhal, and P. Krahenbuhl, "Video compression through image interpolation," in *Proc. Eur. Conf. Comput. Vision*, Munich, Germany, Sep. 2018, pp. 416–431.

[4] A. Habibian, T. V. Rozendaal, J. M. Tomczak, , and T. S. Cohen, "Video compression with rate-distortion autoencoders," in *Proc. Int. Conf. Comput. Vision*, Seoul, Korea, Oct. 2019, pp. 7033-7042.

[5] G. Lu, W. Ouyang, D. Xu, X. Zhang, C. Cai, and Z. Gao, "Dvc: An end-to-end deep video compression framework," in *Proc. IEEE/CVF Conf. Comput. Vision and Pattern Recognit.*, Long Beach, CA, USA, Jun. 2019, pp. 11006–11015.

[6] R. Yang, F. Mentzer, L. V. Gool, and R. Timofte, "Learning for video compression with hierarchical quality and recurrent enhancement," in *Proc. IEEE/CVF Conf. Comput. Vision and Pattern Recognit.*, Virtual, Jun. 2020, pp. 6628–6637.

[7] Y. Liu, P. Du, and Y. Li. "Hierarchical motion-compensated deep network for video compression", in *Proc. SPIE 11730, Big Data III: Learning, Analytics, and Applications*, vol. 117300J (12 Apr. 2021); doi: 10.1117/12.2586459.

[8] R. Yang, F. Mentzer, L. V. Gool, and R. Timofte, "Learning for video compression with recurrent autoencoder and recurrent probability model," *IEEE Trans. Selected Topics in Signal Process.*, vol. 15, no. 2, pp. 388–401, Dec. 2020.

[9] S. Iwai, T. Miyazaki, Y. Sugaya, and S. Omachi, "Fidelity-controllable extreme image compression with generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vision and Pattern Recognit.*, Virtual, Jan. 2021, pp. 8235–8242.

[10] P. Du, Y. Liu, N. Ling, L. Liu, Y. Ren, and M. Hsu. "A generative adversarial network for video compression." in *Proc. SPIE 12097, Big Data IV: Learning, Analytics, and Applications*, 120970E (31 May 2022); doi: 10.1117/12.2618714.

[11] N. Ling, C.-C. J. Kuo, G. J. Sullivan, D. Xu, S. Liu, H. Hang, W. Peng, and J. Liu, "The Future of Video Coding," *APSIPA Trans. on Signal and Inf. Process.*, vol. 11, issue 1, pp. 1-29, Jun. 2022.

[12] R. Yang, R. Timofte, and L. V. Gool, "Perceptual video compression with recurrent conditional gan," in *Proc. Int. Joint Conf. on Artificial Intelligence (IJCAI)*, Messe Wien, Vienna, Austria, Jul. 2022.

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, Long Beach, CA, USA, Dec. 2017, pp. 5998–6008.

[14] J. Johnson, A. Alahi, and F. Li, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vision*, Las Vegas, NV, USA, Oct. 2016, pp. 694–711.

[15] F. Bellard, "Bpg image format," https://bellard.org/bpg/, 2018.

[16] A. Ranjan and M. J. Black, "Optical flow estimation using a spatial pyramid network," in *Proc. IEEE/CVF Conf. Comput. Vision and Pattern Recognit.*, Honolulu, Hawaii, USA, Jun. 2017, pp. 4161–4170.

[17] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, Montreal, CA, Dec. 2014, pp. 2672–2680.

[18] T. Xue, B. Chen, J. Wu, D. Wei, and W. T. Freeman, "Video enhancement with task-oriented flow," *IEEE Trans. Int. Comput. Vision*, vol. 127, no. 8, pp. 1106–1125, Aug. 2019.

[19] F. Bossen, "Common test conditions and software reference configurations," *JCTVC-L1100*, vol. 12, no. 7, Jan. 2013.

[20] J. Ballé, D. Minnen, S. Singh, S. J. Hwang, and N. Johnston, "Variational image compression with a scale hyperprior," in *Proc. Int. Conf. on Learning Representat.*, Vancouver, CA, May 2018.

[21] M. Heusel, H. Ramsauer, T. Unterthiner, B. Unterthiner, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *Proc. Adv. Neural Inf. Process. Syst.*, Long Beach, CA, USA, Dec. 2017.

[22] M. Bińkowski, D. J. Sutherland, M. Arbel, and A. Gretton, "Demystifying mmd gans," in *Proc. Int. Conf. on Learning Representat.*, Montreal, CA, Dec. 2018.

[23] F. Mentzer, G. D. Toderici, M. Tschannen, and E. Agustsson, "High-fidelity generative image compression," in *Proc. Adv. Neural Inf. Process. Syst.*, Virtual, Jan. 2020.

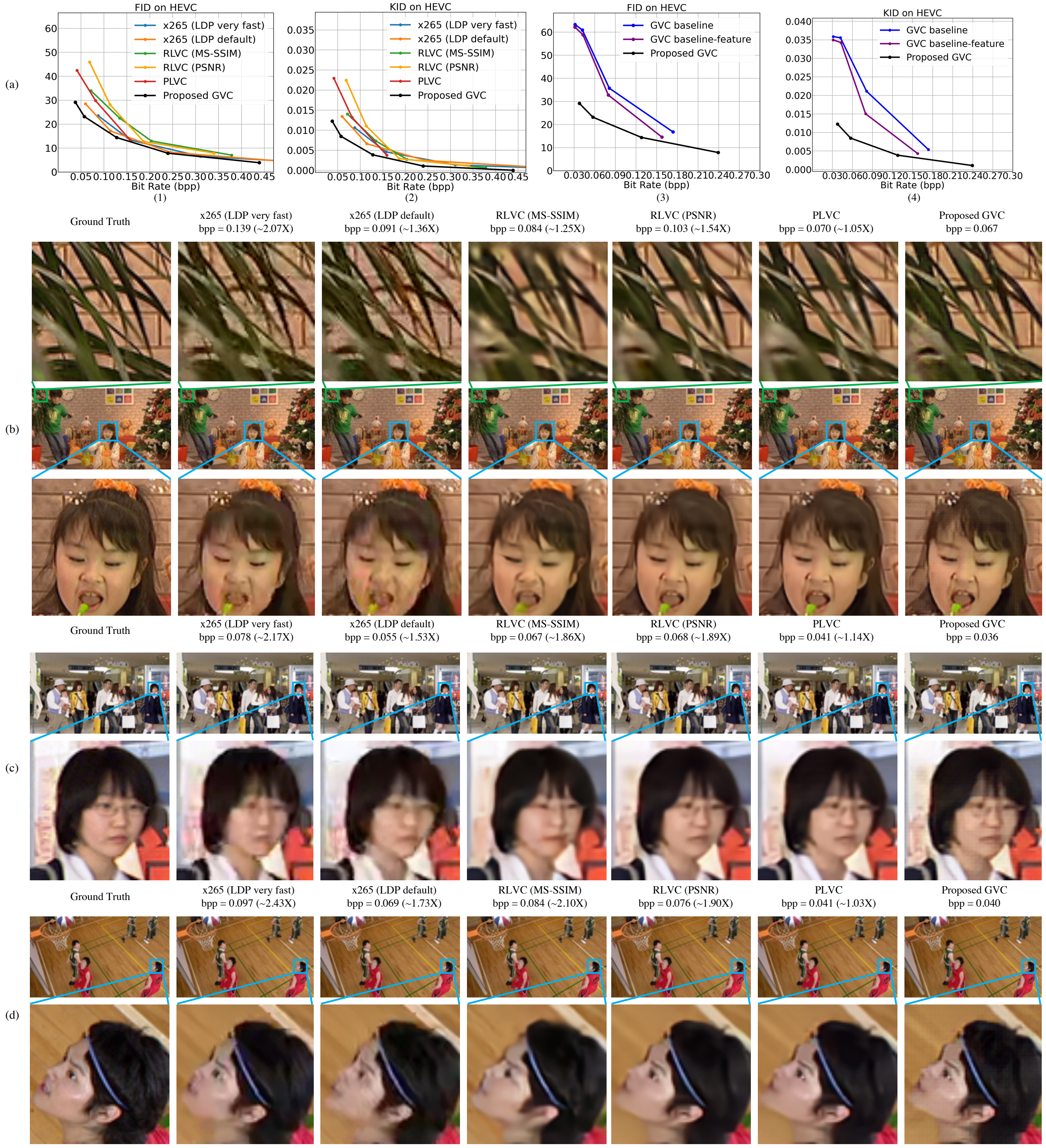

Fig. 3. (a) Rate-distortion FID and KID curves of the HEVC test sequences (Class B, C, D and E) for comparison studies (1) (2) and ablation study (3) (4). Lower FID and KID scores indicate better performance. (b) (c) (d) The visual results of the compared methods from *PartyScene*, *BQMall* and *BasketballDrill* sequences.