CGVC-T: Contextual Generative Video Compression with Transformers

Pengli Du, *Student Member, IEEE*, Ying Liu[†], *Member, IEEE*, Nam Ling, *Life Fellow, IEEE*

***Abstract*—With the high demands for video streaming, recent years have witnessed a growing interest in utilizing deep learning for video compression. Most existing neural video compression approaches adopt the predictive residue coding framework, which is sub-optimal in removing redundancy across frames. In addition, purely minimizing the pixel-wise differences between the raw frame and the decompressed frame is ineffective in improving the perceptual quality of videos. In this paper, we propose a contextual generative video compression method with transformers (CGVC-T), which adopts generative adversarial networks (GAN) for perceptual quality enhancement and applies contextual coding to improve coding efficiency. Besides, we employ a hybrid transformer-convolution structure in the auto-encoders of the CGVC-T, which learns both global and local features within video frames to remove temporal and spatial redundancy. Furthermore, we introduce novel entropy models to estimate the probability distributions of the compressed latent representations, so that the bit rates required for transmitting the compressed video are decreased. The experiments on HEVC, UVG, and MCL-JCV datasets demonstrate that the perceptual quality of our CGVC-T in terms of FID, KID, and LPIPS scores surpasses state-of-the-art learned video codecs, the industrial video codecs x264 and x265, as well as the official reference software JM, HM, and VTM. Our CGVC-T also offers superior DISTS scores among all compared learned video codecs.**

***Index Terms*—Contextual coding, entropy model, generative adversarial network, generative model, perceptual quality, transformers, video compression.**

I. Introduction

WITH the increasing prevalence of video streaming [1] in applications like online meetings and remote home surveillance, there is a growing demand to adopt efficient codecs that can reconstruct videos with higher quality but lower bit rates, over the band-limited Internet. Aiming at storage savings and transmission cost reduction [2], we have witnessed the development of video coding standards in the past two decades [3]–[7], such as Advanced Video Coding (AVC), High Efficiency Video Coding (HEVC), and Versatile Video Coding (VVC). Compared to these handcrafted codecs, deep-learning-based methods have shown their significance in the video coding world. They replace modules such as motion estimation, motion compression, motion compensation, residue compression, and context models in the traditional video codecs by neural networks, and jointly optimize these modules by minimizing the rate-distortion (RD) cost [8]–[14].

Nevertheless, most of these aforementioned learned video compression approaches adopt the mean squared error (MSE) [15] as the distortion loss to train the model, which results in blurred decoded frames. Considering the ability of GAN to capture and reproduce complex patterns in images, GAN [16] was utilized to improve the perceptual quality in image compression. The GAN consists of a generator and a discriminator. In compression scenarios, the generator can be structured as an auto-encoder (AE), including an encoder that compresses the image and a decoder that decompresses the image. Due to its adversarial learning nature, many GAN-based methods [17]–[20] demonstrated that the decompressed images preserve sharp details and achieve higher perceptual quality, especially at low bit rates. Then, researchers extended GAN to learning-based general video compression [21]–[27] and face video compression [28]–[31] by adopting various generative video compression auto-encoders or flexible discriminators.

Later on, the transformers, which adopt the multi-head self-attention mechanism to capture dependencies among sequential video frames, draw much attention. Several learned video codecs [26], [32], [33] demonstrated the effectiveness of transformers in extracting global correlations among frames, especially in the context model and motion compensation modules. However, it is still unclear how to utilize transformers in the encoder and decoder of the learned video codec.

Recently, contextual coding has attracted increased interest, which utilizes contextual information from the neighboring frames to enhance the coding process of the encoder and the decoder. In traditional residue coding, to compress the target frame $\mathbf{X}_t$ at time slot t, a prediction $\mathbf{X}_t^p$ is first generated from decoded reference frames, then only the residue $\mathbf{X}_t - \mathbf{X}_t^p$ is compressed, which has entropy $H(\mathbf{X}_t - \mathbf{X}_t^p)$. In contrast, contextual coding directly compresses $\mathbf{X}_t$ with conditions extracted from $\mathbf{X}_t^p$, resulting in entropy $H(\mathbf{X}_t|\mathbf{X}_t^p) \leq H(\mathbf{X}_t - \mathbf{X}_t^p)$. Hence, contextual coding is expected to achieve higher coding efficiency than residue coding [34]. The recently developed deep contextual video compression (DCVC) series [35]–[40] validated the effectiveness of this coding paradigm.

In this work, we propose a novel contextual generative video compression method with transformers (CGVC-T). Our method is a GAN-based learned video codec, aiming at improving the perceptual quality of decoded frames. Our GAN-based video codec adopts a contextual coding paradigm to efficiently explore the correlations among frames and reduce

[†]Corresponding author.

Pengli Du, Ying Liu, and Nam Ling are with the Department of Computer Science and Engineering, Santa Clara University, CA 95053, USA (email: pdu@scu.edu, yliu15@scu.edu, nling@scu.edu). This work is supported in part by the National Science Foundation under Grant ECCS-2138635 and the NVIDIA Academic Hardware Grant.

0000–0000/00$00.00 © 2021 IEEE

bit rates. To learn richer local and global features, we propose a convolution-transformer hybrid structure in the contextual encoder and decoder. Further, we propose novel probability distribution models with transformer structures for more efficient entropy coding of the motion latent representation and context latent representation. We outline the contributions of our work as follows:

- It is the first time in the literature that contextual coding is employed in a GAN-based video compression model. Our approach not only reduces the bit rates by a large margin compared to residual coding-based learned video codecs, but also improves the perceptual quality of decoded frames. It is beneficial in low bandwidth scenarios and in applications that require reconstructing video texture details.
- This is the first time that a hybrid transformer-convolution structure is adopted in a GAN-based video encoder and decoder. Such a hybrid structure learns more abundant feature representations useful to enhance the perceptual quality of decoded frames.
- Moreover, we propose novel transformer-based entropy models to estimate the conditional probability distribution parameters for the latent features. The entropy models effectively improve the coding efficiency and save more bit rates.

The rest of the paper is organized as follows. In Section II, we introduce related works. In Section III, we elaborate on our proposed CGVC-T method in detail. Section IV presents the experimental results and comparison studies. Section V reveals ablation studies, and Section VI analyzes the computational complexity. Finally, Section VII concludes the paper and highlights future research directions.

II. Related Work

A. Learned Video Compression

A typical learned video compression method, deep video compression (DVC) model [8], replaces modules like motion estimation, motion compression, motion compensation, and residue compression in traditional video codecs with convolutional neural networks (CNN) and optimizes them jointly by minimizing the RD loss. Based on DVC, DVCPro [9] further improved the RD performance by adopting an advanced entropy coding model and a fine-tuned post-processing module. Learned video compression with multi-reference frames (M-LVC) [10] leveraged multiple reference frames rather than one to assist inter-frame prediction. Recurrent learned video compression (RLVC) [11] introduced a recurrent neural network (RNN)-based auto-encoder to compress videos by utilizing hidden correlations among sequential frames. Meanwhile, hierarchical coding structures were developed to code P and B frames [12]–[14]. However, these models are trained by minimizing the MSE, which tends to yield overly smoothed frames and results in unsatisfactory perceptual quality [41].

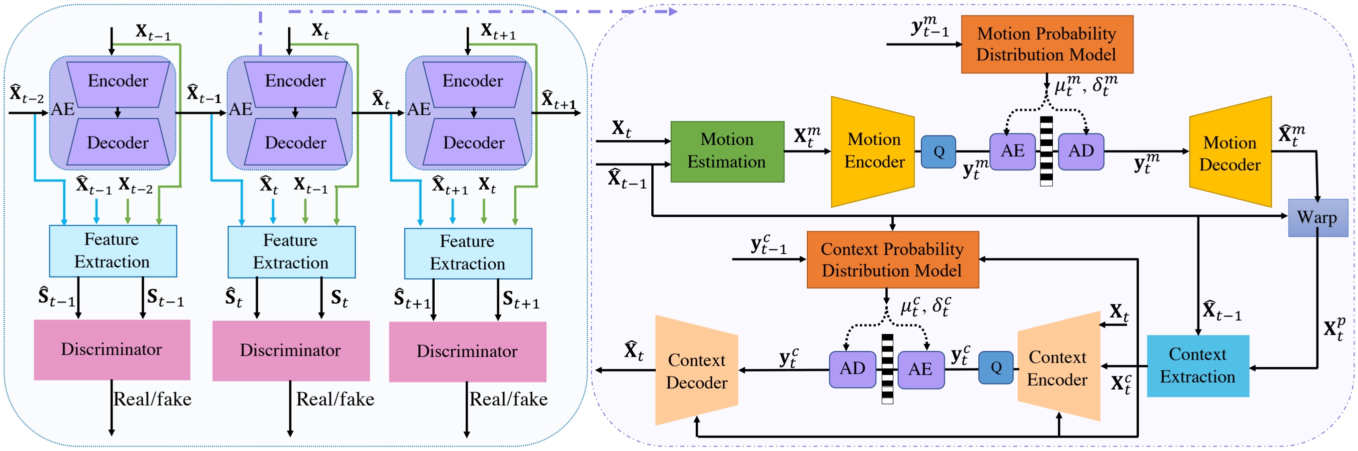
B. GAN-Based Image Compression

Recently, GAN-based image compression methods were proposed [17]–[19] to produce photo-realistic decompressed images with lower bit rates (< 0.1 bpp). Considering that conditional GAN (cGAN) can capture various patterns from data and stabilize the training, many cGAN-based methods [42], [43] adopt the conditions derived from the quantized information to assist the reconstruction of images, aiming to improve the perceptual quality at a specific bit rate. In addition, multi-scale structures [44] were used in the auto-encoder and the discriminator of GAN, which makes the framework more flexible to various contents with different resolutions at both the encoder and decoder side. Instead of using the paired encoder-decoder, fidelity-controllable extreme image compression (FC-EIC) [45] freezes the pre-trained encoder in GAN and fine-tunes a second decoder to effectively suppress undesirable noise and artifacts.

C. GAN-Based Video Compression

GAN-based video codecs were also proposed to improve the perceptual quality of decoded videos. Two discriminators were employed in [46] to handle the adversarial training of spatial and temporal information separately. Multi-level wavelet-based generative adversarial network (MW-GAN) [47] with a pyramid 2D CNN-based discriminator and MW-GAN+ [48] with a 3D CNN-based discriminator were developed to assist in the recovery of frequency information of videos in the wavelet domain by using wavelet packet transform (WPT). Besides, GAN was proposed to compress inter-frame residues [21]. A novel motion compensation approach was also adopted in a GAN [23] for detail synthesis in videos. An end-to-end deep video codec was developed with jointly optimized compression and enhancement modules (JCEVC) [24]. In particular, a dual-path generative adversarial network (DPEG) was adopted to reconstruct video details after compression. Most recently, perceptual learned video compression (PLVC) [25] proposed an RNN-based conditional discriminator, and generative video compression (GVC) [26] proposed a transformer-based conditional discriminator to enhance the perceptual quality of decoded videos by exploiting temporal information. The high visual-fidelity learned video compression (HVFVC) model [27] introduces a confidence-based feature reconstruction method to address poor restoration in newly-emerged regions and adopts a periodic compensation loss to mitigate the checkerboard artifacts. The model is trained with an RD cost that involves a GAN loss.

GAN has been adopted for not only general video compression but also face video compression, improving the perceptual quality of face videos and ensuring important facial details are preserved even at lower bit rates. For example, the Visual-Sensitivity-Based network (VSBNet) [28] proposed to compress extracted facial landmarks and adopted adversarial training to improve the realism of reconstructed frames. A multi-reference prediction network [29] was proposed for generative face video compression, which was trained with perceptual loss and adversarial loss. The compact temporal trajectory representation (CTTR) [30] proposed a spatial-temporal GAN to reconstruct high-fidelity face video frames with strong temporal consistency. The interactive face video coding framework [31] learned and compressed 3D facial representations from 2D face images, and employed GAN



(a) Framework (b) Auto-Encoder (AE)

Fig. 1. (a) The overall framework of the proposed CGVC-T method which includes an auto-encoder and a transformer-based discriminator; (b) the detailed structure of the auto-encoder at time slot t. Q: quantization. AE/AD: Arithmetic encoder/decoder.

to reconstruct high-fidelity talking face video via the motion guidance information.

D. Transformer-Based Image and Video Compression

A transformer employs the self-attention mechanism to learn long-range correlations from sequential inputs [49]. It has shown impressive performance on high-level vision tasks like image classification [50] and segmentation [51], and low-level vision tasks such as image restoration [52] and denoising [53]. Most recently, researchers started to investigate transformer-based image and video compression.

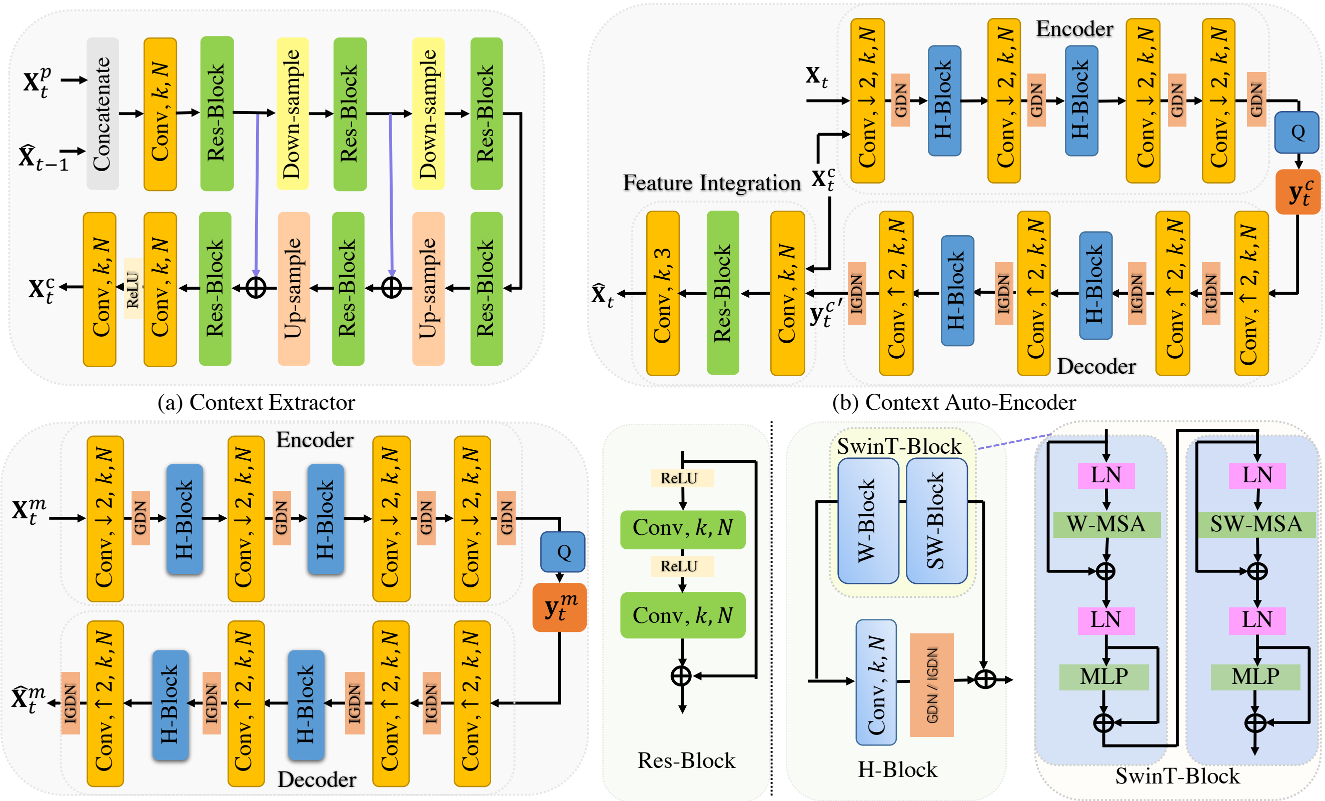
For image compression, novel transformer-based context models were proposed to explore global correlations for more efficient entropy coding, such as Entroformer [54], Contextformer [55], adaptive image compression transformer (AICT) model [56], and spatial-channel auto-regressive context model (SC-AR CM) [57]. Transformer structures were also utilized in the main image encoder and decoder [57] to enable direct interactions between all pixels in an image, facilitating the modeling of complex relationships between distant pixels. Further, transformer-convolution mixed blocks were used in the image encoder and decoder [58]–[60], which effectively combines the ability of CNN and transformers for local and non-local modeling, while maintaining controllable computational complexity. The studies in transformer-based video compression are very limited. Video compression transformer (VCT) [32] used the transformer only in the context model to leverage temporal redundancies and predict the probability distributions for entropy coding. GVC [26] proposed a transformer-based discriminator to explore non-local correlations within video frames. Motion information propagation for video compression (MIP) [33] adopted transformers to propagate previous motion information when coding the current motion latent, which effectively exploits global temporal correlation.

E. Contextual Video Coding

Contextual coding directly compresses the target frame, utilizing contextual information as the condition in the video encoder and decoder. Unlike predictive coding, it does not compress and transmit residue information [61]. Theoretically, it can achieve a lower bit rate than residual coding [34]. Among existing contextual video coding models, DCVC [35] extracts high-dimensional contexts from the feature domain. DCVC with temporal context mining (DCVC-TCM) [36] incorporates feature propagation and multi-scale temporal contexts to further improve the coding efficiency. Based on DCVC-TCM, a feature-based compression architecture [37] was proposed, aiming at generating intermediate features for various downstream human and machine vision tasks, such as video reconstruction, denoising, super-resolution, video action recognition, and video object detection. DCVC with hybrid spatial-temporal entropy model (DCVC-HEM) [38] designed a comprehensive contextual entropy model to leverage both spatial and temporal correlations and improve the prediction of probability distribution to achieve an even lower bit rate. DCVC with diverse context (DCVC-DC) [39] learned hierarchical quality patterns in the spatial domain to further boost the compression ratio. It aims to remove the temporal redundancy of videos and increase the diversity of the context model with a quadtree-based partition. Nevertheless, these methods adopt MSE as the distortion metric and contextual coding has not been utilized in GAN-based video coding which aims to improve perceptual quality. Recently, an offline and online optical flow enhancement strategy [40] was proposed. The strategy was integrated into DCVC and effectively improved the video compression performance.

III. The Proposed CGVC-T Method

Let's consider a sequence of T successive video frames, $\mathbf{X}_t, t = 0, 1, ..., T-1$. The first frame $\mathbf{X}_0$ is an I frame, compressed by BPG [62]. The remaining $T-1$ frames are P frames. We propose a GAN-based contextual generative video compression framework (CGVC-T) to compress the P frames. The CGVC-T is composed of an auto-encoder and a transformer-based discriminator. Fig. 1(a) shows the overall framework of the CGVC-T approach at successive time slots, $t-1$, t, and $t+1$. At time slot t, the auto-encoder that consists of an encoder and a decoder takes the raw frame $\mathbf{X}_t$ and its



(a) Context Extractor

(b) Context Auto-Encoder

(c) Motion Auto-Encoder

(d) Components

Fig. 2. The network structures for (a) the context extractor, (b) the context auto-encoder (CAE), (c) the motion auto-encoder (MAE), and (d) different components: Res-Block, Hybrid-Block (H-Block), and SwinT-Block [65]. The SwinT-Block consists of a regular window-based block (W-Block) and a shifted window-based block (SW-Block). k: kernel size and $k = 3$; N: channels and $N = 128$; →: 2× down-sampling; ←: 2× up-sampling. LN: Layer normalization, W-MSA: regular window-based multi-head self-attention, SW-MSA: shifted window-based multi-head self-attention, MLP: multi-layer perceptron.

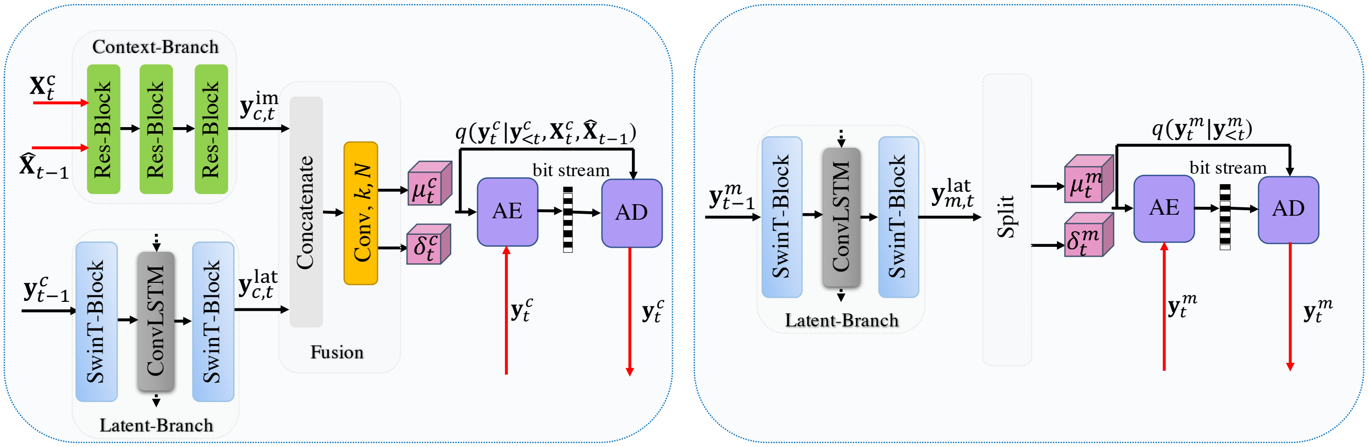
reference frame $\widehat{\mathbf{X}}_{t-1}$ decoded in the previous time slot $t-1$ as inputs, and outputs the decoded frame $\widehat{\mathbf{X}}_t$. The real sample $\mathbf{S}_t$ and fake sample $\widehat{\mathbf{S}}_t$ are then extracted from the ground-truth pair $(\mathbf{X}_{t-1}, \mathbf{X}_t)$ and decoded pair $(\widehat{\mathbf{X}}_t, \widehat{\mathbf{X}}_{t-1})$ respectively, and fed into the discriminator for distinction.

Fig. 1(b) shows the details of the AE at time step t. There are three major components: motion auto-encoder (MAE), context auto-encoder (CAE), and probability distribution model (PDM). Firstly, the motion $\mathbf{X}_t^m$ between the target frame $\mathbf{X}_t$ and its reference frame $\widehat{\mathbf{X}}_{t-1}$ are estimated by using the pyramid optical flow network (SpyNet) [63]. Then, $\mathbf{X}_t^m$ is compressed and quantized into $\mathbf{y}_t^m$ by MAE. The reconstructed motion $\widehat{\mathbf{X}}_t^m$ is utilized to warp $\widehat{\mathbf{X}}_{t-1}$ into the predicted target frame $\mathbf{X}_t^p$. Instead of using residue coding, our proposed CGVC-T adopts contextual coding to improve the coding efficiency. As shown in Fig. 1(b), the context information $\mathbf{X}_t^c$ is extracted from $\mathbf{X}_t^p$ and $\widehat{\mathbf{X}}_{t-1}$, and then fed into both the context encoder and context decoder as a condition to assist the compression of the target frame $\mathbf{X}_t$. Besides, we propose motion PDM (MPDM) and context PDM (CPDM) to predict the probability distribution parameters of the motion $\mathbf{y}_t^m$ and context $\mathbf{y}_t^c$ separately. Then, $\mathbf{y}_t^m$ and $\mathbf{y}_t^c$ are coded into bit streams by entropy coding (arithmetic coding) for transmission. Recall that our contributions are three-fold. We propose generative contextual coding, utilize the transformer-convolution hybrid structures in MAE and CAE, and adopt novel MPDM and CPDM for the distribution estimation of the latent features. The details are as follows.

A. Contextual Coding

For the first time in the literature, this work adopts contextual coding in a GAN-based video codec for coding efficiency enhancement. Unlike residue coding which uses a simple subtraction operation, contextual coding can achieve more bit rate savings [34] with the assistance of extracted contexts. Fig. 2(a) shows the context extractor that generates conditional feature $\mathbf{X}_t^c$ from $\mathbf{X}_t^p$ and the reference frame $\widehat{\mathbf{X}}_{t-1}$ by using CNN-based networks, where Res-Blocks [64] are the dominant components. Without extensive processing, the skip connections in Res-Blocks enable direct information flow among the extracted contexts and allow for the retention of critical details for videos.

Fig. 2(b) shows the structure of the CAE. The context $\mathbf{X}_t^c$ is utilized as the input in both the encoder and decoder of the CAE. In the encoder, the context $\mathbf{X}_t^c$ and raw frame $\mathbf{X}_t$ are concatenated as the inputs, and the compressed and quantized context latent representation is $\mathbf{y}_t^c$. In the decoder, after $\mathbf{y}_t^c$ is decompressed to $\mathbf{y}_t^{c\prime}$, which is the same size as the raw frame, the context $\mathbf{X}_t^c$ is concatenated with $\mathbf{y}_t^{c\prime}$ for the final reconstruction of $\widehat{\mathbf{X}}_t$ to fill in details that were discarded during the encoding process. With the employment of contextual information during the encoding and decoding process, the decoded frame $\widehat{\mathbf{X}}_t$ is expected to preserve more details and achieve higher perceptual quality.



(a) CPDM for the context latent representation $\mathbf{y}_t^c$

(b) MPDM for the motion latent representation $\mathbf{y}_t^m$

Fig. 3. The probability distribution model (PDM) of (a) the context latent representation $\mathbf{y}_t^c$ and (b) the motion latent representation $\mathbf{y}_t^m$ at time slot t. AE/AD: Arithmetic encoder/decoder. CPDM: Context probability distribution model. MPDM: Motion probability distribution model.

B. Hybrid-Block

Inspired by the Swin transformer (SwinT) [65], we propose a Hybrid-Block (H-Block) to leverage the ability of the SwinT-Block [65] to capture global correlations and the power of the Convolution (Conv) layers to gather local information. As shown in Fig. 2(b) and Fig. 2(c), the H-Block is adopted in both the CAE and MAE. Fig. 2(d) shows the detailed structure of the H-Block which consists of a Conv layer and a SwinT-Block, where the SwinT-Block [65] includes a window-based transformer block (W-Block) and a shifted window-based transformer block (SW-Block). By fusing local and global information through addition, the H-Block can extract diverse features from video frames.

Fig. 2(b) shows the detailed structure of the CAE. Inspired by [66], the encoder alternately adopts Conv blocks with 2× down-sampling and H-Blocks. The encoder compresses the concatenation of $\mathbf{X}_t$ and $\mathbf{X}_t^c$, ensuring that the compressed latent feature $\mathbf{y}_t^c$ serves as a comprehensive representation for $\mathbf{X}_t$ by incorporating detailed spatial and global patterns. The decoder is symmetric to the encoder but with 2× up-sampling transpose Conv layer. To minimize the information loss in the current frame $\mathbf{X}_t$ and recover details lost during encoding, the decompressed latent feature $\mathbf{y}_t^{c\prime}$ and context $\mathbf{X}_t^c$ are combined to further enhance the reconstruction of $\widehat{\mathbf{X}}_t$ in the final feature integration process. In addition, for training stability and normalization consistency, we adopt generalized divisive normalization (GDN) and inverse GDN (IGDN) [66]. In Fig. 2(c), similar to CAE, the encoder and decoder of MAE also adopt Conv blocks and H-Blocks alternately. The encoder encodes $\mathbf{X}_t^m$ into $\mathbf{y}_t^m$ and the decoder decodes $\mathbf{y}_t^m$ to $\widehat{\mathbf{X}}_t^m$.

C. Probability Distribution Model

To further compress successive motion $\{\mathbf{y}_t^m\}_{t=1}^{T-1}$ and context latent representations $\{\mathbf{y}_t^c\}_{t=1}^{T-1}$ into bit streams for a video, probability distribution models (PDMs) are proposed for separate entropy coding of context $\mathbf{y}_t^m$ and motion $\mathbf{y}_t^c$.

Fig. 3(a) shows the proposed two-branch CPDM for the context latent representation $\mathbf{y}_t^c$. The first Context-Branch (CB) extracts contextual features $\mathbf{y}_{c,t}^{im}$ from pixel-level inputs: the context $\mathbf{X}_t^c$ and reference frame $\widehat{\mathbf{X}}_{t-1}$. The rich texture details provided by the reference frame $\widehat{\mathbf{X}}_{t-1}$ compensates the information neglected in the extraction process of the context $\mathbf{X}_t^c$. Combining $\widehat{\mathbf{X}}_{t-1}$ and $\mathbf{X}_t^c$, CB is capable of learning more accurate temporal features. In addition, by using Res-Blocks [64], CB introduces more non-linear transforms in the entropy modeling process [67]. The second Latent-Branch (LB) captures the correlated conditions $\mathbf{y}_{c,t}^{lat}$ from the previous latent representation $\mathbf{y}_{t-1}^c$ by adopting SwinT-Block and convolutional long-short-term-memory (ConvLSTM) unit [68]. Due to the recurrent structure of ConvLSTM, $\mathbf{y}_{c,t}^{lat}$ is dependent on the previous latent representations $\mathbf{y}_{<t}^c = [\mathbf{y}_1^c, ..., \mathbf{y}_{t-1}^c]$. A simple CNN layer is used to fuse $\mathbf{y}_{c,t}^{im}$ and $\mathbf{y}_{c,t}^{lat}$ to estimate the probability distribution parameters μ_t^c, σ_t^c for $\mathbf{y}_t^c$.

Due to the temporal dependency among sequential video frames, the latent representations of successive frames are correlated. Therefore, conditioned on the context $\mathbf{X}_t^c$ and the reference frame $\widehat{\mathbf{X}}_{t-1}$ from CB, along with the previous latent representations $\mathbf{y}_{<t}^c$ from LB, the estimated probability distribution of the current latent representation $\mathbf{y}_t^c$ is expected to be more accurate.

The actual and estimated conditional probability mass functions (PMF) of $\mathbf{y}_t^c$ are denoted as $p(\mathbf{y}_t^c|\mathbf{y}_{<t}^c, \mathbf{X}_t^c, \widehat{\mathbf{X}}_{t-1})$ and $q(\mathbf{y}_t^c|\mathbf{y}_{<t}^c, \mathbf{X}_t^c, \widehat{\mathbf{X}}_{t-1})$, respectively. The bit rate of $\mathbf{y}_t^c$ can be approximated by the cross-entropy $H(p, q) = \mathbb{E}_{\mathbf{y}_t^c \sim p}[-\log_2 q(\mathbf{y}_t^c|\mathbf{y}_{<t}^c, \mathbf{X}_t^c, \widehat{\mathbf{X}}_{t-1})]$. The closer q is to p, the smaller the cross-entropy $H(p, q)$ is. Thus, when a more accurate $q(\mathbf{y}_t^c|\mathbf{y}_{<t}^c, \mathbf{X}_t^c, \widehat{\mathbf{X}}_{t-1})$ is applied to arithmetic coding to encode $\mathbf{y}_t^c$ into bit streams, the bit rate is expected to be lower. Let $\mathbf{y}_{it}^c$ be the element at the i-th 3D location of $\mathbf{y}_t^c$. Conditioned on $\mathbf{y}_{<t}^c$, $\mathbf{X}_t^c$ and $\widehat{\mathbf{X}}_{t-1}$, the joint PMF $q(\mathbf{y}_t^c|\mathbf{y}_{<t}^c, \mathbf{X}_t^c, \widehat{\mathbf{X}}_{t-1})$ can be represented as a factorized PMF

$$q(\mathbf{y}_t^c|\mathbf{y}_{<t}^c, \mathbf{X}_t^c, \widehat{\mathbf{X}}_{t-1}) = \prod_{i=1}^{N} q(\mathbf{y}_{it}^c|\mathbf{y}_{<t}^c, \mathbf{X}_t^c, \widehat{\mathbf{X}}_{t-1}), \quad (1)$$

where N is the total number of elements in $\mathbf{y}_t^c$. Following [11], we model $q(\mathbf{y}_{it}^c|\mathbf{y}_{<t}^c, \mathbf{X}_t^c, \widehat{\mathbf{X}}_{t-1})$ as discretized logistic

distribution. Due to the rounding operation in quantization, all latent elements lying in $[\mathbf{y}^c_{it}-0.5,\mathbf{y}^c_{it}+0.5)$ are quantized to $\mathbf{y}^c_{it}$, resulting in the following simplified conditional PMF

$$q(\mathbf{y}^c_{it}|\mathbf{y}^c_{<t},\mathbf{X}^c_t,\widehat{\mathbf{X}}_{t-1}) = \text{Sigmoid}(\mathbf{y}^c_{it}+0.5;\mu^c_{it},\sigma^c_{it}) - \text{Sigmoid}(\mathbf{y}^c_{it}-0.5;\mu^c_{it},\sigma^c_{it}), \tag{2}$$

where $\text{Sigmoid}(\cdot;\mu^c_{it},\sigma^c_{it})$ is the Sigmoid distribution with parameters μ^c_{it} and σ^c_{it}. These distribution parameters are estimated by our proposed CPDM.

Fig. 3(b) shows the MPDM for the motion latent representation $\mathbf{y}^m_t$, which uses the Latent-Branch only to estimate the distribution parameters of the conditional PMF $q(\mathbf{y}^m_t|\mathbf{y}^m_{<t})$. Since ConvLSTM and SwinT-Block can preserve motion dependencies, the bit rate of $\mathbf{y}^m_t$ is expected to be lower as well. We also model $q(\mathbf{y}^m_{it}|\mathbf{y}^m_{<t})$ as discretized logistic distribution, where $\mathbf{y}^m_{it}$ represents the i-th element of $\mathbf{y}^m_t$.

D. Loss Functions

The auto-encoder and the discriminator are trained alternately. Following [26], the auto-encoder is trained by minimizing the loss $\mathcal{L}_{AE}$,

$$\mathcal{L}_{AE} = \lambda_d \times \mathcal{L}_d + \mathcal{L}_f + \mathcal{L}_{vgg} + \mathcal{L}_{adv} + \mathcal{L}_{bpp}. \tag{3}$$

The distortion loss $\mathcal{L}_d$ is the MSE between the ground-truth $\mathbf{X}_t$ and the decoded $\widehat{\mathbf{X}}_t$ summed over $T-1$ P frames,

$$\mathcal{L}_d = \sum_{t=1}^{T-1} \text{MSE}(\mathbf{X}_t,\widehat{\mathbf{X}}_t). \tag{4}$$

The feature loss $\mathcal{L}_f$ (5) is the mean absolute difference (MAD) between the features $\mathbf{f}_t$ and $\widehat{\mathbf{f}}_t$, extracted from $\mathbf{X}_t$ and $\widehat{\mathbf{X}}_t$ by the discriminator. The perceptual loss $\mathcal{L}_{vgg}$ (6) [69] calculates the MSE between L layers of VGG network features extracted from the real sample $\mathbf{S}_t$ and the fake sample $\widehat{\mathbf{S}}_t$. These two training loss terms assist the model in preserving more detailed video content [26].

$$\mathcal{L}_f = \sum_{t=1}^{T-1} \text{MAD}(\mathbf{f}_t,\widehat{\mathbf{f}}_t) \tag{5}$$

$$\mathcal{L}_{vgg} = \sum_{t=1}^{T-1} \text{MSE}\big(\text{VGG}(\mathbf{S}_t),\text{VGG}(\widehat{\mathbf{S}}_t)\big) \tag{6}$$

The Wasserstein GAN (W-GAN) [70] adversarial loss $\mathcal{L}_{adv}$ is defined as

$$\mathcal{L}_{adv} = -\sum_{t=1}^{T-1} \text{D}(\widehat{\mathbf{S}}_t), \tag{7}$$

where $\text{D}(\cdot)$ denotes the discriminator, and the auto-encoder is trained to maximize $\text{D}(\widehat{\mathbf{S}}_t)$ for the fake sample $\widehat{\mathbf{S}}_t$.

The bit rate loss $\mathcal{L}_{bpp}$ (8) is measured by the entropy of both the motion and context latent representations.

$$\mathcal{L}_{bpp} = \mathcal{R}(\mathbf{y}^m_1) + \mathcal{R}(\mathbf{y}^c_1) + \sum_{t=2}^{T-1}\Big[\phi(\mathbf{y}^m_t)+\phi(\mathbf{y}^c_t)\Big] \tag{8}$$

For the first P frame $\mathbf{X}_1$, the entropy of the motion latent representation $\mathcal{R}(\mathbf{y}^m_1)$ and the entropy of the context latent representation $\mathcal{R}(\mathbf{y}^c_1)$ are calculated by using unconditional probability distribution models [66], due to the lack of reference latent representations. For the remaining $T-2$ P frames, the entropy of the context latent representation $\phi(\mathbf{y}^c_t)$ and the entropy of the motion latent representation $\phi(\mathbf{y}^m_t)$ are estimated by the proposed CPDM and MPDM, respectively.

The transformer-based discriminator is trained by maximizing loss function $\mathcal{L}_D$ (9). It aims to output higher values for real samples $\mathbf{S}_t$ and lower values for fake samples $\widehat{\mathbf{S}}_t$.

$$\mathcal{L}_D = \sum_{t=1}^{T-1}\Big[\text{D}(\mathbf{S}_t) - \text{D}(\widehat{\mathbf{S}}_t)\Big] \tag{9}$$

IV. Experimental Results

A. Setups

We train the proposed CGVC-T with Vimeo-90k [71] dataset. In each training sequence, there are $T = 7$ frames (1 I frame and 6 P frames), which are randomly cropped into frames with a 256×256 resolution. We set the hyper-parameter λ_d as 256, 512, 1024, and 2048 to achieve various bit rates. Following [36], the performance is evaluated on the HEVC [72] (Class B, C, D and E), the UVG [73], and the MCL-JCV [74] datasets. HEVC Class B and UVG datasets contain 1080p high-resolution videos, while HEVC class E and MCL-JCV datasets include 720p medium-resolution videos. For low-resolution videos, the class C and D of the HEVC datasets are 480p and 240p. During testing, we expand the group of pictures (GOP) from 7 to 13, as shown in Fig. 4. The first I frame $\mathbf{X}_0$ is followed by 6 P frames $\mathbf{X}_t$, $t = 1, 2, ..., 6$ which are compressed by forward contextual coding. Then, the first I frame $\mathbf{X}_{13}$ of the next GOP is utilized to conduct contextual coding for $\mathbf{X}_t$, $t = 12, 11, ..., 7$ in a backward direction.

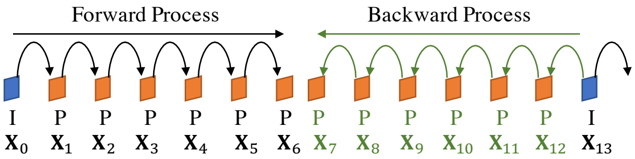


Fig. 4. An illustration of the testing group of pictures (GOP).

B. Compared Methods

We compare our CGVC-T with existing learned video compression methods, such as residue coding methods (RLVC [11], ALVC [14]), contextual coding methods (DCVC [35], DCVC-TCM [36], DCVC-HEM [38], DCVC-DC [39]), and GAN-based video coding approaches (PLVC [25], GVC [26]). RLVC [11] and DCVC series [35], [36], [38], [39] trained two models which we denote as the P model and the M model. The P model adopts the MSE as the distortion loss, which aims to achieve a higher peak signal-to-noise ratio (PSNR). The M model aims to improve the multi-scale structural similarity index (MS-SSIM), by using 1−MS-SSIM as the distortion loss. We also compare the proposed CGVC-T with traditional codecs, including the LDP-very-fast mode of x264 [75] and x265 [76], as well as the official reference software: JM 19.0 [77], HM 16.2 [78], and VTM 16.0 [79]. They are all in IPPP mode.

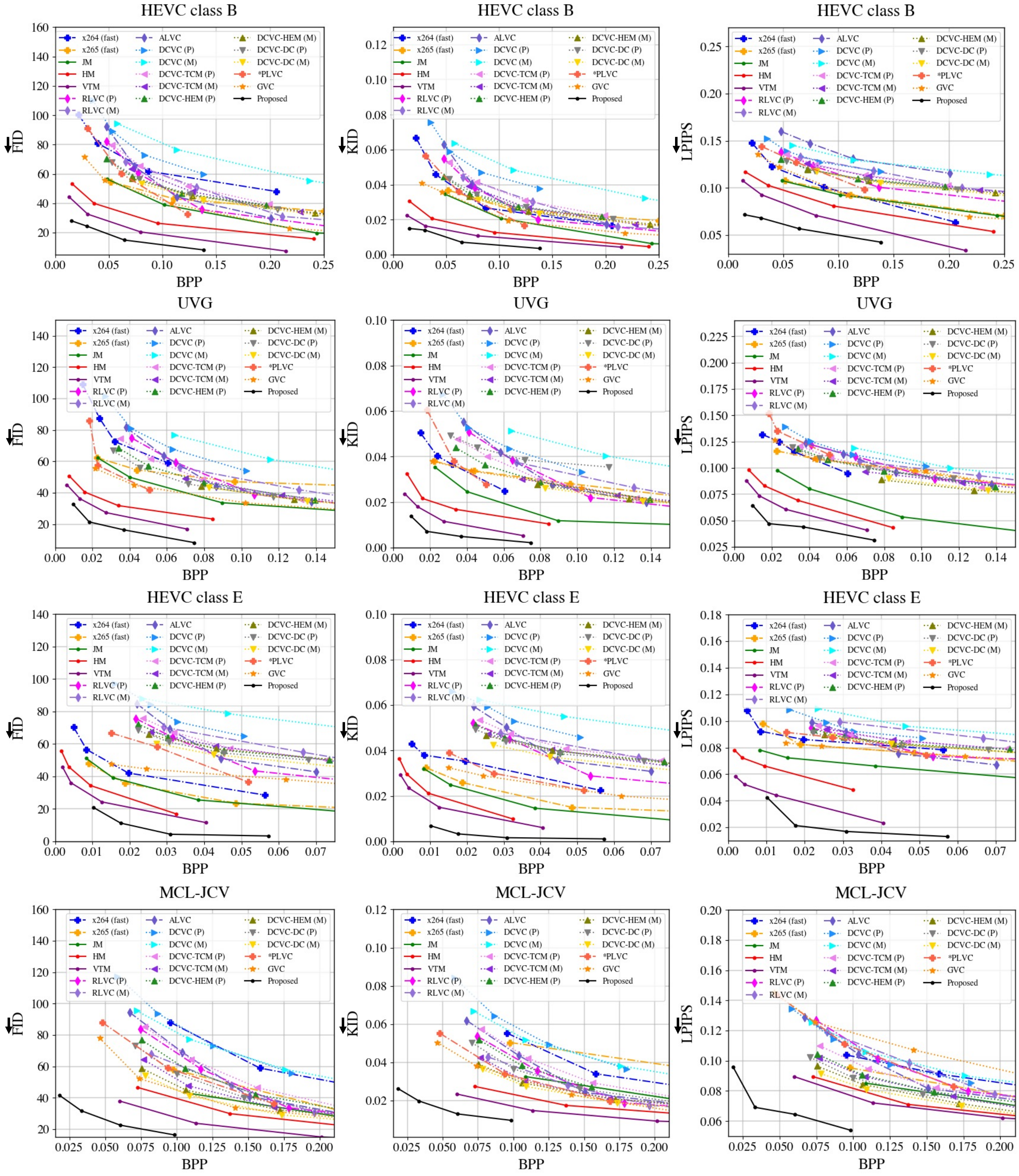


Fig. 5. The rate-distortion performance of high-resolution (1080p) videos (class B of HEVC and UVG dataset) and medium-resolution (720p) videos (class E of HEVC and MCL-JCV dataset). The distortion is measured by perceptual quality metrics: FID, KID, and LPIPS. (↓: the lower the better).

TABLE I
BD-RATE (%) IN TERMS OF FID, KID, AND LPIPS. THE ANCHOR IS VTM.
(THE BEST, SECOND-BEST, AND THIRD-BEST RESULTS ARE MARKED IN RED, GREEN, AND BLUE, RESPECTIVELY.)

BD-rate (%) in terms of FID.

Datasets	HEVC-B	HEVC-C	HEVC-D	HEVC-E	UVG	MCL-JCV	Average
x264	1947.3	1620.7	1006.8	622.5	3432.7	376.4	1501.1
x265	767.0	496.6	1485.1	317.8	2190.0	113.7	895.1
JM	282.4	314.6	391.1	276.1	486.4	113.3	310.6
HM	99.6	**48.8**	86.6	75.6	80.6	62.8	75.7
DCVC (P)	2317.6	1061.8	1086.6	12790.4	2729.6	429.7	3402.6
DCVC (M)	2058.5	1216.6	885.2	13554.1	2153.9	409.8	3379.7
DCVC-TCM (P)	1068.1	812.2	871.5	4949.4	1037.7	211.9	1491.8
DCVC-TCM (M)	763.9	560.7	590.9	3496.8	837.4	139.4	1064.9
DCVC-HEM (P)	933.8	633.2	545.8	5522.3	1039.1	187.2	1476.9
DCVC-HEM (M)	776.1	479.9	355.4	4168.7	801.7	129.9	1118.6
DCVC-DC (P)	858.2	671.6	684.1	5194.2	942.1	159.7	1418.3
DCVC-DC (M)	689.3	382.8	422.6	3250.4	753.4	101.7	933.3
RLVC (P)	430.4	451.3	373.4	1881.6	838.4	147.4	687.1
RLVC (M)	576.7	525.4	411.6	2118.7	866.0	136.2	772.4
ALVC	577.1	522.3	408.5	2782.8	946.4	156.2	898.9
PLVC	438.7	420.5	385.3	1642.7	525.8	163.9	596.2
GVC	307.5	371.4	315.9	1034.8	503.9	97.1	438.5
Proposed	**-43.1**	85.2	**-17.7**	**5.7**	**-54.8**	**-45.8**	**-11.7**

BD-rate (%) in terms of KID.

Datasets	HEVC-B	HEVC-C	HEVC-D	HEVC-E	UVG	MCL-JCV	Average
x264	718.5	2103.7	671.3	3744.9	964.6	295.6	1416.4
x265	1042.5	583.5	535.5	376.2	1834.3	460.7	805.4
JM	190.4	269.8	348.3	233.3	275.1	154.3	245.2
HM	70.9	43.1	98.4	62.2	123.2	51.9	74.9
DCVC (P)	2068.6	1522.7	4214.4	805540.0	4996.1	428.5	136461.7
DCVC (M)	2034.3	1158.8	2539.0	14861.3	3272.9	419.9	4047.7
DCVC-TCM (P)	1435.9	805.8	965.7	8307.6	1594.2	222.8	2222.0
DCVC-TCM (M)	789.7	498.1	582.3	3197.4	1256.6	136.0	1076.7
DCVC-HEM (P)	1307.5	622.2	627.0	9121.3	1587.8	202.1	2244.7
DCVC-HEM (M)	859.0	451.1	376.4	3867.9	1264.7	141.0	1160.0
DCVC-DC (P)	1193.7	671.7	821.9	10112.1	57213.7	176.4	11698.3
DCVC-DC (M)	755.7	360.1	479.7	3279.3	1168.6	113.3	1026.1
RLVC (P)	517.7	427.3	417.5	2040.7	1035.5	120.4	759.9
RLVC (M)	579.2	456.6	400.1	1837.4	1216.6	109.0	766.5
ALVC	816.3	515.9	458.3	4215.5	1338.9	149.7	1249.1
PLVC	489.3	388.0	380.4	1141.2	1249.6	129.9	629.7
GVC	277.3	309.5	370.8	761.9	863.9	61.4	440.8
Proposed	**-31.6**	**30.0**	**-20.6**	**-57.2**	**-63.4**	**-47.8**	**-31.8**

BD-rate (%) in terms of LPIPS.

Datasets	HEVC-B	HEVC-C	HEVC-D	HEVC-E	UVG	MCL-JCV	Average
x264	229.2	203243.4	824.9	10416.9	1414.4	170.3	36049.9
x265	207.9	500.2	165.7	5629.7	1672.7	148.1	1387.4
JM	204.7	328.4	296.7	3320.9	195.5	52.5	733.1
HM	83.3	127.1	118.5	787.7	83.0	13.2	202.1
DCVC (P)	1854.0	1287.1	575.3	2416360.8	2988.7	171.4	403872.9
DCVC (M)	2033.3	4255.4	12199.8	19817.9	2412.2	218.836	6822.9
DCVC-TCM (P)	1069.1	558.2	245.1	25028.9	1568.0	114.4	4763.9
DCVC-TCM (M)	853.9	866.5	611.4	7922.0	1407.7	66.7	1954.7
DCVC-HEM (P)	915.2	376.5	184.6	134261.8	1656.8	91.5	22914.4
DCVC-HEM (M)	805.0	737.2	329.0	7053.2	1188.0	42.4	1692.5
DCVC-DC (P)	850.9	315.8	142.2	209592.8	1939.3	69.6	35485.1
DCVC-DC (M)	716.7	410.6	211.0	6584.2	1226.8	29.4	1529.8
RLVC (P)	532.5	424.6	239.8	46994.0	1702.5	123.8	8336.2
RLVC (M)	836.6	1021.2	795.2	7720.0	1787.9	110.2	2045.2
ALVC	2351.0	8900.5	193.6	9268.8	1694.6	105.1	3752.3
PLVC	568.6	658.6	328.7	2999.5	2211.4	146.1	1152.1
GVC	217.5	552.1	422.5	1370.9	979.2	156.1	616.4
Proposed	**-49.9**	**57.1**	**15.8**	**-66.5**	**-74.6**	**-77.1**	**-32.5**

C. Perceptual Quality Evaluation

We adopt Fréchet Inception Distance (FID) [80], Kernel Inception Distance (KID) [81], Learned Perceptual Image Patch Similarity (LPIPS) [82], and Deep Image Structure and Texture Similarity (DISTS) [83] index as the perceptual quality evaluation metrics. They are more consistent with the human vision system (HVS) [84]. Lower FID, KID, LPIPS, and DISTS scores indicate better quality of decoded frames.

FID assesses the similarity between the distribution of the raw frames and the decoded frames. This similarity is

TABLE II
BD-RATE (%) IN TERMS OF DISTS. THE ANCHOR IS VTM.
(THE BEST, SECOND-BEST, AND THIRD-BEST RESULTS ARE MARKED IN RED, GREEN, AND BLUE, RESPECTIVELY.)

BD-rate (%) in terms of DISTS.							
Datasets	HEVC-B	HEVC-C	HEVC-D	HEVC-E	UVG	MCL-JCV	Average
DCVC (P)	363.4	686.8	920.0	1855.0	673.5	308.0	801.1
DCVC (M)	1272.0	1239.3	2162.4	7538.7	982.2	238.4	2238.8
DCVC-TCM (P)	265.8	453.6	537.9	1286.9	398.8	219.9	527.2
DCVC-TCM (M)	453.0	741.2	824.6	2358.9	-66.3	321.0	772.1
DCVC-HEM (P)	183.0	296.1	449.7	1057.0	366.7	190.2	423.8
DCVC-HEM (M)	439.3	587.9	774.1	1953.3	475.3	275.6	750.9
DCVC-DC (P)	192.4	258.0	333.8	895.6	334.0	156.6	361.7
DCVC-DC (M)	409.4	403.8	492.3	1382.1	489.1	244.7	570.2
RLVC (P)	203.1	302.1	450.2	624.1	434.2	197.2	368.5
RLVC (M)	471.3	774.7	1198.8	1749.4	611.0	341.2	857.7
ALVC	340.7	755.7	381.9	861.6	443.9	192.0	496.0
PLVC	106.7	345.3	415.1	560.1	133.2	188.8	291.5
GVC	48.9	186.4	459.8	328.0	145.7	146.6	219.5
proposed	**-53.7**	**87.5**	**153.0**	**-69.2**	**-66.5**	**72.4**	**20.6**

measured in the feature space of an Inception network, assuming that these features follow Gaussian distributions. KID is similar, but unlike FID, it does not make any assumptions about the distributions in the feature space. LPIPS measures the distance in the feature space of a deep neural network, which is adapted for predicting the similarity of distorted patches [19]. DISTS measures global texture and structure similarity between the original and decoded frames in the VGG feature space, and its hyperparameters are optimized to match the human rating of image quality.

Fig. 5 shows the RD curves in terms of FID, KID, and LPIPS for high and medium resolution videos, comparing the proposed CGVC-T with other approaches. For all bit rate ranges, our CGVC-T achieves the best FID, KID, and LPIPS scores. Although GVC outperforms PLVC and other learned video codecs, especially when the bit rate is under 0.1 bpp, GVC is worse than HM and VTM. In contrast, our CGVC-T outperforms HM and VTM in terms of FID, KID, and LPIPS. This quantitatively validates the superiority of our proposed approach. In addition, our CGVC- T can achieve similar FID, KID, and LPIPS scores with lower bit rates, compared to other methods. For instance, in Fig. 5, when the FID score is 20 for the HEVC class E dataset, VTM (0.023 bpp) and HM (0.029 bpp) need 2.3× and 2.9× the bit rates of our proposed method (0.010 bpp).

In Table I, we report the Bjøntegaard Delta bit rate (BD-rate) [85] results in terms of FID, KID, and LPIPS, while the anchor is VTM. The BD-rate measures the bit rate difference compared to the anchor VTM, where negative values indicate bitrate saving and positive values indicate bitrate increase. The averaged BD-rate result of HEVC, UVG, and MCL-JCV datasets in Table I shows that our CGVC-T achieves 11.7%, 31.8%, and 32.5% bit rate savings in terms of FID, KID, and LPIPS, respectively, when compared to the anchor VTM. Although CGVC-T needs more bit rates than VTM in low-resolution videos (e.g. HEVC class C), it has higher coding efficiency on the medium (e.g. MCL-JCV) and high resolution (e.g. UVG) videos, compared to traditional video codecs and other learned video codecs. For example, our CGVC-T shows -54.8%, -63.4%, and -74.6% BD-rates for the UVG dataset, and -45.8%, -47.8%, and -77.1% BD-rates for the MCL-JCV dataset, in terms of FID, KID, and LPIPS.

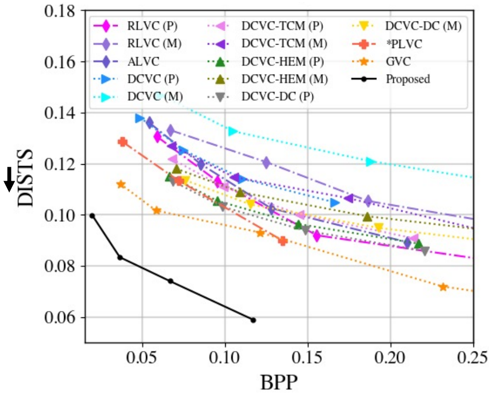
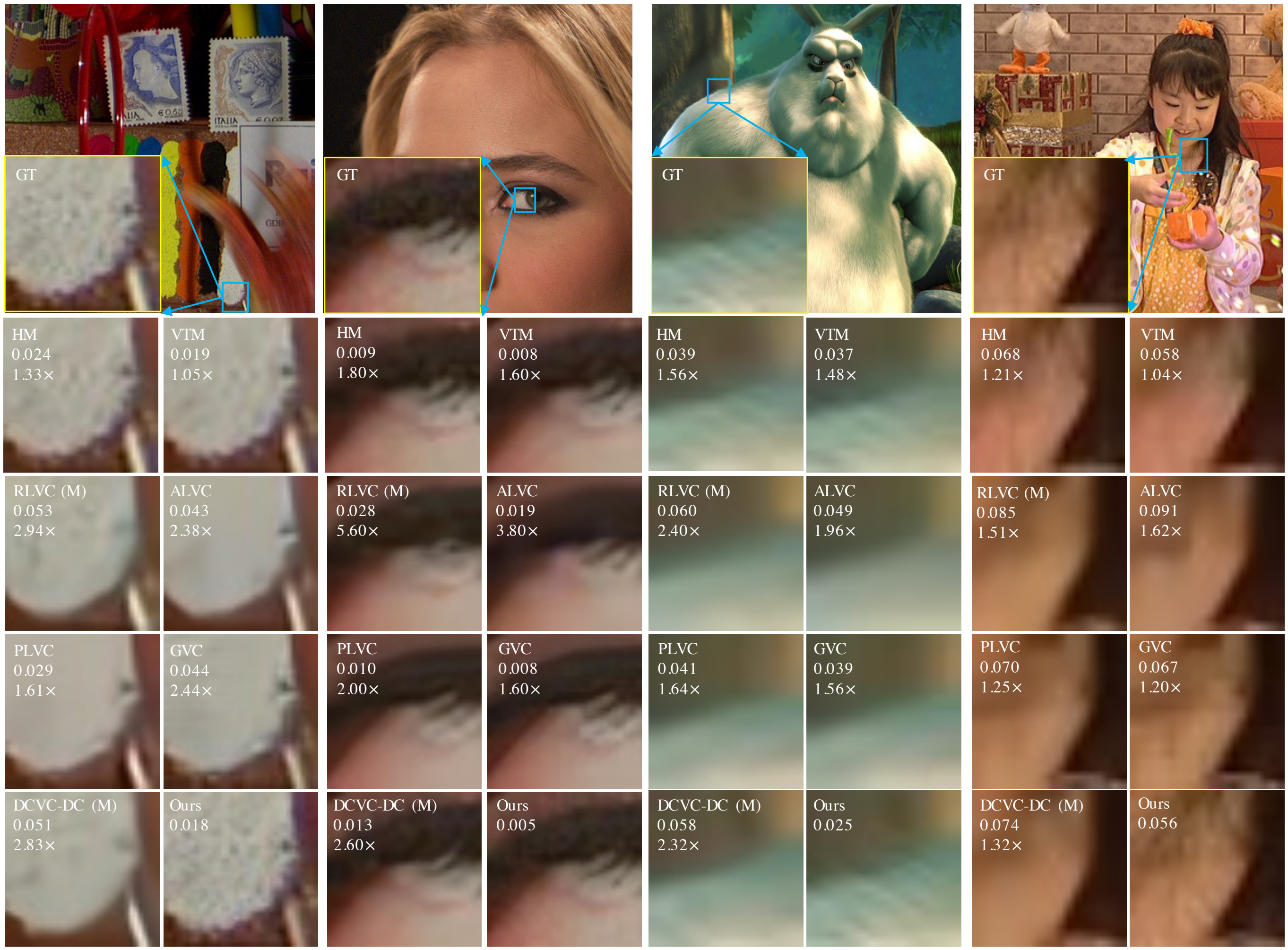


Fig. 6. The average DISTS results across HEVC, UVG, and MCL-JCV datasets for compared learned video codecs. (↓ The lower the better).

Fig. 6 shows the RD curves in terms of DISTS for our proposed CGVC-T and other learned video codecs. We observe that CGVC-T achieves the lowest DISTS scores, which verifies that it preserves texture and structure similarity better than the compared methods. Table II reports the BD-rate in terms of DISTS with VTM as the anchor. Apparently, our proposed CGVC-T achieves the best average BD-rate, while another two GAN-based methods GVC and PLVC achieve the second-best and the third-best average BD-rate, respectively. These results further demonstrate that GAN is superior in balancing bit rate deduction and improving the perceptual quality compared to non-GAN based learned video coding.

D. Objective Quality Evaluation

To show the pixel-domain fidelity, in Table III we provide the BD-rate in terms of PSNR and MS-SSIM for our proposed CGVC-T and existing learned video codecs. The anchor is VTM. We observe that our proposed CGVC-T outperforms existing GAN-based video codecs: GVC and PLVC, as well as the RNN-based ALVC and RLVC, and the very first contextual video coding scheme DCVC. Nevertheless, our proposed CGVC-T is not as competitive as the most recently



(a) HEVC-B *Cactus-1080p* (b) UVG *Beauty-1080p* (c) MCL-JCV *BigBunny-720p* (d) HEVC-C *PartyScene-480p*

Fig. 7. The visual results for high-resolution videos (class B of HEVC *Cactus-1080p* and UVG *Beauty-1080p*), medium-resolution videos (MCL-JCV dataset *BigBunny-720p*), and low-resolution videos (class C of HEVC *PartyScene-480p*), when our method is compared with traditional codecs (HM, VTM), and state-of-the-art learned video codecs (RLVC [11], ALVC [14], DCVC-DC [39], PLVC [25], GVC [26]). GT: GroundTruth. The bpp needed for each method is labeled under the name of video codecs. $f\times$: the method requires f times the bpp of our proposed method.

developed advanced contextual coding schemes DCVC-TCM, DCVC-HEM, and DCVC-DC. The primary reason is that these three approaches optimized their PSNR and MS-SSIM models using the MSE and MS-SSIM loss, respectively, while our proposed CGVC-T was optimized with a perceptual loss. The secondary reason is that these three approaches adopted more sophisticated entropy models or context mining, such as multi-scale temporal contexts or motion-aligned contexts.

E. Visual Results

Fig. 7 shows enlarged areas of decoded frames for compared methods. Compared to other schemes, our CGVC-T preserves richer texture details in the decoded frames at lower bit rates (0.005∼0.056 bpp). For example, in Fig. 7(a), RLVC, ALVC, PLVC, GVC, DCVC-DC, and the traditional video codec HM, need about 2.94×, 2.38×, 1.61×, 2.44×, 2.83×, and 1.33× the bpp of that required by our CGVC-T (0.018 bpp). However, their decoded frames are quite blurry and noisy. Although VTM has a similar bit rate (0.019 bpp) as our CGCV-T (0.018 bpp), its decoded frame retains fewer details than our proposed method. A similar phenomenon is also observed in Fig. 7(b), Fig. 7(c), and Fig. 7(d). Our proposed CGVC-T requires fewer bit rates and recovers more detailed textures in the bunny's fur area in Fig. 7(c) *BigBunny-720p*, and in the girl's hair areas in Fig. 7(d) *PartyScene-480p*. These

TABLE III
BD-RATE (%) IN TERMS OF PSNR AND MS-SSIM, AVERAGED ACROSS THE HEVC, UVG, AND MCL-JCV DATASETS. THE ANCHOR IS VTM. (THE BEST, SECOND-BEST, AND THIRD-BEST RESULTS ARE MARKED IN RED, GREEN, AND BLUE, RESPECTIVELY.)

Methods	BD-rate (%) in terms of PSNR	BD-rate (%) in terms of MS-SSIM
DCVC [35]	991.6 (P)	287.2 (M)
DCVC-TCM [36]	759.7 (P)	125.6 (M)
DCVC-HEM [38]	520.1 (P)	74.3 (M)
DCVC-DC [39]	**460.6** (P)	**54.7** (M)
RLVC [11]	1033.3 (P)	345.1 (M)
ALVC [14]	945.0	496.1
PLVC [25]	1355.2	563.4
GVC [26]	4307.6	1231.3
Proposed	**759.9**	**285.5**

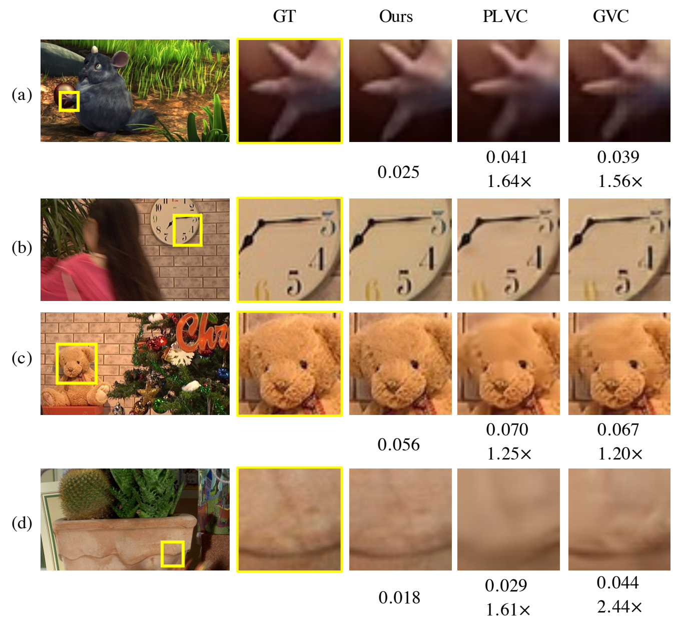


Fig. 8. The visual results for videos (a) MCL-JCV dataset *BigBunny-720p*, (b) and (c) class C of HEVC *PartyScene-480p*, and (d) class B of HEVC *Cactus-1080p*, when our method is compared with other GAN-based methods: PLVC [25] and GVC [26]. GT: GroundTruth. The bpp needed for each method is labeled under the decoded frames respectively. $f\times$: the method requires f times the bpp of our proposed method.

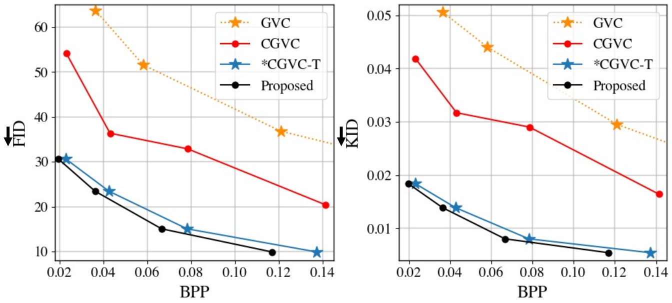


Fig. 9. The average FID and KID results across all datasets, validating the effectiveness of our proposed contextual coding, H-Block design, and PDMs in our approach. (↓: the lower the better.)

visual results demonstrate that our proposed method achieves higher visual quality compared to traditional video codecs and existing learned video codecs for videos of various resolutions.

GAN artifacts can occur when the bit rate is very low. Nevertheless, our proposed CGVC-T effectively mitigates this phenomenon. Fig. 8 shows enlarged areas of sample decoded frames with various types of textures. Fig. 8(a) shows that CGVC-T well preserves the texture of the hand of the mouse in the animation at an extremely low bit rate (0.025 bpp), while GVC and PLVC generate more artifacts or blurriness with even higher bit rates. A similar phenomenon can be observed in Fig. 8(b) for the zoomed numbers on the clock, in Fig. 8(c) for the furry bear, and in Fig. 8(d) for the edge of the flowerpot.

V. Ablation Studies

We conduct two sets of ablation studies to analyze the effectiveness of our proposed modules. The first set of ablation studies is in Sections V.A-V.C, where we validate the effectiveness of the proposed contextual coding module, H-Block, and PDMs for GAN-based video coding. The baseline is a GAN-based approach GVC [26], which adopts residue coding rather than contextual coding, uses CNN-RNN structure instead of H-Blocks in its motion and residue auto-encoders, and employs unconditional factorized probability distribution models (UCF-PDM) [66] for entropy coding instead of our proposed PDMs.

A. The Effectiveness of Contextual Coding

Based on GVC [26], we replaced residue coding with contextual coding, denoted as CGVC. As shown in Fig. 9, the FID and KID results of CGVC are lower than the baseline GVC, which verifies the effectiveness of the contextual coding that uses the context in the encoder and decoder side. On the encoder, the context conveys abundant features from previous frames to improve the compressed features of the current frame. On the decoder, the context assists in fulfilling the details that are lost in the encoding process so that the frame reconstructed from the latent representation is closer to the raw frame.

B. The Effectiveness of the Hybrid-Block

Based on CGVC, we further improve the motion and context auto-encoders with our hybrid transformer-convolution H-Block to form *CGVC-T. We can tell from Fig. 9 that *CGVC-T can achieve significantly lower FID and KID scores than CGVC. This is because the H-Block has the capability of integrating local and global features of video frames.

Besides, to verify the superiority of the H-Block over pure SWIN transformer, we remove the convolution-based lower-branch of the proposed H-Block shown in Fig. 2(d), and only keep the upper-branch SwinT-Block. The resultant model is named CGVC-SW. Table IV provides the BD-rates in terms of FID and KID, with CGVC as the baseline. While CGVC-SW achieves 4.3% and 37.8% BD-rate savings in terms of FID and KID, the H-Block-based *CGVC-T further increases the BD-rate savings to 77.2% (FID) and 84.5% (KID). This demonstrates that the design of a convolutional-transformer hybrid block is beneficial in achieving higher perceptual quality than a pure transformer structure, due to its global and local feature fusion ability.

TABLE IV
Average BD-rate (%) in terms of FID and KID across the HEVC, UVG, and MCL-JCV datasets. The anchor is CGVC.

Methods	BD-rate (%) in terms of FID	BD-rate (%) in terms of KID
CGVC-SW	-4.3	-37.8
*CGVC-T	-77.2	-84.5

C. The Effectiveness of the PDMs

Fig. 9 also shows that our CGVC-T with the proposed probability distribution models MPDM and CPDM needs less bit rate at the same FID and KID, in comparison with *CGVC-T, which adopts the UCF-PDM [66]. The entropy of the context latent representation of the *CGVC-T model is $\mathbb{E}_{\mathbf{y}_t^c \sim p}[-\log_2 q(\mathbf{y}_t^c)] = \mathbb{E}_{\mathbf{y}_t^c \sim p}[-\log_2 \prod_{i=1}^{N} q(\mathbf{y}_{it}^c)]$, and the entropy of the motion latent representation of *CGVC-T is $\mathbb{E}_{\mathbf{y}_t^m \sim p}[-\log_2 q(\mathbf{y}_t^m)] = \mathbb{E}_{\mathbf{y}_t^m \sim p}[-\log_2 \prod_{i=1}^{N} q(\mathbf{y}_{it}^m)]$.

Table V further analyzes the contributions of the Latent-Branch (LB) and Context-Branch (CB) of the two-branch CPDM. Based on *CGVC-T, we first adopted the proposed MPDM and a CPDM with only the LB to estimate the distribution parameters for $\mathbf{y}_t^m$ and $\mathbf{y}_t^c$, denoted as *CGVC-T+MPDM+CPDM-LB. With this model, the conditional entropy of $\mathbf{y}_t^c$ is $\mathbb{E}_{\mathbf{y}_t^c \sim p}[-\log_2 q(\mathbf{y}_t^c | \mathbf{y}_{<t}^c)]$. Then, we adopted both the proposed MPDM and the two-branch CPDM to estimate the distribution parameters for $\mathbf{y}_t^m$ and $\mathbf{y}_t^c$, denoted as *CGVC-T+MPDM+CPDM. With this model, the entropy of $\mathbf{y}_t^c$ is $\mathbb{E}_{\mathbf{y}_t^c \sim p}[-\log_2 q(\mathbf{y}_t^c | \mathbf{y}_{<t}^c, \mathbf{X}_t^c, \widehat{\mathbf{X}}_{t-1})]$. In this ablation study, we trained *CGVC-T, *CGVC-T+MPDM+CPDM-LB, and *CGVC-T+MPDM+CPDM (i.e. our proposed CGVC-T) with the same loss as in (3), with $\lambda_d = 256$.

The results in Table V show that LB is beneficial in bit rate reduction, and CB further decreases the bit rate. Compared to the baseline *CGVC-T, the average decreased bpp of *CGVC-T+MPDM+CPDM-LB is 5.33%. With the assistance of CB, the bpp of *CGVC-T+MPDM+CPDM was reduced more significantly (14.82%). It shows that our proposed PDMs are effective in estimating the probability distributions of the latent representations. Also, these experimental results support the theory that $\mathbb{E}_{\mathbf{y}_t^c \sim p}[-\log_2 q(\mathbf{y}_t^c | \mathbf{y}_{<t}^c, \mathbf{X}_t^c, \widehat{\mathbf{X}}_{t-1})] \leq \mathbb{E}_{\mathbf{y}_t^c \sim p}[-\log_2 q(\mathbf{y}_t^c | \mathbf{y}_{<t}^c)] \leq \mathbb{E}_{\mathbf{y}_t^c \sim p}[-\log_2 q(\mathbf{y}_t^c)]$.

TABLE V
BPP REDUCTION (%). THE BASELINE IS *CGVC-T.

Dataset	*CGVC-T+MPDM +CPDM-LB	*CGVC-T+MPDM+CPDM (The proposed CGVC-T)
HEVC-B	2.22%	7.38%
HEVC-C	4.17%	9.60%
HEVC-D	12.77%	35.77%
HEVC-E	3.5%	9.61%
UVG	5.87%	10.39%
MCL-JCV	3.44%	14.72%
Average	5.33%	14.82%

D. The Proposed Contextual Coding and PDMs for General Learned Video Codecs

This section shows the second set of ablation studies, where we validate the effectiveness of the proposed contextual coding module and PDMs in general learning-based video coding without adversarial learning. The baseline is a well-known learned video codec RLVC [11] that has a CNN-RNN structure.

First, we introduce contextual coding to RLVC and form the contextual RLVC (*CRLVC). The contextual information $\mathbf{X}_t^c$ is extracted the same way as that in Fig. 1, and is fed into the RLVC encoder and decoder. Next, we introduce our proposed PDMs, as shown in Fig. 3, to *CRLVC for entropy coding, which forms CRLVC.

Table VI shows the average BD-rate in terms of PSNR and MS-SSIM, achieved by *CRLVC and CRLVC. The anchor is the original RLVC. We observe that with the proposed contextual coding module, *CRLVC achieved 91.8% and 73.1% BD-rate savings in terms of PSNR and MS-SSIM, respectively. With the proposed probability density models, CRLVC further improves the BD-rate savings by 2.2% (PSNR) and 3.1% (MS-SSIM), respectively. These results demonstrate that our proposed modules can be generalized to universal learned video codecs.

TABLE VI
BD-RATE (%) IN TERMS OF PSNR AND MS-SSIM AVERAGED OVER THE HEVC, UVG, AND MCL-JCV DATASETS. THE ANCHOR IS RLVC.

Methods	BD-rate (%) in terms of PSNR	BD-rate (%) in terms of MS-SSIM
*CRLVC	-91.8	-73.1
CRLVC	-94.0	-76.2

VI. COMPUTATIONAL COMPLEXITY

Table VII shows the computational complexity of compared methods in terms of encoding and decoding time, measured by seconds per frame, and the multiply–accumulate-operations (MACs) of learned video codecs. Besides, we also compare the model size of the learned video codecs. Including our proposed CGVC-T, the encoding and decoding time of compared learned video codecs (RLVC, DCVC-TCM, DCVC-HEM, DCVC-DC, PLVC, and GVC) are tested on the same GPU (Nvidia Tesla V100) on 1080p videos. The runtime of traditional video codecs is tested on CPU as in previous study [9]. Table VII shows the encoding time of our approach (1.24s) is longer than other learned video codecs, but our decoding time (0.39s) is shorter. Although RLVC requires 0.37s for decoding which is less than our method, Fig. 9 indicates it needs more bit rates than our method to achieve the same perceptual quality. Existing contextual coding methods DCVC-TCM (0.47s), DCVC-HEM (0.53s), DCVC-DC (0.77s) need more decoding time than our proposed CGVC-T, due to the adoption of a complex textual-mining entropy model. For GAN-based residue coding methods, PLVC (0.39s) and GVC (0.41s) share a similar decoding speed as our method (0.39s). Although our model size is not the smallest, CGVC-T achieved the best BD-rate savings among all the learning-based video codecs, as shown in Tables I and II.

TABLE VII
COMPUTATIONAL COMPLEXITY AND MODEL SIZE COMPARISON.

Methods	Encoding time	Decoding time	MACs	Model size
HM	92.58s	0.21s	–	–
VTM	743.88s	0.31s	–	–
DCVC-TCM	0.88s	0.47s	2.9T	40.9MB
DCVC-HEM	0.99s	0.53s	3.3T	67.0MB
DCVC-DC	1.05s	0.77s	2.7T	76.0MB
RLVC	0.81s	0.37s	2.5T	163MB
PLVC	0.8s	0.39s	2.5T	163MB
GVC	0.79s	0.41s	2.5T	160MB
Proposed	1.24s	0.39s	2.9T	221MB

s: second, T: Tera operations, MB: Megabyte

VII. Conclusions

In this paper, we propose a contextual generative video compression framework with transformers. It is the first time in the literature that contextual coding has been adopted in GAN-based video coding to enhance coding efficiency. Besides, we propose a hybrid transformer-convolution structure in the motion and context auto-encoders to learn global-local features, such that decoded frames exhibit richer details. Moreover, we proposed PDMs to estimate the probability distribution parameters of the context and motion latent representations from historical data, which further reduced the bit rate. The experimental results and ablation studies demonstrated the effectiveness of our proposed method. In terms of future research, we will investigate the potential of contextual coding to compress motion fields in our GAN-based video codec. In addition, we will extend our work to a GAN-based B-frame coding architecture to further improve coding efficiency.

References

[1] C. Ozcinar, J. Cabrera and A. Smolic, "Visual attention-aware omnidirectional video streaming using optimal tiles for virtual reality," in *IEEE Journal on Emerging and Selected Topics in Circuits and Systems (JETCAS)*, vol. 9, no. 1, pp. 217-230, March 2019.

[2] K. Zhang, L. Zhang, W. -J. Chien, and M. Karczewicz, "Intra-Prediction mode propagation for video coding," in *IEEE Journal on Emerging and Selected Topics in Circuits and Systems (JETCAS)*, vol. 9, no. 1, pp. 110-121, March 2019.

[3] M. Wien, J. M. Boyce, T. Stockhammer, and W. -H. Peng, "Standardization status of immersive video coding," in *IEEE Journal on Emerg. and Selected Topics in Circuits and Systems (JETCAS)*, vol. 9, no. 1, pp. 5-17, March 2019.

[4] T. Wiegand, G. J. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the H.264/AVC video coding standard," in *IEEE Trans. Circuits Syst. Video Technol. (T-CSVT)*, vol. 13, no. 7, pp. 560–576, Jul. 2003.

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

[6] 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," in *IEEE Trans. Circuits Syst. Video Technol. (T-CSVT)*, vol. 31, no. 10, pp. 3736–3764, Aug. 2021.

[7] S. Schwarz et al., "Emerging MPEG standards for point cloud compression," in *IEEE Journal on Emerging and Selected Topics in Circuits and Systems (JETCAS)*, vol. 9, no. 1, pp. 133-148, March 2019.

[8] 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. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 11006–11015.

[9] G. Lu, X. Zhang, W. Ouyang, L. Chen, Z. Gao, and D. Xu, "An end-to-end learning framework for video compression," in *IEEE Trans. on Pattern Analysis and Machine Intelligence (T-PAMI)*, vol. 43, no. 10, pp. 3292-3308, Oct. 2021.

[10] J. Lin, D. Liu, H. Li, and F. Wu, "M-LVC: multiple frames prediction for learned video compression," in *Proc. IEEE/CVF Conf. Comput. Vision and Pattern Recognit. (CVPR)*, Seattle, WA, USA, 2020, pp. 3543-3551.

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

[12] 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. (CVPR)*, Virtual, Jun. 2020, pp. 6628–6637.

[13] 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*, 117300J (12 April 2021).

[14] R. Yang, R. Timofte, and L. Van Gool, "Advancing learned video compression with in-loop frame prediction," in *IEEE Trans. Circuits Syst. Video Technol. (T-CSVT)*, vol. 33, no. 5, pp. 2410-2423, May 2023.

[15] A. Horé and D. Ziou, "Image quality metrics: PSNR vs. SSIM," in *Proc. IEEE Int. Conf. on Pattern Recognit. (ICPR)*, Istanbul, Turkey, 2010, pp. 2366-2369.

[16] 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. (NeurIPS)*, Montreal, Quebec, Canada, Dec. 2014.

[17] E. Agustsson, M. Tschannen, F. Mentzer, R. Timofte, and L. Van Gool, "Generative adversarial networks for extreme learned image compression," in *Proc. IEEE/CVF Int. Conf. Comput. Vision (ICCV)*, Seoul, Korea, Oct. 2019, pp. 221-231.

[18] Y. Pei, Y. Liu, N. Ling, Y. Ren, and L. Liu, "An end-to-end deep generative network for low bit rate image coding," in *Proc. IEEE Int. Symposium on Circuits and Systems (ISCAS)*, Monterey, CA, USA, 2023, pp. 1-5.

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

[20] C. Jia, X. Zhang, S. Wang, S. Wang, and S. Ma, "Light field image compression using generative adversarial network-based view synthesis," in *IEEE Journal on Emerg. and Selected Topics in Circuits and Systems (JETCAS)*, vol. 9, no. 1, pp. 177-189, March 2019.

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

[22] S. Zhang, M. Mrak, L. Herranz, M. G. Blanch, S. Wan, and F. Yang, "DVC-P: deep video compression with perceptual optimizations," in *Proc. IEEE Int. Conf. on Visual Communicat. and Image Process. (VCIP)*, Munich, Germany, 2021, pp. 1-5.

[23] F. Mentzer, E. Agustsson, J. Ballé, D. Minnen, N. Johnston, and G. Toderici, "Neural video compression using gans for detail synthesis and propagation," in *Proc. IEEE/CVF Int. Conf. Comput. Vision (ICCV)*, Tel Aviv, Israel, Oct. 2022, pp. 562-578.

[24] T. Zhao, W. Feng, H. Zeng, Y. Xu, Y. Niu, and J. Liu, "Learning-based video coding with joint deep compression and enhancement," in *Proc. ACM Int. Conf. on Multimedia (ACM-MM)*, Lisbon, Portugal, Oct. 2022, pp. 3045-3054.

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

[26] P. Du, Y. Liu, N. Ling, Y. Ren, and L. Liu, "Generative video compression with a transformer-based discriminator," *Proc. IEEE Picture Coding Symposium (PCS)*, San Jose, CA, USA, 2022, pp. 349-353.

[27] M. Li, Y. Shi, J. Wang, and Y. Huang, "High visual-fidelity learned video compression," in *Proc. ACM Int. Conf. on Multimedia (ACM-MM)*, Ottawa, Canada, Oct. 2023, pp. 8057-8066.

[28] D. Feng, Y. Huang, Y. Zhang, J. Ling, A. Tang, and L. Song, "A generative compression framework For low bandwidth video conference," in *IEEE Int. Conf. on Multimedia and Expo Workshops (ICMEW)*, Shenzhen, China, 2021.

[29] Z. Wang, B. Chen, Y. Ye, and S. Wang, "Dynamic multi-reference generative prediction for face video compression," in *Proc. IEEE Int. Conf. on Image Process. (ICIP)*, Bordeaux, France, 2022, pp. 896-900.

[30] B. Chen, Z. Wang, B. Li, S. Wang, and Y. Ye, "Compact temporal trajectory representation for talking face video compression," in *IEEE Trans. Circuits Syst. Video Technol. (T-CSVT)*, vol. 33, no. 11, pp. 7009-7023, Nov. 2023.

[31] B. Chen, Z. Wang, B. Li, S. Wang, and Y. Ye, "Interactive face video coding: a generative compression framework," *arXiv preprint arXiv*: 2302.09919, 2023.

[32] F. Mentzer, G. Toderici, D. Minnen, S. Caelles, S. J. Hwang, M. Lucic. and E. Agustsson, "VCT: a video compression transformer," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, New Orleans, Louisiana, 2022.

[33] L. Qi, J. Li, B. Li, H. Li, and Y. Lu, "Motion information propagation for neural video compression," in *Proc. IEEE/CVF Conf. Comput. Vision and Pattern Recognit. (CVPR)*, Vancouver, Canada, 2023, pp. 6111-6120.

[34] T. Ladune, P. Philippe, W. Hamidouche, L. Zhang, and O. Déforges, "Optical flow and mode selection for learning-based video coding," in *Proc. IEEE Int. Workshop on Multimedia Signal Process. (MMSPW)*, Tampere, Finland, 2020, pp. 1-6.

[35] J. Li, B. Li, and Y. Lu, "Deep contextual video compression," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, Virtual, Dec. 2021, pp. 18114-18125.

[36] X. Sheng, J. Li, B. Li, L. Li, D. Liu, and Y. Lu, "Temporal context mining for learned video compression," in *IEEE Trans. on Multimedia (T-MM)*, 2022.

[37] X. Sheng, L. Li, D. Liu, and H. Li, "VNVC: a versatile neural video coding framework for efficient human-machine vision," in *IEEE Trans. on pattern analysis and machine intelligence (TPAMI)*, 2024.

[38] J. Li, B. Li, and Y. Lu, "Hybrid spatial-temporal entropy modeling for neural video compression," in *Proc. ACM Int. Conf. on Multimedia (ACM-MM)*, Lisbon, Portugal, Oct. 2022, pp. 1503-1511.

[39] J. Li, B. Li, and Y. Lu, "Neural video compression with diverse contexts," in *Proc. IEEE/CVF Conf. Comput. Vision and Pattern Recognit. (CVPR)*, Vancouver, Canada, 2023, pp. 22616-22626.

[40] C. Tang, X. Sheng, Z. Li, H. Zhang, L. Li, and D. Liu, "Offline and online optical flow enhancement for deep video compression," *arXiv preprint arXiv*: 2307.05092, 2023.

[41] A. Dosovitskiy and T. Brox, "Generating images with perceptual similarity metrics based on deep networks," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, Barcelona, Spain, Dec. 2016, pp. 658–666.

[42] M. J. Muckley, A. El-Nouby, K. Ullrich, H. Jégou, and J. Verbeek, "Improving statistical fidelity for neural image compression with implicit local likelihood models," *arXiv preprint arXiv*: 2301.11189, 2023.

[43] E. Agustsson, D. Minnen, G. Toderici, and F. Mentzer, "Multi-Realism image compression with a contextual generator," in *Proc. IEEE/CVF Conf. Comput. Vision and Pattern Recognit. (CVPR)*, Vancouver, Canada, 2023, pp. 22324-22333

[44] L. Wu, K. Huang and H. Shen, "A gan-based tunable image compression system," in *Proc. IEEE Winter Conf. on Applicat. of Comput. Vision (WACV)*, Snowmass, CO, USA, 2020, pp. 2323-2331

[45] 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. (CVPR)*, Virtual, Jan. 2021, pp. 8235–8242.

[46] V. Veerabadran, R. Pourreza, A. Habibian, and T. Cohen, "Adversarial distortion for learned video compression," in *Proc. IEEE/CVF Conf. Comput. Vision and Pattern Recognit. Workshops (CVPRW)*, Seattle, WA, USA, 2020, pp. 640-644.

[47] J. Wang, X. Deng, M. Xu, C. Chen, and Y. Song. "Multi-level wavelet-based generative adversarial network for perceptual quality enhancement of compressed video," in *Proc. Eur. Conf. Comput. Vision (ECCV)*, Glasgow, United Kingdom, Aug. 2020, pp. 405-421.

[48] J. Wang, M. Xu, X. Deng, L. Shen, and Y. Song, "MW-GAN+ for Perceptual Quality Enhancement on Compressed Video," in *IEEE Trans. Circuits Syst. Video Technol. (T-CSVT)*, vol. 32, no. 7, pp. 4224-4237, Jul. 2022.

[49] 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. (NeurIPS)*, Long Beach, CA, USA, Dec. 2017, pp. 5998–6008.

[50] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers and distillation through attention". in *Proc. Int. Conf. on Machine Learn. (PMLR)*, Virtual, Jul. 2021, pp. 10347-10357.

[51] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: transformer for semantic segmentation," in 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 2021 pp. 7242-7252.

[52] J. Liang, J. Cao, G. Sun, K. Zhang, L. V. Gool, and R. Timofte, "SwinIR: image restoration using swin transformer," in *Proc. IEEE/CVF Int. Conf. on Computer Vision Workshops (ICCVW)*, Montreal, Canada, 2021, pp. 1833-1844.

[53] K. Zhang, Y. Li, J. Liang, J. Cao, Y. Zhang, H. Tang, R. Timofte, and L. V. Gool, "Practical blind denoising via swin-conv-unet and data synthesis," *arXiv preprint arXiv*: 2203.13278, 2022.

[54] Y. Qian, X. Sun, M. Lin, Z. Tan, and R. Jin. "Entroformer: a transformer-based entropy model for learned image compression," in *Proc. Int. Conf. on Learn. Representat. (ICLR)*, virtual, 2021.

[55] A. B. Koyuncu, H. Gao, A. Boev, G. Gaikov, E. Alshina, and E. Steinbach, "Contextformer: A transformer with spatio-channel attention for context modeling in learned image compression," in *Proc. Eur. Conf. Comput. Vision (ECCV)*, Tel Aviv, Israel, Oct. 2022, pp. 447-463.

[56] A. Ghorbel, W. Hamidouche, and L. Morin, "AICT: an adaptive image compression transformer," in *Pro. IEEE Int. Conf. on Image Process. (ICIP)*, Kuala Lumpur, Malaysia, 2023, pp. 126-130.

[57] T. Shen and Y. Liu, "Learned image compression with transformers," in *Proc. SPIE 12522, Big Data V: Learning, Analytics, and Applications*, 1252207 (13 June 2023).

[58] M. Lu, P. Guo, H. Shi, C. Cao, and Z. Ma, "Transformer-based image compression," in *Proc. Data Compress. Conf. (DCC)*, Snowbird, UT, USA, 2022, pp. 469-469.

[59] J. Kim, B. Heo, and J. Lee, "Joint global and local hierarchical priors for learned image compression," in *Proc. IEEE/CVF Conf. Comput. Vision and Pattern Recognit. (CVPR)*, New Orleans, LA, USA, 2022, pp. 5992–6001.

[60] J. Liu, H. Sun and J. Katto, "Learned image compression with mixed transformer-cnn architectures," in *Proc. IEEE/CVF Conf. Comput. Vision and Pattern Recognit. (CVPR)*, Vancouver, Canada, 2023, pp. 14388-14397

[61] F. Brand, J. Seiler, and A. Kaup, "Contextual residual coding: a remedy for bottleneck problems in contextual inter frame coding," *arXiv preprint arXiv*: 2307.12864, 2023.

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

[63] 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.

[64] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[65] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vision (ICCV)*, Montreal, QC, Canada, 2021, pp. 9992-10002.

[66] J. Ballé, V. Laparra, and E. P. Simoncelli, "End-to-end optimized image compression," in *Proc. Int. Conf. on Learning Representat. (ICLR)*, Toulon, France, Apr., 2017.

[67] C. Fu, B. Du and L. Zhang, "SAR image compression based on multi-resblock and global context," in *IEEE Geoscience and Remote Sensing Letters*, vol. 20, pp. 1-5, 2023, Art no. 4002105.

[68] S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, "Convolutional lstm network: a machine learning approach for precipitation nowcasting," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2015, pp. 802–810.

[69] 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.

[70] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Pro. of Machine Learn. Research(PMLR)*, Aug. 2017, pp. 214–223.

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

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

[73] A. Mercat, M. Viitanen, and J. Vanne, "UVG dataset: 50/120fps 4K sequences for video codec analysis and development," in *Proc. ACM Multimedia Systems Conference (MSC)*, 2020, pp. 297–302.

[74] H. Wang, W. Gan, S. Hu, J. Y. Lin, L. Jin, L. Song, P. Wang, I. Katsavounidis, A. Aaron, and C.-C. J. Kuo, "MCL-JCV: a JND-based H.264/AVC video quality assessment dataset," in *Proc. IEEE Int. Conf. on Image Process. (ICIP)*, 2016, pp. 1509–1513.

[75] "x264". https://www.videolan.org/developers/x264.html

[76] "x265". https://www.videolan.org/developers/x265.html

[77] "JM". https://vcgit.hhi.fraunhofer.de/jvet/JM.git

[78] "HM". https://vcgit.hhi.fraunhofer.de/jvet/HM.git

[79] "VTM". https://vcgit.hhi.fraunhofer.de/jvet/VVCSoftware/VTM.git

[80] 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.

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

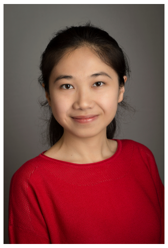
[82] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. on Comput. Vision and Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, 2018, pp. 586-595.

[83] K. Ding, K. Ma, S. Wang, and E. Simoncelli, "Image quality assessment: unifying structure and texture similarity," in *IEEE Trans. on Pattern Analysis and Machine Intelligence (T-PAMI)*, vol. 44, no. 05, pp. 2567-2581, 2022.

[84] Z. Wang, A. C. Bovik, H. R. Sheikh and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," in *IEEE Trans. on Image Process.*, vol. 13, no. 4, pp. 600-612, Apr. 2004.

[85] G. Bjontegaard, "Calculation of average PSNR differences between RD-curves," VCEG-M33, 2001.

Pengli Du is a doctoral student of the Computer Science and Engineering Department at Santa Clara University, CA, USA. She received her B.S. degree in 2016 at the Department of Information Sciences, Beijing Language and Culture University, Beijing, China, and her M.S. degree at the same university in 2019. Her research interests include image and video compression, deep learning, machine learning, pattern recognition, and computer vision. She has published papers in conferences, such as Picture Coding Symposium, and Big Data.



Ying Liu (S'11-M'13) received the B.S. degree in communications engineering from Beijing University of Posts and Telecommunications, Beijing, China, in 2006, and the M.S. and Ph.D. degrees in electrical engineering from The State University of New York, Buffalo, NY, USA, in 2008 and 2012, respectively. She is currently an Assistant Professor with the Department of Computer Science and Engineering, Santa Clara University, Santa Clara, CA, USA. Her main research interests include image and video coding, video coding for machines, deep learning, and computer vision. She serves as an Associate Editor for IEEE Transactions on Circuits and Systems for Video Technology.

Nam Ling (Life Fellow, IEEE) received the B.Eng. degree from the National University of Singapore, Singapore, in 1981, and the M.S. and Ph.D. degrees from the University of Louisiana at Lafayette, Lafayette, LA, USA, in 1985 and 1989, respectively. He was an Associate Dean for the School of Engineering from 2002 to 2010 and was the Chair of the Department of Computer Science and Engineering from 2010 to 2023, Santa Clara University, Santa Clara, CA, USA. He was the Sanfilippo Family Chair Professor and is currently the Wilmot J. Nicholson Family Chair Professor at Santa Clara University. He is/was a Chair/Distinguished/Guest and Consulting Professor with several universities internationally. He has authored or coauthored over 280 publications and seven adopted standard contributions. He has been granted with more than 20 U.S./European/PCT patents. He has delivered more than 120 invited colloquia worldwide. He is an IEEE Fellow due to his contributions to video coding algorithms and architectures. He is also an IET Fellow and an AAIA Fellow. He was named as an IEEE Distinguished Lecturer twice and was also an APSIPA Distinguished Lecturer. He was a recipient of the IEEE ICCE Best Paper Award (First Place) and the Umedia Best/Excellent Paper Award (thrice). He received six awards from Santa Clara University, four at the University Level and two at the School/College level. He was a Keynote Speaker for IEEE APCCAS, VCVP (twice), JCPC, IEEE ICAST, IEEE ICIEA, IET FC Umedia, IEEE Umedia, IEEE ICCIT, ICNLP/SSPS/CVPS, and Workshop at XUPT (twice). He has served as the General Chair/Co-Chair for IEEE Hot Chips, VCVP (twice), IEEE ICME, IEEE VCIP, IEEE SiPS, SocialSec, and Umedia (five times). He was the Honorary Co-Chair for IEEE Umedia 2017. He has also served as the Technical Program Co-Chair for IEEE ISCAS (twice), APSIPA ASC, IEEE APCCAS, IEEE SiPS (twice), DCV, and IEEE VCIP. He was the Technical Committee Chair for IEEE CASCOM TC and IEEE TCMM. He is currently the Chair of the APSIPA U.S. Chapter. He served as a Guest Editor or an Associate Editor for the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS—I: REGULAR PAPERS, the IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING, Journal of Signal Processing Systems (Springer), Multidimensional Systems and Signal Processing (Springer), and other journals. He was a Visiting Professor/Consultant/Scientist for many institutions and companies.