IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 13 March 2024, accepted 27 April 2024, date of publication 2 May 2024, date of current version 21 May 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3396407

RESEARCH ARTICLE

Semantics-Guided and Saliency-Focused Learning of Perceptual Video Compression

BINGYAO LI
Shanghai Business School, Shanghai 200235, China
e-mail: libingyao227@163.com

This work was supported by the 2024 China Society for Drug Regulation Research Project "Research and Development Incentive and Legal Protection Mechanism for Orphan Drugs and Medical Devices."

ABSTRACT In recent years, video compression has emerged as a focal point of considerable interest. Nevertheless, the predominant focus of existing methods lies in the meticulous reconstruction of videos with high fidelity, often at the expense of prioritizing the perceptual visual comfort experienced by human viewers. This paper presents an innovative learnable perceptual video compression method that extends the capabilities of current codecs. It enhances their perceptual coding proficiency by delving into the significance of local semantics and foreground objects in the context of human vision. Incorporating local semantics into the coding system involves the utilization of a region-wise contrastive learning objective, compelling the encoder to extract information pertinent to semantics. To safeguard foreground objects from corruption during compression, we prioritize minimal distortion in the foreground regions. This is achieved by employing an off-the-shelf visual saliency model for the precise detection of these regions. In an effort to augment the representation capacity of the convolution operator employed in our compression network, we introduce a recurrent information-based adaptive convolution block, thereby enhancing compression efficiency. Comprehensive experimental results validate the efficacy of our approach in achieving superior perceptual coding performance.

INDEX TERMS Video compression, perceptual quality, deep learning.

I. INTRODUCTION

The surge in video-sharing platforms and the rise of high-resolution videos have led to a remarkable increase in video data. In the present day, video content constitutes over 80% of internet traffic [1], and this percentage is anticipated to rise even more in the future. This calls for the advancement of video compression technique, which will substantially reduce the cost of video data storage and transportation. The objective of video compression is to maximize the decoded video quality, while minimizing the required bitcost, *i.e.*, the so-called rate-distortion trade-off [2].

There are many efforts on developing video compression standards and traditional codecs, such as H.264 [3], H.265 [4] and H.266 [5]. Traditional video coding methods based on block-wise transformations have undergone extensive examination, incorporating hand-crafted transformations such as Discrete Cosine Transform (DCT), as well as heuristic intra/inter predictions like angular-intra prediction and sparse motion vector-based motion compensation. The traditional codecs, equipped with progressively advancing hand-crafted modules, have achieved significant success over the past decades and have been widely deployed in the industry. Although satisfactory performances have been achieved, when measured by low-level image signal fidelity metrics such as Peak signal-to-noise ratio (PSNR) and Structural Similarity Index Metric (SSIM) [6], these traditional codecs are not superior enough when being evaluated by subjective quality metrics [7]. More seriously, the intra/inter-frame transformations within these codecs are performed in a block-wise manner, which tends to introduce block artifacts that severely degrade the subjective visual comfort.

Recently, learnable codecs [8], [9] emerge, making full use of learnable neural networks (NNs), such as convolutional

The associate editor coordinating the review of this manuscript and approving it for publication was Gangyi Jiang.

neural network (CNN) to perform the transformation. Compared to hand-crated transformations in traditional codecs, the neural network-based transformations here are learnable, data-driven, and supervised under the rate-distortion (RD) trade-off objective. This may break the performance bottleneck of traditional codes by fitting the intrinsic statistics of natural videos with flexible NNs and large-scale parameters. Further, the block artifacts are less in learnable codecs, leading to better subjective quality than traditional ones. But, the information transmitted by these trainable codecs is still not specifically chosen to highlight regions aligned with human vision. Consequently, they remain insufficiently potent in delivering images of high subjective quality.

There are few methods [7], [10], [11], [12] that are designed for perceptual coding, partially solving the above problems. JPD-SE [10] tries to tackle this problem by directly fusing the segmentation map with the images compressed by a traditional codec. Zhu et al. [11] solves this problem by using a pre-trained semantic segmentation network to extract the semantic regions and cluster their labels for enforcing more human priors. However, the hand-crafted semantic extraction scheme within the methods above is not flexible enough, and can not well handle the video contents out of the defined semantic categories. The previous methods also attempt to introduce another generative adversarial network (GAN) [13] loss, facilitating the compression procedure to produce sharp textures. For example, Yang et al. [7] introduces a recurrent conditional discriminator to judge the raw and compressed video conditioned on both spatial and temporal information, aiming to facilitate the learnable produces photo-realistic and sharp textures. Mentzer et al. [12] also follows a similar scheme, but also carefully designs the propagation of high-frequency details across different frames, leading to better results. Although the methods above improve the frame sharpness, they still do not explicitly preserve the semantic component within the videos. The loss of semantics causes the artifacts, such as the distorted object structure, still substantially reduce the human visual comfort.

In this paper, we propose a conditional perceptual video compression framework CPVC. CPVC extracts and compresses the perceptual information within the input videos, conditioning on an off-the-shelf PSNR-oriented video codec and reusing their highly efficient motion-estimation-and-motion-compensation (MEMC) scheme. Considering the human perceptual quality-related information, can be mainly decomposed into two parts, *i.e.*, (1) the semantic information such as the object structure and (2) the human vision salience regions, we also propose two learning objectives to guide the training of CVPC. Specifically, we leverage a patch-wise constrastive learning objective [14] to guide the coding procedure emphasizing the preservation of patch-wise fine-grained semantic information. To align the compression procedure better with human visual saliency, we also produce a saliency-weighted perceptual loss objective, to guide the coding procedure allocating more bits to the video regions of human interest.

Furthermore, the convolution operators in the current learnable video codecs are with fixed parameters. This may be not enough to cope with the highly flexible content within videos. Different frames of different videos exhibit various objects of different properties, requiring various convolutions of different parameters. To cope with this, we also propose a recurrent information-adaptive dynamic convolution to handle the diverse video content, which is inspired by dynamic convolution networks [15]. The dynamic convolutions are adopted in the encoder network of CPVC.

We evaluate the proposed conditional perceptual video compression method CPVC on four video compression datasets, *i.e.*, HEVC class B, HEVC class C, HEVC class D, HEVC class E datasets, proving its superiority to previous video compression methods.

The contribution of this article can be summarized as follows,

1. We propose a conditional perceptual video compression framework CPVC that is built upon off-the-shelf fidelity-based video codecs.

2. A patch-wise contrastive learning and a saliency-weighted perceptual objectives are comprehensively adopted to guide CPVC coding perceptual-related information within videos.

3. A recurrently adaptive dynamic convolution is proposed to enhance the transformation capability of the encoder network of CPVC.

4. Our approach performs favorably against the previous traditional and learnable codecs.

The remainder of the article is structured as follows: In the second section, we conduct a comprehensive and methodical examination of the pertinent scholarly works. The third section delves into the fundamental principles and intricacies of the proposed conditional perceptual video coding framework (CPVC). The fourth section presents the outcomes of our experiments and their meticulous analysis. Finally, in the fifth section, we offer some closing thoughts and conclusions.

II. RELATED WORK

In this section, a brief introduction of the related work in the field of video compression is presented, including both traditional and learnable methods. Finally, we also introduce the recently emerging methods for perceptual video compression.

A. TRADITIONAL VIDEO CODECS

Since the image codecs are the fundamental component of the video codecs, we first give a brief introduction to traditional image codecs. Within the array of coding frameworks, the fundamental techniques in both image and video compression revolve around transform and prediction methodologies. One prominent example is the JPEG [16] standard, renowned as the most widely adopted image compression standard.

The core structure of JPEG encompasses fundamental transform/prediction modules. In the JPEG compression process, the input image undergoes partitioning into non-overlapping 8 × 8 blocks. Each of these blocks undergoes transformation into the frequency domain through block-DCT (BDCT). Subsequently, the DCT coefficients for each transformed block are compressed into a binary stream, achieved through processes such as quantization and entropy coding. This orchestrated sequence of operations facilitates efficient image compression within the JPEG framework.

Over the course of several decades, traditional video compression has undergone significant development, leading to the proposal of various video coding standards. One notable standard is H.264/AVC [3], which emerged between 1999 and 2003 under the auspices of ITUT and ISO/IEC standards. This standard has witnessed remarkable success, finding widespread applications in broadcasting high-definition TV signals, internet streaming, and mobile network videos.

As video resolutions increased and parallel processing architectures became more prevalent, H.265/HEVC [4] was finalized in 2013, offering approximately 50% bitrate savings compared to H.264/AVC. The heightened compression efficiency of H.265/HEVC played a pivotal role in popularizing 4K videos with enhanced fidelity. To outcome the limited search range of H.264 and H.265 codec, there are some methods [17], [18], [19] focus on first downsampling the original video, then compressing the downsampled video with the H.264/H.265 codec, finally upsampling the compressed videos to the original resolution.

The evolution continues with H.266/VVC [5], representing the latest generation of international video coding standards. This standard aims not only to achieve substantial bitrate reduction compared to H.265/HEVC but also to address the diverse needs of current and emerging media applications. These video coding standards adhere to a cohesive hybrid framework, encompassing prediction, transform, quantization, entropy coding, and loop filtering.

B. LEARNABLE VIDEO CODECS

Since the learned video coding technique is derived from learned image coding technique, we first give a introduction to image-based learnable coding methods.

The early approaches [20], [21], [22] commonly embrace a compressive auto-encoder framework. In this setup, a nonlinear transformation is employed to generate a concise latent representation, and an entropy model is devised to estimate its probability distribution. This entropy model-constrained encoder-decoder scheme established the principled framework for later methods. To enhance the nonlinear transformation, Chen et al. [23] introduced non-local attention into the process, resulting in a more compact latent representation. Ma et al. [24] proposed a reversible wavelet-like transformation to mitigate information loss within the nonlinear transformation, which benefits the reconstruction quality of the images. Xie et al. [25] follows the similar spirit, proposing to leverage an enhanced invertible neural networks (INNs) to largely mitigate the information loss problem for better image compression. To cope with the challenge of non-differentiable operations in vector quantization, Agustsson et al. [26] presented a soft-to-hard end-to-end quantization approach. Zhang and Wu. et al. [27] presents a novel a novel Lattice Vector Quantization scheme. There are also some methods to estimate a more accurate and flexible entropy model for latent space. For example, Minnen and Singh [28] mitigate the low efficiency of spatial auto-aggressive model by using a channel-wise auto-aggressive model. Considering that the channel group number (10) is far smaller than spatial resolution of latent feature map, the encoding efficiency is substantially improved. Cheng et al. [29] and Zhu et al. [30] put forth a unified approach using a multivariate Gaussian mixture for learned image compression.

Learnable video codecs extend the learnable video codecs by introducing the inter-frame (P-frame) coding scheme. The prevailing methods typically adopt a learned hybrid video coding framework, leveraging motion-compensated prediction to anticipate the current frame. This involves compressing both the motion vector and residue. Lu et al. [8] introduced the deep video compression (DVC) framework, a notable approach within this paradigm. In DVC, optical flow is employed to represent the motion vector in the motion-compensated prediction. Subsequently, both the motion vector and residue are encoded using image compression methods. Subsequent advancements have focused on refining motion-compensated prediction in learnable video codecs. Notably, Lin et al. [31] introduced Multiple Frames Prediction for Learned Video Compression (M-LVC), which leverages multiple frames as references for improved prediction. Hu et al. [32] addressed pixel-space prediction errors by conducting motion compensation in feature-space. In their work, Liu et al. [33] proposed multiscale motion compensation to capture coarse-to-fine motion vectors. They further suggested a hybrid motion compensation approach, combining pixel-space with feature-space compensation. Agustsson et al. [34] put forth a scale-space flow representation, an intuitive extension of optical flow that introduces a scale parameter. This addition allows the network to better model uncertainty. In summary, the trend has been towards exploring more reference frames as an effective strategy to enhance compensation efficiency.

Beyond the above fully learnable codecs, there are some works [35], [36], [37] that focus on a mixed architecture, *i.e.*, enhancing the traditional codes with neural networks.

In addition to the residual coding scheme, recent approaches have delved into conditional schemes to enhance coding efficiency further. For instance, DCVC [9] introduced a novel approach leveraging conditional coding, utilizing feature domain context as a conditioning factor. Building upon this idea, recent methods such as [38] and [39] have aimed to extract more diverse temporal or spatial contexts to achieve more accurate latent distribution estimation. Our approach aligns with this contextual coding paradigm.

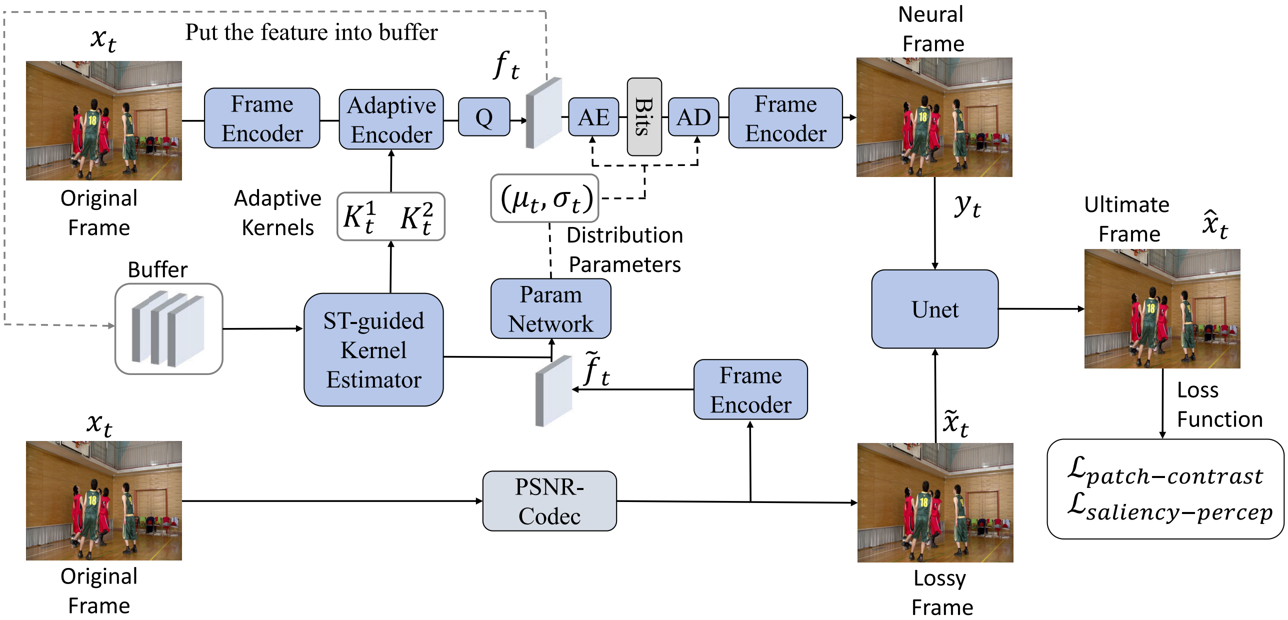

FIGURE 1. **Overview of the proposed Conditional Perceptual Video Compression (CPVC) framework. CPVC is conditioned on PSNR-oriented codecs. The primary emphasis of CPVC lies in encoding perceptual experience-related information, guided by the patch-wise contrastive learning objective $\mathcal{L}patch - contrast$ and the human saliency-weighted perceptual objective $\mathcal{L}saliency - percep$. To enhance the modeling of video data, the convolution kernel of the frame encoder is dynamic, aiming to capture past spatial-temporal information effectively. The videos compressed by the PSNR-oriented codec and neural codec are fused by a UNet to the final frame. Here, "AE" and "AD" refer to the arithmetic encoder and decoder, respectively.**

We specifically model the conditional dependency between the base pixel compression codec and the extended perceptual information compressor, as opposed to focusing on the dependency between motion coding and frame coding.

C. PERCEPTUAL VIDEO CODECS

The perceptual image compression landscape is marked by notable contributions, with HiFiC [40] emerging as a pioneering work. It distinguishes itself by integrating Generative Adversarial Networks and learned compression techniques, resulting in a state-of-the-art generative lossy compression system. In the realm of conveying perceptual features in human face images, Galteri et al. [41] present an innovative approach utilizing segmentation maps as a compact and latent representation. However, a drawback of their method lies in the inherent limitation of the hand-crafted nature of the segmentation map, rendering it less flexible. Extending the paradigm beyond face images, Change et al. [42] push the boundaries by performing perceptual coding for natural images. Their approach involves encoding visual data into compact structure and texture representations, followed by decoding in a deep synthesis fashion, with the goal of achieving superior visual reconstruction quality. Advancements in perceptual compression, namely, training learnable decoders with perceptual and GAN loss function, have further expanded to the video works [7], [43], [44], [45]. For example, Yang et al. [7] introduced a recurrent network designed to reduce inter-frame redundancy. Konuko et al. [43] proposed transporting the compact keypoint representation to drive the prediction of the next frame, with a focus on human face videos. In a departure from previous methods, our approach simultaneously emphasizes the incorporation of learnable semantic information and human visual saliency priors, leading to superior results. This distinction enables our approach to be applied effectively to videos in diverse real-world scenarios, demonstrating superior performance.

III. METHOD

The primary objective of this paper is to construct a perceptual coding framework, denoted as CPVC, which builds upon a pre-existing off-the-shelf PSNR-oriented video codec. This construction involves the integration of two perceptual visual experience-required supervision objectives and advanced video transformation modules.

A. OVERVIEW

We begin by presenting an overview of the proposed CPVC framework, as illustrated in Figure 1. Given an input video sequence $X = x_1, \ldots, x_{t-1}, x_t$, our approach initiates by employing a current lossy codec, such as HEVC, to encode the current frame x_t as $\tilde{x}t$ in P frame mode. This encoding involves prediction from the previously encoded frame $\tilde{x}_{t-1}$. Due to the compression under the PSNR target, the perceptual quality of x_t is intentionally lowered. To mitigate this, we propose extracting and transporting the perceptual information f_t from x_t. The bitcost of f_t is subsequently reduced by leveraging $\tilde{x}_t$ as a condition to estimate the parameters of the entropy model.

The perceptual feature f_t is then decoded into the perceptual frame y_t, which undergoes regularization through

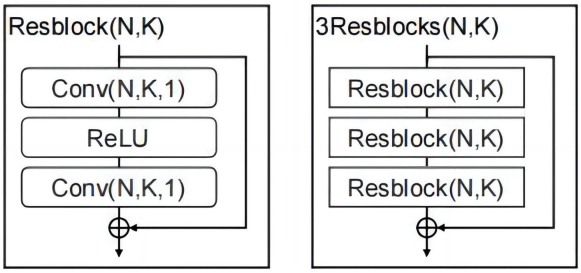

FIGURE 2. The detailed network structures of the Resblock and 3Resblock blocks. "Conv(N,K,S)" denotes the convolution operation with the output channel N, the kernel size K × K and the stride S.

the application of patch-wise contrastive learning objective $\mathcal{L}patch - contrast$ and human saliency-weighted perceptual objective $\mathcal{L}saliency - percep$. These two objectives guide the perceptual coding pathway to prioritize the extraction and transportation of crucial perceptual information over trivial textures. Finally, the perceptual frame y_t and the frame $\tilde{x}_t$ decoded from the PSNR codec are fused using a Unet architecture, yielding the final reconstructed frame $\hat{x}_t$.

To enhance the accuracy of the information extraction process, we introduce a spatial-temporal (ST)-guided kernel estimator. This estimator assimilates past spatial-temporal information to determine the optimal kernel that best fits the current frame.

In the subsequent sections, we delve into the intricate details of each component within our CPVC framework.

B. FRAME ENCODER

This network downsample the spatial resolution of the input frame by four times, which mainly consists of two convolution of kernel size 5 and stride size 2. To enhance its non-linear capability, we append each convolution with the 3Resblocks proposed in FVC [32], as shown in Figure 2. 3Resblocks is built by stacking three residual blocks (Resblock), but also adding a long residue connection for retaining the information. Each Resblock with the input variable *In* and input variable *Out* can be formulated as follows:

$$\begin{aligned}
Res &= In,\\
In_1 &= Conv(In),\\
In_2 &= ReLU(In_1),\\
In_3 &= Conv(In_2),\\
Out &= Res + In_3,
\end{aligned}\tag{1}$$

where ReLU denotes the Rectified Linear Unit [46] that suppresses the negative values, *i.e.*, $ReLU(In_1) = max(0, In_1)$.

C. ST-GUIDED KERNEL ESTIMATOR

Since the frame encoder from the previous FVC framework is designed for modeling the low-level texture patterns, instead of high-level perceptual information. To address this problem, we introduce a ST-guided kernel estimator to estimate two perceptual kernels $K_t^1 \in \mathbb{R}^{128\times3\times3}$ and $K_t^2 \in \mathbb{R}^{128\times3\times3}$ for extracting the perceptual information in a flexible and adaptive manner, as shown in Figure 3. K_t^1 and K_t^2 are group convolutions with the kernel 3 and group size 128.

Let the previous buffered perceptual feature f_{t-1} and the hidden state h_{t-1}, which contains rich perceptual dynamics of video and the current content feature $\tilde{f}_t$ extracted from the PSNR codec stream. First, we employ a ConvLSTM to aggregate the temporal dynamic information T_t:

$$\begin{aligned}
i_t &= \sigma(W_{ii} * f_{t-1} + W_{hi} * h_{t-1} + b_{ii} + b_{hi})\\
j_t &= \sigma(W_{if} * f_{t-1} + W_{hf} * h_{t-1} + b_{if} + b_{hf})\\
g_t &= \tanh(W_{ig} * f_{t-1} + W_{hg} * h_{t-1} + b_{ig} + b_{hg})\\
T_t &= \sigma(W_{io} * f_{t-1} + W_{ho} * h_{t-1} + b_{io} + b_{ho})\\
c_t &= j_t \odot c_{t-1} + i_t \odot g_t\\
h_t &= T_t \odot \tanh(c_t),
\end{aligned}\tag{2}$$

where $*$ denotes the convolution operation. W_{ii}, W_{hi}, b_{ii}, b_{hi}, W_{if}, W_{hf}, b_{if}, b_{hf}, W_{ig}, W_{hg}, b_{ig}, b_{hg}, W_{io}, W_{ho}, b_{io}, b_{ho} are the weights and biases of the convolutions. Hidden state h_t is further used by the next time step of the ConvLSTM operation.

Then, we use a three layer network, consisting of three convolutions with kernel size 3, output channel number 64 and stride size 2, followed by a average pooling operation to extract the spatial content information from $\tilde{f}_t$, producing S_t.

Further, the temporal dynamic information T_t and the spatial content information S_t are concatenated, processed by a five-layer multiple layer perceptron (MLP), producing the kernel parameters of size $2304 = 2 \times 128 \times 3 \times 3$. These parameters will be split into two parts, and reshaped as the content-adaptive kernel K_t^1 and K_t^2.

Given the preliminary visual feature z_t extracted from x_t by the frame encoder, we adopt the above kernels K_t^1 and K_t^2 to produce more high-level perceptual information f_t, which can be formulated as:

$$f_t = Q(z_t + K_t^2 * (ReLU(K_t^1 * z_t))),\tag{3}$$

where * denotes the convolution operation, Q denotes the quantization operation.

D. BITRATE ESTIMATION

We use a parameter network (Param-Net), which consists of five stacked Resblocks, followed by a convolution layer with kernel size 1 and output channel number 256, to produce the parameter of the entropy model, $\mu_t \in \mathbb{R}^{128\times h\times w}$ and $\sigma_t \in \mathbb{R}^{128\times h\times w}$. Then the entropy model is modeled as a Guaussian distribution $p(f_t) \sim \mathcal{N}(\mu_t, \sigma_t^2)$. The entropy is calculated as

$$\begin{aligned}
R(f_t) &= -log_2[CDF(f_t + 0.5) - CDF(f_t - 0.5)],\\
CDF(x) &= \frac{1}{2}\left[1 + \text{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right]\\
\text{erf}(x) &= \frac{1}{\sqrt{\pi}}\int_{-x}^{x} e^{-t^2}dt = \frac{2}{\sqrt{\pi}}\int_{0}^{x} e^{-t^2}dt,
\end{aligned}\tag{4}$$

where CDF denotes the cumulative density function of the Gaussian distribution and erf is the related error function.

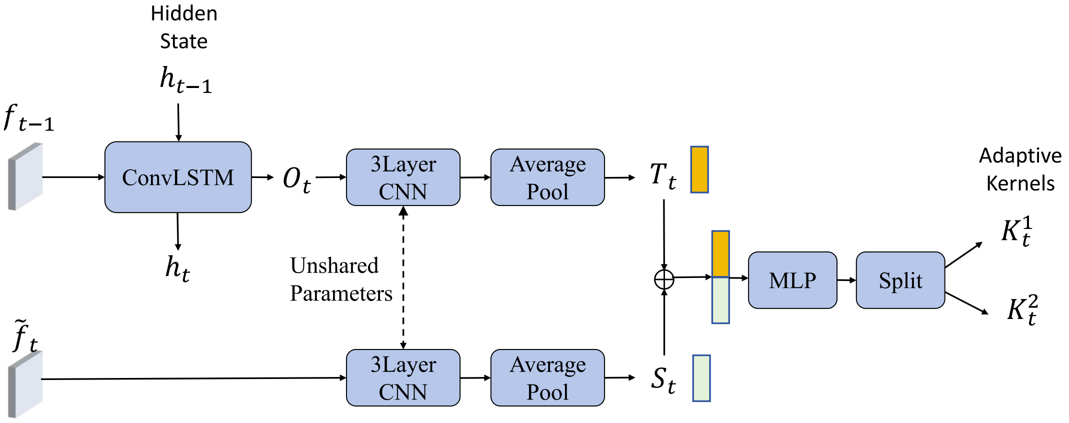

FIGURE 3. Illustration of ST-Guided Kernel Estimator, which makes full use of the past temporal and spatial information to predict the optimal convolution kernel for current frame.

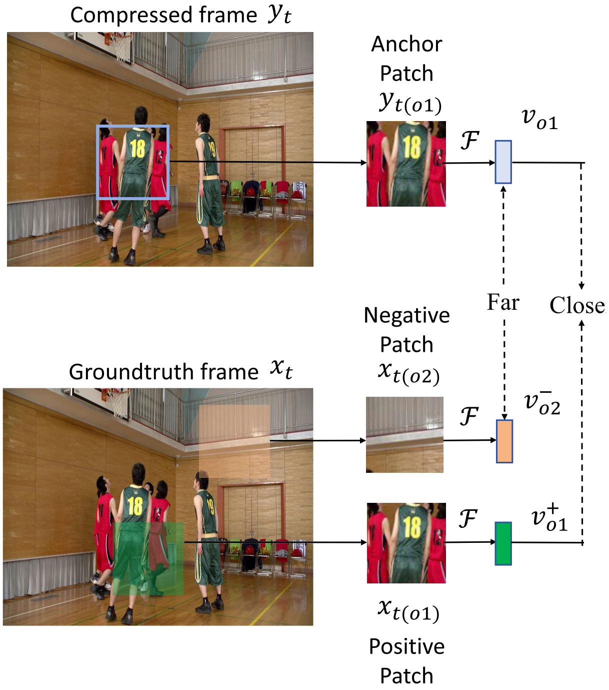

FIGURE 4. Illustration of the patch-wise contrastive learning procedure.

E. LEARNING OBJECTIVES

We regularize the frame y_t decoded from the perceptual stream to be visually comfort from two aspects, (1) the mutual information between the decoded frame and the original frame should be maximized, and (2) the local texture patterns of the human attentive regions between the decoded frame and the original frame should be similar.

For the first objective, we employ the patch-wise contrastive loss function. As shown in Figure 4, given the positive patch $y_t(o_1)$ located in o_1 position of the compressed video y_t, we force its feature to be similar to the positive patch $x_t(o_1)$ located in the same position of the raw frame, but dissimilar to the negative patches such as $x_t(o_2)$ in other location o_2. The features of the above patches are extracted by a small CNN $\mathcal{F}$. $\mathcal{F}$ is implemented as five convolutions of kernel size three, followed by ReLU non-linearity.

$$\begin{aligned} \boldsymbol{v} &= \mathcal{F}(y_t(o_1)), \\ \boldsymbol{v}^+ &= \mathcal{F}(x_t(o_1)), \\ \boldsymbol{v}_n^- &= \mathcal{F}(x_t(o_n)), \end{aligned} \tag{5}$$

Then, the contrastive learning for learning patch semantics can be given by:

$$\mathcal{L}_{patch-contra} = -\log\left[\frac{\exp\left(\boldsymbol{v}\cdot\boldsymbol{v}^+/\tau\right)}{\exp\left(\boldsymbol{v}\cdot\boldsymbol{v}^+/\tau\right)+\sum_{n=2}^{N}\exp\left(\boldsymbol{v}\cdot\boldsymbol{v}_n^-/\tau\right)}\right], \tag{6}$$

where N is the number of negative examples, $\tau = 0.07$ is the temperature hyperparmater.

For the second objective, we improve the perceptual loss [47] by re-weighting it with the saliency map M, which is output from the saliency detection network [48] and binarized by the threshold 0.5.

$$\mathcal{L}_{saliency-percep} = MSE(M \odot VGG16(\hat{x}_t), M \odot VGG16(x_t)), \tag{7}$$

where MSE denotes the mean square loss, $\odot$ denotes the broadcasting multiplication, and VGG16 denotes the pre-trained VGG16 network [49] on ImageNet [50]. We adopt the features output from relu4_3 layer for balancing the local details and the global structure.

The final learning objective is given by,

$$\mathcal{L} = \lambda R(f_t) + \mathcal{L}_{saliency-percep} + \mathcal{L}_{patch-contra} + \mathcal{L}_{adv}, \tag{8}$$

where $\mathcal{L}_{adv}$ denotes the non-saturating adversarial loss [51], λ is the balancing hyper-parameter.

F. OTHER DETAILS

Regarding quantization, to comprehensively optimize the entire model end-to-end, it is essential to employ a differentiable quantization operation. In our framework, we adhere to the approach outlined in [21], approximating the quantization operation by introducing uniform noise during the training phase. When it comes to evaluation, we employ the rounding operation directly.

IV. EXPERIMENTS

A. DATASETS

Several training datasets have been created for learned video compression [52], [53]. Following established methodologies [8], [9], we utilize the popular Vimeo-90k [53] training split in our investigation, where videos are randomly cropped into 256 × 256 patches. To evaluate the efficacy and application range of our proposed video compression method, we utilize HEVC dataset [4]. The HEVC dataset consists of 16 sequences classified into Class B, C, D, and E. Visual quality metrics are computed in the RGB color space.

B. IMPLEMENTATION DETAILS

We use the H.266 implementation VVenC [54] as the PSNR-oriented codec in our framework. We adopt VVenC instead of VTM due to the fast coding speed of VVenC, and this is more practical for the real-world applications. We train four models with varying λ values (specifically, $\lambda = 1$, 2, 4, and 8) to cover a spectrum of coding rates. The corresponding quality parameter (QP) of the H.266 codec is set to 29, 32, 35, and 38, respectively. Our model undergoes a two-stage training process. In the initial stage, we train the model using only two consecutive video frames with the MSE loss for 100,000 steps. This stage serves to establish a robust initial state for the network. In the second stage, we further train the model utilizing the training videos with proposed semantic loss functions for an additional 900,000 steps. For optimization, we employ the Adam optimizer [55] with a batch size of 6 and set the learning rate to 1e-4. The implementation of our model is carried out using PyTorch, and the training process is executed on 2 NVIDIA 4090 GPUs. The entire training duration for our model spans six days.

C. EVALUATION METRICS

We utilize bpp (bits per pixel) as a metric to measure the bits cost for each pixel in every frame. Specifically, the bpp value is calculated with the following formulation,

$$bpp = \frac{bits_{VVC} + bits_{neural}}{T \times H \times W}, \tag{9}$$

where $bits_{VVC}$ and $bits_{neural}$ denote the bits consumed by the VVC codec and our neural frame, respectively. T, H, and W indicates the video length, video frame height, and video frame width, respectively. To quantitatively assess video subjective quality, we incorporate two feature-based metrics, namely LPIPS [56] and DISTS [57], along with two distribution-based metrics, namely FID [58] and KID [59]. We abstain from employing traditional PSNR and SSIM [60] metrics due to their inconsistency with the human visual experience.

D. EXPERIMENTAL RESULTS

1) THE SETTINGS OF THE BASELINE METHODS

All baseline methods are assessed using the GOP size 32 and clip length 96 configuration, a setting widely adopted in recent methodologies [9], [39]. To create compressed videos from H.264 and H.265, we utilize the FFmpeg software [61]. To create compressed videos from H.266, we utilize the VVenC software [54].

a: H.264

The command line for generating H.264 compressed video is provided as follows, “ffmpeg -y -pix_fmt yuv420p -s WxH -r FR -i Video.yuv -vframes N -c:v libx264 -tune zerolatency -crf Q -g GOP -bf 2 -b strategy 0 -sc threshold 0 output.mkv”. In the command, W, H, FR, N, Q, and GOP represent the width, height, frame rate, number of encoded frames, quality, and GOP size, respectively. For our HEVC datasets, N is set to 96. Quality (Q) is configured as 29, 32, 35, 38 in our settings. The GOP size is set to 32, aligning with recent methodologies [39].

b: H.265

The command line for generating H.264 compressed video is provided as follows, “ffmpeg -pix_fmt yuv420p -s WxH -r FR -i Video.yuv -vframes N -c:v libx265 -tune zerolatency -x265-params “crf=Q:keyint=GOP:verbose=1” output.mkv”.

c: H.266

We adopt the open-source VVenC software [54] for performing the compression. The evaluation is performed with the “cfg /experimental/lowdelay_faster.cfg” configuration file.

d: PLVC

We conduct a re-evaluation of the official code under the GOP size 32 and clip length 96 settings to ensure a fair comparison. The compressed frames produced by DCVCDC are saved and subsequently assessed using the four subjective quality metrics adopted in this paper.

e: DCVCDC

To ensure a fair comparison, we re-evaluate the official code under the GOP size 32 and clip length 96 settings. The compressed frames generated by DCVCDC are saved and subsequently assessed using the four subjective quality metrics adopted in this paper.

2) RESULTS

In Table 1, we present the BDBR [62] results of various video compression methods in comparison to H.264 across HEVC Class B, Class C, Class D, and Class E datasets. Notably, our approach demonstrates a bit rate reduction exceeding 60% in the overall results across all benchmark datasets. It is evident that our method surpasses the performance of the advanced traditional codec H.266, the state-of-the-art learnable neural

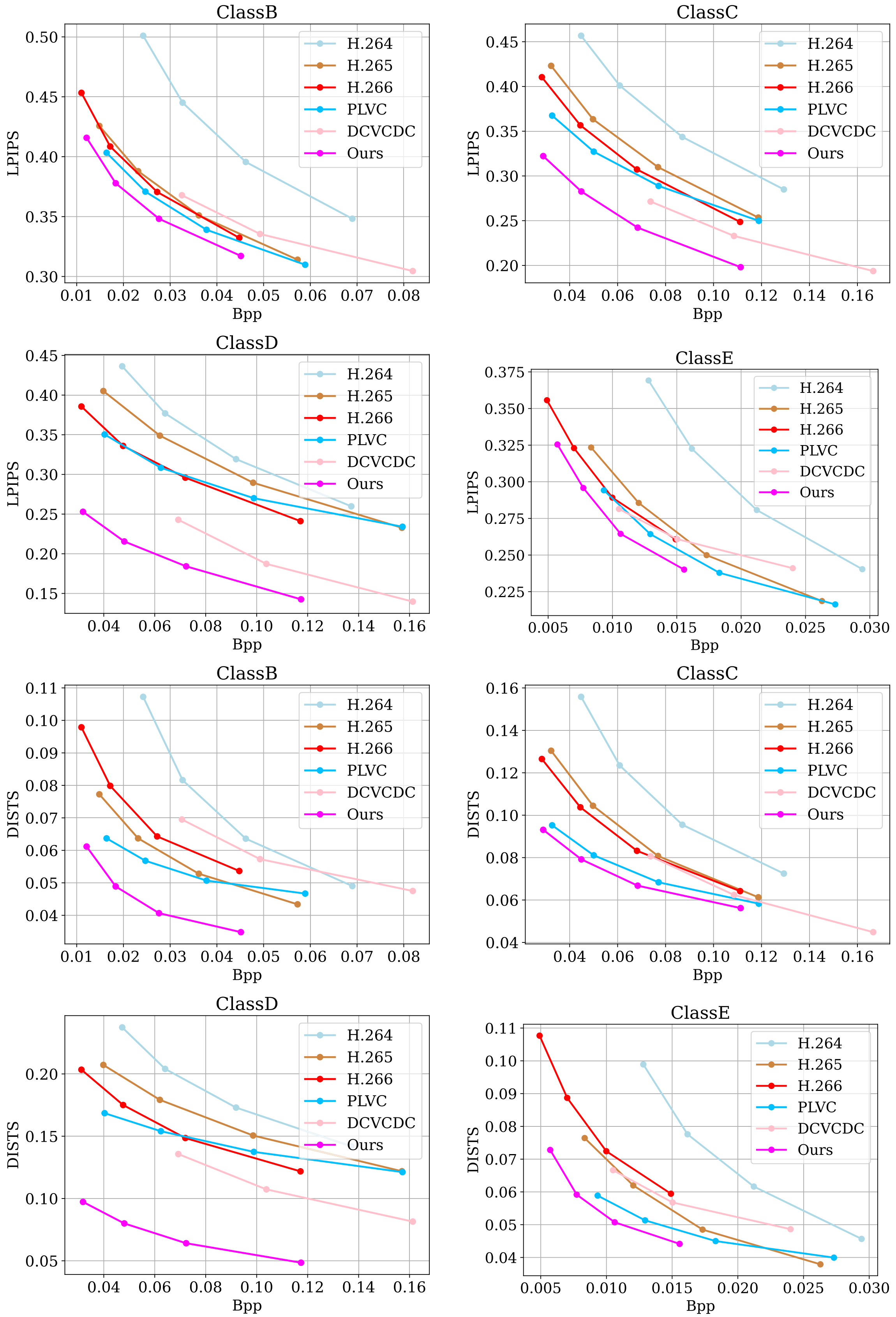

FIGURE 5. Rate-Distortion Curves of different coding methods, when being evaluated with the LPIPS and DISTS metricss.

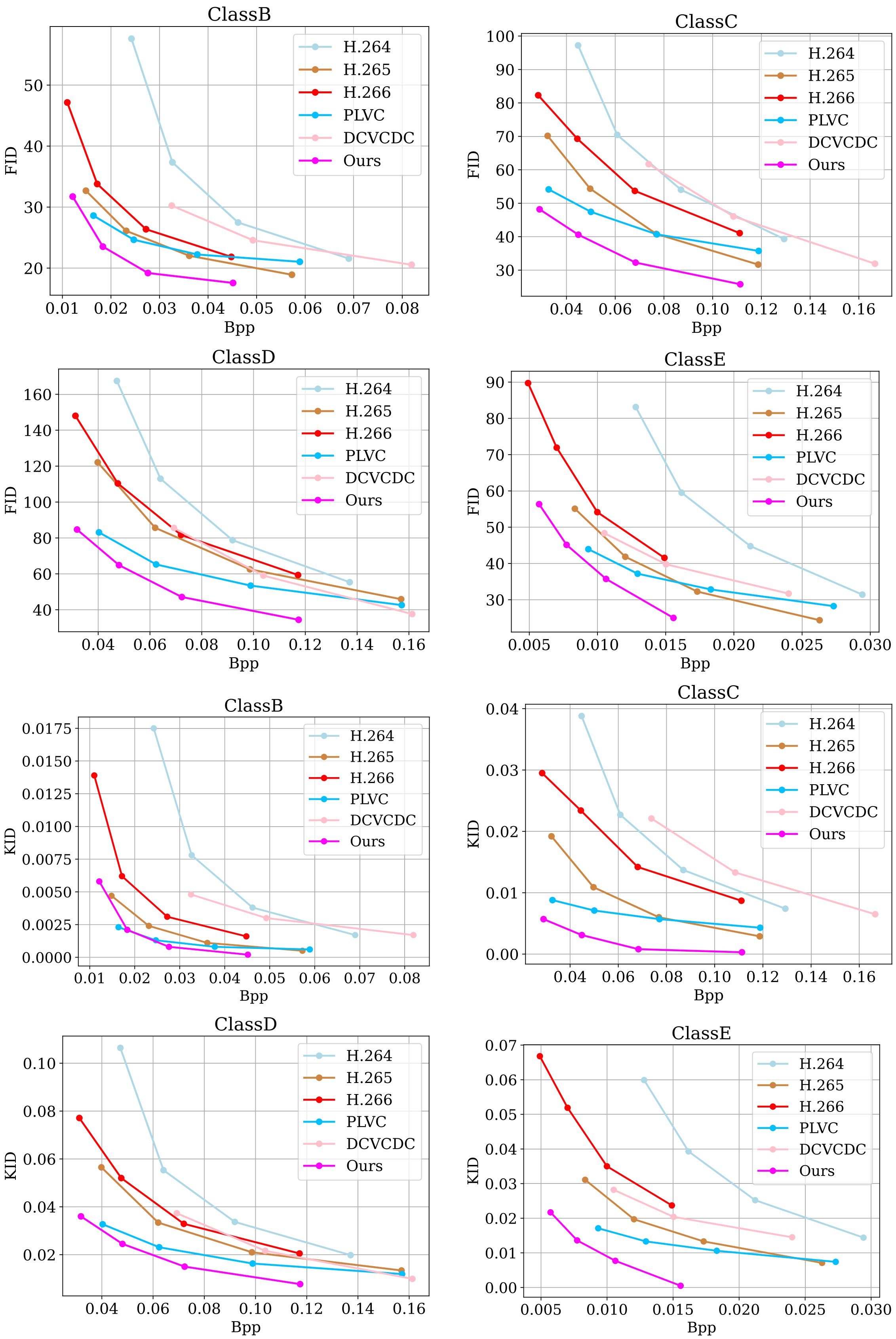

FIGURE 6. Rate-Distortion Curves of different coding methods, when being evaluated with the FID and KID metrics.

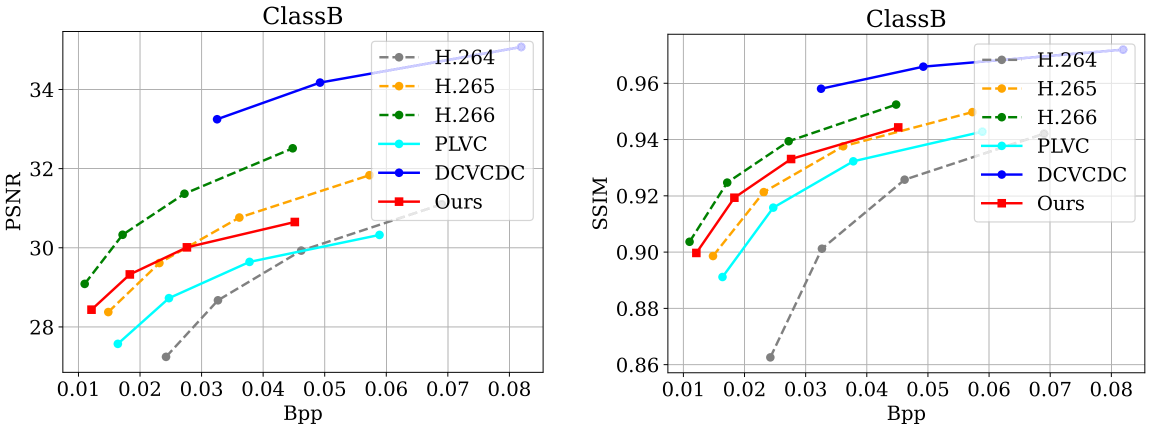

FIGURE 7. Comparison of different codecs in terms of PSNR and SSIM metrics.

TABLE 1. BDBR(%) of four baseline methods (H.265, H.266, PLVC and DCVC) as well as our method, when compared with H.264 on the HEVC Class B, Class C, Class D and Class E datasets. Negative values in BDBR denote bit-rate savings. The smaller BDBR, the more bitrate saved.

		Class B	Class C	Class D	Class E
LPIPS	H.265	-52.90	-34.02	-17.14	-40.99
	H.266	-57.40	-43.52	-39.04	-53.71
	PLVC	-57.25	-48.00	-35.11	-46.51
	DCVCDC	-42.32	-48.92	-54.44	-38.30
	Ours	**-65.80**	**-68.55**	**-77.95**	**-57.68**
DISTS	H.265	-49.11	-35.09	-26.26	-42.64
	H.266	-46.20	-41.60	-46.59	-45.69
	PLVC	-53.59	-56.15	-47.95	-51.07
	DCVCDC	-10.20	-33.24	-53.28	-35.19
	Ours	**-74.12**	**-62.52**	**-90.72**	**-63.68**
FID	H.265	-54.24	-42.38	-26.03	-46.92
	H.266	-51.32	-26.34	-25.38	-46.18
	PLVC	-53.37	-50.26	-47.15	-48.34
	DCVCDC	-12.90	-1.29	-18.29	-36.02
	Ours	**-63.82**	**-66.82**	**-58.96**	**-62.63**
KID	H.265	-59.81	-51.22	-32.57	-52.58
	H.266	-50.88	-23.93	-28.29	-45.00
	PLVC	-68.78	-68.77	-51.78	-63.65
	DCVCDC	-7.58	20.95	-19.75	-38.00
	Ours	**-61.92**	**-93.24**	**-60.24**	**-75.35**

video codec DCVCDC, and the previous perceptual coding method PLVC. For instance, compared to H.264, our proposed method achieves a bit-rate saving of 77.95% on the HEVC Class D dataset, while the corresponding bit-rate savings for recent H.266, DCVCDC, and PLVC are 39.04%, 54.44%, and 35.11%, respectively, when being evaluated with LPIPS. Traditional PSNR codecs, including those the traditional codecs, exhibit poor performance in subjective quality metrics such as LPIPS, as also discussed in [7]. This deficiency motivated our work to enhance traditional codecs by augmenting semantic information and post-enhancing frame quality.

The superior BDBR results of our approach in Table 1 stem from several factors. Firstly, the anchor codec is an older H.264 codec, while our method builds upon the advanced VVC (H.266) codec, which already surpasses H.264 significantly. Secondly, H.264's block-wise compression technique often leads to compression artifacts like blocking and ring artifacts, degrading subjective image quality. In contrast, our approach, employing neural networks with perceptual and adversarial losses, enhances subjective image quality. To quantitatively decompose the performance gain originality, we conduct the following step-by-step experiments. Initially, upgrading the codec from H.264 to VVC reduces the BDBR from -17.14 to -39.04 using LPIPS as the quality metric on the HEVC Class D dataset. Then, integrating the Unet enhancement network with an L1 loss function further reduces the BDBR to 46.23. Subsequently, replacing the L1 loss function with perceptual and adversarial losses substantially reduces the BDBR to 61.11. Finally, introducing contrastive learning-based semantic streams achieves the ultimate reduction of the BDBR to 77.95.

Despite our substantial performance gain with perceptual metrics, our framework indeed demonstrates poor performance on traditional low-level signal distortion metrics like PSNR/SSIM, as illustrated in Figure 7. This observation aligns with the perceptual-distortion trade-off theory [63].

We provide the RD curves of different compression methods in Figure 5 and Figure 6, it is noted that our method outperforms the all other methods by a large margin on all datasets, when being evaluated with four perceptual quality metrics. We observe that DCVCDC significantly outperforms the advanced H.266 codec by a considerable margin in terms

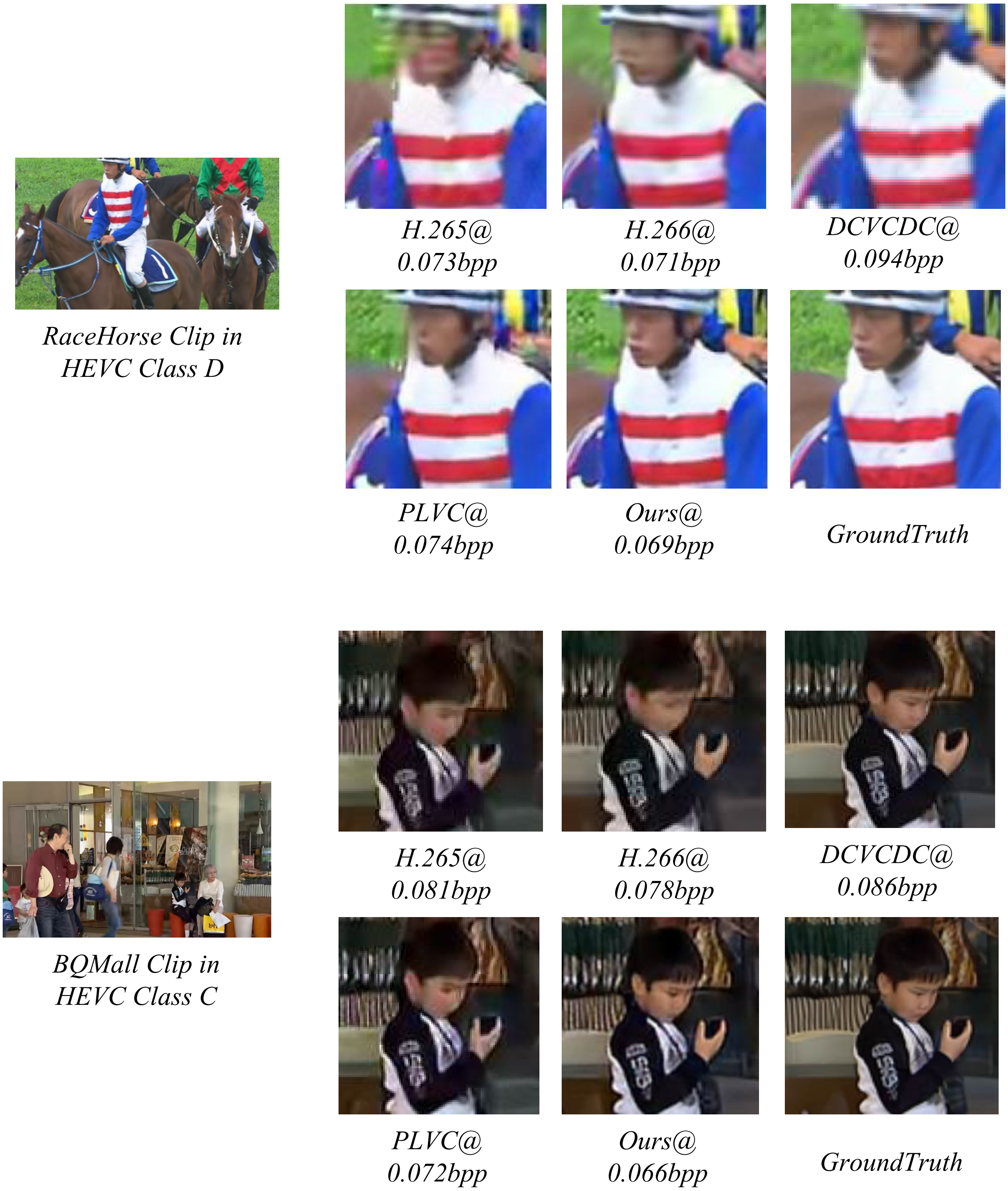

FIGURE 8. **Qualitative comparison of different video compression methods.**

of the PSNR metric, as reported in their paper. However, it exhibits poor performance when evaluated using perceptual metrics. We hypothesize that the reason for this could be that the DCVCDC codec has reached the optimal PSNR point on the PSNR-Perceptual trade-off plane, which corresponds to the least favorable perceptual performance [64].

Finally, we showcase the video frames compressed by various methods in Figure 8. Notably, our model produces noticeably higher-quality reconstructed frames at the same bpp level compared to H.265/H.266. When contrasted with the state-of-the-art neural codec DCVCDC, our approach exhibits sharper object edges and an improved visual experience. Additionally, in comparison to the perceptual codec PLVC, our method preserves more meaningful video contents, including finer details in human facial features.

3) ABLATION STUDY

In this section, we train several variant models to investigate the effectiveness of all proposed modules.

As detailed in Table 2, when simultaneously removing the patch-wise contrastive loss $\mathcal{L}patch - contrast$, replacing the visual saliency weighted perceptual loss $\mathcal{L}saliency - percep$

TABLE 2. Ablation Studies. Smaller BDBR is better, where the average bitcost is lower. The dataset is Class D. The anchor codec for calculating the BDBR is H.265.

$\mathcal{L}_{patch-contrast}$	✗	✗	✗	✓
$\mathcal{L}_{saliency-percep}$	✗	✗	✓	✓
Perceptual loss	✓	✓	✗	✗
Adaptive Kernel	✗	✓	✓	✓
BDBR(LPIPS)	-39.04%	-49.64%	-62.45%	-77.95%

with the plain perceptual loss, and also removing the adaptive kernel design, the resulting model yields the least favorable outcome, with a -39.04% BDBR. Upon reintroducing the adaptive kernel design, the BDBR of the model is restored to -46.64%. Further incorporating the saliency weighted perceptual loss $\mathcal{L}saliency - percep$, the model achieves a -62.45% BDBR. Ultimately, the inclusion of the patch-wise contrastive loss $\mathcal{L}_{patch-contrast}$ results in a -77.95% BDBR.

In summary, all designs presented in this paper are necessary and effective. Notably, the patch-wise contrastive loss emerges as the most significant contributor to the observed improvements.

E. ENCODING SPEED AND MODEL COMPLEXITY

The encoding speed of our approach for a single 1080p frame is 911ms, showcasing a faster performance compared to the recent neural codec DCVCDC (1005ms). Although our approach is slightly slower than the VVenC codec (634ms), it's important to note that we incorporate a neural coding procedure, contributing to this difference in processing time. However, given the substantial bitrate saving achieved by our approach over VVenC, the increase in running time is considered a worthwhile trade-off. The parameter count of the networks within our approach is 9.62M.

V. CONCLUSION

In this paper, we introduce a novel video compression approach tailored for achieving high subjective video quality to enhance the overall human visual experience. Our method is designed within a conditional framework, representing a departure from conventional PSNR-oriented codecs, with a focus on improving perceptual quality. Furthermore, three innovative designs are introduced: a patch-wise contrastive learning objective, a saliency map-weighted perceptual loss, and adaptive kernel convolution. These designs collectively contribute to superior perceptual coding capabilities. Through extensive evaluations on four datasets, our approach demonstrates clear advantages over previous methods.

REFERENCES

[1] G. M. D. T. Forecast, "Cisco networking index: Global mobile data traffic forecast update, 2017–2022," Update, 2019, vol. 2017, p. 2022.

[2] T. Berger, "Rate-distortion theory," in *Wiley Encyclopedia of Telecommunications*. Apr. 2003.

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

[4] 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.

[5] B. Bross, Y.-K. Wang, Y. Ye, S. Liu, J. Chen, G. J. Sullivan, 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, Oct. 2021.

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

[7] R. Yang, R. Timofte, and L. Van Gool, "Perceptual learned video compression with recurrent conditional GAN," 2021, *arXiv:2109.03082*.

[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. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10998–11007.

[9] J. Li, B. Li, and Y. Lu, "Deep contextual video compression," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 18114–18125.

[10] S. Duan, H. Chen, and J. Gu, "JPD-SE: High-level semantics for joint perception-distortion enhancement in image compression," *IEEE Trans. Image Process.*, vol. 31, pp. 4405–4416, 2022.

[11] C. Zhu, G. Lu, R. Xie, and L. Song, "Perceptual video coding based on semantic-guided texture detection and synthesis," in *Proc. Picture Coding Symp. (PCS)*, Dec. 2022, pp. 217–221.

[12] 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. Eur. Conf. Comput. Vis.*, 2022, pp. 562–578.

[13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial networks," *Commun. ACM*, 2020.

[14] T. Park, A. A. Efros, R. Zhang, and J.-Y. Zhu, "Contrastive learning for unpaired image-to-image translation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 319–345.

[15] X. Jia, B. D. Brabandere, T. Tuytelaars, and L. Van Gool, "Dynamic filter networks," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, vol. 29, 2016.

[16] G. K. Wallace, "The JPEG still picture compression standard," *Commun. ACM*, vol. 34, no. 4, pp. 30–44, Apr. 1991.

[17] W. Sun and Z. Chen, "Learned image downscaling for upscaling using content adaptive resampler," *IEEE Trans. Image Process.*, vol. 29, pp. 4027–4040, 2020.

[18] Y. Tian, G. Lu, X. Min, Z. Che, G. Zhai, G. Guo, and Z. Gao, "Self-conditioned probabilistic learning of video rescaling," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 4470–4479.

[19] Y. Tian, Y. Yan, G. Zhai, L. Chen, and Z. Gao, "CLSA: A contrastive learning framework with selective aggregation for video rescaling," *IEEE Trans. Image Process.*, vol. 32, pp. 1300–1314, 2023.

[20] J. Ballé, V. Laparra, and E. P. Simoncelli, "End-to-end optimized image compression," 2016, *arXiv:1611.01704*.

[21] J. Ballé, D. Minnen, S. Singh, S. J. Hwang, and N. Johnston, "Variational image compression with a scale hyperprior," 2018, *arXiv:1802.01436*.

[22] D. Minnen, J. Ballé, and G. Toderici, "Joint autoregressive and hierarchical priors for learned image compression," 2018, *arXiv:1809.02736*.

[23] T. Chen, H. Liu, Z. Ma, Q. Shen, X. Cao, and Y. Wang, "End-to-end learnt image compression via non-local attention optimization and improved context modeling," *IEEE Trans. Image Process.*, vol. 30, pp. 3179–3191, 2021.

[24] H. Ma, D. Liu, N. Yan, H. Li, and F. Wu, "End-to-end optimized versatile image compression with wavelet-like transform," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 3, pp. 1247–1263, Mar. 2022.

[25] Y. Xie, K. L. Cheng, and Q. Chen, "Enhanced invertible encoding for learned image compression," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 162–170.

[26] E. Agustsson, F. Mentzer, M. Tschannen, L. Cavigelli, R. Timofte, L. Benini, and L. V. Gool, "Soft-to-hard vector quantization for end-to-end learning compressible representations," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[27] X. Zhang and X. Wu, "LVQAC: Lattice vector quantization coupled with spatially adaptive companding for efficient learned image compression," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 10239–10248.

[28] D. Minnen and S. Singh, ''Channel-wise autoregressive entropy models for learned image compression,'' in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2020, pp. 3339–3343.

[29] Z. Cheng, H. Sun, M. Takeuchi, and J. Katto, ''Learned image compression with discretized Gaussian mixture likelihoods and attention modules,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7936–7945.

[30] X. Zhu, J. Song, L. Gao, F. Zheng, and H. T. Shen, ''Unified multivariate Gaussian mixture for efficient neural image compression,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 17591–17600.

[31] J. Lin, D. Liu, H. Li, and F. Wu, ''M-LVC: Multiple frames prediction for learned video compression,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3543–3551.

[32] Z. Hu, G. Lu, and D. Xu, ''FVC: A new framework towards deep video compression in feature space,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 1502–1511.

[33] H. Liu, M. Lu, Z. Ma, F. Wang, Z. Xie, X. Cao, and Y. Wang, ''Neural video coding using multiscale motion compensation and spatiotemporal context model,'' *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 8, pp. 3182–3196, Aug. 2021.

[34] E. Agustsson, D. Minnen, N. Johnston, J. Ballé, S. J. Hwang, and G. Toderici, ''Scale-space flow for end-to-end optimized video compression,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8500–8509.

[35] A. Chadha and Y. Andreopoulos, ''Deep perceptual preprocessing for video coding,'' in *Proc. CVPR*, Jun. 2021, pp. 14852–14861.

[36] X. Zhang and X. Wu, ''Attention-guided image compression by deep reconstruction of compressive sensed saliency skeleton,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 13349–13359.

[37] X. Zhang and X. Wu, ''Dual-layer image compression via adaptive downsampling and spatially varying upconversion,'' 2023, *arXiv:2302.06096*.

[38] X. Sheng, J. Li, B. Li, L. Li, D. Liu, and Y. Lu, ''Temporal context mining for learned video compression,'' *IEEE Trans. Multimedia*, vol. 25, pp. 7311–7322, 2022.

[39] J. Li, B. Li, and Y. Lu, ''Neural video compression with diverse contexts,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 22616–22626.

[40] F. Mentzer, G. Toderici, M. Tschannen, and E. Agustsson, ''High-fidelity generative image compression,'' in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 11913–11924.

[41] L. Galteri, M. Bertini, L. Seidenari, T. Uricchio, and A. Del Bimbo, ''Increasing video perceptual quality with GANs and semantic coding,'' in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 862–870.

[42] J. Chang, Z. Zhao, C. Jia, S. Wang, L. Yang, Q. Mao, J. Zhang, and S. Ma, ''Conceptual compression via deep structure and texture synthesis,'' *IEEE Trans. Image Process.*, vol. 31, pp. 2809–2823, 2022.

[43] G. Konuko, G. Valenzise, and S. Lathuilière, ''Ultra-low bitrate video conferencing using deep image animation,'' in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2022, pp. 3515–3520.

[44] Y. Tian, G. Lu, Y. Yan, G. Zhai, L. Chen, and Z. Gao, ''A coding framework and benchmark towards low-bitrate video understanding,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Feb. 20, 2024, doi: 10.1109/TPAMI.2024.3367879.

[45] Y. Tian, G. Lu, G. Zhai, and Z. Gao, ''Non-semantics suppressed mask learning for unsupervised video semantic compression,'' in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 13610–13622.

[46] B. Xu, N. Wang, T. Chen, and M. Li, ''Empirical evaluation of rectified activations in convolutional network,'' 2015, *arXiv:1505.00853*.

[47] J. Johnson, A. Alahi, and L. Fei-Fei, ''Perceptual losses for real-time style transfer and super-resolution,'' in *Proc. Eur. Conf. Comput. Vis.*, Amsterdam, Netherlands, 2016, pp. 694–711.

[48] K. Min and J. Corso, ''TASED-Net: Temporally-aggregating spatial encoder–decoder network for video saliency detection,'' in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2394–2403.

[49] K. Simonyan and A. Zisserman, ''Very deep convolutional networks for large-scale image recognition,'' 2014, *arXiv:1409.1556*.

[50] A. Krizhevsky, I. Sutskever, and G. E. Hinton, ''ImageNet classification with deep convolutional neural networks,'' *Commun. ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[51] M. Lucic, K. Kurach, M. Michalski, S. Gelly, and O. Bousquet, ''Are GANs created equal? A large-scale study,'' in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018.

[52] D. Ma, F. Zhang, and D. R. Bull, ''BVI-DVC: A training database for deep video compression,'' *IEEE Trans. Multimedia*, vol. 24, pp. 3847–3858, 2022.

[53] T. Xue, B. Chen, J. Wu, D. Wei, and W. T. Freeman, ''Video enhancement with task-oriented flow,'' *Int. J. Comput. Vis.*, vol. 127, no. 8, pp. 1106–1125, Aug. 2019.

[54] A. Wieckowski, J. Brandenburg, T. Hinz, C. Bartnik, V. George, G. Hege, C. Helmrich, A. Henkel, C. Lehmann, C. Stoffers, I. Zupancic, B. Bross, and D. Marpe, ''Vvenc: An open and optimized vvc encoder implementation,'' in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, Jul. 2021, pp. 1–2.

[55] D. P. Kingma and J. Ba, ''Adam: A method for stochastic optimization,'' 2014, *arXiv:1412.6980*.

[56] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, ''The unreasonable effectiveness of deep features as a perceptual metric,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 586–595.

[57] K. Ding, K. Ma, S. Wang, and E. P. Simoncelli, ''Image quality assessment: Unifying structure and texture similarity,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 5, pp. 2567–2581, May 2022.

[58] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, ''GANs trained by a two time-scale update rule converge to a local nash equilibrium,'' in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, vol. 30, 2017.

[59] M. Bińkowski, D. J. Sutherland, M. Arbel, and A. Gretton, ''Demystifying MMD GANs,'' 2018, *arXiv:1801.01401*.

[60] Z. Wang, E. P. Simoncelli, and A. C. Bovik, ''Multiscale structural similarity for image quality assessment,'' in *Proc. 37th Asilomar Conf. Signals, Syst. Comput.*, vol. 2, 2003, pp. 1398–1402.

[61] S. Tomar, ''Converting video formats with FFmpeg,'' *Linux J.*, vol. 2006, no. 146, p. 10, 2006.

[62] G. Bjontegaard, *Calculation of Average PSNR Differences Between RD-Curves*, document VCEG-M33, 2001.

[63] Y. Blau and T. Michaeli, ''The perception-distortion tradeoff,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6228–6237.

[64] Y. Blau and T. Michaeli, ''Rethinking lossy compression: The rate-distortion-perception tradeoff,'' in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 675–685.

BINGYAO LI received the master's degree in civil law from Shanghai Jiao Tong University, in 2014. She is currently pursuing the Ph.D. degree with Shanghai University of Finance and Economics.

Additionally, she has spent the last two years as a Professional Lawyer. During her career as a Professional Lawyer, she encountered many criminal cases and tragedies that resulted from the inability to detect criminal actions in a timely manner. She is currently a Lecture with Shanghai Business School. If the video transmission algorithm can be more effectively improved, thereby enhancing the efficiency of automated criminal detection in the cloud, it will greatly reduce such tragedies. Therefore, she began to learn the relevant technologies of video compression and automatic video analysis and developed a strong interest in scientific research.