Deep Perceptual Lossless Coding: A Case Study of Intra-frame Database and Framework

Anonymous

***Abstract*—Perceptually lossless coding (PLC) is a critical technique for high-quality video services, which aims to achieve maximum compression ratio while distortions cannot be perceptible by the human visual system (HVS). However, most existing methods unilaterally consider the perceptual characteristics of HVS and rigidly combine the visual perception with the existing coding system. To address these challenges, this paper proposes a new perceptually lossless coding (PLC) method that jointly optimizes the visual perception and coding system. Specifically, a new block-level video database is built based on Multiple-QP to select the best quantization parameter (QP) for each coding tree unit (CTU), and a fine-grained subjective quality evaluation experiment is adopted to generate accurate perceptually lossless labels. Furthermore, a deep neural network is designed to explore the trade-off relationship between rate and distortion to predict perceptually lossless QP. The experimental results demonstrate that the proposed method achieves an average rate saving of 29.47% compared to the latest method at the same perceptual quality.**

***Index Terms*—Multiple-QP optimization, perceptually lossless coding, versatile video coding (VVC), deep neural network**

I. Introduction

As the internet of video things (IoVT) continues to expand, the data traffic for high-quality video applications, such as cloud games, live streaming, and virtual/augmented reality (VR/AR) has increased nearly tenfold [1]. High-quality video with ultra-high definition (UHD), high frame rate (HFR), and high dynamic range (HDR) can provide a more immersive visual experience, which is crucial to ensuring high-quality video quality of service (QoS). However, high-quality videos contain more pixels and frames, resulting in massive data and a significant burden on transmission bandwidth and storage resources. While hardware resources, such as communication and storage, have improved, they still face challenges, such as long cycle times and high costs. Therefore, high-quality video services, such as HDR video streaming [2], video coding for machine [3], security surveillance [4], UHD video cloud computing [5], urgently need high compression ratio and high-quality video coding technology, particularly under the current resource constraints.

To conserve bandwidth and storage resources, various video coding methods have been developed to improve the compression efficiency of high-quality video, which can be broadly categorized into standard [6] and learning-based methods [7]. Standard method adopts a hybrid coding framework of prediction and transformation to remove data redundancy more efficiently. For instance, Versatile Video Coding (VVC) [8] based on the previous generation of High Efficiency Video Coding (HEVC) [9], have proposed many new technologies, such as wide prediction angle and bidirectional optical flow, to meet the coding requirements of UHD and HFR video. On the other hand, learning-based methods employ hybrid coding methods [10]–[13] combining deep neural network (DNN) with existing coding standards or in an end-to-end manner [14]–[18], which have achieved improved rate-distortion performance on high-resolution images [17]. However, existing coding methods are mainly lossless (perceptually lossless but low compression ratio) and lossy (high compression ratio but perceptual quality cannot be guaranteed), which are difficult to fulfill the compression requirements of high-quality video services. To address this issue, perceptually lossless coding (PLC) considers the characteristics of human visual perception and can provide high-quality reconstructed video without perceptual distortion, while achieving a high compression ratio. As an illustration in Fig. 1, We illustrate a performance comparative example of lossless, perceptually lossless coding and lossy in high-quality video service.

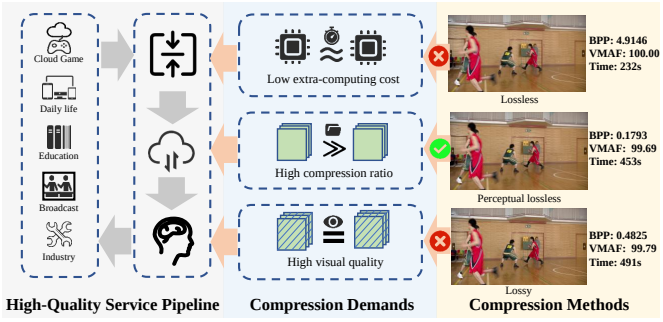


Fig. 1. **Compression results via different compression schemes for high-quality video services.** Lossless uses the $lossless$ mode of VVC, lossy adopts $QP = 22$. The perceptually lossless method is the proposed method.

The main principle of PLC is to model perceptual lossless threshold (PLT) by considering HVS's characteristics and cognitive rules (e.g., luminance adaptation, contrast masking, and pattern complexity), or through a data-driven approach. In the multimedia community, various representative PLC methods have been widely used in video data compression [19]–[23]. However, existing PLC methods have the following limitations: 1) *Inaccurate PLT* leads to a large number of perceptually redundant residuals. 2) Visual threshold-guided methods require *multi-pass coding*, resulting in a dramatic increase in computational complexity. 3) The *characteristics of codec* are not fully considered, resulting in the actual re-

Manuscript received April 19, 2021; revised August 16, 2021.

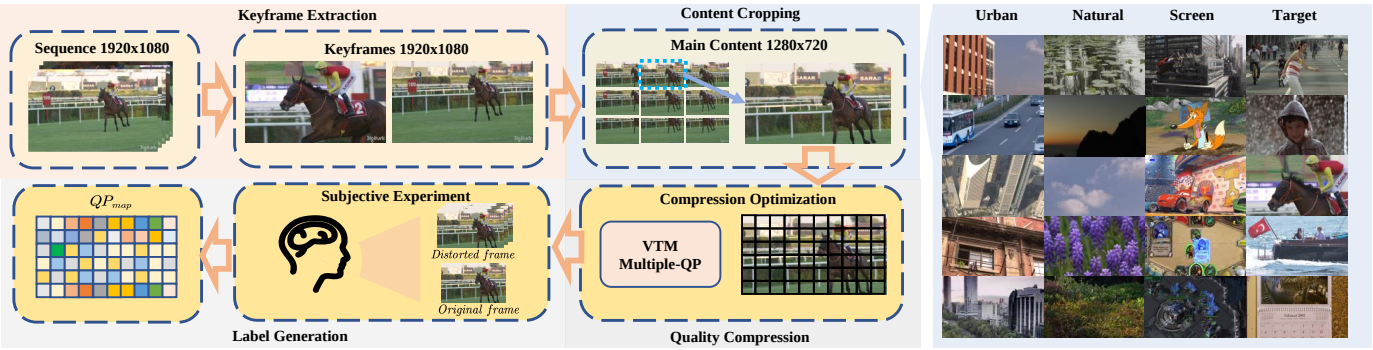


Fig. 2. **Illustrate of database generation process and examples**. The generation of our database mainly includes the generation of frames and labels. The generation of frames includes keyframe extraction and main contents cropping. The generation of labels consists of quality compression and subjective experiment. The generated frame content roughly divided into four categories: Urban, Nature, Screen and target. The detailed description of each part can be found in Section III

construction results not great as expected. Therefore, the most fundamental motivation of this work is to jointly optimize the visual perception characteristics and video coding framework to implement efficient and accurate PLC.

To address the challenges outlined above, we propose a PLC method that comprehensively considers the characteristics of the *coding framework* and *visual perception*. Firstly, in terms of the coding framework, we established an extensive database. In order to fully consider the characteristics of the coding framework and make full use of its coding performance, the *Multiple-QP* tool is turned on to improve the subjective quality [24] when generating distorted frames of different quality. Second, in terms of visual perceptually lossless quality, we use fine-grained quality variance to evaluate subjective experiments to generate accurate perceptually lossless quantization parameter (QP) labels. Finally, we adopt a DNN to model the mapping between the input frame and QP. The main contributions of our work are summarized as follows:

- To research on fine-grained PLC, a novel block-level perceptually lossless video database is established. From the perspective of the coding framework, we use the *Multiple-QP* strategy to obtain QP labels of different quality levels, and from the perspective of human visual perception, we use fine-grained visual quality subjective experiments to select perceptually lossless QP labels.
- In order to achieve efficient and accurate PLC, we propose a QP mapping model based on DNN. According to the relationship between distortion and rate in video coding, we extract rate and distortion features respectively and learn the rate-distortion trade-off relationship.
- The proposed model is integrated into the VVC test model (VTM) for experiments, and the experimental results prove that our method achieves PLC with a higher compression ratio while introducing tolerable complexity.

The rest of this paper is organized as follows. In Section II, we provide a brief review of video lossless, lossy and perceptually lossless compression. Section III details our database establishment process and statistical analysis. Section IV introduces the proposed DNN model framework. Section V describes our experimental protocol, presents the overall performance in both subjective and objective criteria, and shows the ablation results and complexity analysis. Finally, Section VI concludes the paper and proposes some open ideas for future work.

II. Releated Works and Motivations

A. Releated Works

The proposed method is designed specifically for PLC. Therefore, we mainly review some representative methods that are able to achieve PLC. For a detailed overview of standard video coding, the reader is referred to [6], for a detailed deep learning-based overview of video coding, the reader is referred to [25], and for a detail exploration of perceptual optimization in video coding. the reader is referred to [26].

1) ***Lossless Compression****:* Lossless compression refers to techniques for compressing data without losing any information, These methods can be summarized as standard and learning-based methods.

Standard Lossless methods (e.g., HEVC, VVC) ensures lossless compression by disabling lossy modules (e.g., transform, quantization, and loop filtering) [27].

Learning-based methods can be further divided into hybrid and end-to-end methods, hybrid methods use DNN to replace specific modules (e.g., prediction [10], partitioning [11]) of standard methods. End-to-end method mainly uses traditional lossy compression and residual probability estimation [15], [16], [28]. Although lossless methods ensure perceptually lossless, they are difficult to widely adopted limited by the low compression ratio.

2) ***Lossy Compression****:* Lossy compression aims to achieve the highest compression ratio by optimizing the trade-off between rate and distortion. It is widely used in multimedia applications, and standard lossy methods are constantly updated to improve the efficiency of lossy coding. For example, VVC [8], achieves an average of 50% efficiency improvement compared to HEVC [9].

In recent years, learning-based lossy methods have made significant progress, including hybrid and end-to-end methods. Hybrid methods optimize prediction, partitioning, quantization, rate control, and in-loop filter by DNN to improve coding efficiency [25]. End-to-end methods directly use DNN to encode videos [14], [17], [18]. Notably, under some criteria,

end-to-end lossy methods have achieved performance beyond VVC [18]. The lossy method achieves perceptually lossless by adjusting the quality parameters to control the distortion that cannot be distinguished by HVS. Therefore, it is very crucial and necessary to accurately and automatically control the distortion that is imperceptible to HVS.

*3) **Perceptually Lossless Compression**:* Perceptually lossless compression aims to represent video with the smallest file, while still remaining under the visual threshold perceivable by HVS. To achieve this, the visual threshold of HVS has been widely studied, and can be divided into two categories: visual modeling and data-driven methods.

Visual modeling methods [19], [29] research human visual characteristics (e.g., structural sensitivity, fovea) to model visual thresholds. This threshold the guides the existing coding methods to find the best coding parameters through multi-pass coding or preprocessing. In the data-driven method, In data-driven methods [20], [21], [30], the relationship between perceptually lossless QP and the original frame is modeled based on existing perceptually lossless databases [31]–[33].

An innovative mapping between visual thresholds and quality factors using mean square error has been proposed [30], which has achieved improve compression performance in *JPEG* image compression. However, which has shown from *JEPG*, video coding includes prediction coding, which is a non-Markov process. Independent compression for each block can cause error propagation due to rounding, truncation, and filtering [34]. Therefore independent block-level PLC is difficult to migrate to video coding.

B. Motivations

The main method for most visual modeling is to combine various visual characteristics. However, the visual characteristics of HVS have not been thoroughly studied. However, data-driven methods still relies on existing video databases at the sequence-level or frame-level coarse-grained perceptually lossless databases. This prompted us to research block-level fine-grained PLC. The proposed method is conducive to improving PLC performance from the perspective of joint optimization of visual perception and coding systems. First, unlike most PLC methods, the proposed method proposes a video database with block-level perceptually lossless labels, so the performance of the coding system can be more fully exploited. Second, the proposed deep model learns the trade-off relationship between rate and quality features, which is consistent with rate-distortion optimization in video coding. Third, the perceptually lossless rate-distortion relationship is complex and related to video content, which can be fitted by DNN. Therefore, the proposed method deeply fuses the rate and distortion feature, and efficiently integrates visual perception with the coding system.

III. The Proposed Database

In this section, we detail the proposed block-level intra-frame perceptually lossless coding method. We first describe the generation process of our block-level perceptually lossless database, which is beneficial to improve the efficiency of PLC. Since different video contents have different rate-distortion characteristics, it is necessary to select different QP for each coding block.

TABLE I
SOURECE DATABASE STATISTICS INFORMATION.

Database	Resolution	Framerate (fps)	Number
Xiph.org [35]	240p-2160p	25-60	120
SJTU Media [36]	2160p	30	15
MCL-JCV [31]	1080p	30	24-30
Ultra Video Group [37]	1080p	50-120	16
Tencent Video DataSet [38]	2160p	50	86

A. Source data

In order to maintain the authenticity and integrity of our database, we adopt five high-quality and lossless video databases that are currently widely used in the field of video compression to generate our database. The basic statistical information of these databases is shown in Table 1, and the characteristics are described as follows:

Xiph.org Video Test Media [35] is a popular large video data set in the field of video compression. which includes 120 video sequences, covering common resolutions, frame rates and video content in real applications.

SJTU Media Lab proposes a UHD video database [36]. Content and camera settings have been carefully designed to produce diverse and high visual quality videos.

MCL-JCV [31] collects 30 uncompressed sequences which cover a wide range of video features, including multiple video types, multiple targets, and multiple scenes.

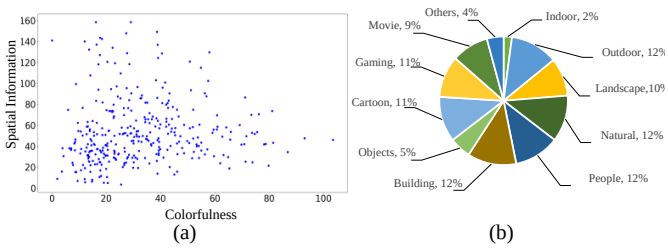
Ultra Video Group (UVG) [37] proposed the first database that provides HFR video (120fps). A higher frame rate makes video content smoother, clearer, and more stereoscopic, and ensures that ghosts are reduced in high-dynamic videos.

Tencent Video DataSet (TVD) [38] is a 4K high-quality video database proposed for learning-based video compression and analysis. The database contains sequences of various static or moving objects, which have been used in the research of neural network video coding (NNVC) by JVET.

The videos in the five databases are very different in resolution ans content, so we have preprocessed the databases to ensure quality of our database and facilitate training. Pre-processing mainly includes two steps of key frame extraction and content cropping. The processed frames are subjected to quality compression and subjective experiments to generate labels. The whole procedure is shown in Fig.3.

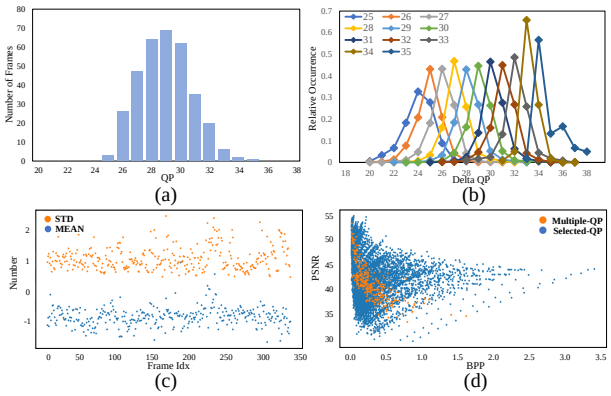
B. Keyframe Extraction

There is a lot of repetitive redundant content and invalid content, such as low dynamic video and pure black/white frames caused by shot switching in videos. Repeated content will lead to a waste of computing resources, falsely high results, and poor model generalization ability in training. Invalid content have little change in rate and distortion under different quality levels, which is not of great research significance. Therefore, in order to avoid the above problems, we manually



(a) (b)

Fig. 3. **Illustration of the database statistic**. (a) The distribution of spatial information and colorfulness of source frames in the proposed database; (b) The distribution of generated frames in a more detailed classification.



(a) (b) (c) (d)

Fig. 4. **Statistics of label**. (a) The distribution of frame-QP. (b) The distribution of CTU-QP offset relative to the frame-QP. (c) The mean and standard deviation of the QP offset in each frame. (d) Rate distortion curves for all distorted versions, with perceptually lossless versions marked with red dots.

extract keyframes from all video sequences to ensure that the keyframes can reflect the main content of the video.

C. Content Cropping

Due to the different resolutions of the source video, we fix the frame size as 1280×720 for the convenience of subsequent research. To keep the data pristine and avoid introducing additional distortion, we use cropping instead of downsampling to generate our database from higher-resolution videos. During the content cropping, we also ensure that the cropped content can reflect the main content in the frame. According to the above selection and cropping rules, total 335 frames are generated. Fig. 2 provides the pipeline of our database generation and some examples in the database.

D. Quality Compression

1) ***Compresssion Optimization****:* VVC adopts a hybrid framework including prediction ($\mathcal{M}_p$), transform($\mathcal{M}_t$), quantization ($\mathcal{M}_q$), and loop filtering ($\mathcal{M}_f$). In the encoding process of each coding unit (CU), the rate-distortion optimization strategy is used to select the best encoding mode. The process can be formulated as:

$$\begin{aligned}&\arg\min_{\mathcal{M}_{p,q,t,f}} \mathcal{J}(CU, QP \mid \mathcal{M}_{p,q,t,f}) = \\ &\mathcal{D}(CU, QP \mid \mathcal{M}_{p,q,t,f}) + \lambda \cdot \mathcal{R}(CU, QP \mid \mathcal{M}_{p,q,t,f}),\end{aligned} \tag{1}$$

where $\mathcal{J}(\cdot)$ denote the calculation function for R-D cost, λ is a Laplacian parameter determined by QP. $\mathcal{R}(\cdot)$ and $\mathcal{D}(\cdot)$ denotes rate cost and distortion respectively.

Numerous studies have shown that maintaining a constant QP policy within each video frame often leads to perceptually suboptimal results. *Multiple-QP* have shown to help improve coding efficiency. The *Multiple-QP* scheme selects a better QP for each CU by brute force search within a range of offset. The RDO process with *Multiple-QP* can be formulated as follows.

$$\begin{aligned}&\arg\min_{\mathcal{M}_{p,q,t,f}, QP_\delta \in S_{QP}} \mathcal{J}_\Delta(CU \mid QP_\delta, \mathcal{M}_{p,q,t,f}) = \\ &\mathcal{D}(CU \mid QP_\delta, \mathcal{M}_{p,q,t,f}) + \lambda \cdot \mathcal{R}(CU \mid QP_\delta, \mathcal{M}_{p,q,t,f}),\end{aligned} \tag{2}$$

where $S_{\Delta QP} = [QP - \Delta QP, QP + \Delta QP]$ is the set of QP candidates. ΔQP is determined by the parameter *MaxDeltaQP* set by the user. In particular, $\mathcal{M}_{p,q,t,f}$ differs only if QP changes, so the R-D cost is determined by QP.

2) ***Compression Detail****:* We generate compressed frames at multiple quality levels to find a perceptually lossless version of each frame. First, a total of 25 quality levels ($S_{QP} = [21, 45]$) are selected empirically, which includes the range of visual distortion from unperceivable to unbearable by human eyes. Second, *VTM16.2* is used as the reference compression platform, and the configuration file selects *encoder_intra_vtm.cfg*, where *MaxDeltaQP* is set to 7 which is the maximum value supported by VVC and can avoid large QP differences results in severe error propagation [34]. Other parameters remain unchanged to ensure consistency with common compression case. Finally, a total of 8710 distorted frames and reference frames were obtained, and the most suitable QP_{map} for each frame was found through subjective experiments. The QP_{map} can be expressed as:

$$QP_{map} = \{QP_{CU_i} \mid CU_i \in I, \arg\min_{\mathcal{M}_{p,q,t,f}, QP_{CU_i} \in S_{QP_{CU_i}}} \mathcal{J}_\Delta\}, \tag{3}$$

where I denotes the input frame that needs to be compressed. QP_{map} is a matrix composed of all CUs' in I. After traversing all QP in S_{QP}, we can get the QP_{map} candidate set $S_{QP_{map}}$ under different distortion levels.

E. Label Generation

The goal of PLC is to get the maximum distortion that HVS cannot perceive, and at the same time the lowest bit rate. Therefore, it can be expressed as follows:

$$\begin{aligned}&\arg\min_{QP_{map} \in S_{QP_{map}}} \mathcal{J}(I \mid QP_{map}) = \lambda \cdot \mathcal{R}(I \mid QP_{map}) + \\ &\mathcal{V}_{Diff}(I, \mathcal{E}nc(I \mid QP_{map})) \ s.t. \ \mathcal{V}_{Diff}(\cdot) = 0\end{aligned}, \tag{4}$$

where $\mathcal{E}nc(\cdot)$ represents the result obtained by compressing input frame I according to QP_{map}, and $\mathcal{V}_{Diff}\{\cdot\}$ represents the difference that the human eye can perceive. Since the reconstructed frames are perceptually lossless, $\mathcal{V}_{Diff}$ is always equal to zero.

Single stimuli and double stimuli are currently widely used subjective testing methods. The purpose of our subjective test

is to use the original image as a reference to find the maximum distortion that is visually perceptually lossless, so the double stimuli method is used.

*1) **Experiment Environment:**:* Subjective experiments are conducted in a controlled laboratory environment. The display devices are two *PHL273V7* 27-inch monitors with 1920×1280 resolution placed side-by-side, and the test software was run on a computer with 64G of memory and a 64-bit Windows operating system. The reference version and the corresponding distorted version were presented side-by-side. Only one subject participated in the test at one time, and the subject was asked to sit in the middle of the two monitors with a distance of 3H [39] from the monitors, keeping the eyes and the middle of the monitors at the same height.

*2) **Experiment Procedure**:* To quickly and accurately find perceptually lossless versions, we employ a binary search strategy [19]. Specifically, a set of images consists of a reference image and corresponding 17 images of different quality levels. First, the images of the intermediate quality level and the reference frame are displayed, and the subjects will be asked to judge whether there is a difference between the two images displayed on the screen. If there is a difference, select the interval with higher quality to repeat the experiment, and vice versa. The process will be repeated until the corresponding perceptually lossless version is found. Since a complete subjective experiment takes a long time, we divide the whole experiment into 3 parts. Each subject completed one part each time, and each part took about 30 minutes. After completing each part, a rest was required before continuing with other parts. 20 subjects (12 males and 8 females), aged 22-26, with normal vision and non-image processing experts, participated in the experiment after training. Finally, we remove outliers and process experimental data according to [19].

F. Database Analysis

*1) **Content Analysis**:* Spatial and temporal information is an important parameter to determine the amount of video compression. Since our research is based on intra-frame compression, we use spatial information (SI) and colorfulness distribution (CF) to quantitatively display the characteristics of our database. As shown in Fig. 3, we can find that our databases are reasonably distributed in the whole range. Furthermore, our database contains urban (indoor, outdoor), nature (natural, landscapes), screen (cartoon, gaming, movie), targets (building, people, objects), and other types, which contain common content in high-quality video, as shown in Fig. 3.

*2) **Label Analysis**:* Based on subjective experiments, we obtained a perceptually lossless label QP_{map}. Fig. 4 shows the statistical information of the label. It can be seen from the figure that the frame-level QP distribute within the interval [23,35], and the average value is 29. The distribution of delta QP shows that most of the CTU-level QP (about 74%) has shifted, which shows the necessity of generating CTU-level labels. In addition, the mean and standard deviation of delta QP shows that the overall QP offsets downward, and the degree of discreteness within a frame further illustrates that the different areas of the content have different rate-distortion characteristics. Therefore, it is necessary to use CTU-level QP to improve compression efficiency. Finally, the rate-distortion curves of all distorted versions and perceptually lossless versions in the database are displayed. According to the statistical information, we can see the rationality and necessity of our database and labels.

IV. The Predictive Network Model

A. Problem Formulation

As mentioned above, our goal is PLC under the joint optimization of perception and coding. Since visual perception is the perception of the frame as a whole, and CTUs are not completely independent in the compression process, in order to accurately and efficiently predict QP_{map}, we use DNN network to directly model the mapping between input frame and QP_{map}.

$$QP^*_{map} = \mathcal{F}_{cnn}(I) \tag{5}$$

B. Network Design

From our database establishment process, PLC can be seen as a rate-distortion trade-off process. Inspired by this, $\mathcal{F}_{cnn}$ is designed to first extract the rate and quality features, then trade-off and fuse in the feature space, and finally decoded them into QP_{map}. Fig.9 provides the overall framework of our network, which mainly includes four modules: rate encoder $\mathcal{F}^R_{Enc}$, distortion encoder $\mathcal{F}^D_{Enc}$, feature trade-off $\mathcal{F}_{Trade-off}$, and feature decoder $\mathcal{F}_{Dec}$.

Rate Encoder. Deep image compression (DIC) uses DNN to extract compact feature map representation of images, which is quantized and dequantized and then passed through a decoder network to reconstruct images. The actual bit rate of DIC depends on the entropy of the quantized feature map, and entropy coding is a lossless process, so it can be considered that the feature map directly determines the rate. Therefore, we use the encoder in the latest DIC method [40] to extract bit rate features, and the optimized entropy coding model of [40] can guide the encoder to learn more sufficient rate features. The rate encoder module can be expressed as:

$$\mathcal{F}^R_{Enc} = Conv(W_R, \mathcal{F}^R_{Enc_L}(\cdots \mathcal{F}^R_{Enc_i}(\mathcal{F}^R_{Enc_1}(I)))), i \in [1, L], \tag{6}$$

where $\mathcal{F}^R_{Enc_i}$ represents a cascade of generalized divisive normalization (GDN) and convolution layer, and GDN will be disabled when i=4. $Conv(\cdot)$ denotes the convolution operation, W_R represents the learned weights. Specifically, it can be formulated as:

$$\mathcal{F}^R_{Enc_i} = GDN(Conv(W_i, X)), \tag{7}$$

where W_i represents the learned weights, and X represents the input feature map.

Distortion Encoder. Learning-based image quality assessment (IQA) methods use DNN to extract perceptual quality features, and a decoder network predicts an image's quality score based on these features. [41] uses both local and global quality features, which coincides with the idea of our method.

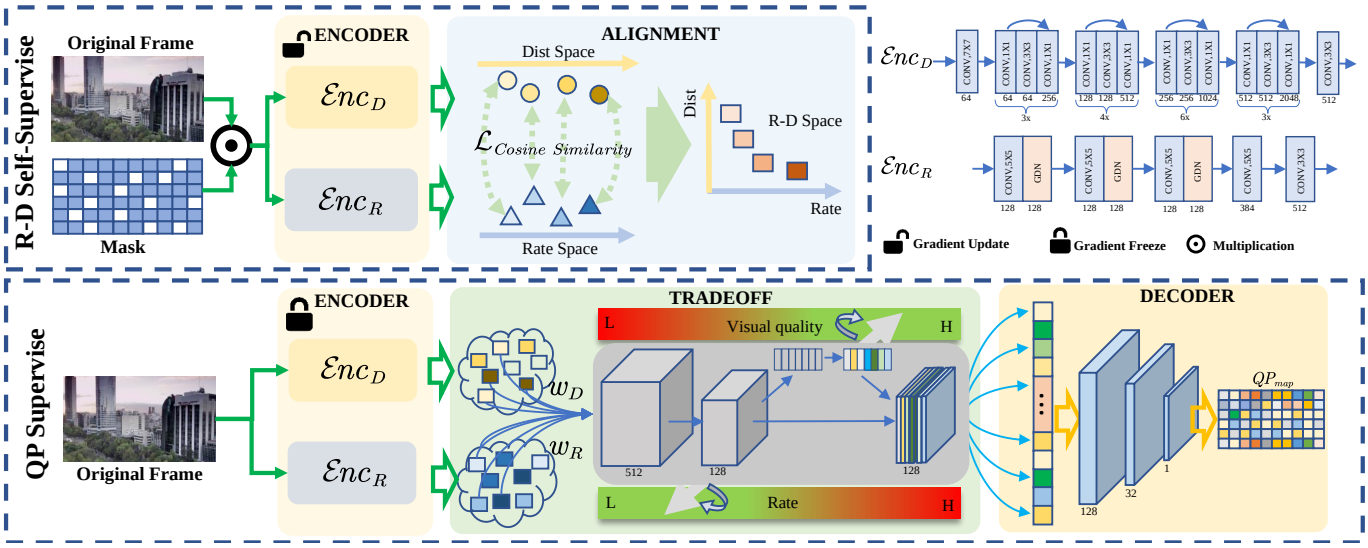


Fig. 5. **Pipeline of the proposed QP_{map} prediction network**. In the pre-training stage, we employ masking and self-supervised to map rate and distortion features into RDO space. In the training stage, the feature encoder network gradients are frozen, rate and distortion features are fusion and trade-off by the trade-off module, and the QP_{map} output by the feature decoder.

Therefore, the encoder in [41] is used as our distorted feature encoder, which can be represented by:

$$\mathcal{F}^D_{Enc} = Conv(W_D, ResNet(I)), \tag{8}$$

where W_D represents the learned weights, and $ResNet(\cdot)$ denotes a ResNet-50 network.

Feature Trade-off. According to the above discussion, the rate and distortion features have different characteristics, and their contribution to perceptually lossless compression is usually unequal. To solve this problem, we design to add a learnable weight to each feature and let the network learn the trade-off relationship between two features. In addition, a channel attention module is added to further weigh the importance of features. Based on this idea, the feature trade-off can be formulated as:

$$\mathcal{F}_{Trade-off} = SE(Conv(W_T, \frac{w_R \cdot \mathcal{F}^R_{Enc} + w_D \cdot \mathcal{F}^D_{Enc}}{w_R + w_D + \epsilon})), \tag{9}$$

where w_R and w_D are learnable weights, ϵ is set as 0.0001 to avoid division by zero. SE is a channel attention module.

Feature Decoder. The decoder module consists of three convolution layers that downsample the fusion feature map to the size of QP_{map}, which can be formulated by:

$$\mathcal{F}^i_{Dec} = Conv(W^i_O, \mathcal{F}_{Trade-off}) + b^i_O, i \in [1, 3], \tag{10}$$

where W^i_O and b^i_O denote weight and bias parameters of i-th convolution layer.

C. Learning Mechanism

In order to achieve a balance between perceptually lossless distortion and rate, we adopt a pre-training and training two-stage learning strategy

R-D Self-Supervise Learning. To balance rate and distortion in PLC, we adopt a self-supervised contrastive learning method to align the distributions of the two features into the same distribution space. Specifically, we perform self-supervised training with masked images before training which is inspired by [42], as shown in Fig. 5. In order to be in line with the coding structure of VVC, we use a 128×128 mask which is the size of CTU to mask the input frame. We used the following similarity function to guide the comparative learning process:

$$\underset{\{\boldsymbol{\theta}^D, \boldsymbol{\theta}^R\}}{\arg\max} \frac{\mathcal{F}^R_{Enc}\{I_m \mid \boldsymbol{\theta}^R\} \cdot \mathcal{F}^D_{Enc}\{I_m \mid \boldsymbol{\theta}^D\}}{max(\|\mathcal{F}^R_{Enc}\{I_m \mid \boldsymbol{\theta}^R\}\|_2 \cdot \|\mathcal{F}^D_{Enc}\{I_m \mid \boldsymbol{\theta}^D\}\|_2, \epsilon)}, \tag{11}$$

where ϵ is a small value to avoid division by zero, I_m denoted masked image. $\boldsymbol{\theta}^D and \boldsymbol{\theta}^R$ are the weights of $\mathcal{F}^D_{Enc} and \mathcal{F}^R_{Enc}$.

Supervision Learning. We regard QP prediction as a regression task and expect neural networks to predict relatively accurate values. In order to make the compression effect that the predicted value can achieve as close as possible to perceptually lossless, we divide the loss function into rate loss and distortion loss. The rate loss corresponds to the predicted QP less than the label, and the distortion loss corresponds to the predicted QP greater than the label. The final loss function can be expressed as follows:

$$\underset{\boldsymbol{\theta}_{Dec}, \boldsymbol{\theta}_{Trade-off}}{\arg\min} \quad \mathcal{L} = \frac{1}{N}\|QP'_{map} - QP^*_{map}\|, \tag{12}$$

where $\boldsymbol{\theta}_{Trade-off}$ and $\boldsymbol{\theta}_{Dec}$ are the weights of $\mathcal{F}_{Trade-off}$ and $\mathcal{F}_{Dec}$, QP'_{map} is the predicted value, and QP^*_{map} denotes the corresponding label.

V. Experimental Evaluation

In this section, we describe the analysis and comparison results in detail to demonstrate the performance of the proposed method from different perspectives.

TABLE II
BITS PER PIXEL (BPP) COMPARISONS UNDER THE SAME VISUAL QUALITY OF JVET TEST SEQUENCES.

Class	Sequence	AVC	HEVC	VVC	JPEG	Mentze2019	Mentze2020	Rhee2022	Bai2021	Hu2021	Li2022	Proposed
Class A 3840x2160	Campfire	7.0419	6.8828	6.2419	4.1807	10.4550	8.7120	7.6143	2.1166	0.5917	0.5052	0.2333
	FoodMarket4	6.5310	6.8862	6.3218	4.2111	8.8301	8.5680	6.1782	1.3373	0.2477	0.2322	0.1065
	Tango2	7.6827	7.8543	7.4846	4.4343	9.5645	9.1020	7.5288	1.6202	0.2260	0.2022	0.0607
	CatRobot1	8.5387	8.5652	8.2462	5.3064	11.5836	10.7880	9.4373	1.9533	0.4630	0.3643	0.1322
	DaylightRoad2	8.9099	8.9470	8.5574	5.5692	11.3589	10.7430	6.7881	2.0099	0.4914	0.3894	0.1475
	ParkRunning3	9.0939	9.2801	8.6010	5.2831	13.9389	11.3220	10.4544	3.2150	1.0039	0.8640	0.4461
	MEAN	**7.9714**	**8.0693**	**7.5755**	**4.8308**	**10.9552**	**9.8725**	**8.0002**	**2.0421**	**0.5040**	**0.4262**	**0.1877**
Class B 1920x1080	BasketballDrive	5.3442	5.4748	5.1768	4.8671	11.0803	10.7280	9.6557	2.0878	0.5536	0.4390	0.2630
	BQTerrace	6.6758	6.2516	5.8881	5.6502	12.0853	11.5680	10.8482	2.3450	1.1097	0.8959	0.4880
	Cactus	6.2826	6.1872	5.8875	6.1798	12.3510	11.8800	11.1391	2.3290	0.8146	0.6919	0.3940
	MarketPlace	8.5334	8.8069	8.2214	5.1259	10.6511	9.9780	8.4347	2.0086	0.5855	0.5939	0.3669
	RitualDance	6.9099	6.8429	6.2786	3.6975	8.8923	7.8750	6.1309	1.4965	0.3695	0.3486	0.2328
	MEAN	**6.7492**	**6.7127**	**6.2905**	**5.1041**	**11.0120**	**10.4058**	**9.4536**	**2.0534**	**0.6866**	**0.5939**	**0.3489**
Class C 832x480	BasketballDrill	5.8859	5.6946	5.2971	6.2392	11.6807	11.2770	10.7153	2.3797	0.9550	0.7725	0.5758
	BQMall	5.6603	5.7384	5.2847	5.8800	11.0674	10.9410	10.2325	2.1776	0.9146	0.7225	0.5942
	PartyScene	8.3094	7.2006	6.6977	7.5959	13.0972	12.8010	12.5762	3.1447	1.9748	1.5038	1.4427
	RaceHorses	6.2044	5.9666	5.4769	5.8443	11.1776	11.0910	10.4556	2.4305	1.1355	0.9068	0.8085
	MEAN	**6.5150**	**6.1501**	**5.6891**	**6.3899**	**11.7557**	**11.5275**	**10.9949**	**2.5377**	**1.2450**	**0.9764**	**0.8553**
Class D 416x240	BasketballPass	5.1814	5.5722	4.7912	5.4654	10.4623	10.0350	9.3503	2.3729	1.0424	0.7450	0.5812
	BlowingBubbles	8.5459	7.4651	6.9371	7.8031	13.7203	13.2360	13.2252	3.1349	2.1119	1.4644	1.3291
	BQSquare	8.2435	7.0109	6.5760	7.5523	12.2654	12.3510	12.0097	3.0548	2.0684	1.4664	1.3163
	RaceHorses	6.6256	6.4184	5.7881	6.2315	11.1776	11.3940	11.0184	2.7491	1.4047	1.0330	0.8613
	MEAN	**7.1491**	**6.6167**	**6.0231**	**6.7631**	**11.9064**	**11.7540**	**11.4009**	**2.8279**	**1.6566**	**1.1772**	**1.0220**
Class E 1280x720	FourPeople	4.3282	4.6951	4.1355	4.2739	8.8674	9.1140	7.1587	1.6363	0.4987	0.3999	0.3200
	Johnny	4.0234	4.3075	3.8687	3.8611	8.4962	8.8560	6.7317	1.5966	0.3789	0.2942	0.1790
	KristenAndSara	3.9878	4.2872	3.8120	3.8678	8.4510	8.8890	6.7881	1.6977	0.3777	0.3042	0.2184
	MEAN	**4.1131**	**4.4299**	**3.9387**	**4.0009**	**8.6049**	**8.9530**	**6.8928**	**1.6435**	**0.4184**	**0.3328**	**0.2391**
Class F 1280x720	BasketballDrillText	5.7396	5.5600	5.1154	6.3501	11.8412	11.0310	10.4966	2.6424	1.0894	0.8517	0.6861
	ChinaSpeed	3.4761	3.9052	3.2316	4.0935	7.7818	7.0680	5.5254	2.4774	0.9327	0.7519	0.5449
	SlideEditing	3.0070	3.0362	2.4694	4.9201	9.9566	7.1700	6.3316	2.9461	1.8267	2.0885	0.6759
	SlideShow	1.1442	1.2703	1.0367	1.4823	6.2564	2.6070	1.7704	1.1683	0.3649	0.2535	0.1548
	MEAN	**3.3417**	**3.4429**	**2.9858**	**4.2115**	**8.9590**	**6.6960**	**6.0310**	**2.3086**	**1.0534**	**0.9864**	**0.5154**
MEAN		**5.9733**	**5.9036**	**5.4171**	**5.2167**	**10.5322**	**9.8681**	**8.7956**	**2.2355**	**0.9273**	**0.7488**	**0.5281**

A. Experimental Protocol

1) Database Description.: To demonstrate the effectiveness and robustness of our method, we conduct experiments on both the proposed dataset and the standard test sequences [43].

Proposed Database: The proposed database contains 335 frames and corresponding CTU-level PLC labels, the detail can be found in Section III. We randomly select image pairs from the dataset as the training set, validating set, and testing set based on the ratio of 8:1:1.

JVET Test Sequence: The official test sequence consists of 26 videos. These videos include multiple resolutions (240p-2160p), multiple frame rates (20fps-60fps), and multiple video types. the detail of these videos can be found in [43]. During testing, we set up the configuration file according to common test conditions (CTC).

2) Evaluation Criteria.: Both objective and subjective evaluation criteria are adopted to evaluate the compression performance and visual quality of different methods.

Objective Criteria: Video Multi-Method Assessment Fusion (VMAF) [44] and Multiscale Structural Similarity (MS-SSIM) [45] are two objective quality evaluation indicators. Because these two indicators are more similar to human visual perception, they are widely used in the field of compression. VMAF is a criterion trained on a large subjective database using support vector machines and mainly uses three indicators: video quality fidelity, detailed loss measure, and temporal information. According to the test manual, we use the 4K model for class A and use the basic model for other classes. All the experiments of VMAF are tested by *FFmpeg*. MS-SSIM measures visual quality by considering the structural distortion of images at different scales, the test experiments are performed by the function *multissim* in *MATLAB* toolbox. For better perceptual quality, the values of VAMF are close to 100, and MS-SSIM is close to 1 while the value of bit per pixel (BPP) is smaller, which means that it acts highly performance in high-quality video service.

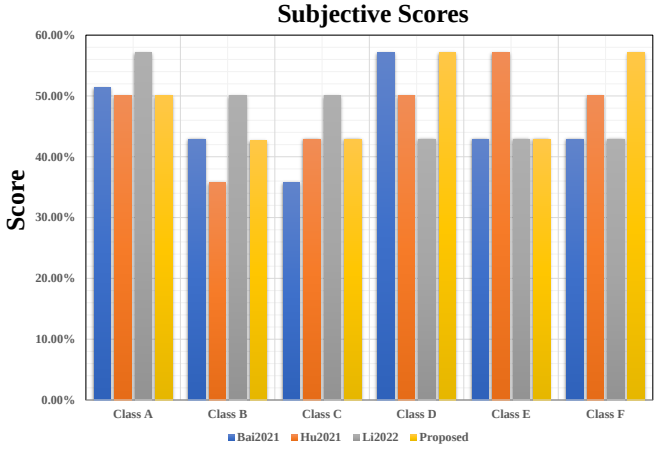


Fig. 6. **Subjective comparison results of the proposed method with three state-of-the-art schemes**. The "Score" obtained by 2AFC refers to the percentage of the video compressed by the PLC method relative to the original version.

Subjective Criteria: To compare visual quality more ac-

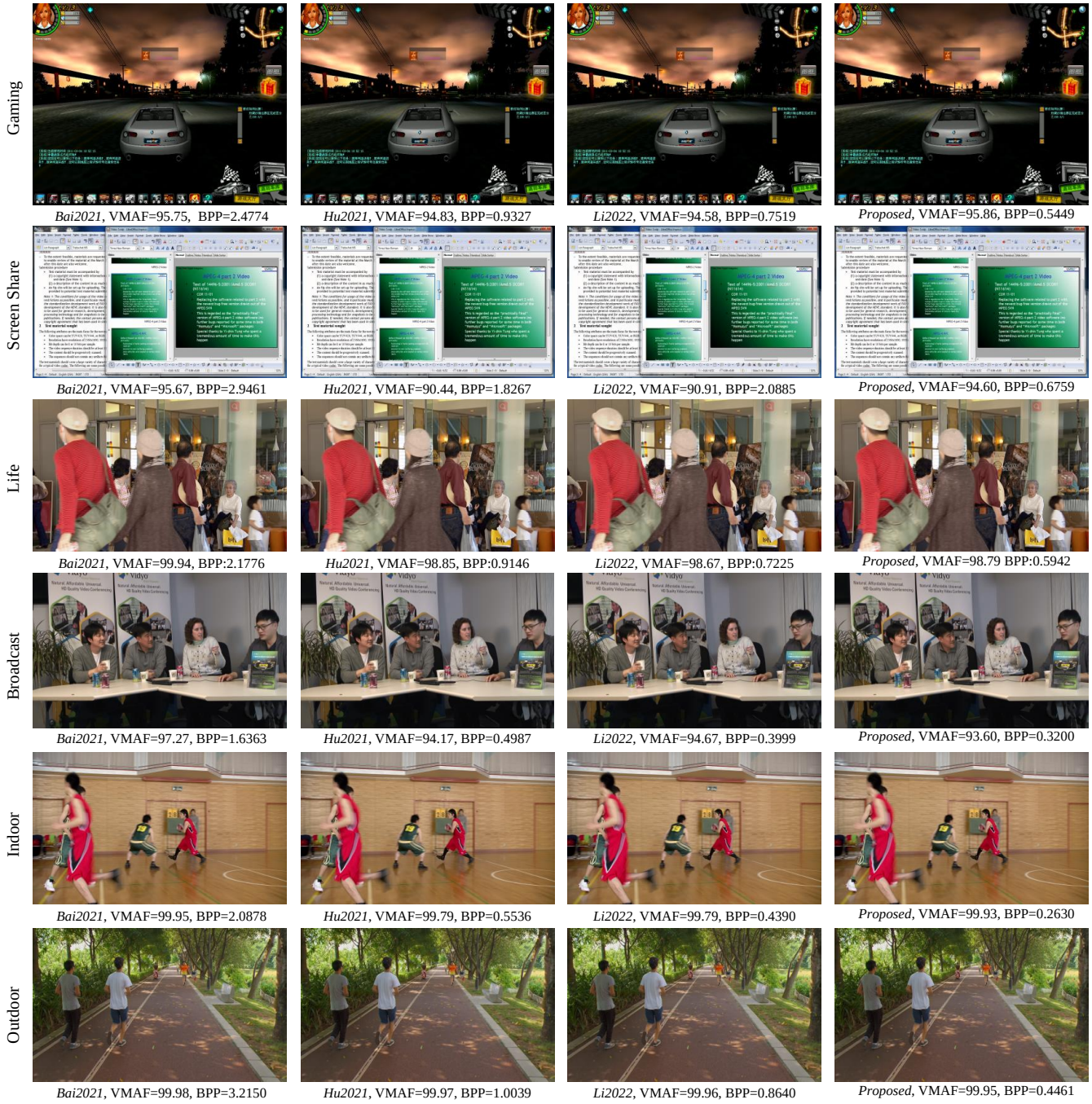


Fig. 7. **Visual comparisons of the compressed frames among four methods**. Each frame is marked with VMAF value and BPP value. **Please zoom in to view the details**

curately and fairly, two-alternative forced choice (2AFC) is employed for subjective evaluation. A total of 20 volunteers (12 males, and 8 females) were invited. The subjective quality comparison results are shown in Fig 6. Specifically, a subject is shown a pair of videos (one is raw video, another is compressed by a perceptually lossless method) and asked to choose the one with better perceptual quality. Each pair is repeated two times with a random order in each trial.

3) Comparison Methods: In order to demonstrate the high quality and compression ratio of our proposed method, we compare a total of 10 methods with standard video coding and deep learning-based lossless and PLC method. For a fair comparison, all the standard methods compress according to the CTC, and the results of learning-based methods obtained by running the released implementation from the related authors.

Lossless Methods. For the lossless method, we use three existing video coding standards AVC, HEVC, and VVC for comparison, all tested in intra and lossless mode. In addition, the lossless mode of JPEG (QF=100) and the three latest lossless intra-frame compression methods *.Mentzer2019* [15], *Mentzer2020* [16], *Rhee2022* [28] based on DNN are com-

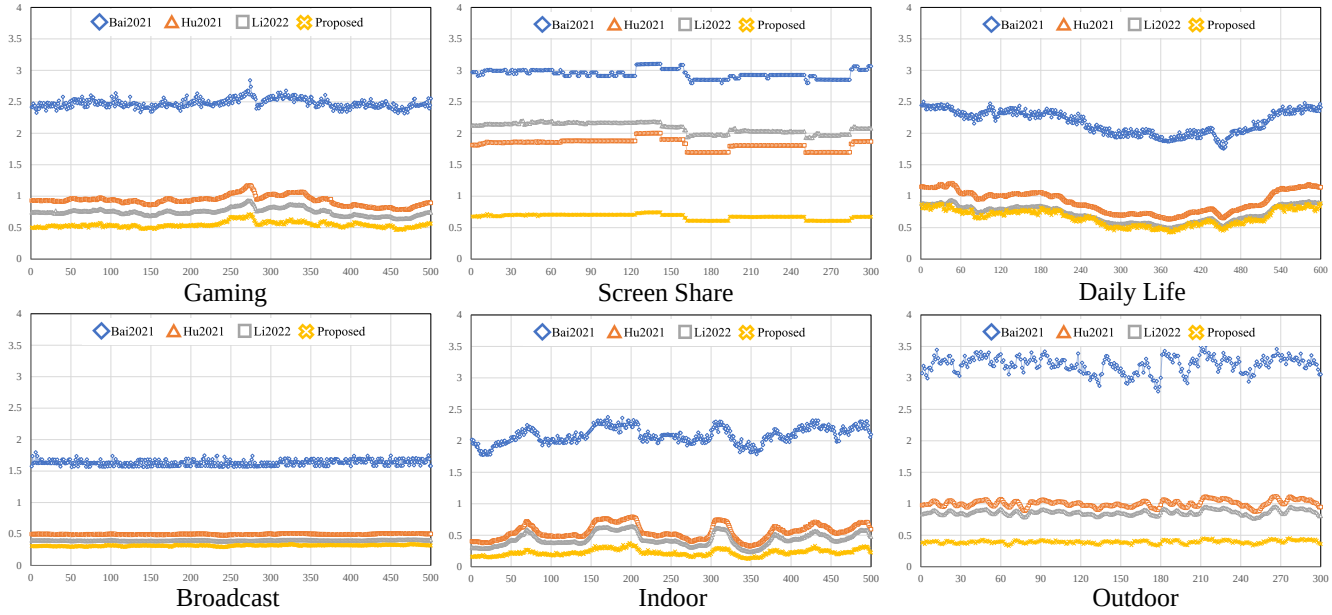


Fig. 8. BPP trend of the compressed sequences among four methods.

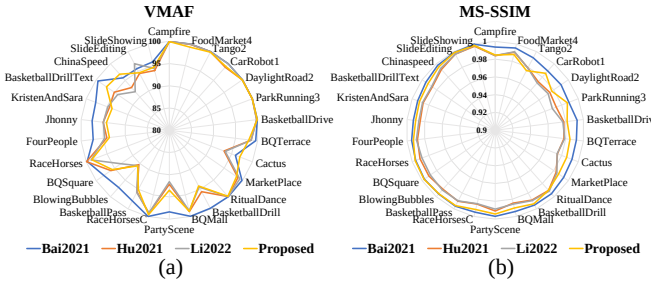


Fig. 9. **Visual quality comparison of standard test sequence among four methods under VMAF and MS-SSIM**.

pared.

PLC Methods. We also selected three recent representative PLC methods, including near-lossless *Bai2021* [14] and two PLC methods *Hu2021* [17], and *Li2022* [18]. In order to ensure the fairness of the comparison, We guarantee that all comparison methods are compared under the same perceptually lossless quality.

4) Implementation Details: All the experiments are conducted on a server equipped with *Xeon W-2265* CPU, 64G memory. We adopt the python toolbox *PyTorch* on the GPU *NVIDIA TITAN RTX 24G*. The input image sample resolution is 1280×720. In the self-supervised pre-training and supervised training stages, the network is trained according to Eq. 11 and Eq. 12, respectively. The learning rate and batch size of both pre-training and training are set to 0.00005 and 8 respectively. Adaptive moment estimation (Adam) is used to optimize the network parameters. We self-supervised pre-training the network for 10 epochs and supervised training for 300 epochs.

B. Analysis on Overall performance

Table 1 shows the compression results on the standard test sequence at the same perceptually lossless quality, and the subjective results of each class are shown in Fig. 6. It can be seen that our method cost the lowest rate to achieve perceptually lossless quality. Moreover, the proposed method can achieve stable performance at different resolutions and content, which proves the robustness of the proposed method. It is worth noting that our method improves coding efficiency more significantly than existing methods at higher resolutions. This is due to the good combination of our model and VVC, which can provide better high-quality video coding services.

Fig. 7 exhibits visual quality comparison results of six common cases in high-quality video applications. The BPP and VMAF results are presented for reference. It can be seen that the test frames compressed by the other three methods have the same perceptual quality. Furthermore, we have also provided a completed BPP trend as shown in Fig. 8. It can be seen that our method maintains a stable advantage over the whole sequence.

Without losing fairness, we also use objective evaluation criteria to make a comprehensive comparison. As shown in Fig. 9, two mainstream objective criteria (VMAF and MS-SSIM) similar to human visual perception are adopted. It can be seen that the near-lossless method *Bai2021* [14] has achieved the highest performance in both indicators. Our method is comparable to the perceptually lossless methods *Hu2021* [17] and *Li2022* [18] have similar performance, but the proposed method obtains fewer bits, in which the average bits saving is 29.47% compared with the latest method *Li2022*.

C. Analysis on Network Module

In this section, we study the effectiveness of each network module designed in Section IV. We generated three models

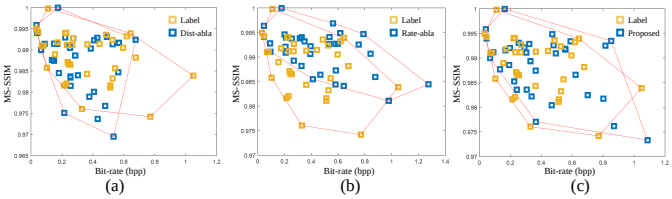


Fig. 10. **Comparative results of ablation experiments.** The overlapping area of each model envelope and label envelope: (a): 0.6647, (b): 0.5419, (c): 0.9087

to verify the effectiveness of our various network modules. Specifically, all experimental settings are as described in Section IV, and we use BPP and MS-SSIM performance as criteria for network performance.

Fig. 10 exhibits the ablation results on rate encoder, distortion encoder, and trade-off module. It can be found that the QP prediction trend of the two feature encoder networks has obvious characteristics. Specifically, the extracted features of the distortion encoder are related to quality, so the prediction results focus on higher quality (smaller QP), resulting in high visual quality while consuming more bits. The rate encoder leads to the opposite result. Therefore, it is necessary to introduce a weighing module to weigh the rate-distortion relationship in PLC. It can be found from Fig. 10 that the compressed results of the prediction QP_{map} are more consistent with the labels after the introduction of the trade-off module.

D. Analysis on Computational Complexity.

Coding complexity is one of the important considerations in high-quality video services, so we conducted a complexity analysis of the proposed PLC method at common resolutions. We divide the whole process into two stages: inference and compression. The inference process is conducted on a *NVIDIA RTX TITAN 24G* GPU. In the compression part, VTM compresses according to the QP_{map} obtained in the inference stage. We also test our models without GPU. Table III shows the average inference and encoding times for six resolutions in the standard test sequence on a computer with a *Xeon W-2265 3.5GHz* CPU and 64G memory. As can be seen from the figure, integrating our proposed method into VTM, the inference stage consumes less than 2% of the coding time, which is tolerable in practical applications.

TABLE III
COMPARATIVE ANALYSIS OF COMPUTATION COMPLEXITY.

Resolution	VTM	Proposed CTU (s)		Proposed GPU (s)	
		Total	Infer	Total	Infer
416×240	15.85	16.22	0.36	16.80	0.95
832×480	63.83	64.95	1.12	64.95	0.98
1024×768	125.40	127.47	2.07	126.40	1.00
1280×720	147.30	149.71	2.41	148.32	1.02
1920×1080	330.75	336.18	5.43	331.77	1.02
3840×2160	1323.00	1342.44	19.44	1324.26	1.26
Complexity	**100.00%**	**101.74%**	**-**	**101.57%**	**-**

VI. Conclusion

In this paper, a novel perceptually lossless coding method is proposed. Specifically, we design a new block-levle database, which provides the label jointly optimized by visual perceptual and coding system. Morever, to establish the mapping between input video frames and quantization parameters (QP), we design a deep neural network to explore rate-distortion trade-off for PLC. The experimental results verify the superiority of this scheme in maintaining high visual quality and obtaining lower bit rate compared with the latest methods.

Many further work can be explored. First of all, we are currently research on the largest coding unit (128×128), and the coding unit QP control can be more refined. In the future, we will try to control QP on a smaller CU to further reduce the code rate of PLC. However, smaller CUs mean more complex modeling scenarios (e.g., multi-type tree partition), which need further research. Second, we currently perform visual lossless encoding in intra-frame, the methods can be also implemented on inter-mode, such as random access and low delay mode.

References

[1] Thomas Barnett, Shruti Jain, Usha Andra, and Taru Khurana. Cisco visual networking index (vni) complete forecast update, 2017–2022. *Americas/EMEAR Cisco Knowledge Network (CKN) Presentation*, 1(1):1–30, 2018.

[2] Vignesh V Menon, Hadi Amirpour, Mohammad Ghanbari, and Christian Timmerer. CODA: Content-aware Frame Dropping Algorithm for High Frame-rate Video Streaming. In *IEEE Data Compression Conference (DCC)*, pages 475–475, 2022.

[3] Lingyu Duan, Jiaying Liu, Wenhan Yang, Tiejun Huang, and Wen Gao. Video coding for machines: A paradigm of collaborative compression and intelligent analytics. *IEEE Transactions on Image Processing*, 29:8680–8695, 2020.

[4] Lei Zhao, Shiqi Wang, Shanshe Wang, Yan Ye, Siwei Ma, and Wen Gao. Enhanced surveillance video compression with dual reference frames generation. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(3):1592–1606, 2021.

[5] Siqi Huang, Jiang Linda Xie, and Muhana Muslam. A cloud computing based deep compression framework for UHD video delivery. *IEEE Transactions on Cloud Computing*, 1(1):1–13, 2022.

[6] Benjamin Bross, Jianle Chen, Jens-Rainer Ohm, Gary J Sullivan, and Ye-Kui Wang. Developments in international video coding standardization after avc, with an overview of versatile video coding (vvc). *Proceedings of the IEEE*, 109(9):1463–1493, 2021.

[7] Yun Zhang, Sam Kwong, and Shiqi Wang. Machine learning based video coding optimizations: A survey. *Elsevier, Information Sciences*, 506(1):395–423, 2020.

[8] Benjamin Bross, Ye-Kui Wang, Yan Ye, Shan Liu, Jianle Chen, Gary J Sullivan, and Jens-Rainer Ohm. Overview of the versatile video coding (VVC) standard and its applications. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(10):3736–3764, 2021.

[9] Gary J Sullivan, Jens-Rainer Ohm, Woo-Jin Han, and Thomas Wiegand. Overview of the high efficiency video coding (HEVC) standard. *IEEE Transactions on circuits and systems for video technology*, 22(12):1649–1668, 2012.

[10] Hongyue Huang, Ionut Schiopu, and Adrian Munteanu. Deep learning based angular intra-prediction for lossless HEVC video coding. In *IEEE Data Compression Conference (DCC)*, pages 579–579, 2019.

[11] Santiago De-Luxán-Hernández, Gayathri Venugopal, Valeri George, Heiko Schwarz, Detlev Marpe, and Thomas Wiegand. A fast lossless implementation of the intra subpartition mode for VVC. In *IEEE International Conference on Image Processing (ICIP)*, pages 1118–1122, 2020.

[12] Changyue Ma, Dong Liu, Xiulian Peng, Li Li, and Feng Wu. Convolutional neural network-based arithmetic coding for HEVC intra-predicted residues. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(7):1901–1916, 2019.

[13] Zhijie Huang, Jun Sun, Xiaopeng Guo, and Mingyu Shang. One-for-all: An efficient variable convolution neural network for in-loop filter of vvc. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(4):2342–2355, 2021.

[14] Yuanchao Bai, Xianming Liu, Wangmeng Zuo, Yaowei Wang, and Xiangyang Ji. Learning scalable ly=-constrained near-lossless image compression via joint lossy image and residual compression. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11946–11955, 2021.

[15] Fabian Mentzer, Eirikur Agustsson, Michael Tschannen, Radu Timofte, and Luc Van Gool. Practical full resolution learned lossless image compression. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10629–10638, 2019.

[16] Fabian Mentzer, Luc Van Gool, and Michael Tschannen. Learning better lossless compression using lossy compression. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6638–6647, 2020.

[17] Yueyu Hu, Wenhan Yang, and Jiaying Liu. Coarse-to-fine hyper-prior modeling for learned image compression. In *AAAI Conference on Artificial Intelligence (AAAI)*, volume 34, pages 11013–11020, 2020.

[18] Jiahao Li, Bin Li, and Yan Lu. Hybrid spatial-temporal entropy modelling for neural video compression. In *ACM International Conference on Multimedia (ACMMM)*, pages 1503–1511, 2022.

[19] Xuelin Shen, Zhangkai Ni, Wenhan Yang, Xinfeng Zhang, Shiqi Wang, and Sam Kwong. Just noticeable distortion profile inference: a patch-level structural visibility learning approach. *IEEE Transactions on Image Processing*, 30:26–38, 2021.

[20] Sanaz Nami, Farhad Pakdaman, Mahmoud Reza Hashemi, and Shervin Shirmohammadi. BL-JUNIPER: A CNN-Assisted Framework for Perceptual Video Coding Leveraging Block-Level JND. *IEEE Transactions on Multimedia*, 1(1):1–16, 2022.

[21] Qin Huang, Haiqiang Wang, Sung Chang Lim, Hui Yong Kim, Se Yoon Jeong, and C-C Jay Kuo. Measure and prediction of HEVC perceptually lossy/lossless boundary QP values. In *IEEE Data Compression Conference (DCC)*, pages 42–51, 2017.

[22] Tao Tian, Hanli Wang, Sam Kwong, and C-C Jay Kuo. Perceptual image compression with block-level just noticeable difference prediction. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 16(4):1–15, 2021.

[23] Jongho Kim, Dae Yeol Lee, Seyoon Jeong, and Seunghyun Cho. Perceptual video coding using deep neural network based jnd model. In *IEEE Data Compression Conference (DCC)*, pages 375–375, 2020.

[24] Miaohui Wang, Jian Xiong, Long Xu, Wuyuan Xie, King Ngi Ngan, and Jing Qin. Rate constrained multiple-QP optimization for HEVC. *IEEE Transactions on Multimedia*, 22(6):1395–1406, 2019.

[25] Dandan Ding, Zhan Ma, Di Chen, Qingshuang Chen, Zoe Liu, and Fengqing Zhu. Advances in video compression system using deep neural network: A review and case studies. *Proceedings of the IEEE*, 109(9):1494–1520, 2021.

[26] Yun Zhang, Linwei Zhu, Gangyi Jiang, Sam Kwong, and C-C Jay Kuo. A survey on perceptually optimized video coding. *ACM Computing Surveys*, 55(12):1–37, 2021.

[27] Benjamin Bross, Tung Nguyen, Heiko Schwarz, Detlev Marpe, and Thomas Wiegand. Lossless coding support in VVC. In *Applications of Digital Image Processing XLIII*, volume 11510, page 115101B, 2020.

[28] Hochang Rhee, Yeong Il Jang, Seyun Kim, and Nam Ik Cho. LC-FDNet: Learned Lossless Image Compression with Frequency Decomposition Network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6033–6042, 2022.

[29] Yiming Li, Hongyi Liu, and Zhenzhong Chen. Perceptually-lossless image coding based on foveated-JND and H. 265/HEVC. *Elsevier Journal of Visual Communication and Image Representation*, 40(1):600–610, 2016.

[30] Miaohui Wang, Zhuowei Xu, Xueqin Liu, Jian Xiong, and Wuyuan Xie. Perceptually Quasi-Lossless Compression of Screen Content Data via Visibility Modeling and Deep Forecasting. *IEEE Transactions on Industrial Informatics*, 18(10):6865–6875, 2022.

[31] Haiqiang Wang, Weihao Gan, Sudeng Hu, Joe Yuchieh Lin, Lina Jin, Longguang Song, Ping Wang, Ioannis Katsavounidis, Anne Aaron, and C-C Jay Kuo. MCL-JCV: a JND-based H. 264/AVC video quality assessment dataset. In *IEEE International Conference on Image Processing (ICIP)*, pages 1509–1513, 2016.

[32] Haiqiang Wang, Ioannis Katsavounidis, Jiantong Zhou, Jeonghoon Park, Shawmin Lei, Xin Zhou, Man-On Pun, Xin Jin, Ronggang Wang, Xu Wang, et al. VideoSet: A large-scale compressed video quality dataset based on JND measurement. *Elsevier Journal of Visual Communication and Image Representation*, 46:292–302, 2017.

[33] Xuelin Shen, Zhangkai Ni, Wenhan Yang, Xinfeng Zhang, Shiqi Wang, and Sam Kwong. A JND dataset based on VVC compressed images. In *IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 1–6, 2020.

[34] Tiesong Zhao, Yuhang Huang, Weize Feng, Yiwen Xu, and Sam Kwong. Efficient VVC Intra Prediction Based on Deep Feature Fusion and Probability Estimation. *IEEE Transactions on Multimedia*, 1(1):1–11, 2022.

[35] C Montgomery and H Lars. Xiph. org Video Test Media (Derf's Collection. *the xiph open source community*, 1994.

[36] Li Song, Xun Tang, Wei Zhang, Xiaokang Yang, and Pingjian Xia. The SJTU 4K video sequence dataset. In *2013 Fifth International Workshop on Quality of Multimedia Experience (QoMEX)*, pages 34–35, 2013.

[37] Alexandre Mercat, Marko Viitanen, and Jarno Vanne. UVG dataset: 50/120fps 4K sequences for video codec analysis and development. In *ACM Multimedia Systems Conference (MMSys)*, pages 297–302, 2020.

[38] Xiaozhong Xu, Shan Liu, and Zeqiang Li. Tencent video dataset (TVD): A video dataset for learning-based visual data compression and analysis. *arXiv preprint arXiv:2105.05961*, 2021.

[39] Radiocommunication Sector. The Present State of Ultra-high Definition Television. *International Telecommunication Union*, 2015.

[40] Yichen Qian, Ming Lin, Xiuyu Sun, Zhiyu Tan, and Rong Jin. Entroformer: A Transformer-based Entropy Model for Learned Image Compression. In *International Conference on Learning Representations (ICLR)*, pages 1–1, 2022.

[41] Shaolin Su, Qingsen Yan, Yu Zhu, Cheng Zhang, Xin Ge, Jinqiu Sun, and Yanning Zhang. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3667–3676, 2020.

[42] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16000–16009, 2022.

[43] Xiang Li Jill Boyce, karsten Suehring. JVET common test conditions and software reference configurations. *JVET-J1010*, 2018.

[44] Zhi Li, Anne Aaron, Ioannis Katsavounidis, Anush Moorthy, and Megha Manohara. Toward a practical perceptual video quality metric. *The Netflix Tech Blog*, 6(2):2, 2016.

[45] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers*, 2(1):1398–1402, 2003.