Referenceless Rate-Distortion Modeling with Learning from Bitstream and Pixel Features

Yangfan Sun
University of Missouri-Kansas City
Kansas City, Missouri
ysb5b@umsystem.edu

Li Li
University of Missouri-Kansas City
Kansas City, Missouri
lil1@umkc.edu

Zhu Li
University of Missouri-Kansas City
Kansas City, Missouri
lizhu@umkc.edu

Shan Liu
Tencent America
Palo Alto, California
shanl@tencent.com

ABSTRACT

Generally, adaptive bitrates for variable Internet bandwidths can be obtained through multi-pass coding. Referenceless prediction-based methods show practical benefits compared with multi-pass coding to avoid excessive computational resource consumption, especially in low-latency circumstances. However, most of them fail to predict precisely due to the complex inner structure of modern codecs. Therefore, to improve the fidelity of prediction, we propose a referenceless prediction-based R-QP modeling (PmR-QP) method to estimate bitrate by leveraging a deep learning algorithm with only one-pass coding. It refines the global rate-control paradigm in modern codecs on flexibility and applicability with few adjustments as possible. By exploring the potentials of bitstream and pixel features from the prerequisite of one-pass coding, it can reach the expectation of bitrate estimation in terms of precision. To be more specific, we first describe the R-QP relationship curve as a robust quadratic R-QP modeling function derived from the Cauchy-based distribution. Second, we simplify the modeling function by fastening one operational point of the relationship curve received from the coding process. Third, we learn the model parameters from bitstream and pixel features, named them hybrid referenceless features, comprising texture information, hierarchical coding structure, and selected modes in intra-prediction. Extensive experiments demonstrate the proposed method significantly decreases the proportion of samples' bitrate estimation error within 10% by 24.60% on average over the state-of-the-art.

CCS CONCEPTS

• **Information systems** → *Multimedia information systems*; *Multimedia streaming*;

KEYWORDS

Rate-distortion modeling; referenceless rate-distortion model; machine learning; transcoding; video processing

1 INTRODUCTION

The knowledge of bitrate and corresponding video quality is the necessary prerequisite to make the optimal bitrate allocation strategies for Internet-based video services, otherwise, it may cause a large of unnecessary waste or deficiency on clients' bandwidths. However, due to the inner complexity of modern codecs, e.g., high efficient video coding (HEVC) [27] and advanced video coding (AVC) [32], it has become a challenge to assess the bitrate and video quality in an accurate and fast way.

Many research focused on the characteristics of bitrate and quantization parameter (R-QP) of block-level rate-control paradigm [3] [31], which require a lot of inner algorithm adjustments for different codecs to execute bit allocation. It might cause fluctuation of temporal qualities due to insufficient bit assignment on few last coding units (CUs) of coding frames. To solve this issue, we adopt a global-based rate-control paradigm that considers each video clip (or frame) as a basic unit [4] [28] [29]. It can work with most modern codecs without the need for excessive block-level adjustments, that provide a global strategy of bit assignment to prevent inconsistent video quality. Moreover, parallel implementation can be achieved on video clips (or frames) in proportion to available computational resources.

The global rate-control paradigm certainly has many inherent advantages over the block-level paradigm. However, previous attempts have shown the fundamental problem of describing the characteristics of factors in bitrate allocation of the global paradigm. This bottleneck is mainly caused by the following reasons: 1) The global multi-pass coding method can establish the actual R-QP relationship curve, but excessive computational cost is needed. 2) Insufficient content or coding information was adopted to describe the R-QP relationship, e.g., Xu et al. [34] and Santamaria et al. [26] only used pixel information (original frames). Covell et

al. [4] and Sun et al. [28] simply took statistic coding domain data into account. 3) The linear modeling function is insufficient to fit nonlinear R-QP relationship. While considering situations of multiple resolutions or frame sizes, the fitting performance of the linear model [4] deteriorates as the resolution of frame size increases [28].

Therefore, to solve this issue, we propose a referenceless prediction-based R-QP modeling method (PmR-QP) throughout only one-pass coding needed as a prerequisite to extract the bitstream and pixel information. It exploits the potentials of hybrid referenceless features to track down the quantization processes, which are used to signal redundant information elimination. Then, leveraging a deep learning algorithm as the replacement of actual coding [11] [10], PmR-QP learns the corresponding oblique relationships between the extracted representatives and bitrate-quality information. To be specific, firstly, we develop an optimized R-QP modeling function to characterize the relationship between bitrate and QP. Secondly, we enhance and unify the features of bitstream in multiple coding domains to learn the content-dependent R-QP model parameters from scratch. Currently, we successfully validate the efficiency of PmR-QP in intra-predicted frames, which occupy the majority proportion of bits in videos. The contributions of this paper list as followed,

- We derive a quadratic R-QP modeling function from the Cauchy-based distribution to characterize the relationship between bitrate and corresponding QP, which is used to directly control video quality. The quadratic modeling function can fit the non-linear R-QP relationship better than the previous linear rate-control modeling function [4]. We have it tested to prove its feasibility in real cases prevailingly.

- We fasten an operational point on the R-QP relationship curve from the one-pass coding in passing that no additional computational cost is needed. As the means of model parameter elimination, it can greatly improve the proceeding of deep learning in inferring speed and estimated accuracy.

- We significantly explore the potentials of bitstream and pixel information from multi-levels coding domains, e.g., reconstruction, hierarchical segmentation, and macro-block intra-prediction. To concatenate them in the proposed network, we modify these features to a uniform type and structure. To the best of our knowledge, no previous works have used the homogeneous scheme because of inconsistency in different coding domains.

- Performance experiments and ablation studies on DIV2K dataset [1], demonstrate the PmR-QP method outperforms the state-of-the-art linear modeling solution.

The remainder of the paper is organized as follows. Section 2 will introduce related researches. In Section 3, we will discuss the proposed R-QP modeling function, and hybrid bitstream features in details. The proposed network and hyper-parameters will be discussed in Section 4. Section 5 will show the detailed experimental results and Section 6 will present the conclusions and future plans.

2 RELATED WORK

The knowledge of rate and distortion (R-D) relationship is essential for rate control. It decides how many bits should be provided to obtain minimal distortion subject to the budget of bits. Ou et al. [24] considered a similarity index as a quality metric for R-D model to correlate bitrate allocation with human perception. Gao et al. [6] proposed a Nash bargaining solution for optimizing a structural similarity index (SSIM)-based CTU-level R-D scheme. Both of these methods need the actual R-D relationship from multiple passes of coding. Because the R-D data is recorded from real samples, the accuracy can be assured. However, the excessive computational cost needs to be reduced while being applied to latency-sensitive scenarios. Therefore, differing from actual coding, the idea of R-D estimation was proposed for more practical video applications.

Estimation methods can be divided into two categories depending on whether adopting a modeling function to calibrate the estimated results or not. Non-modeling-based methods implement end-to-end frameworks to predict R-D relationship directly, while modeling-based methods derive R-D relationship from modeling functions. Many researchers have leveraged deep learning algorithms to estimate R-D relationship due to their availability in different video applications [16] [36] [12] [35]. For non-modeling-based methods, Xu et al. [34] and Santamaria et al. [26] proposed CNN-based R-D estimated methods. They both adopted original frames as references to estimate the R-D relationship explicitly. In [34], SSIM maps were attributed as distortion and learned through a novel CNN in separate with the number of bits. Santamaria et al. [26] followed a similar framework to estimate the number of bits (pixel-wise) and absolute distortion mappings instead of SSIM maps individually, through a neural network with two pipelines. The notion of nonlinearity for R-D estimation was proposed in [26] and activated function was improved by adding Parametric Rectified Linear Unit (PReLU) [8] to achieve nonlinear fitting. These solutions diminished the complexity of codex over multi-passes coding.

Then, modeling-based R-D estimation methods were proposed by Covell et al. [4] and Sun et al. [28]. Covell et al. [4] used statistic coding representatives to predict bitrate and constant rate factor (CRF) implicitly through a linear logarithmic R-CRF model. Sun et al. [28] optimized the R-CRF model to second-order function, which described the nonlinearity of R-CRF relationship better. Both of them only employed pure statistical coding information, which has been proved its insufficiency to describe video content. On the other hand, the disadvantage of current non-modeling-based methods was the mere adoption of pixel information (frames). Neither of them used these intersectional domains data to study the R-D relationship. The boundedness of single domain data might mislead the algorithms to make a global decision.

3 PROPOSED METHOD

3.1 Overall Framework

In this section, we elaborate upon the framework of PmR-QP method. As aforementioned, QP is adopted as the quality metric of intra-frames and used to directly control bitrate by the employed codec. Fig. 1 shows the details of the framework, whose objective is to predict the content-dependent R-QP model parameters $\hat{p}_n$ initially, then derive the corresponding bitrate through the proposed model with the given QP. As shown, the complete framework can be divided into three subtasks, R-QP modeling function $m(.)$, referenceless features extraction $E_b(.)$, and network training $t(.)$. With
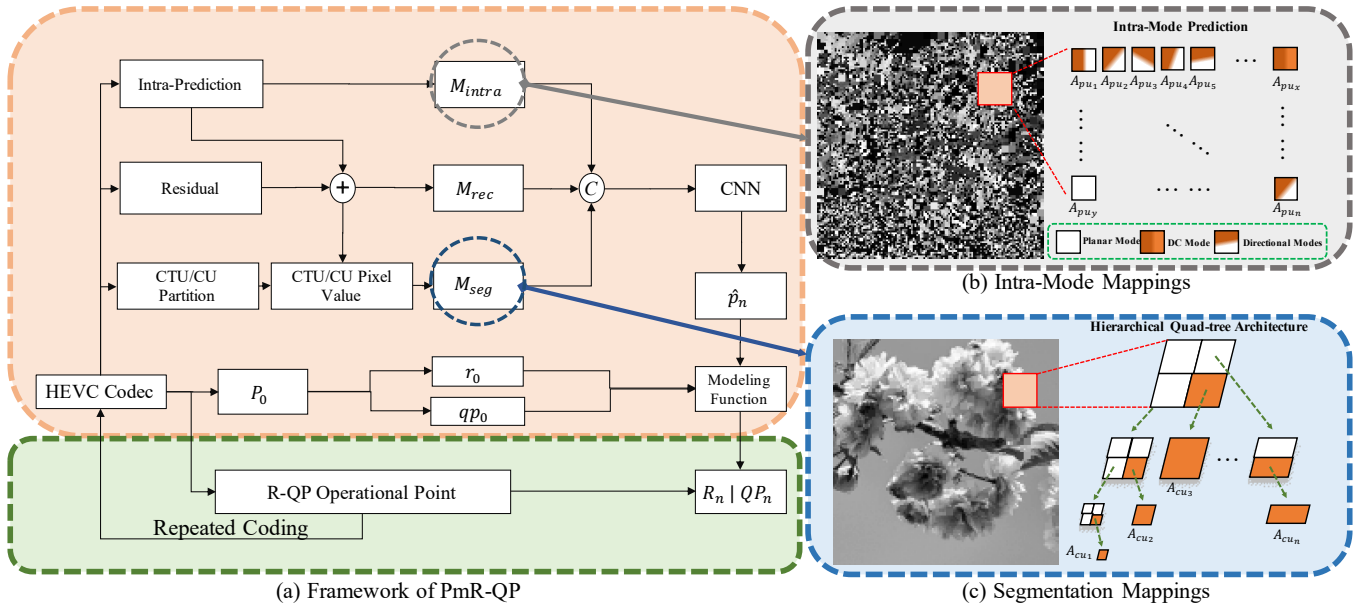
Figure 1: (a) Framework of PmR-QP compared with multi-pass coding method. (b) Extraction procedure of intra-mode mappings M_{intra}. (c) Extraction procedure of segmentation mappings M_{seg}.

the given QP and trained R-QP model parameters $\hat{p_n}$, we can derive the predicted bitrate $\hat{R}$ as followed,

$$\hat{R} = m(QP, \hat{p_n}). \tag{1}$$

$\hat{p_n}$ are learned from the concatenated hybrid referenceless features,

$$\hat{p_n} = t(cat(M_{rec}, M_{seg}, M_{intra}), \Theta), \tag{2}$$

where Θ denotes the set of network variables (weights and biases). M_{rec}, M_{seg}, and M_{intra} represent the hybrid referenceless features (reconstructed, segmentation, and intra-mode mappings). They are extracted from the target intra-frames in multiple coding domains,

$$[M_{rec}, M_{seg}, M_{intra}] = E_b(\gamma), \tag{3}$$

In summary, Eq. (1), Eq. (2), and Eq. (3) can jointly merge to the PmR-QP method, denoted as $F_p(.)$,

$$\hat{R} = F_p(\gamma). \tag{4}$$

3.2 Proposed R-QP Modeling Function

The existing study in [13] clarified the knowledge of Discrete cosine transform (DCT)'s coefficients' probability distribution is critical at the derivation of the relationship between bitrate and Quantization step (Qstep) (associated with QP). Here, to explore an accurate description of the bitrate and QP relationship, we formulate the R-QP modeling function derived from the entropy of DCT's coefficients.

Generally, bit allocation strategy splits bits into two groups: header bits R_h and residual bits R_r (dominant fraction of total bit consumption R_{total}). We can assume that

$$R_{total} \approx R_r. \tag{5}$$

As known, R_r is related to the entropy of DCT's coefficients. Due to the property of residual bits, the entropy of DCT's coefficients

is extremely sensitive to quantization. Therefore, the approximate correlation of total bits R_{total} and the entropy of DCT's coefficients $H(Q)$ at varying Q (Qstep) can be represented as,

$$H(Q) \approx R_{total}(Q) \approx R_r(Q). \tag{6}$$

In [13], the probability distribution function (PDF) of Cauchy distribution [5] is proven that a better description is on actual data than Gaussian [7] and Laplacian [22] distributions. Then, the entropy of quantized DCT's coefficients in informative theory can be extended based on Cauchy-PDF,

$$H(Q) = -\frac{1}{\pi} \sum_{n=-\infty}^{\infty} tan^{-1} \frac{\gamma Q}{\gamma^2 + (n^2 - 0.25)Q^2}$$
$$\times log_2 [\frac{1}{\pi} tan^{-1} \frac{\gamma Q}{\gamma^2 + (n^2 - 0.25)Q^2}], \tag{7}$$
$$n = \pm 1, \pm 2 ..., \pm N,$$

where, nQ denotes as quantization level, while γ is the variable of zero-mean Cauchy-PDF. A linear modeling function between H(Q) and Q [30] was proposed to simplify Eq. (7), but hardly to characterize the non-linear $H(Q) - Q$ relationship, especially when γ increases at a large margin. The hypothesis quadratic $H(Q) - Q$ relationship can be suggested given by the observation of Fig. 2,

$$ln(Q) \propto ln(H(Q))^2 + ln(H(Q)) + c, \tag{8}$$

where c represents a constant. The transformation between QP and Q can follow the equation below,

$$QP = 6 \cdot log_2 Q + 4, \tag{9}$$

Since we need to assess the relationship between bitrate and QP eventually, Eq. (6), Eq. (8) and Eq. (9) can be joint derived that
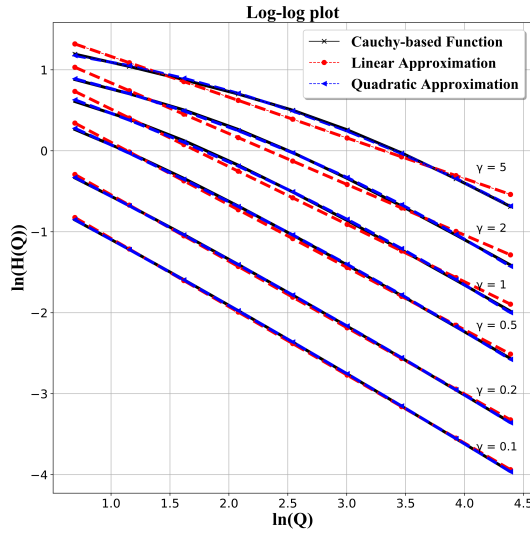
Figure 2: H(Q)-Q log-log plot of Cauchy-PDF-based function approximation.

the entropy of DCT's coefficients in varying QP has a quadratic logarithmic changing trend in approximation,

$$QP \propto \frac{6[ln(R(QP))^2 + ln(R(QP))] - 4}{ln(2)}. \tag{10}$$

Therefore, based on previous R-D modeling functions [9] [18] [19] [4], we devise a quadratic logarithmic modeling function from Eq. (10),

$$QP = \alpha(\gamma)ln(R(QP))^2 + \beta(\gamma)ln(R(QP)) + \mu(\gamma), \tag{11}$$

where $\alpha(\gamma)$, $\beta(\gamma)$, and $\mu(\gamma)$ denote as content-dependent model parameters related to intra-prediction frame γ.

The proposed quadratic R-QP modeling function can fit a wide range of QP settings to achieve many practical uses precisely. But the increasing number of model parameters results in prohibitive levels of training complexity compared with the linear approximation [4]. To overcome this issue, we further explore the potential of coding information to simplify the modeling function. An operational R-QP point P_0 is encoded along with bitstream data that have been ignored previously. We decide to involve it by fastening the proposed modeling function on P_0. P_0 would not deteriorate the fitting capacity of function since it is from actual coding. Meanwhile, the freedom of fastened function is limited fractionally with fewer model parameters to learn. Assume the values of bitrate and QP at P_0 are r_0 and qp_0, respectively, then R-QP modeling function can be eliminated one modeling parameter as followed,

$$
\begin{aligned}
QP = \alpha^*(\gamma)[ln(R(QP))^2 - ln(r_0)^2] + \\
\beta^*(\gamma)[ln(R(QP)) - ln(r_0)] + qp_0,
\end{aligned} \tag{12}
$$

where $\alpha^*(\gamma)$ and $\beta^*(\gamma)$ are the model parameters pending to predict from network after simplification. This proposed modeling function successfully balances the trade-off of training difficulty and predicted precision, evaluated in Sec. 5.3.2.

3.3 Hybrid Referenceless Features

The core execution of compression techniques in most modern codecs is to eliminate overlapped or redundant information in terms of spatial, temporal, statistical, and visual domains [17]. Bits can be saved due to the eliminating processes without abandoning any relevant data. Therefore, as long as we detect the exact number of bits saved subject to corresponding levels of quality in compression, the relevance of bitrate and QP can be received in different levels of quantization. However, due to the complex inner architectures of modern coding standards, such as HEVC, it is extremely difficult to approximate the saved bits [2]. Due to the successes of deep learning algorithms in many multimedia applications, it is natural to come up with a learning-based method as the fundamental core to track down the saved bits by exploring the features of information redundancies in different coding procedures. However, the type and structure of data in different coding procedures is inconsistent to analyze, which leads that most previous researches only concentrated on the studies of local procedures partially.

In this paper, we bring out the unification of the type and structure of data in different coding domains. It takes advantage of the global features extraction of bitstream and pixel information, as hybrid referenceless features, to significantly improve the estimated performance. In detail, the hybrid referenceless features include the components of texture information, hierarchical coding structure information, and intra-predicted modes. We unify them into pictures by mapping them to two-dimension planes, as shown in Fig. 3.

Fig. 3 demonstrates hybrid referenceless features from two intra-predicted frame samples with different levels of texture intricacy. We visualize features by coding in different QP settings (QP={10, 26, 38}) to study the quantized sensitiveness of each feature individually. Initial assumption indicates the referenceless features mirror the progression of quantization that leads the possibility to learn the property of R-QP through the exploration of referenceless features. The observation shows that underlying layers (segmentation and intra-prediction) of coding respond to the changes of QP more unmistakably. Even though their descriptions of the multifaceted nature of images and the distribution of high and low frequencies are not as good as the reconstructed image, they can increase the quantized sensitiveness of the referenceless features. The following paragraphs will explain the generalization of each feature step by step.

3.3.1 ***Reconstructed Mappings****.* Reconstructed Mappings M_{rec} possess the homologous structure with corresponding original images since the quantization would not damage the integrity of reconstructed Mappings M_{rec} but simply modify the number of bits for each symbol [21]. They are defined as the representatives in pixel-domain to preserve the texture information in low-distorted details. As the replacement of original images, reconstructed mappings M_{rec} are extracted to describe the content complexity and distribution of high- and low-frequency information. As shown in Section 5, they dominate the circumstances of a single feature as input. It proves the efficiency of M_{rec}, especially lacking original data, e.g., video transcoding [33].
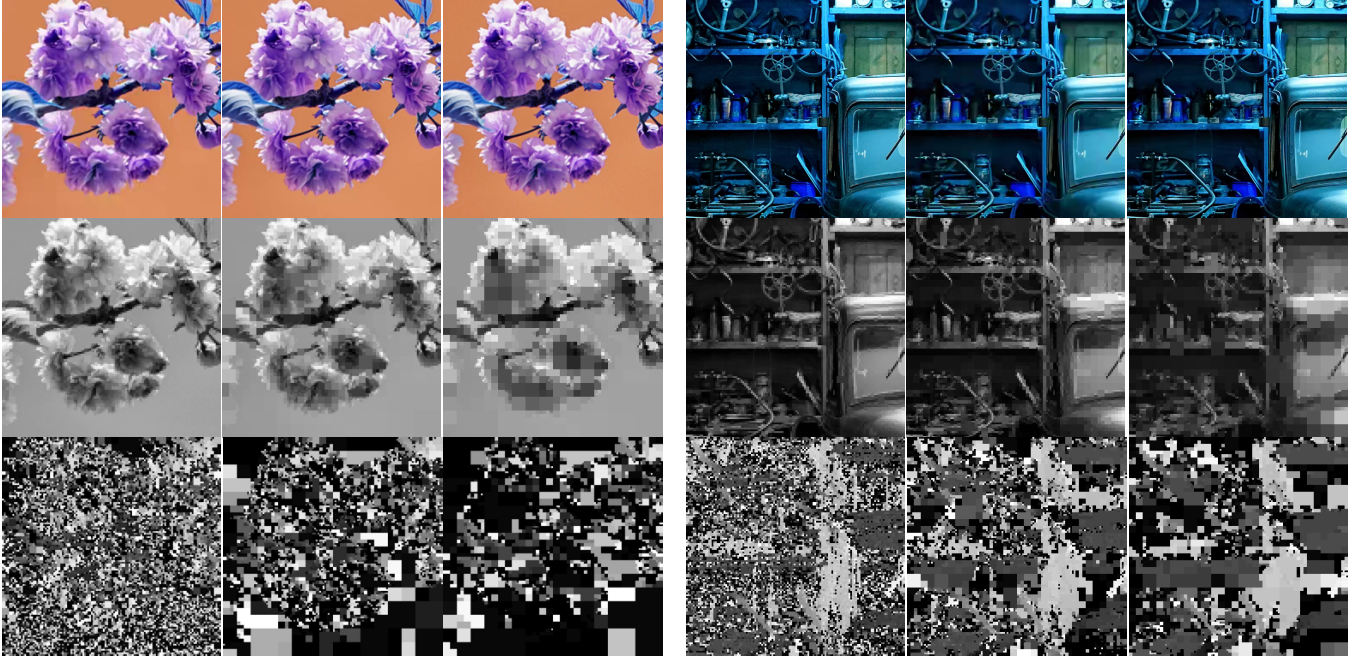
Figure 3: Hybrid bitstream features in QP settings $\in \{10, 26, 38\}$ in different levels of texture complexity, including M_{reg}, M_{seg}, and M_{intra}, respectively.

3.3.2 ***Segmentation Mapping.*** The quad-tree coding tree units (CTUs) architecture is employed with variable sizes of units [23] in HEVC codec, which can be partitioned into hierarchical coding units (CUs) and further divided into predicted units (PUs). The availability of larger block sizes in a quad-tree partitioning structure decides the most significant improvement of coding efficiency compared with previous codec generation [14], also preserving dominant bits saving. The knowledge of CU partition might help track the arrangement of bits at macroblock (MB) level [25]. For instance, a larger-size CU requires less bit per pixel (bpp) than a smaller-size CU, while a deeper depth of CU requires more bpp.

To explore the partitioning information, we extract and visualize CU information, denoted them as segmentation mappings M_{seg}, as shown in Fig. 1(c). To be specific, three types of partitioning factors are utilized, including the average pixel esteem, size, and location of each CU. We first create a sharing-sized blank image with the original frame and split it into number of different scaled areas according to the corresponding size and location of CUs. The average pixel esteem is assigned as the shared pixel values for the complete scaled areas, as shown

$$M_{seg} = r([A_{cu_1}]_{w_1 \times h_1}, [A_{cu_2}]_{w_2 \times h_2}, ..., [A_{cu_n}]_{w_n \times h_n}), \quad (13)$$

where,

$$A_{cu_n} = \sum_{i}^{w_n} \sum_{j}^{h_n} \frac{V_{p(i,j)}}{w_n h_n}. \quad (14)$$

Here, the sets of $\{w_1, w_2...w_n\}$ and $\{h_1, h_2...h_n\}$ represent the width and height of corresponding CUs, respectively. $r(.)$ is the reshaping operator and $V_{p(i,j)}$ denotes as the pixel value at location of $\{i, j\}$ in M_{rec}. The operator $[.]_{w \times h}$ is to assign the average pixel

esteem A_{cu_n} from Eq. (14) to the complete area CU_n with the size of $\{w_n \times h_n\}$. Until now, we finish projecting the partitioning information of intra-predicted frame to the blank image, The blocking effect can be observed in M_{seg}, which corresponds to the distribution of frequency information. It is worth noting that we merge some CUs visually due to the identical A_{cu_n} between them, which is not shown in the HEVC partitioning. Even though, it matches better with the strategy of actual bit allocation.

3.3.3 ***Intra-Mode Mappings.*** Intra-prediction is an operation in video coding to eliminate pixel similarity for intra-frames [27]. It executes the predicted mode selection for each PU based on the least distortion principle to reference samples. 33 angular modes for both luma and chroma channel and two non-directional modes (DC and planar) are involved in HEVC/H.265 [27], which exceeds the number of modes in AVC/H.264. Therefore, HEVC/H.265 explicitly provides better compression efficiency on erasing the spatial information redundancies than AVC/H.264.

To generate intra-mode mappings, we first number predicted modes $Pred_i$ from 0 to 34 sequentially. To project modes information into pictures, we then evenly distribute different values in the interval of [0, 238] as the regional pixel esteem according to their serial numbers, as Fig. 1(b) shown. The value interval follows pixel value distribution of common pictures. Differing from segmentation mappings M_{seg}, we fix the scale of PUs as the square of 16 with observing a slight effect on performance. It is noted that pixels within an identical unit share an uniform mode. At last, we cluster and reshape the comprehensive PUs to the uniform size of the rest of referenceless features,

$$M_{intra} = r([A_{pu_1}]_{16 \times 16}, [A_{pu_2}]_{16 \times 16}, ..., [A_{pu_n}]_{16 \times 16}), \quad (15)$$

where n is the quantity of PUs in the intra-mode mappings. The complete extraction of M_{intra} is visualized in Fig. 1(b).

4 ARCHITECTURE OF PROPOSED NETWORK

We propose a convolutional neural network (CNN) to learn the connection between R-QP relationship and hybrid referenceless features, which can be seen as a typical regression problem. To highlight our main contributions on modeling optimization and features enhancement, we give up the adoption of deeper and more effective neural network but only 5-layers architecture, which is shown in Fig. 4. We adopt *Adam* optimizer [15] to train at the learning rate of 0.0001. *ReLu* as activated function is configured at the output of first four layers. To normalize the input and output, we use the image normalization and *StandardScaler* to process the complete referenceless features and model variables separately. Batch size sets up to 10 and the number of epochs is 100. The loss function is designed as follow,

$$L(\Theta) = \frac{1}{n} \sum_{i=1}^{n} ||\hat{p_n} - p_n||^2. \tag{16}$$

where $\hat{p_n}$ and p_n are the predicted model parameters and ground truth (n is number of parameters). It trains at NVIDIA GeForce GTX 1080 GPU around 22 hours for each task.

5 EXPERIMENTAL RESULTS

5.1 Simulation Setup

DIV2K dataset [1] includes 900 high-resolution images in varying scales. To fulfill the experiments, we crop 800 images into the desirable patches (512x512 and 768x768) stochastically as training/validation sets and the rest of images as test set. The ratio of training and validating samples is 90% and 10%. As shown in Fig. 4, we train the network parameters Θ strictly in the single quantized level ($QP = 10$), which are applicable for circumstances in the rest of quantized levels ($QP \in [10, 38]$) of identical content frames. As the labels of network, model parameters p_n are calculated by the least square method [20]. To validate the generalization of PmR-QP, we test on two frame resolutions and execute the features extraction at HM 16.9 platform [27].

We take the related bitrate estimation error δ as the measure of accuracy in bitrate estimation as followed,

$$\delta = \frac{R(QP) - \hat{R}(QP)}{R(QP)} \times 100\%. \tag{17}$$

5.2 Performance of PmR-QP in Bitrate Estimation

We first evaluate the entire performance improvement of PmR-QP compared with the linear modeling solution. Table 1 shows the accuracy of bitrate estimation by the proposed and linear prediction-based modeling method. Here, PmR-QP employs the optimized quadratic R-QP modeling function to learn their parameters $\hat{p_n}$ from the overall hybrid referenceless features. For a fair comparison of modeling function and training features, the linear solution is trained by the proposed network as well. The results in Table 1 show that PmR-QP achieves 87.92%, 79.11% and 60.55% of samples' estimated error δ within 30%, 20%, and 10% in 512x512, respectively.

Table 1: Comparison of PmR-QP and the linear modeling solution in bitrate estimation

Frame Size	Model	Features	P_0	Estimated Error δ		
				30%	20%	10%
512x512	Linear	M_{rec}	-	80.70%	63.14%	35.81%
	PmR-QP	All	✓	87.92%	79.11%	60.55%
	Improvement			**+7.22%**	**+15.97%**	**+24.74%**
768x768	Linear	M_{rec}	-	83.56%	67.40%	38.61%
	PmR-QP	All	✓	90.43%	81.97%	63.07%
	Improvement			**+6.87%**	**+14.57%**	**+24.46%**

14.33%, 22.90%, and 29.83% precision improves in each error region, which shows that PmR-QP outperforms the linear solution. In 768x768, the prediction performance of PmR-QP is better than in 512x512 that 90.43%, 81.97%, and 63.07% in each error region. 5.65%, 13.88%, and 23.57% rises are bought by PmR-QP to the linear one. In general, PmR-QP significantly promotes the accuracy of bitrate estimation by 26% (within 10% error) on average in both resolutions over the linear solution.

5.3 Ablation Studies of PmR-QP Method

Table 2: Performance of R-QP modeling function

Frame Size	Model	Features	P_0	Estimated Error δ		
				30%	20%	10%
512x512	Linear	All	✓	90.01%	75.19%	43.45%
	Quadratic			87.92%	79.11%	60.55%
	Improvement			**-2.09%**	**+3.92%**	**+17.1%**
768x768	Linear	All	✓	91.64%	78.24%	45.38%
	Quadratic			90.43%	81.97%	63.07%
	Improvement			**-1.21%**	**+3.73%**	**+17.69%**

5.3.1 Improvements in Quadratic R-QP Modeling Function. To show the superiority of proposed R-QP modeling function, we maintain the identical configuration for the rest optimizations. The comparison of linear and quadratic modeling function is shown in Table 2, the proportions of samples' estimation error within 30%, 20%, and 10% are 87.92%, 79.11%, and 60.55% in 512x512 along with 90.43%, 81.97%, and 63.07% in 768x768 by utilizing quadratic modeling function. Compared with the linear function, the increasing precision is -2.09%, 3.92%, and 17.10% in 512x512 along with -1.21%, 3.73%, and 17.69% in 768x768. It proves the fidelity of quadratic modeling function at characterizing R-D relationships, especially in preciser estimation scenarios.
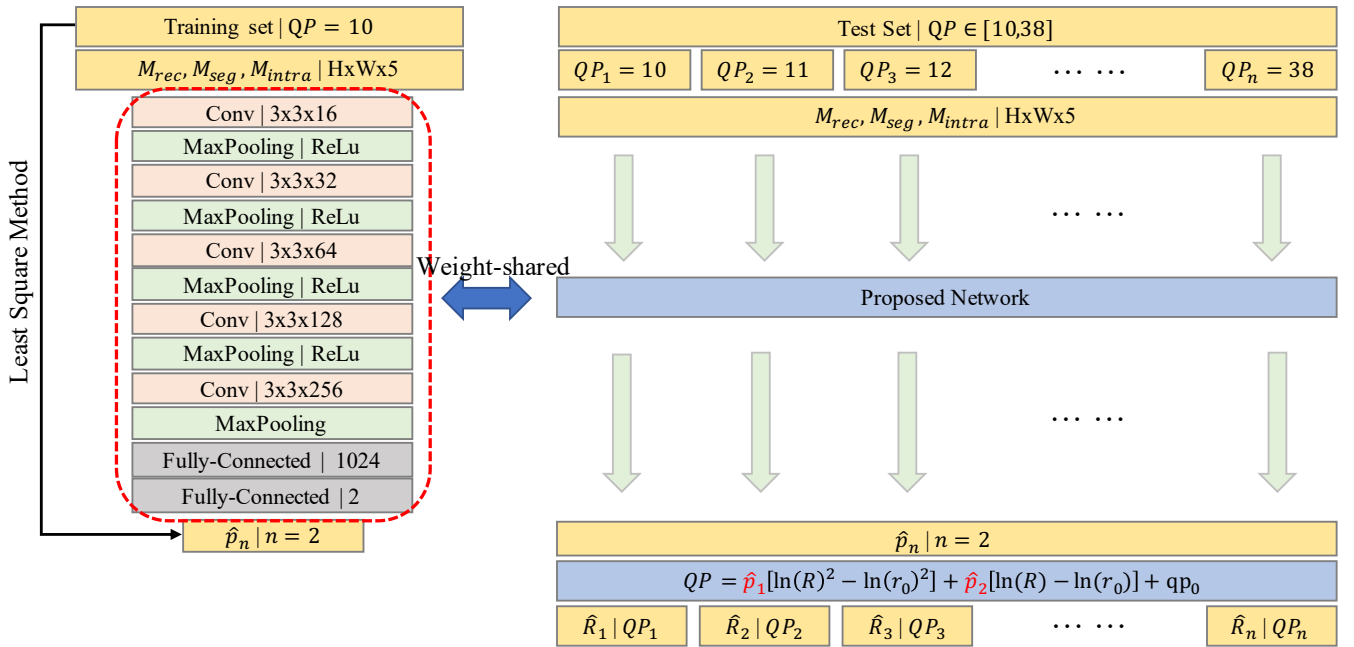
Figure 4: Architecture of proposed network and the complete procedures of training and testing.

Table 3: Comparison of simplifying by operational point P_0 or not

Frame Size	Model	Features	P_0	Estimated Error δ		
				30%	20%	10%
512x512	Quadratic	All	-	73.59%	56.21%	30.72%
			✓	87.92%	79.11%	60.55%
	Improvement			**+14.33%**	**+22.90%**	**+29.83%**
768x768	Quadratic	All	-	84.78%	68.09%	39.50%
			✓	90.43%	81.97%	63.07%
	Improvement			**+5.65%**	**+13.88%**	**+23.57%**

5.3.2 Improvements in Modeling Function Simplification.
As known, higher-order models can improve the estimation accuracy of the R-QP relationship, however, blindly improving the number of orders would complicate the learning-based estimated process, even though they can fit complex data well due to the increasing number of model parameters p_n. The notion of model simplification is to inherit an operational point from the one-pass coding, using it to reduce model parameters p_n. Table 3 shows a comparison of quadratic modeling function based estimated methods with or without model simplification. In 512x512, the improvement by adopting model simplification is 14.33%, 22.90%, and 29.83%, while the estimated error δ is below 30%, 20%, and 10%, respectively. In 768x768, the improvement is 5.65%, 13.88%, and 23.57% corresponding to each error region. Model simplification brings the most

significant promotion in bitrate estimation, which is 9.99%, 18.39%, and 26.7% on average in each error region. It indicates that most deterioration of estimation essentially originates from the capacity of neural network training.

5.3.3 Improvements in the Hybrid referenceless features.
This section discusses the performance of hybrid referenceless features. In prior researches, pixel domain features or bitstream features are adopted more often due to the inconsistency between the two different domains. We find a method to extract and unify multi-domains coding features and observe their R-QP estimation performance in Table 4, which demonstrates the superiority of entire features combination over other ways to combine features. It indicates the existence of non-overlapping information from different domains to achieve the performance-boosting of estimation. Compared with texture domain features M_{rec}, the entire features combination can bring out 2.98% and 6.54% improvements on average in each resolution. Meanwhile, it also shows a better improvement compared with other coding bitstream domains, e.g., M_{seg} or M_{intra}. Note that our method at feature extraction always performs better in higher resolution scenarios.

Besides, we dig out more facts from the comparison of different domain features. In single domain level, M_{rec} outperforms M_{seg} or M_{intra} in estimation, which reveals a fact that richer information is provided by texture of frames. Even though, M_{seg} and M_{intra} show a very close outcome due to partial over-lapping texture information in these coding domains. Moreover, M_{seg} replenishes the knowledge of partitioning structure and M_{intra} describes the similarity of neighboring blocks. They represent redundancy in different aspects. By combining them with M_{rec} one by one, we can observe obvious growths and the best performance while combining

Table 4: Comparison of different combinations of hybrid referenceless features

Frame Size	Model	Patterns	Bitrate Estimated Error δ		
			30%	20%	10%
512x512	Quadratic	M_{seg}	82.24%	71.07%	51.17%
		M_{intra}	80.88%	70.22%	51.01%
		$\{M_{seg}, M_{intra}\}$	82.80%	72.78%	52.18%
		M_{rec}	84.96%	76.41%	56.96%
		$\{M_{rec}, M_{seg}\}$	86.54%	77.32%	58.03%
		$\{M_{rec}, M_{intra}\}$	86.53%	77.53%	58.43%
		All	**87.61%**	**79.11%**	**60.55%**
768x768	Quadratic	M_{seg}	81.92%	71.88%	52.64%
		M_{intra}	79.89%	69/88%	51.81%
		$\{M_{seg}, M_{intra}\}$	84.13%	73.80%	53.43%
		M_{rec}	84.48%	75.18%	56.19%
		$\{M_{rec}, M_{seg}\}$	87.16%	78.53%	60.32%
		$\{M_{rec}, M_{intra}\}$	88.82%	79.37%	60.42%
		All	**90.43%**	**81.97%**	**63.07%**

all. It is noted that only M_{seg} as input can achieve 51.1% and 52.64% (within 10% error) in each resolution, which adopts the lowest network and coding complexity.

5.4 Observation on Different Behaviors in Quantization

We investigate R-QP curves by adopting different aforementioned optimized methods to further claim the outperformance of PmR-QP. We chose two typical data samples in different quantifying changing circumstances. It is observed that the best fitting actual R-QP curve is given by PmR-QP employed the entire hybrid referenceless features and simplified R-QP modeling function compared with other methods, especially significantly outperform the conventional method. The worst performance is provided by PmR-QP without simplifying R-QP function, which expresses the deterioration of results originated from the training process. However, due to fitting limits of linear modeling function, model simplification cannot bring any obvious gains. The performance of the linear modeling based solution is as good as the quadratic method in uniform quantifying changes but much worse in non-uniform circumstances.

6 CONCLUSION

In this paper, we propose a referenceless PmR-QP to predict bitrate in given frames' quality information precisely throughout only one-pass coding. This method is built on the root of the global rate-control paradigm, which takes advantage of inborn systemic superiority over block-level paradigm in Internet-based multimedia services. Moreover, it efficiently tackles the defects of prior global rate-control methods. In detail, first, we derive the quadratic R-QP modeling function from Cauchy-based distribution on the entropy of DCT's coefficients, which is better at fitting the relationship between rate and level of quantization than linear function and applicable in most real cases. Second, efficiently utilizing the coding information, we involve an operational point to simplify the proposed modeling function. Third, PmR-QP significantly enhances the description of characteristics between R-QP and the frames' content information by exploring and unifying bitstream features from multiple coding domains. Extensive experiments and ablation studies demonstrate the global improvements in PmR-QP, and the performance-boosting from each optimized step. Generally, PmR-QP can achieve 24.60% decreases on samples' bitrate estimating error lower than 10% on average compared with the state-of-the-art.

In the future, we intend to expand our work to the inter-prediction level by exploring features of motion estimation. Likely, the similarity of successive inter-frames can be represented by the accumulated motion estimation features since they have been applied in other compressed video tasks. It is possible to aggregate these features as elements of hybrid referenceless features to learn R-QP information of intra- and inter-frames.

REFERENCES

[1] Eirikur Agustsson and Radu Timofte. 2017. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 126–135.

[2] Frank Bossen, Benjamin Bross, Karsten Suhring, and David Flynn. 2012. HEVC complexity and implementation analysis. *IEEE Transactions on Circuits and Systems for Video Technology* 22, 12 (2012), 1685–1696.

[3] Tihao Chiang and Ya-Qin Zhang. 1997. A new rate control scheme using quadratic rate distortion model. *IEEE Transactions on circuits and systems for video technology* 7, 1 (1997), 246–250.

[4] Michele Covell, Martin Arjovsky, Yao-chung Lin, and Anil Kokaram. 2016. Optimizing transcoder quality targets using a neural network with an embedded bitrate model. *Electronic Imaging* 2016, 2 (2016), 1–7.

[5] Jimmie D Eggerton and Mandyam D Srinath. 1986. Statistical distributions of image DCT coefficients. *Computers & electrical engineering* 12, 3-4 (1986), 137–145.

[6] Wei Gao, Sam Kwong, Yu Zhou, and Hui Yuan. 2016. SSIM-based game theory approach for rate-distortion optimized intra frame CTU-level bit allocation. *IEEE Transactions on Multimedia* 18, 6 (2016), 988–999.

[7] Nathaniel R Goodman. 1963. Statistical analysis based on a certain multivariate complex Gaussian distribution (an introduction). *The Annals of mathematical statistics* 34, 1 (1963), 152–177.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*. 1026–1034.

[9] Zhihai He, Yong Kwan Kim, and Sanjit K Mitra. 2001. Low-delay rate control for DCT video coding via/spl rho/-domain source modeling. *IEEE transactions on Circuits and Systems for Video Technology* 11, 8 (2001), 928–940.

[10] Tianchi Huang, Rui-Xiao Zhang, Chenglei Wu, Xin Yao, Chao Zhou, Bing Yu, and Lifeng Sun. 2019. Generalizing Rate Control Strategies for Realtime Video Streaming via Learning from Deep Learning. In *Proceedings of the ACM Multimedia Asia on ZZZ*. 1–6.

[11] Tianchi Huang, Rui-Xiao Zhang, Chao Zhou, and Lifeng Sun. 2018. Qarc: Video quality aware rate control for real-time video streaming based on deep reinforcement learning. In *Proceedings of the 26th ACM international conference on Multimedia*. 1208–1216.

[12] Chuanmin Jia, Shiqi Wang, Xinfeng Zhang, Shanshe Wang, and Siwei Ma. 2017. Spatial-temporal residue network based in-loop filter for video coding. In *2017 IEEE Visual Communications and Image Processing (VCIP)*. IEEE, 1–4.

[13] Nejat Kamaci, Yucel Altunbasak, and Russell M Mersereau. 2005. Frame bit allocation for the H. 264/AVC video coder via Cauchy-density-based rate and distortion models. *IEEE Transactions on Circuits and Systems for Video Technology* 15, 8 (2005), 994–1006.

[14] Il-Koo Kim, Junghye Min, Tammy Lee, Woo-Jin Han, and JeongHoon Park. 2012. Block partitioning structure in the HEVC standard. *IEEE transactions on circuits and systems for video technology* 22, 12 (2012), 1697–1706.

[15] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[16] Thorsten Laude and Jörn Ostermann. 2016. Deep learning-based intra prediction mode decision for HEVC. In *2016 Picture Coding Symposium (PCS)*. IEEE, 1–5.

[17] Zhibin Lei, Wu Chou, Jialin Zhong, and Chin-Hui Lee. 2000. Video segmentation using spatial and temporal statistical analysis method. In *2000 IEEE International Conference on Multimedia and Expo. ICME2000. Proceedings. Latest Advances in the Fast Changing World of Multimedia (Cat. No. 00TH8532)*, Vol. 3. IEEE, 1527–1530.

[18] Bin Li, Houqiang Li, Li Li, and Jinlei Zhang. 2014. λ domain rate control algorithm for high efficiency video coding. *IEEE Trans. Image Processing* 23, 9 (2014), 3841–3854.

[19] Bin Li, Jizheng Xu, Dong Zhang, and Houqiang Li. 2013. QP refinement according to Lagrange multiplier for High Efficiency Video Coding. In *2013 IEEE International Symposium on Circuits and Systems (ISCAS2013)*. IEEE, 477–480.

[20] Etienne PD Mansard and ER Funke. 1980. The measurement of incident and reflected spectra using a least squares method. In *Coastal Engineering 1980*. 154–172.

[21] John D McCarthy, M Angela Sasse, and Dimitrios Miras. 2004. Sharp or smooth?: comparing the effects of quantization vs. frame rate for streamed video. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 535–542.

[22] F Müller. 1993. Distribution shape of two-dimensional DCT coefficients of natural images. *Electronics Letters* 29, 22 (1993), 1935–1936.

[23] Jens-Rainer Ohm, Gary J Sullivan, Heiko Schwarz, Thiow Keng Tan, and Thomas Wiegand. 2012. Comparison of the coding efficiency of video coding standards—including high efficiency video coding (HEVC). *IEEE Transactions on circuits and systems for video technology* 22, 12 (2012), 1669–1684.

[24] Tao-Sheng Ou, Yi-Hsin Huang, and Homer H Chen. 2011. SSIM-based perceptual rate control for video coding. *IEEE Transactions on Circuits and Systems for Video Technology* 21, 5 (2011), 682–691.

[25] N Purnachand, Luis Nero Alves, and Antonio Navarro. 2012. Fast motion estimation algorithm for HEVC. In *2012 IEEE Second International Conference on Consumer Electronics-Berlin (ICCE-Berlin)*. IEEE, 34–37.

[26] M SANTAMARIA GOMEZ, Ebroul Izquierdo, Saverio Blasi, Marta Mrak, et al. 2018. Estimation of Rate Control Parameters for Video Coding Using CNN. (2018).

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

[28] Yangfan Sun, Mouqing Jin, Li Li, and Zhu Li. 2018. A Machine Learning Approach to Accurate Sequence-Level Rate Control Scheme for Video Coding. In *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 1013–1017.

[29] Yangfan Sun, Li Li, and Zhu Li. 2020. YOCO: Light-Weight Rate Control Model Learning. In *2020 27th IEEE International Conference on Image Processing (ICIP)*. IEEE, 1013–1017.

[30] NG Ushakov. 2001. Density of a probability distribution.

[31] Shanshe Wang, Siwei Ma, Shiqi Wang, Debin Zhao, and Wen Gao. 2013. Quadratic ρ-domain based rate control algorithm for HEVC. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 1695–1699.

[32] Thomas Wiegand, Gary J Sullivan, Gisle Bjontegaard, and Ajay Luthra. 2003. Overview of the H. 264/AVC video coding standard. *IEEE Transactions on circuits and systems for video technology* 13, 7 (2003), 560–576.

[33] Jun Xin, Chia-Wen Lin, and Ming-Ting Sun. 2005. Digital video transcoding. *Proc. IEEE* 93, 1 (2005), 84–97.

[34] Bin Xu, Xiang Pan, Yan Zhou, Yiming Li, Daiqin Yang, and Zhenzhong Chen. 2017. CNN-based rate-distortion modeling for H. 265/HEVC. In *2017 IEEE Visual Communications and Image Processing (VCIP)*. IEEE, 1–4.

[35] Hua Yang and Kenneth Rose. 2007. Advances in recursive per-pixel end-to-end distortion estimation for robust video coding in H. 264/AVC. *IEEE Transactions on Circuits and Systems for Video Technology* 17, 7 (2007), 845–856.

[36] Ming Yang, Ying Xie, Jian Yu, Zhe Wang, and Tao Wu. 2018. Using Deep Learning Neural Network for Block Partitioning in H. 265/HEVC. In *11th EAI International Conference on Mobile Multimedia Communications*. European Alliance for Innovation (EAI), 204.