Joint Optimization for SSIM-Based CTU-Level Bit Allocation and Rate Distortion Optimization

Yang Li and Xuanqin Mou, *Senior Member, IEEE*

***Abstract*—Structural similarity (SSIM)-based distortion D_{SSIM} is more consistent with human perception than the traditional mean squared error D_{MSE}. To achieve better video encoding quality, many studies on optimal bit allocation (OBA) used D_{SSIM} as the distortion metric. However, the MSE-based rate distortion optimization (RDO) was still used in these studies. The inconsistency between the optimization goals of OBA and RDO results in a non-optimal SSIM-based encoding performance. To solve this problem, we propose an accurate coding tree unit level D_{SSIM}-D_{MSE} model, which enables performing the SSIM-based RDO with simpler R-D_{MSE} cost scaled by the SSIM-based Lagrangian parameter λ_{SSIM}. Moreover, based on this model, the R-D_{SSIM} model can be accurately estimated based on the joint relationship of R-D_{SSIM}-λ_{SSIM}. With the accurate R-D_{SSIM} model, the SSIM-based OBA problem is then solved. Accordingly, the SSIM-based OBA and SSIM-based RDO are unified together in our scheme, called SOSR. Compared with the HEVC reference encoder HM16.20, SOSR saves 5%, 11%, and 17% bitrate under the same SSIM in the commonly used all-intra, hierarchical and non-hierarchical low-delay-B configurations, which is superior to existing state-of-the-art SSIM-based OBA schemes.**

***Index Terms*—Optimal bit allocation, rate distortion optimization, SSIM.**

I. Introduction

High Efficiency Video Coding (HEVC) standard has achieved significant compression performance improvement compared with the previous H.264/AVC [1]. However, due to the widely available applications such as video on demanding, video streaming, and video chatting, the burden of video transmission and storage is still growing. Faced with this situation, how to control the encoding to achieve the minimum possible distortion with the limited bits becomes a fundamental challenge.

HEVC encoding is controlled by many encoding parameters (e.g., quantization parameter (QP) and Lagrange multiplier λ), as well as a large number of encoding modes (e.g., block partition mode and prediction mode) [2]. In practice, encoder usually select the best combination of parameters and modes in the steps as follows. First, the encoder determines how many bits (R) are allocated to each encoding unit to achieve the minimal distortion (D) according to the R-D property of each unit, known as optimal bit allocation (OBA). The appropriate encoding parameters such as λ are then determined aiming at achieving the allocated bits for each unit. By applying the determined parameters, the encoder traverses all possible modes to encode an unit, and the mode with the minimum R-D cost (D+λR) is selected as the best mode, known as rate distortion optimization (RDO). These steps can be performed at different levels from the group of pictures, frame, to coding tree unit (CTU), where the CTU-level optimization greatly affect the encoding performance and is therefore investigated in this study.

In many studies such as [3]–[8] including the HEVC reference encoder HM [9], the mean squared error (MSE) is used as the distortion metric, denoted by D_{MSE}. In this way, the optimization goals of OBA and RDO are consistent, both aiming to minimizing the average MSE of a frame. Thereby, OBA and RDO can be solved uniformly by the Lagrangian optimization method [10], [11]. Usually, such an unified optimization is referred to as rate control collectively in many MSE-based studies [5], [8], [12]. However, MSE measures the pixel-wise difference between the encoded and the original videos. It has been validated to be poorly correlated with the human perception [13]. Minimizing MSE will not achieve an optimal perceptual quality. To solve this problem, many perceptual quality metrics such as the well-known Structural SIMilarity index (SSIM) [13] have been adopted into distortion measure (denoted by D_{SSIM}) in recent studies [14]–[25]. SSIM evaluates the similarity of luminance, contrast, and structures between two images, to which human perception is highly sensitive, thereby achieving better consistency with human perception than MSE.

Specifically, Ou *et al.* [14] established the R-D_{SSIM} model for the macroblocks of H.264/AVC, based on which an SSIM-based OBA scheme was proposed. In [15], Wang *et al.* proposed a macroblock-level maximum distortion descend method to solve the SSIM-based OBA problem of H.264/AVC. In [16], Gao *et al.* proposed a Nash bargaining game-based OBA scheme by considering SSIM as the utility of each CTU in HEVC. In [17], Zhou *et al.* proposed an R-D_{SSIM} model for CTUs based on the discrete cosine transform (DCT)-domain SSIM index [20], based on which the SSIM-based OBA was solved. By using SSIM in OBA, better perceptual quality has been achieved in these studies. However, it is worth noting that the MSE-based RDO was still used in above studies, which is inconsistent with the objective of SSIM-based OBA. Accordingly, the encoding mode with the minimum R-D_{MSE} cost rather than that with the minimum R-D_{SSIM} cost will be

Manuscript received August 24, 2020; revised February 22, 2021; accepted March 04, 2021. This work was supported in part by the National Key Research and Development Program of China under Grant 2016YFA0202003, and in part by the National Natural Science Foundation of China under Grant 62071375. *(Corresponding author: Xuanqin Mou.)*

Yang Li and Xuanqin Mou are with the Institute of Image Processing and Pattern Recognition, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: liyang2012@stu.xjtu.edu.cn; xqmou@mail.xjtu.edu.cn).

Code of this study is available at http://gr.xjtu.edu.cn/web/xqmou/sosr.

This paper has been accepted by IEEE Transactions on Broadcasting. Digital Object Identifier 10.1109/TBC.2021.3068871.

selected. The resulting encoding is not optimal in terms of the R-D_{SSIM} performance.

Therefore, to achieve better R-D_{SSIM} performance, the optimization goals of OBA and RDO should have to be based on SSIM consistently. However, there is a problem that prevents this strategy, that is, calculating R-D_{SSIM} cost for RDO is too time-consuming. Since HEVC has a large number of possible modes to encode a CTU [1], using D_{SSIM} in RDO will bring a huge increase in mode decision time, as can be found in [26]–[28]. An effective solution to this problem is to establish a D_{SSIM}-D_{MSE} model, based on which the complex R-D_{SSIM} cost can be mapped to a simpler R-D_{MSE} cost. In general, there are two widely used D_{SSIM}-D_{MSE} models. In [21], Yeo *et al.* approximated D_{SSIM} as a local variance-normalized D_{MSE}. Besides, in [20], Wang *et al.* proposed to approximate D_{SSIM} as the transform coefficients-normalized D_{MSE} based on the DCT-domain SSIM index [29]. In addition, there are also some other related studies, such as [30] that proposed a hyperbolic model of the D_{SSIM} and quantization step. All these models help to improve the R-D_{SSIM} performance of RDO without increasing computational burden. In particular, the Yeo's model and the DCT-domain model have led to a series of SSIM-based RDO schemes for H.264/AVC and HEVC, such as [22]–[24], [31] that are based on Yeo's model and [18]–[20], [25] that are based on the DCT-domain model. However, there are two limitations in these studies. First, our experimental results will show that accuracy of the widely used Yeo's model and the DCT-domain model is less than satisfactory. In addition, these studies typically only focused on the SSIM-based RDO without studying the R-D_{SSIM} relationship, so that the SSIM-based OBA has not been solved.

In this study, these problems are comprehensively solved through the following steps. Specifically, an accurate CTU-level D_{SSIM}-D_{MSE} model is proposed by a theoretical derivation with reasonable simplifications, based on which the SSIM-based RDO can be performed by a simpler R-D_{MSE} cost scaled by an SSIM-related Lagrangian multiplier. After performing the SSIM-based RDO, the resulted R, D_{SSIM}, and the applied Lagrangian multiplier of each CTU are used to accurately estimate the R-D_{SSIM} model. At last, the accurate R-D_{SSIM} model is used to solve the SSIM-based OBA problem of the next frame. In this way, the SSIM-based OBA and SSIM-based RDO are unified, called SOSR.

The contribution of the proposed model lies in three aspects. 1) We propose an HEVC scheme that unifies SSIM-based OBA and SSIM-based RDO together, resulting in better R-D_{SSIM} performance. 2) A CTU-level D_{SSIM}-D_{MSE} model is proposed, which is more accurate than the two widely used models. This model will benefit related SSIM-based studies in the future. 3) Our scheme enables the R-D_{SSIM} model to be accurately calculated by the SSIM-based R-D-λ joint relationship, which directly benefits solving the SSIM-based OBA problem.

The rest of the paper is organized as follows. Section II introduces the background. Section III describes the proposed SOSR scheme. Experimental results and discussions are presented in Section IV. Finally, Section V concludes this paper.

II. Backround

A. SSIM

SSIM measures the luminance similarity, contrast similarity, and structural similarity between the pristine image $\mathbf{x}$ and distorted image $\mathbf{y}$. Specifically, the similarity is calculated pixelwise, which is defined in [13] as follows:

$$\text{SSIM}_i = \frac{2\mu_{x_i}\mu_{y_i} + C_1}{\mu_{x_i}^2 + \mu_{y_i}^2 + C_1} \cdot \frac{2\sigma_{xy_i} + C_2}{\sigma_{x_i}^2 + \sigma_{y_i}^2 + C_2}, \tag{1}$$

where i indicates the i-th pixel in a frame, C_1 and C_2 are constants to prevent dividing by zero, and μ, σ^2, σ_{xy} are mean, variance, and covariance, respectively, which are calculated by

$$\begin{aligned} \sigma_{x_i}^2 &= \sum\nolimits_{l=1}^{L} \omega_l (x_{i+l} - \mu_{x_i})^2, \\ \sigma_{y_i}^2 &= \sum\nolimits_{l=1}^{L} \omega_l (y_{i+l} - \mu_{y_i})^2, \\ \sigma_{xy_i} &= \sum\nolimits_{l=1}^{L} \omega_l (x_{i+l} - \mu_{x_i})(y_{i+l} - \mu_{y_i}), \\ \mu_{x_i} &= \sum\nolimits_{l=1}^{L} \omega_l x_{i+l}, \quad \mu_{y_i} = \sum\nolimits_{l=1}^{L} \omega_l y_{i+l}, \end{aligned} \tag{2}$$

where $L = 121$ and $\omega_l (l = 1, 2, \cdots L)$ represents an 11×11 Gaussian filter [13].

SSIM is a quality index ranging from 0 to 1, with larger values indicating better quality. Thus, the SSIM-based distortion of a unit (a frame or a CTU) can be calculated as:

$$D_{\text{SSIM}} = 1 - \frac{1}{\text{M}} \sum_{i \in \text{unit}} \text{SSIM}_i, \tag{3}$$

where M is the number of pixels in the unit and '$i \in$ unit' indicates all the pixels located in the unit.

B. OBA and RDO

For encoding optimization, distortion is usually assumed to be a differentiable function of the encoding bits, which can be expressed as $D_k = D_k(R_k)$ for the k-th CTU in a frame. Accordingly, the OBA problem can be formulated as follows:

$$\arg\min_{R_k} \sum_{k=1}^{N} D_k(R_k), \text{ s.t.}, \sum_{k=1}^{N} R_k \le R_c, \tag{4}$$

where R_k is the to-be-allocated bits, N is the number of CTU in a frame, and R_c is the constrained bits. This problem can be solved by minimizing the unconstrained problem as follows by the Lagrangian optimization method [10], [11]:

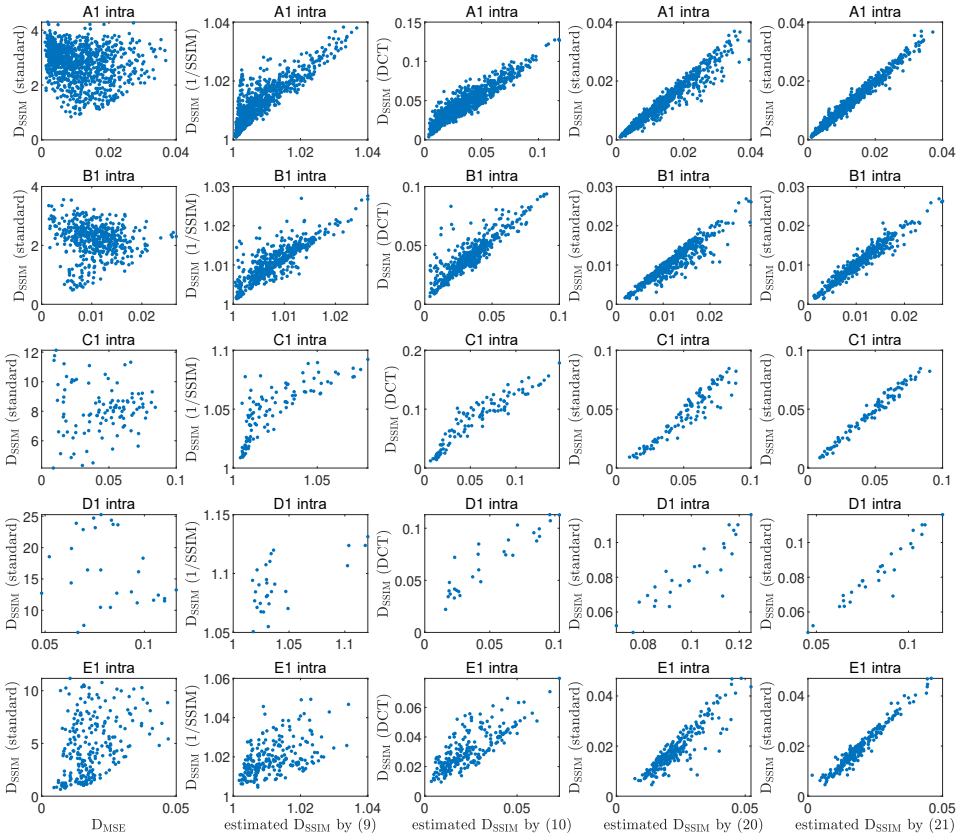
$$\arg\min_{R_k} \sum_{k=1}^{N} (D_k(R_k) + \lambda R_k). \tag{5}$$

To achieve the allocated R_k, the lagrangian multiplier λ_k is determined for the k-th CTU based on the association between R-D-λ:

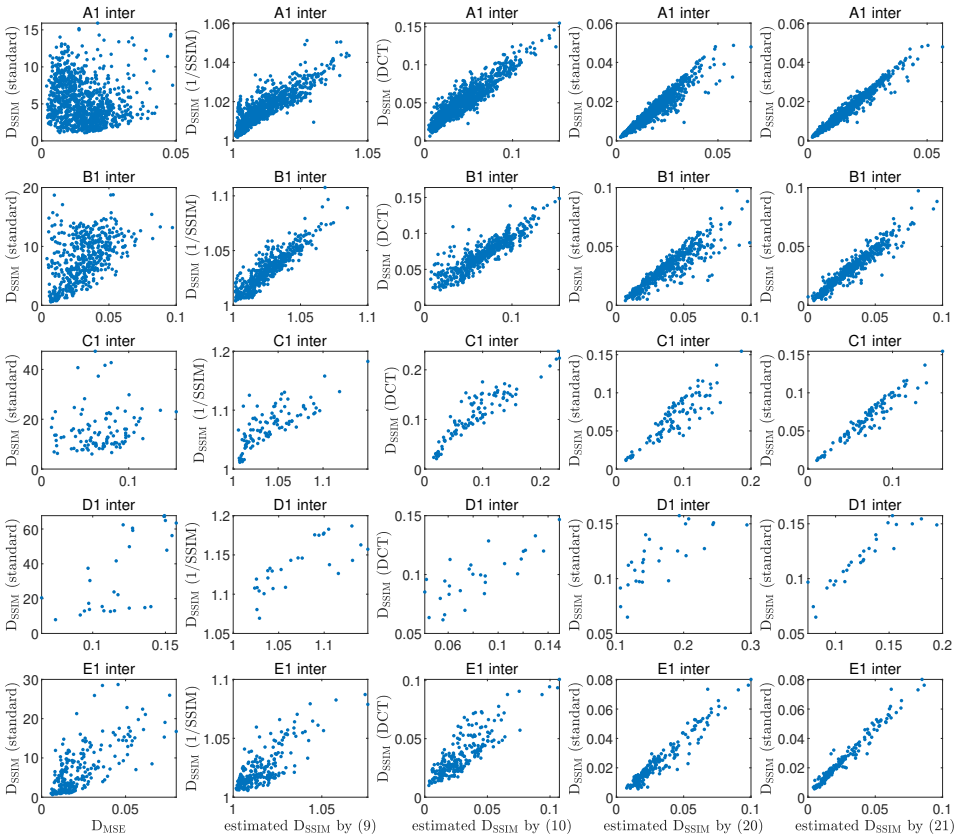
$$\lambda_k = -\frac{\partial D_k(R_k)}{\partial R_k}. \tag{6}$$

Using λ_k as the encoding parameter, RDO searches the best mode m that has the minimum R-D cost to encode the CTU, i.e.,

$$\arg\min_{m} D_k(m) + \lambda_k \cdot R_k(m). \tag{7}$$



(a)



(b)

Fig. 1. Comparison of the CTU-level D_{SSIM}-D_{MSE} models, including the Yeo's model (9), the DCT-domain model (10), the proposed model (20), and (21) improved by the regression method. The data are collected from the 100-th frame of five videos encoded in AI and LDB configurations at QP 32. Specifically, A1: Traffic1600p, B1:Kimono1080p, C1:BasketballDrill480p, D1:BasketballPass240p, E1:FourPeople720p.

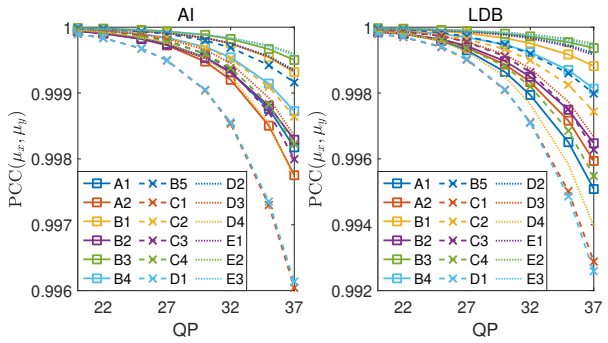


Fig. 2. The average PCC between μ_x and μ_y of all pixels in a frame. The plots show the average values of the first 100 frames for eighteen 8-bit test videos in AI and LDB configurations. The video numbers are in accordance with that in the common test configuration [32].

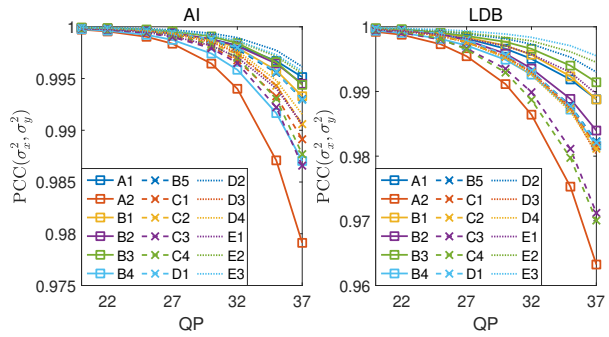


Fig. 3. The average PCC between σ_x^2 and σ_y^2 of all pixels in a frame. The experimental setting is consistent with that in Fig. 2.

In many existing SSIM-based OBA studies [14]–[17], the distortion metrics of OBA and RDO are D_{SSIM} and D_{MSE}, respectively. Accordingly, the optimization goal (5) of OBA and the optimization goal (7) of RDO are inconsistent. As a result, the corresponding R-D_{SSIM} performance is not optimal.

To solve this problem, some studies proposed to map the R-D_{SSIM} cost to the R-D_{MSE} cost that has a lower complexity [21]–[24]. Specifically, based on a D_{SSIM}-D_{MSE} model denoted by $f(\cdot)$, the R-D_{SSIM} cost is equivalent to a modified R-D_{MSE} cost as follows:

$$\begin{aligned} &D_{SSIM} + \lambda_{SSIM} \cdot R \\ =&f(D_{MSE}) + \lambda_{SSIM} \cdot R, \end{aligned} \tag{8}$$

where the Lagrangian multiplier $\lambda_{SSIM} = -\dfrac{\partial D_{SSIM}}{\partial R}$.

C. D_{SSIM}-D_{MSE} model

However, modeling the D_{SSIM}-D_{MSE} relationship is not easy. In the first column of Fig. 1, we illustrate the actual values of D_{SSIM}-D_{MSE} of ten example frames that were encoded by HM16.20 in all-intra (AI) and low-delay-B (LDB) configurations. It can be seen that there is no evident one-to-one mapping between D_{SSIM} and D_{MSE}. This is because that D_{SSIM} captures the structural degradation of the local regions, whereas D_{MSE} calculates the pixel-wise error. Therefore, the D_{SSIM}-D_{MSE} relationship depends on the image content, which varies over different CTUs in a frame.

To solve this problem, Yeo *et al.* developed a D_{SSIM}-D_{MSE} model in [21], where D_{SSIM} is expressed as the variance-normalized D_{MSE} as follows:

$$D_{SSIM} = 1 + \frac{D_{MSE}}{(2\sigma_x^2 + C_2)}, \tag{9}$$

where 1/SSIM of a block is used as D_{SSIM} in [21], σ_x^2 is the variance of the original image block. This model has low computation complexity. However, its modeling accuracy for HEVC is less than satisfactory, which can be seen in the second column of Fig. 1.

In recent study [17], Zhou *et al.* adopted a DCT-domain D_{SSIM}-D_{MSE} model for HEVC, which can be expressed as

$$D_{SSIM} = \frac{D_{MSE}}{S^2}, \tag{10}$$

where

$$S = \frac{\frac{1}{T}\sum_{t=1}^{T}\sqrt{\frac{\sum_{i=1}^{N_t-1}(X_t(i)^2+Y_t(i)^2)}{N_t-1} + C_2}}{E\left(\sqrt{\frac{\sum_{i=1}^{N_t-1}(X_t(i)^2+Y_t(i)^2)}{N_t-1} + C_2}\right)}. \tag{11}$$

In this model, $X_t(i)$ and $Y_t(i)$ are the i-th DCT coefficients in the t-th subblock of the original and reconstructed CTUs, respectively, N_t is the number of DCT coefficients of the subblock, and $E(\cdot)$ denotes the expectation operation in the whole frame. This model was proposed in [20], which is derived based on the DCT-domain SSIM index [29]. However, as shown in the third column of Fig. 1, this model also fails to accurately model the D_{SSIM}-D_{MSE} relationship in some videos.

To solve this problem, an accurate D_{SSIM}-D_{MSE} model will be proposed in the next section, and its modeling performance has been shown in Fig. 1 for comparison.

III. The Proposed Joint Optimization Scheme

In this section, we proposed a joint optimization scheme of SSIM-based OBA and SSIM-based RDO. Specifically, an accurate CTU-level D_{SSIM}-D_{MSE} model is proposed first. Then, based on the proposed model, SSIM-based OBA and SSIM-based RDO are unified in our scheme.

A. Relationship between D_{SSIM} and D_{MSE}

To explore the CTU-level relationship between D_{SSIM} and D_{MSE}, we empirically use the same assumption as [21] and [17], that is $\mu_{x_i} = \mu_{y_i}$. This assumption is easy to comprehend because that the coding process will reserve luminance information as much as possible. We use some sample experiments to validate the assumption in terms of Pearson correlation coefficient (PCC) [33] that is in accordance with the SSIM calculation. As shown in Fig. 2, the PCCs between μ_{x_i} and μ_{y_i} exceed 0.99 for different videos even at QP=37. With this assumption, the SSIM value at the i-th pixel as in (1) can be rewritten as follows:

$$\text{SSIM}_i \approx 1 - \frac{\sigma_{x_i}^2 + \sigma_{y_i}^2 - 2\sigma_{xy_i}}{\sigma_{x_i}^2 + \sigma_{y_i}^2 + C_2} \tag{12}$$

TABLE I
RELATIVE ERROR BETWEEN THE ACTUAL D_{SSIM} CALCULATED BY (3) AND THE VALUES CALCULATED BY (15).

	A	B	C	D	E	avg.
AI	0.0%	0.1%	0.1%	0.6%	0.2%	0.2%
LDB	0.1%	0.2%	0.5%	1.1%	0.3%	0.5%

Substituting (2) into (12), SSIM can be calculated as

$$\begin{aligned}\text{SSIM}_i &= 1 - \frac{\sum_{l=1}^{L}\omega_l(x_{i+l}-y_{i+l})^2}{\sigma_{x_i}^2+\sigma_{y_i}^2+C_2}\\ &= 1-(x_i-y_i)^2\sum_{l=1}^{L}\frac{\omega_l}{\sigma_{x_{i-l}}^2+\sigma_{y_{i-l}}^2+C_2}\\ &= 1-e_i^2\cdot W_i,\end{aligned} \tag{13}$$

where e_i^2 represents the pixel-level squared error and W_i denotes the weight related to the content around the pixel. Specifically,

$$\begin{aligned}e_i^2 &= (x_i-y_i)^2,\\ W_i &= \sum_{l=1}^{L}\frac{\omega_l}{\sigma_{x_{i-l}}^2+\sigma_{y_{i-l}}^2+C_2}.\end{aligned} \tag{14}$$

In the calculation of W_i, the variance σ_y^2 of the distorted frame is unavailable before encoding. In practice, we approximate it by σ_x^2. The PCCs between σ_x^2 and σ_y^2 are illustrated in Fig. 3. Although not as high as that between μ_x and μ_y, the PCCs exceed 0.96 for all test videos, which justifies that the approximation is available.

Based on (13) and (3), D_{SSIM} of a CTU can be calculated as

$$D_{\text{SSIM}} = \frac{1}{M}\sum_{i\in\text{CTU}} e_i^2\cdot W_i. \tag{15}$$

In Table I, we measure the difference between the actual D_{SSIM} calculated by (3) and the estimated values calculated by (15). The results show that the simplifications used in (15), i.e., $\mu_x=\mu_y$ and $\sigma_x^2=\sigma_y^2$, only bring less than 0.5% error.

According to (15), D_{SSIM} approximates a pixel-weighted MSE, where the weights depend on the image content around each pixel. Therefore, to obtain a CTU-level D_{SSIM}-D_{MSE} model, it is necessary to simplify the weighting from pixels to sub-block and then to CTU.

Specifically, first, note that σ_x^2 in W_i are calculated based on Gaussian weighting in a local area as in (2), and then further filtered by Gaussian in (14), implying that W_i will be similar within a small region. We calculate the average coefficients of variation (the ratio of the standard deviation to the mean) of W_i in a 4×4 block. The average results for the videos from class A to class E are 0.22, 0.22, 0.17, 0.19, 0.25, respectively, which verify the similarity of W_i in local areas. Thereby, we use the average W_i of a 4×4 sub-block for weighting as a simplification of the pixel weighting in (15), i.e.,

$$D_{\text{SSIM}} \approx \frac{1}{M}\sum_{t=1}^{T}\left(\sum_{i\in\text{blk}_t} e_i^2\cdot\frac{1}{16}\sum_{i\in\text{blk}_t}W_i\right), \tag{16}$$

where blk_t denotes the t-th 4×4 subblock in a CTU. Table II lists the relative error of (16) compared to (15). The results show that the simplification brings only 3% and 4% error in average in AI and LDB configurations, respectively.

TABLE II
RELATIVE ERROR OF (16) COMPARED TO (15). THE EXPERIMENTAL SETTING IS CONSISTENT WITH THAT IN FIG. 2.

	A	B	C	D	E	avg.
AI	3%	4%	3%	3%	3%	3%
LDB	4%	6%	4%	3%	4%	4%

TABLE III
PCC BETWEEN MSE AND QUANTIZATION STEP FOR THE SUBBLOCKS.

	A	B	C	D	E	avg.
AI	0.93	0.91	0.88	0.87	0.83	0.88
LDB	0.91	0.89	0.85	0.84	0.81	0.86

Secondly, in [34], Wang and Kwong proposed that MSE of a macroblock increases linearly with the quantization step in H.264/AVC. A similar model can also be found in [35], where MSE is proposed to be proportional to the quantization step in the coding unit level of HEVC. We find that the relationship still holds for the subblocks of HEVC, which can be described as:

$$\sum_{i\in blk_t} e_i^2 \approx \rho_t Q_t, \tag{17}$$

where ρ_t is the linear model parameter related to image content, and Q_t is the applied quantization step of the subblock. We calculate the PCC between $\sum_{i\in blk_t} e_i^2$ and Q_t for different videos. The results in Table III show that in AI and LDB configurations, the corresponding average PCCs are 0.88 and 0.86, respectively, which verifies the rationality of (17).

Based on (17), D_{MSE} of a CTU can be calculated as

$$D_{\text{MSE}} = \frac{1}{M}\sum_{t=1}^{T}\rho_t Q_t, \tag{18}$$

while the D_{SSIM} in (16) can be rewritten as:

$$D_{\text{SSIM}} = \frac{1}{M}\sum_{t=1}^{T}\left(\rho_t Q_t\cdot\frac{1}{16}\sum_{i\in\text{blk}_t}W_i\right). \tag{19}$$

Since the same quantization step is usually applied for all the sub-blocks in a CTU, the following CTU-level D_{SSIM}-D_{MSE} relationship can be obtained based on (18) and (19):

$$\begin{aligned}D_{\text{SSIM}} &= \frac{\sum_{t=1}^{T}\left(\rho_t\cdot\frac{1}{16}\sum_{i\in blk_t}W_i\right)}{\sum_{t=1}^{T}\rho_t}D_{\text{MSE}},\\ &= \Theta\cdot D_{\text{MSE}},\end{aligned} \tag{20}$$

where the Θ is used to represent the slope. In (20), ρ_t is related to the image content. Due to the similarity in the image content between the collocated CTUs, we calculate ρ_t as $\frac{\sum_{i\in blk_t} e_i^2}{Q_t}$ based on the encoding results of its encoded-collocated block for simplification.

To compensate for the error caused by the simplification, the least squares regression between the collocated CTUs is applied, which can be expressed as:

$$D_{\text{SSIM}} = \theta\cdot\Theta\cdot D_{\text{MSE}} + \eta, \tag{21}$$

TABLE IV
COMPARISON BETWEEN YEO'S MODEL (9), DCT-DOMAIN MODEL (10), THE PROPOSED (20), AND THE ENHANCED MODEL (21) WITH THE LEAST SQUARES REGRESSION IN TERMS OF PCC.

	AI				LDB			
class	Yeo	DCT	(20)	(21)	Yeo	DCT	(20)	(21)
A	0.77	0.88	0.95	0.99	0.78	0.88	0.91	0.93
B	0.66	0.81	0.93	0.98	0.64	0.79	0.86	0.91
C	0.8	0.84	0.94	0.99	0.82	0.85	0.93	0.97
D	0.85	0.8	0.95	0.99	0.85	0.77	0.93	0.95
E	0.87	0.8	0.98	0.99	0.88	0.8	0.97	0.98
avg.	0.77	0.82	0.95	0.99	0.77	0.81	0.91	0.94

TABLE V
COMPARISON BETWEEN YEO'S MODEL (9), DCT-DOMAIN MODEL (10), THE PROPOSED (20), AND THE ENHANCED MODEL (21) WITH THE LEAST SQUARES REGRESSION IN TERMS OF RELATIVE ERRORS.

	AI				LDB			
class	Yeo	DCT	(20)	(21)	Yeo	DCT	(20)	(21)
A	15%	27%	4%	3%	17%	27%	8%	6%
B	13%	22%	6%	3%	15%	21%	11%	7%
C	31%	23%	12%	4%	33%	23%	21%	12%
D	51%	28%	17%	4%	49%	26%	28%	16%
E	23%	18%	15%	4%	24%	18%	18%	6%
avg.	27%	23%	11%	4%	28%	23%	18%	10%

where θ and η are the linear parameters updated between the collocated CTUs.

As shown in the 4-th column of Fig. 1, the proposed (20) achieves better modeling performance than Yeo's model and the DCT-domain model for the test videos. And, the proposed (21) further improves the performance. To quantitatively evaluate the modeling accuracy of the proposed model, we calculate the PCC and the relative error between the actual D_{SSIM} of the CTUs in a frame and the estimated values calculated by different models. The results are summarized in Table IV and Table V, respectively. The conclusion can be drawn from the results, that is, the proposed model (21) and (20) achieved the best and second best accuracy, respectively.

B. Joint SSIM-based OBA and SSIM-based RDO

In the above subsections, an accurate CTU-level D_{SSIM}-D_{MSE} model is established. Benefit from the proposed model, SSIM-based OBA and SSIM-based RDO can be unified in the proposed SOSR scheme as follows.

1) SSIM-based OBA: To solve the SSIM-based OBA problem, we adopt the widely used hyperbolic function as the R-D_{SSIM} model, which is expressed as

$$D_{\text{SSIM}} = \alpha \cdot R^{\beta}, \tag{22}$$

where α and β are the model parameters. To verify its effectiveness, we compare it with the exponential [10] and logarithmic [17] models. Table VI summarizes the R-D_{SSIM} modeling accuracy of the three models in terms of R squared. It can be seen from the results that the adopted model has the best modeling accuracy.

Substituting (22) into (6), the R-λ_{SSIM} model can be obtained:

$$\lambda_{\text{SSIM}} = -\alpha\beta \cdot R^{\beta-1}. \tag{23}$$

With (23) in hand, we can search the optimal λ_{SSIM} by the Bisection method [36] to meet the bits constraint in (4).

2) SSIM-based RDO: After the optimal λ_{SSIM} is calculated, the SSIM-based RDO process can be re-written as follows by substituting the proposed D_{SSIM}-D_{MSE} model (21) into (8):

$$\begin{aligned} &\arg\min_{m} D_{\text{SSIM}} + \lambda_{\text{SSIM}} \cdot R \\ =&\arg\min_{m} (\theta\Theta D_{\text{MSE}} + \eta) + \lambda_{\text{SSIM}} \cdot R \\ =&\arg\min_{m} D_{\text{MSE}} + \frac{1}{\theta\Theta}\lambda_{\text{SSIM}} \cdot R \end{aligned} \tag{24}$$

TABLE VI
R-D_{SSIM} MODELING PERFORMANCE COMPARISON BETWEEN THE EXPONENTIAL, LOGARITHMIC, AND HYPERBOLIC MODELS IN TERMS OF THE R SQUARED.

	AI			LDB		
class	exponential	logarithmic	hyperbolic	exponential	logarithmic	hyperbolic
A	0.98	0.96	0.98	0.80	0.84	0.92
B	0.94	0.95	0.96	0.79	0.84	0.87
C	0.96	0.97	0.98	0.85	0.88	0.92
D	0.97	0.97	0.97	0.85	0.89	0.94
E	0.85	0.86	0.89	0.73	0.76	0.81
avg.	0.94	0.94	0.96	0.81	0.85	0.89

According to (24), the SSIM-based RDO can be achieved based on the R-D_{MSE} cost with the Lagrangian multiplier λ_{MSE}, which is a scaled λ_{SSIM}, i.e.,

$$\lambda_{\text{MSE}} = \frac{1}{\theta\Theta}\lambda_{\text{SSIM}}. \tag{25}$$

In this way, there is no need to calculate D_{SSIM} for all the candidate modes, while the R-D_{SSIM} cost is still minimized in the RDO process.

3) Joint R-D_{SSIM}-λ_{SSIM}-based R-D_{SSIM} Parameter Estimation: After encoding the CTU with the mode selected by the SSIM-based RDO, the actual D_{SSIM} and R are generated. Thereby, in the joint R-D-λ relationship:

$$\begin{cases} D_{\text{SSIM}} = \alpha \cdot R^{\beta}, \\ \lambda_{\text{SSIM}} = \theta\Theta\lambda_{\text{MSE}} = -\alpha\beta \cdot R^{\beta-1}, \end{cases} \tag{26}$$

only α and β are unknown. Thus, they can be uniquely solved as follows:

$$\begin{cases} \alpha = \dfrac{D_{\text{SSIM}}}{R^{-\theta\Theta\lambda_{\text{MSE}}R/D_{\text{SSIM}}}}, \\ \beta = -\dfrac{\theta\Theta\lambda_{\text{MSE}}R}{D_{\text{SSIM}}}. \end{cases} \tag{27}$$

The solved parameters will be used to solve the SSIM-based OBA (Section III-B1) for the subsequent frame.

In [8], the MSE-based R-D-λ joint relationship is used for R-D_{MSE} model estimation, which is similar as the method in (26). It is worth noting that without the λ_{SSIM}-λ_{MSE} relationship proposed in (25), the value of λ_{SSIM} that is associated with the R and D_{SSIM} generated by encoding is unknown, and then the R-D-λ joint relationship in (26) cannot be uniquely solved. Therefore, to the best of our knowledge, this is the first time the R-D-λ joint relationship exploited in the SSIM-based studies.

IV. Experiments And Discussions

A. Experiment Setup

This section presents the experimental comparison to demonstrate the advantage of the proposed SOSR scheme. In particular, we have implemented the SOSR into HM16.20. Since main goal of this paper is to solve the inconsistency of the optimization goals of OBA and RDO in existing SSIM-based OBA studies, the state-of-the-art SSIM-based OBA studies, i.e., Gao's scheme [16] and Zhou's scheme [17], are the main competitors of our study. In addition, Li's scheme [8], which is a state-of-the-art rate control scheme based on MSE, is used as a reference for encoding performance comparison.

For performance evaluation, we use the same setup as [16] and [17]. Specifically, AI and LDB (both hierarchical and non-hierarchical) configurations are adopted in performance evaluation. In each configuration, the video sequences from class A to class E are encoded at four QPs (22, 27, 32, and 37) [32]. Then, the resulted bitrates are set as the bit constraints for an OBA scheme. For performance evaluation, JCTVC-K0103 [5] and JCTVC-M0036 [6] are set as the anchor schemes for intra and inter encoding, respectively. By comparing a scheme to the anchor, the R-D_{SSIM} performance of the scheme is evaluated in terms of SSIM-based Bjøntegaard delta bit rate (BDBR) and Bjøntegaard delta SSIM (BD-SSIM) [37]. BDBR calculates the relative increase of bitrate at the same SSIM compared with the anchor, while BD-SSIM calculates the absolute gain of SSIM at the same bitrate. The negative BDBR and positive BD-SSIM indicate an improved R-D_{SSIM} performance. Besides, different implementations of SSIM will yield different BD-SSIM values. Thus, BDBR that is a relative measure is more credible to verify the R-D_{SSIM} performance than BD-SSIM. In this paper, the standard implementation of SSIM [38] is adopted.

Besides, the video sequences in class F contain unnatural screen content, while SSIM is only designed to evaluate natural image quality. Therefore, class F is usually excluded in the studies dedicated to optimizing the SSIM-based video quality, such as [16]–[18], [25]. In addition, compared with the LDB configuration, random access configuration can achieve better encoding quality, leaving less room for quality improvement. Thus, it is also usually excluded in the SSIM-based studies [16]–[18], [25]. Therefore, the video sequences in class F and the random access configuration are not included in the R-D performance comparison in main body of this paper. Instead, the corresponding results of the proposed SOSR scheme as well as the results in the low-delay-P configuration are provided in the Appendix.

B. R-D Performance Comparison

For intra encoding, Table VII summarizes the R-D_{SSIM} performance of different schemes. Compared with JCTVC-K0103, the proposed SOSR scheme achieves 9.6% BDBR saving and 0.0019 BD-SSIM improvement, both of which are better than the results of Gao's scheme. In addition, the proposed scheme is also compared with JCTVC-M0257 that is the default intra OBA scheme of HM16.20. As shown in Table VII, the BDBR and BD-SSIM gains of the proposed model are 5.2% and 0.0008, respectively, which further verify the effectiveness of the proposed model.

TABLE VII
R-D performance of intra encoding in terms of BDBR (SSIM) and BD-SSIM. Configuration: AI.

	anchor: K0103 [5]				anchor: M0257 [39]	
	BDBR (SSIM)		BD-SSIM		BDBR (SSIM)	BD-SSIM
class	Gao [16]	SOSR	Gao [16]	SOSR	SOSR	SOSR
A	-2.0	-15.1	0.0010	0.0009	-7.4	0.0004
B	-2.0	-11.0	0.0009	0.0014	-6.1	0.0007
C	-4.1	-8.4	0.0023	0.0033	-3.5	0.0013
D	-4.1	-3.3	0.0022	0.0030	-1.1	0.0010
E	-1.6	-10.1	0.0003	0.0011	-7.8	0.0008
avg.	-2.7	-9.6	0.0013	0.0019	-5.2	0.0008

TABLE VIII
R-D performance of inter encoding compared with Li [8] in terms of BDBR (SSIM) and BD-SSIM. Anchor: JCTVC-M0036 [6].

	hierarchical LDB				non-hierarchical LDB			
	BDBR (SSIM)		BD-SSIM		BDBR (SSIM)		BD-SSIM	
class	Li	SOSR	Li	SOSR	Li	SOSR	Li	SOSR
A	-5.4	-19.0	0.0003	0.0010	-8.4	-28.1	0.0005	0.0017
B	-4.9	-15.1	0.0006	0.0019	-9.8	-22.0	0.0016	0.0035
C	-1.8	-8.6	0.0006	0.0031	-3.1	-11.5	0.0012	0.0050
D	-1.6	-8.2	0.0011	0.0053	-3.0	-13.9	0.0022	0.0104
E	-4.6	-6.9	0.0004	0.0006	-5.2	-11.7	0.0005	0.0011
avg.	-3.7	-11.2	0.0006	0.0026	-5.9	-17.4	0.0012	0.0043

For inter encoding, Table VIII summarizes the R-D performance comparison with Li's [8] scheme, which is implemented in HM16.20 based on their source codes. The results show that the proposed SOSR scheme presents better performance than Li's scheme in terms of both BDBR and BD-SSIM. Specifically, SOSR saves more than 11% and 17% encoding bits at the same SSIM respectively in hierarchical and non-hierarchical LDB configurations, while Li's scheme saves only 3.7% and 5.9%, respectively. In Table IX, comparison with Zhou's scheme [17] is presented. Zhou's scheme was implemented based on HM16.19. Because there is little difference in inter encoding performance between this platform and the HM16.20 we used, we refer to the results in their paper for comparison. Moreover, their results are evaluated based on the SSIM implemented in DCT-domain [20]. Therefore, for fair comparison, our method is also evaluated based on the DCT-domain SSIM. As can be seen from Table IX, the proposed SOSR achieves better BDBR and competitive BD-SSIM compared with Zhou's scheme. Moreover, it can be seen from Table VIII and Table IX that under the standard implementation of SSIM and the DCT-domain SSIM, the proposed SOSR shows similar excellent BDBR performance, which further validates the superior performance of our method.

C. Analysis of the Proposed Model

Compared with other similar schemes, the main innovation of our proposed SOSR scheme lies in the novel D_{SSIM}-D_{MSE} model, the joint optimization of SSIM-based OBA and SSIM-based RDO, as well as the accurate R-D_{SSIM} model estimation. This subsection analyzes the contribution of each of these innovations to the encoding performance.

TABLE IX
R-D PERFORMANCE OF INTER ENCODING COMPARED WITH ZHOU [17] IN TERMS OF BDBR (SSIM) AND BD-SSIM THAT ARE CALCULATED BASED ON THE DCT-DOMAIN SSIM. ANCHOR: JCTVC-M0036 [6].

	hierarchical LDB				non-hierarchical LDB			
	BDBR (SSIM)		BD-SSIM		BDBR (SSIM)		BD-SSIM	
class	Zhou	SOSR	Zhou	SOSR	Zhou	SOSR	Zhou	SOSR
A	-5.8	-13.4	0.0030	0.0038	-11.7	-18.4	0.0067	0.0059
B	-4.9	-12.9	0.0027	0.0037	-13.9	-17.1	0.0076	0.0061
C	-4.9	-9.3	0.0027	0.0043	-12.3	-12.3	0.0074	0.0067
D	-12.2	-11.4	0.0074	0.0058	-22.8	-19.6	0.0139	0.0116
E	-5.1	-6.8	0.0027	0.0010	-9.4	-12.2	0.0064	0.0018
avg.	-6.6	-10.8	0.0037	0.0039	-14.0	-15.9	0.0084	0.0064

TABLE X
R-D_{SSIM} PERFORMANCE OF THE PROPOSED SOSR SCHEME WITH THREE DIFFERENT D_{MSE}-D_{SSIM} MODELS IN TERMS OF BDBR (SSIM).

class	Yeo (9)	DCT (10)	proposed (21)
A	-13.6	-16.2	-19.0
B	-11.2	-13.1	-15.1
C	-5.0	-6.7	-8.6
D	-4.8	-6.0	-8.2
E	-6.3	-6.8	-6.9
avg.	-7.8	-9.4	-11.2

First, in the SOSR scheme, we replace the proposed D_{SSIM}-D_{MSE} model with the Yeo's model (9) and the DCT-domain model (10), respectively. Table X shows the corresponding SSIM-based BDBR comparison in the hierarchical LDB configuration. As can be seen from the results, the proposed SOSR scheme can achieve R-D_{SSIM} performance improvement based on all the three models. Similar to the accuracy comparison of the three models in Table IV, the Yeo's model has the worst R-D_{SSIM} performance, achieving 7.8% BDBR savings, and the DCT-domain model has the second best R-D_{SSIM} performance, achieving 9.4% BDBR savings. The proposed D_{SSIM}-D_{MSE} model performs best, achieving 11.2% BDBR savings, which highlights the significance of the proposed model in R-D_{SSIM} performance improvement.

Secondly, we disabled the SSIM-based RDO in the proposed SOSR scheme. Specifically, after allocating bits by the SSIM-based OBA as described in Section III-B1, the MSE-based default scheme of HM16.20 [6] is adopted to calculated λ_{MSE}, which is used to achieve the allocated bits and is used for RDO. In this way, the proposed SOSR degenerates into the conventional method as [14]–[17] that combines the SSIM-based OBA and MSE-based RDO. As shown in Table XI, after the SSIM-based RDO is disabled, the encoding performance is greatly reduced. The difference in encoding performance between different models is also reduced. Specifically, compared with the results in Table X, the BDBR savings of SOSR with the Yeo's model, DCT-domain model, and the proposed D_{SSIM}-D_{MSE} model are reduced to 2.3%, 4.4%, and 4.4%, respectively.

As has been discussed, the performance degradation is due to the inconsistent optimization goals of OBA and RDO at this time. Moreover, the default scheme of HM uses the traditional regression method for R-λ_{MSE} model estimation [6]. As shown in Fig. 4, compared with the R-D_{SSIM}-λ_{SSIM} relationship-based estimation in (26) we used, the traditional method brings larger error between the target encoding bits and the actual encoding bits. Accordingly, the allocated bits will not be accurately achieved, thereby deviating from the optimal encoding.

TABLE XI
R-D_{SSIM} PERFORMANCE OF THE PROPOSED SOSR SCHEME WITH THREE DIFFERENT D_{MSE}-D_{SSIM} MODELS IN TERMS OF BDBR (SSIM) WHEN THE SSIM-BASED RDO IS DISABLED.

class	Yeo (9)	DCT (10)	proposed (21)
A	-5.5	-9.6	-9.9
B	-1.5	-4.7	-4.8
C	-1.3	-2.6	-2.4
D	-3.1	-4.1	-4.3
E	-1.9	-2.9	-2.9
avg.	-2.3	-4.4	-4.4

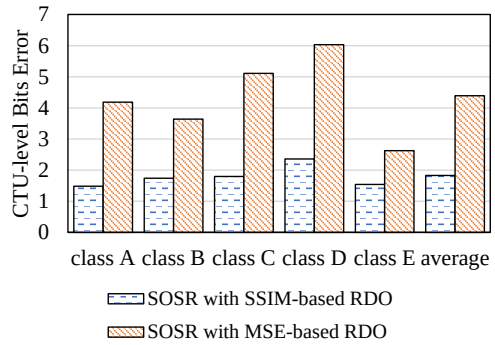


Fig. 4. CTU-level relative error between the actual encoding bits and the targeted bits for the proposed SOSR with SSIM-based RDO and SOSR with MSE-based RDO.

In summary, we can conclude that the three innovations all contribute to the R-D_{SSIM} performance improvement of the proposed SOSR scheme.

D. SSIM vs. PSNR

In addition to SSIM, the MSE-based peak signal-to-noise ratio (PSNR) is also widely used to evaluate quality of an encoded video. Therefore, in this subsection, we calculate the PSNR-based BDBR and BD-PSNR of different schemes for reference. The results are shown from Table XII to Table XIV for AI, hierarchical LDB and non-hierarchical LDB, respectively. As shown in these tables, the proposed SOSR is inferior

TABLE XII
R-D PERFORMANCE COMPARISON FOR INTRA ENCODING IN TERMS OF BDBR (PSNR) AND BD-PSNR. SYMBOL '—' INDICATES THAT THE RESULT WAS NOT PROVIDED IN THE CORRESPONDING PAPER. CONFIGURATION: AI.

	anchor: K0103 [5]				anchor: M0257 [39]	
	BDBR (PSNR)		BD-PSNR		BDBR (PSNR)	BD-PSNR
class	Gao [16]	SOSR	Gao [16]	SOSR	SOSR	SOSR
A	—	4.1	0.08	-0.22	3.2	-0.18
B	—	5.2	0.06	-0.20	2.8	-0.12
C	—	3.1	0.15	-0.17	2.0	-0.12
D	—	2.9	0.19	-0.20	1.5	-0.10
E	—	6.4	0.07	-0.32	3.8	-0.20
avg.	—	4.3	0.11	-0.22	2.7	-0.14

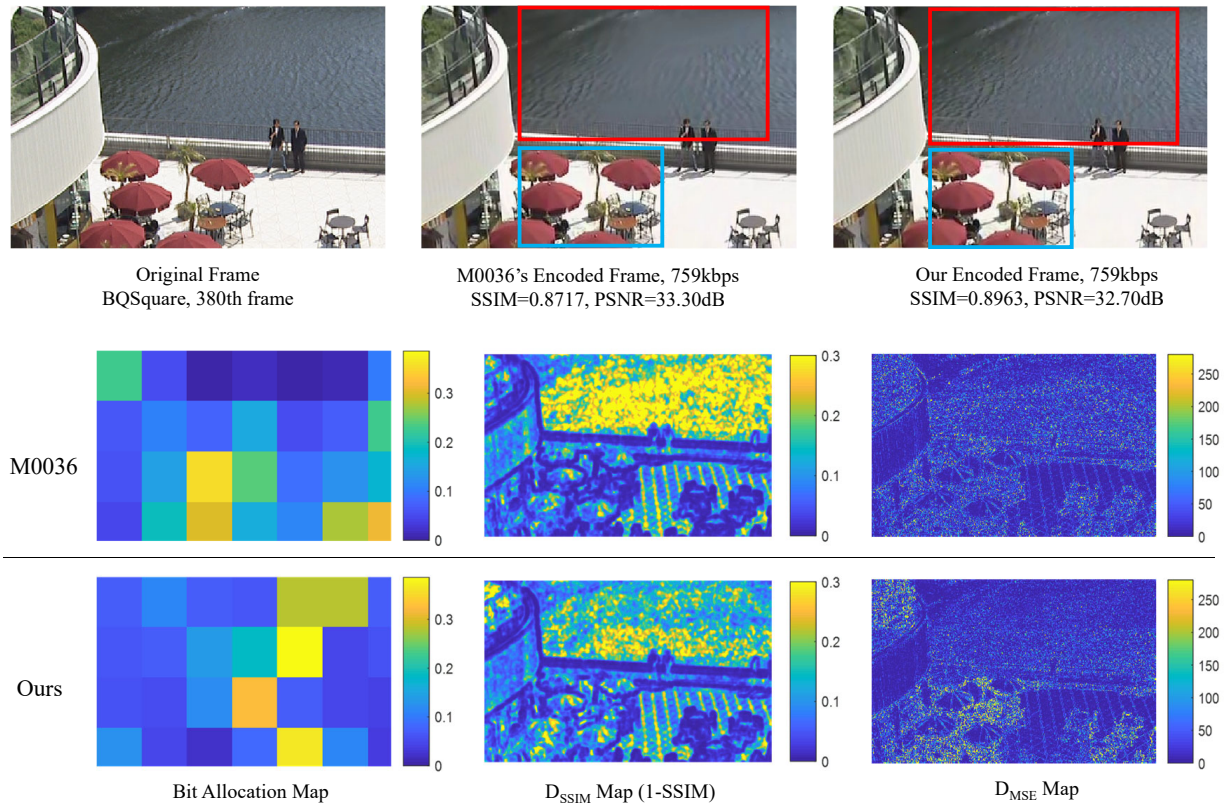


Fig. 5. Visual comparison of frames encoded by JCTVC-M0036 [6] and the proposed SOSR. The D_{SSIM} map is calculated by 1-SSIM. The D_{MSE} map illustrates the pixel-wise D_{MSE}. Example frame: BQSquare, 380th frame, 759kbps, non-hierarchical LDB.

TABLE XIII
R-D PERFORMANCE COMPARISON FOR INTER ENCODING IN TERMS OF BDBR (PSNR) AND BD-PSNR. ANCHOR: JCTVC-M0036 [6]. CONFIGURATION: HIERARCHICAL LDB.

	BDBR (PSNR)			BD-PSNR		
class	Li [8]	Zhou [17]	SOSR	Li [8]	Zhou [17]	SOSR
A	-2.1	-3.2	0.2	0.07	0.12	-0.04
B	-2.6	-3.0	3.3	0.06	0.11	-0.08
C	-0.8	-2.7	1.1	0.03	0.09	-0.05
D	-0.7	-3.3	1.7	0.03	0.13	-0.08
E	-5.8	-3.2	4.0	0.17	0.12	-0.06
avg.	-2.4	-3.1	2.3	0.07	0.11	-0.06

TABLE XIV
R-D PERFORMANCE COMPARISON FOR INTER ENCODING IN TERMS OF BDBR (PSNR) AND BD-PSNR. ANCHOR: JCTVC-M0036 [6]. CONFIGURATION: NON-HIERARCHICAL LDB.

	BDBR (PSNR)			BD-PSNR		
class	Li [8]	Zhou [17]	SOSR	Li [8]	Zhou [17]	SOSR
A	-2.6	-5.2	4.3	0.08	0.23	-0.22
B	-4.9	-2.9	11.6	0.12	0.1	-0.23
C	-2.2	-5.5	6.1	0.09	0.25	-0.21
D	-1.8	-5.8	3.5	0.07	0.26	-0.13
E	-4.3	-5.5	7.6	0.13	0.25	-0.19
avg.	-3.2	-5.0	6.6	0.10	0.22	-0.20

to other schemes in terms of BDBR (PSNR) and BD-PSNR. Compared with HM16.20, our scheme has 2.7%, 2.3%, 6.6% bits increase under the same PSNR in AI, hierarchical LDB, and non-hierarchical LDB configurations, respectively. This is not surprising, because PSNR is calculated based on MSE, which is not the optimization objective of our scheme.

Fig. 5 illustrates the difference in encoding results caused by different optimization objectives, where an example frame is respectively encoded by our scheme and JCTVC-M0036 at the same bitrate. JCTVC-M0036 minimizes D_{MSE} of a frame and is the default inter rate control scheme of HM16.20. Therefore, the encoded frame has better PSNR than the frame encoded by our scheme. However, we can find that the frame encoded by [6] does not have a good visual quality compared with that encoded by our scheme.

Specifically, as shown in the bit allocation map, JCTVC-M0036 allocates a large number of bits to the region marked in blue box and a small number of bits to region marked in red box. Among them, the blue boxed region is highly textured, while the red boxed region has simpler structures. After encoding with JCTVC-M0036, all these regions do not have significant D_{MSE} distortion. However, the water ripple shown in the red boxed region with a lot of structural information, has become smooth after encoding. These structural distortions will attract the attention of the viewer and lead to the decline of subjective perceived quality.

On the other hand, we can see that D_{SSIM} map correctly identifies these structural distortions. Correspondingly, our

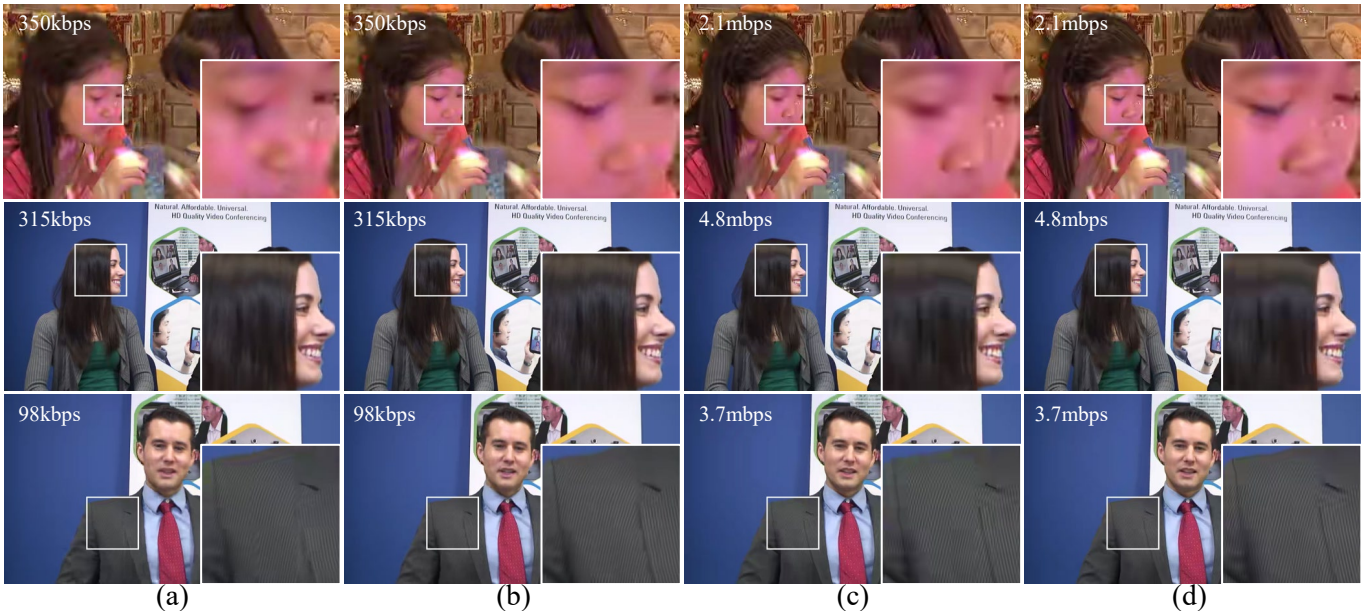


Fig. 6. Visual quality comparison of frames encoded by the proposed SOSR scheme, Gao [16], and Li [8] at different bitrate. The frames from the first to the last row are BlowingBubbles 48th frame, KristenAndSara 60th frame, and Johnny 60th frame. (a) Li [8], non-hierarchical LDB. (b) SOSR, non-hierarchical LDB. (c) Gao [16], AI. (d) SOSR, AI.

TABLE XV
ENCODING TIME COMPARISON.

	Gao [16]	Li [8]	Zhou [17]	SOSR
AI	101.0%	-	-	100.6%
LDB	-	101.3%	102.7%	102.2%

scheme allocates more bits to the water ripple region. It can be clearly seen that structure of this region is preserved and the visual quality is improved. At the same time, because the total bits are constrained, the bits allocated to the blue boxed region by our scheme are reduced, so this region has larger D_{MSE} distortion. However, because of the visual masking effect [40], quality of the blue boxed region is still good, whether based on the visual observation or the D_{SSIM} map. Thus, the frame encoded by our scheme has better visual quality. Therefore, it is reasonable to use D_{SSIM} as the distortion metric in the optimization objective.

In Fig. 6, we further compare the visual quality performance of the proposed SOSR with two available encoders Gao [16] and Li [8]. Both Gao [16] and Li [8] achieve better R-D_{SSIM} performance than the default HM. Fig. 6 demonstrates that the frames encoded by the proposed SOSR has better visual quality than the other two schemes. In particular, SOSR better retains the structural information as shown in the zoomed-in regions in Fig. 6, while the other schemes bring some severe distortions, such as blurring, which reduces the visual quality.

E. Complexity Comparison

To evaluate the computational complexity of our scheme, Table XV compares its encoding time with the default HM encoder (i.e., JCTVC-M0257 for AI and JCTVC-M0036 for LDB). The test was carried out on an AMD 3900X processor with 32GB memory. The results show that our scheme and the other three competitors have brought only a small increase in encoding time. In addition, compared with the other two SSIM-based schemes (i.e., Gao [16] and Zhou [17]), the time increase of our scheme is slightly smaller.

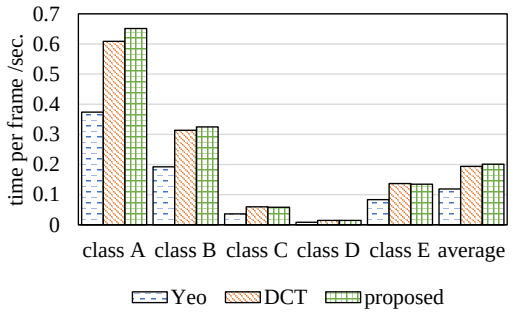


Fig. 7. Complexity comparison of three different D_{SSIM}-D_{MSE} models implemented in HM16.20, including Yeo's model (9), DCT-domain model (10), and the proposed model (21).

Besides, we also measure the complexity of the D_{SSIM}-D_{MSE} model estimation which is a main component of the proposed SOSR. Fig. 7 illustrates the average processing time for a frame of different models. As can be seen from the results, Yeo's model (9) has the lowest complexity. Besides, since the fast DCT algorithm exists [41], the DCT-domain model (10) has the second highest complexity, while the proposed model (21) has similar but just a little larger time-consuming to that of the DCT-domain model.

V. CONCLUSION

In this study, a scheme called SOSR is proposed to unify the SSIM-based optimal bit allocation (OBA) and SSIM-based

rate-distortion optimization (RDO) for HEVC. To achieve this goal, an accurate CTU-level D_{SSIM}-D_{MSE} model is first proposed, which has been validated to be more accurate than two widely used models. With this model, the SSIM-based RDO can be performed based on the low-complexity R-D_{MSE} cost with an SSIM-related Lagrangian multiplier, which is determined by the SSIM-based OBA. Moreover, the joint relationship of SSIM-based R-D-λ can be exploited to achieve more accurate R-D_{SSIM} model estimation, which further benefit the solving of OBA. In this way, OBA and RDO are optimized based on SSIM consistently in this study. Experimental results have validated that the proposed SOSR has a superior R-D_{SSIM} performance compared with other state-of-the-art studies in three commonly used configurations. According to our experimental analysis, the accurate D_{SSIM}-D_{MSE} model and the SSIM-based joint optimization of OBA and RDO both contribute to the R-D_{SSIM} performance improvement.

APPENDIX

TABLE A1
R-D PERFORMANCE OF SOSR FOR VIDEOS OF CLASS F IN DIFFERENT CONFIGURATIONS.

	PSNR-based performance		SSIM-based performance	
	BDBR(PSNR)	BD-PSNR	BDBR(SSIM)	BD-SSIM
AI	4.7	-0.2244	-2.3	-0.0007
hier. LDB	3.4	-0.3114	-5.8	0.0021
non-hier. LDB	-1.9	0.8839	-20.1	0.0054
random access	-5.3	0.4311	-19.2	0.0022

TABLE A2
R-D PERFORMANCE OF SOSR IN HIERARCHICAL LOW DELAY P CONFIGURATION.

	PSNR-based performance		SSIM-based performance	
class	BDBR(PSNR)	BD-PSNR	BDBR(SSIM)	BD-SSIM
A	0.0	-0.0363	-10.2	0.0026
B	3.2	-0.0885	-9.0	0.0018
C	1.9	-0.0668	-7.0	0.0024
D	1.6	-0.0688	-8.1	0.0049
E	5.0	-0.0885	-5.9	0.0007
Avg.	2.3	-0.0698	-8.1	0.0025

TABLE A3
R-D PERFORMANCE OF SOSR IN NON-HIERARCHICAL LOW DELAY P CONFIGURATION.

	PSNR-based performance		SSIM-based performance	
class	BDBR(PSNR)	BD-PSNR	BDBR(SSIM)	BD-SSIM
A	1.5	-0.1176	-14.5	0.0042
B	8.2	-0.1776	-13.8	0.0038
C	5.6	-0.1757	-12.0	0.0050
D	3.2	-0.1167	-13.7	0.0098
E	9.0	-0.2151	-7.4	0.0009
Avg.	5.9	-0.1605	-12.3	0.0047

TABLE A4
R-D PERFORMANCE OF SOSR IN RANDOM ACCESS CONFIGURATION.

	PSNR-based performance		SSIM-based performance	
class	BDBR(PSNR)	BD-PSNR	BDBR(SSIM)	BD-SSIM
A	-3.4	0.1446	-7.7	0.0016
B	-2.0	0.0719	-4.3	0.0011
C	-2.2	0.1288	-4.5	0.0014
D	-1.9	0.1394	-4.9	0.0033
E	-1.4	0.0489	-4.9	0.0007
avg.	-2.2	0.1067	-5.3	0.0016

It can be seen from Table A1 that except for the BD-SSIM in all intra configuration, our scheme has achieved significant improvements in both SSIM-based BDBR and BD-SSIM in all other configurations. However, it should be noted that the videos of class F contain unnatural screen content, while SSIM is only designed to evaluate natural image quality. Therefore, the rationality of SSIM for videos of class F needs to be further studied.

From Table A2 to Table A4, we can see that the proposed SOSR achieves 12.3%, 8.1%, and 5.3% SSIM-based BDBR savings in non-hierarchical low delay P, hierarchical low delay P, and random access configurations. Moreover, SOSR also achieves a PSNR-based BDBR savings of 2.2% in random access configuration. This is because the temporal spacing between frames in the same hierarchical level in the random access configuration is larger than that in LDB configuration. Thereby, the regression-based R-λ_{MSE} model estimation [6] used by HM is more inaccurate. In contrast, the accurate R-D-λ joint relationship-based modeling method used in SOSR helps to solve this problem.

REFERENCES

[1] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, "Overview of the high efficiency video coding (HEVC) standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1649–1668, 2012.

[2] T. Wiegand, H. Schwarz, A. Joch, F. Kossentini, and G. J. Sullivan, "Rate-constrained coder control and comparison of video coding standards," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 688–703, 2003.

[3] T. Chiang and Y.-Q. Zhang, "A new rate control scheme using quadratic rate distortion model," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 7, no. 1, pp. 246–250, 1997.

[4] H. Choi, J. Yoo, J. Nam, D. Sim, and I. V. Bajic, "Pixel-wise unified rate-quantization model for multi-level rate control," *IEEE Journal of Selected Topics in Signal Processing*, vol. 7, no. 6, pp. 1112–1123, dec 2013.

[5] B. Li, H. Li, L. Li, and J. Zhang, "Rate control by R-lambda model for HEVC," *ITU-T SG16 Contribution, JCTVC-K0103*, pp. 1–5, 2012.

[6] B. Li, H. Li, and L. Li, "Adaptive bit allocation for R-lambda model rate control in HM," *ITU-T SG16 Contribution, JCTVC-M0036*, pp. 1–7, 2013.

[7] W. Gao, S. Kwong, and Y. Jia, "Joint machine learning and game theory for rate control in high efficiency video coding," *IEEE Transactions on Image Processing*, vol. 26, no. 12, pp. 6074–6089, dec 2017.

[8] S. Li, M. Xu, Z. Wang, and X. Sun, "Optimal bit allocation for CTU level rate control in HEVC," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 11, pp. 2409–2424, nov 2017.

[9] "HEVC Reference Software 16.20," https://hevc.hhi.fraunhofer.de/svn/svn\-_HEVCSoftware/tags/HM-16.20/, Sep. 2018.

[10] G. J. Sullivan, T. Wiegand *et al.*, "Rate-distortion optimization for video compression," *IEEE Signal Processing Magazine*, vol. 15, no. 6, pp. 74–90, 1998.

[11] A. Ortega and K. Ramchandran, "Rate-distortion methods for image and video compression," *IEEE Signal Processing Magazine*, vol. 15, no. 6, pp. 23–50, 1998.

[12] H. Guo, C. Zhu, M. Xu, and S. Li, "Inter-block dependency-based CTU level rate control for HEVC," *IEEE Transactions on Broadcasting*, vol. 66, no. 1, pp. 113–126, 2019.

[13] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli *et al.*, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[14] T.-S. Ou, Y.-H. Huang, and H. H. Chen, "SSIM-based perceptual rate control for video coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 21, no. 5, pp. 682–691, may 2011.

[15] C. Wang, X. Mou, and L. Zhang, "Optimization of the block-level bit allocation in perceptual video coding based on MINMAX," *arXiv preprint arXiv:1511.04691*, 2015.

[16] W. Gao, S. Kwong, Y. Zhou, and H. Yuan, "SSIM-based game theory approach for rate-distortion optimized intra frame CTU-level bit allocation," *IEEE Transactions on Multimedia*, vol. 18, no. 6, pp. 988–999, jun 2016.

[17] M. Zhou, X. Wei, S. Wang, S. Kwong, C.-K. Fong, P. Wong, W. Yuen, and W. Gao, "SSIM-based global optimization for CTU-level rate control in HEVC," *IEEE Transactions on Multimedia*, 2019.

[18] S. Wang, A. Rehman, K. Zeng, and Z. Wang, "SSIM-inspired two-pass rate control for High Efficiency Video Coding," in *2015 IEEE 17th International Workshop on Multimedia Signal Processing (MMSP)*. IEEE, oct 2015.

[19] S. Wang, A. Rehman, Z. Wang, S. Ma, and W. Gao, "SSIM-motivated rate-distortion optimization for video coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 4, pp. 516–529, 2012.

[20] ——, "Perceptual video coding based on SSIM-inspired divisive normalization," *IEEE Transactions on Image Processing*, vol. 22, no. 4, pp. 1418–1429, 2013.

[21] C. Yeo, H. L. Tan, and Y. H. Tan, "On rate distortion optimization using SSIM," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 7, pp. 1170–1181, jul 2013.

[22] W. Dai, O. C. Au, W. Zhu, P. Wan, W. Hu, and J. Zhou, "SSIM-based rate-distortion optimization in H.264," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 7343–7347.

[23] J. Qi, X. Li, F. Su, Q. Tu, and A. Men, "Efficient rate-distortion optimization for HEVC using SSIM and motion homogeneity," in *2013 Picture Coding Symposium (PCS)*. IEEE, 2013, pp. 217–220.

[24] C. Yeo, H. L. Tan, and Y. H. Tan, "SSIM-based adaptive quantization in HEVC," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 1690–1694.

[25] A. Rehman and Z. Wang, "SSIM-inspired perceptual video coding for HEVC," in *2012 IEEE International Conference on Multimedia and Expo*. IEEE, 2012, pp. 497–502.

[26] Z.-Y. Mai, C.-L. Yang, L.-M. Po, and S.-L. Xie, "A new rate-distortion optimization using structural information in H.264 I-frame encoder," in *International Conference on Advanced Concepts for Intelligent Vision Systems*. Springer, 2005, pp. 435–441.

[27] C.-L. Yang, R.-K. Leung, L.-M. Po, and Z.-Y. Mai, "An SSIM-optimal H.264/AVC Inter frame encoder," in *IEEE International Conference on Intelligent Computing and Intelligent Systems*, vol. 4. IEEE, 2009, pp. 291–295.

[28] Y.-H. Huang, T.-S. Ou, P.-Y. Su, and H. H. Chen, "Perceptual rate-distortion optimization using structural similarity index as quality metric," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 20, no. 11, pp. 1614–1624, 2010.

[29] S. S. Channappayya, A. C. Bovik, and R. W. Heath, "Rate bounds on SSIM index of quantized images," *IEEE Transactions on Image Processing*, vol. 17, no. 9, pp. 1624–1639, Sep. 2008.

[30] F. Cen, Q. Lu, and W. Xu, "SSIM based rate-distortion optimization for intra-only coding in HEVC," in *2014 IEEE International Conference on Consumer Electronics (ICCE)*. IEEE, 2014, pp. 17–18.

[31] J. Ban, H. Lai, and X. Lin, "A novel method rate distortion optimization for HEVC based on improved SSIM," in *2016 9th International Symposium on Computational Intelligence and Design (ISCID)*, vol. 2. IEEE, 2016, pp. 260–263.

[32] F. Bossen *et al.*, "Common test conditions and software reference configurations," *JCTVC-L1100*, vol. 12, pp. 1–4, 2013.

[33] R. Taylor, "Interpretation of the correlation coefficient: a basic review," *Journal of Diagnostic Medical Sonography*, vol. 6, no. 1, pp. 35–39, 1990.

[34] H. Wang and S. Kwong, "Rate-distortion optimization of rate control for H.264 with adaptive initial quantization parameter determination," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 1, pp. 140–144, 2008.

[35] Q. Cai, Z. Chen, D. O. Wu, and B. Huang, "Real-time constant objective quality video coding strategy in high efficiency video coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 7, pp. 2215–2228, 2019.

[36] L. Li, B. Li, H. Li, and C. W. Chen, "λ-domain optimal bit allocation algorithm for High Efficiency Video Coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 1, pp. 130–142, jan 2018.

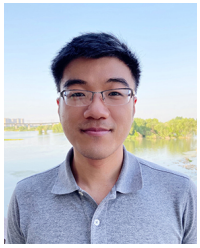
[37] G. Bjøntegaard, "Calculation of average PSNR differences between RD-curves," *VCEG Contribution,VCEG-M33*, 2001.

[38] Z. Wang, "The SSIM index for image quality assessment," https://www.ece.uwaterloo.ca/~z70wang/research/ssim/ssim.m, Feb. 2003.

[39] M. Karczewicz and X. Wang, "Intra frame rate control based on SATD," in *JCT-VC M0257, 13th Meeting*, 2013, pp. 1–5.

[40] J. Ross and H. D. Speed, "Contrast adaptation and contrast masking in human vision," *Proceedings of the Royal Society of London. Series B: Biological Sciences*, vol. 246, no. 1315, pp. 61–70, 1991.

[41] J. Makhoul, "A fast cosine transform in one and two dimensions," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 1, pp. 27–34, 1980.



Yang Li received his B.Sc degree in electronic science and technology from the Shandong Normal University, Jinan, China, in 2012. He is currently pursuing the Ph.D. degree with the Institute of Image Processing and Pattern Recognition, Xi'an Jiaotong University. His research interest includes perceptual video coding.

Xuanqin Mou (Senior Member, IEEE) has been with the School of Electronic and Information Engineering, Institute of Image Processing and Pattern Recognition (IPPR), Xi'an Jiaotong University, Xi'an, China, since 1987, where he has been an Associate Professor and a Professor, since 1997 and 2002, respectively. He is currently the Director of IPPR and the Director of the National Data Broadcasting Engineering and Technology Research Center. He has authored or coauthored over 200 peer-reviewed journal or conference papers. He was a recipient of the Yung Wing Award for Excellence in Education, the KC Wong Education Award, the Technology Academy Award for Invention by the Ministry of Education of China, and the Technology Academy Awards from the Government of Shaanxi Province, China. He served as a member of the 12th Expert Evaluation Committee for the National Natural Science Foundation of China, and the Executive Committee of the China Society of Image and Graphics, the Executive Committee of the Chinese Society for Stereology, and also serves as the Director of the Intelligent Imaging Society for Chinese Stereology.