IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 15, 2019, accepted April 1, 2019, date of current version May 9, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2910245

Perceptual-Based HEVC Intra Coding Optimization Using Deep Convolution Networks

XUEBIN SUN[1], HAN MA[2], WEIXUN ZUO[1], AND MING LIU[1], (Senior Member, IEEE)
[1]Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Hong Kong 999077
[2]Department of Precision Instrument, Tsinghua University, Beijing 100084, China

Corresponding author: Ming Liu (eelium@ust.hk)

This work was supported in part by the National Natural Science Foundation of China under Grant U1713211, and in part by the Research Grant Council of Hong Kong SAR Government, China, under Project 11210017 and Project 21202816. The work of M. Liu was supported by the Shenzhen Science, Technology and Innovation Commission (SZSTI) Project under Grant JCYJ20160428154842603.

ABSTRACT In this paper, we propose a novel perceptual-based intra coding optimization algorithm for the High Efficiency Video Coding (HEVC) using deep convolution networks (DCNs). According to the saliency map, the algorithm can intelligently adjust bit rate allocation between the salient and non-salient regions of the video. The proposed strategy mainly consists of two techniques, saliency map extraction, and intelligent bit rate allocation. First, we train a DCN model to generate the saliency map that highlights semantically salient regions. Compared with the texture-based region of interest (ROI) extraction techniques, our model is more consistent with the human visual system (HVS). Second, based on the saliency map, a modified rate-distortion optimization (RDO) method is designed to adaptively adjust bit rate allocation. As a result, the quality of the salient regions will be improved by allocating more bits while allocating fewer bit rates for the non-salient regions. The experimental results demonstrate that our approach can deal with multiple types of video to enhance the visual experience. For conventional videos, the proposed method achieves 0.64-dB PSNR improvement for the salient regions and saves 3.02% bit rate on average compared with HM16.7. Moreover, for conversational videos, the proposed method can significantly reduce the bit rate by 8.65% without dropping the quality of important regions.

INDEX TERMS Perceptual-based, HEVC, saliency map, DCN, HVS.

I. INTRODUCTION

The High Efficiency Video Coding (HEVC) standard, approved by the Joint Collaborative Team on Video Coding (JCT-VC), greatly outperforms previous standards H.264/AVC in terms of coding bit rate and video quality [1]. The superior compression efficiency is owed to adopting several novel techniques, such as the flexible hierarchical coding structure and multiple prediction modes [2]. HEVC provides three block concepts, which consist of coding unit (CU), prediction unit (PU), and transform unit (TU) [3]. CU is the basic coding unit, with a size ranging from 8×8 to 64×64, which can be split into the PU and TU [4]. The PU is used for prediction and always takes the size of its CU in intra-coding [5]. The TU is the unit for transformation and quantization to a size that does not exceed the CU size [6]. Moreover, to improve the accuracy of intra-prediction, HEVC adopts 35 prediction modes for each PU [7]. To obtain the most suitable size and mode, each CU, PU, and TU block recursively performs a RDO evaluation. This ''try all and select the best'' philosophy can maximize the coding efficiency.

However, the RDO method optimizes the coding performance only based on the conventional objective metric [8], which ignores the perceptual characteristics of the video content. Actually, the quality of different parts of a frame of a video should be different. The salient regions should maintain higher quality, while non-salient regions should be of low quality to save the bandwidth.

Hence, more bits should be allocated to salient areas of a frame of picture, and fewer bits for non-salient areas, to which people pay less attention. Take a news broadcast video for example, it is reasonable to improve the quality of the announcer's regions. Because we pay less attention to the background, we can allocate fewer bits to those regions to save the bandwidth. Therefore, it is highly desirable to

The associate editor coordinating the review of this manuscript and approving it for publication was Madhu S. Nair.

(a) (b)

FIGURE 1. Experiment results of HEVC algorithm and the proposed method for Akiyo video. (a) HEVC coding result. (b) The proposed method coding result.

develop a perceptual-based rate control algorithm, that can largely save the bit rates for some special applications. Figure 1 shows the experimental results of the proposed perceptual-based method compared with the standard HEVC algorithm for the Akiyo video sequence under low bit rates. It can be observed that compared with HEVC, the proposed method obtains clearer facial features and a better visual experience.

Recently, there has been a growing interest in perceptual video coding optimization. More specifically, Wu *et al.* [9] have proposed a medical ultrasound video coding method with HEVC based on ROI map. They develop an effective an ROI extraction technique based on image-textural features. According to the ROI map, the quantization parameter (QP) is adaptively adjusted for ROIs and non-ROIs. Yang *et al.* [10] use the Prewitt filter to extract perceptual features, which are utilized to optimize the RDO process by perceptually adjusting the Lagrangian multiplier. To our best knowledge, the existing perceptual-based video coding approaches merely use deep learning methods to extract the saliency map [11]. Compared with traditional ROI extraction methods, deep learning methods are more consistent with the human visual system.

In this paper, we design a DCN model to locate multiple regions of saliency within each frame of a video. Model training needs only be done offline. Moreover, a perceptual feature guided RDO approach is proposed for HEVC. The proposed method adaptively adjusts the Lagrangian multiplier in the RDO process according to the perceptual characteristics of the video content. Experimental results demonstrate that the proposed method can significantly improve the perceptual coding performance, compared with the original RDO process in HEVC.

The rest of the paper is organized as follows. Section II presents an overview of related works. Section III gives a detailed description of the proposed perceptual-based video coding optimization algorithm. Experimental results are discussed in section IV. Section V concludes with future areas for research.

II. RELATED WORKS

Over the past decade, a large number of scholars focused on perceptual-based image or video coding optimization algorithms. According to the application field, the methods can be classified into three main categories: conversational or surveillance video coding, medical image or video coding and conventional video coding.

A. CONVERSATIONAL OR SURVEILLANCE VIDEO CODING

For conversational or surveillance videos, the background is often fixed and most attention is paid to the people or objects in the foreground. Therefore, there is no need using the same metrics to code the foreground and background. Deng *et al.* [12] propose an ROI-based bit allocation approach for HEVC to improve the subjective quality of conversational videos. In their method, the robust SURF cascade face detector is employed to extract the regions of interest (ROI) in a conversational video. According to the extracted ROI results, they develop a perceptual rate-distortion model to improve the subjective quality. Xu *et al.* [13] introduce a novel hierarchical coding method for conversational videos based on HEVC. In contrast to the previous ROI-based HEVC optimization algorithms, their method allows unequal importance to facial features by generating a pixel-wise weight map. Experimental results demonstrate that the visual quality of the face, in particular, facial features, can be enhanced by their approach. In order to reduce the coding cost of a surveillance video, Xing *et al.* [14] present a method to encode the foreground objects and the background separately. They perform a ROI extraction method following the block partition in the HEVC's quadtree structure. By subtracting the ROIs, they get the background-layer video. Their method achieves a significant total bit-rate saving and remarkable bit-rate cost reduction on ROIs. In order to improve the perceptual video quality, Goswami *et al.* [15] propose a low complexity skin tone detection technique for ROI coding in HEVC. Experimental results indicate their method improves the detection precision with low complexity, which paves the way to ROI-based HEVC optimization.

B. MEDICAL IMAGE OR VIDEO CODING

In medicine, ultrasound or endoscopic images contain large regions of black background, which are used for recording patient information and has no use for doctors' diagnosis. Thus, these regions can be coded with fewer bit rates. Sanchez and Hernández-Cabronero [16] present a novel graph-based rate control method for ROI coding in the pathology image. Their approach is designed for block-based predictive transform coding methods, by compressing the non-ROI with a lossy method while ROI with a lossless method. Bartrina-Rapesta *et al.* [17] design an ROI coding approach that is able to prioritize multiple ROIs at different priorities, ranging from lossy to lossless coding. The insight of their method is owed to the combination of RDO with a strategy allocation. Yee *et al.* [18] propose an ROI-based medical image compression method based on better portable graphics (BPG). They apply lossless BPG compression algorithm to the ROI areas, and lossy BPG for non-ROI regions. Compared with traditional image compression techniques, the compression rate is improved between 10-25%.

C. CONVENTIONAL VIDEO CODING

For conventional video, the perceptual-based optimization method is designed by enhancing the quality of important regions and decreasing the quality of unimportant regions. These methods can reduce the bit rate without influencing visual experience. Meddeb *et al.* [19] introduce a novel ROI-based rate control (RC) algorithm. They designed a tile-based rate control method and implement it to HEVC. Experimental results show that their method achieves a better representation of the ROI while respecting the global rate constraint. Zeng *et al.* [20] propose a perceptual sensitivity-based rate control algorithm for HEVC based on the human visual system observation theory. The perceptual sensitivity is measured by a mean squared error (MSE) metric. After this, a bit allocation technique is performed. More bits are allocated to those regions with higher perceptual sensitivity. Experimental results indicate that their method is able to improve the perceptual coding performance. Zhang *et al.* [21] propose a novel rate control scheme for ROI mode coding based on a discrete Fourier transform coefficient model and radial basis function neuron network. Their method performance outperforms conventional algorithms, especially for the sequence with obvious ROI details.

D. SUMMARY AND ANALYSIS

The aforementioned algorithms are performed based on the attention mechanism, that prominent texture regions might attract more attention than sparse texture regions. Undoubtebly, they can be used for special needs and improve the visual experience. However, most of them use texture or color features to analyze the perceptual importance regions of an image. They rarely use deep learning methods to extract the saliency map, which is more consistent with human perception system. Additionally, few of them make a comprehensive performance evaluation for both conversational and conventional videos. In this paper, we propose a perceptual-based HEVC optimization method based on DCN to improve the perceptual performance.

III. PROPOSED PERCEPTUAL-BASED CODING OPTIMIZATION METHOD

The purpose of the perceptual-based video coding method is to improve the coding quality of the salient regions of a video. The key is to extract the perceptual maps of the input video according to human visual characteristics, and effectively use these perceptual maps to guide the whole video coding. In order to improve the coding quality of salient regions, a rate-distortion optimization scheme guided by saliency map features is designed. Therefore, the bit rates can be adaptively allocated based on the perceptual saliency map of each frame. The proposed perceptual-based HEVC optimization algorithm mainly consists of two techniques: saliency map extraction using DCN and the bit rate allocation approach. The two techniques will be discussed in detail as follows.

A. EXTRACT SALIENCY MAP WITH DCN

Deep networks have been successfully applied to a variety of applications, such as image classification [22], object detection [23], and semantic segmentation [24]. In the paper, our purpose is to train a DCN to detect the salient regions of a frame [25]. Unlike the traditional object detection model, our model merely needs to precisely segment objects boundaries, while it is critical to approximately identify and locate multiple objects of a frame. In a normal DCN model, a set of 3D feature maps are learned to recognize an individual class [26]. For example, given an $n \times n$ image, the parameters of the layer, l, need to be calculated as follows:

$$\sum_{l \in L} d_l \times C \times \frac{n}{k^l} \times \frac{n}{k^l}, \tag{1}$$

where d_1 represents the number of features of layer l, and k is the max pooling stride size. C donates the number of the class. For such networks, learning a model with this many parameters would need a large amount of computation. It is very important to ensure real-time in video compression. With the aim of reducing the complexity of the algorithm, some optimization strategies are explored.

Most CNN models are designed for classification or recognition tasks. For example, ImageNet is designed for recognition of various animals, flowers, and other objects. In our model, there is no need to distinguish every category of every species. By folding similar sets of classes into a more general class, we dramatically reduce the number of classes. It is obvious that most images contain only a few classes, so building a separate feature map for each class is computationally inefficient. As long as objects of these combined classes have a similar structure, they will be classified in the same general category. The map produced will be almost identical. Moreover, many classes have similar lower-level characteristics, even when the number of classes is relatively small. Therefore, to reduce the number of parameters, the parameters are shared across the feature maps for different classes.

The architecture of the saliency map extraction network is presented in Fig. 2. Our model is designed on the basis of classic VGG network [27], that largely consists of convolution layers. Just before the final output layer, we perform global average pooling on the convolution feature maps instead of softmax in the case of categorization. Besides this, to improve the detection performance of all objects, we use the sigmoid active function.

Let Z_l^c represent the total sum of the activations of layer l for all feature maps for a given class c. $f_k(x, y)$ donates the activation of unit k in the last convolutional layer at spatial location (x, y). Thus, Z_l^c can be obtained by the following formula:

$$Z_l^c = \sum_k \omega_k^c \sum_{x,y} f_k(x, y) = \sum_k \sum_{x,y} \omega_k^c f_k(x, y). \tag{2}$$

We define $S_c(x, y)$ to indicate the importance of the activation at the spatial grid leading to the classification of an image to class c. As a result, we can obtain the multi-structure

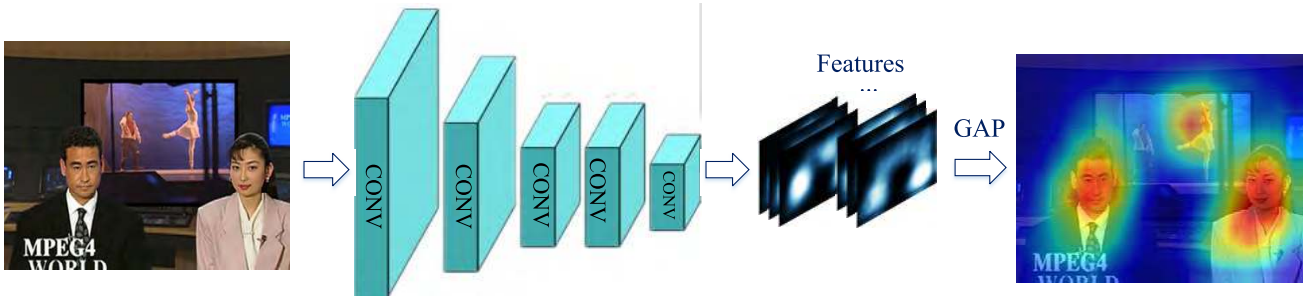


FIGURE 2. The DCN structure for saliency map extraction.

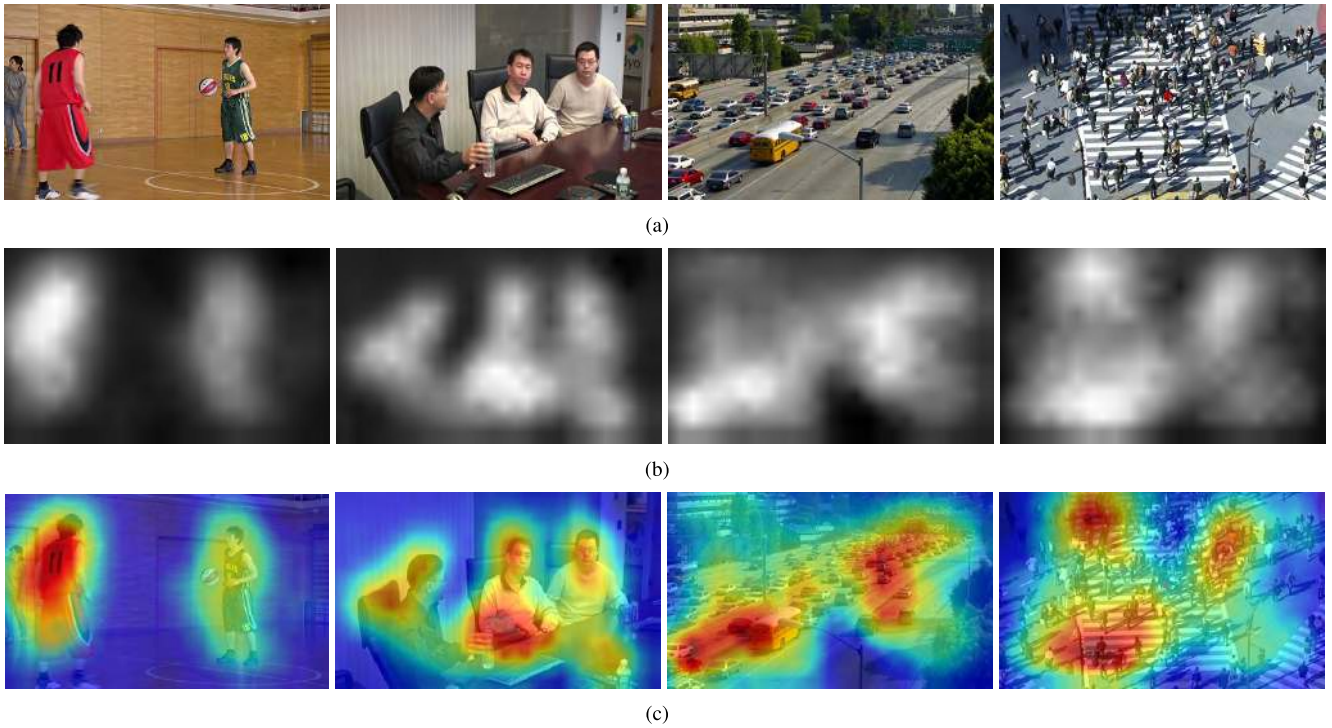


FIGURE 3. Experiment result of the proposed method for four video sequences. (a) The four different test frames (BaketballPass, Vidyo1, Traffic, PeopleOnStreet), (b) the saliency map of each frame, and (c) the heat map of each frame.

saliency map.

$$S_c(x, y) = \sum_k \omega_k^c f_k(x, y). \tag{3}$$

As shown in Figure 2, the global average pool outputs the average value of each unit's characteristic graph in the last convolution layer. These values are weighted and are used to generate the final output. We compute the weighted sum of the feature maps for the final convolution layer to obtain the saliency map.

The model is trained with the Caltech data set consisting of 256 classes of man-made and natural objects, common plants and animals, buildings, etc. We tested the CNN model on the video test sequences recommended by JCT-VC. Figure 3 illustrates four different frames, along with the corresponding saliency maps and heat maps generated by our method. It can be seen that the proposed method can deal with multiple types of video from simple to more complex ones. The salient content of the image can be accurately captured whether it includes the human body, animal or objects. When encoding, it is important that we should not degrade the quality of human bodies or other main objects. Moreover, the salient region is arbitrary shape and gradually-changed, which will avoid the square effect when incorporating it into HEVC.

B. INTERGRATING THE SALIENCY MAP WITH HEVC

The efficient coding performance of HEVC comes from its more flexible division of the coding unit and higher precision angle prediction mode. In order to evaluate the compression efficiency of each candidate configuration, the RDO technique is utilized. The rate-distortion cost is expressed as the following equation:

$$RD_{cost} = SSE + \lambda \times Bit, \tag{4}$$

TABLE 1. Performance of the proposed algorithm versus the original algorithm HM16.7.

Sequence		PSRN difference for each region Δ PSNR (dB) whole	salient	others	Bit rate Δ BR%	Coding Time Δ T%
Class A [2560 × 1600]	PeopleOnStreet	-0.24	0.47	-0.37	-4.40	2.65
	Traffic	-0.13	0.24	-0.25	-3.38	2.22
Class B [1920 × 1024]	BasketballDrive	-0.14	0.07	-0.15	-3.41	5.22
	Cactus	-0.19	0.10	-0.20	-0.74	4.49
	ParkScene	-0.17	0.23	-0.20	-3.69	4.64
Class C [1280 × 704]	FourPeople	-0.25	0.47	-0.30	-3.55	5.16
	Johny	-0.24	0.28	-0.28	-4.29	4.22
	SlideEditing	-0.23	0.35	-0.35	-1.36	5.86
	Vidyo1	-0.18	0.24	-0.26	-2.47	5.75
	Vidyo3	-0.26	0.18	-0.30	-3.80	5.83
Class D [832 × 448]	BasketballDrill	-0.31	0.56	-0.34	-7.02	5.88
	BQMall	-0.26	0.37	-0.29	-4.38	5.26
	PartyScence	-0.35	0.27	-0.38	-5.03	4.59
	RaceHorses	-0.28	0.14	-0.30	-4.89	5.59
Class E [384 × 192]	BasketballPass	-0.12	0.63	-0.24	-0.94	12.19
	BlowingBubbles	-0.36	0.34	-0.39	-2.48	10.87
	BQsquare	-0.28	0.78	-0.32	-2.42	10.20
	RaceHorses	-0.22	0.47	-0.26	-1.67	10.64
Class F Conversational Video [320 × 256]	Akiyo	-0.12	0.39	-0.45	-1.09	10.86
	Foreman	-0.18	0.83	-0.12	-5.69	12.02
	Mother-daughter	-0.21	1.15	-0.30	-1.22	12.36
	News	-0.12	1.64	-0.32	-0.16	11.44
	Silent	-0.32	0.59	-0.42	-8.26	10.61
	Stefan	-0.22	0.21	-0.26	-0.10	10.32
Average		**-0.22**	**0.46**	**-0.29**	**-3.02**	**7.45**

where SSE represents the sum of the squared residual, and Bit denotes the coding bit rate. λ is the Lagrangian multiplier, which acts as a weighting factor between distortion and bits:

$$\lambda = \alpha \times 2^{(\frac{QP-12}{3})}, \tag{5}$$

where α is a constant defined according to experiments, QP represents the quantization parameter. This ''try all and select the best'' philosophy can maximize the coding efficiency; however, the RDO method ignores the perceptual characteristics of the video content. It can be seen that λ plays a very important role in the optimization of the coding performance. A lager λ will result in a lower bit rate and higher distortion, and vice versa [28]. Unfortunately, λ is only a function of QP, which hardly pays attention to human visual system (HVS) perception.

Therefore, we incorporated the saliency map into the RDO process to guide the adjustment of the Lagrangian multiplier. In the previous section, we used deep convolution neural networks to obtain the saliency map, which is an eight-bit grayscale image. The larger value region indicates the current region tends to have important information, and vice versa. Corresponding to the HEVC division method, the saliency map is also divided into 64 × 64 coding tree units (CTU). The sum of all pixels in each CTU of the saliency map is calculated, represented by $Saliency_{CTU}$. Hence, the larger $Saliency_{CTU}$ should be encoded with high quality by allocating more bits. The content-based RDO is modified as follows:

$$RD_{cost} = SSE + k \times \lambda \times Bit \tag{6}$$

$$k = m - (m - n) \times \frac{Saliency_{CTU} - Saliency_{min}}{Saliency_{max} - Saliency_{min}}, \tag{7}$$

where $Saliency_{CTU}$ represents the importance value of the current CTU, while $Saliency_{max}$ and $Saliency_{min}$ denotes the respective maximum and minimum importance value of the saliency map of the current frame, respectively. As can be observed, the value of k belongs to the closed interval [m,n]. m and n are empirically determined, relating to the bit rate and distortion. m determines the coding quality and bit rate of non-salient regions. The larger the value is the fewer bit rates the non-salient regions is allocated to encode. On the contrary, n is related to the coding performance of salient regions. The smaller the value of n, the higher the image quality of salient regions. However, a small value of n impose a burden on the bit rate.

By inducing the parameter k, the bits for each CTU can be adaptively adjusted according to video content. Especially, the CTU located at the important region will be assigned a smaller Lagrangian multiplier, so that it will allocate more bits to encode. On the contrary, the CTU located at the unimportant region will be assigned with the larger Lagrangian multiplier, as it can tolerate a larger amount of distortions.

For conventional videos, our aim is to improve the coding quality of salient regions. Besides, the quality of non-salient regions should not decrease too much and the bit rate of the

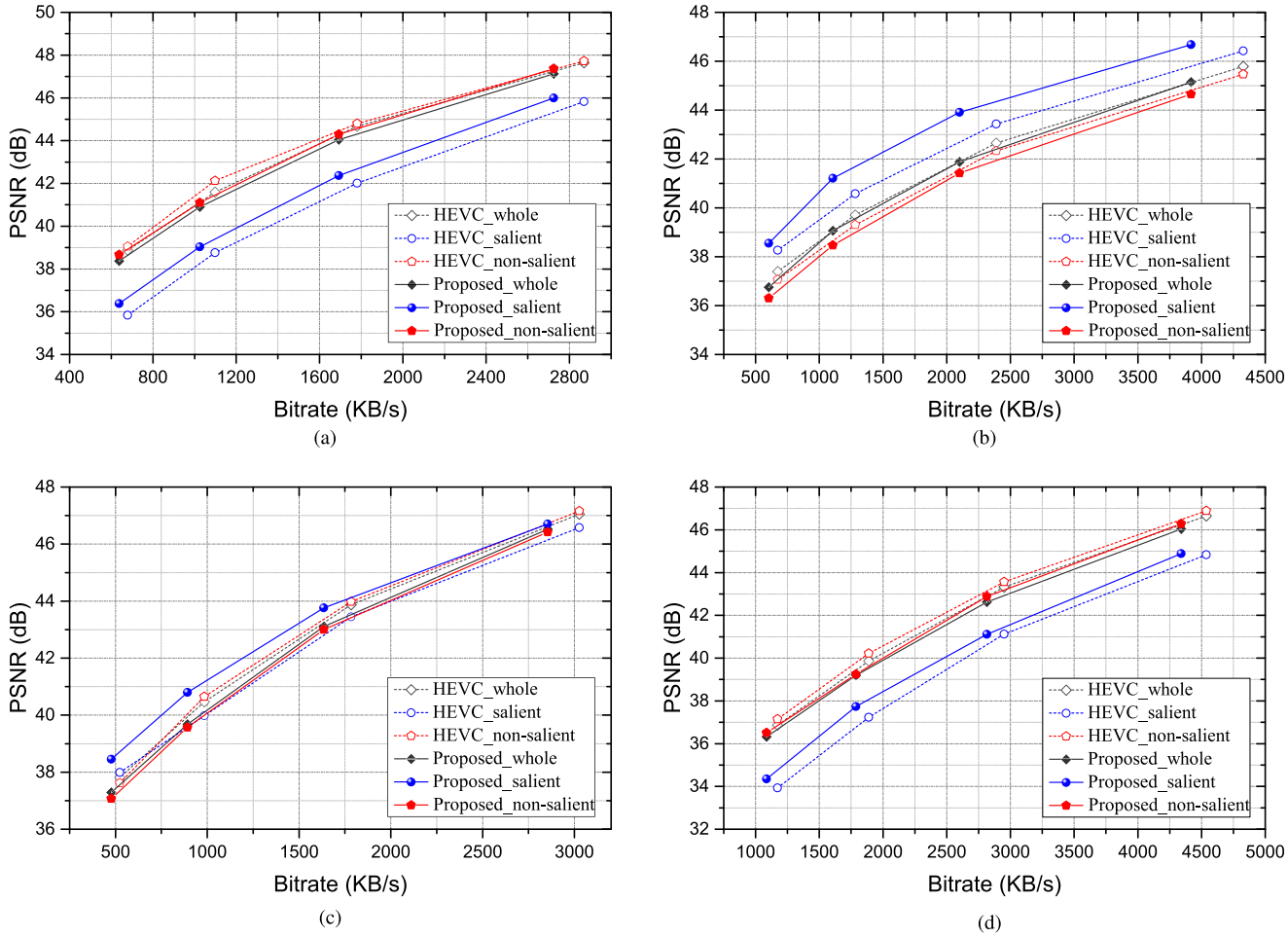


FIGURE 4. RD curves of four test video sequences: (a) Akiyo, (b) Foreman, (c) Mother-daughter, (d) News.

coding should not increase. With that aim, m and n are empirically determined as 2 and 0.5. For conversational videos, we define a larger m to save the bandwidth and improve the visual experience simultaneously.

IV. EXPERIMENTAL RESULTS

A. EXPERIMENTAL CONDITIONS AND EVALUATION METRICS

In order to evaluate the performance, the proposed method is implemented on the recent HEVC reference software (HM16.7). The experiments are run on the Intel Core i5-6300HQ CPU @2.30 GHz with a 4-GB memory. Twenty-four test sequences from class A to class F, recommended by JCT-VC [29], are used to carry out the experiments. Class A to class E belong to conventional videos, while class F mainly consists of conversational videos. The experiments are configured under an all-Intra configuration. For each sequence, 100 frames are tested with different quantization parameters (QPs): 22, 27, 32, and 37.

The unmodified HM16.7 encoder is used as the anchor. The coding performance of the proposed method is measured in terms of bit rate, coding time and PSNR [30]. The PSNR difference for the whole image, salient regions, and non-salient regions are calculated between the proposed method and the standard HEVC algorithm, respectively. Additionally, to assess the bit rate and coding time performance, we consider the calculation of ΔBR and ΔT. These metrics are defined as follows:

$$\Delta PSNR = PSNR_{proposed} - PSNR_{HM16.7} \tag{8}$$

$$\Delta BR = \frac{BitRate_{proposed} - BitRate_{HM16.7}}{BitRate_{HM16.7}} \times 100\% \tag{9}$$

$$\Delta T = \frac{T_{proposed} - T_{HM16.7}}{T_{HM16.7}} \times 100\%, \tag{10}$$

where $PSNR_{proposed}$, $BitRate_{proposed}$ and $T_{proposed}$ represent the PSNR, bit rate, and encoding time of the proposed algorithm, respectively. $PSNR_{HM16.7}$, $BitRate_{HM16.7}$ and $T_{HM16.7}$ depict the PSNR, bit rate, and encoding time in HM16.7, respectively. ΔBR represents bit rate increase, and ΔT denote the total encoding time changing.

B. EXPERIMENTAL RESULT

Table 1 depicts the performance of the proposed perceptual-based HEVC optimization algorithm compared

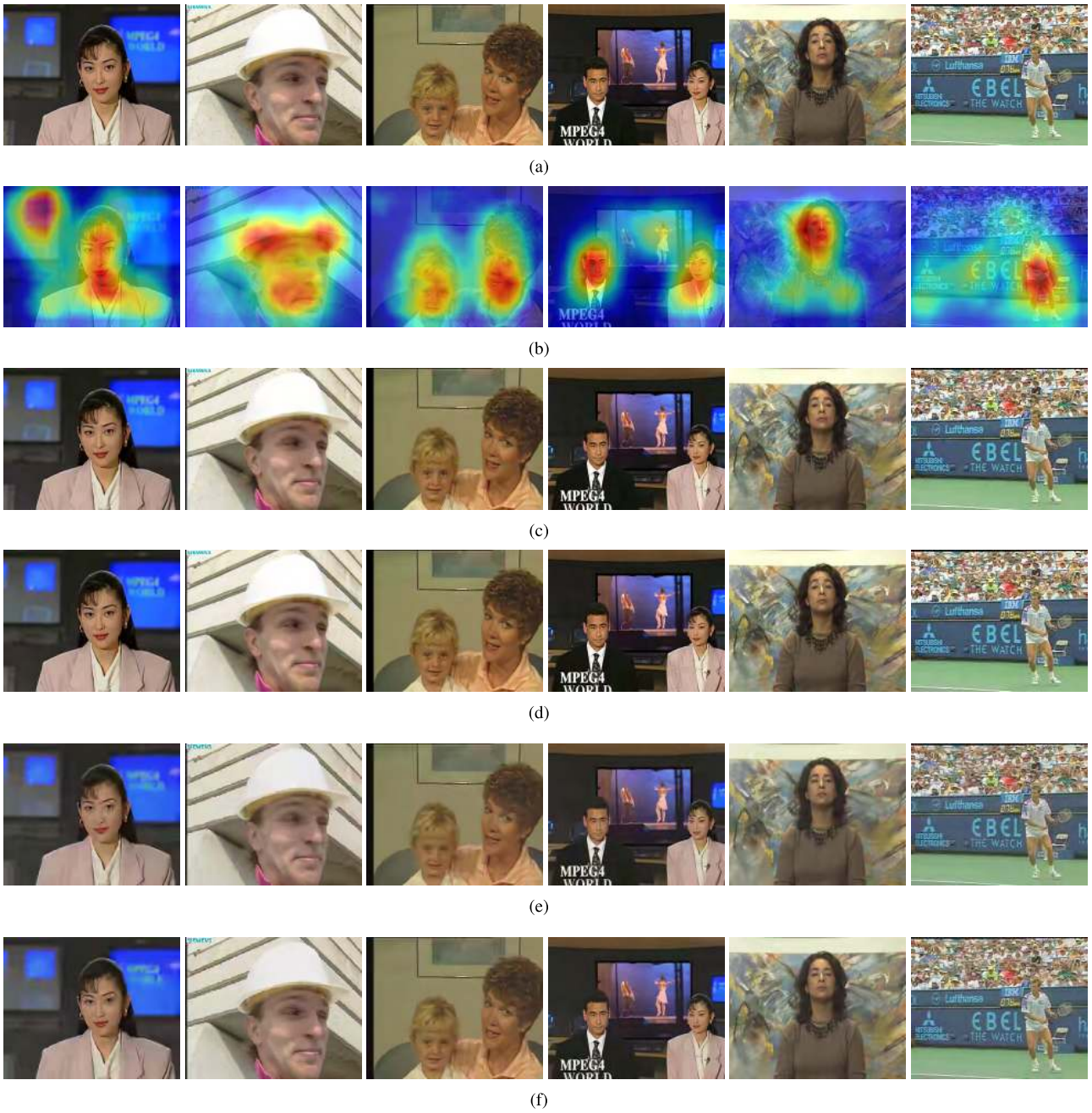


FIGURE 5. Experimental results of the proposed method compared with HEVC for six video sequences: (a) original video sequence (Akiyo, Foreman, Mother-daughter, News, Silent, Stefan), (b) heat map, (c) HEVC coding results with QP = 32, (d) proposed method coding result with QP = 32, m = 6 and n = 0.5, (e) HEVC coding results with QP = 37 and (f) proposed method coding result with QP = 37, m = 4 and n = 0.

with the standard HM16.7 under the same setting. The value for m and n is set to 2 and 0.5, respectively. The QP is set to 32. From the experiment results, we can see that the proposed algorithm achieves average 0.46 dB PSNR improvement in salient regions, 0.29 dB PSNR reduction in non-salient regions and 0.22 dB PSNR reduction of the whole video. Speaking generally, we pay more attention to the salient regions, while the non-salient regions hardly catch our eyes so much. The PSNR reduction in non-salient regions has little effect on the visual experience. Additionally, the bit rate is dropped by 3.02% and coding time is increased by 7.45% on average. In our experiment, we make use of reduced-dimension processing to the high-resolution image and projects it to low-dimensional space to extract the importance map using DCN. Therefore, for high-resolution video, there is only a small increase in coding time. For low-resolution video, however, as most of the time is spent on saliency map extraction, the coding time increases.

TABLE 2. Performance of the proposed algorithm under different conditions of parameters.

Sequence	Parameter (m=3, n=0)				Parameter (m=3, n=0.5)				Parameter (m=6, n=0.5)			
	Δ PSNR (dB)			Δ BR%	Δ PSNR (dB)			Δ BR%	Δ PSNR (dB)			Δ BR%
	whole	salient	others		whole	salient	others		whole	salient	others	
Akiyo	-0.12	1.07	-0.48	7.10	-0.34	0.32	-0.68	-3.64	-0.63	0.27	-1.04	-6.65
Foreman	-0.31	1.94	-0.27	-1.48	-0.37	0.81	-0.32	-10.63	-0.65	0.69	-0.6	-13.80
Mother	-0.26	1.77	-0.36	5.99	-0.44	1.17	-0.54	-5.79	-0.79	0.82	0.88	-9.35
News	-0.08	1.82	-0.30	7.52	-0.29	1.50	-0.51	-2.70	-0.68	0.51	-0.87	-5.29
Silent	-0.32	0.88	-0.48	7.32	-0.51	0.28	-0.49	-13.24	-0.73	0.22	-0.72	-14.10
Stefan	-0.19	0.30	-0.25	4.70	-0.48	0.18	-0.31	-0.96	-0.44	0.04	-0.48	-5.83
Average	**-0.21**	**1.30**	**-0.36**	**5.19**	**-0.41**	**0.71**	**-0.48**	**-6.16**	**-0.65**	**0.43**	**-0.77**	**-8.65**

Experimental results demonstrate that the proposed algorithm is indeed able to improve the visual quality in salient regions and save the bit rate, with nearly the same quality of the whole video.

For conversational videos, such as class F, the background is single and usually contains few information. In this case, we can define a larger m. The bandwidth can be saved by dropping the coding quality of the background. However, this method will not reduce the visual experience, as people pay less attention to the background. Table 2 depicts the performance of the proposed method compared with standard HM16.7 for class F when changing m and n. It can be seen the parameter n mainly affects the image quality of salient regions. The smaller the value of n is, the higher the obtained quality of the salient regions. On the other hand, the higher m is, the lower the bitrate needed for coding the video. However, a bigger value of the m will drop the image quality of the non-significant areas. It can be concluded that with the configuration of $m = 6$ and $n = 0.5$, the bit rate can be saved by 8.65% on average, while still maintaining the high quality of the salient regions.

To evaluate the coding performance intuitively, Figure 4 shows the rate-distortion (RD) curves of four test sequences in class F compared with HM16.7. The parameter is configured with $m = 6$ and $n = 0.5$. It can be observed that the RD curves of the whole regions (blacke solid and dotted lines) of each video sequence almost overlap together. This means that the proposed method nearly obtained the same overall coding quality for each frame. However, using the proposed method, we get a higher PSNR value at the salient regions compared with HM16.7. The improvement is obtained at the expense of reducing the coding quality of non-salient regions. Experiment results indicate that the proposed approach achieves almost the same coding quality on the whole image from a low to high bit rate compared with HM 16.7. Meanwhile, it significantly improves the quality of the salient regions of the reconstruction videos and enhances our viewing experience.

To further evaluate the performance subjectively, Figure 5 shows the reconstructed frame of the six conversational video sequences encoding by HM16.7 and the proposed algorithm. As shown, under the condition of high bit rate with QP = 32, it is hard for us to distinguish between the videos encoded by the standard HM16.7 method and the proposed method. However, the proposed perceptual-based method saves 8.65% bitrate under the condition of $m = 6$ and $n = 0.5$. On the other hand, under the low bit rates with QP = 37, the parameter is configured with $m = 4$ and $n = 0$. At this condition, our method needs nearly the same bit rates as the standard HEVC algorithm. Remarkably, our method obtains clearer facial details (e.g., the nose for Akiyo sequence, and the eyes for foreman and Mother-daughter sequences).

V. CONCLUSION AND DISSCUSION

In this work, to improve the perceptual coding efficiency, we developed a saliency-map-guided RDO technique for HEVC. The technique is performed by adaptively adjusting the Lagrangian multiplier. As a result, there is a new balance between the bit rate and distortion. More specifically, we train a DCN model to identify multiple semantic regions at any scale and generate a saliency map. This provides sufficient information to perform the intelligent bit rate allocation method. To the best of our knowledge, our method is the first to use deep convolution networks in HEVC to improve the quality of salient regions of a video. This technology can specially be used for conversational or surveillance videos, which can largely reduce the bandwidth without degrading the visual experience. Future studies will concentrate on perceptual-based medical image or video coding optimization.

ACKNOWLEDGMENTS

The authors would like to thank the editors and anonymous reviewers for their valuable comments. Further thanks to Aaditya Prakash for sharing his method.

REFERENCES

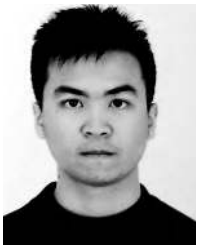
[1] T. Wiegand *et al.*, "Special section on the joint call for proposals on high efficiency video coding (HEVC) standardization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 12, pp. 1661–1666, Dec. 2010.

[2] W.-J. Han *et al.*, "Improved video compression efficiency through flexible unit representation and corresponding extension of coding tools," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 12, pp. 1709–1720, Dec. 2010.

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

[4] I.-K. Kim, J. Min, T. Lee, W.-J. Han, and J. Park, "Block partitioning structure in the HEVC standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1697–1706, Dec. 2012.
[5] J. Vanne, M. Viitanen, T. D. Hamalainen, and A. Hallapuro, "Comparative rate-distortion-complexity analysis of HEVC and AVC video codecs," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1885–1898, Dec. 2012.
[6] J.-R. Ohm and G. J. Sullivan, "High efficiency video coding: The next frontier in video compression [Standards in a Nutshell]," *IEEE Signal Process. Mag.*, vol. 30, no. 1, pp. 152–158, Jan. 2013.
[7] J. Lainema, F. Bossen, W.-J. Han, J. Min, and K. Ugur, "Intra coding of the HEVC standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1792–1801, Dec. 2012.
[8] A. Ortega and K. Ramchandran, "Rate-distortion methods for image and video compression," *IEEE Signal Process. Mag.*, vol. 15, no. 6, pp. 23–50, Nov. 1998.
[9] Y. Wu, P. Liu, Y. Gao,and K. Jia, "Medical ultrasound video coding with H.265/HEVC based on ROI extraction," *PLoS One*, vol. 11, no. 11, 2016, Art. no. e0165698.
[10] A. Yang, H. Zeng, J. Chen, J. Zhu, and C. Cai, "Perceptual feature guided rate distortion optimization for high efficiency video coding," *Multidimensional Syst. Signal Process.*, vol. 28, no. 4, pp. 1249–1266, 2017.
[11] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2921–2929.
[12] X. Deng, M. Xu, and Z. Wang, "A ROI-based bit allocation scheme for HEVC towards perceptual conversational video coding," in *Proc. 6th Int. Conf. Adv. Comput. Intell. (ICACI)*, Oct. 2013, pp. 206–211.
[13] M. Xu, X. Deng, S. Li, and Z. Wang, "Region-of-interest based conversational HEVC coding with hierarchical perception model of face," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 3, pp. 475–489, Jun. 2014.
[14] P. Xing, Y. Tian, T. Huang, and W. Gao, "Surveillance video coding with quadtree partition based ROI extraction," in *Proc. Picture Coding Symp. (PCS)*, Dec. 2013, pp. 157–160.
[15] P. Goswami, P. V. Srikanth, and J. Rahiman, "Low complexity in-loop skin tone detection for ROI coding in the HEVC encoder," in *Proc. 22nd Nat. Conf. Commun. (NCC)*, Mar. 2016, pp. 1–6.
[16] V. Sanchez and M. Hernández-Cabronero, "Graph-based rate control in pathology imaging with lossless region of interest coding," *IEEE Trans. Med. Imag.*, vol. 37, no. 10, pp. 2211–2223, Oct. 2018.
[17] J. Bartrina-Rapesta, J. Serra-Sagristà, and F. Aulí-Llinàs, "JPEG2000 ROI coding through component priority for digital mammography," *Comput. Vis. Image Understand.*, vol. 115, no. 1, pp. 59–68, Jan. 2011.
[18] D. Yee, S. Soltaninejad, D. Hazarika, G. Mbuyi, R. Barnwal, and A. Basu, "Medical image compression based on region of interest using better portable graphics (BPG)," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Oct. 2017, pp. 216–221.
[19] M. Meddeb, M. Cagnazzo, and B. Pesquet-Popescu, "ROI-based rate control using tiles for an HEVC encoded video stream over a lossy network," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2015, pp. 1389–1393.
[20] H. Zeng, A. Yang, K. N. Ngan, and M. H. Wang, "Perceptual sensitivity-based rate control method for High Efficiency Video Coding," *Multimedia Tools Appl.*, vol. 75, pp. 10383–10396, Oct. 2015.
[21] Z. Zhang, T. Jing, J. Han, Y. Xu and F. Zhang, "A new rate control scheme for video coding based on region of interest," *IEEE Access*, vol. 5, pp. 13677–13688, 2017.
[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2012, pp. 1097–1105.
[23] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.
[24] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Apr. 2017.
[25] A. Prakash, N. Moran, S. Garber, A. Dilillo, and J. Storer, "Semantic perceptual image compression using deep convolution networks," in *Proc. Data Compress. Conf. (DCC)*, Apr. 2017, pp. 250–259.
[26] M. Li, W. Zuo, S. Gu, D. Zhao, and D. Zhang, "Learning convolutional networks for content-weighted image compression," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3214–3223.
[27] K. Simonyan and A. Zisserman. (2014). "Very deep convolutional networks for large-scale image recognition." [Online]. Available: https://arxiv.org/abs/1409.1556
[28] K. Rouis, M.-C. Larabi, and J. B. Tahar, "Perceptually adaptive Lagrangian multiplier for HEVC guided rate-distortion optimization," *IEEE Access*, vol. 6, pp. 33589–33603, 2018.
[29] F. Bossen, *Common Test Conditions and Software Reference Configurations*, document JCTVC-L1100, Jan. 2013.
[30] G. Bjontegaard, "Calculation of average PSNR difference between RD-curves," Tech. Rep., 2001.

XUEBIN SUN received the bachelor's degree from the Tianjin University of Technology, in 2011, and the master's and Ph.D. degrees from Tianjin University, in 2014 and 2018, respectively. He is currently a Research Associate with the Department of Electronic and Computer Engineering, Robotics Institute, The Hong Kong University of Science and Technology (HKUST), Hong Kong. His research interests include video coding optimization, digital image processing, deep learning, and point cloud processing algorithms.



HAN MA is currently pursuing the degree with the Department of Precision Instrument, Tsinghua University, Beijing, China. His research interests include image processing and deep learning.

WEIXUN ZUO received the B.E. degree in optical information science and technology from Anhui University, in 2016, and the M.Sc. degree in electronic engineering from The Hong Kong University of Science and Technology, where he is currently a Research Assistant with the Robotics Institute. His research interests include deep reinforcement learning, SLAM, and mobile robot.

MING LIU received the B.A. degree in automation from Tongji University, in 2005, and the Ph.D. degree from the Department of Mechanical and Process Engineering, ETH Zürich, in 2013, supervised by Prof. R. Siegwart. He is currently with the ECE Department, the CSE Department, and the Robotics Institute, The Hong Kong University of Science and Technology. His research interests include dynamic environment modeling, deep-learning for robotics, 3D mapping, machine learning, and visual control. He received twice the Innovation Contest Chunhui Cup Winning Award, in 2012 and 2013. He received the Wu Weijun AI Award, in 2016. He was the Program Chair of the IEEE-RCAR 2016, the Program Chair of the International Robotics Conference in Foshan, in 2017, and the Conference Chair of ICVS 2017.