Deep-Learned Perceptual Quality Control for Intelligent Video Communication

Xuebin Sun, Miaohui Wang, *Member, IEEE*, Rongfu Lin, Yuxiang Sun, and Shing Shin Cheng, *Member, IEEE*

***Abstract*—With the development of video technology, a large amount of video data generated from video conferences, sports events, live broadcasts and network classes flows into our daily lives. However, ultra-high-definition video transmission is still a challenge due to the limited network bandwidth and instability, which further affects the quality of video service closely linked with consumer electronic video display. To address this challenge, we propose a deep-learned perceptual quality control approach, which can significantly improve the video quality and visual experience at the same bandwidth. The proposed scheme mainly involves saliency region extraction, perceptual-based bits allocation, and video enhancement. Firstly, we exploit a multi-scale deep convolutional network module to predict the static saliency map that semantically highlights the salient regions. Secondly, we develop a recurrent neural network model to extract the dynamic saliency regions. Finally, a three-level rate allocation scheme is developed based on the resulted saliency guidance, which is more reasonable by taking into account the visual characteristics of human eyes. Experimental results on a large dataset show that our method achieves an average gain of 1.5dB on the salient regions without introducing an extra bandwidth burden, which significantly improves the visual experience and paves the way to intelligent video communication.**

***Index Terms*—HEVC, perceptual quality control, static saliency, dynamic saliency.**

I. Introduction

TO REDUCE the impact of the COVID-19 epidemic, video conferences and online classrooms have become ubiquitous [1], [2], [3], due to the fact that many countries have been forced to shut down the public face-to-face communication. In addition, more and more people use short videos sharing their highlight moments via consumer electronics (Smartphone, Tablet, Laptop, *etc*). However, video communication has the characteristics of large data volume, and it is sensitive to network delay and jitter. Therefore, perceptual quality control is essential to consumer products based on video storage and transmission [4], [5], [6].

Recently, human vision has been widely investigated in visual signal processing [7]. It shows that human eyes pay unbalanced attention to different video scenes, and have different visual perception sensitivities to different image areas. Further, human eyes are selective for the brightness, texture, motion regions in videos, and the sensitivity and tolerance to different content are different. In general, for still videos (e.g., video conferencing and remote classrooms), face areas attract more attention [8], but for dynamic videos, attention is more focused on moving objects. Thus, when the salient regions have higher perceptual quality, it will significantly improve reviewers' visual experience.

Rate-distortion optimization (RDO) [9] based on an objective distortion metric has been adopted to optimize the video coding efficiency [10], [11]. However, the distortion metrics (e.g., sum of square error (SSE) or mean absolute differences (MAD)) have a low correlation with human vision. Moreover, the Lagrange operator is only related to the quantization parameter (QP), ignoring the perceptual characteristics of video content. In other video content. In other words, RDO fails to fully consider the factors of human vision when optimizing the coding efficiency. Since human eyes are the ultimate receiver for video applications, it is of great significance to consider the human visual characteristics to improve coding efficiency.

To relieve the burden of ultra-high-definition video communication among in consumer electronics, a deep learning-based perceptual quality control scheme for HEVC is proposed in this paper. Videos, such as sports events, live broadcasts, conferences, entertainment programs, *etc.*, are first encoded into the bitstream by a perceptual quality control-based encoder. When receiving the bitstream, a typical consumer device decodes it, which is further enhanced by a deep-learned module. Fig. 1 illustrates the coding results of *Host* and *Video conference* by the proposed scheme and the HEVC algorithm [12]. For the *Host* video, face regions attract more attention as illustrated in Fig. 1(a), and the profile of the connecting guests is more important. Compared with HEVC illustrated in Fig. 1(b), the proposed algorithm obtains high visual quality in the face region, especially in the eye regions,

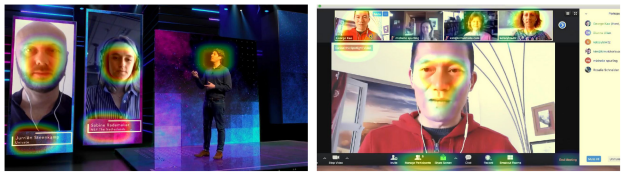
Manuscript received 2 August 2022; accepted 6 September 2022. Date of publication 14 September 2022; date of current version 24 October 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 62106152 and Grant 61701310; in part by the Natural Science Foundation of Guangdong Province under Grant 2022A1515011245 and Grant 2019A1515010961; and in part by the Natural Science Foundation of Shenzhen City under Grant 20220809160139001 and Grant 20200805200145001. *(Corresponding authors: Miaohui Wang; Shing Shin Cheng.)*

Xuebin Sun and Miaohui Wang are with the Guangdong Key Laboratory of Intelligent Information Processing, Shenzhen Institute of Artificial Intelligence and Robotics for Society, Shenzhen University, Shenzhen 518060, China (e-mail: sunxuebin@szu.edu.cn; wang.miaohui@gmail.com).

Rongfu Lin and Shing Shin Cheng are with the Department of Mechanical and Automation Engineering, The Chinese University of Hong Kong, Hong Kong, China (e-mail: rongfulin@126.com; sscheng@cuhk.edu.hk).

Yuxiang Sun is with the Department of Mechanical Engineering, The Hong Kong Polytechnic University, Hong Kong, China (e-mail: yx.sun@polyu.edu.hk; sun.yuxiang@outlook.com).

Digital Object Identifier 10.1109/TCE.2022.3206114



(a) Visual attention regions

(b) HEVC compression results

(c) Perceptual-based coding results

Fig. 1. Experimental results of the HEVC algorithm and the proposed method for *Host* and *Video conference* sequences: (a) Heat map of the gaze locations, (b) HEVC (*Host*: 44512 bits, *Video conference*: 64184 bits), and (c) Deep-learned perceptual quality control method (*Host*: 44000 bits, *Video conference*: 63392 bits). Best viewed by zooming in.

as described in Fig. 1(c). Similarly, the foreground human attracts more attention than the background wall for video conferences. By zooming in, it can be observed that the proposed scheme obtains higher coding quality in the face region. The most important thing is that the bit costs are almost the same, which means that our algorithm significantly increases the visual experience without the additional bandwidth burden, paving the way for intelligent video communication.

To realize perceptual quality control, one needs to investigate a reasonable rate allocation strategy, and simultaneously regulate the output bit-rate to maintain satisfying visual quality. In this paper, we first develop a compression-based saliency detection method, and propose a new rate regulation and quality control method according to the significance map. Furthermore, due to the fact that humans will feel uncomfortable at their boundaries where the encoding quality of salient regions is too high and the coding quality of non-salient regions is too low, we hence investigate an effective video enhancement module to address this difficulty. Our main contributions are summarized as follows.

- We have investigated a deep-learned static and dynamic salient extraction method for video compression. Compared with the traditional saliency advances, the proposed multi-scale network is more compatible with video coding on consumer electronics in considering the human visual characteristics. From the perspective of the computation efficiency, the deep models proposed in this paper is a suitable choice.
- Based on the resulted saliency guidance, a quality control method is further designed, where a new tanh-based rate allocation is developed by considering the group of picture level, frame level and coding tree unit level. To our knowledge, this is the earliest exploration to model visual attention with tanh-based rate allocation in perceptual quality control by considering visual saturation effect.
- To alleviate the mutation of coding quality at the boundary between the salient and non-salient regions, a time-varying and space-varying recurrent neural network for video enhancement is explored to further improve the compressed video quality for consumer electronic video display and transmission. In general, it is a new attempt to further improve the visual performance for deep-learned perceptual quality control, which have rarely been studied before.

Subjective and objective results demonstrate that the proposed method can significantly improve the performance of perceptual coding,[1] which can be specially used for various video applications, including video conference, sport broadcasting, and online education.

The remainder of this paper is organized as follows. In Section II, the related works are briefly reviewed. The proposed methodology is introduced in Section III. Experimental results are presented in Section IV, and the conclusion is given in Section V.

II. Related Works

The proposed scheme is specially designed for deep-learned perceptual quality control for intelligent video communication. Therefore, we mainly review some representative perceptual video coding and rate control methods in this section.

A. Perceptual Video Coding

To achieve perceptual video coding, visual attention models [22], [23], [24] (e.g., saliency detection) have been investigated to improve compressed video quality.

For dynamic videos with intense motion, a moving object is regarded as a salient region [25], [26], [27]. For surveillance videos, the foreground objects attract more attention [13], [28], [29]. In some cases, a stationary saliency model can be jointly considered with the motion vector (MV) field to calculate the dynamic saliency [30]. Based on the saliency detection, some early perceptual-based intelligent video coding algorithms have been developed in recent years. For instance, human face is usually considered as a saliency region that needs high-quality compression [14], [31]. Similarly, Li *et al.* [15] proposed to allocate different bits to different blocks according to their saliency values. Hadizadeh and Bajić [16] introduced a saliency-aware video coding technique based on the Itti-Koch-Niebur (IKN) saliency model. Zhu *et al.* [17] proposed an enhanced HEVC method by introducing a learning-based attention mechanism to extract the spatial and temporal saliency maps. To our knowledge, the above perception-based coding methods were designed for still videos, so they were unsuitable for videos with intense motion, in which the viewer's attention may be focused on moving objects. Compared with the traditional methods, a

[1]The implementation can be downloaded from https://github.com/simaniu/Perceptual-Quality-Control.

TABLE I
SUMMARY OF PERCEPTUAL-BASED CODING ALGORITHMS

Methods	Static Saliency		Dynamic Saliency		Rate	Post-
	Traditional	Deep	Traditional	Deep	Control	Processing
Xue *et al.*[13]	×	×	√	×	×	×
Xu *et al.*[14]	√	×	×	×	√	×
Li *et al.* [15]	√	×	×	×	√	×
Hadizadeh *et al.*[16]	√	×	×	×	×	×
Zhu *et al.* [17]	×	√	√	×	×	×
Zhou *et al.*[18]	×	×	×	×	√	×
Li *et al.*[19]	×	×	×	×	√	×
Chen *et al.*[20]	×	×	×	×	√	×
Wang *et al.*[21]	×	×	×	×	√	×
Proposed	×	√	×	√	√	√

deep learning-based method is more competitive in extracting salient regions for intense motion modeling, which can be more consistent with the human visual system (HVS).

B. Perceptual Rate Control

As discussed above, there are few studies on the deep-learned saliency-driven rate control for HEVC as summarized in Table I. The main reason is that a perceptual-based rate control method is more difficult than a perceptual video coding method.

Recently, several studies have been developed on perceptual rate control in video communication [18], [19], [20], [21]. For instance, Zhou *et al.* [18] proposed a Structural Similarity (SSIM)-based coding tree unit (CTU)-level rate control for HEVC, where mean square error (MSE) was replaced by SSIM in the RDO process. Li *et al.* [19] proposed a modified R-λ model to make a balance between coding quality and bandwidth burden. Chen and Pan [20] designed an optimized RDO strategy for H.265/HEVC, where the CTU-level rate allocation was formulated as a decision-making problem. To satisfy low-bit applications, Wang *et al.* [21] introduced an optimal CTU-level bit allocation for HEVC based on the motion vectors and complexity. Currently, most of the RDO-based rate control methods are designed by exploring an improved R-λ model or a new bit-rate allocation to obtain a more accurate or lower complexity rate control. However, the existing coding methods are the lack of taking human perception information into account, which still leave space for exploring perceptual-based rate control solutions.

III. PROPOSED PERCEPTUAL QUALITY CONTROL METHOD

In this paper, we address the problem of deep-learned saliency-driven perceptual quality control for HEVC. The architecture of the proposed framework is presented in Fig. 2, where the green parallelograms represent the encoder and decoder modules, and the yellow parallelograms highlight the proposed saliency module, rate control module and enhancement module. We describe them in detail as follows.

A. Static Saliency Extraction Module

A saliency map is one of the most-used methods to represent the weight distribution that attracts human visual attention.

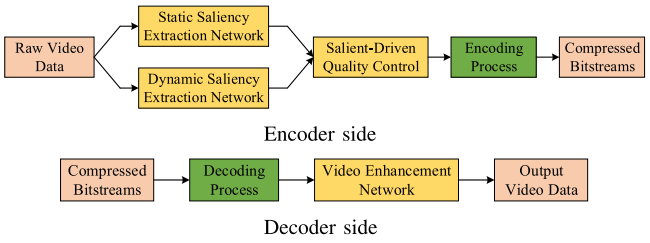


Fig. 2. Block diagram illustration of the proposed deep-learned perceptual quality control for intelligent video communication: the yellow parallelograms represent the modification that we have made, and the pink parallelograms indicate the input-output data.

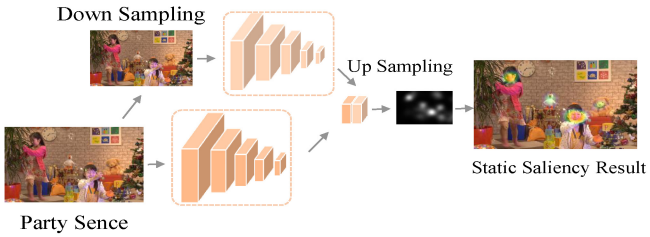


Fig. 3. The multi-scale DCN structure for the static saliency extraction.

In this section, we investigate a new static saliency method based on a multi-scale deep convolutional network (DCN). As illustrated in Fig. 3, the proposed DCN model takes an RGB image on two different scales. Let $I(x, y)$ represent the input image, and $I'(x, y)$ be half the size of the original one. Each image is processed by a feature encoder, and the latent features are fused to generate the final saliency map.

Let Y_k denote a three-dimensional array containing the responses of the neurons at the k-th layer. The size of Y_k is $m_k \times n_k \times d_k$, where $m_k \times n_k$ represents the spatial size of the receptive field, and d_k denotes the template for which the neuron is tuned. Higher-layer neural responses encoded more meaningful semantic representations than lower-layer neural responses. The encoder is designed based on the VGG-16 network [32] by removing the final pooling and fully connected layers.

Let Y_c represent the neural response of the last convolutional layer, which is useful for object detection. A 1×1 convolution layer is added after Y_c, which has only one filter to detect whether the response in Y_c belongs to a salient region or not. The output of Y_c with a saliency detection filter is represented as Y_s. Finally, to obtain a saliency map, we use a linear interpolation to upscale Y_s to match the input image size, and scale pixel values to the range of 0 and 1.

The value of the ground-truth saliency is quantized to the interval [0, 1], which can be considered as the probability distribution that an observer pays attention to each pixel. If using `Softmax` in the last layer, it will introduce a polynomial distribution to the prediction result. However, the observer may focus on multiple points, so each prediction is more reasonable as an independent one. Thus, for the last layer, we adopt a `Sigmoid` operation. In this way, the predicted outcome can be viewed as the probability of an independent random binary variable. A binary cross-entropy is adopted as the loss

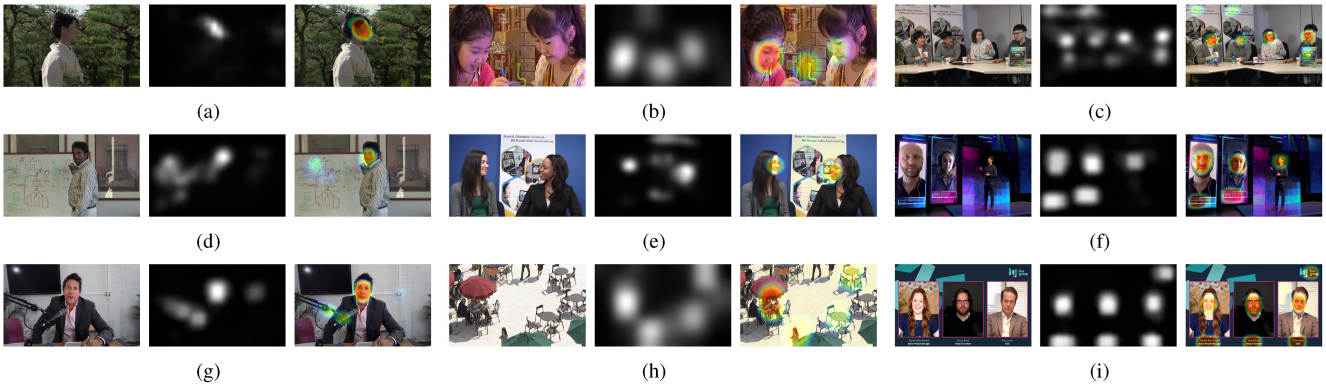


Fig. 4. Experiment results of the static saliency: (a) *Kimono*, (b) *BlowingBubbles*, (c) *FourPeople*, (d) *Vidyo3*, (e) *KristenAndSara*, (f) *Host*, (g) *Broadcaster*, (h) *BQSquare*, and (i) *VirtualConference*.

function, which is defined as follows:

$$Loss = -\frac{1}{N}\sum_{j=1}^{N}\Big(S_j log\Big(\hat{S}_j\Big) + \big(1 - S_j\big) log\Big(1 - \hat{S}_j\Big)\Big), \quad (1)$$

where S and $\hat{S}$ represent the predicted saliency map and its corresponding ground-truth, respectively. We train and evaluate the proposed saliency model on the OSIE dataset [33]. The Adam optimization is used to train our model with a learning rate of 0.0001.

The proposed multi-scale DCN model is tested on the video dataset recommended by JCT-VC [34]. Fig. 4(i) illustrates the saliency results of nine different video sequences. As seen, static salient regions can be accurately extracted, where the salient area gradually changes with an arbitrary shape. This will alleviate the blocking effect in the bit-rate allocation.

B. Dynamic Saliency Extraction Module

In certain video scenes, moving objects attract more attention [35]. If these objects can be encoded with high quality, the visual experience can be significantly improved. In this section, a moving object extraction network model is designed based on this observation.

As illustrated in Fig. 5, the proposed model learns long-term spatial-temporal features directly from the training data in an end-to-end fashion. The offline trained model can accurately propagate the initial motion by memorizing and updating the target features, including appearance, position, and scale, which can automatically propagate the temporal motion in a whole video. Since long short-term memory (LSTM) can learn long-term correlation, it is adopted in modeling the dynamic saliency for video data. In the segmentation part, we use convLSTM to track the characteristics of spatial-temporal data at different scales to improve the performance on small objects.

Firstly, the convolutional encoder processes a frame x_t to extract the feature graph $\widetilde{x_t}$, which is defined as follows:

$$\widetilde{x_t} = Encoder(x_t). \quad (2)$$

Then $\widetilde{x_t}$ is sent as the input of convLSTM. The internal states c_t and h_t are automatically updated when new observations $\widetilde{x_t}$ are given, which capture new features from a salient object.

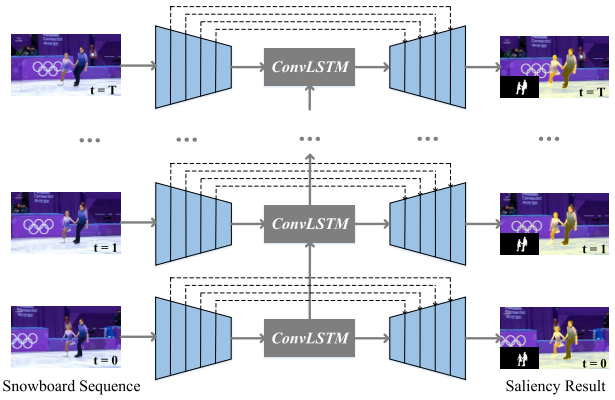


Fig. 5. Network structure for the dynamic saliency extraction.

c_t and h_t is calculated by the following formula:

$$c_t, h_t = ConvLSTM(\widetilde{x_t}, c_{t-1}, h_{t-1}). \quad (3)$$

The output h_t is passed to the decoder to obtain a full resolution segmentation result $\widetilde{y_t}$, which is calculated by Eq. (4). We also add a skip connection between the encoder and the decoder, which allows shallow convolution features F_x can be introduced.

$$\widetilde{y_t} = Decoder(h_t, F_x). \quad (4)$$

Combining shallow-layer and deep-layer features is more beneficial to generate segmentation masks. In the training process, the binary cross-entropy loss is calculated between $\widetilde{y_t}$ and y_t. Back propagation is used to train the parameters of the encoder, decoder, and convLSTM modules.

We use the cross-entropy loss and the logarithm of the *Jaccard* index as the loss functions to train the segmentation network. Experimental results are shown in Fig. 6. As seen, the proposed method can accurately extract dynamic objects. As human eyes pay more attention to moving objects than background regions, these salient regions should be allocated more bits and encoded with higher quality.

C. Perceptual Quality-Driven Quality Control Approach

Video communication is one of the most challenging research areas in consumer electronics, such as smartphones,

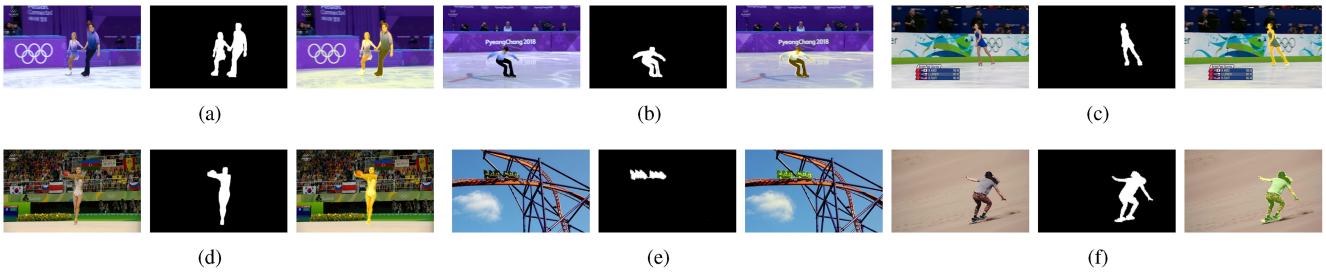


Fig. 6. Experimental results of the dynamic salient region extraction: (a) *Pair figure-skating*, (b) *Men's figure-skating*, (c) *Women's figure-skating*, (d) *Artistic gymnastics* (e) *Rollercoaster*, and (f) *Snowboard*.

tablets, laptops, *etc*. The current video coding systems rarely take into account the human visual system (HVS), where people may pay more attention to specific areas and moving objects. There are two main challenges to achieve perceptual quality control: one is how to efficiently predict the salient region for human eyes (see Secs. III-A and III-B), and the other is how to allocate the bit-rate to these salient regions and perform the quality control. In this section, we will provide the detailed descriptions of the proposed rate control approach.

In video coding, the traversal of every possible combination of coding unit (CU), prediction unit (PU), and transform unit (TU) is determined by a so-called rate-distortion optimization (RDO), which is the fundamental of a modern video codec. RDO can guarantee that a codec can find the optimal solution based on two basic facts: 1) the relationship between rate (R) and distortion (D) is a convex function, and 2) the number of the coding parameter space is discrete and countable. A detailed mathematical derivation and analysis for RDO is referred to [36].

The perceptual quality-driven quality control algorithm aims to minimize distortion while maintaining the target bit rate, which can be expressed by minimizing the distortion D according to the number of bits R used for the target R_t:

$$\{Para\}_{opt} = \arg\min_{\{Para\}} D \quad s.t. \quad R \le R_t. \tag{5}$$

The relationship between D, R, and $Para$ have been well defined and introduced in HEVC [37]. Simply put, a hyperbolic R-D model has been witnessed in HEVC [11], which is a convex function. In other words, the optimal solution of R-D can be theoretically guaranteed.

The proposed perceptual bit allocation is designed for three levels: a GOP layer, a picture layer, and a CU layer. A new picture-level salient weighting S_{Pic} is designed by

$$S_{Pic} = \frac{\sum_{\{i=1,\ j=1\}}^{\{i\le width,\ j\le height\}} S(i,j)}{width \times height}. \tag{6}$$

Thus, the number of target bits of the current picture $Target_{Pic}$ is computed by

$$Target_{Pic} = \frac{Target_{GOP} - Coded_{GOP}}{\sum_{UncodePics} S_{Pic}} \times S_{PicCur}, \tag{7}$$

where $Target_{GOP}$ represents the bit budget of the current GOP, and $Coded_{GOP}$ denotes the number of bits encoded in the current GOP. S_{PicCur} and S_{Pic} represent the saliency map of the current frame and that of the uncoded frames, respectively.

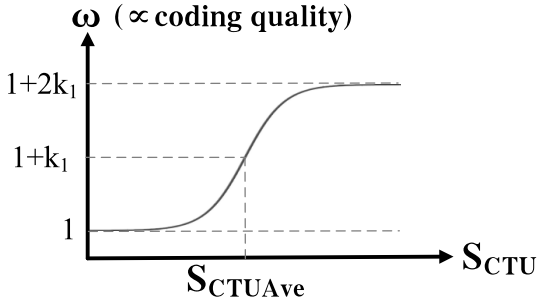


Fig. 7. The relation curve between S_{CTU} and ω.

The CTU-level saliency is indicated by ω_i, which represents the bit weight of the current CTU. ω_i is calculated by a modified *tanh* function

$$\begin{aligned}\omega_i &= 1 + k_1\left(tanh\left(\frac{k_2(S_{CTU} - S_{CTUAve})}{S_{CTUAve}}\right) + 1\right) \\ &= 1 + k_1\left(\frac{1 - exp\left(\frac{-2k_2(S_{CTU} - S_{CTUAve})}{S_{CTUAve}}\right)}{1 + exp\left(\frac{-2k_2(S_{CTU} - S_{CTUAve})}{S_{CTUAve}}\right)} + 1\right),\end{aligned} \tag{8}$$

where S_{CTU} and S_{CTUAve} represent the saliency value for the current CTU and the average saliency value of the current frame, respectively. S_{CTU} is calculated according to the saliency guidance by adding the value of all pixels belonging to it. k_1 determines the limit value of the saliency result for the salient and no-salient regions, while k_2 determines the changing speed of the saliency value from salient regions to non-salient regions. Here, we set k_1=0.5 and k_2= 1, empirically. As illustrated in Fig. 7, the higher the significance value of S_{CTU}, the larger the weight of ω. ω has a positive correlation with the compressed video quality.

In a traditional R-λ model, bpp determines the final QP value. In our method, we assign bits based on the weight map of the proposed saliency model. Firstly, bits per weight (bpw) is initialized:

$$\overline{bpw} = \frac{T}{\sum_{n=1}^{N} \omega_n}, \tag{9}$$

where T represents the target bit for the current frame, and ω_n represents the weight of the n-th CTU. N denotes the index number of CTU in the current frame. The relationship between

T_{sal} and T_{no_sal} is defined as follows:

$$\begin{cases} T_{sal} + T_{no\_sal} = T \\ T_{sal} = \dfrac{\sum_{n\in n_{sal}} \omega_n \overline{bpw}}{\sum_{n\in n_{no\_sal}} \omega_n \overline{bpw}} T_{non\_sal}, \end{cases} \tag{10}$$

where n_{sal} represents the index of the salient CTU, and the weighting value of which is larger than 1. n_{no_sal} denotes the index of the non-salient CTU, and the weighting value of which is smaller than 1. Then, the target bit T_{sal} for the salient regions can be calculated by

$$T_{sal} = \frac{\sum_{n\in n_{sal}} \omega_n}{\sum_{n\in n_{sal}} \omega_n + \sum_{n\in n_{no\_sal}} \omega_n} T. \tag{11}$$

Meanwhile, T_{no_sal} is calculated by

$$T_{no\_sal} = \frac{\sum_{n\in n_{no\_sal}} \omega_n}{\sum_{n\in n_{sal}} \omega_n + \sum_{n\in n_{no\_sal}} \omega_n} T. \tag{12}$$

Then, taking into consideration of a salient map, the target bits for the salient and no-salient regions can be obtained as follows. bpw_j for the j-th CTU can be computed by

$$bpw_j = \begin{cases} \left(T_{sal} - T_{sal}^{encoded}\right) / \sum_{n\in \hat{n}_{sal}} \omega_n \\ \left(T_{no\_sal} - T_{no\_sal}^{encoded}\right) / \sum_{n\in \hat{n}_{no\_sal}} \omega_n \end{cases}, \tag{13}$$

where $T_{sal}^{encoded}$ and $T_{no\_sal}^{encoded}$ represent the encoded bits for salient and no-salient regions, respectively. $\hat{n}_{sal}$ and $\hat{n}_{no\_sal}$ denote the current and its subsequent CTUs for the salient and non-salient regions, respectively.

Let T_j denote the target bits for the j-th CTU. bpp_j can be calculated by

$$bpp_j = \frac{T_j}{N_j} = \frac{\omega_j \cdot bpw_j}{N_j}, \tag{14}$$

where ω_j represents the weight value of the j-th CTU. It can be observed that CTUs with larger bpw and ω will be allocated more bits. Therefore, the salient regions are emphasized with more target bits. Then, rate control can be performed with bpp_j known for each CTU. Additionally, we adjust the boundary constraints of λ_j and QP_j so that a salient region has a higher priority in bit allocation.

Generally, based on the proposed saliency model, we employ bpw to estimate the perceived weight for each *CTU*. Then bpp, λ and QP are calculated successively based on bpw. After encoding a CTU, the related model parameters, such as α and β, are updated. Thus, the proposed perceptual quality control scheme sequentially obtains a QP value for each CTU in a frame. The difference between our method and the traditional R-λ is that we use a perceptual model and the weight of each pixel in bpw to estimate the value of bpp for each CTU. The larger the weight and bpw, the higher the bits allocated, which further results in better visual quality. The high-level sketch of the proposed approach is described in Algorithm 1.

Algorithm 1 Perceptual Quality-Driven Rate Control Scheme

Input:
Input video sequence;
Saliency map sequence;
Target bit-rate R_{tar};
Output:
QP and λ for each CTU.
1: Calculate the salient value S_{Pic} for each frame.
2: Calculate the target bits for the current frame $Target_{Pic}$.
3: **while** $i \le N$ **do**
4: Calculate the weighting factor ω for each CTU according to Eq. (8).
5: Calculate the target bits for the salient region T_{sal} and no-salient region T_{no-sal}.
6: Calculate the target bits and bpp for the current CTU.
7: Calculate λ and QP for the current CTU.
8: $i = i + 1$
9: **end while**

D. Video Enhancement Network

In perceptual quality control, one of the key problems is that if the coding quality of salient region is too high and that of non-salient region is too low, human eyes will feel uncomfortable at the boundary. To alleviate the mutation of coding quality at the boundary between the salient and non-salient regions, a time-varying and space-varying RNN for video enhancement is designed to further improve the performance of perceptual-driven rate control for video communication, as illustrated in Fig. 8.

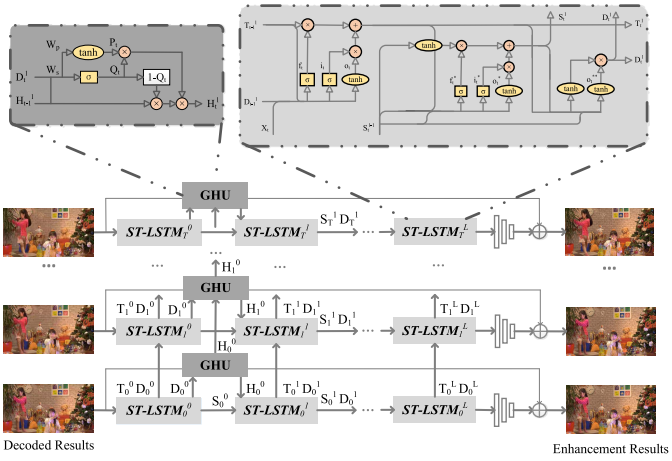


Fig. 8. Video enhancement network.

In the proposed enhancement module, both the space-varying and time-varying features are captured by ST-LSTM, which is built by connecting the temporal and spatial memories based on a gated structure in a cascade manner. By adding more nonlinear layers in the periodic transition and increasing the depth of the network structure from one state to another, ST-LSTM has a stronger feature extraction ability compared with convLSTM. An ST-LSTM cell contains a dual memory, a temporal memory, and a spatial memory module. The ST-LSTM of the k-layer is updated as follows.

$$T_t^k = f_t \odot T_{t-1}^k + i_t \odot o_t \tag{15}$$

$$S_t^k = f_t^* \odot tanh\left(W * S_t^{k-1}\right) + i_t^* \odot o_t^* \tag{16}$$

$$D_t^k = o_t^{**} \odot tanh\left(W * \left[D_t^k, S_t^k\right]\right), \tag{17}$$

where $*$, $\odot$ and *tanh* represent the convolution, element-wise multiplication and activation functions, respectively. W denotes the convolution filters. The subscript t represents a time step, and k represents the k-th hidden layer stacked in the causal LSTM network. The temporal memory depends on its previous state, which is controlled by a forgetting gate f_t, an input gate i_t and an input modulation gate O_t. The spatial memory S_t^k is determined by its previous layer state S_t^{k-1}. For the bottom layer (k=1), the uppermost space memory at $(t-1)$ is allocated to S_t^{k-1}, which is significantly different from the original LSTM. The final output is jointly determined by the dual storage state.

Since a recursion depth along the spatial-temporal transition path is significantly increased, the cascade memory is superior to the simple tandem structure of a spatial-temporal LSTM [38]. Each pixel in the final generated frame will have a larger input acceptance field at each time step, which provides the prediction model with greater modeling capabilities for short-term video dynamic changes and sudden changes.

Due to the long transition, the temporal memory may forget the appearance of outdated frames. Such a loop architecture is still unresolved, especially for videos with periodic motion or frequent occlusion. We need an information highway to learn the frame-skip relationship. A gradient highway unit (GHU) is utilized to prevent the rapid disappearance of long-term gradients [38], [39]. The formulation of GHU is defined as follows:

$$P_t = \tanh\left(W_{pd} * D_t^l + W_{ph} * H_{t-1}^l\right) \tag{18}$$

$$Q_t = \sigma\left(W_{sd} * D_t^l + W_{sh} * H_{t-1}^l\right) \tag{19}$$

$$H_t^l = P_t \odot Q_t + (1 - Q_t) \odot H_{t-1}^l, \tag{20}$$

where W represents a convolution kernel, and P_t denotes a switch gate to learn the transformation between D_t and the hidden state H_{t-1}^l. H_t^l represents the input for the next layer. *MaxPool* and *Conv* denote the `MaxPooling` and convolution operation, respectively.

We use the *Euclidean* loss between the output $I(x, y)$ and the ground-truth $I_*(x, y)$ as the loss function:

$$L = ||I(x, y) - I_*(x, y)||_2^2, \tag{21}$$

IV. Experimental Results

A. Experimental Condition and Evaluation Metrics

To verify the performance of the proposed method, we have implemented it on the HEVC reference software HM 16.7. The experiments have been conducted a computing platform with CPU@3.9G Hz and GPU@24G RAM. The BD-BR and BD-PSNR results are obtained based on four QPs = {22, 27, 32, 37}. The overall coding performance is measured by $\Delta PSNR$:

$$\Delta PSNR = PSNR_{proposed} - PSNR_{HM16.7} \tag{22}$$

where $PSNR_{proposed}$ and $PSNR_{HM16.7}$ represent the PSNR results of the proposed method and HM16.7.

In addition, the overall computational complexity is measured by ΔT:

$$\Delta T = \frac{T_{proposed} - T_{HM16.7}}{T_{HM16.7}} \times 100\%, \tag{23}$$

where $T_{proposed}$ and $T_{HM16.7}$ represent the coding time of the proposed method and HM16.7, respectively.

In video communication, RC aims to achieve the most accurate bit rate as required. Therefore, the bit rate error is an important evaluation index, which is represented by bit-rate error (BRE)(%), and defined by

$$BRE = \frac{|TBR - ABR|}{TBR} \times 100\%, \tag{24}$$

where TBR and ABR represent the target and actual bits, respectively.

B. Coding Performance on Conventional Dataset

1) Comparison With Standard HEVC Algorithm: Table II provides the overall BD-BR and BD-PSNR results of our method compared with HEVC. Class A to Class E are the standard test sequences recommended by JCT-VC, while Class F and Class G are built by ourselves. The proposed method obtains an average quality increment by 1.50 dB in the salient regions and a decrease by 0.30 dB in the no-salient regions. Generally, human eyes pay more attention to the salient regions, while the remaining areas attract less attention. The PSNR improvement in salient regions enhances the visual experience. As for the rate control, the actual bit rate is almost equal to the target bit rate, which means that the visual perception quality can be significantly improved without increasing the bandwidth burden.

2) R-D Curve Performance: To show the overall R-D performance, we also provide the R-D curves as illustrated in Fig. 9. It can be seen that the proposed method obtains a higher PSNR value on the salient regions from the low bit-rate to the high bit-rate compared with HM16.7. This improvement is achieved at the cost of reducing the quality of non-salient regions. Experimental results demonstrate that our method can improve the perceptual video quality without increasing the bandwidth burden.

3) Comparison With Other Perceptual Quality-Based Methods: In this section, we also compare the proposed algorithm with recent perceptual quality-based coding algorithms in terms of BD-BR and BD-PSNR. It can be seen from Table III that our method obtains the average BD-BR and BD-PSNR results for the salient regions are −21.22% and 1.30 dB, respectively. The performance of the proposed quality control scheme outperforms other competing algorithms in the salient regions. The main reason can be that the proposed scheme takes into account the visual characteristics of the human eyes, which allocates the bit-rate more reasonably.

C. Coding Performance on the Eye-Tracking Dataset

Apart from the conventional video dataset, we also verify the coding performance of the proposed scheme on the eye-tracking dataset [40], [41]. Experimental results are evaluated

TABLE II
CODING PERFORMANCE OF THE SALIENT AND NON-SALIENT REGIONS COMPARED WITH HM 16.7 UNDER THE SAME BIT-RATE

Class	Sequence	Target Bitrate (kbps)	Actual Bitrate (kbps)	BRE (%)	Δ PSNR (dB) whole	non-salient	salient	Coding Time Δ T (%)
A	PeopleOnStreet	68820	68824	0.0058	-0.1771	-0.1916	1.3821	-2.73
[2560 × 1600]	Traffic	64708	64710	0.0031	-0.2071	-0.2184	1.4436	-3.31
	BasketballDrive	15589	15591	0.0128	-0.2119	-0.2220	0.7762	-3.87
B	BQTerrace	44340	44340	0	-0.1829	-0.2033	1.4472	-4.45
[1920 × 1024]	Cactus	31160	31162	0.0064	-0.1391	-0.1536	1.0822	-2.35
	ParkScene	35638	35638	0	-0.1390	-0.1543	1.6485	-1.19
	BasketballDrill	6647	6650	0.0451	-0.1785	-0.1915	1.3692	-1.96
C	BQMall	9365	9370	0.0533	-0.2050	-0.2383	1.9049	-3.85
[832 × 448]	PartyScene	20215	20219	0.0197	-0.2368	-0.2523	1.9473	-5.32
	RaceHorses	11100	11110	0.0900	-0.1266	-0.1407	1.8465	-2.06
	BasketballPass	1508	1509	0.0663	-0.0257	-0.1178	1.6039	-2.07
D	BlowingBubbles	2595	2597	0.0770	-0.0990	-0.1282	1.3843	-2.95
[384 × 192]	BQSquare	4483	4484	0.0223	-0.0994	-0.1849	1.8288	-1.56
	RaceHorses	2573	2575	0.0777	-0.1277	-0.1765	1.6877	-6.34
	FourPeople	11315	11317	0.0176	-0.3170	-0.3390	0.8580	-3.49
	Johnny	6255	6257	0.0319	-0.2822	-0.2958	0.8084	-5.62
E	Kristen	7970	7972	0.0250	-0.3131	-0.3431	0.9652	-3.54
[1280 × 704]	Vidyo1	7136	7138	0.0280	-0.2770	-0.3065	1.2890	-4.96
	Vidyo3	8257	8258	0.0121	-0.2759	-0.3130	1.2067	-5.64
	Host	4252	4253	0.0235	-0.3142	-0.3804	2.0059	-3.34
F	Broadcaster	4933	4932	0.0202	-0.3711	0.3847	1.8509	-2.92
[1280 × 704]	VirtualConference3	4109	4110	0.0243	-0.5303	-0.7098	1.1580	-0.61
	VirtualConference4	5782	5784	0.0345	-0.4235	-0.5848	1.4660	-3.48
	Pair figure-skating	2516	2522	0.2384	-0.4918	-0.8952	1.2600	-6.20
G	Artistic gymnastics	8252	8250	0.0242	-0.2358	-0.3965	2.5977	-0.57
[1280 × 704]	Men's figure-skating	2642	2648	0.2271	-0.1863	-0.3215	1.8061	-6.76
	Women's figure-skating	3704	3708	0.1079	-0.3203	-0.4239	1.9357	-3.42
Average		**14661**	**14664**	**0.0447**	**-0.2406**	**-0.3023**	**1.5023**	**-3.48**

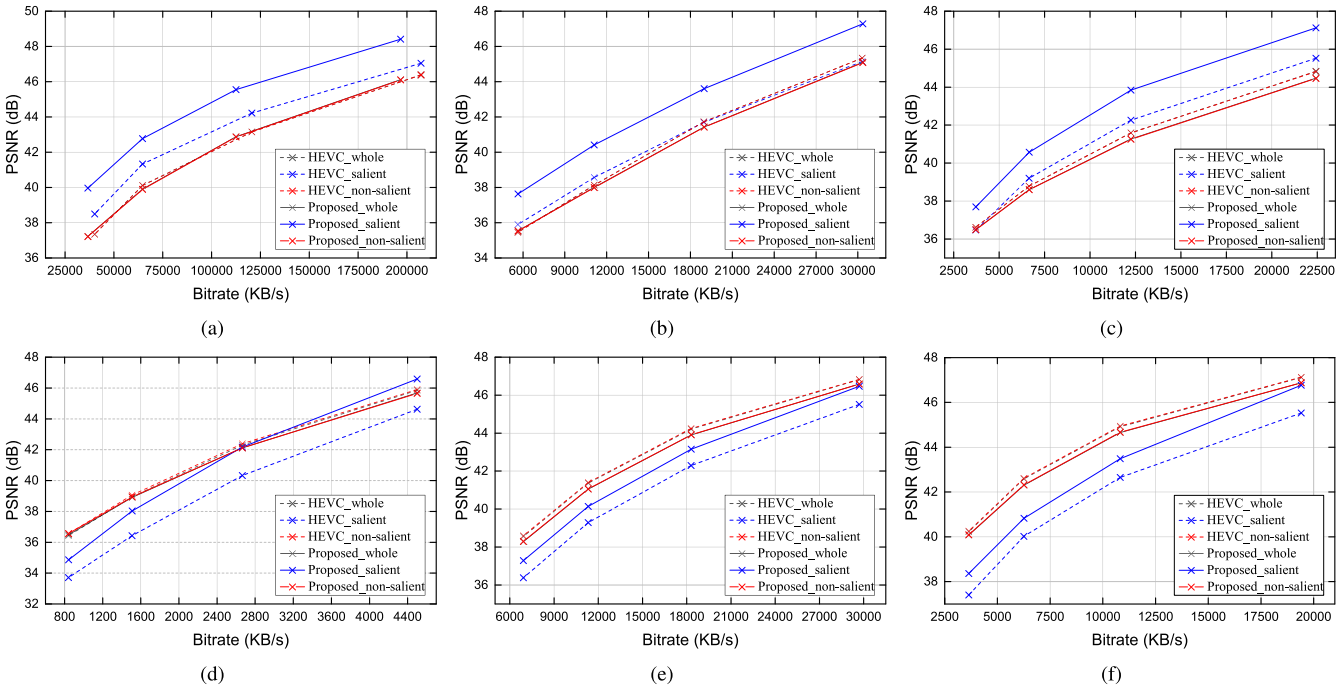


Fig. 9. R-D curves of six video sequences: (a) *Traffic*, (b) *RaceHorses*, (c) *BasketballDrill*, (d) *BasketballPass*, (e) *FourPeople*, and (f) *Johnny*.

in terms of BD-PSNR and BD-EWPSNR. EWPSNR takes into account human perceptual information, which is defined as follows:

$$EWPSNR = 10log_{10}\left(\frac{255^2}{EWMSE}\right), \tag{25}$$

where *EWMSE* is defined by

$$EWMSE = \frac{\sum_{x=1}^{W}\sum_{y=1}^{H}\left(\omega_{x,y}\cdot\left(F'_{x,y}-F_{x,y}\right)^2\right)}{W\cdot H\cdot\sum_{x=1}^{W}\sum_{y=1}^{H}\omega_{x,y}}, \tag{26}$$

TABLE III
CODING PERFORMANCE COMPARISON WITH THE HADIZADEH'S AND ZHU'S METHODS IN TERMS OF BD-BR AND BD-PSNR

Sequence	Hadizadeh *et al.* [16]				Zhu *et al.* [17]				Perceptual-based Method			
	Whole Image		Salient Region		Whole Image		Salient Region		Whole Image		Salient Region	
	BD-BR (%)	BD-PSNR (dB)	BD-BR (%)	BD-PSNR (dB)	BD-BR (%)	BD-PSNR (dB)	BD-BR (%)	BD-PSNR (dB)	BD-BR (%)	BD-PSNR (dB)	BD-BR (%)	BD-PSNR (dB)
PeopleOnStreet	5.94	-0.21	-7.36	0.61	12.27	-0.63	-7.61	0.51	9.03	-0.42	-18.28	1.19
Traffic	14.74	-0.89	-3.08	0.19	13.77	-0.76	-10.07	0.69	-0.43	0.02	-27.12	1.63
BasketballDrive	16.51	-0.68	-3.71	0.25	19.76	-0.78	-11.14	0.53	7.91	-0.20	-22.69	1.18
Kimono	8.56	-0.39	-11.57	0.22	8.78	-0.43	-10.10	0.50	2.89	-0.10	-19.75	0.42
BasketballDrill	10.38	-0.61	-9.28	0.31	13.03	-0.62	-3.12	0.57	5.76	-0.24	-24.72	1.46
RaceHorses	5.61	-0.48	-6.89	0.60	14.66	-0.40	-13.93	0.65	2.69	-0.15	-29.34	1.87
BasketballPass	6.01	-0.32	-4.62	0.33	15.91	-0.56	-6.22	0.43	1.58	-0.08	-22.02	1.67
BQSquare	5.84	-0.49	-4.18	0.59	6.96	-0.18	-13.02	0.64	2.02	-0.15	-18.26	1.85
FourPeople	12.67	-0.83	-3.63	0.53	14.32	-0.83	-12.29	0.76	5.36	-0.29	-13.09	0.87
Johnny	15.87	-0.77	-6.82	0.70	11.69	-0.55	-9.47	0.50	6.33	-0.25	-16.91	0.89
Average	**10.21**	**-0.56**	**-6.11**	**0.43**	**12.83**	**-0.57**	**-11.56**	**0.59**	**4.31**	**-0.19**	**-21.22**	**1.30**

TABLE IV
PERFORMANCE OF THE PROPOSED ALGORITHM VERSUS THE HADIZADEH'S AND ZHU'S METHODS ON THE EYE-TRACKING DATASET

Test Sequence	Zhu *et al.* [17]		Hadizadeh *et al.* [16]		Perceptual-based Method	
	BD-PSNR (dB)	BD-EWPSNR (dB)	BD-PSNR (dB)	BD-EWPSNR (dB)	BD-PSNR (dB)	BD-EWPSNR (dB)
Bus	-0.39	0.47	-0.61	0.24	-0.42	0.44
City	-0.24	0.64	-0.45	0.16	-0.16	1.27
Crew	-0.15	0.33	-0.34	0.02	-0.34	1.79
Stefan	-0.16	0.72	-0.5	0.42	-0.46	0.85
Harbor	-0.19	0.44	-0.34	0.32	-0.33	1.03
Soccer	-0.47	0.11	-0.56	-0.03	-0.42	1.41
Tempete	-0.26	0.59	-0.47	0.28	-0.49	0.32
Foreman	-0.26	0.50	-0.51	0.08	-0.25	0.70
Hall Monitor	-0.05	0.39	-2.64	-1.66	-0.20	0.65
Flower Garden	-0.17	0.50	-0.24	0.52	-0.50	0.58
Mobile Calendar	-0.19	0.73	-0.43	0.54	-0.72	0.81
Mother Daughter	-0.39	-0.14	-0.54	-0.31	-0.46	0.52
Average	**-0.24**	**0.44**	**-0.64**	**0.05**	**-0.40**	**0.80**

where $\omega_{x,y}$ represents the perceived weight at the pixel position of (x,y), and $F'_{x,y}$ and $F_{x,y}$ denote the pixel values before and after encoding.

Table IV provides the BD-PSNR and BD-EWPSNR results. As seen, the proposed method obtains the EWPSNR gain by 0.80 dB compared with HEVC, 0.75 dB compared with Hadizadeh's method, and 0.36 dB compared with Zhu's method.

D. Subjective Visual Quality

To demonstrate the subjective performance, Fig. 10 and Fig. 11 provide several reconstructed frames encoded by HM 16.7 and the proposed method. The experimental results show that the proposed method greatly improves the quality of salient regions. For instance, the proposed method preserves more details on the human faces than HM 16.7. Also, our method has higher coding quality on moving objects. It indicates that the proposed perceptual quality control method can effectively improve the coding quality for salient regions at the same bandwidth.

E. Computational Complexity

Table V provides the average computational complexity in terms of the running time (Sec.) and the frame per second (FPS) in predicting a saliency map. As seen, the proposed saliency detection method may not directly satisfy some real-time applications (e.g., video conference or online classroom). However, some optimization methods can be used to solve this problem: 1) Due to the fact that video frames are very similar to each other in a short period of time, several adjacent frames can share the same saliency map to save the computing time. 2) Since specific saliency values are not that sensitive in perceptual quality control, one can reduce the image resolution of the input video frame when predicting the saliency guidance. 3) Some model optimization approaches can be further used to reduce the computational complexity of deep models, including knowledge distillation, weight quantization, pruning, and network structure optimization.

TABLE V
AVERAGE COMPUTATIONAL COMPLEXITY IN TERMS OF THE RUNNING TIME (SEC.) AND THE FRAME PER SECOND (FPS).

Video Resolution	Static Saliency		Dynamic Saliency		Enhancement	
	Time	FPS	Time	FPS	Time	FPS
1920×1024	0.249	4.02	1.168	0.86	2.047	0.49
1280×704	0.156	6.41	0.405	2.47	0.935	1.07
832×448	0.124	8.06	0.185	5.41	0.400	2.50
384×192	0.096	10.42	0.058	17.24	0.086	11.63

(a) *Kimono*

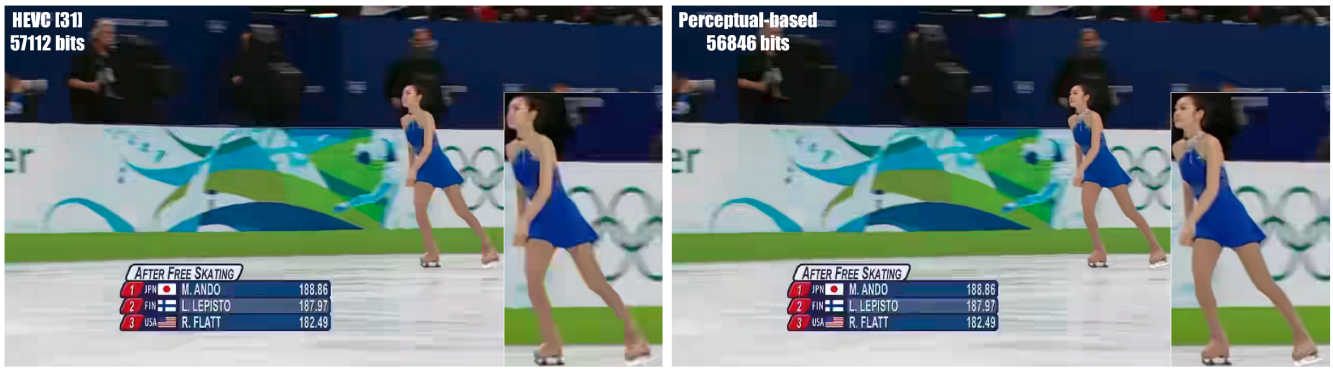
(b) *Broadcaster*

(c) VirtualConference

(d) PartyScene

Fig. 10. Qualitative demonstration for the comparison of subjective coding performance between proposed method and HEVC on static videos.

However, it is worth noting that our method can still be used to compress a raw video in an offline manner. Furthermore, in the experiments, the salient region extraction and video coding have been run in parallel on a GPU and a CPU, separately. They communicate through a shared folder. As for each frame, saliency detection takes less computation time



(a) *Women's figure-skating*

(b) *Pair figure-skating*

(c) *Rollercoaster*

Fig. 11. Qualitative comparisons of the subjective performance between the proposed method and HEVC on moving objects.

than encoding, and hence saliency detection will not increase the overall running time as illustrated in Table II.

V. Conclusion and Discussion

In this paper, we proposed a deep-learned perceptual quality control method for intelligent consumer electronic video display and transmission. Specifically, we develop a multi-scale saliency extraction network to extract the saliency regions for static videos, while exploring an LSTM-based network to extract the moving objects for dynamic videos. Based on the static and dynamic saliency guidance, we further design a three-level rate control scheme: more bits are allocated to the salient regions to preserve higher visual quality, and fewer bits are allocated to the non-salient regions to save more bits. This strategy can significantly improve the human visual experience without increasing bandwidth. To avoid the perceptual quality degradation caused by the difference between the salient and no-salient regions, we also propose an effective video enhancement module. Subjective and objective experiments validate the effectiveness of the proposed method in ensuring the perceptual video quality under a target bit-rate.

Video compression is currently tightly linked with consumer electronic video display and transmission. Currently, most intelligent-terminal products are equipped with video communication modules, such as smartphone, surveillance, laptop, drone, *etc*. Perceptual quality control can ensure the quality of video service without increasing the transmission bandwidth. For various video-based applications of consumer electronics, the proposed method can be deployed both on the server side and the client side. Therefore, we believe that our method can be useful for many video-based applications serving a large number of consumers.

REFERENCES

[1] E. De la Torre, R. Rodriguez-Sanchez, and J. L. Martínez, "Fast video transcoding from HEVC to VP9," *IEEE Trans. Consum. Electron.*, vol. 61, no. 3, pp. 336–343, Aug. 2015.

[2] D. Kobayashi, K. Nakamura, T. Onishi, H. Iwasaki, and A. Shimizu, "A 4K/60p HEVC real-time encoding system with high quality HDR color representations," *IEEE Trans. Consum. Electron.*, vol. 64, no. 4, pp. 433–441, Nov. 2018.

[3] G. Kulupana, D. S. Talagala, H. K. Arachchi, M. Akinola, and A. Fernando, "Concealment support and error resilience for HEVC to improve consumer quality of experience," *IEEE Trans. Consum. Electron.*, vol. 67, no. 2, pp. 107–118, May 2021.

[4] A. Singhadia, M. Mamillapalli, and I. Chakrabarti, "Hardware-efficient 2D-DCT/IDCT architecture for portable HEVC-compliant devices," *IEEE Trans. Consum. Electron.*, vol. 66, no. 3, pp. 203–212, Aug. 2020.

[5] M. J. Garrido, F. Pescador, M. Chavarrias, P. J. Lobo, and C. Sanz, "A 2-D multiple transform processor for the versatile video coding standard," *IEEE Trans. Consum. Electron.*, vol. 65, no. 3, pp. 274–283, Aug. 2019.

[6] M. J. Garrido, F. Pescador, M. Chavarrias, P. J. Lobo, and C. Sanz, "A high performance FPGA-based architecture for the future video coding adaptive multiple core transform," *IEEE Trans. Consum. Electron.*, vol. 64, no. 1, pp. 53–60, Feb. 2018.

[7] M. Wang, X. Liu, W. Xie, and L. Xu, "Perceptual redundancy estimation of screen images via multi-domain sensitivities," *IEEE Signal Process. Lett.*, vol. 28, pp. 1440–1444, 2021.

[8] H. Xun, C. Shen, X. Boix, and Z. Qi, "SALICON: Reducing the semantic gap in saliency prediction by adapting deep neural networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2015, pp. 262–270.

[9] H. Yuan, Q. Wang, Q. Liu, J. Huo, and P. Li, "Hybrid distortion-based rate-distortion optimization and rate control for H.265/HEVC," *IEEE Trans. Consum. Electron.*, vol. 67, no. 2, pp. 97–106, May 2021.

[10] M. Wang and B. Y. Yan, "Lagrangian multiplier based joint three-layer rate control for H.264/AVC," *IEEE Signal Process. Lett.*, vol. 16, no. 8, pp. 679–682, Aug. 2009.

[11] M. Wang, K. N. Ngan, and H. Li, "Low-delay rate control for consistent quality using distortion-based lagrange multiplier," *IEEE Trans. Image Process.*, vol. 25, no. 7, pp. 2943–2955, Jul. 2016.

[12] K. Goswami, J. H. Lee, and B. G. Kim, "Fast algorithm for the high efficiency video coding (HEVC) encoder using texture analysis," *Inf. Sci.*, vols. 364–365, pp. 72–90, Oct. 2016.

[13] H. Xue, Y. Zhang, and Y. Wei, "Fast ROI-based HEVC coding for surveillance videos," in *Proc. 19th Int. Symp. Wireless Pers. Multimedia Commun. (WPMC)*, 2016, pp. 299–304.

[14] 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.

[15] S. Li, M. Xu, X. Deng, and Z. Wang, "A novel weight-based URQ scheme for perceptual video coding of conversational video in HEVC," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, 2014, pp. 1–6.

[16] H. Hadizadeh and I. V. Bajić, "Saliency-aware video compression," *IEEE Trans. Image Process.*, vol. 23, no. 1, pp. 19–33, Jan. 2013.

[17] S. Zhu, C. Liu, and Z. Xu, "High-definition video compression system based on perception guidance of salient information of a convolutional neural network and HEVC compression domain," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 7, pp. 1946–1959, Jul. 2020.

[18] M. Zhou *et al.*, "SSIM-based global optimization for CTU-level rate control in HEVC," *IEEE Trans. Multimedia*, vol. 21, no. 8, pp. 1921–1933, Aug. 2019.

[19] W. Li, P. Ren, E. Zhang, and F. Zhao, "Rate control for HEVC intra-coding with a CTU-dependent distortion model," *Signal Image Video Process.*, vol. 13, no. 1, pp. 17–25, Aug. 2019.

[20] Z. Chen and X. Pan, "An optimized rate control for low-delay H.265/HEVC," *IEEE Trans. Image Process.*, vol. 28, no. 9, pp. 4541–4552, Sep. 2019.

[21] P. Wang, C. Ni, G. Zhang, and K. Li, "R-Lambda model based CTU-level rate control for intra frames in HEVC," *Multimedia Tools Appl.*, vol. 78, no. 1, pp. 125–139, 2019.

[22] Y. Zeng, Y. Zhuge, H. Lu, L. Zhang, M. Qian, and Y. Yu, "Multi-source weak supervision for saliency detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 6074–6083.

[23] T. Yubing, F. A. Cheikh, F. F. E. Guraya, H. Konik, and A. Trémeau, "A spatiotemporal saliency model for video surveillance," *Cogn. Comput.*, vol. 3, no. 1, pp. 241–263, 2011.

[24] L. Wang, H. Lu, X. Ruan, and M.-H. Yang, "Deep networks for saliency detection via local estimation and global search," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3183–3192.

[25] P. Tokmakov, C. Schmid, and K. Alahari, "Learning to segment moving objects," *Int. J. Comput. Vis.*, vol. 127, no. 3, pp. 282–301, 2019.

[26] W. Wang, J. Shen, R. Yang, and F. Porikli, "Saliency-aware video object segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 1, pp. 20–33, Jan. 2018.

[27] A. Faktor and M. Irani, "Video segmentation by non-local consensus voting," in *Proc. BMVC*, vol. 2, 2014, p. 8.

[28] P. Xing, Y. Tian, T. Huang, and W. Gao, "Surveillance video coding with quadtree partition based ROI extraction," in *Proc. Picture Coding Symp. (PCS)*, 2013, pp. 157–160.

[29] C.-L. Zhao, M. Dai, and J.-Y. Xiong, "Region-of-interest based rate control for UAV video coding," *Optoelectron. Lett.*, vol. 12, no. 3, pp. 216–220, 2016.

[30] K. Singh and S. R. Ahamed, "Low power motion estimation algorithm and architecture of HEVC/H.265 for consumer applications," *IEEE Trans. Consum. Electron.*, vol. 64, no. 3, pp. 267–275, Aug. 2018.

[31] B. Xiong, X. Fan, C. Zhu, X. Jing, and Q. Peng, "Face region based conversational video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 7, pp. 917–931, Jul. 2011.

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

[33] J. Xu, M. Jiang, S. Wang, M. S. Kankanhalli, and Z. Qi, "Predicting human gaze beyond pixels," *J. Vis.*, vol. 14, no. 1, pp. 97–97, 2014.

[34] F. Bossen *et al.*, "Common test conditions and software reference configurations," in *Proc. JCTVC*, vol. 12, 2013, p. 9.

[35] X. Ning, L. Yang, Y. Fan, J. Yang, and T. Huang, "YouTube-VOS: Sequence-to-sequence video object segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 603–619.

[36] G. J. Sullivan and T. Wiegand, "Rate-distortion optimization for video compression," *IEEE Signal Process. Mag.*, vol. 15, no. 6, pp. 74–90, Nov. 1998.

[37] B. Li *et al.*, "Adaptive bit allocation for R-lambda model rate control in 715 HM," in *Proc. JCTVC M0036 13th Meeting Joint Collaborative Team Video Coding ITU-T SG1 6WP3 ISO/IEC JTC1/SC*, vol. 29, 2013, pp. 18–26.

[38] Y. Wang, Z. Gao, M. Long, J. Wang, and P. S. Yu, "PredRNN++: Towards a resolution of the deep-in-time dilemma in spatiotemporal predictive learning," in *Proc. PMLR*, 2018, pp. 5123–5132.

[39] X. Sun, S. Wang, M. Wang, S. S. Cheng, and M. Liu, "An advanced LiDAR point cloud sequence coding scheme for autonomous driving," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 2793–2801.

[40] H. Hadizadeh, M. J. Enriquez, and I. V. Bajic, "Eye-tracking database for a set of standard video sequences," *IEEE Trans. Image Process.*, vol. 21, no. 2, pp. 898–903, Feb. 2012.

[41] X. Sun, X. F. Yang, S. Wang, and M. Liu, "Content-aware rate control scheme for HEVC based on static and dynamic saliency detection," *Neurocomputing*, vol. 411, pp. 393–405, Oct. 2020.