The Journal of Supercomputing (2025) 81:464
https://doi.org/10.1007/s11227-025-06954-z

Perceptual QP optimization for VVC with dual hybrid neural networks

Javier Ruiz Atencia[1] · **Otoniel Mario López Granado**[1] · **Manuel Pérez Malumbres**[1] · **Miguel Onofre Martínez-Rach**[1]

Accepted: 14 January 2025

Abstract

This paper introduces a dual hybrid neural network model combining convolutional neural networks (CNNs) and artificial neural networks (ANNs) to optimize the quantization parameter (QP) for both 64×64 and 32×32 blocks in the versatile video coding (VVC) standard, enhancing video quality and compression efficiency. The model employs CNNs for spatial feature extraction and ANNs for structured data handling, addressing the limitations of current heuristic and just noticeable distortion (JND)-based methods. A dataset of luminance channel image blocks, encoded with various QP values, is generated and preprocessed, and the dual hybrid network structure is designed with convolutional and dense layers. The QP optimization is applied at two levels: the 64×64 model provides a global QP offset, while the 32×32 model refines the QP for further partitioned blocks. Performance evaluations using model error metrics like mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), as well as perceptual metrics like weighted PSNR (WPSNR), MS-SSIM, PSNR-HVS-M, and VMAF, demonstrate the model's effectiveness. While our approach performs competitively with state-of-the-art algorithms, it significantly outperforms in VMAF, the most advanced and widely adopted perceptual quality metric. Furthermore, the dual-model approach yields better results at lower resolutions, whereas the single-model approach is more effective at higher resolutions. These results highlight the adaptability of the proposed models, offering improvements in both compression efficiency and perceptual quality, making them highly suitable for practical applications in modern video coding.

Keywords Hybrid network · CNN · Perceptual · QP · VVC · AdaptiveQP · QPA · HVS

O. M. L. Granado, M. P. Malumbres, M. O. Martínez-Rach have contributed equally to this work.

Extended author information available on the last page of the article

Published online: 05 February 2025

1 Introduction

In today's digital age, the surge of high-quality video content has become more apparent than ever, driven by widespread streaming services, social media platforms, and various multimedia applications. Videos are produced, uploaded, and consumed in massive volumes across networks that are still bound by bandwidth constraints. This scenario, combined with the rapid evolution of resolutions-from full HD to 4K and beyond-further underscores the need for advanced video compression strategies. Effective video compression not only saves storage space but also ensures smooth video delivery over bandwidth-limited networks. Consequently, the research and development of new techniques to improve compression efficiency without compromising perceived video quality has become a central focus in the field of video coding.

Among the most recent advances in compression technology is the versatile video coding (VVC) standard [1], also known as H.266. Building upon the success of its predecessors (H.264/AVC and H.265/HEVC), VVC typically achieves the same subjective video quality at around half the bit rate compared to earlier standards. While these advancements are substantial, challenges remain. One of the most critical tasks in any video codec is selecting the quantization parameter (QP), which directly influences the trade-off between bit rate and visual quality at the block level. Inappropriate QP assignments across different spatial regions can lead to noticeable distortions and/or inefficient use of bits.

Several strategies have been proposed to address this challenge. Traditional approaches rely on heuristic rules or basic visual sensitivity metrics for perceptual optimization. For instance, JND-based (just noticeable distortion) methods leverage fundamental insights into the human visual system (HVS) to dynamically adjust the QP such that distortions remain below the typical detection thresholds of human observers. However, conventional JND schemes and even the sophisticated reference algorithms in VVC's test model (VTM) [2], such as AdaptiveQP [3, 4] and quantization parameter adaptation (QPA) [5], are still limited by their reliance on fixed rules and predefined sensitivity metrics. As a result, they may not fully exploit the complex interactions between spatial content, motion, and human perception in diverse video sequences.

To bridge this gap, our work introduces a novel dual hybrid neural network framework that integrates convolutional neural networks (CNNs) for extracting spatial features from video blocks and artificial neural networks (ANNs) for structured data processing. By harnessing the unique advantages of these two architectures, our model refines the QP assignment for both 64×64 and 32×32 blocks in VVC, seeking to deliver competitive compression ratios while maximizing perceptual quality metrics such as WPSNR, MS-SSIM, PSNR-HVS-M, and especially VMAF.

The remainder of this paper is organized as follows. Section 2 presents the state-of-the-art approaches in rate control and perceptual QP optimization, highlighting both traditional methods and recent neural-network-based solutions. Section 3 details our methodology, including dataset preprocessing and our dual-model architecture. Section 4 reports experimental results and provides a comparative analysis

with existing methods. Finally, Sect. 5 concludes the paper and points to future research directions.

2 Related work

In VVC, AdaptiveQP [3, 4] and QPA [5] stand out as built-in perceptual mechanisms in VTM. AdaptiveQP uses local variance in luminance blocks to increase QP in highly textured regions, leveraging the masking effect where noise or artifacts can be concealed by high detail. QPA goes further, estimating a QP offset at both the CTU (128×128) and CU (64×64) levels. Despite their effectiveness, their heuristics do not necessarily capture complex motion or perception-related phenomena (e.g., occlusions, fast scene changes, or local object details).

Just noticeable distortion (JND) methods have also been integrated into many codecs to better mimic the HVS [6], indicating that some distortions remain imperceptible in areas with higher texture or in the presence of luminance masking. In HEVC, for example, researchers proposed JND-based rate control schemes that attempt to distribute bits more intelligently across frames and regions [7–10]. Zhou et al. [11] introduced a JND-based perceptual rate control approach that mathematically models the relationship between JND factors and allocated bits, demonstrating notable improvements in subjective quality. These approaches underscore the importance of accounting for human perception to achieve higher coding efficiency, yet they often rely on hand-crafted or semi-empirical models of visual sensitivity.

Machine learning (ML) and deep learning (DL) have become increasingly prominent in video coding research, as they can learn complex mappings between video features and optimal encoding decisions without relying solely on predefined or piecewise functions. Particularly in rate control, deep reinforcement learning (DRL) has gained traction for dynamic video sequences, where spatio-temporal content shifts rapidly and conventional prediction models fail. Zhou et al. [12] proposed a DRL-based rate control method for HEVC to minimize distortion, buffer underflows/overflows, and quality fluctuations in scenes containing fast motion, occlusions, or abrupt changes. This technique models rate control as a Markov decision process (MDP), allowing the learned agent to adapt QP selection at both frame and CTU levels.

Other recent works emphasize global optimization with advanced rate-distortion (R-D) models. For example, in [13], a decision-tree-based scheme for VVC UHD coding was proposed, coupling R-D modeling with visual features to refine bit allocation in ultrahigh-resolution videos. A relevant concept is the shift from pixel- or block-based fidelity metrics to more perceptually aligned measures like SSIM or VMAF. Zhou et al., in another study [14], tackled SSIM-based global optimization for CTU-level rate control, casting bit allocation into a convex optimization problem and demonstrating the potential for higher perceptual quality gains over standard λ-domain approaches. Moreover, Wei et al. provided a comprehensive review of state-of-the-art rate control [15], illustrating how $R - \delta$ modeling, $R - Q$ modeling, and ML-based methods each have their own set of advantages and challenges.

Notably, rate control is essential for balancing bit rate usage and video quality under practical network constraints. Historically, solutions such as TM5 for MPEG-2 [16] and VM8 for MPEG-4 [17] employed empirical or polynomial models to predict bit rate-distortion trade-offs. In H.264/AVC and H.265/HEVC, more advanced approaches like JVT-G012 [18] and JCTVC-H0213 [19] appeared, either applying $R - Q$ relationships or resorting to λ-domain models, respectively, to derive quantization parameters. Although these methods substantially improved rate-distortion performance, they commonly rely on simplistic assumptions about how spatial detail and motion complexity affect perceived quality.

While each of the aforementioned strategies brings valuable insights into rate control and QP optimization, they either focus on a single scale (e.g., CTU-level adaptation) or rely heavily on handcrafted metrics and specific assumptions. Our proposed dual hybrid neural network addresses these limitations by integrating:

1. Spatial feature extraction via CNNs: Convolutional layers capture detailed spatial patterns (textures, edges, etc.) that can guide QP decisions based on local complexity and perceptual sensitivity.
2. Structured data handling via ANNs: Additional block-level statistics and encoder-side features (e.g., motion vectors, coding cost, or variance measures) are processed in fully connected layers, allowing the model to fuse numeric and visual cues.
3. Two-stage modeling: We explore two operating modes for QP optimization. In the first, we apply a QP offset solely at the 64×64 block level. In the second, the QP offset determined for the 64×64 block is further refined at the 32×32 block level. This hierarchical approach allows us to compare a single-scale offset application with a more fine-grained adaptation that handles local variations more precisely.
4. Comprehensive evaluation: Our framework is validated against multiple perceptual quality metrics (WPSNR, MS-SSIM, PSNR-HVS-M, VMAF), with a particular focus on VMAF, which correlates more strongly with human visual perception than classic metrics like PSNR.

Thus, we extend existing literature by creating a robust pipeline that merges neural feature extraction with structured data modeling to realize efficient and perceptually guided QP optimization for VVC.

3 Methodology

This section details the methodology used for designing and evaluating the proposed hybrid model. First, the generation and preprocessing of the dataset consisting of 64×64 and 32×32 pixel image blocks using the luminance channel are described. Next, the architecture of the hybrid neural network, which integrates a convolutional neural network (CNN) for image processing and an artificial neural

network (ANN) for handling structured data, is presented. Finally, the normalization and preprocessing techniques applied to the data before being fed into the neural network are explained.

3.1 Dataset preparation

A dataset of 64×64 and 32×32 pixel image blocks, using only the luminance channel, has been developed. This dataset is designed to train and evaluate a hybrid convolutional neural network model. The block datasets were extracted from some image databases like ESPL synthetic image database [20], USC-SIPI image database [21], TESTIMAGE [22] and Kodak image dataset [23] as well as from some images randomly selected from the video sequences of VVC common test conditions [24].

For dataset generation, the VVC reference software, known as VTM [2], was used and modified to partition the video exclusively into square blocks of 64×64 and 32×32 pixels and store them into a CSV (comma-separated values) file. Each video sequence was encoded in All-intra mode, with a wide range of QP values from 12 to 47. This value is stored in the dataset as $\mathrm{QP}_{\mathrm{base}}$ and will be one of the input elements to the neural networks.

While the VVC standard allows the use of QP offsets at smaller block sizes (e.g., 16×16 or 8×8), preliminary tests indicate that the additional overhead in the bitstream significantly diminishes any rate-distortion benefits. Hence, we chose not to include blocks smaller than 32×32 in our approach.

During the encoding process, each frame is partitioned into 64×64 and 32×32 blocks, and rate-distortion optimization (RDO) is used to decide how to encode in a way that minimizes visual distortion (i.e., loss of quality) while controlling the amount of data needed to represent that block (i.e., bitrate). This cost function is mathematically formalized as follows:

$$\min_{\mathbf{p}_k} \; D_k^{\mathrm{SSE}}(\mathbf{p}_k) + \lambda \cdot R_k(\mathbf{p}_k) \tag{1}$$

where D_k^{SSE} denotes the sum of squared errors (SSE) for a block $\mathbf{B}_k$, $R_k(\mathbf{p}_k)$ is the rate for a block $\mathbf{B}_k$, λ is the lagrange multiplier, which depends on the QP value, and $\mathbf{p}_k$ is the vector of encoding decisions for the block $\mathbf{B}_k$.

At this stage, further modifications were made to the VTM reference software. In the RDO, a weighted SSE distortion metric based on the WPSNR (weighted peak signal-to-noise ratio) [5] was used instead of the conventional SSE distortion measure. Therefore, Eq. (1) is modified as follows:

$$\min_{\mathbf{p}_k} \; w_k \cdot D_k^{\mathrm{SSE}}(\mathbf{p}_k) + \lambda \cdot R_k(\mathbf{p}_k) \tag{2}$$

where w_k is the weighting factor for a $\mathbf{B}_k$. In addition, a process of searching for the perceptually optimal QP value has been conducted. For this purpose, we have added a new stage to the encoding process that allows us to use a range of QP values around the $\mathrm{QP}_{\mathrm{base}}$. The variable that controls this QP offset is called Delta QP (ΔQP), and it is defined as:

$$\Delta QP \in \{-6, -5, \ldots, 5, 6\} \tag{3}$$

This means that, for each block, a total of thirteen different QP values are evaluated. For each of these encodings, the weighted RDO is performed (Eq. 2), and after all the encoding processes, the ΔQP value that minimizes the cost of the RDO is considered the ground truth and is stored in the dataset, along with the block pixels in the luminance channel. Figure 1 summarizes the entire process described for a given QP_{base} value.

In addition to the QP base and blocks per frame (BPF) data, we have expanded the structured data inputs for the ANN with additional features derived from the mean directional variance (MDV) metric proposed by Ruiz-Coll et al. [25]. The MDV provides a 12-element vector for each image block, capturing local directional variance. These vectors are crucial in analyzing the texture of video blocks and provide insight into the spatial complexity of the content. To illustrate that, we have selected four 64×64 image blocks and their associated polar diagram of the MDV metric, as shown in Fig. 2. As can be seen, the MDV metric, represented as a polar diagram, perfectly captures the image directivity. For example, in Fig. 2b and f,

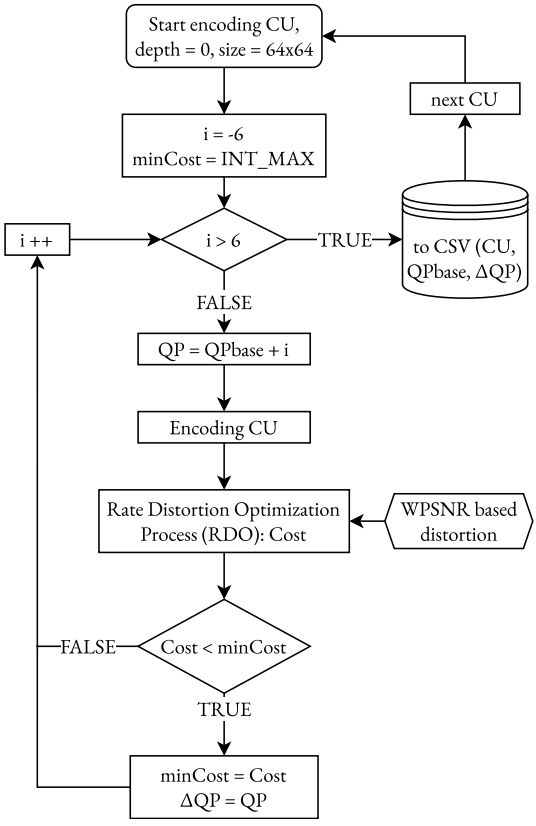

Fig. 1 Flow chart of image database extraction

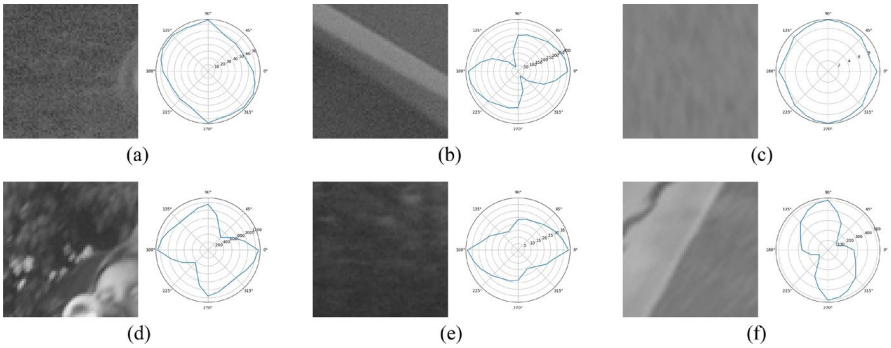

Fig. 2 Example of four 64×64 image blocks and their corresponding MDV polar diagrams. The left side of each pair shows the image block, while the right side depicts the polar diagram representing the MDV values. These diagrams illustrate the distribution of local directional variance across twelve angular segments

there is a minimum present in the direction of the most important axis in the image. However, in Fig. 2a and f, there is no minimum present, indicating that the image is either smooth or contains homogeneous texture.

Rather than using the MDV data to classify blocks into plain, edge, or texture categories, as in [26], we have extracted statistical measures from the MDV vector to enhance the learning process of our network. Specifically, for each block, we calculate the minimum, maximum, mean, and variance of the MDV vector elements. These statistical values are then stored in a CSV as additional structured inputs to the ANN. This approach allows the network to utilize richer, more informative structured data, facilitating improved predictions of QP offsets. By incorporating this additional information, the network is better equipped to converge on an optimal solution, particularly in cases where the block exhibits complex textures or directional patterns.

After processing the CSV, the dataset is stored in a pandas DataFrame (Python) with the following columns and data types:

- `QP_base` (int): Initial quantization parameter value.
- `QP_delta` (int): Optimal ΔQP value for the block.
- `BPF` (float): Number of blocks per frame. This is needed because the WPSNR metric is frame-size dependent.
- `minMDV` (float): Minimum MDV value.
- `maxMDV` (float): Maximum MDV value.
- `meanMDV` (float): Mean MDV value.
- `varMDV` (float): Variance of MDV values.
- `pix_xxxx` (int): Luminance value of the pixel xxxx.

In Table 1, we present the structured dataset corresponding to the example blocks presented in Fig. 2. These examples demonstrate the diversity in the block characteristics, such as ΔQP and MDV statistic values.

Table 1 Example structure of the dataset for 64×64 blocks from Fig. 2

Variable	Block (a)	Block (b)	Block (c)	Block (d)	Block (e)	Block (f)
QP_base	33	20	34	27	42	17
QP_delta	+5	−1	−3	−2	+4	−4
BPF	2025	2025	2025	506.25	506.25	99.04
minMDV	57.90	38.95	8.17	448.02	15.84	505.99
maxMDV	76.19	408.41	9.58	1268.48	38.72	316.07
meanMDV	67.24	275.61	8.85	949.80	25.84	18336.01
varMDV	51.87	13519.55	0.25	60428.35	47.95	
pix_0001	56	99	95	69	37	131
pix_0002	56	106	94	80	39	131
...	..	..	..	..	..	..
pix_4095	76	42	98	97	29	84
pix_4096	69	40	90	97	29	85

3.2 Hybrid neural network proposal

We propose two independent hybrid neural networks: one dedicated to processing 64×64 blocks and another for 32×32 blocks. This approach simplifies the architecture while allowing for the independent optimization of hyperparameters for each block size. The decision to implement separate networks for each block size arose from experimental results showing that resizing smaller blocks (32×32) to match the input size of the original 64×64 network yielded lower performance, both in terms of prediction accuracy and generalization. A network has not been developed for 128×128 blocks, because in the RDO process in VVC encoding, the maximum size of the transform is 64×64, and this does not allow us to obtain the optimal Delta QP as a ground truth.

Although we have used only two block sizes 64×64 and 32×32, our method may be also easily extended to 16×16 and 8×8 block sizes. However, there is a performance constraint when using lower block sizes since there would be much more coding blocks in every CTU partition that would require an extra bit rate cost to store Delta QPs of each coding block. We have carried out tests with a specific network for 16×16 blocks. However, the experimental results showed that signaling the Delta QP in the bitstream for such small blocks introduces an excessively high overhead in the bit rate, without a substantial improvement in quality.

Each developed neural network consists of two main subnetworks: a convolutional neural network (CNN) to process the pixel data and an artificial neural network (ANN) to handle the structured input data. Both subnetworks are integrated into dense layers to produce the final output. Figure 3 shows the architecture of our proposed models.

We initially explored popular pre-trained architectures, such as EfficientNet [27] and MobileNetV3-Small [28], due to their well-documented performance in image

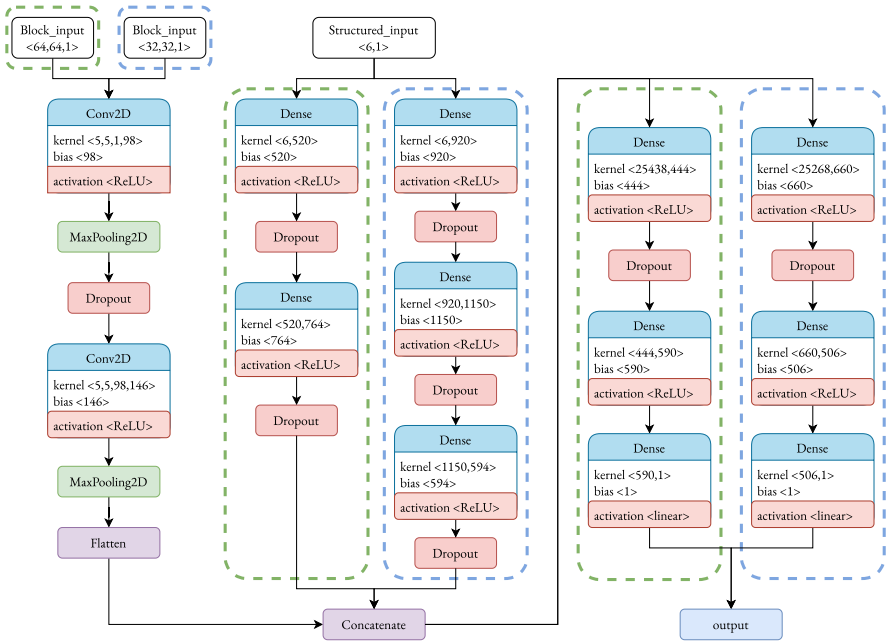

Fig. 3 Diagram of our proposed dual hybrid CNN+Ann model. The green dotted part corresponds to the neural network for blocks of size 64×64, while the blue dotted part corresponds to blocks of size 32×32. What is not dotted is common to both networks

classification tasks and their relatively low computational cost. However, these models did not provide a competitive advantage in our specialized setup, where only luminance channel blocks are processed and a large-scale pre-trained feature hierarchy is less beneficial. We adapted these networks to accept single-channel inputs, adjusted the block image sizes, and fine-tuned the final layers to predict the QP offset. Nonetheless, both EfficientNet and MobileNetV3-Small, originally developed for multi-channel color images and large-scale classification tasks, exhibited overfitting on our smaller, domain-specific dataset, leading to suboptimal performance.

We therefore conducted a direct comparison between the pre-trained architectures and our simplified two-layer CNN. Table 2 shows that our simpler approach achieves lower MSE, indicating better predictive accuracy for the specific task of QP estimation. In addition, its reduced model size provides practical advantages for

Table 2 MSE values for pre-trained CNN architectures and our simple CNN proposal for 64×64 block sizes

	EfficientNet [27]	MobileNetV3-Small [28]	Two-layer CNN (ours)
Train	3.544	3.310	2.067
Validation	4.426	4.352	2.012
Test	5.255	4.443	2.078

video coding applications, where computational overhead is a key constraint. Our simpler architecture, with just two convolutional layers, converged more efficiently to a lower validation MSE, demonstrating that a task-specific design can outperform off-the-shelf pre-trained models in specialized scenarios.

To design the architecture of our neural networks and to search for optimal hyperparameters, we utilized Keras with TensorFlow as the backend. Additionally, Keras Tuner was employed to perform an extensive hyperparameter search for both neural networks, ensuring that our model configuration was both efficient and effective. This combination allowed us to streamline the development process, leveraging the robust features of Keras and the comprehensive tuning capabilities of Keras Tuner.

The architecture for both models, designed for 64×64 and 32×32 block sizes, shares many components, with differences indicated in the diagram (Fig. 3). The parts enclosed in green dotted lines correspond to the model dedicated to 64×64 blocks, while the blue dotted lines represent the model for 32×32 blocks.

For both models, the input layers expect image blocks of shape (64, 64, 1) for the 64×64 model and (32, 32, 1) for the 32×32 model. Two convolutional layers are applied to extract spatial features from the images, followed by MaxPooling layers to reduce the dimensionality of the feature maps. Dropout layers are incorporated to prevent overfitting. The specific number of filters and kernel sizes for these layers are presented in Fig. 3.

Parallel to the image input, the structured data, which includes the $\mathrm{QP}_{\mathrm{base}}$, Blocks per Frame (BPF), and the four statistical features from the MDV vector (minimum, maximum, mean, and variance), is processed by an ANN. The input layer for the structured data has a shape of (6, 1). The structured data is passed through dense layers, which are also depicted in the figure, with Dropout layers included to further prevent overfitting.

The outputs from both the CNN and ANN are concatenated into a single vector, which then passes through several dense layers to refine the QP offset prediction. The architecture of these dense layers and their respective units is also detailed in Fig. 3. The final output layer produces a single value, $\Delta\mathrm{QP}$, using a linear activation function. This value is constrained within the range of $[-6, 6]$ to control the QP offset for the corresponding blocks.

Preprocessing is applied to the data before feeding it into the network. Image blocks are normalized by dividing each pixel value by 255, as the input bit depth at the encoder is set to 8 bits per pixel (Eq. 4). Additionally, the structured data is normalized using the StandardScaler method to standardize the features by removing the mean and scaling to unit variance (Eq. 7).

$$\mathrm{Input}_{\mathrm{pixel}} = \frac{\mathrm{Data}_{\mathrm{pixel}}}{255} \tag{4}$$

$$\mu = \frac{1}{N} \sum_{i=1}^{N} \mathrm{Data}_{\mathrm{struct}} \tag{5}$$

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\text{Data}_{\text{struct}} - \mu)^2} \tag{6}$$

$$\text{Input}_{\text{struct}} = \frac{\text{Data}_{\text{struct}} - \mu}{\sigma} \tag{7}$$

3.3 QP selection strategy: single- versus dual-model approach

In our proposed method, two configurations for QP selection are available depending on whether one or both hybrid neural network models are used. These options allow different levels of flexibility and fine-tuning during the encoding process, as described below:

3.3.1 Option A: single-model (SM)

In this scenario, the 64×64 hybrid neural network is employed to predict the optimal QP value at the CU level. The VTM defaults to a maximum CU size of 64×64 pixels.

During encoding, a CU of size 64×64 and its associated structured data are passed through the neural network model, which returns a QP offset or ΔQP. For example, let's assume the model provides a QP offset of -2 (Fig. 4a2). If the CU is subsequently partitioned into smaller units (e.g., 32×32 or smaller blocks), the QP offset remains -2 for all partitions (Fig. 4a3). Thus, the 64×64 model controls

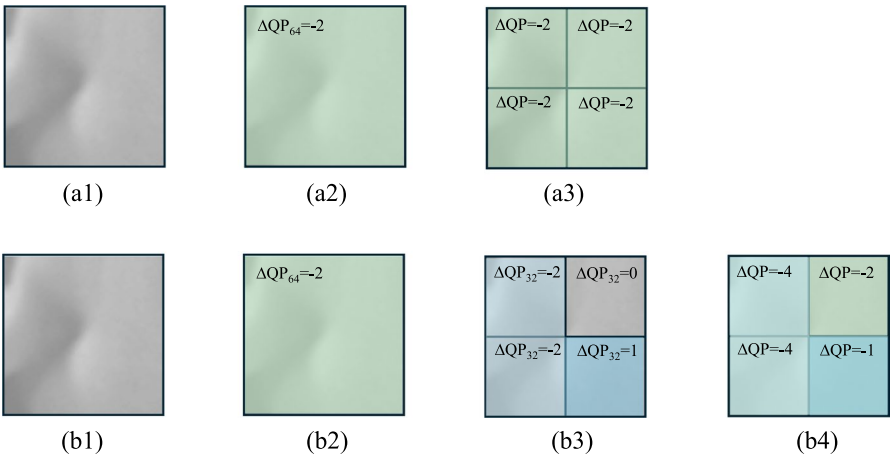

Fig. 4 Illustration of the QP offset strategies for 64×64 and 32×32 Coding Units (CUs). In (a), only the 64×64 model is used, with the same QP offset applied across the entire CU. In (b), both the 64×64 and 32×32 models are employed, allowing different QP offsets for each partitioned 32×32 block. The total QP offset is the sum of the 64×64 offset and the individual 32×32 offsets

the QP for both the large 64×64 CU and any smaller partitions that result from the division of this CU.

3.3.2 Option B: dual-model (DM)

In this approach, both hybrid neural networks are employed. The 64×64 model still determines the QP at the initial CU level, but when the CU is partitioned into smaller 32×32 blocks, the 32×32 model refines the QP selection further.

At the 64×64 CU level, the process begins similarly to option A. The QP offset from the 64×64 model is applied (Fig. 4b2). If the 64×64 CU is partitioned into four 32×32 blocks, each of these smaller CUs is then passed through the 32×32 model, which provides an additional QP offset for each block (Fig. 4b3). For example, the 32×32 model might return the following offsets for the four blocks: -2, 0, -2, 1. The final QP for each 32×32 CU is determined by adding the QP offsets from both models (Fig. 4b4).

To comply with the limitations set by the VTM, the resulting QP values are constrained within a range of *QPbase* - 6 to *QPbase* + 6. This means that for a base QP of 32, the QP values will stay between 26 and 38, maintaining perceptual quality while introducing local QP adjustments based on both the 64×64 and 32×32 models.

4 Results and discussion

In this section, we present both the performance of the proposed hybrid neural networks and their integration into the VVC reference software comparing it with respect to the native perceptual coding algorithm. In order to evaluate the performance of the proposed hybrid neural networks, we have trained, validated and tested using a collection of CUs of 32×32 and 64×64 sizes extracted from the dataset described in Sect. 3.1. About 3.5 million blocks of each size were used for training, validation, and testing of the proposed hybrid neural networks, with splits of 70%, 20%, and 10%, respectively. In order to verify the operation after integrating them into the VTM reference software, all sequences of common test conditions described for the VVC standard [24] have been used.

In addition, to better understand the dataset characteristics and the behavior of the Delta QP target variable, we analyzed its distribution for blocks of 32×32 and 64×64 sizes. As shown in Fig. 5, the majority of Delta QP values fall within the range of -4 to -1 for both block sizes. This indicates that most blocks require a reduction in QP (less compression) to achieve better perceptual rate-distortion optimization. The sharp drop-off at the extreme values (-6 and 6) highlights the constraints imposed by the VVC bitstream syntax, which limits the Delta QP range. Interestingly, a noticeable peak at $\Delta QP = 6$ is observed. This occurs when blocks are nearly perfectly predicted, resulting in a residual block of zero. In such cases, the algorithm assigns the highest possible QP value since further compression does not alter the block's content. This behavior is common in highly uniform, dark regions or synthetic/artificial images

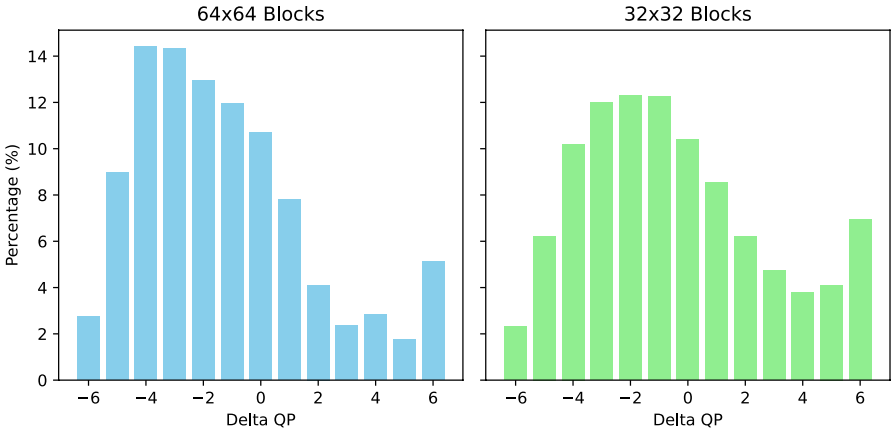

Fig. 5 Distribution of Delta QP values for the labelled datasets with 64×64 blocks and 32×32 blocks (color figure online)

4.1 Hybrid neural network performance

In Fig. 6, we show the evolution of the training and validation losses for both of our proposed hybrid models (64×64 and 32×32) over 100 training epochs. The plot provides valuable insights into each model's performance. Initially, both losses decrease rapidly, indicating effective learning and generalization of the architectures. However, around epoch 50 for the 64×64 model and epoch 75 for the 32×32 model, while the training loss continues to decrease, the validation loss stabilizes and fluctuates, suggesting potential overfitting, which may affect performance on new data.

Choosing the best epoch for each model strikes a balance between minimizing training loss and avoiding overfitting. For the 64×64 block size model, we selected epoch 53, as this point provides sufficient training without significantly compromising generalization ability. For the 32×32 block size model, epoch 74 was chosen, as it offered the best trade-off between learning and generalization for this network. This careful selection for both models ensures robust performance across unseen

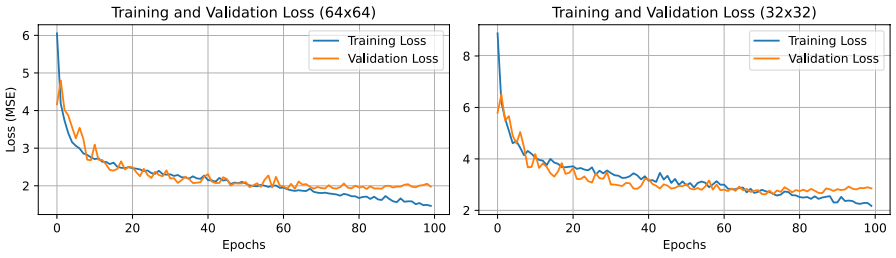

Fig. 6 Training and validation loss. The plot shows the mean squared error (MSE) loss for both training and validation datasets across 100 epochs

Table 3 Loss values for the proposed models

	64 × 64 Block size			32 × 32 Block size		
	MSE	RMSE	MAE	MSE	RMSE	MAE
Train	2.067	1.438	0.994	2.780	1.667	1.199
Validation	2.012	1.418	0.986	2.780	1.667	1.198
Test	2.078	1.442	1.002	2.764	1.662	1.198

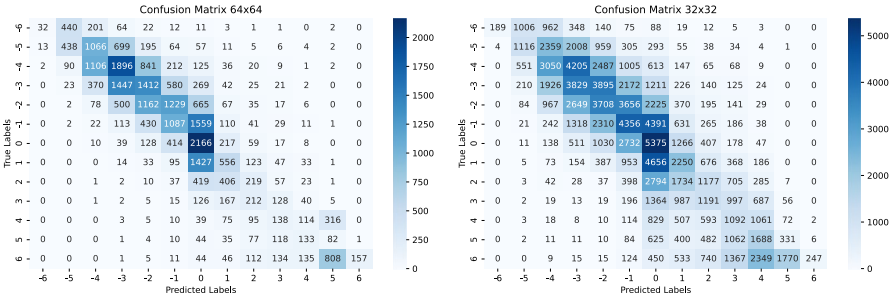

Fig. 7 Confusion matrices for the test dataset. The matrix on the left corresponds to the 64 × 64 network, while the matrix on the right corresponds to the 32 × 32 network

blocks of images, optimizing the balance between learning and generalization. As shown in Table 3, the performance metrics, such as MSE, RMSE, and MAE, for both models are practically identical in the different partitions of the data set.

The confusion matrices for the test dataset shown in Fig. 7, provide a comprehensive overview of the performance of both models in predicting the optimal ΔQP values. The diagonal elements of each matrix represent correctly predicted instances, showing that both models perform well overall. High values along the diagonal, particularly for ΔQP values like −1, 0, and 1, reflect strong performance in accurately predicting these values, indicating that the networks are proficient in predicting most of the optimal ΔQP values.

However, when comparing the two models, a notable difference can be observed. The 64 × 64 block size model exhibits a sharper diagonal, suggesting higher accuracy and fewer mispredictions compared to the 32 × 32 model, whose diagonal is slightly wider. This suggests that the 64 × 64 model makes more precise predictions, with less variance, as evidenced by fewer off-diagonal entries in the confusion matrix. The results in Table 3, where the 64 × 64 model consistently achieves lower loss values, further reinforce this observation.

Although misclassifications are inevitable given that the models operate in a regression framework rather than pure classification, these errors are generally small. In most cases, mispredictions are only one or two positions away from the correct value, demonstrating that the models' errors remain localized. This close proximity between true and predicted values even in misclassifications underscores the models' robustness, as their predictions seldom deviate far from the correct optimal QP value.

4.2 Integration of hybrid neural network in VTM reference software

Once both neural network models were trained and evaluated, they were integrated into the VVC reference software, VTM (version 17.0) [2], to perform inference on 32×32 and 64×64 CU blocks during encoding. To import the models, we used the TensorFlow C API (version 2.16.0).

Following the integration of both models, we evaluated the implementation using the sequences specified in the VTM common test conditions [24], which include video sequences of varying resolutions, from 240p to 4K.

For the comparative analysis, we selected two state-of-the-art perceptual algorithms presented in VTM, AdaptiveQP [3, 4] and QPA [5], and compared their performance to our proposed algorithms. For AdaptiveQP, we evaluated two depth levels, applying the algorithm at both the 64×64 (AQP2) and 32×32 (AQP4) block levels, to ensure a fair comparison with our two-model approach. For our algorithm, we conducted tests using both configurations: the single-model approach (SM), where only the 64×64 model is used, and the dual-model approach (DM), where the 64×64 model is combined with the 32×32 model for further partitioned blocks. This comparison allows us to directly assess the impact of utilizing different block sizes and model configurations on the coding performance.

Upon executing the tests for base QP values of 22, 27, 32, 37, and 42, we obtained the following values of the Bjøntegaard delta rate (BD-Rate) [29], comparing the results of both models against the executions that did not apply any perceptual mechanism as a reference.

Table 4 shows the results for the perceptual objective metrics. The weighted peak signal-to-noise ratio (WPSNR) is the primary metric used for training our model, as it emphasizes perceptual relevance by assigning weights to different image regions based on their visual importance. The MS-SSIM (multi-scale structural similarity) metric [30] is designed to assess structural fidelity by evaluating luminance, contrast, and structural information across multiple scales, aligning with how HVS processes images at various resolutions. PSNR-HVS-M [31], an improved version of PSNR, incorporates the contrast sensitivity function (CSF) and inter-coefficient contrast masking, which better captures the HVS's response to different spatial frequencies and textures. Finally, video multi-method assessment fusion (VMAF) [32], developed by Netflix, combines multiple perceptual quality metrics, including visual information fidelity (VIF), and machine learning models to predict subjective quality scores. Unlike the other metrics, VMAF is trained using real subjective test data, providing strong alignment with human perception, and has been shown to have a high correlation with mean opinion score (MOS) values for both HD and UHD content.

The results presented in Table 4 offer a comprehensive view of the performance of different algorithms, including QPA, AdaptiveQP at two block sizes (denoted as AQP2 for 64×64 and AQP4 for 32×32), and our proposed models (SM and DM), across various perceptual metrics and resolutions. An initial general observation reveals that AdaptiveQP falls short in most cases, while QPA exhibits strong performance across a majority of metrics and resolutions. However, when it comes to the VMAF metric, widely regarded as the most advanced and widely used perceptual

Table 4 BD-Rate results comparing default QPA and Adaptive QP algorithms (AQP2 and AQP4) with our proposed models (SM and DM) across different perceptual metrics and resolutions

Metric	Resolution	QPA [5]	AQP2 (64 × 64) [4]	AQP4 (32 × 32) [4]	SM (ours)	DM (ours)
WPSNR	240p	−9.447	3.252	2.811	−4.814	−7.495
	480p	−9.590	1.251	1.634	−7.854	−7.721
	720p	−1.864	3.003	4.137	−1.894	1.729
	1080p	−9.508	4.852	7.735	−8.313	−5.893
	2160p	−2.909	6.896	8.794	−2.515	1.437
MS-SSIM	240p	−9.597	−0.356	−1.286	−9.641	−9.535
	480p	−9.455	−4.487	−3.940	−9.104	−6.692
	720p	−9.215	−2.144	−0.591	−8.766	−5.811
	1080p	−8.927	−2.209	−1.019	−7.825	−5.383
	2160p	−1.776	−1.516	0.268	1.262	7.166
PSNR-HVS-M	240p	−7.856	3.523	3.309	−2.823	−4.954
	480p	−6.414	1.426	1.815	−5.137	−4.824
	720p	0.745	4.193	5.407	0.492	3.045
	1080p	−6.060	4.601	5.837	−4.766	−3.771
	2160p	−0.756	6.436	8.197	−0.226	3.981
VMAF	240p	−12.602	8.663	11.259	−52.470	−49.079
	480p	−0.406	5.244	7.507	−9.452	−5.045
	720p	0.896	7.644	9.698	−14.262	−12.577
	1080p	0.159	7.488	9.775	−19.436	−13.229
	2160p	1.336	9.685	12.239	−11.738	−5.336

AQP2, Adaptive QP with depth level up to 64 × 64 blocks; AQP4, adaptive QP with depth level up to 32 × 32 blocks; SM (single-model), proposed method with depth level up to 64 × 64 blocks; DM (dual-model), proposed method with depth level up to 32 × 32 blocks

metric today, our proposed algorithms outperform QPA, making this the most noteworthy result.

Our algorithm performs substantially better than AdaptiveQP across most metrics and resolutions. While AdaptiveQP produces acceptable results for MS-SSIM metric, it struggles overall, especially in WPSNR and VMAF metrics, which are more reflective of perceptual quality. The comparison between QPA and our models shows more competitive results. Although QPA generally achieves better performance in WPSNR, MS-SSIM, and PSNR-HVS-M, our algorithm significantly outperforms QPA in VMAF across all video resolutions. This is a key takeaway, as VMAF has become a standard for video quality assessment, particularly because it is machine-learning-based and trained on subjective test data, making it highly aligned with real human perception.

The comparison between our single-model approach (SM) and the dual-model approach (DM) reveals an interesting pattern. At higher resolutions (such as 1080p and 2160p), SM tends to outperform DM, suggesting that adjusting the QP at smaller block sizes (i.e., 32 × 32 blocks) is not as beneficial when the overall

sequence resolution is high. For example, in 2160p MS-SSIM, SM achieves a BD-Rate of −2.515, while DM performs worse, with a BD-Rate of 1.437, indicating a loss in performance with the dual-model approach. This suggests that at higher resolutions, the extra complexity introduced by fine-tuning QP at smaller block sizes does not translate into better performance and might, in fact, contribute to overhead in the bitstream, reducing overall efficiency.

Conversely, at lower resolutions, such as 240p and 480p, the dual-model approach (DM) performs better, as shown by its lower BD-Rate values across various metrics compared to SM. For example, in 480p WPSNR, DM achieves −7.721, compared to −7.854 for SM, indicating a slight improvement. This trend suggests that adjusting the QP at smaller block sizes is more impactful at lower resolutions, where finer granularity in QP optimization can lead to better perceptual quality.

While QPA demonstrates strong results in most perceptual metrics, particularly WPSNR, MS-SSIM, and PSNR-HVS-M, the standout finding is that our proposed models significantly outperform QPA in VMAF, the most widely recognized metric in the industry today. For instance, in 1080p VMAF, DM achieves an impressive BD-Rate of −13.229, while QPA only manages 0.159. Similarly, at 2160p, DM reaches −5.336 in VMAF, while QPA yields 1.336, again highlighting the superiority of our approach in this metric.

The significance of VMAF cannot be overstated, as it integrates various perceptual features through machine learning models, making it particularly well suited to reflect real human viewing experiences. If QPA were to consistently outperform our models across all metrics, we would not have achieved such strong results overall. However, the fact that our algorithm surpasses QPA in VMAF means that we are highly competitive in terms of perceptual quality, especially in terms of what matters most in current industry standards.

It is also important to highlight that VMAF does not perform optimally at low resolutions, such as 240p. This is evident from the relatively high BD-Rate values across all algorithms for this metric. For instance, in 240p VMAF, QPA produces a BD-Rate of −12.602, while DM reaches −49.079, and AQP4 shows 11.259, all of which indicate anomalies. The primary reason for this is that the VMAF model has been trained for 1080p sequences, and although scaling is applied during the evaluation of lower resolution sequences, the results at 240p suggest that VMAF does not provide reliable data for very low resolutions. This inconsistency must be kept in mind when interpreting VMAF results at 240p, though the overall trend across higher resolutions remains highly favorable for our algorithm.

Finally, AdaptiveQP performs poorly in most cases compared to both our algorithm and QPA. However, it does show slightly better results for the MS-SSIM metric, indicating that AdaptiveQP is more aligned with preserving structural similarity across multiple scales. For example, in 240p MS-SSIM, AQP2 achieves a BD-Rate of −0.356, which is more competitive compared to its performance in WPSNR or VMAF. Nevertheless, AdaptiveQP's performance is generally inferior, especially in VMAF and PSNR-HVS-M, where our models and QPA consistently outperform it across all resolutions.

The analysis of Table 4 demonstrates that our proposed models (SM and DM) perform substantially better than AdaptiveQP and provide competitive performance

compared to QPA, particularly when evaluated with the VMAF metric. VMAF is the most important and relevant metric in today's industry, and the fact that our models outperform QPA in this metric highlights the success of our approach. Additionally, SM performs better at higher resolutions, while DM is more effective at lower resolutions, suggesting that fine-tuning QP at smaller block sizes is more beneficial at lower resolutions. Finally, VMAF's limited accuracy at low resolutions and AdaptiveQP's poor performance overall, except for MS-SSIM, are also critical factors to consider when interpreting the results.

5 Conclusion

In this study, we proposed a dual hybrid neural network model that combines convolutional neural networks (CNNs) and structured data inputs to optimize the quantization parameter (QP) for both 64×64 and 32×32 blocks in the versatile video coding (VVC) standard. By leveraging the strengths of CNNs for spatial feature extraction and artificial neural networks (ANNs) for structured data handling, our model aimed to enhance QP prediction accuracy and improve video compression efficiency. The performance of the proposed models was evaluated using a range of perceptual metrics, including WPSNR, MS-SSIM, PSNR-HVS-M, and VMAF.

The results show that our hybrid models provide significant improvements over existing perceptual coding algorithms. In particular, our models outperform AdaptiveQP across all metrics and resolutions and demonstrate highly competitive performance compared to QPA. While QPA generally performs better on traditional metrics such as WPSNR and MS-SSIM, our models excel when evaluated using VMAF, the most advanced and widely recognized metric for perceptual video quality. This highlights the effectiveness of our approach in aligning compression optimization with real-world perceptual quality standards, as VMAF is trained on subjective data and closely correlates with human visual experience.

Further analysis indicates that the single-model approach, which uses only the 64×64 neural network, performs better at higher resolutions (1080p and 2160p). In contrast, the dual-model approach which integrates both the 64×64 and 32×32 networks, is more effective at lower resolutions (240p and 480p). This suggests that fine-tuning QP at smaller block sizes is beneficial for lower resolutions but may introduce unnecessary overhead at higher resolutions, where larger block sizes dominate and finer adjustments are less critical.

In future work, we plan to further optimize our models by exploring the impact of additional structured data inputs and advanced hyperparameter tuning techniques. We also aim to refine the dual-model approach by investigating how the model could better balance QP adjustments across different block sizes, especially at higher resolutions where overhead from smaller block adjustments could be minimized. Moreover, we will explore training the model on different video resolutions and content types to generalize the approach across a broader range of scenarios.

Acknowledgements This research was funded by MICIU/AEI/10.13039/501100011033 and by "ERDF A way of making Europe" under grant PID2021-123627OB-C55 and by the Valencian

Ministry of Innovation, Universities, Science and Digital Society (Generalitat Valenciana) under grant CIAICO/2021/278.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

References

1. ITU-T, ISO/IEC JTC 1/SC29 WG5: Versatile video coding. ITU-T Rec. H.266 version 3, 2023 (2023)
2. Chen J, Ye Y, Kim SH (2021) Algorithm description for versatile video coding and test model 11 (VTM 11). In: 20th Meeting of the Joint Video Experts Team (JVET). Doc: JCTVC-T2002
3. McCann K, Rosewarne C, Bross B, Naccari M, Sharman K (2014) High efficiency video coding (HEVC) test model 16 (HM 16) encoder description. In: 18th Meeting of the Joint Collaborative Team on Video Coding (JCT-VC), Sapporo. Document: JCTVC-R1002
4. Naccari M, Mrak M (2014) Chapter 5 - Perceptually optimized video compression. Academic Press Library in Signal Processing, vol 5, pp 155–196. Elsevier. https://doi.org/10.1016/B978-0-12-420149-1.00005-3 . https://www.sciencedirect.com/science/article/pii/B9780124201491000053
5. Helmrich CR, Bosse S, Siekmann M, Schwarz H, Marpe D, Wiegand T (2019) Perceptually optimized bit-allocation and associated distortion measure for block-based image or video coding. In: 2019 Data Compression Conference (DCC), pp 172–181. https://doi.org/10.1109/DCC.2019.00025
6. Wang G, Wang H, Li H, Yu L, Yin H, Xu H, Ye Z, Song J (2024) A survey on just noticeable distortion estimation and its applications in video coding. J Vis Commun Image Represent 98:104034. https://doi.org/10.1016/j.jvcir.2023.104034
7. Yan Y, Xiang G, Li Y, Xie X, Yan W, Bao Y (2020) Spatiotemporal perception aware quantization algorithm for video coding. In: 2020 IEEE International Conference on Multimedia and Expo (ICME), pp 1–6. https://doi.org/10.1109/ICME46284.2020.9102882
8. Xiang G, Zhang X, Huang X, Yang F, Zhu C, Jia H, Xie X (2022) Perceptual quality consistency oriented CTU level rate control for HEVC intra coding. IEEE Transact Broadcast 68(1):69–82. https://doi.org/10.1109/TBC.2021.3120916
9. Jin S, Guan X, Liu Z (2023) VVC adaptive QP offset algorithm based on visual perception. In: Wang G, Chen L (eds.) Third International Conference on Signal Image Processing and Communication (ICSIPC 2023), vol 12916, p 129161. https://doi.org/10.1117/12.3005138. International Society for Optics and Photonics, SPIE
10. Zhang M, Zhang Z, Li Y, Cheng R, Jing H, Liu Z (2024) CTU-level adaptive QP offset algorithm for V-PCC using JND and spatial complexity. IEICE Transact Fundam Electron Commun Comput Sci. https://doi.org/10.1587/transfun.2024EAL2021
11. Zhou M, Wei X, Kwong S, Jia W, Fang B (2020) Just noticeable distortion-based perceptual rate control in HEVC. IEEE Transact Image Process 29:7603–7614. https://doi.org/10.1109/TIP.2020.3004714
12. Zhou M, Wei X, Kwong S, Jia W, Fang B (2021) Rate control method based on deep reinforcement learning for dynamic video sequences in HEVC. IEEE Transact Multimed 23:1106–1121. https://doi.org/10.1109/TMM.2020.2992968
13. Zhou M, Wei X, Jia W, Kwong S (2023) Joint decision tree and visual feature rate control optimization for VVC UHD coding. IEEE Transact Image Process 32:219–234. https://doi.org/10.1109/TIP.2022.3224876

14. Zhou M, Wei X, Wang S, Kwong S, Fong C-K, Wong PHW, Yuen WYF, Gao W (2019) SSIM-based global optimization for CTU-level rate control in HEVC. IEEE Transact Multimed 21(8):1921–1933. https://doi.org/10.1109/TMM.2019.2895281
15. Wei X, Zhou M, Wang H, Yang H, Chen L, Kwong S (2024) Recent advances in rate control: from optimization to implementation and beyond. IEEE Transact Circuits Syst Video Technol 34(1):17–33. https://doi.org/10.1109/TCSVT.2023.3287561
16. Wang L (2000) Rate control for MPEG video coding. Signal Process: Image Commun 15(6):493–511. https://doi.org/10.1016/S0923-5965(99)00009-0
17. Khan IU, Ansari MA, Saeed SH, Khan K (2018) Evaluation and analysis of rate control methods for h.264/avc and mpeg-4 video codec. Int J Electr Comput Eng (IJECE) 8(2):1273–1280. https://doi.org/10.11591/ijece.v8i2.pp1273-1280
18. Li Z (2003) Adaptive basic unit layer rate control for JVT. In: JVT 7th Meeting, Pattaya, Mar2003
19. Choi H, Nam J, Yoo J, Sim D, Bajic I (2012) Rate control based on unified RQ model for HEVC. ITU-T SG16 Contribution, JCTVC-H0213, 1–13
20. Kundu D, Evans BL (2015) Full-reference visual quality assessment for synthetic images: a subjective study. In: 2015 IEEE International Conference on Image Processing (ICIP), pp 2374–2378. https://doi.org/10.1109/ICIP.2015.7351227
21. University of Southern California, Signal and Image Processing Institute: The USC-SIPI Image Database. Accessed: 2024-05-10. http://sipi.usc.edu/database/database.php
22. Asuni N, Giachetti A (2014) TESTIMAGES: a Large-scale archive for testing visual devices and basic image processing algorithms. In: Smart Tools and Apps for Graphics - Eurographics Italian Chapter Conference. https://doi.org/10.2312/stag.20141242
23. Kodak: The Kodak Color Image Dataset. Accessed: 2024-05-10. http://r0k.us/graphics/kodak/
24. Bossen F, Boyce J, Suehring K, Li X, Seregin V (2020) VTM common test conditions and software reference configurations for SDR video. In: 20th Meeting of the Joint Video Experts Team (JVET). Doc. JVET-T2010
25. Ruiz-Coll D, Fernández-Escribano G, Martínez JL, Cuenca P (2016) Fast intra mode decision algorithm based on texture orientation detection in HEVC. Signal Process: Image Commun 44:12–28. https://doi.org/10.1016/j.image.2016.03.002
26. Atencia JR, López-Granado O, Pérez Malumbres M, Martínez-Rach M, Coll DR, Fernández Escribano G, Van Wallendael G (2024) A hybrid contrast and texture masking model to boost high efficiency video coding perceptual rate-distortion performance. Electronics 13(16):3341. https://doi.org/10.3390/electronics13163341
27. Tan M, Le QV (2019) EfficientNet: rethinking model scaling for convolutional neural networks. ArXiv **abs/1905.11946**
28. Howard A, Sandler M, Chen B, Wang W, Chen L, Tan M, Chu G, Vasudevan V Zhu Y, Pang R, Adam H, Le Q (2019) Searching for MobileNetV3. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp 1314–1324. IEEE Computer Society, Los Alamitos, CA, USA. https://doi.org/10.1109/ICCV.2019.00140
29. Bjontegaard G (2001) Calculation of average PSNR differences between RD-curves. In: Proc. of the ITU-T Video Coding Experts Group - Thirteenth Meeting
30. Wang Z, Simoncelli EP, Bovik AC (2003) Multiscale structural similarity for image quality assessment. In: The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003, vol 2, pp 1398–14022. https://doi.org/10.1109/ACSSC.2003.1292216
31. Ponomarenko N, Silvestri F, Egiazarian K, Carli M, Astola J, Lukin V (2007) On between-coefficient contrast masking of dct basis functions. In: Proceedings of the Third International Workshop on Video Processing and Quality Metrics for Consumer Electronics, VPQM 2007, Scottsdale, Arizona, USA, 25-26 January p 4 (2007). Contribution: organisation=sgn,FACT1=1
32. Li Z, Aaron A (2016) Toward a practical perceptual video quality metric. https://medium.com/netflix-techblog/toward-a-practical-perceptual-video-quality-metric-653f208b9652. Accessed: 2024-09-10

Authors and Affiliations

Javier Ruiz Atencia[1] · Otoniel Mario López Granado[1] · Manuel Pérez Malumbres[1] · Miguel Onofre Martínez-Rach[1]

✉ Javier Ruiz Atencia
javier.ruiza@umh.es

Otoniel Mario López Granado
otoniel@umh.es

Manuel Pérez Malumbres
mels@umh.es

Miguel Onofre Martínez-Rach
mmrach@umh.es

[1] Computer Engineering department, Miguel Hernández University of Elche, Avd. de la Universidad s/n, 03202 Elche, Alicante, Spain