Machine Learning Based Efficient QT-MTT Partitioning Scheme for VVC Intra Encoders

Alexandre Tissier, Wassim Hamidouche, Souhaiel Belhadj Dit Mdalsi, Jarno Vanne, Franck Galpin, Daniel Menard

Machine Learning based Efficient QT-MTT Partitioning Scheme for VVC Intra Encoders

Alexandre Tissier, Wassim Hamidouche, Souhaiel Belhadj Dit Mdalsi, Jarno Vanne, Franck Galpin and Daniel Menard

***Abstract*—The next-generation Versatile Video Coding (VVC) standard introduces a new Multi-Type Tree (MTT) block partitioning structure that supports Binary-Tree (BT) and Ternary-Tree (TT) splits in both vertical and horizontal directions. This new approach leads to five possible splits at each block depth. It thereby improves the coding efficiency of VVC over that of the preceding High Efficiency Video Coding (HEVC) standard, which only supports Quad-Tree (QT) partitioning with a single split per block depth. However, MTT also has brought a considerable impact on encoder computational complexity. This paper proposes a two-stage learning-based technique to tackle the complexity overhead of MTT in VVC intra encoders. In our scheme, the input block is first processed by a Convolutional Neural Network (CNN) to predict its spatial features through a vector of probabilities describing the partition at each 4×4 edge. Subsequently, a Decision Tree (DT) model leverages this vector of spatial features to predict the most likely splits at each block. Finally, based on this prediction, only the N most likely splits are processed by the Rate-Distortion (RD) process of the encoder. In order to train our CNN and DT models on a wide range of image contents, we also propose a public VVC frame partitioning dataset based on existing image dataset encoded with the VVC reference software encoder. Our solution relying on the top-3 configuration reaches 47.4% complexity reduction for a negligible bitrate increase of 0.79%. A top-2 configuration enables a higher complexity reduction of 70.4% for 2.49% bitrate loss. These results emphasize a better trade-off between VTM intra-coding efficiency and complexity reduction compared to the state-of-the-art solutions. The source code of the proposed method and the training dataset are made publicly available at GitHub.**

***Index Terms*—VVC, MTT, complexity reduction, CNN, DT.**

I. Introduction

The emerging video formats such as 4K/8K and 360-degree videos alongside the explosion of IP video traffic [1] pushed organizations such as ISO/IEC, ITU-T Joint Video Experts Team (JVET) and Alliance for Open Media (AOM) to propose new video compression standards. AOM developed the AV1 codec released in 2018 as a successor to VP9 and JVET developed Versatile Video Coding (VVC) ITU-T H.266 | MPEG-I - Part 3 (ISO/IEC 23090-3) [2, 3] in July 2020 as a successor to High Efficiency Video Coding (HEVC).

These two standards share the same hybrid video coding structure. Therefore, during standardization, different coding tools were integrated to improve the intra and inter predictions, the in-loop filtering, or to enhance the partitioning of the block of pixels. The selection of specific coding tools leads to different coding efficiencies and computational costs. Several comparison studies were conducted between VVC and AV1 [4] based on subjective and objective quality metrics including Peak Signal-to-Noise Ratio (PSNR) and Video Multi-method Assessment Fusion (VMAF). This latter is a Machine Learning (ML)-based quality metric leveraging detail loss metric, visual information fidelity measure, and averaged temporal frame difference. Both works conclude that VVC outperforms AV1 in terms of both objective and subjective quality scores. However, the computational cost of AV1 fluctuates around the VVC complexity depending on the considered coding configuration.

The VVC reference software, named VVC Test Model (VTM), implements all normative VVC coding tools allowing rate-distortion-complexity evaluation and conformance testing. As the successor to HEVC, the VTM implementation brings 25.32% and 36.95% bitrate reductions at the expense of a significant increase in encoder computational complexity estimated to 2630% and 859% compared to the HEVC test Model (HM) 16.22 [5] in All Intra (AI) and Random Access (RA) configurations, respectively.

Compared to HEVC, which is based on a Quad-Tree (QT) block partitioning, VVC integrates a nested Multi-Type Tree (MTT) partitioning scheme allowing in addition to QT, horizontal and vertical Binary-Tree (BT) and Ternary-Tree (TT) splits [6]. This new partitioning scheme is the most efficient tool integrated in VVC [7] with 8.5% coding efficiency gain reached in RA configuration compared to HEVC. Nevertheless, this coding efficiency improvement is brought at the expense of a significant complexity increase. At each level of the hierarchical partitioning process, up to five splits are tested by the encoder, compared to the one split (i.e., QT split) for HEVC. Authors in [8] have shown that disabling BT and TT splits, decreases the encoding time by 91.7% in AI. Therefore, the partitioning process offers the highest opportunity in terms of complexity reduction compared to other coding tools. In [9], up to 97.5% complexity reduction was reported when only the optimal split was tested by the intra encoder at each level of the partitioning process, compared to an exhaustive search. To reach a real-time VVC encoding, the complexity of the QT-MTT partitioning process must be significantly reduced. This work aims to maximize the complexity reduction while minimizing the Bjøntegaard Delta Bit Rate (BD-BR) loss. A large body of literature has investigated the problem of block partitioning for HEVC [10–14] and VP9 [15]. For instance, Xu *et al.* [16] proposed a Convolutional Neural Network (CNN) to predict a hierarchical partition map to avoid exploring improbable block depths.

A. Tissier, W. Hamidouche, S. Belhadj Dit Mdalsi and D. Menard are with Univ. Rennes, INSA Rennes, CNRS, IETR - UMR 6164, 20 Avenue des Buttes de Coesmes, 35708 Rennes, France. (emails: whamidou@insa-rennes.fr and dmenard@insa-rennes.fr).

J. Vanne is with with Tampere University, Korkeakoulunkatu 10, Tampere, 33720, Finland (email: jarno.vanne@tuni.fi).

F. Galpin is with with InterDigital, Cesson-Sévigné, 35510, France.

However, those approaches cannot be directly applied to VVC as the new partitioning scheme is significantly different and more complex with the multiplication of available partitions. Therefore, predicting the optimal block depth becomes more challenging due to the high number of divisions leading to the same depth. Recently, a number of complexity reduction methods tailored for VVC have emerged through fast encoder strategies [17], probabilistic approach [18] or machine learning-based approach [19]. These techniques can leverage adaptive resolution as pre-processing [20], or the prediction of the intra perdition mode [21] and the partitioning mode [19]. Deep learning techniques [22] provided a breakthrough in video coding, especially in the complexity reduction domain. Li *et al.* [22] proposed a CNN that predicts the most probable split through a multi-classification at each block depth. One of the drawbacks of this technique is the time overhead related to the CNN inference at each depth. Indeed, the CNNs computational complexity must be carefully controlled to not annihilate the gain obtained by the complexity reduction technique. Authors in [23] have recently proposed a CNN that processes a block to predict its partition through a vector of probabilities. The average probability over the split edges is then compared to a fixed threshold to decide whether to perform the split or not at each depth. The main drawbacks of this solution are, first, its local decision, which does not leverage all probabilities of the block edges, and second, comparing the average probability of the split edges to a fixed threshold.

In this paper, we propose an efficient complexity reduction technique for the QT-MTT partitioning. Our approach combines a moderate complexity CNN that extracts spatial features of the block of pixels with multi-class classifiers to derive the best partitions for testing in the Rate-Distortion Optimization (RDO) process. A single CNN is used to process a 64×64 block and predicts the probability of each boundary of all the 4×4 pixel blocks within the input 64×64 block. At each level of the hierarchical partitioning process, a multi-class classifier is used to predict from the set of boundary probabilities the N-most likely splits to explore by the encoder. One classifier is trained for each size of the 16 different sub-blocks. At each depth, the number of explored partitions N can be adjusted from one to the total number of possible splits. Controlling the parameter N allows exploring many trade-offs between complexity reduction and quality loss. The proposed solution with top-$N = 3$ configuration reaches 47.4% complexity reduction for a negligible BD-BR loss of 0.79%, while the top-$N = 2$ configuration enables in average a complexity reduction of 70.4% for 2.49% BD-BR loss.

The rest of this paper is organized as follows. Section II introduces the partitioning background and the state-of-the-art of complexity reduction techniques. Section III describes the proposed two-stage method that combines two machine learning algorithms. The used dataset to train our models is detailed in Section IV, followed by the experimental setup and the training process for both CNN and Decision Tree (DT) models, presented in Section V. Section VI presents the performance of the two machine learning models and a comparison of our method against state-of-the-art techniques in terms of complexity reduction and BD-BR loss. The complexity of the ML techniques are assessed and analysed in Section VII. Finally, Section VIII concludes the paper.

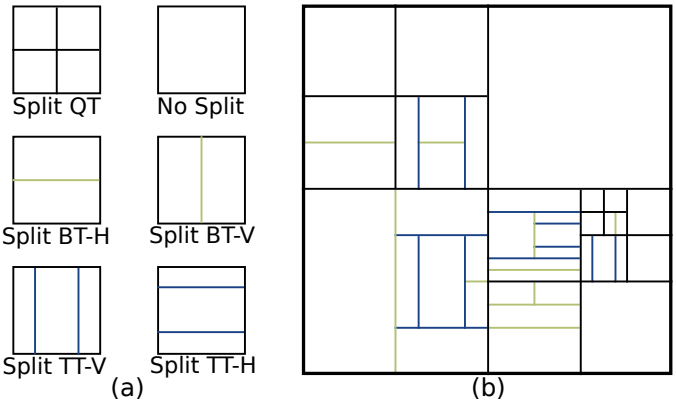

Fig. 1. CTU partitioning in VVC. (a) VVC split types. (b) Example of a CTU partition in MTT.

II. Background and Related work

In this section, we first describe the frame partitioning in the VVC standard, and then we give a brief review of complexity reduction techniques for both HEVC and VVC encoders.

A. Frame partitioning in VVC

The new block partitioning scheme proposed in VVC is the most efficient coding tool integrated into the standard [24]. The partitioning process starts from a root block named Coding Tree Unit (CTU), i.e., a block of 128×128 pixels in the VTM AI configuration. The blocks resulting from the partitioning process are named Coding Units (CUs) and may have a size between 64×64 and 4×4. Fig. 1(a) presents the split modes supported by VVC. As HEVC, QT divides a CU into four equal sub-CUs. Moreover, VVC allows a rectangular shape for CU with its novel splits BT and TT. The BT divides a CU into two sub-CUs while the TT divides a CU into three sub-CUs with the ratio 1:2:1. Both BT and TT can split a CU horizontally or vertically. Fig. 1(b) presents the splits of a CTU processed by the VTM RDO. The RDO process relies on an exhaustive search that calculates the Rate-Distortion (RD)-cost for each CU, then selects the CTU partition that results in the lowest RD-cost. In AI configuration, the VTM forces the first split of the CTU to be a QT. Additional constraints are also applied to CUs, such as QT split is not allowed after a BT or a TT split. Excluding the VTM restrictions, the encoder performs all possible splits on each CU to select the optimal partition with the lowest RD-cost.

B. Existing techniques for encoder complexity reduction

A brief review of complexity reduction methods proposed to speed up the reference software encoders of the HEVC and VVC standards is provided in this section. Authors in [9] have studied the complexity reduction opportunities in the VTM3.0 under the AI configuration. Several tools have been identified to reduce the encoding complexity, such as the

partitioning process, the intra mode prediction, and multiple transform selection. This study showed that the partitioning process has the highest impact on encoder computational complexity with 97.5% complexity reduction, followed by the intra mode prediction with 65.2% and the multiple transform selection with 55.2%. This study motivated the work of this paper that tackles the complexity reduction of the partitioning process. Meanwhile, several works have studied the complexity reduction of the intra mode prediction [25] and the multiple transform selection [26]. To reduce the complexity of the partitioning process, state-of-the-art solutions rely on different approaches to compute features relevant to this problem, including handcrafted and learning-based techniques. The computed features are then fed to a classifier such as DT, Support Vector Machine (SVM), or dense layers. It should be noted that the performance of the following techniques is provided under the AI coding configuration unless otherwise specified.

1) HEVC complexity reduction techniques: The complexity reduction of the HEVC recursive block partitioning has been widely investigated in the literature. Min *et al.* [27] defined a complexity score metric that predicts the spacial complexity of a block. The complexity score is computed as the l1 norm of the differences between the luminance pixel value at a position and the mean luminance value. This complexity score is computed for the two sub-blocks obtained by dividing the block into horizontal, vertical, or two diagonals. Then, the difference between the two complexity values of the resulting sub-blocks is computed. This value is then compared to a predefined threshold to decide whether the block should be split, not split, or undetermined. This solution reaches 52% of complexity reduction at the expense of a slight BD-BR increase of 0.8%. Correa *et al.* [10] used three sets of a decision tree trained with features such as RD-cost, gradient, the sum of pixels, or variance of pixels. These decision trees predict whether the CU, Prediction Unit (PU), or Transform Unit (TU) must be split or not. By predicting these partitioning specific to HEVC, this solution achieves a 65% of computational complexity reduction for 1.36% BD-BR loss. CNNs have already been exploited for HEVC encoding complexity reduction [11, 16]. Xu *et al.* [16] first used a CNN to predict a hierarchical CU partition map which provides an efficient representation of the CU partitioning in intra mode. A Long-Short Term Memory (LSTM) network was then integrated to predict the partition in inter prediction mode. In AI configuration, this solution reduces the complexity by 62% for 2.25% BD-BR increase. Under the low delay P coding configuration, it reaches 54.2% of computational complexity reduction for 1.5% BD-BR increase. Li *et al.* [11] claimed a real-time CTU partition prediction based on their previous proposed complexity reduction method presented in [16]. They simplified the CNN by pruning the weights at different levels to perform multiple approximations. These different configurations allow complexity control at CTU level by selecting the proper CNN. An estimation of the CNN run time is performed at the frame level to enhance the stability of complexity control. The CNN run time speed up is improved by 17 to 20 times with a complexity control error of 2%.

2) Joint Exploration Model (JEM) complexity reduction techniques: The JEM software was developed in early 2014 to study the potential coding gain behind developing a new standard with coding performance beyond HEVC. The JEM software is based on HM with new coding tools such as Multiple Transform Selection (MTS), and BT proposed to enhance the partitioning efficiency at the cost of higher computational complexity. Wang *et al.* [18] proposed a novel RD-cost estimation scheme relying on the motion divergence field. Based on the estimated RD-cost, a probabilistic framework is developed to skip unnecessary splits. The proposed algorithm reduces the complexity by 54.7% for a 1.15% BD-BR increase in RA configuration. The same authors proposed in [28] the choice of a dynamic parameter at CTU level based on neighboring partitions as the first step. In the second step, QT and BT decision tree classifiers predict the probability of the different splits to derive the proper partition. These two techniques enable 67.6% complexity reduction for a 1.34% BD-BR increase in AI configuration. Jin *et al.* [29] designed a multi-classification CNN that predicts the partition depth of a 32×32 CU. This method enables skipping unnecessary splits for the partition depth predicted outside the candidate depth range. As a result, the encoder computational complexity is reduced by 42.8% for a BD-BR increase of 0.65%.

3) VVC complexity reduction techniques: Although VVC was recently standardized, several techniques have already been proposed to tackle the problem of encoder complexity. Indeed, the extension of the partitioning process with QT, BT, and TT splits considerably increases the computational encoding complexity. Predicting a split at each CU becomes a real challenge due to the number of partition possibilities that have increased significantly compared to HEVC. This situation raises the need for a lightweight partitioning process that decreases the encoder complexity while preserving its coding efficiency for live applications. Lei *et al.* [34] proposed a two-step look-ahead prediction method for the intra prediction mode and the partitioning process. This solution computes the rate-distortion cost of only 7 intra modes out of 67, and if a CU has multiple partitions, the RD-cost of partitioned CU is computed from the parent CU. Moreover, this solution approximates the RD-costs of different partition directions in order to skip unnecessary directions. Intra mode technique combined with partitioning prediction technique allows a computational complexity reduction of 45.8% for a 1.03% BD-BR increase. Amestoy *et al.* [19] proposed a cascade framework through random forest classifiers to determine the probability of each split. To improve the accuracy of their classifiers, the impact of each feature is evaluated, such as Quantization Parameter (QP), variance, and mean of the block gradients. Furthermore, the thresholds applied to the different classifiers are optimized. A risk interval was proposed to limit the RD-cost increase by computing both splits of the classifier output when the probability falls in this risk interval. In RA configuration, this solution enables from 30.1% to 61.5% of complexity reduction for respectively 0.61% to 2.22% BD-BR increase depending on the risk interval configuration. Another cascade framework was proposed by Yang *et al.* [30] with one binary classifier for each split mode. The authors used three categories of features,

TABLE I
MAIN FEATURES OF THE EXISTING STATE-OF-THE-ART COMPLEXITY REDUCTION TECHNIQUES IN ALL INTRA CODING CONFIGURATION. CR: COMPLEXITY REDUCTION

Solution	Handcrafted	Decision Tree	Neural network	Software	CR (%)↑	BD-BR (%)↓	CR/BD-BR ratio↑
Biao *et al.* [27]	✓	✗	✗	HM 10.0	52.0%	0.80%	65.00
Xu *et al.* [16]	✗	✗	✓	HM 16.5	62.0%	2.25%	27.55
Wang *et al.* [28]	✓	✓	✗	HM 13.0 QTBT	67.6%	1.34%	50.44
Jin *et al.* [29]	✗	✗	✓	JEM 3.1	42.8%	0.65%	65.84
Yang *et al.* [30]	✓	✓	✗	VTM 2.0	52.0%	1.59%	32.70
Chen *et al.* [31]	✓	✓	✗	VTM 4.0	51.2%	1.62%	31.60
Li *et al.* [22]	✗	✗	✓	VTM 7.0	45.8%	1.32%	34.69
Zhao *et al.* [32]	✓	✗	✓	VTM 7.0	39.4%	0.87%	45.28
Saldanha *et al.* [33]	✓	✓	✗	VTM 10.0	48.8%	1.01%	48.31
Tissier *et al.* [23]	✗	✗	✓	VTM 10.2	54.0%	1.40%	38.57
Ours (Top-3)	✗	✓	✓	VTM 10.2	47.4%	0.79%	60.00

including global texture information, local texture information, and context information, as input for their classifiers. A fast intra mode decision using a one-dimensional gradient descent search is combined with the CTU structure decision. This fast partitioning solution achieves 52% of complexity reduction for a 1.59% BD-BR increase. The intra mode decision technique enhances the complexity reduction to 62.5% for 1.93% BD-BR loss. Chen *et al.* [31] also used a supervised learning method. Instead of a cascade framework, they designed six SVM models depending on the CU sizes, which are trained to skip vertical and horizontal splits. SVM features are derived from entropy, texture contrast, and Haar wavelet of the current CU. This solution reaches a 51.2% of computational complexity reduction for a 1.62% BD-BR increase. Li *et al.* [22] proposed a deep learning approach to predict the CTU partition with a binary or multi-classification at each CU depth. To train the CNN, they designed an adaptive loss function that defines penalty weights based on a split proportion that penalizes the high difference between the RD-cost for the split predicted and the minimum RD-cost of the parent CU. Based on the prediction accuracy of the CU size, an adaptive threshold is compared to the output of the CNN. This technique enables the reduction of the complexity by 45.8% for a 1.32% BD-BR increase. Another CNN model is proposed by Zhao *et al.* in [32]. First, the standard deviation of CU pixels is compared to an adaptive threshold based on QP and CU depth in order to classify a block into complex or homogeneous CU. The CUs defined as homogeneous are no more split. Second, for complex CU, a CNN is trained to predict whether or not the current CU must be early terminated. This solution achieves a 39.4% computational complexity reduction for a 0.87% BD-BR increase. Saldanha *et al.* [33] presented a configurable partitioning decision using a LightGBM (LGBM) model. Five LGBM binary classifiers are trained offline, exploiting hand-crafted features such as QP, width, or variance. The classifiers predict a split probability which is compared to a threshold to skip the split. This technique obtains several trade-offs with a complexity reduction from 35.2% to 61.3% and a BD-BR increase from 0.46% to 2.43%.

Table I summarizes the features and performance of the VVC complexity reduction techniques mentioned above. Compared to the state-of-the-art solutions, the contributions of our paper are summarized as follows: 1) Training decision tree classifiers by CU size instead of performing a simple thresholding decision. 2) Dataset balancing and enhancing with integrating more screen content video sequences. 3) Assess the complexity of the CNN and DT models on different CPU and GPU platforms as a first step toward integrating our model in professional VVC encoders. The proposed solution outperforms state-of-the-art methods enabling the highest ratio between complexity reduction and BD-BR loss.

III. PROPOSED TWO-STAGE VVC PARTITION PREDICTION SOLUTION

As described in Section I, VVC introduces BT and TT splits at the cost of a high increase in computational complexity. In addition, this recursive partitioning process computes the rate-distortion cost for a set of coding tools for each CU. The proposed partitioning prediction technique is composed of a CNN for spatial features extraction followed by a set of multi-class classifiers based on DTs to predict the appropriate partitioning decision at different depths of the partitioning tree. Fig. 2 illustrates the flow diagram of the proposed partitioning prediction solution.

A. Overall presentation

Our method is a top-down solution that early terminates unlikely splits, i.e., it follows the hierarchical process of the VVC encoder and skips split modes that have a low probability of belonging to the optimal partitioning. Our complexity reduction method comprises two components: spatial features extraction and DT classifiers. The features extraction block consists of a CNN that processes an input luminance block $\mathbf{B}$ to predict a vector of probabilities $\tilde{\mathbf{p}}$ of splits at the 4×4 block edges:

$$\tilde{\mathbf{p}} = f_\theta(\mathbf{B}). \quad (1)$$

where f_θ is a parametric function of the CNN with trainable parameters θ and $\mathbf{B}$ is the input block of size 68×68.

The block $\mathbf{B}$ consists of the current CU of size 64×64 padded with four rows and four columns on the top and left of the block resulting in a block size of 68×68. The left and top neighbor pixels are used as reference samples in the intra prediction through the Multiple Reference Line (MRL) intra prediction [35]. Indeed, their pixels are used to derive the intra mode of the current block. Therefore, it is necessary to

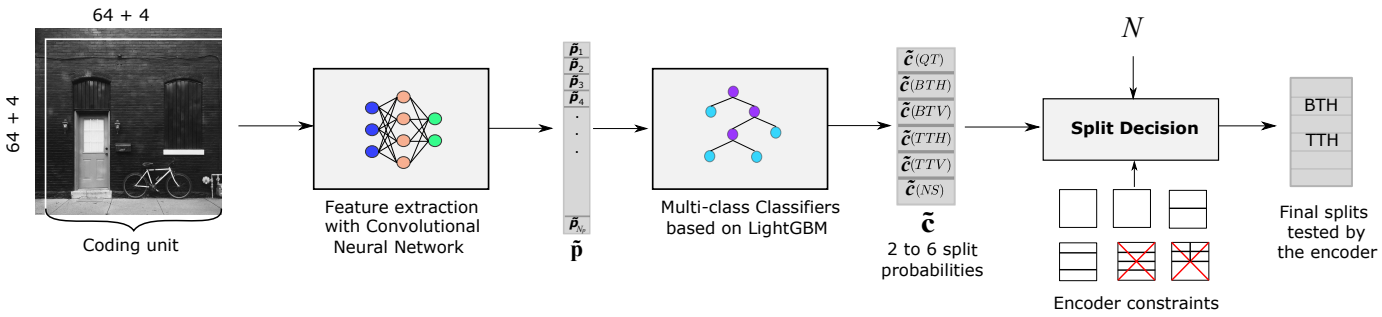

Fig. 2. Workflow diagram of the proposed block partitioning scheme for glsvvc in AI coding configuration. A CNN first processes the input luminance block to predict $\tilde{\mathbf{p}}$, a vector of N_p probabilities describing all edges at each 4×4 sub-block. The vector $\tilde{\mathbf{p}}$ is then processed by a decision tree LightGBM to predict the probabilities of the six partitioning modes through the vector $\tilde{\mathbf{c}}$. The top-N splits with highest probabilities are tested by the RD process of the encoder to select the optimal split in terms of RD-cost.

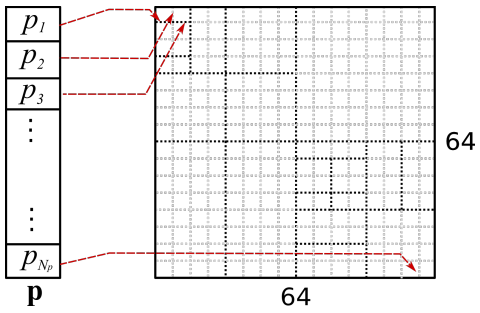

Fig. 3. Correspondence between the CNN output vector and a 64×64 CU partition.

include these samples in the features extraction stage, i.e., the CNN. The CNN predicts for each $N_B \times N_B$ CU a vector of N_p probabilities of a split at each 4×4 edge of the block. The length N_p of the predicted probabilities vector $\tilde{\mathbf{p}}$ is computed as follows:

$$N_p = \frac{N_B}{2}\left(\frac{N_B}{4} - 1\right). \tag{2}$$

In the case of CU size of 64×64, N_p is the length of the vector $\tilde{\mathbf{p}}$ which is equal to 480. Fig. 3 presents the connections between the probability vector components and the 480 4×4 edges of an input CU. As this figure highlights, the first $\mathbf{p}_1$ and the third $\mathbf{p}_3$ components of the vector correspond to the horizontal bottom edges of the first and second 4×4 blocks, respectively. For an accurate partitioning prediction, these two components must have a value (probability) close to 1 as a TT split is performed to split the sub-block. The last component of the vector $\mathbf{p}_{N_p}$ has a value (probability) close to 0 since no split is performed at this final @vertical edge of the CU as illustrated in Fig. 3.

The classifiers are then fed with the probabilities vector $\tilde{\mathbf{p}}$ derived from the CNN to predict the split decision to perform at each tree depth

$$\tilde{\mathbf{c}} = g_{\omega_i}(\tilde{\mathbf{p}}), \quad \forall i \in \{1, \ldots, M\}, \tag{3}$$

where g_{ω_i} is a parametric function of classifier i, ω_i is its trainable parameters, and M is the number of considered classifiers.

A separate multi-class classifier based on DT model is applied for each CU size. The classifier takes as an input the probability vector $\tilde{\mathbf{p}}$ and predicts a vector $\tilde{\mathbf{c}}$ of six probabilities that correspond to the six possible split modes performed by the VVC encoder at each tree depth. This approach results in M separate DT models that are trained separately to enhance the prediction performance and enable better model convergence the with a reduced number of trainable parameters. Once the split probabilities are derived for the CU, a selection of the highest probabilities is made based on the selected configuration. The configuration specifies the N value, which defines the number of tested splits by the encoder. Thus, the encoder tests only the N first splits corresponding to the highest probabilities (Top-N) predicted by the DT model. The encoder then skips the remaining splits with lower probabilities than the top-N candidates to reduce the encoding computational complexity. The proposed spatial features extraction CNN model and the DT classifiers are described in more detail in the following two sections.

B. Spatial features extraction CNN model

Fig. 4 presents the adopted CNN architecture, which is inspired by the ResNet network [36]. The orange layers represent convolution layers with 3×3 kernel (Conv 3×3), whereas the yellow layers denote convolutions with 1×1 kernel (Conv 1×1) that transform the input feature map matrix to match the dimension of the next layer which are then summed up (green plus). The red layers correspond to the max pooling (Max Pool) that subsample the input features map with a window of 2×2 by selecting the maximum over four values. The last layer in purple is a fully connected layer (Dense) that predicts the 480 components vector. The sigmoid activation function is used to predict the output values within the interval $[0, 1]$. All these layers result in a network with 226,088 trainable parameters. The input consists of 68×68 luminance pixels of the CU currently processed plus the QP value that highly influences the final partition. The QP is provided as an external input to the last fully connected layer. The QP parameter is crucial for saving the memory bandwidth as it leads to only one model shared for all QP values instead of adopting one model by QP value. Therefore, our model can be efficiently used with a rate control mechanism that adapts the QP value to the target bandwidth. The output is a 480 components vector representing the fine-grain partitions (4×4) of the 64×64

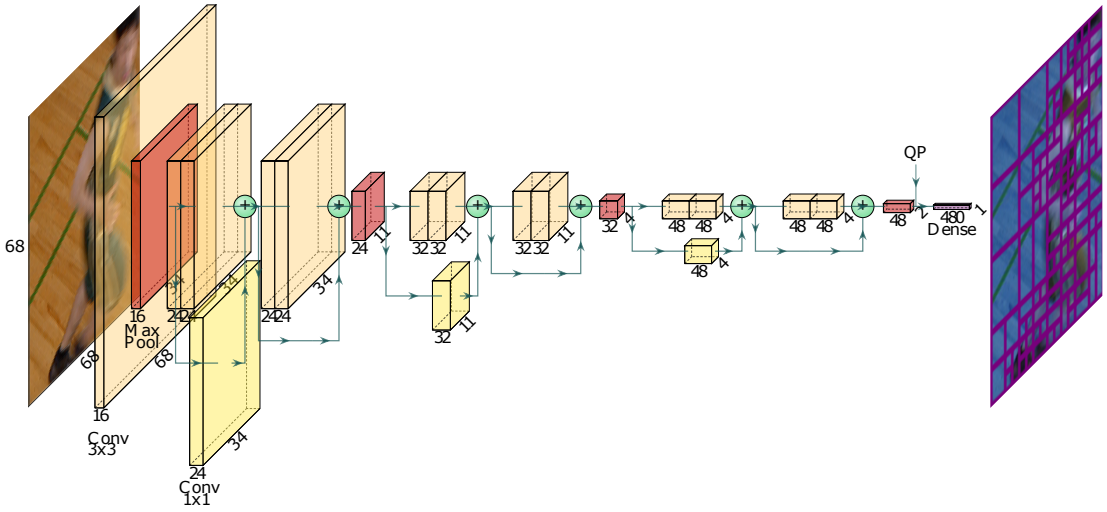

Fig. 4. The CNN architecture with convolution layers in orange and yellow, max-pooling layers in red and fully connected layer in purple.

CU. This solution has the advantage of predicting the whole CU spatial features in one shot, which is very convenient for reducing the complexity overhead and latency introduced by this step. Moreover, the architecture of the network is less deep than state of the art CNN architectures [37, 38] and thus will require less time to predict the output vector.

C. Multi-class classifiers based on DT models

The decision approach adopted in our previous work [23] relied only on the probability at the spatial location of a specific split. The pixel level characteristics of the input CU are well exploited by the CNN and are represented in the output vector $\tilde{\mathbf{p}}$. Nevertheless, the partitioning decision considered in [23] uses only information in the split edges (local), while the DT may benefit from information of all edges in the CU (the whole vector of probabilities, i.e., global). The maximum convolution kernel size may limit the CNN extracting these global features in a CU.

The vector of probabilities $\mathbf{p}$ can be considered as spatial features relevant to the block partitioning process. Therefore, the CNN inference is carried out once for each CU of size 64×64 to predict the corresponding probability vector $\mathbf{p}$, then a specific DT model processes this vector at each level of the partitioning tree to derive a set of N more likely splits to explore by the encoder. To predict split probabilities at various CU depths, we consider multiple models covering all possible sizes of the rectangular sub-blocks from 64×64 to 4×4 excluded. Table II illustrates that the sixteen CU sizes can be further split into two to six different partition modes, including the no split mode. The DT model is fed with the probability vector $\tilde{\mathbf{p}}$ and the QP value. The probability vector is then cropped into a sub-vector that includes only the probabilities of the sub-block edges. The DT model performs a multi-class classification by predicting a probability vector $\tilde{\mathbf{c}}$ of six components corresponding to the probabilities of the six possible splits. Therefore, the probabilities of non-possible splits are set to zero during the training process.

Several machine learning models were tested to solve this multi-class classification problem including DT, random forest, SVM with different kernels, and LGBM model [39]. However, this latter was considered based on its excellent classification performance and low complexity at both training and inference stages.

TABLE II
POSSIBLE SPLITS ACCORDING TO THE CU SIZE

Height \ Width	64	32	16	8	4
64	QT	-			
32	-	All	BT, TT	BT	BTH
16		BT, TT	All	TTH	TTH
8		BT, TTV		BT	BTH
4		BTV, TTV		BTV	-

IV. PROPOSED DATASET FOR TRAINING

In this section, we present the dataset in its soft and hard representations and its balancing process.

A. Dataset for the learning process

The lack of a public dataset providing encoded blocks with the VTM, and their corresponding partitions drives us to construct our training dataset to optimize the proposed models' weights. Our work focuses on AI configuration, thus the temporal relationship between adjacent frames is not considered. Five public image datasets were selected including Div2k [40], 4K images [41], jpeg-ai [42], HDR Google [43] and flickr2k [44]. The resulting dataset presents a high diversity of still image contents. However, since these datasets

TABLE III
BREAKDOWN OF OUR DATASET BY RESOLUTION.

Resolution	240p	480p	720p	1080p	4K	8K	Total
Nb images	500	500	579	2557	654	418	5208

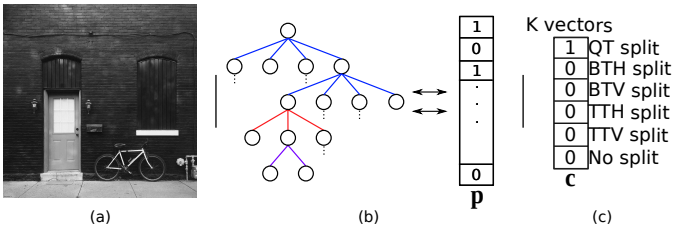

Fig. 5. Representation of our dataset. (a) is a luminance 68×68 input block $\mathbf{B}$. (b) is the optimal partitioning of the block represented by a tree and transformed to the soft representation, i.e. a 480 elements vector $\mathbf{p}$. (c) is the hard representation, K vectors $\mathbf{c}$ of 6 probabilities, that define the optimal split for each of the K blocks in the tree. K is the decision tree depth.

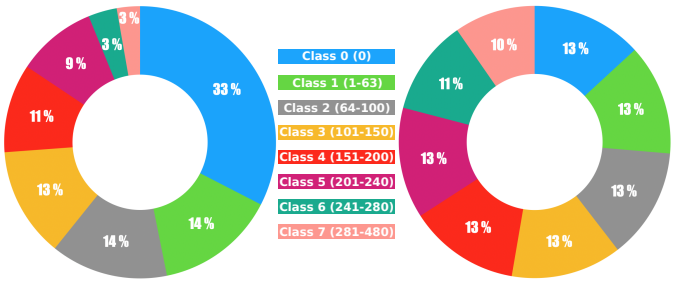

Fig. 6. Breakdown of our dataset by partition classes. (a) Unbalanced dataset. (b) Balanced dataset.

include more images in high-resolution (Full HD and 4K resolutions), a set of high-resolution images are downscaled to lower resolutions with a bilinear filter and added to the dataset. The resulting dataset includes around 5208 images at different resolutions as detailed in Table III, which gives the number of pictures per resolution. Images of similar resolution are then concatenated to build a pseudo-video sequence. This latter is encoded with the VTM encoder in AI configuration at different QP values, QPs $\in \{22, 27, 32, 37\}$.

It should be noted that the VTM encoder includes multiple speed-up techniques, reducing the complexity brought by the VVC partitioning process [17]. However, to achieve a high coding efficiency by testing more partitioning configurations, these speed-up techniques have been disabled to build our dataset. Compared to the VTM anchor, disabling these speed-up techniques leads to more accurate partitioning configurations, enhancing the coding efficiency. Nevertheless, a higher encoding time is needed to create the ground truth, but only one encoding pass is required, so increasing the encoding time is not critical at this stage. The VTM in AI configuration relies on the dual-tree tool that performs separate partitioning for luminance and chrominance components. The partitioning information of both components is recorded, while only the prediction of luminance partitioning is considered in this paper since it takes most of the encoding complexity with more than 85% of the total encoding time [8].

The optimal partitions computed by the VTM encoder is first saved as a tree. Therefore, to integrate the dataset into our proposed method, two data representations are defined as soft and hard representations, as illustrated in Fig. 5. Fig. 5-(a) shows a block $\mathbf{B}$ processed by the VTM encoder which is a 64×64 luminance block plus 4 rows and columns on top and left of the block. These extra pixels are required for the MRL intra prediction tool. Fig. 5-(b) illustrates the partition tree of the this block which is converted to a 480 components vector. This soft representation of the tree depicts the 64×64 block luminance partition in 4×4 blocks in a single vector $\mathbf{p}$.

The hard representation of our dataset is a succession of K vectors that defines all splits at each depth. Fig. 5-(c) presents one of those vector $\mathbf{c}$ which defines the optimal split for a specific CU size. The vector size is set to six as the maximum number of splits defined by VVC. For instance, the 64×64 CU size has only two possible splits with QT and no split, and thus the four remaining components of the vector are set to zero. Instead, for a CU size of 32×32, all splits are available, so the vector size is 6 with QT, the two BTs, the two TTs and no split.

B. Dataset processing

Our dataset includes more than 26 million instances of 64×64 block partition, excluding any Common Test Conditions (CTC) sequence to ensure a fair comparison of our method against state-of-the-art techniques. To analyze the dataset disparity, we first randomly select 2 million instances. Furthermore, to address the data heterogeneity issue, several classes are defined with eight levels of depth partition, from no partition to deep partition. Fig. 6 gives the number of instances in the dataset at each of these eight depth levels, which are defined based on the number of edges activated in the 64×64 CUs. Fig. 6-(a) represents the class repetitions through 2 million instances before the dataset balancing. It can be noticed that class 0 without any split contains more than a third of the 2 million 64×64 CUs instances. In contrast, the last two classes, which illustrate deep partition, are under-represented and must be increased.

Fig. 6-(b) illustrates the depth distribution of block partitioning over the eight classes after the dataset balancing. 1,200,000 instances (64×64 block) were selected at each QP value with a homogeneous representation over the eight classes. The two last classes are slightly under-represented as highly deep partitions are derived by the encoder only for extremely complex blocks encoded at low QP values.

The hard representation is composed of sub-datasets separated by the different CU sizes. These sub-datasets are also balanced to enhance their representation. Indeed, we individually balanced each sub-dataset over the available splits and the QPs. For instance, the 4×16 CU size dataset is balanced between the proportion of BTH, TTH and no split but also among the 4 QPs.

V. Training Process and Experimental Setup

A. CNN model

The CNN is trained from scratch with the proposed dataset described in Section IV, relying on the Keras framework [45] running on top of the Tensorflow module [46]. The weights θ of the CNN are updated at each batch iteration with the ADAM stochastic gradient descent optimizer [47]. The loss function is defined to optimize those weights by minimizing the mean

squared error between the predicted probability vector $\tilde{\mathbf{p}}$ and the corresponding ground truth vector $\mathbf{p}$ as follows:

$$\mathcal{L}_{cnn} = ||\mathbf{p} - \tilde{\mathbf{p}}||_2^2. \quad (4)$$

The batch size is set to 128 instances (64×64 CUs) and the learning rate is equal to 10^{-3}. The training is performed on a hundred epochs with a random shuffle of the dataset at each epoch. The CNN training was carried out on a RTX 2080 Ti Graphics Processing Unit (GPU).

B. DT LGBM model

The DT models are implemented under the LGBM framework [39] version 2.3.1. This latter is a gradient boosting framework based on a decision tree developed by Microsoft. LGBM has many advantages, such as low memory usage, the capacity to handle large-scale data, and low inference computational time. This last advantage is essential for our problem as the inference is carried-out at each CU level.

LGBM is a DT method that sums the predictions of all the trees to reach high accuracy. The trees are optimized in a stage-wise way by adding or updating a new tree based on the error of the whole ensemble learned so far. For DT classification, the used cross-entropy loss function is defined as follows:

$$\mathcal{L}_{dt} = -\sum_{i=1}^{6} \mathbf{c}_i \log \tilde{\mathbf{c}}_i, \quad (5)$$

where c_i is the vector of ground truth split probabilities and $\tilde{c}_i$ is the vector of split probabilities predicted by the model.

C. Evaluation procedure and implementation details

All experiments are conducted with the VVC Test Model (VTM) version 10.2 in AI coding configuration. We consider test video sequences defined in the VVC Common Test Conditions (CTC) [48]. The CTC video sequences are separated into seven classes as follows: A1 (3840×2160), A2 (3840×2160), B (1920×1080), C (832×480), D (416×240), E (1280×720), and F (832×480 to 1920×1080). These video sequences are encoded at four QP values: 22, 27, 32, and 37.

The proposed solution is assessed in terms of coding efficiency and computational complexity. The coding efficiency is measured with the BD-BR metric [49] that computes the bitrate loss over four QPs in percentage with respect to the anchor (i.e., VTM10.2) for the same PSNR objective quality. The BD-BR is calculated across the three components, Y, U, and V. The computational complexity reduction compared to the anchor is assessed by computing the Δ Encoding Time (ΔET) as follows:

$$\Delta ET = \frac{1}{4} \sum_{QP_i \in \{22,27,32,37\}} \frac{T_R(QP_i) - T_C(QP_i)}{T_R(QP_i)}, \quad (6)$$

where T_R is the reference encoding time of the VTM anchor, and T_C is the encoding time of the VTM with the proposed complexity reduction method.

Our complexity reduction technique was integrated into the VTM10.2 encoder, which is developed in C++ programming language. The CNN is built and trained with Python under the Keras framework, and then the model is converted to C++ code with the frugally deep framework [50]. The DT models are also trained in Python, and then converted to C++ with the LGBM framework [39].

All encoding operations were carried out sequentially on an Intel Xeon E5-2603 v4 processor running at 1.70 GHz under Ubuntu 16.04.5 operating system (OS).

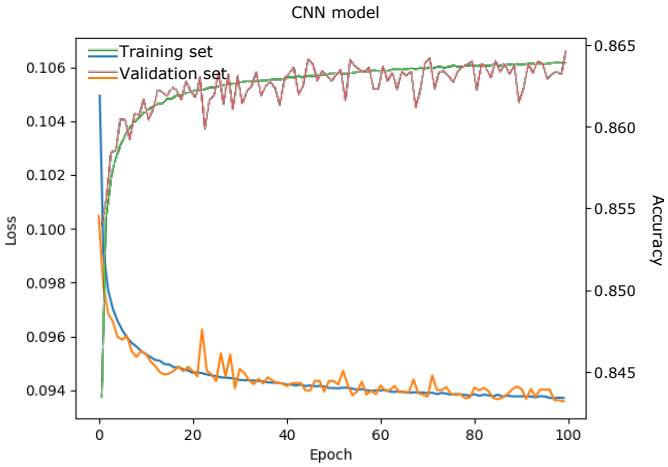

Fig. 7. Decreasing loss (in X) and increasing accuracy (in Y) as a function of epoch for the training and validation set.

VI. Experimental results

In this section, the performance of the proposed method is assessed and analysed. The CNN performance is analysed through its accuracy and Receiver Operating Characteristic (ROC) curves. The multi-class DT classifiers' accuracy is presented through its top-N accuracy. The complexity reduction proposed method is then assessed in terms of both computational complexity reduction and BD-BR loss compared to the VTM. Multiple configurations of our method are explored depending on the tested top-N DT LGBM output splits. Several existing techniques have investigated the complexity reduction of the new MTT partitioning. We compare our proposed solution with four best-performing state-of-the-art methods, including solutions proposed by Saldanha *et al.* [33], Chen *et al.* [31], Yang *et al.* [30], and Li *et al.* [22]. Finally, we assess the inference overheads of both CNN and DT models.

A. CNN performance

Fig. 7 shows the loss and accuracy of the CNN model versus the training epochs on both training and validation datasets. The blue and orange curves correspond to the losses computed by Eq. (4) over a hundred training epochs. The green and red curves represent the binary accuracy computed after a shrinkage with a fixed threshold of 0.5. The accuracy is computed between the ground truth and the predicted vector of probabilities. The two curves in blue and green are from the training set, and those in orange and red are from the validation set. The goal of the training is to update the model weights to minimize the loss at each iteration. We notice that the loss

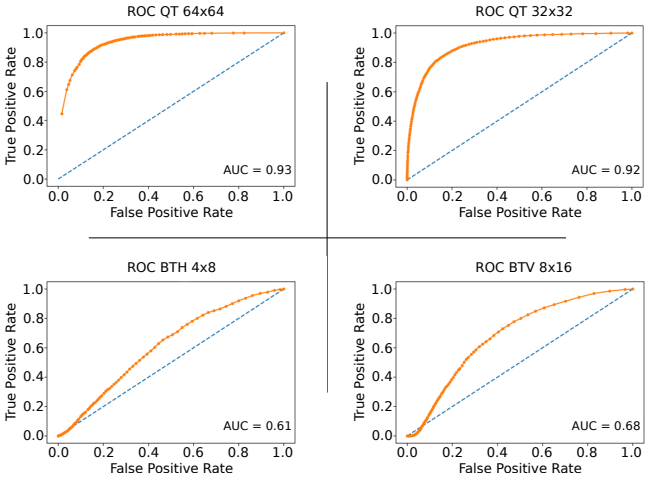

Fig. 8. Several CNN ROCs for different splits and sizes on the CTC video sequences.

curves decrease from 0.105 to 0.094 over 100 epochs for both training and validation sets. The validation curve follows the training curve, which means the model generalizes well on the validation set. The model accuracy reaches more than 86% of true prediction, i.e., 86% of the estimated probabilities with a threshold of 0.5 are equal to the ground truth.

The CNN prediction performance is also analyzed under VTM with the ROC curves. The ROC curves represent the true positive rate versus the false positive rate. The split probabilities required to plot these ROCs are determined by averaging all probabilities at the exact spatial position of the split. Fig. 8 presents the ROC curves of QT, BTH and BTV splits at different CU sizes computed on the CTC sequences. The QT ROC curves show that the average probability is able to reach 0.8 of true positive for approximately false positive rate of 0.1, for both 64×64 and 32×32 CU sizes. Instead, the BT ROC curves are closer to the random guess curve, which is the diagonal one in blue. Indeed, the CNN pays more attention to those probabilities. Moreover, the high CU sizes more easily determine the split choice as it concerns many pixels. The ROC area under the curve values are given as a single score to compare the results of the graphs. As shown by the curves, the ROC area under the curve confirms the results by reaching scores higher than 0.9 for the QT and less than 0.7 for the BT splits.

Visual illustrations of 64×64 CU partitions from ground truth and CNN prediction are given on the first and second rows of Fig. 9, respectively. On the top row, the optimal partitions, derived by the VTM anchor, are the ground truth. On the bottom row, the partitions predicted by the CNN are displayed with color codes based on their probabilities. The color code varies from red to blue with a probability ranging between 0.3 and 1, and the gray color represents probabilities below 0.3. The partitions for QP 37 is shallow with a maximum of three depths and two CUs of size 32×32. Its CNN prediction is relevant, with each edge defined as a split with a probability higher than 0.9. For the other edges that are not defined as a split, the probability values are lower than 0.3 except for six edges predicted with probabilities between 0.3 and 0.4. Compared to the QP 37 partitions, the QP 27 configuration is partitioned deeper and is harder to predict efficiently. The figure shows that each edge determined as a split has a probability higher than 0.3 except the TT split, and some CUs are predicted deeper.

The CNN output is not directly used in the VTM. It is considered as spatial features and used as an input for the DT LGBM models. The DT LGBM will benefit from all the probabilities that impact the partition.

TABLE IV
PERFORMANCE THE DT LGBM MODELS WITH THEIR DEFINED SIZE AND NUMBER OF OUTPUT THROUGH TOP-1, TOP-2 AND TOP-3 ACCURACY COMPUTED ON THE VALIDATION SET.

Width	Height	#classes	Top-1 acc.	Top-2 acc.	Top-3 acc.
64	64	2	91.69%	-	-
32	32	6	59.94%	78.38%	88.87%
32	16	5	58.50%	77.96%	89.85%
16	32	5	56.55%	77.48%	89.24%
32	8	4	54.80%	80.22%	94.11%
8	32	4	54.91%	80.45%	94.40%
32	4	3	66.64%	87.80%	-
4	32	3	68.80%	87.38%	-
16	16	6	50.95%	71.27%	84.60%
16	8	4	62.89%	82.74%	94.05%
8	16	4	62.36%	83.25%	94.39%
16	4	3	68.96%	90.12%	-
4	16	3	68.95%	88.54%	-
8	8	3	81.46%	93.05%	-
8	4	2	82.16%	-	-
4	8	2	86.26%	-	-
Average		2	86.70%	-	-
		3	70.96%	89.38%	-
		4	58.74%	81.67%	94.24%
		5	57.53%	77.72%	89.55%
		6	55.45%	74.83%	86.74%
		-	67.24%	82.97%	91.19%

B. DT LGBM performance

The second part of the proposed solution includes DT LGBM models that take advantage of all spatial features predicted by the CNN to derive split probabilities. As presented in Section IV, multiple models are trained to handle the different CU sizes. The input element range starts from one probability plus the QP for 4×8 or 8×4 CU sizes to 480 probabilities plus the QP for 64×64 CU size. The output is a vector of six classes with QT, the two BTs, the two TTs and no split. For CUs with fewer splits available due to VTM restrictions, a mask is applied on the predicted vector to restrict the unsuited splits.

Table IV presents the accuracy of the DT LGBM models on the CTC dataset. Three values are reported to analyse the DT LGBM results based on the top-N accuracy with $N \in \{1, 2, 3\}$. Top-N accuracy is a metric that measures how often the correct class falls in the top-N highest predicted probabilities. Top-1 accuracy reaches, on average, 67.24% of correct predictions. Based on the number of available splits, the prediction accuracy is different, decreasing from 86.7% to 55.45% for 2 and 6 classes, respectively. The highest top-1 accuracy is achieved with the 64×64 block size binary classification between QT and no split with 91.69%. All DT

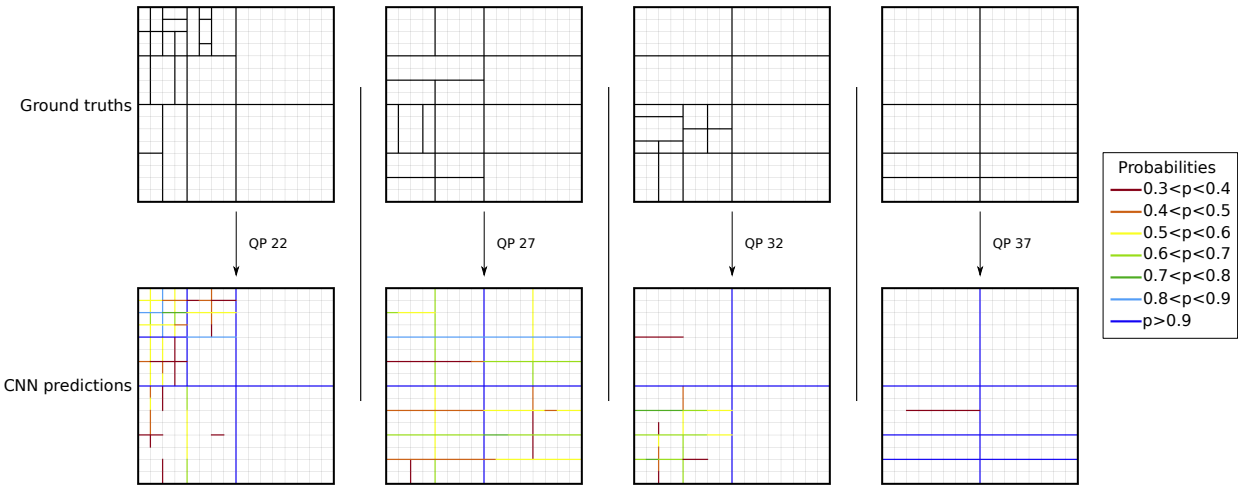

Fig. 9. Ground truth partitions on top and their corresponding CNN predictions on the bottom for the sequence MarketPlace with QPs 22, 27, 32 and 37.

LGBM models have more than 50% top-1 accuracy even with multi-class classification. The top-2 accuracy is given for models with at least three classes, which reaches, on average, 82.97%. The lowest accuracy is 71.27% obtained with the smallest block size available for six classes decision, i.e., 16×16. Finally, top-3 accuracy achieves 91.19% accuracy on average over the model with at least four classes. By skipping half of the available splits, the top-3 configuration achieves 86.74% accuracy for six classes. This accuracy performance reduces the complexity by a factor of two, with high confidence in predicting the right split. These results exhibit that the accuracy of DT models with three classes depends on the available splits. Higher accuracies are reached for the 8×8 block size model in top-1 and top-2 by at least +12% and +3%, respectively, compared with other models with three classes. In addition to the no split mode, two splits are in competition: either BT and TT in the same direction or BT horizontal and vertical. As shown by the results, the direction of a split is easier to predict than the difference between BT and TT in the same direction.

C. Complexity reduction performance under the VTM

The two-stage proposed model is integrated in the VTM10.2 encoder configured in AI setting. For a fair comparison, the reference used to compute the BD-BR loss and the complexity reduction is the classical VTM encoder, including multiple native speed-up techniques [17] for the tree partitioning process. These speed-up techniques significantly reduce the execution time with a slight BD-BR degradation compared with an exhaustive RDO process. Thus, these experiments exhibit the gain provided by our approach compared with the common configuration of the VTM encoder.

We compare our method to the best-performing state-of-the-art methods in complexity reduction and BD-BR loss. These state-of-the-art methods rely on different VTM versions; meanwhile, the tree partitioning tool has not significantly changed during the standardization.

Table V presents the BD-BR and complexity reduction performance of our method compared with the state-of-the-art techniques proposed by Saldanha *et al.* [33], Chen *et al.* [31], and Li *et al.* [22]. In this table, two configurations are presented based on the top-2 and top-3 configurations. The results are illustrated for the CTC sequences from class A1 to class F. The average is given for each class independently, and all the sequences from class A1 to class E. However, class F is not considered in the average computation as it includes specific video sequences such as screen content.

On average, our top-3 configuration through all sequences reaches 0.79% BD-BR loss for a complexity reduction of 47.4%. Compared with Li *et al.* [22] method, our solution achieves better performance in both BD-BR and complexity reduction. Chen *et al.* [31] solution reduces 4.2% more complexity but at the expense of a significant increase in BD-BR loss of 0.83% compared with our top-3 configuration. This low gain in the complexity reduction is not relevant compared with the loss in BD-BR, which is almost doubled. The method proposed by Saldanha *et al.* [33] enhances the complexity reduction by 1.5% but for a 0.22% more BD-BR loss. The second configuration with top-2 achieves, on average, 2.49% BD-BR loss with 70.4% of complexity reduction.

High-resolution videos are the main interest in the VVC development. Tackling the complexity reduction of high-resolution videos is particularly important as the encoding time is proportional to the sequence resolution. Classes A and B represent high-resolution sequences with 4K and full HD, respectively. Our top-3 and top-2 configurations are able to reach 47.6% and 68.9% complexity reductions for 0.66% and 1.80% BD-BR losses on class A sequences, respectively. In both BD-BR and complexity reduction metrics our method is better than Chen *et al.* [31] and Li *et al.* [22] techniques. Compared with these two techniques, our top-3 configuration solution achieves, on average, a lower BD-BR loss of 0.69% and 0.85% and a higher computational complexity reduction by 1.5% and 2.3%, respectively. Saldanha *et al.* [33] solution obtains slightly lower complexity reduction with a a 0.12% BD-BR loss compared to our top-3 configuration. Concerning the class B sequences, our top-3 configuration can halve the complexity with 51% of complexity reduction for a 0.8% BD-

TABLE V
ΔET↑ AND BD-BR↓ PERFORMANCE OF THE PROPOSED SOLUTION IN COMPARISON WITH STATE OF THE ART TECHNIQUES IN AI CODING CONFIGURATION.

Class	Sequence	Saldanha *et al.* [33], VTM10.0		Chen *et al.* [31], VTM4.0		Li *et al.* [22], VTM7.0		Ours top-3, VTM10.2		Ours top-2, VTM10.2	
		BD-BR	ΔET	BD-BR	ΔET	BD-BR	ΔET	BD-BR	ΔET	BD-BR	ΔET
Class A1	Tango2	0.71%	53.1%	1.38%	48.7%	1.52%	40.8%	0.50%	48.5%	1.48%	80.9%
	FoodMarket4	0.69%	46.5%	0.84%	30.3%	1.26%	42.2%	0.46%	46.7%	1.4%	67.5%
	Campfire	0.83%	43.8%	1.29%	49.1%	2.02%	48.7%	0.72%	48.2%	2.02%	67.4%
	Average	0.74%	47.8%	1.17%	42.7%	1.6%	43.9%	0.56%	47.8%	1.63%	71.9%
Class A2	CatRobot1	0.9%	44.1%	1.99%	50.4%	2.16%	45.4%	0.95%	46.6%	2.56%	62.9%
	DaylightRoad2	1.1%	53.7%	2%	54.1%	1.16%	49.4%	1.00%	52.1%	2.47%	73.7%
	ParkRunning3	0.45%	41.4%	0.8%	40.6%	1.15%	41.7%	0.32%	43.3%	0.91%	60.9%
	Average	0.82%	46.4%	1.6%	48.4%	1.49%	45.5%	0.76%	47.4%	1.98%	65.9%
Class B	MarketPlace	0.6%	54.5%	-	-	0.8%	46.6%	0.46%	55.0%	1.29%	75.7%
	RitualDance	1.09%	53.5%	-	-	1.07%	44.9%	0.80%	51.4%	2.41%	73.4%
	Cactus	1.04%	50%	1.73%	49.8%	1.12%	49.3%	0.84%	49.5%	2.6%	72.8%
	BasketballDrive	1.26%	57.1%	1.54%	50.1%	1.64%	52%	0.88%	52.1%	2.51%	76.0%
	BQTerrace	1.11%	48.3%	1.4%	56%	1.11%	45.6%	1.02%	47.0%	2.76%	75.1%
	Average	1.02%	52.7%	1.56%	52%	1.15%	47.7%	0.80%	51.0%	2.31%	74.6%
Class C	RaceHorses	0.75%	46.9%	1.35%	50.6%	0.96%	46.5%	0.61%	45.3%	1.97%	69.2%
	BQMall	1.4%	51%	2.12%	58.9%	1.17%	49.8%	0.94%	46.8%	3.08%	71.1%
	PartyScene	0.77%	48.3%	1.01%	51%	0.61%	45.2%	0.54%	43.7%	2.16%	69.3%
	BasketballDrill	1.52%	40.3%	2.05%	54.8%	1.63%	39.3%	1.48%	44.7%	4.69%	67.7%
	Average	1.11%	46.6%	1.63%	53.8%	1.09%	45.2%	0.89%	45.1%	2.97%	69.3%
Class D	RaceHorses	0.72%	45%	1.28%	54.7%	1.2%	41.6%	0.60%	43.9%	2.29%	66.5%
	BQSquare	0.57%	40.7%	0.75%	52.8%	0.74%	44.5%	0.55%	44.0%	2.43%	70.0%
	BlowingBubbles	0.82%	43.7%	1.4%	54.9%	0.92%	41.6%	0.60%	40.2%	2.37%	64.7%
	BasketballPass	1.32%	49.3%	1.77%	53.1%	1.41%	44.5%	0.84%	44.5%	2.86%	67.2%
	Average	0.86%	44.7%	1.3%	53.9%	1.07%	43.1%	0.64%	43.1%	2.49%	67.1%
Class E	FourPeople	1.71%	54.1%	2.71%	56.3%	1.33%	49.9%	1.10%	49.9%	3.46%	73.1%
	Johnny	1.65%	55.3%	2.77%	55.8%	2.33%	48.2%	1.34%	50.5%	3.71%	72.9%
	KristenAndSara	1.26%	53%	2.17%	52.8%	1.76%	50.5%	0.97%	49.0%	3.32%	71.8%
	Average	1.54%	54.1%	2.55%	55%	1.81%	49.5%	1.14%	49.8%	3.5%	72.6%
	Average	1.01%	48.8%	1.62%	51.2%	1.32%	45.8%	0.79%	47.4%	2.49%	70.4%
Class F	ArenaOfValor	-	-	-	-	-	-	1.00%	44.9%	3.28%	66.2%
	BasketballDrillText	-	-	2.09%	56.3%	-	-	1.57%	43.3%	4.69%	67.7%
	SlideEditing	-	-	0.52%	45.4%	-	-	0.95%	46.1%	4.14%	69.7%
	SlideShow	-	-	2.11%	45.8%	-	-	1.39%	44.4%	4.56%	68.0%
	Average	-	-	1.57%	49.2%	-	-	1.23%	44.7%	4.17%	67.9%

BR loss. The closest to our method is Saldanha *et al.* [33] solution which achieves slightly more complexity reduction score but for an increase of 0.22% BD-BR loss. Our top-2 configuration enables 74.6% complexity reduction for 2.31% BD-BR loss. The highest performance in complexity reduction and BD-BR loss is achieved for the *MarketPlace* sequence. Indeed, the top-3 and top-2 configurations reach 55% and 75.7% complexity reductions for 0.46% and 1.29% BD-BR losses, respectively.

For lower resolution classes C to E, the top-3 configuration achieves less than 1% BD-BR loss for a maximum of 50.5% complexity reduction. Compared to Saldanha *et al.* [33], our method has approximately the same complexity reduction but obtains lower BD-BR loss under classes C and D. For class E, their solution achieves 4.3% higher complexity reduction but with an increase of 0.4% BD-BR loss. Li *et al.* [22] has a higher BD-BR loss for lower complexity reduction for classes C and D compared with our top-3 configuration. For class E, Li *et al.* [22] solution achieved the same complexity reduction but at the cost of 0.53% BD-BR loss compared with our method. Chen *et al.* [31] almost doubled the BD-BR loss for a complexity reduction increase of 3.8% compared with our top-3 configuration through low resolution classes. In the case of the top-2 configuration, the performance is lower on low resolutions than on high-resolution. This approach on low-resolution contents, including classes C, D, and E reaches the complexity of high-resolution contents (classes A1, A2, and B) of 69.66% for a higher BD-BR loss of 2.98% which is higher by approximately 1% compared with the BD-BR loss of high resolutions contents (1.97%). Low-resolution contents result in deeper partitions, so the more the complexity is reduced, the less space is available to skip splits. Therefore, as the global partition is composed of more splits, the impact on BD-BR is more significant at high complexity reductions.

Class F has specific sequences with screen contents such as slides or gaming content. Our method still achieves 44.7% complexity reduction for 1.23% BD-BR loss. The work of Chen *et al.* [31] is the only one that reported results on class F with three out of four sequences. Our solution performs a slightly lower complexity reduction with a less BD-BR loss.

Fig. 10 illustrates the performance of our method in complexity reduction versus BD-BR jointly with state-of-the-art methods in a 2D plan averaged on the CTC classes excluding class F. Saldanha *et al.* [33] proposed different trade-

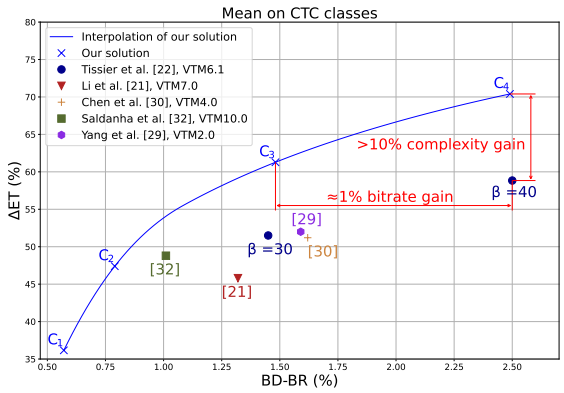

Fig. 10. Complexity reduction versus BD-BR performance comparison between the proposed method and state of the art techniques on CTC classes except class F (AI configuration). An interpolation cure over our four configurations is plotted in a blue dot line.

offs depicted in this figure. In addition to the configurations presented in Table V, two new configurations are included. The selection of the top-3 modes for the multi-classification with 6 outputs (classes) and top-4 modes for the other DT LGBM models is defined as C_1. The top-3 configuration is defined as C_2. The top-3 for all DT LGBM models except the multi-classification with 6 outputs for which top-2 is used in configuration C_3, and finally, top-2 for all DT models is defined as C_4. These configurations reach different trade-off points between complexity reduction and BD-BR loss, allowing us to draw an interpolation curve over these four configurations. This interpolation helps us to compare the results since the rate-distortion-complexity trade-off points are not linear. As explained in Table V, the figure confirms that our solution outperforms the best-performing state-of-the-art techniques. The points representing the state-of-the-art solutions are below our approach's interpolation curve.

To illustrate the gains brought by introducing DT models and dataset balancing, we show in Fig. 10 two configurations of our previous solution [23] relying on decision thresholds with $\beta = 30$ and $\beta = 40$. Our C_3 and C_4 configurations are better than the $\beta = 30$ and $\beta = 40$ solutions with a gain in complexity reduction of approximately 10% for a similar BD-BR loss. Compared with $\beta = 40$ which reaches 58.8% complexity reduction, our solution C_3 affords a significant BD-BR gain of 1.02%.

D. Ablation study

Table VI shows the contributions of the DT LGBM models and the dataset balancing on the top of the baseline model that considers only the CNN prediction with a simple threshold $\beta = 30$ [23]. We can notice that using the DT LGBM models enhance the ΔET/BD-BR ratio on average from 35.24 to 54.18. Furthermore, dataset balancing enables consistent gains, reaching the highest ΔET/BD-BR ratios over all CTC classes.

TABLE VI
ABLATION STUDY OF THE PROPOSED METHOD. PERFORMANCE IN TERMS OF ΔET↑, BD-BR↓ AND THE RATIO ΔET/BD-BR↑. THE BASELINE MODEL IS THE CNN PREDICTION WITH A THRESHOLD $\beta = 30$, BASELINE + DT INCLUDES THE DT MODELS FOR THE PREDICTION, AND OURS USES THE DATASET BALANCING. THE BEST RATIO IS HIGHLIGHTED IN BOLD.

Class	Baseline ($\beta = 30$) [23]	Baseline + DT	Ours (C_3)
A1	1.55/62.9 (40.58)	0.61/47.5 (79.17)	0.56/47.8 (**85.35**)
A2	1.47/60.0 (40.81)	0.80/46.4 (58.00)	0.76/47.4 (**62.36**)
B	1.41/61.1 (43.33)	0.85/51.0 (60.00)	0.80/51.0 (**63.75**)
C	1.20/37.9 (29.37)	0.96/45.7 (47.60)	0.89/51.1 (**57.41**)
D	0.83/32.5 (39.15)	0.70/43.3 (61.86)	0.64/43.1 (**67.34**)
E	2.29/54.4 (23.75)	1.23/43.9 (35.69)	1.14/49.8 (**43.68**)
Ave.	1.45/51.1 (35.24)	0.86/46.6 (54.18)	0.79/47.4 (**60.00**)
F	1.61/36.3 (22.54)	1.34/24.5 (18.28)	1.23/44.7 (**36.34**)

VII. COMPLEXITY ANALYSIS

In this section we assess the complexity overhead of the CNN and DT inferences under the VTM. Then, optimizations are proposed to reduce the complexity of the ML prediction models.

A. Complexity overhead

Table VII presents the time spent in the CNN and in the DT predictions for the C_2 configuration in comparison with the VTM10.2 anchor encoding time. These values are obtained by computing the ratio between the CNN or DT time and the VTM reference encoding time. The execution time of the CNN inference is, on average lower than the execution time of the DT model. Indeed, even if the CNN inference is more complex, the DT infers at each CU size unlike the CNN which is carried-out only once for each 64×64 CU. The run-time ratio of the CNN is higher for the Ultra High Definition (UHD) classes, primarily class A1. This is caused by the early skip methods integrated into the VTM that lead to shallow partition, resulting in a faster encoding process. Moreover, the results for the DT inference time is homogeneous through all the CTC classes. On average, the run-time of both the CNN and DT is under 2%, taken together; they require less than 3% of the encoding time. These results show that the prediction times are negligible, especially since the CNN and DT models can be optimized, accelerated, or run in parallel with the encoding. Indeed, the execution time of the CNN can be significantly reduced by targeting a GPU or on multi-core processors with optimizations, as presented below.

B. CNN inference optimisation

CNN-based complexity reduction methods have achieved outstanding results for the VTM encoder. However, CNN complexity needs to be carefully optimized to minimize the complexity overhead of the prediction. The proposed CNN processes as input a luminance block of size 68×68 with ten convolution layers and a final fully connected layer. The proposed CNN model consists of 226 088 training parameters. The execution time of the CNN inference is evaluated on Central Processing Unit (CPU) and GPU platforms as detailed below.

TABLE VII
COMPLEXITY OVERHEAD OF THE CNN AND DT LGBM (IN %) FOR C_2 CONFIGURATION COMPARED TO THE RUN TIME OF THE VTM ANCHOR.

	CNN coml. overhead	DT coml. overhead	Total
Class A1	2.7%	2.0%	4.7%
Class A2	1.2%	1.7%	2.9%
Class B	1.0%	1.7%	2.7%
Class C	0.6%	1.6%	2.2%
Class D	0.5%	1.6%	2.1%
Class E	1.4%	1.8%	3.2%
Class F	1.2%	1.6%	2.8%
Average	1.2%	1.7%	2.9%

TABLE VIII
INFERENCE TIME UNDER THE TENSORFLOW FRAMEWORK UNDER DIFFERENT CPU AND GPU PLATFORMS. THE INFERENCE TIME IS COMPUTED ON A 64×64 BLOCK , AND THE FRAME RATE ON A FULL HD RESOLUTION VIDEO.

	Platform	Inference time	Frame rate (fps)
CPU	Xeon W-2125 (8 cores - 4 GHz)	2.13 ms	0.92
	Ryzen 5 2600X (6 cores - 3.6 GHz)	2.53 ms	0.78
	I7-8700 (6 cores - 3.2 GHz)	2.66 ms	0.74
	I5-10300H (4 cores - 2.5 GHz)	3.36 ms	0.59
GPU	TF-TRT 2.4.0	259.49 μs	7.6
	TF-TRT 2.0.0 (FP32)	135.17 μs	14.5
	TF-TRT 2.0.0 (FP16)	120.80 μs	16.2

a) Optimisation on Central Processing Unit: As presented previously, the frugally-deep framework was considered to generate the C++ source code for the CNN inference, which is then integrated into the VTM. The proposed solution computed one CNN inference for each 64×64 block of pixels to predict the partitions. The CNN execution time depends on the targeted CPU.

The CNN inference compiled without specific compilation options and with one thread lasts 153ms for a 64×64 block. The inference complexity is also studied under the TensorFlow framework with different CPUs.

Table VIII lists the different CPUs and GPUs used to infer with our CNN through the TensorFlow framework and provides their respective inference times. The inference time depends mainly on the CPU clock rate. Indeed, with the Xeon W-2125 (8 cores - 4 GHz), the inference time to predict the partition of a 64×64 block is 2.13ms. The slowest CPU is the I5-10300H (4 cores - 2.5 GHz) with 3.36ms per inference. The inference time under the frugally-deep framework is at least $50\times$ higher than under the TensorFlow framework, which includes single instruction multiple data optimizations.

b) Optimisation on Graphics Processing Unit: The GPU is more adapted to train and infer CNN as most of the computations are matrix based and can be computed in parallel. Furthermore, the CUDA Application Programming Interface (API) that manages parallel computing along the cuDNN framework that optimizes standard operation for CNN are available to improve the inference execution time. Different versions of TensorFlow and CUDA are tested, impacting the performance of the CNN. For these experiments, the GPU selected is the Nvidia GTX 1650 Max Q. Under the TensorFlow framework specialized for GPU version 2.0.0 and the CUDA version 10.0, the CNN infers at $254.65\mu s$ for a 64×64 block partition, resulting in a 7.7 Frames per Second (fps) on full HD resolution.

Optimizing the CNN model is proposed to improve the inference execution time. The TensorRT framework proposed by Nvidia defined different optimizations like layer or tensor fusion to optimize the GPU memory, bandwidth, and precision refinement by quantizing the CNN model.

Table VIII details the inference time under the TensorRT framework with the GPU Nvidia GTX 1650 Max Q with different versions of the TensorRT framework with 32-bit and 16-bit floating-point data types. The fastest configuration is the TF-TRT 2.0.0 with FP16, which can predict the partition of the 64×64 block at $120.8\mu s$, which leads to 16.2 fps on a full High Definition (HD) sequence, while with FP32 configuration, the inference reaches $135.17\mu s$. These optimizations show impressive results using a GPU with TensorRT. Moreover, a dedicated processor can be used for inference, such as neural processing units proposed by Huawei or tensor processing units offered by Google, to reach even higher performance. Another option is to change the CNN architecture to limit the number of parameters and computations to reduce its inference run time.

VIII. CONCLUSION

In this paper we have proposed a two stage CNN and DT method to reduce the complexity of the VTM encoder in AI configuration. The CNN is designed to predict a CU partition through a vector of probabilities based on the local activity of pixels in a block. This vector considered as spatial features is then passed as input to the DT model that predicts the split probabilities at each CU depth. The DT method integrated after the CNN benefits from all the probabilities inside the computed CU instead of taking only the probabilities at the spatial location of the split. Depending on the selected configuration, a top-N probability selection is performed on the DT output to skip the unlikely splits.

Our proposed method outperforms state-of-the-art techniques regarding the trade-off between complexity reduction and coding loss. Concerning the top-3 configuration, our proposal assessed on the VVC CTC sequences enabled on average 47.4% complexity reduction for a negligible BD-BR loss of 0.79%. The top-2 configuration was able to reach a higher complexity reduction of 69.8% for 2.57% BD-BR loss. The complexity of the ML model has been carefully optimized. The CNN is able to predict the partition of the 64×64 block partition at $120.8\mu s$, resulting in 16.2 fps on a full HD sequence. These promising results motivated us to extend our method to the RA configuration. This extension, investigated as future work, will require taking advantage of the motion flow among adjacent frames.

ACKNOWLEDGMENTS

This work is supported by the Hubert Curien Partnerships (PHC) Maghreb 2021, No 45988WG (Eco-VVC project), and Région Bretagne through the TRISTRAM collaborative project and the Allocations de Recherche Doctorale (ARED) program.

REFERENCES

[1] Cisco, "Cisco visual networking index: Forecast and trends, 2017-2022," in *https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white-paper-c11-741490.html*, Feb. 2019.

[2] B. Bross, Y. Wang, Y. Ye, S. Liu, J. Chen, G. Sullivan, and J. Ohm, "Overview of the versatile video coding (vvc) standard and its applications," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 10, pp. 3736–3764, 2021.

[3] W. Hamidouche, T. Biatek, M. Abdoli, E. François, F. Pescador, M. Radosavljević, D. Menard, and M. Raulet, "Versatile video coding standard: A review from coding tools to consumers deployment," *IEEE Consumer Electronics Magazine*, vol. 11, no. 5, pp. 10–24, 2022.

[4] D. García-Lucas, G. Cebrián-Márquez, and P. Cuenca, "Rate-distortion/complexity analysis of HEVC, VVC and AV1 video codecs," *Multimedia Tools and Applications*.

[5] F. Bossen, X. Li, and K. Suehring, "AHG report: Test model software development (AHG3)," *JVET-T0003*.

[6] Y. Huang, J. An, H. Huang, X. Li, S. Hsiang, K. Zhang, H. Gao, J. Ma, and O. Chubach, "Block partitioning structure in the vvc standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 10, pp. 3818–3833, 2021.

[7] E. François, M. Kerdranvat, R. Jullian, and C. Chevance, "VVC PER-TOOL PERFORMANCE EVALUATION COMPARED TO HEVC," p. 14.

[8] M. Saldanha, G. Sanchez, C. Marcon, and L. Agostini, "Complexity analysis of VVC intra coding," in *2020 IEEE International Conference on Image Processing (ICIP)*. IEEE, pp. 3119–3123.

[9] A. Tissier, A. Mercat, T. Amestoy, W. Hamidouche, J. Vanne, and D. Menard, "Complexity reduction opportunities in the future VVC intra encoder," in *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*. IEEE, pp. 1–6.

[10] G. Correa, P. Assuncao, L. Agostini, and L. da Silva Cruz, "Fast hevc encoding decisions using data mining," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 4, pp. 660–673, 2015.

[11] T. Li, M. Xu, X. Deng, and L. Shen, "Accelerate CTU partition to real time for HEVC encoding with complexity control," *IEEE Transactions on Image Processing*, vol. 29, pp. 7482–7496.

[12] A. Mercat, F. Arrestier, M. Pelcat, W. Hamidouche, and D. Menard, "Machine learning based choice of characteristics for the one-shot determination of the HEVC intra coding tree," in *2018 Picture Coding Symposium (PCS)*, pp. 263–267.

[13] L. Shen, Z. Zhang, and Z. Liu, "Effective cu size decision for hevc intracoding," *IEEE Transactions on Image Processing*, vol. 23, no. 10, pp. 4232–4241, 2014.

[14] ——, "Adaptive inter-mode decision for hevc jointly utilizing inter-level and spatiotemporal correlations," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 10, pp. 1709–1722, 2014.

[15] S. Paul, A. Norkin, and A. Bovik, "Speeding up VP9 intra encoder with hierarchical deep learning based partition prediction," *IEEE Transactions on Image Processing*, vol. 29, pp. 8134–8148.

[16] M. Xu, T. Li, Z. Wang, X. Deng, R. Yang, and Z. Guan, "Reducing complexity of HEVC: A deep learning approach," *IEEE Transactions on Image Processing*, vol. 27, no. 10, pp. 5044–5059.

[17] A. Wieckowski, J. Ma, H. Schwarz, D. Marpe, and T. Wiegand, "Fast partitioning decision strategies for the upcoming versatile video coding (VVC) standard," in *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 4130–4134.

[18] Z. Wang, S. Wang, J. Zhang, S. Wang, and S. Ma, "Probabilistic decision based block partitioning for future video coding," *IEEE Transactions on Image Processing*, vol. 27, no. 3, pp. 1475–1486.

[19] T. Amestoy, A. Mercat, W. Hamidouche, D. Menard, and C. Bergeron, "Tunable VVC frame partitioning based on lightweight machine learning," *IEEE Transactions on Image Processing*, vol. 29, pp. 1313–1328.

[20] K. Choi, T. V. Le, Y. Choi, and J. Y. Lee, "Low-complexity intra coding in versatile video coding," *IEEE Transactions on Consumer Electronics*, vol. 68, no. 2, pp. 119–126, 2022.

[21] X. Dong, L. Shen, M. Yu, and H. Yang, "Fast intra mode decision algorithm for versatile video coding," *IEEE Transactions on Multimedia*, vol. 24, pp. 400–414, 2022.

[22] T. Li, M. Xu, R. Tang, Y. Chen, and Q. Xing, "Deepqtmt: A deep learning approach for fast qtmt-based cu partition of intra-mode vvc," *IEEE Transactions on Image Processing*, vol. 30, pp. 5377–5390, 2021.

[23] A. Tissier, W. Hamidouche, J. Vanne, F. Galpin, and D. Menard, "CNN oriented complexity reduction of VVC intra encoder," in *2020 IEEE International Conference on Image Processing (ICIP)*. IEEE, pp. 3139–3143.

[24] M. Wang, J. Li, L. Zhang, K. Zhang, H. Liu, S. Wang, S. Kwong, and S. Ma, "Extended coding unit partitioning for future video coding," *IEEE Transactions on Image Processing*, vol. 29, pp. 2931–2946.

[25] Q. Zhang, Y. Wang, L. Huang, and B. Jiang, "Fast CU partition and intra mode decision method for h.266/VVC," *IEEE Access*, vol. 8, pp. 117 539–117 550.

[26] T. Fu, H. Zhang, F. Mu, and H. Chen, "Two-stage fast multiple transform selection algorithm for VVC intra coding," in *2019 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, pp. 61–66.

[27] Biao Min and R. Cheung, "A fast CU size decision algorithm for the HEVC intra encoder," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 5, pp. 892–896.

[28] Z. Wang, S. Wang, J. Zhang, S. Wang, and S. Ma, "Effective quadtree plus binary tree block partition decision for future video coding," in *2017 Data Compression Conference (DCC)*, pp. 23–32.

[29] Z. Jin, P. An, L. Shen, and C. Yang, "CNN oriented fast QTBT partition algorithm for JVET intra coding," in *2017 IEEE Visual Communications and Image Processing (VCIP)*, pp. 1–4.

[30] H. Yang, L. Shen, X. Dong, Q. Ding, P. An, and G. Jiang, "Low-complexity ctu partition structure decision and fast intra mode decision for versatile video coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 6, pp. 1668–1682, 2020.

[31] F. Chen, Y. Ren, Z. Peng, G. Jiang, and X. Cui, "A fast CU size decision algorithm for VVC intra prediction based on support vector machine," *Multimedia Tools and Applications*.

[32] J. Zhao, Y. Wang, and Q. Zhang, "Adaptive CU split decision based on deep learning and multifeature fusion for h.266/VVC," *Scientific Programming*, vol. 2020, pp. 1–11.

[33] M. Saldanha, G. Sanchez, C. Marcon, and L. Agostini, "Configurable fast block partitioning for vvc intra coding using light gradient boosting machine," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2021.

[34] M. Lei, F. Luo, X. Zhang, S. Wang, and S. Ma, "Look-ahead prediction based coding unit size pruning for VVC intra coding," in *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 4120–4124.

[35] B. Bross, P. Keydel, H. Schwarz, D. Marpe, T. Wiegand, L. Zhao, X. Zhao, X. Li, S. Liu, Y. Chang, H. Jiang, P. Lin, C. Kuo, C. Lin, and C. Lin, "Multiple reference line intra prediction," *JVET-L0283*.

[36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *arXiv:1512.03385 [cs]*. [Online]. Available: http://arxiv.org/abs/1512.03385

[37] M. Tan and Q. V. Le, "EfficientNet rethinking model scaling for convolutional neural networks," *arXiv:1905.11946 [cs, stat]*. [Online]. Available: http://arxiv.org/abs/1905.11946

[38] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv:1409.1556 [cs]*. [Online]. Available: http://arxiv.org/abs/1409.1556

[39] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T. Liu, "LightGBM: A highly efficient gradient boosting decision tree," *Advances in Neural Information Processing Systems 30*, p. 9.

[40] E. Agustsson and R. Timofte, "NTIRE 2017 challenge on single image super-resolution: Dataset and study," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, pp. 1122–1131.

[41] E. Makov, *Dataset image 4k*. [Online]. Available: https://www.kaggle.com/evgeniumakov/images4k

[42] I. Jtc and I.-T. Sg, "Call for evidence on learning-based image coding technologies (JPEG AI)," p. 15.

[43] S. Hasinoff, D. Sharlet, R. Geiss, A. Adams, J. Barron, F. Kainz, J. Chen, and M. Levoy, "Burst photography for high dynamic range and low-light imaging on mobile cameras," *ACM Transactions on Graphics*, vol. 35, no. 6, pp. 1–12.

[44] R. e. a. Timofte, "NTIRE 2017 challenge on single image super-resolution: Methods and results," p. 12.

[45] F. Chollet et al., *Keras*. [Online]. Available: https://keras.io

[46] F. Abadi et al., "TensorFlow: Large-scale machine learning on heterogeneous distributed systems," p. 19.

[47] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv:1412.6980 [cs]*. [Online]. Available: http://arxiv.org/abs/1412.6980

[48] F. Bossen, J. Boyce, K. Suehring, X. Li, and V. Seregin, "JVET common test conditions and software reference configurations for SDR video," *JVET document, JVET-M1010*.

[49] G. Bjontegaard, "Calculation of average PSNR differences between RD-curves," *VCEG-M33*.

[50] T. Hermann, "Frugally deep." [Online]. Available: https://github.com/Dobiasd/frugally-deep