A CNN-Based Fast Inter Coding Method for VVC

Zhaoqing Pan, *Senior Member, IEEE*, Peihan Zhang, Bo Peng, *Member, IEEE*, Nam Ling, *Fellow, IEEE*, and Jianjun Lei, *Senior Member, IEEE*

***Abstract*—The Versatile Video Coding (VVC) achieves superior coding efficiency as compared with the High Efficiency Video Coding (HEVC), while its excellent coding performance is at the cost of several high computational complexity coding tools, such as Quad-Tree plus Multi-type Tree (QTMT)-based Coding Units (CUs) and multiple inter prediction modes. To reduce the computational complexity of VVC, a CNN-based fast inter coding method is proposed in this paper. First, a multi-information fusion CNN (MF-CNN) model is proposed to early terminate the QTMT-based CU partition process by jointly using the multi-domain information. Then, a content complexity-based early Merge mode decision is proposed to skip the time-consuming inter prediction modes by considering the CU prediction residuals and the confidence of MF-CNN. Experimental results show that the proposed method reduces an average of 30.63% VVC encoding time, and the Bjøontegaard Delta Bit Rate (BDBR) increases about 3%.**

***Index Terms*—Versatile Video Coding (VVC), Quad-Tree plus Multi-type Tree (QTMT), early Merge mode decision, CNN.**

I. Introduction

With the increase of video resolutions, the demand for more effective video coding technologies has increased rapidly. To solve this issue, the Joint Video Expert Group (JVET) has developed the latest video coding standard, called Versatile Video Coding (VVC) [1]. By introducing a series of new high-complexity coding technologies, VVC has achieved a giant coding performance improvement on the basis of High Efficiency Video Coding (HEVC) [2]–[7]. However, the extremely high computational complexity becomes a bottleneck for the VVC to be applied in real-time multimedia applications.

To improve the coding efficiency, the Quad-Tree plus Multi-type Tree (QTMT) partition structure is adopted in VVC. In the Coding Unit (CU) encoding process, the CU is recursively split into sub-CUs according to the QTMT, and the best partition mode is determined by the minimum Rate Distortion (RD) cost. The QTMT partition structure allows the CU with a square or a rectangle shape, which dramatically increases the encoding complexity. Hence, simplifying the QTMT-based CU partition process can significantly decrease the computational complexity of VVC. Besides, the advanced prediction modes have been introduced to VVC for improving the inter prediction accuracy, such as affine motion compensation prediction, adaptive motion vector resolution, bi-directional optical flow, and so on. These advanced prediction techniques also increase the computational complexity of VVC. In order to reduce the VVC encoding complexity, these modes can be conditionally skipped.

To reduce the computational complexity of CU encoding process, many fast CU encoding methods have been proposed. These methods can be roughly classified into two categories, namely statistical analysis-based methods [8]–[14] and CNN-based methods [15], [16]. The statistical analysis-based methods simplify the CU encoding process by building the relationship between the statistical features and the CU mode parameter. In [8], Tang *et al.* proposed a fast CU encoding method for intra and inter coding, in which the CU partition process is early terminated by using the edge features extracted by the canny edge detector. In addition, the three-frame difference is used to measure the motion activity of the CU content in the inter coding. In [9], Chen *et al.* regarded that the uniform area is usually encoded in large size CUs, and the non-uniform area is usually encoded in small size CUs. Based on this analysis, the CU partition process is early terminated according to the variance and gradient information of the CU. In [10], Cui *et al.* simplified the CU partition process by using the direction gradient information. In [11], Saldanha *et al.* utilized the variance and the best angular intra prediction mode of the current CU to skip the horizontal or vertical partition. In [12], Yang *et al.* proposed a fast intra coding scheme consisting of a cascade decision structure-based fast QTMT partition decision method and gradient descent-based fast intra mode decision method. In [13], Dong *et al.* proposed an adaptive mode pruning method to skip the non-promising modes and a mode-dependent termination method to skip the intra predictions of remaining depth levels. Although these statistical analysis-based fast CU encoding methods can improve the computational efficiency, the effectiveness of the statistical features depends on researchers' experience. To avoid designing features artificially, the CNN-based methods have emerged. These methods learn the features for CU size decision by convolution operations automatically. In [15], Tang *et al.* proposed a shape adaptive CNN-based fast CU partition decision for intra coding to handle CUs with various sizes by utilizing the variable size pooling layer. In [16], Tissier *et al.* trained a CNN

Manuscript received April 15, 2021; revised May 26, 2021; accepted May 30, 2021. Date of publication June 7, 2021; date of current version June 28, 2021. This work was supported in part by the National Key R&D Program of China under Grant 2018YFE0203900; in part by the National Natural Science Foundation of China under Grants 61931014, 61722112, 61520106002, and 61971232; in part by the Natural Science Foundation of Tianjin under Grant 18JCJQJC45800; and in part by the Natural Science Foundation of Jiangsu Province of China under Grant BK20201391. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Xun Cao. *(Corresponding author: Peihan Zhang.)*

Zhaoqing Pan, Peihan Zhang, Bo Peng, and Jianjun Lei are with the School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China (e-mail: zqpan3-c@my.cityu.edu.hk; peihan_zhang@tju.edu.cn; bpeng@tju.edu.cn; jjlei@tju.edu.cn).

Nam Ling is with the Department of Computer Engineering, Santa Clara University, Santa Clara, CA 95053 USA (e-mail: nling@scu.edu).

Digital Object Identifier 10.1109/LSP.2021.3086692

to predict a vector that contains the probabilities of having edges at the 4×4 border of the block in each 64×64 CU to skip the unlikely partition of intra coding. Since the current CNN-based methods mainly focus on the intra-frame CU size decision by extracting the texture features of the CU, they are not suitable for inter-frame CU size decision, as the temporal correlations among the inter frames are ignored.

To realize the reduction of the computational complexity of inter coding for VVC, a CNN-based fast inter coding method is proposed in this paper. First, by using CNN to analyze the CU texture and motion activity characteristics, a fast CU size decision method is proposed for early terminate the CU encoding process. Second, an early Merge decision method is proposed to speed up the inter mode selection process based on the confidence of CNN and the residuals predicted by Merge mode. The main contributions of this paper are summarized as follows:

1) A CNN-based fast CU encoding method is proposed for VVC inter coding. To the best of our knowledge, our work is the first attempt to speed up VVC inter coding based on deep learning.
2) To early terminate the CU partition process, a multi-information fusion CNN (MF-CNN) is proposed with the luma component, residuals, and motion field of the current CU as references.
3) The proposed fast inter coding method is integrated into VVC, and the experimental results demonstrate that the proposed method can significantly reduce the encoding time with negligible rate-distortion performance degradation.

II. Proposed Method

To reduce the computational complexity of VVC inter coding, a fast inter coding method is proposed in this paper, which consists of a CNN-based CU partition early termination method and a content complexity-based early Merge mode decision method. The details are introduced as follows.

A. MF-CNN-Based CU Partition Early Termination

The QTMT partition structure allows the CU to be partitioned flexibly, but the computational complexity is increased dramatically. To early terminate the QTMT-based CU encoding process, the CU partition problem is modeled as a two classification problem in this paper, that "partition" is regarded as 1, and "non-partition" is regarded as 0. To achieve the partition classification, an MF-CNN is proposed. The architecture of the proposed MF-CNN is shown in Fig. 1. The input of the proposed MF-CNN includes the luma component, residuals, and bi-directional motion field of the CU. The luma component reflects the texture complexity of the CU. The residuals and motion field are used to measure the complexity and intensity of the motion. Note that the residuals and the motion field are obtained by performing motion estimation on the currently processed frame in advance. In order to extract the useful features for CU partition encoding early termination, the Asymmetric Kernel Convolution Group (AKCG) [17] is adopted to learn the horizontal and vertical directions features from the luma component,

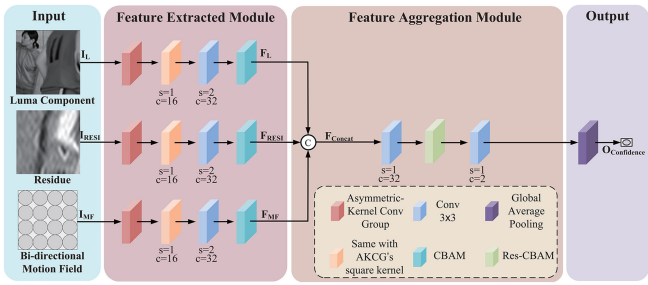

Fig. 1. Structure of proposed MF-CNN. Taking the CU's luma component, residue, and motion field as input, the network outputs the decision whether to terminate the CU partition encoding process. For convolution layers, "s" represents stride, and "c" represents channel.

TABLE I
ASYMMETRIC-KERNEL CONVOLUTION GROUP

CU Size	Kernel Size			Stride
Type-1	9×5	7×7	5×9	1
Type-2	7×3	5×5	3×7	1
Type-3	5×1	3×3	1×5	1

TABLE II
CHARACTERISTIC OF THE MERGE MODE. (UNIT:%)

Class	Sequence	T_m	P_m	A_α	A_β
B	BQTerrace	43.89	74.57	95.50	96.52
C	BasketballDrill	32.54	60.57	92.95	97.69
D	BQSquare	45.94	72.77	95.71	98.80
E	FourPeople	36.43	73.03	96.13	98.16
Average		39.70	70.23	95.10	97.79

residue, and motion field of the current encoding CU. Next, two convolution layers are used to learn the correlations among the learned features by the AKCG. Then three Convolution Block Attention Modules (CBAMs) [18] are used to highlight each previously learned feature and these features are further fused by cascade and convolution layers. To further increase the learning ability of the proposed network, the residual attention module (RES-CBAM) is used. Finally, a global average pooling layer is applied to acquire the binary classification results.

The proposed MF-CNN is a fully convolutional network [19], [20] so that it can handle CUs with various sizes. To fit different sizes of CU and obtain higher prediction accuracy, three AKCGs with different kernel sizes are used for three types of MF-CNNs, and they are defined as that: Type-1∈{128×128, 128×64, 64×128}, Type-2∈{64×64, 64×32, 32×64, 64×16, 16×64}, and Type-3∈{32×32, 32×16, 16×32, 16×16}. Each type of MF-CNN determines whether the partition process of CUs of different sizes can be early terminated. The details of the AKCG are reported in Table I. To avoid spending too much inference time on small CUs, only CUs whose width and height are both larger than 16 are predicted by the MF-CNN.

The loss function used to train the network consists of two parts. The first part is the basic cross-entropy loss. The second part is utilized to impose more penalties on the probability of incorrect prediction or a larger RD cost according to the goal of RD optimization in VVC [21]. Finally, the loss function L is

defined as

$$L = -\frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{2} y_i \log \hat{y_i} + \frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{2} \hat{y_i}(\frac{r_i}{r_{min}} - 1), \quad (1)$$

where N denotes the batch size. The ground-truth label and prediction value vector are denoted as y_i and $\hat{y_i}$. r_1 denotes the minimum RD cost of current CU obtained in non-partition inter prediction modes, and r_2 denotes the minimum RD cost of current CU obtained in partition modes. r_{min} denotes the smaller value between r_1 and r_2.

To further increase the prediction accuracy of the proposed CU partition early termination method, a confidence threshold scheme is proposed. Commonly, a larger confidence threshold, a higher prediction accuracy. While a large threshold could also decrease the loss of the coding efficiency and computational complexity reduction. To achieve a trade-off between the coding efficiency and computational complexity, a group of T_n (which range from 0.8 to 1 with the step of 0.05) has been tested and the experimental results show that when T_n is set to 1 for the CU with the size of 128x128, and 0.95 for the other CUs, the proposed MF-CNN achieves the best performance.

B. Content Complexity-Based Early Merge Mode Decision

In VVC, the Merge mode consists of not only translation Merge mode but also novel affine Merge mode and triangle prediction Merge mode. The Merge mode effectively improves the coding performance while consuming less computational complexity. Besides, the natural video sequence contains sufficient background areas and static areas, and these areas with simple content are suitable for encoding in Merge mode.

Statistical Analysis for the Merge Mode: In order to analyze the computational complexity and the utilization rate of the Merge mode, two metrics including the percentage of Merge mode to the total encoding time of all inter prediction modes, and the percentage of Merge mode selected as the best inter prediction mode, are counted. Four VVC standard video sequences are tested, and the statistical results are shown in Table II. T_m represents the percentage of predicting time used by Merge mode, and P_m represents the percentage that Merge is selected as the best prediction mode. It can be seen that the Merge mode only consumes about 39.70% of the total modes predicting time, and an average of 70.23% CUs select the Merge mode as their best prediction modes. From these values, it is concluded that a lot of time can be saved if Merge mode can be determined in advance.

Early Merge Mode Decision Method: The natural video sequence contains a lot of simple content, such as background and homogenous regions. In video coding process, the CUs with simple content are usually encoded in a large size mode such as Merge mode [22]. The previous work [22] has proved that the prediction residuals of these simple content CUs have a high probability to be transformed and quantized to zero. In addition, the CUs with simple content are usually not to be partitioned during the QTMT-based CU encoding process. Thus, the prediction residuals and CU partition information can be used to evaluate the content complexity of a CU.

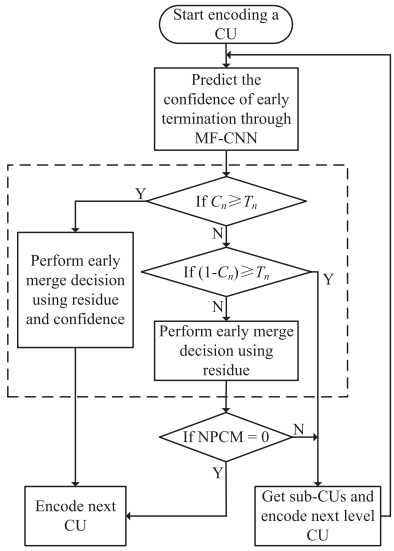

Fig. 2. Flowchart of the proposed CNN-based fast inter-frame coding method.

In order to analyze the relationship between the simple content CU and the Merge mode, four video sequences are encoded, and the statistical results are listed in Table II. It can be seen from A_α that when the residuals equal to zero, an average of 95.10% CUs correctly select the Merge mode as their best inter prediction modes. Based on this analysis, the best inter prediction mode of the CUs M is determined by

$$M = \begin{cases} Merge, \text{if } \phi_1 = 0 \text{ and } \phi_2 = 0 \\ non - Merge, \text{ otherwise} \end{cases}, \quad (2)$$

where ϕ_1 and ϕ_2 denote the residuals of the current CU after affine Merge mode and translation Merge mode encoding, respectively.

In addition, the A_β in Table II represents that when the residuals equal to zero and the CU will not be partitioned according to the MF-CNN, an average of 97.79% CUs correctly select the Merge mode as the best inter prediction mode. Thus, the best inter prediction mode of the non-partition CUs, M_n, is determined by

$$M_n = \begin{cases} Merge, \text{ if } \phi_1 = 0 \text{ and } \phi_2 = 0 \text{ and } C_n > T_m \\ non - Merge, \text{otherwise} \end{cases}, \quad (3)$$

where C_n denotes the confidence of terminating the partition of the CU obtained by the MF-CNN, and T_m is set to 0.98 in this paper.

C. Overview of the Proposed Fast Inter Coding Method

The overall flowchart of the proposed method is shown in Fig. 2. First, the C_n is obtained through the MF-CNN. Then, the C_n is compared with the threshold T_n. If C_n is larger than the T_n, the partition of the current CU is early terminated, and the confidence-based early Merge mode decision is used to skip the time-consuming inter prediction modes. Otherwise, the confidence of partitioning the CU is calculated as $1 - C_n$. If $1 - C_n$ is larger than the T_n, the current CU will be partitioned

TABLE III
PERFORMANCE OF THE COMPARISON AND THE PROPOSED METHOD. (UNIT: %)

Class	Sequence	Tang[7]		$Proposed_p$		$Proposed_o$	
		BDBR	TS	BDBR	TS	BDBR	TS
A1	Campfire	2.87	28.21	2.80	30.08	3.17	38.23
	FoodMarket4	1.41	32.21	1.59	42.90	1.74	46.12
	Tango2	3.75	29.29	3.68	34.05	4.03	38.56
A2	CatRobot1	4.18	30.12	5.59	30.62	6.45	36.84
	DaylightRoad2	4.84	29.63	4.43	29.20	5.63	35.47
	ParkRunning3	3.05	31.54	1.61	21.30	2.10	26.45
B	MarketPlace	3.30	34.19	3.22	36.47	4.33	33.64
	RitualDance	5.49	38.53	2.97	31.23	3.55	34.17
	BasketballDrive	5.22	35.11	2.96	32.39	3.30	37.28
	BQTerrace	1.95	34.64	0.98	13.80	1.90	20.21
	Cactus	2.93	31.18	5.20	25.42	5.72	29.36
C	BasketballDrill	1.74	15.30	1.59	24.38	2.29	29.23
	BQMall	0.31	8.29	2.35	22.41	2.69	27.48
	PartyScene	2.08	20.56	1.84	14.94	2.22	20.80
	RaceHorses	3.15	20.66	2.23	22.55	3.02	26.39
D	BasketballPass	10.10	24.26	1.56	21.18	1.85	26.97
	BlowingBubbles	2.26	29.25	2.29	16.97	3.03	22.15
	BQSquare	3.37	33.09	0.84	9.69	1.61	14.86
	RaceHorses	6.21	33.85	2.24	20.33	2.92	24.20
E	FourPeople	0.27	11.59	1.76	25.26	2.31	33.77
	Johnny	1.96	17.68	1.69	24.92	3.53	35.22
	KristenAndSara	1.87	15.18	2.11	26.21	2.58	36.50
	Average	**3.29**	**26.56**	**2.52**	**24.83**	**3.18**	**30.63**

into sub-CUs. Otherwise, only the early Merge mode decision method is used to accelerate inter encoding. Finally, if the number of partition modes in the CU's mode candidate list (NPCM) is zero, it means the current CU finishes the encoding process. On the contrary, the CU will be partitioned into sub-CUs.

III. EXPERIMENTS

A. Experimental Settings

The experiments are carried out on VTM 6.0 under JVET common test conditions [23] with Random Access (RA) configuration, and four Quantization Parameters (QPs) {22, 27, 32, 37} are utilized to encode the standard test video sequences. All the results are obtained by encoding 32 frames and the coding performance of this paper is measured with widely employed Bjøontegaard Delta Bit Rate (BDBR) [24] and encoding time saving rate (TS). The calculation formula of encoding time saving rate is defined as

$$TS = \frac{1}{4} \sum_{QP_i = \in \{22,27,32,37\}} \frac{T_{orig}(QP_i) - T_{prop}(QP_i)}{T_{orig}(QP_i)}, \quad (4)$$

where T_{orig} indicates the encoding time of VTM 6.0 and T_{prop} represents the encoding time of the proposed method.

The proposed MF-CNN is trained by utilizing the Tensorflow [25]. 83 videos are collected from [26] to create a largescale dataset based on VVC inter coding at four QPs. For different QPs and CU sizes, the models are trained independently. Moreover, all models are optimized by using Adam. The initial learning rate is set to 10^{-4} for 100 epochs and reduce to 90% every 100 000 steps. All experiments are conducted on a hardware platform with Intel i7-8700 K processor @3.70 GHz and 32 GB RAM and GeForce GTX 1080Ti GPU.

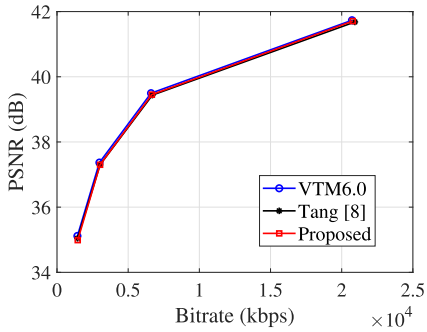

Fig. 3. RD curves comparison.

B. Experimental Results

Table III shows the results of the proposed method compared to VTM 6.0. $Proposed_o$ represents the overall fast inter coding method and $Proposed_p$ represents the CU partition early termination method. It is observed that the overall method can save the encoding time ranges from 14.86% to 46.12%, 30.63% on average with 3.18% BDBR increase. The CU partition early termination method saves the encoding time ranges from 9.69% to 42.90%, 24.83% on average with 2.52% BDBR increase. Compared with the original VTM 6.0, the time overhead introduced by the MF-CNN is 2.14%, and it has been considered in the results. The experimental results demonstrate that the proposed overall method achieves effectively encoding time-saving on average with a negligible BDBR increase. To make a fair comparison, the statistical analysis-based method [8] is re-implemented on VTM 6.0, and the results are also presented in Table III. According to the results, it can reduce the complexity from 8.29% to 38.53%, 26.56% on average with 3.29% BDBR increase. Compared to the method in [8], the proposed method achieves better complexity reduction with less BDBR performance loss.

To intuitively show the RD performance, the RD curves of the original VTM 6.0, the statistical analysis-based method [8], and the proposed method are shown in Fig. 3. The values of bitrates and PSNR denote the total average results of all the video sequences. It is observed that the proposed method achieves similar RD performance with VTM 6.0 and the statistical analysis-based method [8].

IV. CONCLUSION

This paper proposes a fast inter coding method for VVC. First, an MF-CNN-based CU partition early termination method is designed to simplify the QTMT-based CU partition process by fully leveraging the texture and motion activity characteristics. Second, a content complexity-based early Merge mode decision method is proposed to skip the time-consuming inter prediction modes by utilizing the residuals and confidence. Experimental results demonstrate that the proposed method can effectively reduce the complexity of VVC encoding with a negligible BDBR increase.

REFERENCES

[1] B. Bross, J. Chen, and S. Liu, Versatile Video Coding (Draft 6), Document JVET-J1001, Jun. 2019.

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

[3] L. Jia, K. Jia, and X. Fan, "Adaptive lagrangian multiplier for quantization parameter cascading in HEVC hierarchical coding," *IEEE Signal Process. Lett.*, vol. 27, pp. 1220–1224, 2020.

[4] L. Yu, L. Shen, H. Yang, L. Wang, and P. An, "Quality enhancement network via multi-reconstruction recursive residual learning for video coding," *IEEE Signal Process. Lett.*, vol. 26, no. 4, pp. 557–561, Apr. 2019.

[5] H. Guo, C. Zhu, M. Xu, and S. Li, "Inter-block dependency-based CTU level rate control for HEVC," in *IEEE Trans. Broadcast*, vol. 66, no. 1, pp. 113–126, Mar. 2020.

[6] Y. Gao, C. Zhu, S. Li, and T. Yang, "Source distortion temporal propagation analysis for random-access hierarchical video coding optimization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 2, pp. 546–559, Feb. 2019.

[7] Z. Pan, W. Yu, J. Lei, N. Ling, and S. Kwong, "TSAN: Synthesized view quality enhancement via two-stream attention network for 3D-HEVC," *IEEE Trans. Circuits Syst. Video Technol.*, early access, pp. 1–14, Feb. 2021, doi: 10.1109/TCSVT.2021.3057518.

[8] N. Tang *et al.*, "Fast CTU partition decision algorithm for VVC intra and inter coding," in *Proc. APCCAS*, Nov. 2019, pp. 361–364.

[9] J. Chen, H. Sun, J. Katto, X. Zeng, and Y. Fan, "Fast QTMT partition decision algorithm in VVC intra coding based on variance and gradient," in *Proc. VCIP*, Dec. 2019, pp. 1–4.

[10] J. Cui, T. Zhang, C. Gu, X. Zhang, and S. Ma, "Gradient-based early termination of CU partition in VVC intra coding," in *Proc. DCC*, Mar. 2020, pp. 103–112.

[11] M. Saldanha, G. Sanchez, C. Marcon, and L. Agostini, "Fast partitioning decision scheme for versatile video coding intra-frame prediction," in *Proc. ISCAS*, Oct. 2020, pp. 1–5.

[12] 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 Trans. Circuits Syst. Video Technol.*, vol. 30, no. 6, pp. 1668–1682, Jun. 2019.

[13] X. Dong, L. Shen, M. Yu, and H. Yang, "Fast intra mode decision algorithm for versatile video coding," *IEEE Trans. Multimedia.*, early access, pp. 1–15, Jan 2021, doi: 10.1109/TMM.2021.3052348.

[14] G. Fu, L. Shen, H. Yang, X. Hu, and P. An, "Fast intra coding of high dynamic range videos in SHVC," *IEEE Signal Process. Lett.*, vol. 25, no. 11, pp. 1665–1669, Nov. 2018.

[15] G. Tang, M. Jing, X. Zeng, and Y. Fan, "Adaptive CU split decision with pooling-variable CNN for VVC intra encoding," in *Proc. VCIP*, Dec. 2019, pp. 1–4.

[16] A. Tissier, W. Hamidouche, J. Vanney, F. Galpinz, and D. Menard, "CNN oriented complexity reduction of VVC intra encoder," in *Proc. ICIP*, Oct. 2020, pp. 3139–3143.

[17] Z. Chen, J. Shi, and W. Li, "Learned fast HEVC intra coding," *IEEE Trans. Image Process.*, vol. 29, pp. 5431–5446, 2020.

[18] S. Woo, J. Park, J. Lee, and I. Kweon, "CBAM: Convolutional block attention module," in *Proc. ECCV*, pp. 3–19, 2018.

[19] B. Peng, J. Lei, H. Fu, Y. Jia, Z. Zhang, and Y. Li, "Deep video action clustering via spatio-temporal feature learning," *Neurocomputing.*, pp. 1–9, Feb 2021, doi: 10.1016/j.neucom.2020.05.123.

[20] J. Lei, X. Li, B. Peng, L. Fang, N. Ling, and Q. Huang, "Deep spatial-spectral subspace clustering for hyperspectral image," *IEEE Trans. Circuits Syst. Video Technol.*, early access, Sep 2020, doi: 10.1109/TCSVT.2020.3027616.

[21] T. Li, M. Xu, R. Tang, Y. Chen, and Q. Xing, "Deep-QTMT: A deep learning approach for fast QTMT-based CU partition of intra-mode VVC," 2006, *arXiv:13125*.

[22] Z. Pan, S. Kwong, M. Sun, and J. Lei, "Early MERGE mode decision based on motion estimation and hierarchical depth correlation for HEVC," *IEEE Trans. Broadcast*, vol. 60, no. 2, pp. 405–412, Jun. 2014.

[23] K. Suehring and X. Li, "JVET common test conditions and software reference configurations," Document JVET-G1010, Aug. 2017.

[24] G. Bjontegaard, "Calculation of average PSNR differences between RD-curves," document VCEG-M33, ITU, Geneva, Switzerland, Apr. 2001.

[25] M. Abadi *et al.*. "TensorFlow: Large-scale machine learning on heterogeneous distributed systems," 2016, *arXiv, 1603.04467*:

[26] Xiph-org, "Xiph-org Video Test Media," 2017. [Online]. Available: https/ /media.xiph.org/video/derf