Versatile Video Coding based Quality Scalability with Joint Layer Reference

Xiem HoangVan, Sang NguyenQuang, and Fernando Pereira, *Fellow, IEEE*

***Abstract*—Scalability is an essential coding feature for adaptive video streaming applications, notably considering the growing heterogeneity of the transmission, display and consumption environments. Versatile video coding (VVC) is the emerging video coding standard, targeting offering higher compression efficiency regarding previous standards to further facilitate already available and novel video applications, notably at higher spatial resolutions. In this context, this paper proposes the first VVC-based quality scalability extension, targeting to offer higher compression efficiency than the native VVC quality scalability solution. The proposed Quality Scalable Versatile Video Coding (QS-VVC) solution is designed based on a layered coding approach with one base layer (BL) and one or more enhancement layers (EL). To achieve higher compression performance, a novel joint layer referencing approach is proposed where the base and enhancement layers decoded information are jointly exploited to create a new EL coding reference. Experimental results shown that the proposed QS-VVC codec outperforms the most relevant benchmarks, notably VVC-based simulcasting, native VVC quality scalability, and the previous Scalable High Efficiency Video Coding (SHVC) standard.**

***Index Terms*— Versatile video coding, quality scalability, joint layer reference**

I. Introduction

Nowdays, digital video is the key data component in a wide range of applications from video telephony and video surveillance to mobile and Internet streaming, and TV broadcasting [1]. In these application scenarios, at least the network and terminal characteristics are rather heterogeneous, e.g. in terms of bandwidth, display capabilities and complexity resources; moreover, some of these characteristics may dynamically vary along time, thus asking for flexible solutions, notably in terms of video coding. Therefore, it is critical to be able to have highly adaptive video coding streams in order the best quality of experience is offered to the users for the resources available at any instant time. This increasingly important requirement asks for scalable video streams where, from a single, unique coded bitstream, appropriate and efficient sub-streams may be easily extracted to address the relevant network and terminal constrained or dynamically changing conditions [2]. Depending on the application domain, different forms of video coding scalability may be more relevant, notably: i) *temporal scalability*, which refers to sub-streams successively increasing the frame rate; ii) *spatial resolution scalability*, which refers to sub-streams successively increasing the frame spatial resolution; and iii) *quality scalability*, which refers to sub-streams successively increasing the frame quality for a target frame rate and spatial resolution.

Versatile video coding (VVC) is the emerging video coding standard, recently developed by the Joint Video Experts Team (JVET) of ISO/IEC MPEG and ITU-T VCEG to meet the growing demand for compression efficiency in current and emerging video applications [3]. It is expected that this novel video coding standard provides the same perceptual quality as the most efficient video coding solutions in the market, notably the HEVC standard [4], at around half the bitrate [5]. In addition, VVC is also expected to offer native flexible, high-level syntax mechanisms for resolution adaptivity, scalability and multi-view features [3].

Since scalability capabilities are fundamental in the VVC framework, this paper proposes an improved quality scalable video coding solution, designed as a VVC extension, based on a multi-layer coding approach, called Quality Scalable Versatile Video Coding (QS-VVC). In this design, a novel joint layer referencing approach is proposed where the base and enhancement layers already decoded information are effectively combined using a set of spatio-temporal features to improve the overall compression performance, notably regarding the native VVC quality scalability solution. This requires the EL decoded picture buffer (DPB) to also include the new joint layer reference (JLR) frame. The rate-distortion (RD) performance results show that the proposed QS-VVC solution outperforms the relevant benchmarks, notably VVC simulcasting (where the base and enhancement layer frames are independently coded), the native VVC quality scalability solution and the previous SHVC standard.

To target its objectives, this paper is organized as follows: Section II offers a brief overview of the relevant background work, notably SHVC and VVC. Next, Section III describes the proposed QS-VVC solution and the joint layer reference frame creation process. Section IV reports and analyses the QS-VVC performance and, finally, Section V presents the main conclusions and ideas for future work.

Manuscript received August 31, 2020; revised November 03, 2020. This research is funded by Vietnam National Foundation for Science and Technology Development (NAFOSTED) under grant number 102.01-2020.15. The associate editor coordinating the review of this manuscript was Dr. Mateo Naccari. (*Corresponding author: Xiem HoangVan.*)

X. HoangVan, Sang NguyenQuang are with the Faculty of Electronics and Telecommunications, VNU-University of Engineering and Technology, Hanoi 10000, Vietnam (email: xiemhoang@vnu.edu.vn).

F. Pereira is with Instituto Superior Técnico and Instituto de Telecomunicações, Lisbon, Portugal (email: fp@lx.it.pt).

II. Background Work

This section will briefly review SHVC, the most recent scalable coding standard and the VVC standard itself, including its native scalability capabilities.

A. Scalable High Efficiency Video Coding Standard

In the past, the SHVC standard has been designed to address the three forms of scalability mentioned above [2]. In SHVC, the adopted layered coding structure and high level syntax (HLS) approach build on top of the core HEVC coding tools [4]. In this context, SHVC compresses the video with one base layer (BL) and one or more enhancement layers (ELs), always using the HEVC coding tools, thus limiting the complexity and offering a high degree of HEVC compatibility. The SHVC standard adopts a multi-layer coding framework where the BL decoded picture, resampled if necessary, notably for spatial scalability, is used as an additional reference picture for EL prediction beyond the EL reference pictures [6]. In comparison with HEVC, the SHVC extension is limited to syntax changes at the slice level and above, this means without additional, specific coding modes; this increases the compatibility between SHVC and HEVC and eases its implementation and deployment. In terms of RD performance, SHVC brings a rate penalty of around 14.3% and 24.3% compared with single-layer HEVC for Random Access and Low Delay configurations, respectively, notably for two-layer quality scalability [2].

B. Versatile Video Coding Standard

The VVC standard still adopts the block-based hybrid coding architecture as the preceding HEVC standard [4]. VVC has been developed with two main objectives in mind: i) to specify a video coding solution with a compression performance substantially beyond the HEVC standard; and ii) to be highly versatile for effective use in a broadened range of applications, notably offering mechanisms for resolution adaptivity, region-based access, scalability, coding of various chroma sampling formats, and flexible bitstream handling [3]. To offer the targeted additional compression efficiency, VVC includes multiple new coding tools, as detailed in [3].

In terms of scalability, and differently from the past, VVC targets to offer scalability capabilities from version 1 and not in future extensions such as SHVC with HEVC. This happens by supporting these capabilities as an intrinsic function where an HLS approach is again adopted to provide temporal, spatial and also quality scalabilities. For example, the available reference picture resampling (RPR) tool, which allows resampling a reference picture to be used for inter prediction when that reference picture has a different resolution than the current picture to be coded, is also appropriate to offer spatial scalability without any need for additional signal processing-level coding tools. For quality scalability, the same HLS flexibility offers a native solution where the BL and EL reconstructed frames are used as references for EL coding, thus available in the DPB.

III. A Quality Scalable VVC Based Solution

This section presents the key technical novelties of this paper, notably the proposed QS-VVC solution and its joint layer reference creation process.

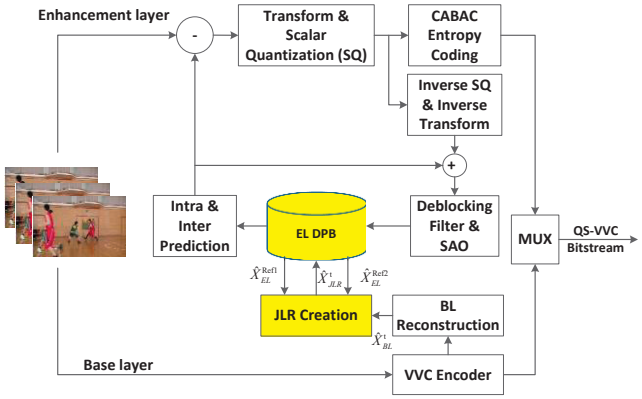

Fig. 1. Proposed QS-VVC encoder architecture.

A. QS-VVC Architecture and Walkthrough

Fig. 1 illustrates the proposed QS-VVC encoder architecture for a two-layer case where the layered coding structure is deployed with one BL and one EL; naturally, this architecture may be easily extended in the same manner for two or more ELs. The QS-VVC architecture adopts a multi-layer scalable coding process where coding proceeds as follows:

- **BL coding and decoding**: First, the BL frame is entirely VVC encoded and decoded before the EL starts to be coded. In this step, a higher quantization parameter (QP) value is typically used to obtain a coarser quality BL decoded picture, thus typically requiring a lower bitrate.
- **JLR creation**: Second, since the set of scalable coding references plays a key role on the final EL compression efficiency, a new high quality reference frame is created in this step by jointly exploiting the available BL and EL decoded frames; this process is described in Section III-B.
- **EL encoding**: Finally, to achieve higher quality decoded video, VVC is used to encode the EL frame with a lower QP. However, to achieve higher EL compression performance, instead of using only the BL and EL references for coding the EL frame as specified for the native VVC scalable coding approach [3], the proposed QS-VVC solution also exploits the new JLR frame, now available at the EL DPB. In this way, a common rate-distortion optimization (RDO) process will be applied to the four EL coding references available at the EL DPB to find the optimal coding reference and associated information.

In summary, the proposed QS-VVC solution provides quality scalability by adopting a layered coding approach as commonly happens for prior scalable video coding standards [2]. However, to improve the EL compression performance, a new JLR frame is added to the VVC EL DPB, created as proposed in the next sub-section.

B. Joint Layer Reference Frame Creation

Since the EL (decoded) reference frames, $\hat{X}_{EL}^{Ref1}$, $\hat{X}_{EL}^{Ref2}$ and BL decoded frame, $\hat{X}_{BL}^{t}$, exist at different time instants, it is proposed here to start the JLR creation process by interpolating a Motion Compensated Temporal Interpolation (MCTI) frame for the current (BL) time instant by performing MCTI with the available EL reference frames. The new EL MCTI frame is then adaptively combined with the BL decoded frame, available for the same time instant, using a block-level fusion process as depicted in Fig. 2.

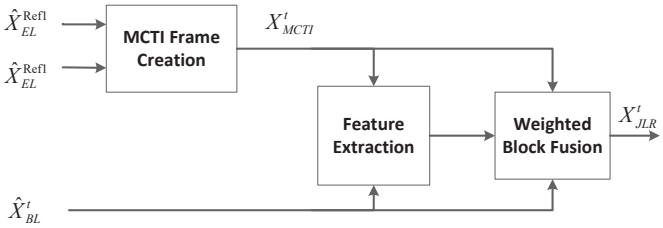

Fig. 2. Joint layer reference frame creation architecture.

1) MCTI Frame Creation

In the JLR frame creation process, the MCTI frame intends to express the EL temporal correlation. For this, two EL reference frames (one in the past and another in the future of the EL frame to code) are employed to estimate the target EL frame to be coded. Several interpolation methods can be used with this purpose, such as bi-linear interpolation, frame averaging, block copying [7] or deep learning-based methods [8]. Among them, the MCTI process proposed and detailed in [9] was selected due to its good trade-off between the interpolated frame quality and the associated computational complexity. The MCTI creates an interpolated frame based on motion estimation and compensation, as commonly adopted in conventional frame rate-up conversion.

2) MCTI and BL Decoded Frames Fusion

The JLR frame is obtained by fusing the BL decoded frame, expressing the spatial correlation regarding the EL frame to code, with the MCTI frame, expressing the temporal correlation again regarding the EL frame to code. To better perform this fusion, a spatio-temporal block-based fusion method is proposed where a set of spatio-temporal features are firstly extracted from the two frames to fuse to create a so-called block fusion map. This map is then used to fuse the MCTI and BL decoded frames at block-level as shown in Fig. 2. The fusion process includes two main steps, notably the Feature Extraction and the Weighted Block Fusion.

a. Feature Extraction

In the proposed JLR creation process, a number of discriminative features is adopted to block-level control the proposed frame fusion process; these features are adopted based on their correlation with the "ground truth" decision where the original picture would be used to compute the ideal block fusion weights between the MCTI and BL decoded frames. First, inspired by the fact that the MCTI frame quality is highly dependent on the temporal correlation between the EL references, two temporal discriminative features, Motion Vector Amplitude (MVA), F_{MVA}, and the Sum of Squared Differences between the two motion compensated EL references, F_{SSD}, are adopted. Second, since the BL decoded quality is usually affected by the quantization error and the video content itself, two spatial discriminate features are adopted, notably the BL block VARiance, F_{VAR}, and the BL RESidue energy, F_{RES}. In detail, the adopted fusion features to be computed at encoder and decoder based on already available information are:

• **MV Amplitude:** The MV field obtained during the MCTI creation process, which represents the temporal correlation between the two EL references, is used to compute the block-level motion vector amplitude as:

$$F_{MVA} = \sqrt{mv_x^2 + mv_y^2} \tag{1}$$

• **Sum of Squared Differences (SSD) between the two motion compensated EL references:** The similarity between the two motion compensated EL references is appropriate to estimate how good the MCTI interpolated frame is, here expressed by the SSD computed as:

$$F_{SSD} = \sum_{x=0}^{N-1}\sum_{y=0}^{N-1}\left(\hat{X}_{EL}^{Ref1}(x - mv_x, y - mv_y) - \hat{X}_{EL}^{Ref2}(x + mv_x, y + mv_y)\right)^2 \tag{2}$$

where (mv_x, mv_y) is the relevant block-level motion vector.

• **BL block VARiance:** Since the BL decoded frame quality is mainly affected by the quantization noise, driven by the quantization parameter, and the video content itself, whichcannot be changed, the block content 'complexity' is here assessed by the block variance computed as:

$$F_{VAR} = \frac{1}{N \times N}\sum_{x=0}^{N-1}\sum_{y=0}^{N-1}\left(\hat{X}_{BL}^{t}(x,y) - \mu_{BL}\right)^2 \tag{3}$$

where μ_{BL} is the average luminance intensity for the current block computed as:

$$\mu_{BL} = \frac{1}{N \times N}\sum_{x=0}^{N-1}\sum_{y=0}^{N-1}\hat{X}_{BL}^{t}(x,y) \tag{4}$$

• **BL RESidue energy:** As the quantization noise has a critical impact on the BL decoded quality, the BL residue energy is here computed as:

$$F_{RES} = \sum_{x=0}^{N-1}\sum_{y=0}^{N-1}\left(\hat{X}_{BL}^{t}(x,y) - P_{BL}^{t}(x,y)\right)^2 \tag{5}$$

where $P_{BL}^{t}(x,y)$ is the BL prediction block already available at both the encoder and decoder.

This set of features is made available to the Weighted Block Fusion module to finally control the JLR frame creation.

b. Weighted Block Fusion

To create a high accuracy JLR frame, X_{JLR}^{t}, the available BL decoded frame, $\hat{X}_{BL}^{t}$, is adaptively combined with the interpolated MCTI frame, X_{MCTI}^{t}. In this case, a weighted block-level frame fusion process [10] is adopted due to its trade-off between computational complexity and accuracy. The linear weighting block-level combination is performed at block-level as follows:

$$X_{JLR}^{t} = W_{Block} \times \hat{X}_{BL}^{t} + (1 - W_{Block}) \times X_{MCTI}^{t} \tag{6}$$

In this fusion process, a weighting fusion factor, W_{Block}, is used to control the balance between the two frames being combined. Naturally, the larger is the weighting fusion factor, the more reliable/better should be the quality of the corresponding reference frame and the higher its contribution for the created JLR fused frame. Since the spatio-temporal features defined above should express well how reliable are the MCTI and BL decoded frames in terms of EL prediction power, it is reasonable to estimate the weighting fusion factor W_{Block}, based on those features. In this paper, it is proposed to jointly exploit the expressiveness of the extracted features to create a block-level weighting fusion map. Since F_{MVA} and F_{SSD} are created based on the MCTI information and F_{VAR} and F_{RES} are created based on BL information, it is appropriate to

compute the weighting fusion factor in (6) as follows since this factor directly weights the BL decoded frame:

$$W_{Block} = \frac{\overline{F_{MVA}} + \overline{F_{SSD}}}{\overline{F_{MVA}} + \overline{F_{SSD}} + \overline{F_{VAR}} + \overline{F_{RES}}} \quad (7)$$

To better exploit all features information, a normalization process is applied to obtain the normalized features, $\overline{F_{MVA}}$, $\overline{F_{SSD}}$, F_{VAR}, $\overline{F_{RES}}$, for each block; for example, it comes for $\overline{F_{MVA}}$:

$$\overline{F_{MVA}} = \frac{F_{MVA}}{Max\ value\ of\ F_{MVA}\ in\ current\ frame} \quad (8)$$

Finally, the weighted computed in (7) is used to obtain JLR frames as in (6). Notice that if F_{MVA} and F_{SSD} are low, this implies that the temporal correlation is high, meaning that $\hat{X}_{EL}^{MCTI}$ is reliable, and thus the weighting fusion factor, W_{Block}, multiplying the BL decoded frame should be low and vice-versa. The JLR frame is then inserted into the EL DPB at the last position of List0 to be used as an additional reference for more efficiently coding the EL frames.

IV. Performance Assessment

This section discusses the RD performance for the QS-VVC solution with respect to the most relevant benchmarks.

A. Test Material and Conditions

The proposed QS-VVC solution has been implemented on top of the VVC reference software, the so-called *VVC Test Model (VTM)*, version 8.0 [11]. Although two quality layers are adopted in this paper, more quality layers may be included in a similar way. Eight video sequences from the VVC Common Test Conditions (CTC) have been selected for performance assessment, notably from classes A, B, C, and D [12]; their characteristics are included in Table I.

To assess the QS-VVC RD performance, four QP pairs are used, i.e., QPs for BL and EL are {(37, 33), (32, 28), (27, 23), (22, 18)}. Moreover, the Random Access (RA) test configuration is used to allow hierarchical coding. The following benchmarks are adopted for comparison:

1. VVC simulcasting (**VVC-SIM**), where the BL and ELs are independently coded using VVC;
2. Native VVC quality scalability (**VVC-Scalable**), where the HLS-level VVC capabilities are used to obtain two quality layers with different QPs;
3. **VVC–MCTI**, where MCTI is used to create an additional (to VVC) reference frame for EL coding;
4. VVC single layer (**VVC-SL**), where only the EL is coded using VVC;
5. **SHVC**, where the BL and ELs are coded in a scalable way using the SHVC standard.

TABLE I: Summary of Test Sequences

Resolution	Sequence (Abbreviation)	Number of frames	Frame rate
A	PeopleonStreet (A1)	150	30 Hz
2560×1600	Traffic (A2)	150	30 Hz
B	Kimono (B1)	240	24 Hz
1920×1080	ParkScene (B2)	240	24 Hz
C	RaceHorses (C1)	300	30 Hz
832×480	BasketballDrill (C2)	500	50 Hz
D	RaceHorses (D1)	300	30 Hz
416×240	BlowingBubbles (D2)	500	50 Hz

The computational complexity assessment has been performed in personal computers with processor Intel® Core™ i7-4800MQ @2.7 GHz, RAM of 8 GB and Microsoft Visual Studio 2017 Community.

B. Compression Performance and Complexity Analysis

Table II reports the BD-rate savings [13] for VVC-SIM, VVC-scalable, VVC-MCTI, VVC- SL and the proposed QS-VVC, always using SHVC as reference: in this table, negative BD-Rate values correspond to rate savings. Table II also reports the encoding time variation (*ETV*) in % between the proposed QS-VVC (ET_{QS-VVC}) and the native VVC-Scalable ($ET_{VVC-Scalable}$). From the results in Table II, the key conclusions are:

- Overall, the proposed QS-VVC solution achieves better RD performance than all the relevant scalable coding benchmarks, e.g., SHVC, VVC-SIM and VVC-Scalable;
- Compared to native VVC-Scalable, QS-VVC achieves around 3% bitrate savings, on average. This is a relevant rate saving since VVC is a very optimized codec; this gain is due to the additional JLR frame for EL coding;
- Compared to VVC-MCTI, QS-VVC also achieves around 3% bitrate savings, on average; this gain is mainly due to the proposed fusion process;
- Compared to SHVC, QS-VVC achieves a BD-rate saving of 16.40% while VVC-Scalable only saves 13.62%;
- Compared to VVC-SL, there is still a RD performance penalty to offer quality scalability, already less than 20% for QS-VVC; this leaves room for further research targeting to offer more efficient scalability functions;
- Finally, QS-VVC introduces less than 5% of encoding time increase, on average, in comparison to VVC-scalable.

In summary, QS-VVC brings added value to the quality scalability video coding state-of-the-art, notably the most recent standard solution as represented by VVC-Scalable.

TABLE II: BD-Rate (%) and Encoding Time Variation (%)

Seq.	VVC-SIM	VVC-Scalable	VVC-MCTI	VVC-SL	Proposed QS-VVC	ETV
A1	3.36	-19.94	-20.19	-33.85	-24.62	5.38
A2	-3.31	-9.96	-9.95	-36.92	-14.22	9.10
B1	2.91	-12.03	-11.96	-33.09	-17.69	6.19
B2	-0.76	-10.12	-10.80	-34.66	-12.36	6.35
C1	5.03	-13.99	-14.03	-30.54	-15.16	2.48
C2	-8.33	-21.47	-21.56	-41.50	-23.4	3.49
D1	2.28	-14.61	-14.59	-34.52	-15.85	2.90
D2	4.02	-6.82	-6.82	-35.65	-7.91	3.34
Avg.	**0.65**	**-13.62**	**-13.74**	**-35.09**	**-16.4**	**4.90**

V. Conclusion

This paper proposes a novel VVC-based quality scalable video coding solution offering significant RD performance gains regarding all the relevant benchmarks, notably the native VVC quality scalability solution. The proposed QS-VVC solution offers quality scalability capabilities by adopting a layered coding approach, extending the VVC capabilities with a new JLR frame to be added to the DPB. Future work will consider the design of a deep learning-based improved JLR creation process targeting to reduce the compression penalty associated to scalable coding regarding non-scalable coding.

REFERENCES

[1] Cisco Systems, "Cisco Visual Networking Index: Forecast and Trends, 2017–2022", Cisco Systems White Paper, 2019 (online at https://www.cisco.com/c/en/us/solutions/collateral/serviceprovider/visual-networking-index-vni/white-paper-c11-741490.pdf)

[2] J. M. Boyce, Y. Ye, J. Chen and A. K. Ramasubramonian, "Overview of SHVC: Scalable Extensions of the High Efficiency Video Coding Standard," in *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 1, pp. 20-34, Jan. 2016.

[3] J. Chen, Y. Ye, and S. H. Kim, "Algorithm Description for Versatile Video Coding and Test Model 8 (VTM 8)", Document: JVET-Q2002-v3, ITU-T SG 16 and ISO/IEC JTC 1/SC 29/WG 11 Meeting, Brussels, Belgium, Jan. 2020.

[4] G. J. Sullivan, J. Ohm, W. Han and T. Wiegand, "Overview of the High Efficiency Video Coding (HEVC) Standard," in *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1649-1668, Dec. 2012.

[5] N. Sidaty, W. Hamidouche, O. Déforges, P. Philippe, and J. Fournier, "Compression Performance of the Versatile Video Coding: HD and UHD Visual Quality Monitoring", *Picture Coding Symposium* (PCS), Ningbo, China, Nov. 2019.

[6] G. J. Sullivan, J. M. Boyce, Y. Chen, J. Ohm, C. A. Segall and A. Vetro, "Standardized Extensions of High Efficiency Video Coding (HEVC)," in *IEEE Journal of Selected Topics in Signal Processing*, vol. 7, no. 6, pp. 1001-1016, Dec. 2013.

[7] W. Bao, W. Lai, C. Ma, X. Zhang, Z. Gao and M. Yang, "Depth-Aware Video Frame Interpolation," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR*), Long Beach, CA, USA, 2019.

[8] T. Fu, X. Zheng, S. Wang and S. Ma, "Composite Long-Term Reference Coding for Versatile Video Coding (VVC)," *IEEE International Conference on Image Processing (ICIP)*, Taipei, Taiwan, 2019.

[9] X. HoangVan, J. Ascenso, and F. Pereira, "Improving Scalable Video Coding Performance with Decoder Side Information" *Picture Coding Symposium (PCS),* San Jose, CA, USA, Dec. 2013.

[10] V. Chaudhary, V. Kumar, "Block-based Image Fusion using Multi-scale Analysis to Enhance Depth of Field and Dynamic Range", *Signal, Image and Video Processing*, vol. 12, pp. 271-279, Feb. 2018.

[11] "VTM8.0," Feb. 2020. [Online]. Available: https://vcgit.hhi.fraunhofer.de/jvet/VVCSoftware_VTM.

[12] J. Boyce, K. Suehring, X. Li, and V. Seregin, "JVET Common Test Conditions and Software Reference Configurations," Document JVET-J1010-v1, ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11 10th Meeting: San Diego, US, Apr. 2018.

[13] G. Bjontegaard, "Calculation of Average PSNR Differences between RD Curves," Document VCEG-M33, 13th ITU-T VCEG Meeting, Austin, TX, USA, Apr. 2001.