The High-Level Syntax of the Versatile Video Coding (VVC) Standard

Ye-Kui Wang, Robert Skupin, Miska M. Hannuksela, *Member, IEEE*, Sachin Deshpande, *Senior Member, IEEE*, Hendry, Virginie Drugeon, Rickard Sjöberg, Byeongdoo Choi, Vadim Seregin, Yago Sanchez, Jill M. Boyce, *Fellow, IEEE*, Wade Wan, and Gary J. Sullivan, *Fellow, IEEE*

(Invited Paper)

***Abstract*— Versatile Video Coding (VVC), a.k.a. ITU-T H.266 | ISO/IEC 23090-3, is the new generation video coding standard that has just been finalized by the Joint Video Experts Team (JVET) of ITU-T VCEG and ISO/IEC MPEG at its 19th meeting ending on July 1, 2020. This paper gives an overview of the VVC high-level syntax (HLS), which forms its system and transport interface. Comparisons to the HLS designs in High Efficiency Video Coding (HEVC) and Advanced Video Coding (AVC), the previous major video coding standards, are included. When discussing new HLS features introduced into VVC or differences relative to HEVC and AVC, the reasoning behind the design differences and the benefits they bring are described. The HLS of VVC enables newer and more versatile use cases such as video region extraction, composition and merging of content from multiple coded video bitstreams, and viewport-adaptive 360° immersive media.**

***Index Terms*— Versatile video coding (VVC), H.266, MPEG-I, high-level syntax (HLS), network abstraction layer (NAL) unit, parameter sets, subpictures, slices, tiles, temporal scalability, scalability, profiles, tiers, levels, hypothetical reference decoder (HRD), video usability information (VUI), supplemental enhancement information (SEI).**

I. Introduction

CODED video content consists of a series of data structures that contain header syntax and supplemental information in addition to the compressed bits that directly represent color component samples. The handling of these data structures forms the system interface for operation of an encoder or decoder within a system environment, and this interface needs to support the functionalities that will be used by the system to enable the features of the application.

The Versatile Video Coding (VVC) standard (Rec. ITU-T H.266 | ISO/IEC 23090-3) [1] and the associated Versatile Supplemental Enhancement Information (VSEI) standard (Rec. ITU-T H.274 | ISO/IEC 23002-7) [2] have been designed for use in a maximally broad range of applications, including both the traditional uses such as television broadcast, video conferencing, or playback from storage media, and also newer and more advanced use cases such as adaptive bit rate streaming, video region extraction, composition and merging of content from multiple coded video bitstreams, multiview video, scalable layered coding, and viewport-adaptive 360° immersive media.

In the video coding standardization community, and specifically in the Joint Video Experts Team (JVET) of the ITU-T and ISO/IEC that developed the VVC standard, the system interface is modeled as a network abstraction layer (NAL), and the data structures carried by this interface are known as NAL units. A coded video bitstream consists of a series of NAL units. The NAL units that represent the values of color component samples are called the video coding layer (VCL) NAL units or coded slice NAL units. The compressed data within the VCL NAL units is called the slice data, and the slice data in each slice consists of a series of coded elemental regions of the picture called coding tree units (CTUs) that are sent according to a particular scanning order. CTUs are square regions of coded video pictures that contain samples of the luma and also the chroma color planes (when chroma is present, i.e., except with monochrome video). The CTUs in a coded picture have a size selected by the encoder, and the

Manuscript received August 25, 2020; revised February 3, 2021; accepted February 14, 2021. Date of publication April 5, 2021; date of current version October 4, 2021. This article was recommended by Associate Editor W. Zhu. *(Corresponding author: Ye-Kui Wang.)*

Ye-Kui Wang is with ByteDance, San Diego, CA 92122 USA (e-mail: yekui.wang@bytedance.com).

Robert Skupin and Yago Sanchez are with Fraunhofer Institute for Telecommunications, Heinrich Hertz Institute (HHI), 10587 Berlin, Germany (e-mail: robert.skupin@hhi.fraunhofer.de; yago.sanchez@hhi.fraunhofer.de).

Miska M. Hannuksela is with Nokia Technologies, 33100 Tampere, Finland (e-mail: miska.hannuksela@nokia.com).

Sachin Deshpande is with Sharp Labs of America, Camas, WA 98607 USA (e-mail: sdeshpande@sharplabs.com).

Hendry is with LG Electronics, San Diego, CA 92131 USA (e-mail: dr.hendry@lge.com).

Virginie Drugeon is with Panasonic, 63225 Langen, Germany (e-mail: virginie.drugeon@eu.panasonic.com).

Rickard Sjöberg is with Ericsson, 164 83 Stockholm, Sweden (e-mail: rickard.sjoberg@ericsson.com).

Byeongdoo Choi is with Tencent America LLC, Palo Alto, CA 94306 USA (e-mail: bdchoi@tencent.com).

Vadim Seregin is with Qualcomm Technologies, Inc., San Diego, CA 92121 USA (e-mail: vseregin@qti.qualcomm.com).

Jill M. Boyce is with Intel, Hillsboro, OR 97124 USA (e-mail: jill.boyce@intel.com).

Wade Wan is with Broadcom Inc., Irvine, CA 92618 USA (e-mail: wade.wan@broadcom.com).

Gary J. Sullivan is with Microsoft, Redmond, WA 98052 USA (e-mail: garysull@microsoft.com).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSVT.2021.3070860.

Digital Object Identifier 10.1109/TCSVT.2021.3070860

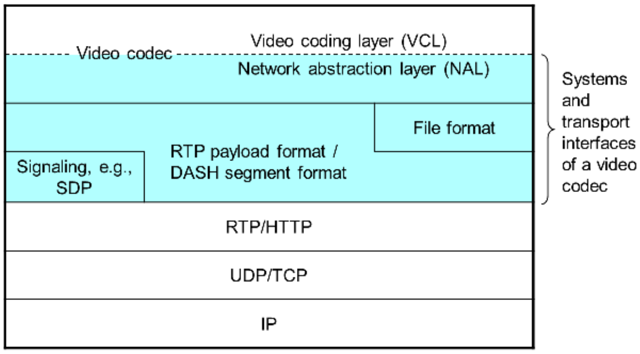
Video codec
Video coding layer (VCL)
Network abstraction layer (NAL)
File format
RTP payload format / DASH segment format
Signaling, e.g., SDP
RTP/HTTP
UDP/TCP
IP
Systems and transport interfaces of a video codec

Fig. 1. A typical protocol stack of an IP-based video application system.

CTUs in a picture all have the same size except for being truncated at the right and bottom edges of a picture if the picture size is not divisible by the CTU size. For VVC, the CTU size is up to 128 × 128 for the luma, and for high-resolution video, bigger sizes tend to work better. In HEVC, the same concept applies, with a maximum CTU size of 64 × 64. In AVC and several earlier standards dating back to Rec. ITU-T H.261 in 1988, the equivalent elemental structure is called a macroblock and its size is always 16 × 16. For purposes of this paper, we will refer to macroblocks as CTUs.

The structuring of the bitstream into NAL units and all of the syntax outside the slice data, including the headers that precede the slice data within the VCL NAL units and all of the syntax in the non-VCL NAL units, is considered the high-level syntax (HLS).

HLS topics thus include the basic structure of the bitstream and coded data structures, sequence and picture level parameters signaling, syntax for random access and stream adaptation, decoded picture management (including reference picture management), profile and level signaling, the bitstream buffering data flow model, high-level picture partitioning (slicing, tiling, etc.), temporal scalability, extensibility and backward compatibility, data loss resilience, signaling of supplemental information, etc.

The HLS of a video coding design forms much of the systems and transport interface to the application environment, e.g., as illustrated in Fig. 1. Depending on the context, the word "codec" may refer to either an encoder or a decoder (or some combination of encoders and decoders), or to a video coding format.

Conversational applications, such as video telephony and video conferencing, are typically based on the internet protocol (IP) [3], [4], user datagram protocol (UDP) [5], and real-time transport protocol (RTP) [6] transport mechanisms. For sending and/or receiving video with RTP/UDP/IP, an RTP payload format, such as the Advanced Video Coding (AVC) RTP payload format [7], the Scalable Video Coding (SVC) RTP payload format [8], or the High Efficiency Video Coding (HEVC) RTP payload format [9], is needed, which defines the format and rules for encapsulation of coded data units into RTP streams and packets, as well as specifies new and uses existing session description protocol (SDP) [10] parameters for session negotiation purposes. The session negotiation process between a sender and a receiver establishes agreed parameters, e.g., picture resolution, coding format, and format configuration, etc., to use for the communication.

Streaming applications, on the other hand, are typically based on the IP, transmission control protocol (TCP) [11], and hypertext transfer protocol (HTTP) [12] transport methods, and typically rely on a file format such as the ISO base media file format (ISOBMFF) [13]. One such streaming system is dynamic adaptive streaming over HTTP (DASH) [14]. For using a video codec with ISOBMFF and DASH, a media file format specific to the video codec, such as the AVC file format and the HEVC file format [15], would be needed for encapsulation of the video content in ISOBMFF tracks and in DASH representations and segments. Important information about the video bitstreams, e.g., the profile, tier, and level, and other properties, would need to be exposed as file format level metadata and/or DASH media presentation description (MPD) for content selection purposes, e.g., for selection of appropriate media segments both for initialization at the beginning of a streaming session and for stream adaptation during the session.

The HLS design needs to be able to provide information about the video bitstreams, e.g., to be signaled in SDP or DASH MPD, for session negotiation, content selection, and enabling various optimizations of application systems. Generally, the HLS of a video coding specification should be designed to:

- enable the use of the video content in different application systems, including the ones described above and others (e.g., traditional digital television broadcast based on the MPEG-2 transport system [16]);
- provide flexible random accessing and stream adaptation capabilities while keeping high coding efficiency;
- provide data loss resilience while retaining high coding efficiency;
- ensure interoperability; and
- provide extensibility and backward compatibility.

For VVC, the HLS specifically includes the NAL unit syntax (including the NAL unit header), decoding capability information (DCI), operating point information (OPI), video parameter set (VPS), sequence parameter set (SPS), picture parameter set (PPS), adaptation parameter set (APS), video usability information (VUI), supplemental enhancement information (SEI), access unit delimiter (AUD), picture header (PH), slice header (SH), end of sequence (EOS), end of bitstream (EOB), etc. CTU-level and lower-level coding tool syntax are not considered HLS.

VVC inherited much of its HLS design from the preceding AVC and HEVC standards. These include the NAL-unit-based syntax structure, the hierarchical syntax and data unit structure, the VUI and SEI message mechanism, and the video buffering model based on a hypothetical reference decoder (HRD). The hierarchical syntax and data unit structures consist of sequence-level parameter sets, multi-picture-level or picture-level parameter sets, slice-level header parameters, and lower-level parameters. The VUI payload and additional SEI messages are specified in the VSEI standard [2], an accompanying standard that was developed in conjunction with VVC. Key HLS features in VVC that are new or significantly different compared to AVC and HEVC (and their extensions) include the following: subpictures and rectangular slices (Section VII), support of picture resolution changes at inter-coded pictures (Section IX), APSs (Section IV-C), PHs (Section IV-D), gradual decoding refresh (GDR, Section V), direct signaling of reference picture lists (RPLs, Section VI-B), and significantly simplified multi-layer support (Section X).

The remainder of this paper is organized as follows. Section II presents a brief introduction to the basics of most HLS aspects and the similarities on these aspects between VVC, HEVC and AVC. Sections III through XII discuss various particular HLS features of VVC and their differences compared to earlier video coding standards. Finally, Section XIII concludes the paper.

For simplicity, unless otherwise stated, in the text below "VVC" refers to the first version of VVC (which includes some features to enable future extensibility). At the time of writing this paper, no substantial changes to the HLS design concepts are planned for future versions of the standard.

II. HLS Basics and Similarities in HLS Between VVC, HEVC and AVC

VVC has inherited many aspects of the HLS designs of AVC [17] and HEVC [18]. This section offers an overview of the HLS aspects that have such similarities, while pointing out significant differences. More details on these can be found in [19]–[21].

A. Bitstream Structure

As in AVC and HEVC, a bitstream in VVC consists of one or more coded video sequences (CVSs). A CVS is independently coded from other CVSs. Each CVS consists of one or more layers, each of which is a representation of the video with a specific quality or spatial resolution, or a representation of some component interpretation property, e.g., as depth or transparency maps or perspective views. In another dimension, each CVS consists of one or more access units (AUs), and each AU consists of one or more picture units (PUs) of different layers. Fig. 2 illustrates an example structure of a CVS. A coded layer video sequence (CLVS) is a layer-wise CVS that consists of a sequence of PUs in the same layer. If a bitstream has multiple layers, a CVS in the bitstream has one or more CLVSs for each layer. Otherwise, a CVS is identical to a CLVS. In Fig. 2, an AU is represented by a vertical group of PUs belonging to the same time instant, while a CLVS is represented by a horizontal group of PUs belonging to the same layer that spans across several AUs.

Each PU contains one coded picture, and each coded picture comprises one or more coded slice NAL units, which are also referred to as VCL NAL units. In addition to coded slice NAL units, a PU may contain non-VCL NAL units, such as parameter sets and SEI NAL units. A simplified structure of a PU is presented in Fig. 3. NAL units are typically ordered in a PU as follows (some NAL units are optional): DCI, VPS, SPS, PPS, prefix APS, PH, prefix SEI, VCL, suffix SEI, suffix APS, and EOS. Additionally, an AUD NAL unit and OPI NAL unit are allowed only at the start of an AU and an EOB NAL unit is allowed only at the end of an AU.

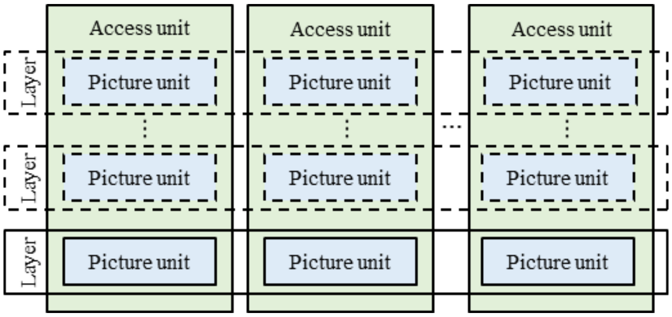

Fig. 2. Structure of coded video sequence. Dashed boxes indicate optionally present structures.

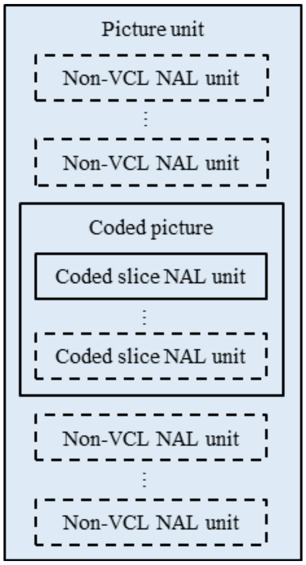

Fig. 3. Simplified structure of picture unit. Dashed boxes indicate optionally present structures.

B. NAL Units

A VVC NAL unit consists of a two-byte NAL unit header and a NAL unit payload, as in HEVC. However, the syntax of the NAL unit header is slightly different compared to HEVC. The NAL unit header in both HEVC and VVC includes a layer identifier (ID), a NAL unit type, and a temporal ID, which are important for both the decoding process and systems usage. For instance, some systems may utilize the information carried in the NAL unit header to perform actions such as accessing a bitstream from a specific point onwards or stream adaptation through (temporal) layer pruning during network transmission [20], [22].

C. Random Access Support

Random access refers to starting to access and decode a bitstream from an access unit that is not the first access unit of the bitstream in decoding order. To support tuning in and channel switching in broadcast/multicast and multiparty video conferencing, seeking in local playback and streaming, as well as stream adaptation in streaming, it is sensible to include random access points (RAPs) that control prediction dependencies on AUs preceding each particular RAP in the bitstream, at application-specific frequencies.

VVC includes signaling of intra RAP (IRAP) pictures as in HEVC and can also identify GDR pictures in the NAL unit header through dedicated NAL unit types, wherein the GDR picture signaling is new in VVC. The different IRAP picture types can be used to match the stream access point types as defined in the ISO base media file format (ISOBMFF) [13], which are utilized for random access support in DASH [14].

As with several prior video coding standards, VVC supports the ability for an encoder to change the ordering of pictures, so that the order in which the pictures appear in the bitstream, which is the order in which the pictures are processed by the decoder and is known as the decoding order. After decoding, the pictures are then reordered again within the decoder so they can be output from the decoder in the original order in which they were input to the encoder (e.g., from a camera), which is known as the output order.

VVC also includes the concept of identification of leading pictures of an IRAP picture, as in HEVC, which are pictures that follow the IRAP picture in decoding order but precede the IRAP picture in output order. There are two types of leading pictures: random access decodable leading (RADL) pictures and random access skipped leading (RASL) pictures. RADL pictures are decodable when the decoding starts at the associated IRAP picture; RASL pictures are not guaranteed to be decodable when the decoding started at the associated IRAP picture and are usually discarded [19].

Some pictures types, such as broken link access (BLA) and temporal sublayer access (TSA), as well as picture type differentiation into reference pictures and sublayer non-reference pictures (e.g. TRAIL_N, TRAIL_R), that were in HEVC were not carried over to VVC. This was partly because some of these properties could be indicated by other syntax combinations or through proper bitstream handling in systems, and there was also a desire to simplify the HLS by specifying fewer NAL unit types in VVC than in HEVC – e.g., there are only five instead of six bits for the NAL unit type field in the VVC NAL unit header.

In addition to providing RAP support for the decoding of sequences of whole pictures, VVC additionally supports RAPs for sequences of rectangular regions of pictures known as subpictures, as further described in Section VII-C.

D. Video Parameter Set (VPS)

VVC bitstreams may contain a video parameter set (VPS) containing information about layers and output layer sets (OLS) that is necessary for operation of the decoding process of scalable bitstreams. An OLS is a set of layers in the bitstream among which one or more layers are specified to be output from the decoder (other layers identified in the OLS may also need to be decoded in order to decode the output layers, although they themselves are not specified to be output). Much of the information contained in the VPS can be used in systems for purposes such as session negotiation and content selection. The VPS was introduced for handling multi-layer bitstreams. For single-layer VVC bitstreams, the presence of the VPS in the CVS is optional, since the information contained in the VPS is not necessary for the operation of the decoding process of the bitstream. Its absence in a CVS is indicated by referencing a VPS ID equal to 0 in the SPS, in which case simple default values are inferred for the VPS parameters. This is different from HEVC, where the VPS is always required to be present (either within the bitstream or conveyed by some external means, such as the system application configuration).

E. Sequence Parameter Set (SPS)

The SPS conveys sequence-level information shared by all pictures in an entire CLVS. This includes profile, tier, and level (PTL) indicators, picture format (the color sampling format, maximum picture width, maximum picture height, and bit depth), feature/tool control flags, coding/prediction/transform block structures and hierarchies, candidate RPLs that may be referenced by the encoder, etc. In most applications, there would be only one or a few SPSs for an entire bitstream, and thus there would be no need to update an SPS (i.e., to send a new SPS using the SPS ID of an existing SPS but with different values for certain parameters) within the bitstream. Pictures from a particular layer that refer to SPSs with different SPS IDs or with the same SPS ID but with different SPS content belong to different CLVSs. As in AVC and HEVC, SPSs can be transported in-band (i.e., transported together with the coded pictures), out-of-band (i.e., not transported together with the coded pictures), or using a mixture of in-band and out-of-band signaling.

F. Picture Parameter Set (PPS)

The PPS conveys picture-level information that is shared by all slices of a picture, and may also be shared across multiple pictures. This includes feature/tool on/off flags, picture width and height, default RPL sizes, configurations of tiles and slices, etc. While by design two consecutive pictures can refer to two different PPSs, which consequently may lead to a large number of PPS being used within a CLVS, in practice, it is expected that the number of PPSs for an entire bitstream is not high, since the PPS is designed to carry parameters that do not change frequently and are likely to apply to multiple pictures. Therefore, typically there is no need to update a PPS within a CLVS or even within an entire bitstream. For parameters that could apply to multiple pictures but are expected to change frequently from picture to picture, a new type of parameter set called an adaptation parameter set (APS), which is discussed in Section IV-C, was introduced for VVC. Like SPSs, PPSs can be transported in-band, out-of-band, or using a mixture of in-band and out-of-band signaling. One basic design principle regarding which picture-level parameters should be included in the PPS versus which should be in the APS is the frequency at which such parameters are likely to change, so that frequently changing parameters are not included into the PPS, in order to avoid requiring PPS updates, which would disallow out-of-band transmission of PPSs in typical use cases.

G. Slice Header (SH)

The SH conveys information for a particular slice, which is the set of the CTUs in a picture that is conveyed in a particular scanning order within a single VCL NAL unit. The SH in VVC was designed aiming to convey information that applies only to the slice and may be different for other slices of the same picture. Information that applies to all slices of a picture is conveyed in another header called the PH, which is explained in more detail in Section IV-D.

H. Reference Picture Management

Reference picture management is a core functionality that is necessary for any video coding scheme that uses inter prediction with multiple reference pictures. It manages the storage and removal of reference pictures into and from the decoded picture buffer (DPB), and puts the reference pictures in a proper order in the RPLs. When a picture is referenced for inter-picture prediction (either as a temporal reference or an inter-layer reference), an index into an RPL is used to select which picture in the DPB is being referenced.

The reference picture management designs of VVC, HEVC and AVC all share the concepts of a DPB, the use of up to two RPLs, and the concept of marking reference pictures as "used for short-term reference", "used for long-term reference" and "unused for reference". The details of the VVC reference picture management scheme differ from that of both AVC and HEVC, but are closer to what is done in HEVC. This is explained in more detail in Section VI.

I. Temporal Scalability Support

VVC includes support of temporal scalability in a similar way as in HEVC. Such support includes the signaling of a temporal ID in the NAL unit header, the restriction that pictures of a particular temporal sublayer cannot be used for inter prediction referencing by pictures of a lower temporal sublayer, a specified sub-bitstream extraction process for extracting a temporal subset of the bitstream, and the requirement that each sub-bitstream extraction output of an appropriate input bitstream must be a conforming bitstream. Media-aware network elements (MANEs) can utilize the temporal ID in the NAL unit header for stream adaptation purposes based on this temporal scalability.

J. High-Level Picture Partitioning

VVC includes four different high-level picture partitioning schemes, namely subpictures (presented in detail in Section VII-C), slices (presented in detail in Section VII-B), tiles, and wavefront parallel processing (WPP).

Slices in VVC have a different form from the slices specified in AVC and HEVC, but the concept that each slice contains a region of the coded picture and is encapsulated in its own NAL unit was kept, as well as the property that in-picture prediction (including intra sample prediction, motion information prediction, and coding mode prediction) and entropy coding dependencies across slice boundaries are disabled. Thus, a slice can be reconstructed independently from the other slices within the same picture (although there may still exist some relatively minor sample value interdependencies near the boundaries of a slice due to loop filtering operations). As in AVC and HEVC, there are also three basic types of slices in VVC: I, P and B slices. I slices use only intra prediction, P slices may use both intra prediction and inter prediction with one RPL, and B slices may use both intra prediction and inter prediction with two RPLs.

Tiles were introduced in HEVC, and are very similar in VVC. Horizontal and vertical boundaries are used to partition a picture into tile columns and rows. A tile row spans from the left edge of the picture to the right edge of the picture, and likewise a tile column runs from the top of a picture to the bottom of the picture. The number of tiles in a picture can be derived simply as the number of tile columns multiplied by the number of tile rows. The number and positions of the tiles are specific to the individual picture, and thus the tile structure can change from picture to picture within a CLVS. The scan order of CTUs in a slice is established to be local within a tile (in the order of a CTU raster scan of a tile) before proceeding to scan the CTUs of the next tile when applicable. In a manner similar to slices, tiles break in-picture prediction dependencies as well as entropy decoding dependencies. However, tiles do not need to be included into individual NAL units. Each tile can be processed by one processor/core by an encoder or decoder, and the inter-processor/inter-core communication required for in-picture prediction between processors for decoding neighboring tiles is limited to conveying the shared SH data in cases where slice is spanning more than one tile, and the sharing of reconstructed samples and metadata near the edges of the tiles for operation of in-loop filtering after completing the prediction and residual decoding stages of the tile decoding process.

In WPP, a tile is partitioned into single rows of CTUs. Although the term WPP is not used explicitly in the HEVC specification, the WPP functionality can be enabled by a flag called sps_entropy_coding_sync_enabled_flag in the SPS. In WPP, the entropy coding statistics for one WPP partition can be based on data from CTUs in other WPP partitions within the same picture but this sharing of data is limited to a diagonal cascading of dependencies. Parallel processing is possible through parallel decoding of CTU rows, where the start of the decoding of each next CTU row is delayed by at least one CTU, so as to ensure that necessary data related to the CTUs above the CTU being processed is available before the subject CTU is decoded. Using this staggered start (which appears like a wavefront when represented graphically), parallelization is possible with up to as many processors/cores as the number of CTU rows in the tile. Because in-picture prediction between neighboring CTU rows within a picture is permitted, the required inter-processor/inter-core communication to enable in-picture prediction can be substantial, so the parallel processes need to be relatively closely coupled. However, the use of WPP does not incur additional NAL units and typically does not incur a large penalty in coding efficiency.

When more than one tile or WPP partition is included in a slice, an entry point byte offset for starting the decoding of each tile or WPP partition other than the first one in the slice may be signaled in the slice header to facilitate convenient allocation of coded data to different processing cores in parallel decoding and to enable the decoding of CTUs in raster-scan order of the entire picture if desired, which was asserted to be what some hardware decoder designs would do even when the bitstream order of the CTUs is not in raster-scan order.

K. Profile, Tier, and Level (PTL)

In order to restrict the feature set to what is needed for a particular group of applications, video coding standards define *profiles*, which are defined decoder feature sets to be supported for interoperability with encoders that use these features. In addition to profiles, VVC (like HEVC) also defines *levels* and *tiers*. A level imposes restrictions on the bitstream (values of syntax elements and their arithmetic combinations), related to spatial resolution, pixel rate, bit rate values and variations, etc., with higher values of level corresponding to higher complexity limits. Tiers modify the bit rate value and variation limits for each level. The Main tier is intended for most applications, while the High tier is designed to address video contribution applications that have significantly higher bit rate values than video distribution applications. Each of profile, tier, and level affects the implementation and decoding complexities, and a combination of the three specifies an interoperability point for bitstreams and decoders.

L. Hypothetical Reference Decoder (HRD)

The HRD specifies a buffering model, including operations for both the coded picture buffer (CPB) and DPB. While the DPB is a memory buffer that holds uncompressed pictures for referencing or output reordering after the pictures have been decoded, the CPB is a buffer of compressed bits that temporarily holds the coded data coming from the bitstream until it is removed from the CPB for decoding. Through the parameters and the operations, the HRD directly imposes constraints on different timing, buffer sizes, and bit rate values, which indirectly imposes constraints on bitstream characteristics and statistics. For example, the level definitions rely on the HRD process. There are five basic HRD parameters, namely the initial CPB removal delay, CPB size, bit rate, initial DPB output delay, and DPB size. Bitstream conformance and decoder conformances (output order decoder conformance and timing decoder conformance) are also specified as part of the HRD specification. Although the name includes the word "decoder", the HRD is typically implemented along with the encoder, to guarantee that the generated bitstream is conforming.

AVC specifies AU-based HRD operations only. HEVC and VVC HRD additionally specify decoding unit (DU) based HRD operations. The DU based HRD operation was introduced to provide better support for ultralow-delay applications. It specifies a conforming behavior for encoders to send a slice of a picture before the encoding of other slices of the same picture, as well as for decoders to be able to start decoding a received slice before receiving other slices of the same picture. Decoders are allowed to operate the HRD at the conventional AU level even when the DU-level HRD parameters are present. More details on the HRD operation for AVC can be found in [17] and [23], and more about the HRD operation for HEVC can be found in [18] and [24].

M. VUI and SEI

The VUI is a syntax structure sent as part of the SPS (and in HEVC, possibly also in the VPS). The VUI carries information that does not affect the operation of the signal processing steps of the decoding process, but can be important for proper interpretation of the video pictures after they are decoded. For example, the VUI may indicate how the samples of the decoded pictures are intended to be converted into light for display with an accurate rendition of brightness and color hue and saturation.

SEI assists in processes related to decoding, display or other purposes. Like the VUI, the SEI does not affect the signal processing operations within the decoding process. SEI syntax for various purposes is carried in syntax structures called SEI messages, and one or more SEI messages are carried within NAL units called SEI NAL units. SEI messages can have a very high level of scope like that of the VUI, or may have a narrower scope, such as applying to an individual picture or slice. As their name implies, SEI messages are intended to be *supplemental* to the video content. Decoder support of SEI messages is generally optional, and even if a decoder uses an SEI message, in most cases the decoder is not required to use an SEI message exactly in the way it is described in the standard. However, SEI messages do affect bitstream conformance (e.g., if the syntax of an SEI message in a bitstream does not follow the specification, then the bitstream is not conforming to the standard) and some SEI messages are needed for specifying the HRD operations and HRD-based bitstream conformance requirements.

III. Profiles, Tiers and Levels Information

A. Profile; Tier, Level (PTL)

VVC v1 defines six profiles, as described in Table I. These profiles were each defined to address a broad range of applications, and they have logical "nesting" relationships to each other.

In Table I, the term "4:2:0" refers to video with chroma color planes that have half the width and half the height of the luma plane, which is the most common format used for encoding camera-captured content in consumer applications. "4:4:4" refers to using chroma color planes that have the same width and the same height as the luma plane, as commonly used for graphics, display monitors and computer desktop rendering. "4:2:2" is a less common format with chroma that has half the width of the luma but the same height, e.g., as has often been used in studios for interlace-captured video. Monochrome video has only a single color plane, which is referred to as the luma plane (although it may not actually represent a luma signal – for example, a monochrome picture may represent a depth map for 3D video applications or

TABLE I
THE SIX PROFILES DEFINED IN VVC V1

Profile name	Remarks
Main 10	• Monochrome and 4:2:0 chroma sampling only • 8 to 10 bits bit depth • The bitstream has only one layer
Main 10 Still Picture	Same as Main 10 profile but the bitstream contains only a single picture (no inter-picture prediction)
Main 10 4:4:4	Same as Main 10 profile but additionally supports 4:2:2 and 4:4:4 chroma sampling and the palette and adaptive color transform coding tools
Main 10 4:4:4 Still Picture	Same as Main 10 4:4:4 profile but the bitstream contains only a single picture (no inter-picture prediction)
Multilayer Main 10	Same as Main 10 profile but the bitstream can have more than one layer
Multilayer Main 10 4:4:4	Same as Main 10 4:4:4 profile but the bitstream can have more than one layer

a transparency map for the overlay of decoded video on a background).

Note that a decoder that conforms to a still picture profile shall be capable of decoding the first picture of a typical video profile bitstream, since the first picture of a bitstream is ordinarily an intra-coded picture. Such a decoder could also decode other IRAP pictures that are extracted as snapshots from video bitstreams. Having this subset relationship between the video and still picture profiles provides the ability to share encoder and decoder modules for use in different applications (and there has been a converging trend to the point where most newer video cameras can also be used for still-image photography and vice versa).

The PTL information is signaled using a syntax structure that can be included in the VPS (in which case, the PTL structure applies to one or more OLSs) or the SPS (in which case, the PTL structure applies to the OLS that contains only the layer referring to that SPS). This includes a PTL to which the bitstream conforms, as well as additional PTLs for the temporal sublayer representations, each of which is a self-contained subset of the bitstream.

B. Externally Specified Sub-Profiles

In addition to the profiles defined in the VVC specification as shown in Table I, VVC also allows the encoder to signal bitstream conformance to sub-profiles. A sub-profile is an interoperability subset indicator, similar to a profile, that imposes further restrictions on an existing indicated profile. Such sub-profiles would be defined outside the VVC specification and indicated using an identifier code that is registered as specified by Rec. ITU-T T.35 in order to avoid having multiple defined meanings for the same sub-profile code value [25]. External organizations may define their own sub-profiles which they believe are sufficient to fulfill the needs of their specific applications. The sub-profile indicator syntax element within the PTL structure enables signaling within a bitstream that the bitstream conforms to such externally defined restrictions.

C. General Constraints Information (GCI)

In addition to the PTL information, the PTL syntax structure may also optionally include a general constraints information (GCI) syntax structure, which contains a list of constraint flags and non-flag syntax elements indicating specific constraint properties of the bitstream. When present, a GCI syntax element value greater than 0 indicates that the bitstream is constrained in a particular way, typically to indicate that a particular coding tool is not used in the bitstream, whereas the value 0 signals that the associated constraint may not apply, such that the associated coding tool is allowed (but not required) to be used in the bitstream (if its use is supported in the indicated profile).

The GCI structure contains several types of constraint syntax elements, including:

- Flags for general bitstream restrictions, such as indicating that only intra coding is being used, that all layers are coded independently or that the bitstream contains only one AU;
- Fields constraining the bit depth and chroma format of the coded pictures;
- Flags indicating that certain NAL unit types are not allowed to be present within the bitstream;
- Flags constraining the ways that the pictures can be partitioned into slices, tiles, and subpictures within the bitstream;
- Flags constraining the size of CTUs, as well as the size and type of partitioning trees;
- Flags constraining the use of particular intra coding tools;
- Flags constraining the use of particular inter coding tools;
- Flags constraining the transform, quantization, and residual coding tools; and
- Flags constraining aspects of in-loop filters.

The purpose of the GCI syntax structure is to enable the simple discovery of configuration information about the features needed for decoding the bitstream and to allow the signaling of interoperability points which impose restrictions beyond those specified by the PTL, with a finer granularity than allowed by previous video coding standards. Similar to sub-profiles, use of the GCI syntax structure could allow interoperability to be defined for decoder implementations that do not support all features of a VVC profile but address the needs of particular applications. Decoder implementations may examine the GCI syntax elements to check if a bitstream avoids the use of particular features, in order to determine how to configure the decoding process and identify whether the bitstream is decodable by the decoder. Decoder implementations that support all features of a VVC profile can ignore the GCI syntax element values, as such decoders will be capable of decoding any bitstream conforming to the indicated PTL.

Unlike the sub-profile indicator whose semantics are defined externally to the VVC specification, the semantics of the GCI syntax elements are defined within the VVC specification. Sub-profiles can also be used in combination with the GCI, with a sub-profile imposing constraints on the values of the GCI syntax elements. The use of the GCI, either together with a sub-profile indicator or instead of it, can avoid the possibility that the meaning of the sub-profile indicator might be unrecognized by a decoder (as there is no requirement that the meaning of a sub-profile indicator needs to be published).

IV. New NAL Units and Syntax Structures

A. Decoding Capability Information (DCI)

The DCI NAL unit contains bitstream-level PTL information. It includes one or more PTL syntax structures that can be used during session negotiation between sender and receiver of a VVC bitstream. When the DCI NAL unit is present in a VVC bitstream, each OLS in the CVSs of the bitstream shall conform to the PTL information carried in least one of the PTL structures in the DCI NAL unit.

In AVC and HEVC, the PTL information for session negotiation is available in the SPS and for the HEVC layered coding extension, it is available in the VPS. This design of conveying the PTL information for session negotiation in AVC and HEVC has a disadvantage because the scope of SPS and VPS is only within a CVS, instead of applying to the whole bitstream. Because of this, sender-receiver session initiation may suffer from re-initialization during bitstream streaming at every new CVS. DCI solves this problem by providing a way to indicate whole-bitstream-level information, so that the conformance to the indicated decoding capability can be guaranteed until the end of the entire bitstream.

B. Operating Point Information (OPI)

The decoding processes of HEVC and VVC have similar input variables to set the decoding operating point, i.e., the target OLS and the highest temporal sublayer of the bitstream to be decoded, through a decoder API. However, in scenarios where layers and/or sublayers of the bitstream are removed during transmission or a device does not expose the decoder API to the application, it could occur that a decoder may not be correctly informed about the operating point for decoder to process the given bitstream. Hence, the decoder may not be able to conclude on the properties of pictures in the bitstream, e.g., proper buffer allocation for decoded pictures as well as whether individual pictures are output or not. In order to address this issue, VVC adds a way of indicating these two variables within the bitstream through the newly introduced operating point information (OPI) NAL unit. In the AUs at the beginning of the bitstream and its individual CVSs, the OPI NAL unit informs the decoder about the target OLS and the highest temporal sublayer of the bitstream to be decoded.

In the case when the OPI NAL unit is present and the operating point is also provided to the decoder via decoder API information (e.g., the application may have more updated information about the target OLS and sublayer), the decoder API information takes precedence. In absence of both a decoder API and any OPI NAL unit in the bitstream, suitable fallback choices are specified in the VVC standard.

C. Adaptation Parameter Set (APS)

Adaptation parameter sets (APSs) convey picture- and/or slice-level information that may be shared by multiple slices of a picture, and/or by slices of different pictures, but could change frequently from picture-to-picture and for which the total number of variants could be very high thus not suitable for inclusion into the PPS. Three types of parameters are included in APSs: adaptive loop filter (ALF) parameters, luma mapping with chroma scaling (LMCS) parameters, and quantization scaling list parameters. Depending on the type of data they carry, APSs can be carried in two distinct NAL unit types, either preceding or succeeding the associated slices as a prefix or suffix. Using suffix NAL units could be helpful for ALF parameters, since a typical way for an encoder to operate, especially in low-delay use cases, would be to use the statistics of the current picture to generate the ALF parameters to apply to subsequent pictures in decoding order.

D. Picture Header (PH)

The use of a PH is a rather simple concept that was used in older standards such as MPEG-2 but was not used in AVC and HEVC since the PPS and SH were sufficient to convey the picture-level information in those standards. In VVC, a PH structure is present for each PU. A PH is present either in a separate PH NAL unit or as syntax included in the slice header (SH). The PH can only be included in the SH if the entire PU contains only one slice. To simplify the design, within a CLVS, PHs can only be either all in PH NAL units or all in SHs. When the PHs are in the SHs, there is no PH NAL unit in the CLVS.

The PH is designed for two objectives. The first is to help reduce the signaling overhead of SHs for pictures containing multiple slices per picture, by carrying syntax elements that have the same value for all slices of a picture, thus avoiding the repetition of these syntax elements in each SH and avoiding moving syntax into the PPS that might change frequently from picture to picture. These include IRAP/GDR picture indications, inter and intra slice type allowance flags, and information related to picture order count (POC), RPLs, deblocking filter, sample adaptive offset (SAO), ALF, LMCS, quantization scaling lists, quantization parameter (QP) delta, weighted prediction, coding block partitioning, virtual boundaries, identification of a picture for collocated picture prediction properties, etc. The second purpose is to help the decoder to identify the first slice of each coded picture that contains multiple slices. Since one and only one PH is present for each PU, when the decoder receives a PH NAL unit, it easily knows that the next VCL NAL unit is the first slice of a picture.

V. Random Access Support and Picture Types

VVC supports three types of IRAP pictures which have distinct NAL unit types: two types of IDR pictures (one type without leading pictures and one type that may have associated RADL pictures) and one type of CRA picture. IDR pictures are conventionally referred to as closed group-of-pictures (GOP) RAPs whereas CRA pictures are conventionally referred to as open-GOP RAPs. These are basically the same as in HEVC.

A key difference in random access support between VVC and HEVC is that GDR is specified in VVC in a way that affects the required decoding process rather than being only a metadata property indication. This difference was necessary to enable efficient GDR operation in the context of the rest of the VVC design, and it also enables improved system support, such as by allowing a CVS to start with a GDR picture. Using the GDR feature (in AVC, HEVC, or VVC), the decoding of a bitstream can start from an inter-coded picture, whereby only a region of the inter-coded picture can be correctly decoded without referring to previous pictures. Although when beginning the decoding process with the decoding of a GDR picture, some areas of the picture cannot be correctly decoded, after decoding a number of additional pictures referred to as the recovery period, the entire picture for the recovery point and all subsequent pictures in output order would be correctly decoded.

AVC and HEVC support GDR with metadata using a recovery point SEI message for the signaling of GDR RAPs and recovery points. In VVC, a new NAL unit type is specified for indication of GDR pictures and the recovery point is signaled in the PH syntax structure. A CVS and a bitstream are allowed to start with a GDR picture. This means that it is allowed for an entire bitstream to contain only inter-coded pictures without a single intra-coded picture. GDR enables encoders to smooth the bit rate of a bitstream by distributing intra-coded slices or blocks in multiple pictures as contrasted with intra coding entire pictures, thus allowing significant end-to-end delay reduction, which is especially important for ultralow-delay applications like wireless display, online gaming, and remote-control drone operation.

Another new feature in random access support in VVC is that it allows random accessibility at the subpicture level, instead of always at picture level, as discussed in Section VII-C.

VI. Reference Picture Management

A. Picture Order Count (POC)

In HEVC and VVC, the POC is a variable that is derived as an output order indicator and is basically used as a picture ID for identification of pictures in many parts of the decoding process, including DPB management, part of which is reference picture management. To minimize signaling overhead bit costs while maintaining robustness against data losses, the most significant bits (MSBs) of the POC value may not be sent in the bitstream, and instead only the differences between POC values are often really necessary for proper operation of the decoding process.

With the use of a PH in VVC, the POC least significant bits (LSBs), which are used for deriving the POC value and have the same value for all slices of a picture, are signaled in the PH, as contrasted with HEVC where they are signaled in the SH. VVC also allows the signaling of the POC MSB cycle value in the PH, to enable the derivation of the POC value without tracking POC MSBs in a way that relies on the POC information of earlier decoded pictures. This, for example, allows the mixing of IRAP and non-IRAP pictures within an AU in multi-layer bitstreams. An additional difference between the POC signaling in HEVC and VVC is that in VVC there is POC LSB information that is signaled for every picture, including IDR pictures. In HEVC the POC LSBs are not signaled for IDR pictures, which saves a few bits for the IDR pictures but turned out to show some disadvantages for enabling the mixing of IDR and non-IDR pictures within an AU during later development of the multi-layer extensions of HEVC. The signaling of POC LSB information for IDR pictures also makes it easier to support the merging of IDR pictures and non-IDR pictures from different bitstreams into a single coded picture; without this, the handling of the POC LSBs in the merged picture would need some more complicated design.

B. Reference Picture List Signaling and Reference Picture Marking

In VVC, for all types of slices (i.e., B, P, and I slices), two RPLs, called list 0 and list 1 (herein referred to as RPL 0 and RPL 1), are directly signaled and derived, without using an RPL initialization or modification process. The RPLs are not based on reference picture sets [19] as in HEVC or based on the sliding window plus memory management control operation processes [26] as in AVC.

Reference picture marking is directly based on RPLs 0 and 1, indicating both active and inactive entries in the RPLs, where only the active entries may be used by reference indices in inter prediction of the current picture. Fig. 4 shows various syntax structures and elements related to reference picture management signaling that are included in the SPS, PPS, PH, and SH. In Fig. 4, the syntax elements that are always present are shown in solid rectangles and those that are conditionally present are shown in dotted rectangles. A number of predefined candidate RPL syntax structures (i.e. the ref_pic_list_struct (listIdx, rplsIdx) syntax structures) may be signaled in the SPS, for RPL 0 and for RPL 1 (if different from those for RPL 0), for use by referencing them in the PH or SH.

The syntax elements in the PPS indicate the default number of active entries for RPL 0 and RPL 1, a flag to control the presence of RPL 1 syntax in the ref_pic_lists( ) structure, and a flag that specifies whether the RPL information is included in the PH or SH. Information for the derivation of the two RPLs (i.e. the ref_pic_lists( ) structures) is signaled either in the PH, if all slices of the picture have the same RPLs, or in the SH. Instead of referencing a predefined candidate RPL structure (identified via rpl_idx[i]), a new RPL structure (i.e. a ref_pic_list_struct(i, sps_num_ref_pic_lists[i]) structure) can also be directly signaled in the PH and SH.

Each RPL syntax structure includes information for a number of reference picture entries for the particular

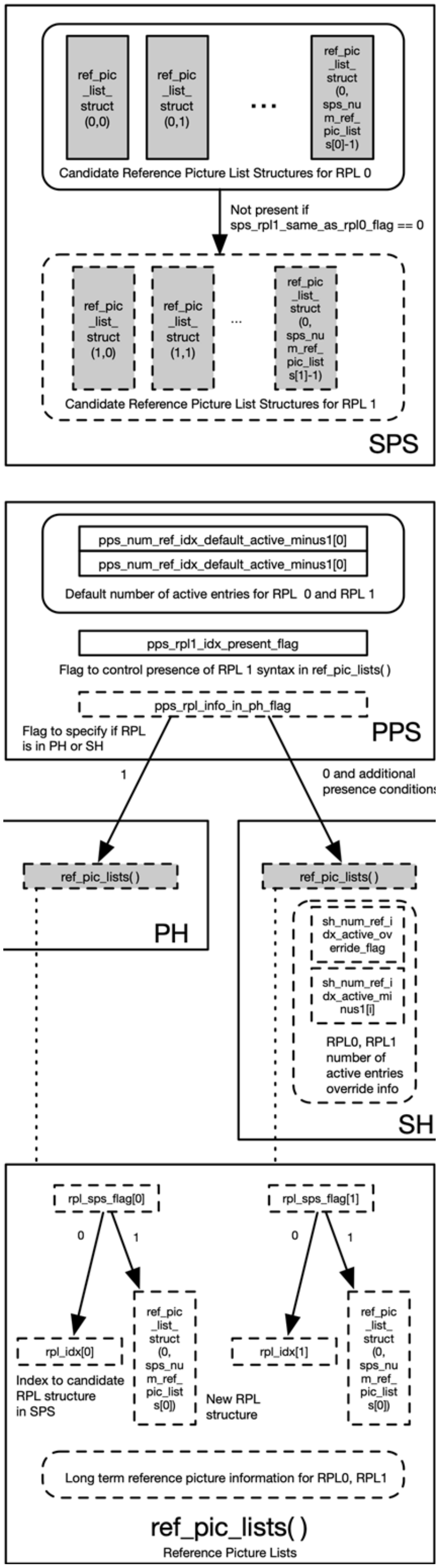
ref_pic _list_ struct (0,0)
ref_pic _list_ struct (0,1)
...
ref_pic _list_ struct (0, sps_num_ref_pic_lists[0]-1)
Candidate Reference Picture List Structures for RPL 0
Not present if sps_rpl1_same_as_rpl0_flag == 0
ref_pic _list_ struct (1,0)
ref_pic _list_ struct (1,1)
...
ref_pic _list_ struct (0, sps_num_ref_pic_lists[1]-1)
Candidate Reference Picture List Structures for RPL 1
SPS
pps_num_ref_idx_default_active_minus1[0]
pps_num_ref_idx_default_active_minus1[0]
Default number of active entries for RPL 0 and RPL 1
pps_rpl1_idx_present_flag
Flag to control presence of RPL 1 syntax in ref_pic_lists( )
pps_rpl_info_in_ph_flag
Flag to specify if RPL is in PH or SH
PPS
1
0 and additional presence condition
ref_pic_lists( )
PH
ref_pic_lists( )
sh_num_ref_idx_active_override_flag
sh_num_ref_idx_active_minus1[i]
RPL0, RPL1 number of active entries override info
SH
rpl_sps_flag[0]
0
1
rpl_sps_flag[1]
0
1
rpl_idx[0]
Index to candidate RPL structure in SPS
ref_pic _list_ struct (0, sps_num_ref_pic_lists[0])
New RPL structure
rpl_idx[1]
ref_pic _list_ struct (0, sps_num_ref_pic_lists[0])
Long term reference picture information for RPL0, RPL1
ref_pic_lists( )
Reference Picture Lists

Fig. 4. Reference picture management signaling.

RPL. A reference picture entry in the RPL is either a short-term reference picture entry, a long-term reference picture entry, or an inter-layer reference picture entry. The default numbers of active entries for RPL 0 and RPL 1 are signaled in the PPS (i.e. pps_num_ref_idx_default_active_minus1[i]) and can be overridden (using the syntax elements sh_num_ref_idx_active_override_flag and sh_num_ref_idx_active_minus1[i]) in the SH.

VII. High-Level Picture Partitioning

VVC inherited the concepts of tiles and WPP from HEVC, with some minor to moderate differences. The basic concept of slices was also kept in VVC but was designed in a rather different form. VVC is the first video coding standard that includes subpictures as a feature, which can provide a functionality that was previously specified in a version 2 extension of HEVC using metadata and encoder constraints for what is known as motion-constrained tile sets (MCTSs), but is designed in a different way to have better coding efficiency and to be friendlier for usage in application systems. More details of these differences are described below.

A. Tiles and Wavefront Parallel Processing (WPP)

As described in Section II-J and as in HEVC, a picture can be split into tile rows and tile columns in VVC, and in-picture prediction across tile boundaries is disallowed. However, the syntax for the signaling of tile partitioning has been simplified, by using a unified syntax design for both the uniform and the non-uniform tile partitioning schemes. The WPP design in VVC has two differences compared to HEVC: i) The signaling of entry point offsets for WPP in the SH is optional in VVC, while it is mandatory in HEVC; and ii) The CTU lag between the ability to start the decoding of consecutive rows for WPP is reduced from the two CTUs needed in HEVC to only one. The latter change is somewhat related to the expectation that CTUs in VVC would typically be larger than in HEVC.

B. Slices

In VVC, the conventional slices consisting of an arbitrary number of CTUs have been removed. The main reasoning behind this architectural change is as follows. The advances in video coding since 2003 (the publication year of AVC v1) have been such that slice based error concealment has become practically impossible, due to the ever-increasing number and efficiency of in-picture and inter-picture prediction mechanisms. An error-concealed picture is the decoding result of a transmitted coded picture for which there has been some data loss (e.g., loss of some slices) for the coded picture or a corruption of some reference picture(s) that are used for reference by the coded picture (e.g., that a reference picture was lost or was itself an error-concealed picture), so that the decoded picture result is not error-free. For example, when one of the multiple slices of a picture is lost, it may be error-concealed using an interpolation of the border sample values of neighboring slices. While more advanced prediction mechanisms provide significantly higher coding efficiency, they also make it harder to estimate the content and quality of an error-concealed picture, which was already a hard problem with the use of simpler prediction mechanisms. Advanced

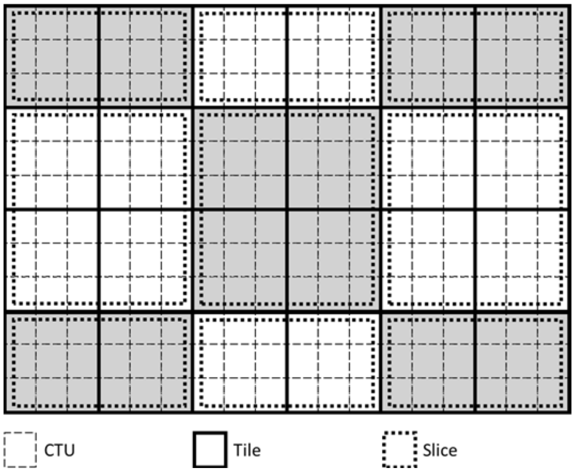

Fig. 5. A picture with 18 by 12 luma CTUs that is partitioned into 24 tiles and 9 rectangular slices.

in-picture prediction mechanisms also cause the coding efficiency loss due to splitting a picture into multiple slices to become more significant. Furthermore, network conditions have become significantly better while at the same time the techniques for dealing with packet losses in the transmission protocol layers have been significantly improved. As a result, very few implementations have recently used slices for the previously common requirement of matching a prescribed maximum transmission unit size. Instead, substantially all applications where low-delay data loss resilience is required (e.g., video telephony and video conferencing) now rely on system/transport-level loss resilience (e.g., retransmission, forward error correction) and/or picture-based loss resilience tools (feedback based loss resilience, insertion of IRAPs, scalability with higher protection level of the base layer, and so on). Considering all the above, it has become rare that a picture that cannot be correctly decoded is passed to the decoder, and when such a rare case occurs, the system can typically afford to wait for an error-free picture to be decoded and available for display without resulting in frequent and long periods of picture freezing that are experienced by end users.

Slices in VVC have two modes: rectangular slices and raster-scan slices. Rectangular slices, as indicated by their name, each cover a rectangular region of the picture. Typically, a rectangular slice consists of a number of complete tiles, e.g., as shown in Fig. 5.

Alternatively, it is also possible for a rectangular slice to be a subset of a tile consisting of one or more consecutive, complete CTU rows within the tile, e.g., as shown in upper-right quadrant of Fig. 6. A raster-scan slice consists of one or more complete tiles in tile raster scan order, and hence the region covered by a raster-scan slice could have a non-rectangular shape (e.g. as shown in Fig. 7), although it could also happen to have the shape of a rectangle. The concept of slices in VVC is therefore strongly linked to or based on the concept of tiles instead of CTUs as in prior standards.

The layout of rectangular slices (including the position and size of each of the slices) is signaled in the PPS, based on the layout of tiles, by signaling the tile position of the first tile in the slice, as well as the width and height of the slice in units of tiles for slices covering multiple tiles, or by signaling the number of slices in one tile and the number of CTU rows in each slice for tiles covering multiple slices. Information on the number of tiles included in a raster-scan slice is signaled in the SH, which might be beneficial for low-delay applications where the number of tiles to be included in a slice might not be known beforehand when the PPS is generated.

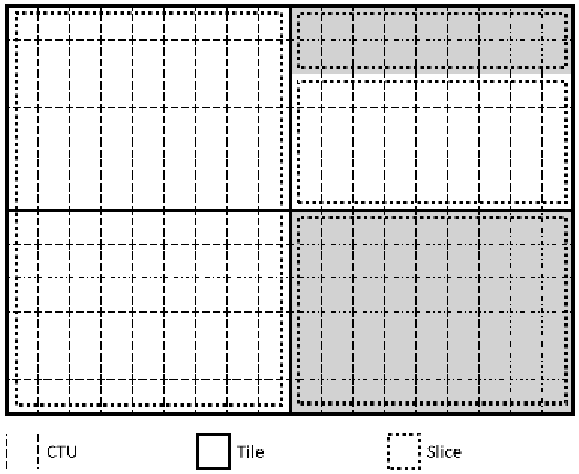

Fig. 6. A picture partitioned into 4 tiles and 4 rectangular slices (note that the top-right tile is split into two rectangular slices).

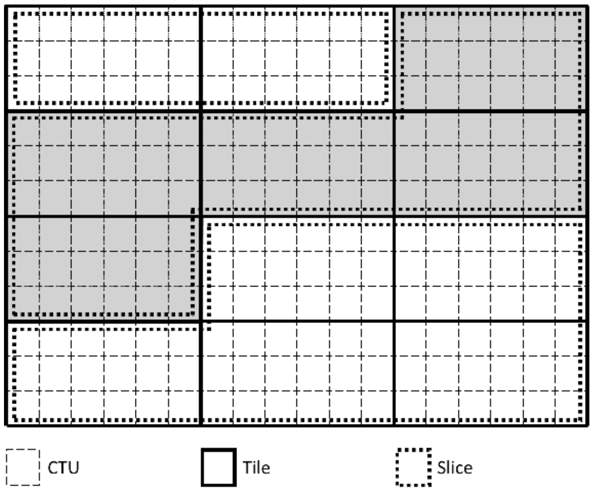

Fig. 7. A picture with 18 by 12 luma CTUs that is partitioned into 12 tiles and 3 raster-scan slices.

C. Subpictures

Concepts of subpictures were proposed previously for AVC in [27] and for HEVC in [28] but were not adopted in these prior standards. During the standardization of VVC, a more comprehensive design of the subpicture feature was developed.

1) Subpicture Concept and Functionality: In VVC, each subpicture consists of one or more complete rectangular slices that collectively cover a rectangular region of the picture. An example is shown in Fig. 8, in which each subpicture consists of exactly one slice (although this is not a requirement).

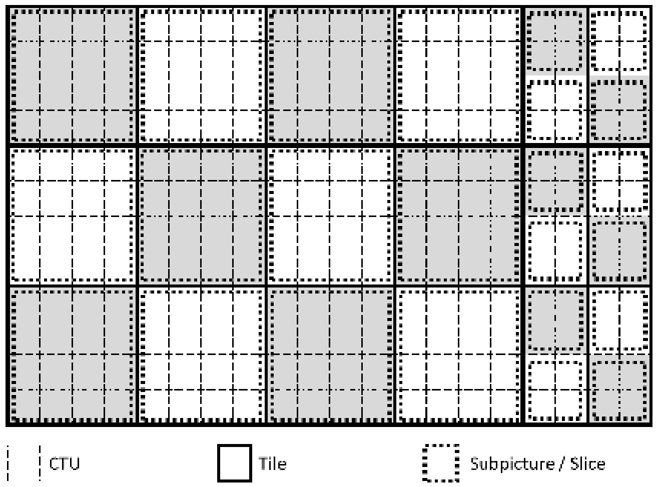

Fig. 8. A picture partitioned into 18 tiles, 24 slices and 24 subpictures.

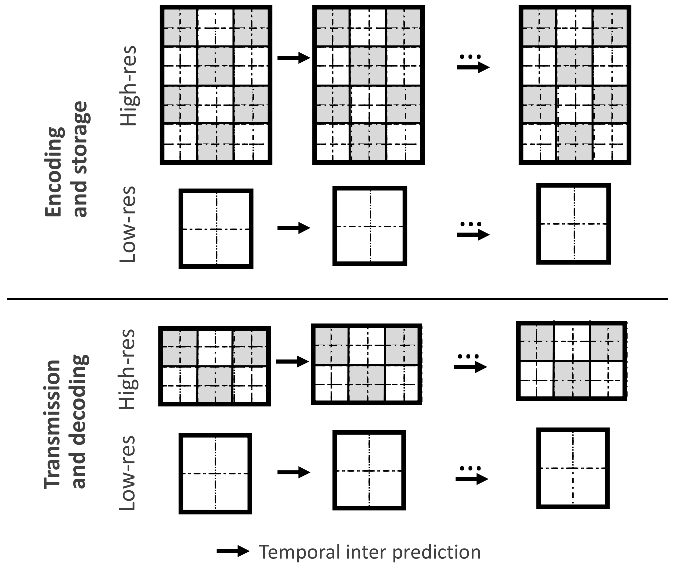

Fig. 9. A typical subpicture-based viewport-dependent 360° video delivery scheme.

A subpicture may be either specified to be extractable (i.e., coded independently of other subpictures of the same picture and earlier pictures in decoding order) or not extractable. Regardless of whether a subpicture is extractable or not, the encoder can control whether in-loop filtering (including deblocking, SAO, and ALF) is applied across the subpicture boundaries individually for each subpicture.

Functionally, extractable subpictures are similar to the MCTSs specified in HEVC. They both enable independent coding and extraction of an area of a sequence of coded pictures, for use cases like viewport-dependent 360° video streaming and region of interest (ROI) applications.

The properties of subpictures and tiles interact through their relationship to rectangular slices. The only direct coupling of the concepts of subpictures and tiles is a requirement that either or both of the following conditions must be true: 1) all CTUs in a subpicture must belong to the same tile; or 2) all CTUs in a tile must belong to the same subpicture. It is possible to have multiple subpictures within a tile (e.g., as shown in Fig. 8) and it is also possible to have multiple tiles in a subpicture, and both of these cases could even occur within the same picture. The left and right vertical boundaries of subpictures are always aligned with the vertical boundaries of some tiles, but since it is possible for a rectangular slice to cover only some of the rows of a tile, it is possible for the top or bottom horizontal boundaries of a subpicture to not coincide with the horizontal boundaries of any tile. Thus, unlike with tiles, the partitioning line segment that forms a top or bottom boundary of a subpicture does not need to continue all the way to the edges of the picture. Unlike with tiles, the division of the pictures into subpictures persists throughout each entire CLVS, in order to simplify the inter-picture prediction relationships for extractable subpictures.

In streaming of 360° video, a.k.a. omnidirectional video, at any particular moment only a subset (i.e., the current viewport) of the entire omnidirectional video sphere would be rendered to the user, while the user can turn or tilt their head at any time to change their viewing orientation and consequently their current viewport. While it is desirable to have at least some lower-quality representation of the area not covered by the current viewport available at the client and ready to be rendered to the user just in case the user suddenly changes their viewing orientation to somewhere else on the sphere, a high-quality representation of the omnidirectional video is only needed for the current viewport that is being rendered to the user at any given moment. Splitting the high-quality representation of the entire omnidirectional video into subpictures at an appropriate granularity enables such an optimization as shown in Fig. 8 with 12 high-resolution subpictures on the left-hand side and the remaining 12 subpictures of the omnidirectional video in lower resolution on the right-hand side.

Another example subpicture-based viewport-dependent 360° video delivery scheme is shown in Fig. 9, wherein a higher-resolution representation of the full video scene consists of subpictures, while a lower-resolution representation of the scene does not use subpictures and can be coded with less frequent RAPs than the higher-resolution representation. The client receives the full video scene in the lower-resolution while for the higher-resolution video, the client only receives and decodes the subpictures that cover their current viewport.

2) Differences Between Subpictures and MCTSs: There are several important design differences between extractable subpictures and MCTSs. First, the subpictures feature in VVC allows motion vectors of a coding block to point outside of the subpicture even when the subpicture is extractable, by applying sample padding at subpicture boundaries in this case, similarly as at picture boundaries. Second, additional changes were introduced for the selection and derivation of motion vectors in the merge mode and in the decoder motion vector refinement process of VVC. This allows higher coding efficiency compared to the motion constraints applied at the encoder for MCTSs (which have been rather complex constraints to specify and test). Third, rewriting of SHs (and PH NAL units, when present) is not needed when extracting of one or more extractable subpictures from a sequence of pictures to create a sub-bitstream that is a conforming bitstream. In sub-bitstream extractions based on HEVC MCTSs, rewriting of SHs is needed. Note that in both HEVC MCTSs extraction and VVC subpictures extraction, rewriting of SPSs and PPSs is needed. However, typically there are only a few parameter sets in a bitstream, while each picture has at least one slice, and the SHs are prefixed to the slice data within the slice NAL units; therefore the rewriting of SHs can be a significant burden for application systems. Fourth, slices of different subpictures within a picture are allowed to have different NAL unit types. This is the feature often referred to as mixed NAL unit types or mixed subpicture types within a picture, discussed in more detail below. Fifth, VVC specifies HRD and level definitions for subpicture sequences, thus the conformance of the sub-bitstream of each extractable subpicture sequence can be ensured by encoders. The fact that the subpictures of VVC are rectangular is a simplification relative to the possible region shapes that could be expressed using MCTSs.

3) Mixed Subpicture Types Within a Picture: In AVC and HEVC, all VCL NAL units in a picture need to have the same NAL unit type. VVC introduces the option to mix subpictures with certain different VCL NAL unit types within a picture, thus providing support for random access not only at the picture level but also at the subpicture level. In VVC VCL NAL units within a subpicture are still required to have the same NAL unit type.

The capability of random accessing from IRAP subpictures is beneficial for 360° video applications. In viewport-dependent 360° video delivery schemes similar to the one shown in Fig. 9, the content of spatially neighboring viewports largely overlaps, i.e. only a fraction of subpictures in a viewport is replaced by new subpictures during a viewport orientation change, while most subpictures remain in the viewport. Subpicture sequences that are newly introduced into the viewport must begin with IRAP slices but significant reduction in overall transmission bit rate can be achieved when the remaining subpictures are allowed to carry out inter-picture prediction at viewport changes.

The indication of whether a picture contains just a single type of NAL units or more than one type is provided in the PPS referred to by the picture (i.e., using a flag called pps_mixed_nalu_types_in_pic_flag). A picture may consist of subpictures containing IRAP slices and subpictures containing trailing slices at the same time. A few other combinations of different NAL unit types within a picture are also allowed, including leading picture slices with RASL and RADL NAL unit types, which allows the merging of subpicture sequences with open-GOP and close-GOP coding structures extracted from different bitstreams into one bitstream.

4) Subpicture Layout and ID Signaling: The layout of subpictures in VVC is signaled in the SPS, is thus constant within a CLVS. Each subpicture is signaled by the position of its top-left CTU and its width and height in number of CTUs, therefore ensuring that a subpicture covers a rectangular region of the picture with CTU granularity. The order in which the subpictures are signaled in the SPS is the index of each subpicture within the picture.

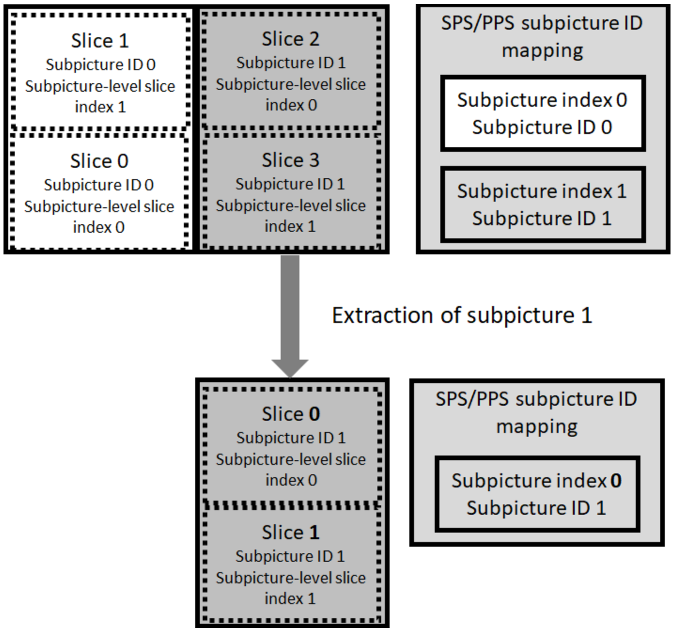
Slice 1
Subpicture ID 0
Subpicture-level slice index 1
Slice 2
Subpicture ID 1
Subpicture-level slice index 0
Slice 0
Subpicture ID 0
Subpicture-level slice index 0
Slice 3
Subpicture ID 1
Subpicture-level slice index 1
SPS/PPS subpicture ID mapping
Subpicture index 0
Subpicture ID 0
Subpicture index 1
Subpicture ID 1
Extraction of subpicture 1
Slice 0
Subpicture ID 1
Subpicture-level slice index 0
Slice 1
Subpicture ID 1
Subpicture-level slice index 1
SPS/PPS subpicture ID mapping
Subpicture index 0
Subpicture ID 1

Fig. 10. Extraction of one subpicture from a bitstream containing two subpictures and four slices.

For enabling extraction and merging of subpicture sequences without rewriting of SHs or PHs, the slice addressing scheme in VVC is based on subpicture IDs and a subpicture-specific slice index to associate slices to subpictures. In the SH, the subpicture ID of the subpicture containing the slice and the subpicture-level slice index are signaled. Note that the value of subpicture ID of a particular subpicture can be different from the value of its subpicture index. A mapping between the two is either signaled in the SPS or PPS (but never both) or implicitly inferred. When present, the subpicture ID mapping needs to be rewritten or added when rewriting the SPSs and PPSs during the subpicture sub-bitstream extraction process. The subpicture ID and the subpicture-level slice index together indicate to the decoder the exact position of the first decoded CTU of a slice within the DPB slot of the decoded picture. After sub-bitstream extraction, the subpicture ID of a subpicture remains unchanged while the subpicture index may change. Even when the raster-scan CTU address of the first CTU in a slice in the subpicture has changed compared to the value in the original bitstream, the unchanged values of subpicture ID and subpicture-level slice index in the respective SH would still correctly determine the position of each CTU in the decoded picture of the extracted sub-bitstream. Fig. 10 illustrates the usage of subpicture ID, subpicture index and subpicture-level slice index to enable subpicture extraction with an example containing two subpictures and four slices.

Similar to subpicture extraction, the signaling for subpictures allows merging several subpictures from different bitstreams into a single bitstream by only rewriting the SPSs and PPSs, provided that the different bitstreams are coordinately generated (e.g., using distinct subpicture IDs but otherwise mostly aligned SPS, PPS and PH parameters such as CTU size, chroma format, coding tools, etc.).

While subpictures and slices are signaled independently in the SPS and PPS, respectively, there are inherent reciprocal constraints between the subpicture and slice layouts in order to form a conforming bitstream. First, the presence of subpictures requires using rectangular slices and forbids raster-scan slices. Second, the slices of a given subpicture shall be consecutive NAL units in decoding order, which means that the subpicture layout constrains the order of coded slice NAL units within the bitstream.

D. Order of VCL NAL Units Within a Picture

Slices (VCL NAL units) in a picture are ordered based on their slice addresses and the subpicture indices they belong to. Slices with smaller subpicture indices are placed first. If there are multiple slices within the same subpicture, the slices within a subpicture are ordered according to the slice address. This results in slices of a picture being ordered such that each CTU shall have its entire left and top boundaries consisting of a picture boundary or boundaries of previously decoded CTU(s). This constraint avoids decoders having to store and reorder slices before processing the slices of a picture as they are received.

VIII. In-Loop Filtering Across Subpicture, Slice, Tile, and Virtual Boundaries

VVC specifies flags at different levels for controlling the in-loop filtering processes, including deblocking, SAO, and ALF, across subpicture, slice, tile, and virtual boundaries.

Each subpicture is associated with a SPS-level flag for controlling of in-loop filtering across subpicture boundaries. All slices within a picture are associated with a PPS-level flag for controlling of in-loop filtering across slice boundaries. All tiles within a picture are associated with a PPS level flag for controlling of in-loop filtering across tile boundaries. These flags basically work based on a everyone-can-veto rule: for any particular boundary that is a boundary (or a part thereof) of a subpicture, slice, and/or tile boundary, when any one of the above associated flags indicates that filtering across the boundary is turned off, regardless of the values of the other flags, filtering across the boundary is turned off.

The virtual boundary signaling in VVC allows turning off in-loop filtering at signaled positions within the coded pictures that do not have to be aligned with the CTU boundaries. Also, virtual boundaries do not introduce further in-picture prediction breaks such as introduced by slices and tiles when used for this purpose. Virtual boundaries are useful in at least two applications. First, for coding of 360° video, when the 360° video uses a particular projection format that introduces discontinuities, e.g., at the unaligned face boundaries of a cubemap projection, virtual boundaries allow disabling of in-loop filtering across these boundaries without the need for content scaling to align projection discontinuities and CTU boundaries. Second, when the GDR feature is used, the boundary between the refreshed region (i.e., the correctly decoded region) and the unrefreshed region in a recovering picture that is between a GDR picture and its recovery point can be signaled as a virtual boundary, to disable in-loop filtering across the boundary, thus avoiding decoding mismatch for some samples at or near the boundary. This can be useful when the application determines to display the correctly decoded regions during the GDR process.

IX. Picture Resolution Change Within a Sequence

In AVC and HEVC, the spatial resolution of pictures cannot change unless a new sequence using a new SPS starts, with an IRAP picture. VVC enables picture resolution change within a sequence at a position without encoding an IRAP picture, which is always intra-coded. This feature is sometimes referred to as reference picture resampling (RPR), as the feature needs resampling of a reference picture used for inter prediction when that reference picture has a different resolution than the current picture being decoded.

In order to allow reusing the motion compensation module of existing implementations, the scaling ratio is restricted to be larger than or equal to 1/2 (2 times downsampling from the reference picture to the current picture), and less than or equal to 8 (8 times upsampling). The horizontal and vertical scaling ratios are derived based on picture width and height, and the left, right, top and bottom scaling offsets specified for the reference picture and the current picture.

RPR allows resolution change without the need of coding an IRAP picture, which causes a momentary bit rate spike in streaming or video conferencing scenarios, e.g., to cope with network condition changes. RPR also can be used in application scenarios wherein zooming of the entire video region or some region of interest is needed. The scaling window offsets are allowed to be negative to support a wider range of zooming-based applications [29]. Negative scaling window offsets also enable extraction of subpicture sequences out of a multi-layer bitstream while keeping the same scaling window for the extracted sub-bitstream as in the original bitstream [30].

Differently from the spatial scalability in the scalable extension of HEVC, where picture resampling and motion compensation are applied in two different stages, RPR in VVC is carried out as part of the same process on a block level where the derivation of sample positions and motion vector scaling are performed during motion compensation.

In an effort to limit the implementation complexity, a change of picture resolution within a CLVS is disallowed when the pictures in the CLVS have multiple subpictures per picture. Furthermore, decoder motion vector refinement, bi-directional optical flow, and prediction refinement with optical flow are not applied when RPR is used between the current picture and the reference pictures. The collocated picture for the derivation of temporal motion vector candidates is also restricted to have the same picture size, scaling window offsets, and CTU size as the current picture.

For support of RPR, some other aspects of the VVC design have been made different from HEVC. First, the picture resolution and the corresponding conformance and scaling windows are signaled in the PPS instead of in the SPS, while in the SPS the maximum picture resolution and corresponding conformance window are signaled. In applications, the maximum picture resolution with the corresponding conformance window offsets in the SPS can be used as the intended or desired picture output size after cropping. Second, for a single-layer bitstream, each picture store (a slot in the DPB for storage of one decoded picture) occupies the buffer size as required for storing a decoded picture having the maximum picture resolution. This prevents decoders from being required to support rapid memory reallocation while operating the decoding process.

X. Multi-Layer and Scalability Support

Having the ability to use inter-picture prediction from reference pictures of different sizes than the current picture by means of RPR in the VVC core design allows VVC to easily support bitstreams containing multiple layers of different resolutions, e.g., two layers with standard definition and high definition resolutions, respectively. In a VVC decoder, such functionality can be integrated without the need of any additional signal-processing-level coding tools, as the upsampling functionality needed for spatial scalability support can be provided by reusing the RPR upsampling filter. Nevertheless, additional HLS designs to enable the scalability support of a bitstream are needed.

Scalability is supported in VVC but is included only in the multi-layer profiles. Different from the scalability supports in any earlier video coding standards, including extensions of AVC and HEVC, the design of VVC scalability has been made friendly to single-layer decoder implementations as much as possible. The decoding capability for multi-layer bitstreams is specified in a manner as if there was only a single layer in the bitstream. For example, the decoding capability, such as the DPB size, is specified in a manner that is independent of the number of layers in the bitstream to be decoded. Basically, a decoder designed for single-layer bitstreams does not need significant changes to be able to decode multi-layer bitstreams.

Compared to the designs of multi-layer extensions of AVC and HEVC, the HLS aspects have been significantly simplified at the sacrifice of some flexibility. For examples, 1) an IRAP AU is required to contain a picture for each of the layers present in the CVS, which avoids the need of specifying a layer-wise startup decoding process [31], and 2) a much simpler design for POC signaling, as summarized in Section VI-A, instead of the complicated POC resetting mechanism [32], is included in VVC, to make sure that the derived POC values are the same for all pictures in an AU.

Like in HEVC, the information about layers and layer dependency is included in the VPS. The information of OLSs is provided for signaling of which layers are included in an OLS, which layers are output, and other information such as PTL (see Section III-A) and HRD parameters (see Section XI) associated with each OLS. Similar to HEVC, there are three modes of operations to output either all layers, only the highest layer, or particular indicated layers in a custom output mode.

There are some differences between the OLS design in VVC and in HEVC. First, in HEVC the layer sets are signaled, then OLSs are signaled based on the layer sets, and for each OLS the output layers are signaled. The design in HEVC allowed a layer to belong to an OLS that was neither an output layer nor a layer required for decoding an output layer. In VVC, the design requires any layer in an OLS to be either an output layer or a layer required for decoding an output layer. Therefore, in VVC OLSs are signaled by indicating the output layers of an OLS and then other layers belonging to an OLS are simply derived by the layer dependencies indicated in the VPS. Furthermore, VVC requires each layer to be included in at least one OLS.

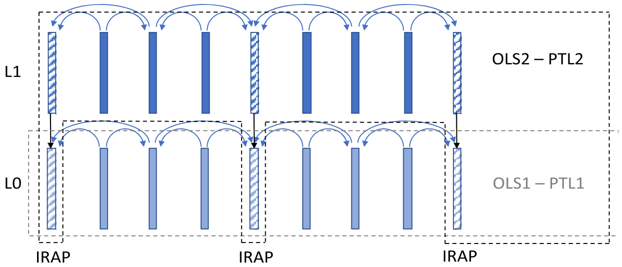

Fig. 11. An example of a bitstream with two OLSs where OLS2 has vps_max_tid_il_ref_pics_plus1[1][0] equal to 0.

Another difference in the VVC OLS design is that, contrary to HEVC, for which an OLS consists of all NAL units that belong to the set of identified layers mapped to the OLS, VVC may exclude some NAL units that belong to non-output layers mapped to an OLS. More specifically, an OLS for VVC consists of the set of layers that are mapped to the OLS with non-output layers including only IRAP or GDR pictures with ph_recovery_poc_cnt equal to 0 or pictures from the sublayers that are used for inter-layer prediction. This allows indicating an optimal level value for a multi-layer bitstream considering only all the "necessary" pictures of all sublayers within the layers that form the OLS, where "necessary" herein means needed for output or decoding. Fig. 11 shows an example of a two-layer bitstream with vps_max_tid_il_ref_pics_plus1[1][0] equal to 0, i.e. a sub-bitstream for which only IRAP pictures from layer L0 are kept when OLS2 is extracted.

Taking into account some scenarios for which it is beneficial to allow different RAP periodicity at different layers, similarly as in AVC and HEVC, AUs are allowed to have layers with non-aligned RAPs. For faster identification of RAPs in a multi-layer bitstream, i.e. AUs with a RAP at all layers, the access unit delimiter (AUD) was extended compared to HEVC with a flag indicating whether the AU is an IRAP AU or GDR AU. Furthermore, the AUD is mandated to be present at such IRAP or GDR AUs when the VPS indicates multiple layers. However, for single layer bitstreams as indicated by the VPS or bitstreams not referring to a VPS, the AUD is completely optional as in HEVC because in this case RAPs can be easily detected from the NAL unit type of the first slice in the AU and the respective parameter sets.

To enable sharing of SPSs, PPSs, and APSs by multiple layers and at the same time to make sure that bitstream extraction process does not throw away parameter sets needed by the decoding process, a VCL NAL unit of a first layer can refer to an SPS, PPS, or APS with the same or a lower layer ID value, as long as all OLSs that include the first layer also include the layer identified by the lower layer ID value.

Due to that the combination of picture resolution change within a sequence and multiple subpictures per picture is

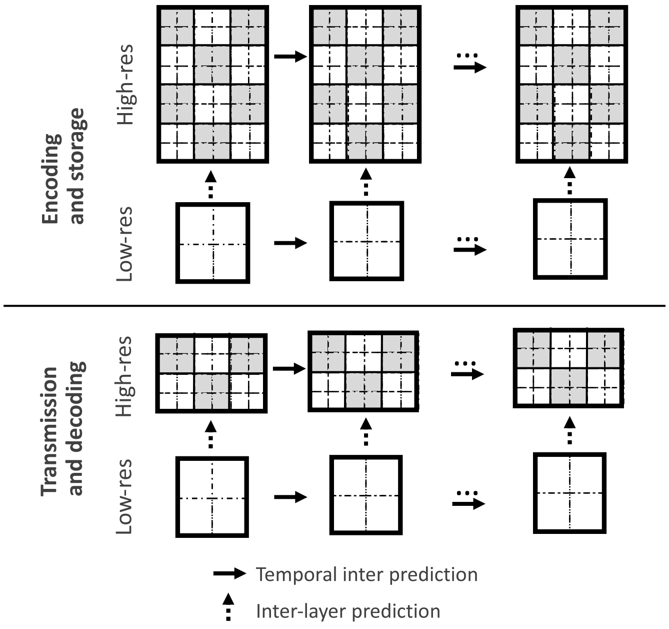

Fig. 12. A viewport-dependent 360° video delivery scheme based on a combination of subpictures and spatial scalability with inter-layer prediction.

not supported in VVC, the combination of spatial scalability and multiple subpictures per picture for all layers is also not supported. VVC does support the combination of SNR scalability and multiple subpictures per picture for all layers, as well as a special combination of spatial scalability and multiple subpictures per picture wherein the reference layer with lower resolution does not use multiple subpictures per picture, which enables an improved viewport-dependent 360° video delivery scheme as shown in Fig. 12 (vs the scheme shown in Fig. 9 without inter-layer prediction).

XI. Hypothetical Reference Decoder (HRD)

As for previous video coding standards, VVC includes the concept of HRD, which is based on the leaky bucket model [33]. The CPB operations basically rely on three parameters: namely transmission bit rate R, buffer size B, and initial decoder buffer fullness or initial delay. The signaling of HRD information in VVC differs from HEVC and previous video coding standards to provide additional support for, for instance, subpictures and splicing operations as explained in the following.

Following the design in HEVC, the HRD parameters are provided by two types of parameters. The sequence-level HRD parameters of VVC may be signaled within the VPS (when they apply to a multi-layer OLS) or the SPS (typically for a single-layer bitstream) and include information such as transmission bit rate and buffer size or whether the HRD operation is constant bit rate (CBR) or variable bit rate (VBR). In addition, the bitstream may contain buffering period (BP), picture timing (PT) and decoding unit information (DUI) SEI messages, where the BP SEI message primarily contains information about initial CPB removal delays, the PT SEI message primarily contains information about CPB removal delay and DPB output delay for AU based HRD operation and the DUI SEI message primarily contains information about CPB removal delay and DPB output delay for DU based HRD operation (although such information may also be contained within the PT SEI message as an alternative). For signaling of the PT information, VVC introduced a flag (general_same_pic_timing_in_all_ols_flag) to enable signaling and use of only one, non-scalable-nested PT SEI message for an AU that applies to all OLSs.

Similar to the HRD operation in HEVC [24], VVC provides support for alternative sets of initial buffering parameters at RAPs. However, compared to HEVC, the syntax of VVC regarding alternative sets provides a more generic solution that not only supports random access at CRA AUs but also random access at dependent RAP (DRAP) AUs as explained in XI-A. In addition, further signaling has been included into VVC into the BP SEI message and PT SEI message that eases bitstream splicing operation as discussed in XI-B. Another important change in VVC is the use of scalable nesting SEI messages compared to HEVC. First, scalable nesting SEIs are not used for indicating HRD operation of temporal sublayers as was the case in HEVC, but the BP, PT and DUI SEI messages have been extended as discussed in XI-C. Second, the extraction process in VVC normatively specifies that BP, PT and DUI SEI messages are substituted with the corresponding SEI messages initially included in a nesting SEI message when layers are dropped due to the extraction process. This is described more in detail in XI-D. Finally, VVC provides further information related to HRD for subpictures with a new SEI message called subpicture level information SEI message as discussed in XI-E.

A. Alternative Timing for Random Access

The alternative timing information in VVC provides HRD information related to random access at CRA or DRAP AUs when the AUs containing the associated RASL pictures or all AUs in between an IRAP AU and a DRAP AU, respectively, are removed from the bitstream. The alternative timing information is conveyed in BP SEI messages and PT SEI messages. The BP SEI message of the IRAP AU contains the initial buffering information for normal operation, but in addition, it indicates whether the alternative timing information is used (bp_use_alt_cpb_params_flag). If so, additional timing information is carried in the PT SEI message of the AU directly following the IRAP AU in decoding order. The information consists of a delta value to be considered for the initial buffering delay before removing the IRAP AU from the CPB, as well as two offsets to be used for AUs following the IRAP AU in order to compute the CPB and DPB removal delays.

B. Bitstream Splicing

Additional HRD information is provided in VVC to support bitstream splicing and editing operations. As for HEVC, the CPB removal time of an AU with a BP SEI message is indicated both in the PT SEI message as a delta to the CPB removal time of the previous BP and also in the BP SEI message as a delta to the previous non-discardable AU. A splicer needs to run an HRD conformance check and keep track of timing information to avoid buffer overflows or underflows for the spliced bitstream. In order to ease splicing operations HEVC and VVC include a syntax element (bp_concatenation_flag) that allows running a simpler splicing operation. When the flag is equal to 1, the removal time of the first AU after the splicing point might be greater than the value indicated in the PT SEI message or BP SEI message. The indication allows a splicer to run fewer checks without risking buffer underflows with the signaled removal time in the PT SEI message or BP SEI message. When bp_concatenation_flag is equal to 1 and the initial CPB removal delay requires removing the first AU of the spliced bitstream later than signaled in the PT SEI message or BP SEI message, the later CPB removal time takes precedence. When bp_concatenation_flag is equal to 0, the value for the CPB removal time indicated in the PT SEI message and BP SEI message is always used and therefore needs to be accurate, as an access unit being available afterwards would lead to a buffer underflow. VVC allows more information to be provided to further ease splicing operations. The PT SEI message may include a flag (pt_delay_for_concatenation_ensured_flag) that identifies AUs as potential splicing points. This flag indicates that if a second bitstream is spliced after an AU of a first bitstream that is marked as a potential splicing point, and the initial buffering delay information of the first AU of the second bitstream is smaller than a value (bp_max_initial_removal _delay_for_concatenation) indicated in the BP SEI message in the first bitstream, it is guaranteed that bp_concatenation_flag equal to 0 can be used without incurring a buffer underflow.

C. HRD for Temporal Sublayers

In HEVC, HRD signaling for temporal sublayers was achieved by using the scalable nesting (SN) SEI message for nesting of appropriate BP, PT and DUI SEI messages that apply to a particular temporal sub-bitstream. HRD parameters in temporal scalable VVC bitstreams are signaled in the PT SEI message either with different CPB removal delays for each value of the highest decoded temporal sublayer or with a CPB removal delay delta. Since the DPB output delays are the same (except for a possible offset) for all temporal representations of a temporal scalable bitstream, the PT SEI message contains only one DPB output delay for all sublayers. DPB output offsets for temporal sub-bitstreams may be signaled in the BP SEI message when the sequence of DPB output times of the pictures should be shifted for a temporal sub-bitstream. With this new HRD signaling for temporal sublayers, HRD conformance for temporal scalable bitstreams does not require the usage of scalable nesting SEI messages, and consequently, the scalable nesting SEI message in VVC has been designed without the capability for nesting of SEI messages for temporal subsets of bitstreams.

D. General Sub-Bitstream Extraction Processes

Similar to HEVC, the VVC specification includes a sub-bitstream extraction process that allows to extract the sub-bitstream corresponding to a particular operation point (i.e., an OLS and the included temporal sublayers). While in HEVC, the extraction process is part of the decoding process, i.e. the decoder would need to discard NAL units not relevant for the operation point when present in the bitstream, the VVC design assumes that the bitstream fed to the decoder does not contain NAL units not belonging to the indicated operation point, i.e., when needed, NAL units not relevant for the operation point are discarded by the extractor that is not part of the decoder. A notable difference compared to HEVC is that in VVC the handling of scalable nested HRD SEI messages is normatively specified, e.g., the extracted bitstream carries correct HRD timing parameters for the target operation point. The process includes removing the original BP SEI messages, and, if any, DUI SEI messages and inserting the appropriate SEI messages originally included in a scalable nesting SEI message when the target operation point does not include all layers in the bitstream. PT SEI messages have, however, a special handling; the above operations are only needed when it is not indicated that each PT SEI message applies to all OLSs.

E. HRD for Subpictures

VVC allows HRD conformance testing of individual independently coded subpictures, i.e., the bitstream portion related to individual subpictures can be extracted to form a valid bitstream and the conformance of that bitstream to an HRD model can be tested. Conformance testing requires definition of a complete HRD model for such a subpicture sub-bitstream and VVC allows to carry the necessary information, in addition to other HRD parameters, by means of a new SEI message referred to as the subpicture level information (SLI) SEI message.

The SLI SEI message provides level information for subpicture sequences, which is needed for deriving the CPB size and the bit rate values of an HRD model that describes the processing of a subpicture bitstream by a decoder. While an original bitstream consisting of multiple subpictures may adhere to the limits defined by a particular level as indicated in the parameter sets of the bitstream (e.g., Level 5.1 for 4K at 60Hz), a subpicture sub-bitstream of that particular bitstream may correspond to a lower level (e.g., Level 3 for 720p at 60Hz). Furthermore, VVC allows expressing the level of a particular subpicture sub-bitstream by means of fractions of reference levels, which allows a more fine granular level signaling than in HEVC, and this information also provides guidance to systems that merge multiple subpictures into a single joint bitstream on how much each subpicture sub-bitstream will contribute towards the level limits of a merged bitstream. Additional properties such as the bitstream exhibiting a constant bit rate or, in case of multi-layer bitstreams, the level contribution of layers without subpicture partitioning applied (e.g., the lower-resolution pictures as shown in Fig. 12) can also be signaled in the SLI SEI message, thus enabling the derivation of a compete HRD model for each individual subpicture sequence also for such scenarios. A further part of the effort to allow conforming subpicture sub-bitstreams is to apply the numerous conformance related constraints that have a picture scope also with a subpicture scope, such as constraints on minimum compression ratio or the bin-to-bit ratio for the VCL NAL units that belong to an individual subpicture.

In HEVC, the sub-bitstream extraction and conformance testing of independently coded regions (i.e., MCTS) requires the parameter sets for such sub-bitstreams to be carried in a nested form by means of an MCTS extraction information set SEI message. VVC adds a new subpicture sub-bitstream extraction process that allows the generation of a conforming bitstream from an independently coded subpicture by removing the unnecessary NAL units associated to other subpictures and actively rewriting the relevant parts of the parameter sets to correctly reflect the properties of the subpicture sub-bitstream. For instance, this process includes rewriting the level indicators and HRD parameters in VPSs and SPSs, as well as the picture sizes, partitioning information, conformance window offsets, scaling window offsets, and virtual boundary positions in the appropriate section of the respective parameter sets. The subpicture sub-bitstream extraction process rewrites some of the information in the parameter sets of the bitstream based on the information provided in the SLI SEI message.

XII. VUI and SEI

The VUI syntax structure and most SEI messages used with VVC are not specified in the VVC specification, but rather in the VSEI specification. The SEI messages necessary for HRD conformance testing are specified in the VVC specification. VVC v1 defines five SEI messages relevant for HRD conformance testing and VSEI v1 specifies 20 additional SEI messages. The SEI messages carried in the VSEI specification do not directly impact conforming decoder behavior and have been defined so that they can be used in a codec-agnostic manner, allowing the VSEI standard to be used in the future with other video coding standards in addition to VVC. Rather than referring specifically to VVC syntax element names, the VSEI specification refers to variables which have values that are set within the VVC specification. Another reason for having these aspects written in a separate standard is that the core coding technology document and the VSEI document can be drafted, maintained and extended by different groups of editors who do not need to be familiar with both aspects and do not need to manage editing work on an excessively large single document. (Currently, the ITU publication for AVC is about 800 pages and HEVC is about 700 pages, whereas VVC v1 is about 500 pages, and VSEI v1 is about 75 pages that would otherwise need be included within VVC.)

Compared to HEVC, the VUI syntax structure of VVC focuses only on information relevant for proper rendering of the pictures and does not contain any timing information or bitstream restriction indications. In VVC, the VUI is signaled within the SPS, which includes a length field before the VUI syntax structure to signal the length of the VUI payload in bytes. This makes it possible for a decoder to easily jump over the information, and more importantly, allows convenient future VUI syntax extensions by directly adding new syntax elements to the end of the VUI syntax structure, in a similar manner as SEI message syntax extension.

TABLE II
List of SEI Messages in VVC v1

Name of SEI message	Purpose of SEI message
SEI messages specified in the VVC specification	
Buffering period	Initial CPB removal delays for HRD
Picture timing	CPB removal delays and DPB output delays for HRD
Decoding unit information	CPB removal delays and DPB output delays for DU based HRD
Scalable nesting	Mechanism to associate SEI messages with specific output layer sets, layers or sets of subpictures
Subpicture level information	Information about levels for subpicture sequences
SEI messages specified in the VSEI specification	
Filler payload	Filler data for adjusting the bit rate
User data registered by Rec. ITU-T T.35	Conveying user data; can be used as container for data by other organizations
User data unregistered	
Film grain characteristics	Model for film grain synthesis
Frame packing arrangement	Information about how stereoscopic video is coded in the bitstream, e.g., by packing the two pictures for each time instance of the two views into one picture
Parameter sets inclusion indication	Indication of whether the sequence contains all the required NAL units for decoding
Decoded picture hash	Hash of the decoded pictures for error detection
Mastering display color volume	Description of the color volume of a display used to author the content
Content light level information	Upper bounds for the nominal target brightness light level of the content
Dependent RAP indication	Indicates a picture using only the preceding IRAP picture for inter prediction referencing
Alternative transfer characteristics	Preferred alternative value for the transfer characteristics of the content
Ambient viewing environment	Characteristics of the nominal ambient viewing environment for the display of the content, can be used to assist the receiver in processing content depending on the local viewing environment
Content color volume	Color volume characteristics of the associated picture
Equirectangular projection	Indication of the projection format applied, including information needed for remapping of the content onto a sphere for rendering in omnidirectional video applications
Generalized cubemap projection	
Sphere rotation	Information on rotation angles for conversion between the global and local coordinate axes, for use in omnidirectional video applications

TABLE II
(Continued.) LIST OF SEI MESSAGES IN VVC v1

Region-wise packing	Information needed for remapping of the cropped decoded pictures, involving region-wise operations like repositioning, resizing and rotation, onto projected pictures, for use in omnidirectional video applications
Omnidirectional viewport	Coordinates of one or more regions corresponding to viewports recommended for display, for use in omnidirectional video applications
Frame-field information	Indicates how the associated picture should be displayed, its source scan type, and whether it is a duplicate of a previous picture
Sample aspect ratio information	Information about aspect ratio of the color component samples of the associated picture

The VUI syntax structure contains the following information:

- The content being interlaced or progressive;
- Whether the content contains frame-packed stereoscopic video or projected omnidirectional video;
- Sample aspect ratio;
- Whether the content is appropriate for overscan display;
- Color description, including color primaries, matrix and transfer characteristics, which is particularly important to be able to signal ultra high definition (UHD) vs high definition (HD) color space as well as high dynamic range (HDR);
- Chroma location compared to luma (for which the signaling was clarified for progressive content compared to HEVC).

When the SPS does not contain any VUI, the information is considered unspecified and has to be conveyed via external means or specified by the application if the content of the bitstream is intended for rendering on a display.

Table II lists all the SEI messages specified for VVC v1, as well as the specification containing their syntax and semantics. Of the 20 SEI messages specified in the VSEI specification, many were inherited from HEVC (for example, the filler payload and both user data SEI messages). Some SEI messages are essential for correct processing or rendering of the coded video content. This is for example the case for the mastering display color volume, the content light level information or the alternative transfer characteristics SEI messages which are particularly relevant for HDR content. Other examples include the equirectangular projection, sphere rotation, region-wise packing or omnidirectional viewport SEI messages, which are relevant for signaling and processing of 360° video content.

New SEI messages that were specified for VVC v1 include the frame-field information SEI message, the sample aspect ratio information SEI message, and the subpicture level information SEI message.

The frame-field information SEI message contains information to indicate how the associated picture should be displayed (such as field parity or frame repetition period), the source scan type of the associated picture and whether the associated picture is a duplicate of a previous picture. This information used to be signaled in the picture timing SEI message in previous video coding standards, together with the timing information of the associated picture. However, it was observed that the frame-field information and timing information are two different kinds of information that are not necessarily signaled together. A typical example consists in signaling the timing information at the systems level, but signaling the frame-field information within the bitstream. It was therefore decided to remove the frame-field information from the picture timing SEI message and signal it within a dedicated SEI message instead. This change also made it possible to modify the syntax of the frame-field information to convey additional and clearer instructions to the display, such as the pairing of fields together, or more values for frame repetition.

The sample aspect ratio SEI message enables signaling different sample aspect ratios for different pictures within the same sequence, whereas the corresponding information contained in the VUI applies to the whole sequence. It may be relevant when using the reference picture resampling feature with scaling factors that cause different pictures of the same sequence to have different sample aspect ratios.

The subpicture level information SEI message provides information of levels for the subpicture sequences, as described in Section XI-E.

XIII. CONCLUSION

The high-level syntax (HLS) is an integral part of a video coding standard. It provides an interface to the application environment, provides basic functionalities such as random accessing and stream adaptation capabilities, as well as information needed for session negotiation, content selection, and for enabling various optimizations of application systems. The HLS of VVC inherited many aspects from the preceding video coding standards AVC and HEVC, but also included substantial new features as well as changes to the inherited features, partly to address the needs of the newer and more advanced use cases. To fully enable the use of VVC in various application systems, the standardization of more elements of the VVC systems and transport interface, i.e., the extensions to the file format, RTP payload format, and MPEG-2 transport stream standards for encapsulation of VVC video in these environments, is needed. These standardization activities are all currently ongoing in MPEG and IETF. Further work to develop SEI messages for conveying additional information associated with video bitstreams for VVC and other video contexts is also under way in JVET.

ACKNOWLEDGMENT

The authors would like to thank all the JVET experts who have contributed to the VVC HLS design. The archives of JVET, JCT-VC, JVET and VCEG documents can be found online at https://www.jvet-experts.org/, http://phenix.int-evry.fr/jct/index.php, https://www.itu.int/wftp3/av-arch/jvt-site/, and https://www.itu.int/wftp3/av-arch/video-site/, respectively.

REFERENCES

[1] *Versatile Video Coding*, Standard ITU-T H.266, ISO/IEC 23090-3, 2020.
[2] *Versatile Supplemental Enhancement Information Messages for Coded Video Bitstreams*, Standard ITU-T H.274, ISO/IEC 23002-7, 2020.

[3] J. Postel, "Internet protocol," Standard 791, IETF RFC, Sep. 1981.

[4] S. Deering and R. Hinden, "Internet protocol, version 6 (IPv6) specification," Standard 8200, IETF RFC, Jul. 2017.

[5] J. Postel, "User datagram protocol," Standard 768, IETF RFC, Aug. 1980.

[6] H. Schulzrinne, S. Casner, R. Frederick, and V. Jacobson, "RTP: A transport protocol for real-time applications," Standard 3550, IETF RFC, Jul. 2003.

[7] Y.-K. Wang, R. Even, T. Kristensen, and R. Jesup, "RTP payload format for H.264 video," Standard 6184, IETF RFC, May 2011.

[8] S. Wenger, Y.-K. Wang, T. Schierl, and A. Eleftheriadis, "RTP payload format for scalable video coding," Tech. Rep. 6190, IETF RFC, May 2011.

[9] Y.-K. Wang, Y. Sanchez, T. Schierl, S. Wenger, and M. M. Hannuksela, "RTP payload format for high efficiency video coding (HEVC)," Standard 7798, IETF RFC, Mar. 2016.

[10] M. Handley, V. Jacobson, and C. Perkins, "SDP: Session description protocol," Standard 4566, IETF RFC, Jul. 2006.

[11] J. Postel, "Transmission control protocol," Standard 793, IETF RFC, Sep. 1981.

[12] R. Fielding and J. Reschke, "Hypertext transfer protocol (HTTP/1.1): Message syntax and routing," Standard 7230, IETF RFC, Jun. 2014.

[13] *Information Technology—Coding of Audio-Visual Objects—Part 12: ISO Base Media File Format*, Standard ISO/IEC 14496-12, Dec. 2020.

[14] *Information Technology—Dynamic Adaptive Streaming Over HTTP (DASH)—Part 1: Media Presentation Description and Segment Formats*, Standard ISO/IEC 23009-1, Dec. 2019.

[15] *Information Technology—Coding of Audio-Visual Objects—Part 15: Carriage of Network Abstraction Layer (NAL) Unit Structured Video in the ISO Base Media File Format*, Standard ISO/IEC 14496-15, Sep. 2019.

[16] *Information Technology—Generic Coding of Moving Pictures and Associated Audio Information—Part 1: Systems*, Standard ITU-T H.222.0, ISO/IEC 13818-1, Aug. 2018.

[17] *Advanced Video Coding, (AVC)*, Standard ITU-T H.264, ISO/IEC 14496-10, Jun. 2019.

[18] *High Efficiency Video Coding, (HEVC)*, Standard ITU-T H.265, ISO/IEC 23008-2, Nov. 2019.

[19] R. Sjöberg *et al.*, "Overview of HEVC high-level syntax and reference picture management," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1858–1870, Dec. 2012.

[20] T. Schierl, M. M. Hannuksela, Y.-K. Wang, and S. Wenger, "System layer integration of high efficiency video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1871–1884, Dec. 2012.

[21] R. Sjöberg and J. Boyce, "HEVC high-level syntax," in *High Efficiency Video Coding*, V. Sze, M. Budagavi, and G. J. Sullivan, Eds. Cham, Switzerland: Springer, 2014, ch. 2, pp. 13–48.

[22] Y.-K. Wang, M. M. Hannuksela, S. Pateux, A. Eleftheriadis, and S. Wenger, "System and transport interface of SVC," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 9, pp. 1149–1163, Sep. 2007.

[23] J. Ribas-Corbera, P. A. Chou, and S. L. Regunathan, "A generalized hypothetical reference decoder for H.264/AVC," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 7, pp. 674–687, Jul. 2003.

[24] S. Deshpande, M. M. Hannuksela, K. Kazui, and T. Schierl, "An improved hypothetical reference decoder for HEVC," *Proc. SPIE*, vol. 8666, Feb. 2013, Art. no. 866608, doi: 10.1117/12.2009907.

[25] *Procedure for the Allocation of ITU-T Defined Codes for Non Standard Facilities*, Standard ITU-T T.35, Feb. 2000.

[26] Y.-K. Wang *et al.*, *On Reference Picture Management for VVC*, document JVET-M0128, 13th Meeting of ITU-T/ISO/IEC Joint Video Experts Team (JVET), Jan. 2019.

[27] M. M. Hannuksela and Y.-K. Wang, *New Image Segmentation Method*, document VCEG-O46, 15th VCEG Meeting, Pattaya, Thailand, Dec. 2001.

[28] M. Coban, Y.-K. Wang, and M. Karczewicz, *AHG4: Support of Independent Sub-Pictures*, document JCTVC-I0356, 9th JCT-VC Meeting, Geneva, Switzerland, Apr./May 2012.

[29] J. Samuelsson, S. Deshpande, and A. Segall, *AHG9: On Scaling Window Offsets*, document JVET-R0114, 18th JVET Meeting, Teleconf., Apr. 2020.

[30] Y.-K. Wang, Z. Deng, K. Zhang, and L. Zhang, *AHG9/AHG8: On Reference Picture Resampling*, document JVET-S0048, 19th JVET Meeting, Teleconf., Jun./Jul. 2020.

[31] M. M. Hannuksela, *MV-HEVC/SHVC HLS: Layer-Wise Startup of the Decoding Process*, document JCTVC-M0206, 13th JCT-VC Meeting, Incheon, South Korea, Apr. 2013.

[32] Y. Chen, Y.-K. Wang, and A. K. Ramasubramonian, *MV-HEVC/SHVC HLS: Cross-Layer POC Alignment*, document JCTVC-N0244, 14th JCT-VC Meeting, Vienna, Austria, Jul./Aug. 2013.

[33] A. R. Reibman and B. G. Haskell, "Constraints on variable bit-rate video for ATM networks," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 2, no. 4, pp. 361–372, Dec. 1992.

Ye-Kui Wang received the B.S. degree in industrial automation from the Beijing Institute of Technology in 1995 and the Ph.D. degree in information and telecommunication engineering from the Graduate School in Beijing, University of Science and Technology of China, in 2001.

He is currently a Principal Scientist with ByteDance, San Diego, CA, USA. His earlier working experiences and titles include the Chief Scientist of Media Coding and Systems at Huawei Technologies, San Diego, the Director of Technical Standards at Qualcomm, and a Principal Member of Research Staff at Nokia Corporation. His research interests include video coding, storage, transport, and multimedia systems. He has been an active contributor to various multimedia standards, including video codecs, file formats, RTP payload formats, and multimedia streaming and application systems, developed by various standardization organizations, including ITU-T VCEG, ISO/IEC MPEG, JVT, JCT-VC, JCT-3V, 3GPP SA4, IETF, AVS, DVB, ATSC, and DECE. He has been chairing the development of OMAF at MPEG, and has been an editor for several standards, including VVC, OMAF, all versions of HEVC, VVC file format, HEVC file format, layered HEVC file format, ITU-T H.271, SVC file format, MVC, RFC 6184, RFC 6190, RFC 7798, 3GPP TR 26.906, and 3GPP TR 26.948. He has coauthored about 1000 standardization contributions, over 60 academic articles, and about 500 families of patent applications (out of which 336 U.S. patents have been granted as of February 2021).

Robert Skupin received the Dipl.-Ing. (FH) degree in electrical engineering from H-BRS, Sankt Augustin, Germany, in 2009, and the M.S. degree in computer engineering from the Technische Universität Berlin, Germany, in 2014.

He has been with the Video Communication and Applications Department, Fraunhofer Heinrich-Hertz Institute, Berlin, Germany, since 2009. His research interests include efficient coding, storage, and transport of traditional video data as well as emerging formats in AR/VR. He has actively contributed to H.265/HEVC, its scalable extensions and, recently, H.266/VVC as well as related system standards, such as the ISO Base Media File Format, MPEG-2 Transport Stream, and the Omnidirectional MediA Format.

Miska M. Hannuksela (Member, IEEE) received the M.Sc. degree in engineering and the D.Sc. degree in technology from the Tampere University of Technology, Finland, in 1997 and 2010, respectively.

He has been with Nokia, Tampere, Finland, since 1996 in different roles, including research manager/leader positions in the areas of video and image compression, end-to-end multimedia systems, as well as sensor signal processing and context extraction. He is currently working as a Bell Labs Fellow and the Head of Video Research with Nokia Technologies. He has published more than 180 journals and conference papers and more than 1000 standardization contributions in JVET, JCT-VC, JVT, MPEG, 3GPP, and DVB. He has been an editor in several video and systems standards, including High Efficiency Image File Format (HEIF), Omnidirectional Media Format, RFC 3984, RFC 7798, as well as some parts of H.264/AVC, H.265/HEVC, and ISO Base Media File Format. He has granted patents from more than 130 patent families. His research interests include video compression, multimedia communication systems and formats, user experience and human perception of multimedia, and sensor signal processing. He has several best paper awards and received an award of the best doctoral thesis of the Tampere University of Technology in 2009 and the Scientific Achievement Award nominated by the Centre of Excellence of Signal Processing, Tampere University of Technology, in 2010. He was an Associate Editor of IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS OF VIDEO TECHNOLOGY from 2010 to 2015.

Sachin Deshpande (Senior Member, IEEE) received the B.E. degree from the Government College of Engineering Pune, the M.Tech. degree from IIT Mumbai, and the Ph.D. degree from the University of Washington, Seattle, WA, USA. He has been with Sharp Labs of America since January 2000, where he is currently a Distinguished Scientist. He has contributed several technologies to ISO/IEC and ITU-T Versatile Video Coding (VVC), High Efficiency Video Coding (HEVC), Scalable High Efficiency Video Coding (SHVC), MultiView High Efficiency Video Coding (MV-HEVC), MPEG Omnidirectional MediA Format (OMAF), MPEG Systems, and ATSC 3.0 International standards. He has several publications in reputed peer-reviewed journals, conferences, and books. He holds more than 160 U.S. patents. He received Sharp Laboratories "Department Inventor of the Year" award three times. He was a recipient of Sharp Corporation Excellence Award for outstanding development and contribution to next-generation video coding technologies and standard. He is also the Chair of MPEG Omnidirectional MediA Format Ad-Hoc group. He was the Vice-Chair of the ATSC S33 Specialist Group on Management and Protocols and the Chair of the ATSC S33-2 Service Announcement Ad-Hoc Group. He is also an Editor of ISO/ IEC 23090-2 and 23090-7 and has been the Editor of multiple ATSC 3.0 standards (A332, A333, and A338).

Hendry received the B.S. degree from the University of Indonesia, Jakarta, Indonesia, in 2002, and the M.S. and Ph.D. degrees from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Republic of Korea, in 2005 and 2011, respectively.

He was a Staff Engineer with Qualcomm, San Diego, CA, USA, and a Senior Staff Engineer with Huawei Technologies, San Diego. He is currently a Principal Engineer with LG Electronics, San Diego. Since 2005, he has been actively involved in the development of various multimedia standards, including the MPEG-21 Multimedia Framework, MPEG Multimedia Application Format (MAF), High efficiency Video Coding (HEVC) standard and its multi-layer extensions, ISO Base Media File Format, MPEG-2 Transport Stream, and recently, the Versatile Video Coding (VVC) standard. His research interests include video coding, storage, transport, and multimedia systems, for which he has coauthored many standardization technical contributions and patents. He was an editor for several MPEG standards, including MPEG-21 IPMP Components Base Profile, MPEG Music Application Format, MPEG Professional Archival Application Format, MPEG Multimedia Middleware Parts 2, 3, and 8, and MPEG-2 Transport Stream for carriage of HEVC.

Virginie Drugeon received the engineering degree from the École Nationale Supérieure des Télécommunications (ENST), Paris, France, in 2006. She has been working since 2007 as an Engineer for Panasonic Europe in Germany in the area of video coding and transmission. She has participated in the Joint Collaborative Team on Video Coding (JCT-VC) and Joint Video Experts Team (JVET) for the standardization of the video coding standards High Efficiency Video Coding (HEVC) and Versatile Video Coding (VVC), as well as to Digital Video Broadcasting (DVB) for the standardization of 3D Television (3DTV) and Ultra High Definition Television (UHDTV) broadcast, including High Dynamic Range and High Frame Rate. She is also the Chair of the Digital Video Broadcasting Technical Module subgroup on Audio/Video Coding (DVB TM-AVC subgroup).

Rickard Sjöberg received the M.Sc. degree in computer science from the Royal Institute of Technology (KTH), Stockholm, Sweden, in 1997. He has been with Ericsson, Stockholm, Sweden, and U.K., since 1996, primarily working with still image and video coding research, real-time video codec implementations, and subjective video encoding optimizations. He is currently an Expert of Video Coding with Digital Representation and Interaction, Ericsson Research, working as the Technical Leader of Ericsson's video coding research. Since 1997, he has also been an active participant in video codec standardization and has coauthored more than 300 standard contribution documents to the developments of the H.264/AVC, HEVC, and VVC video codec specifications.

Byeongdoo Choi received the B.S., M.S., and Ph.D. degrees in electronics engineering and computer science from Korea University in 2001, 2003, and 2007, respectively.

He is currently a Principal Researcher/Manager of Media Standards with Tencent America LLC. Until 2017, he had been a Director with Samsung Electronics Company Ltd., to lead media standardization and develop video-based algorithms and solutions. He was also a Principal Researcher with Sharp Labs of America from 2017 to 2018. He joined the Fraunhofer Institute for Telecommunication, Heinrich-Hertz-Institut (HHI), Berlin, Germany, as a Visiting Scholar, from 2007 to 2008. Since 2009, he has actively contributed to MPEG, JVET, JCT-VC, and JCT-3V for VVC, HEVC, MV-HEVC, SHVC, OMAF, and MPEG-I standardization. He has led multiple ad-hoc groups for media standards with chairmanship or editorship. He is also a Board of Director Member and the Vice-Chairman of Technical Working Group with 8K Association. He has been the author and the inventor of about 50 international academic articles, over 50 granted international patents, and about 300 technical standard contributions in the fields of video signal processing, video compression, visual quality enhancement, and multimedia communications.

Vadim Seregin received the M.S. degree (Hons.) in physics and the Ph.D. degree in mathematical modeling, numerical methods, and software systems from Moscow State University, Russia, in 2002 and 2005, respectively. From 2006 to 2011, he was with the Multimedia Platform Laboratory, Digital Media and Communications Research and Development Center, Samsung Electronics Company Ltd., Suwon, Republic of Korea. He is currently with Qualcomm Technologies Inc., San Diego, CA, USA. He has been an active contributor to HEVC, HEVC extensions, and VVC standardization activities. His research interests include image and video coding, video processing, and transmission.

Yago Sanchez received the M.Sc.-Ing. degree in telecommunications engineering from the Tecnun-Universidad de Navarra, Spain, in September 2009. From 2009 to 2019, he worked as a Researcher with the Image Communication Group of Prof. Thomas Wiegand, Technische Universität Berlin, and was a Guest Researcher with Fraunhofer Heinrich-Hertz Institute (HHI), Berlin, Germany. He is currently working as a Researcher with the Video Communication and Applications Department, Fraunhofer Heinrich-Hertz Institute. In 2013, he was visiting the End2End Mobile Video Research group of Alcatel Lucent-Bell Labs, USA, where he was doing research on mobile video delivery optimization for low-delay HTTP streaming. In addition, he is a coauthor of IETF RFC 7798, and the IETF RTP Payload Format for H.266/VVC Video. He has been an active participant in standardization activities in IETF, JCT-VC, JVET, MPEG, and 3GPP. His research interests include adaptive streaming services for IPTV and OTT services, adaptive streaming services for mobile TV over LTE networks, and virtual reality 360° video.

Jill M. Boyce (Fellow, IEEE) received the B.S. degree in electrical engineering from the University of Kansas in 1988 and the M.S.E. degree in electrical engineering from Princeton University in 1990.

She is currently an Intel Fellow and the Chief Media Architect with Intel, responsible for defining media hardware architectures for Intel's video hardware designs. She represents Intel at the Joint Video Exploration Team (JVET) of ITU-T SG16 and ISO/IEC MPEG. She also serves as an Associate Rapporteur for ITU-T VCEG. She is also an Editor of the VSEI specification. She was formerly the Director of Algorithms with Vidyo, Inc., where she led video and audio coding and processing algorithm development. She was formerly the VP of Research and Innovation Princeton for Technicolor, formerly Thomson. She was formerly with Lucent Technologies Bell Labs, AT&T Labs, and Hitachi America. She was an Associate Editor of IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY (CSVT) from 2006 to 2010.

Wade Wan received the B.S. degree in biomedical engineering and electrical engineering from Johns Hopkins University in 1995, and the M.S. and Ph.D. degrees in electrical engineering and computer science from the Massachusetts Institute of Technology in 1998 and 2002, respectively. He is currently a Distinguished Engineer of Video Technology with Broadcom Inc., Irvine, CA, USA, and has designed features, such as video decoding, post-processing, including HDR and colorimetry conversions, digital video recording, transcoding, and container format support that have been deployed in millions of set-top box products worldwide. He has been an active contributor to video coding, multimedia systems, and application standards and has been the principal member driving Broadcom's participation in various standardization organizations, including the Joint Video Team (JVT), Joint Collaborative Team on Video Coding (JCT-VC), and Joint Video Experts Team (JVET) of MPEG and ITU-T SG16 VCEG, the Alliance for Open Media (AoM), the Society of Cable Telecommunications Engineers (SCTE), and Digital Video Broadcasting (DVB).

Gary J. Sullivan (Fellow, IEEE) received the B.S. and M.Eng. degrees from the University of Louisville in 1982 and 1983, respectively, and the Ph.D. degree from the University of California, Los Angeles, in 1991.

He is currently a Video and Image Technology Architect with Microsoft Research. He has been a longstanding chairman/co-chairman of various video and image coding standardization activities in ITU-T VCEG, ISO/IEC MPEG, ISO/IEC JPEG, and in their joint collaborative teams since 1996. He has led the development of the Advanced Video Coding (AVC) standard (ITU-T H.264 | ISO/IEC 14496-10) the High Efficiency Video Coding (HEVC) standard (ITU-T H.265 | ISO/IEC 23008-2), the Versatile Video Coding (VVC) standard (ITU-T H.266 | ISO/IEC 23090-3), and various other projects. At Microsoft, he has been the Originator and a Lead Designer of the DirectX Video Acceleration (DXVA) video decoding feature of the Microsoft Windows operating system. He is also a fellow of SPIE. He received the IEEE Masaru Ibuka Consumer Electronics Award, the IEEE Consumer Electronics Engineering Excellence Award, two IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY (TCSVT) best paper awards, and the SMPTE Digital Processing Medal. The team efforts that he has led have been recognized by three Emmy awards.