IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 12 August 2022, accepted 13 September 2022, date of publication 22 September 2022, date of current version 3 October 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3208599

RESEARCH ARTICLE

Group-Based Adaptive Rendering System for 6DoF Immersive Video Streaming

SOONBIN LEE, (Graduate Student Member, IEEE),
JONG-BEOM JEONG, (Graduate Student Member, IEEE),
AND EUN-SEOK RYU, (Senior Member, IEEE)
Department of Computer Science Education, Sungkyunkwan University (SKKU), Seoul 03063, South Korea

Corresponding author: Eun-Seok Ryu (esryu@skku.edu)

This work was supported in part by the National Research Foundation of Korea (NRF) Grant funded by the Korean Government through Ministry of Science and ICT (MSIT) under Grant 2022R1F1A1074935, and in part by the Institute of Information and Communications Technology Planning and Evaluation (IITP) Grant funded by the Korean Government through MSIT (Development of low latency VR/AR streaming technology based on 5G edge cloud) under Grant 2020-0-00231-003.

ABSTRACT The Moving Picture Experts Group (MPEG) has started an immersive media standard project to enable multi-view video and depth representation in three-dimensional (3D) scenes. The MPEG Immersive Video (MIV) standard technology is intended to provide a limited 6 degrees of freedom (DoF) based on depth map-based image rendering (DIBR). The 6DoF immersive video system is still challenging because multiple high-quality video streams require high bandwidth and computing resources. This paper proposes a group-based adaptive rendering method for 6DoF immersive video streaming. With group-based MIV, each group can be transmitted independently, which enables adaptive transmission depending on the user's viewport. The proposed method derives weights from groups for view synthesis and allocates high-quality bitstreams according to a given viewport. This paper also discussed the results of the group-based approach in the MIV, and the advantages and drawbacks of this approach are detailed. In addition, pixel rate constraint analysis has been introduced to facilitate deployment with existing video codecs. On end-to-end evaluation metrics with TMIV anchor, the proposed method saves average 37.26% Bjontegaard-delta rate (BD-rate) on the peak signal-to-noise ratio (PSNR).

INDEX TERMS Virtual reality, metaverse, MPEG immersive video (MIV), adaptive streaming.

I. INTRODUCTION

With the current demand and interest in virtual reality (VR), the necessity for efficient VR technology is critical because of the large amount of data that has to be processed in the systems. Low latency and high-resolution are significant factors in increasing the quality of experience (QoE) of users. Moreover, the demand for technology to provide users with higher DoF is also growing. In these media markets and technological movements, MPEG has established an immersive media standard project to facilitate the compression, sharing, and distribution of immersive media between various devices and platforms. Specifically, the MPEG-immersive (MPEG-I) visual group is developing a standard for coding immersive media, called MPEG Immersive Video (MIV). The MPEG-I defined three types of DoF for users [1]. First, 3DoF supports only experiences where viewers are limited to rotational movements around pitch, yaw, and roll. Second, 3DoF+ supports a restricted movement of the user's head, which is an intermediate approach to 6DoF. Finally, 6DoF supports free viewpoint, which means full movement of the user. The MIV project was launched at the 125th MPEG meeting to discuss and evaluate the 3DoF+ and 6DoF coding technologies [2], [3].

In the 6DoF videos, the motion parallex feature can be achieved by using the DIBR technique with depth map information and associated camera parameters. Given a target view to generate, DIBR replaces the textures from the input

The associate editor coordinating the review of this manuscript and approving it for publication was Lei Wei.

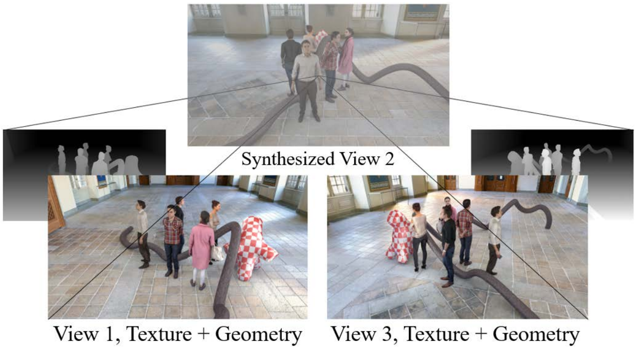


FIGURE 1. Example of DIBR representation.

videos with their new positions corresponding to the depth maps [4]. The MIV standard is designed to provide the capability to compress multiview video plus depth (MVD) representation [5], [6]. Because 6DoF should support the number of target views and movement of users, it requires high bandwidth and computing resources. Several video codecs with high efficiency video coding (HEVC), such as MV-HEVC or 3D HEVC, were developed to compress the MVD of the video data efficiently. However, those codecs would require a number of decoder instantiations, exceeding the capability of devices.

Figure 1 shows an example of DIBR. A key feature of the MIV encoder is packing the input videos into atlases, which are compact representations of the picture. The atlases have minimal pixel redundancies with multiple input views, allowing usage of the existing video codec with certain pixel rate constraints. However, due to the inter-view dependency, it becomes difficult to partially handle the generated atlas bitstream. In light of this issue, this paper investigate the feasibility of adaptive immersive video streaming. The proposed method utilizes a group-based approach that separates all input views into groups. Each group can be transmitted independently, and these sub-bitstreams enable adaptive streaming depending on the user's viewport. Overall, this paper presents an adaptive rendering system that can lead to viewport adaptive streaming that reduces bandwidth without significantly impacting the QoE.

In summary, the main contributions are as follows:

- This paper investigate MIV technology with a group-based approach and run a number of experiments to evaluate their performance, including an aspect of video codec compatibility.
- This paper implement view weighting techniques in a real system that calculates the contribution of each view and enables viewport-adaptive 6DoF streaming.
- This paper show that the proposed method achieves an efficient bitrate allocation under the target bitrate, while the group-based approach has tradeoffs between coding efficiency and resolution constraints.

The rest of the paper is organized as follows: This paper first introduce the background and related work in Section 2 regarding MIV standardization and group-based MIV. This paper then introduce the proposed adaptive rendering system and the view weighting calculation method in Section 3. In Section 4, this paper present the experimental setup and the metrics used in our evaluation. This paper also discuss the results of group-based MIV and a proposed adaptive rendering system. Our comprehensive set of evaluations included empirical evaluations based on common test conditions (CTCs) from the MPEG-I community.

II. BACKGROUND AND RELATED WORK

This section describes the MIV standard's immersive video technology in general. In addition, this paper briefly discusses group-based TMIV and multiview streaming research.

A. MIV STANDARDIZATION

The MPEG-I project (ISO/IEC 23090) is launched the first phase standardization in 2018 for 3DoF technology, which provides three dimensional degrees of freedom, including part-2 omnidirectional media format (OMAF) [7]. Subsequently, discussions were held on 6DoF technology, which provides full movement of the user in three-dimensional space, followed by a call for proposal (CfP) at the 125th meeting to define and standardize part-12 MPEG immersive video [2], [8], [9], [10], [11]. The concept of 6DoF system architecture contains pre-processing and post-processing modules for removing inter-view redundancy. Several studies implemented 6DoF systems by down-sampling multiple videos and eliminating the correlation among [4], [12]. Based on the proposed responses, MPEG-I proposed TMIV as a reference software for 6DoF video compression. TMIV supports pre-processing and post-processing for transmitting multiview videos to compress 6DoF videos more efficiently. In a scenario of streaming, 6DoF technologies must consider several representations of multiple views, requiring multiple decoder instantiations [13], [14], [15]. This makes it difficult to deploy 6DoF technology on edge devices such as mobile phones.

The MIV standard enables video codecs to handle multiple inputs through the inter-view redundancy removal process. With the development of the MIV standard, a test model for immersive video (TMIV) has been implemented as a reference software to comply with the MIV specification [12], [16]. The TMIV encoder extracts the patches from the input views and aggregates them for generating atlases. Notably, the atlases have two types of patches: basic view and additional view [17]. The basic view is complete as a single patch, and the additional views are multiple patches have no pixel redundancies with the basic view. The texture and geometry atlases can be encoded separately as videos using the existing video codec such as HEVC or versatile video coding (VVC), and the bitstreams are multiplexed together with metadata sub-bitstream to generate the MIV-compliant bitstream. The proposed algorithms are implemented in the

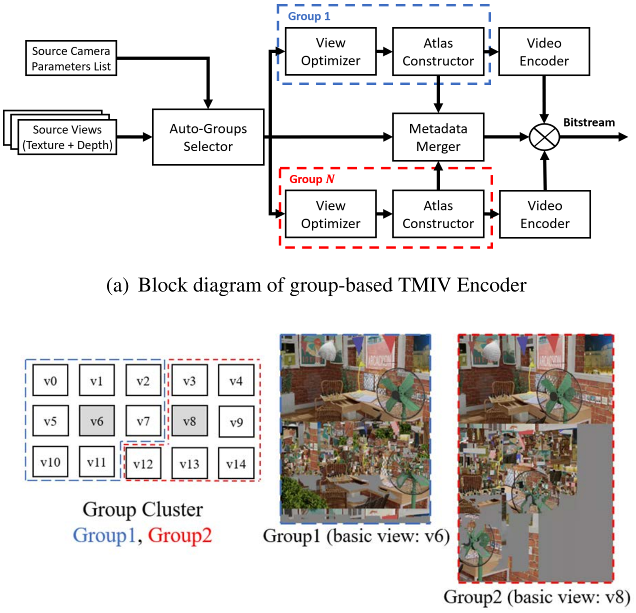
Algorithm 1 MIV Interview Redundancy Removal Algorithm

```
Input: Multiple texture videos with corresponding depth maps,
  camera parameters
Output: Pruned videos ($Atlases$)
  $V_S$ : set of all source views, e.g. $S = \{1, 2, 3, \ldots N\}$
  // Calculate the cost and allocate basic views as $B$
  $V_B$ : set of basic views
  $V_A \leftarrow V_S - V_B$ : set of additional views
  $Atlas \leftarrow V_B$ // Initialize atlas
  for $i \leftarrow 1$ to $B$ do
    for $j \leftarrow 1$ to $A$ do
      $V_i \oslash V_j = Patch_{ij}$ // $\oslash$ is pruning operator
      $Atlas \leftarrow Atlas \cup Patch_{ij}$ // Patch packing
    end for
  end for
```

TMIV software and the corresponding HEVC reference software. As mentioned previously, the main principle of MIV is to take several reference views that capture most of the information in the videos from specific positions of view, while supplementary information is collected into patches. The MIV specification allows for only two textures and two geometry atlases, enabling compatibility with existing codecs such as HEVC.

To facilitate deployment on any device, the pixel rate constraints are defined under CTCs in the MPEG-I group. Algorithm 1 illustrates the inter-view redundancy removal process used for generating atlases. The source views that are captured simultaneously from different positions are inputted to the TMIV. Then TMIV determines which source views are basic views and additional views.

From the source views, the views that contain most of the information in the scene are selected as the basic views. The remaining views are defined as additional views, and the basic and additional views are input to the pruner. The pruner removes the inter-view redundancy between the basic views and the additional views. Using the DIBR technique with a depth map, 2D textures are represented in 3D space. If two points from different views have the same position in the 3D space, one point is removed from its view. Overlapping information is removed from the additional views, and the pixels are extracted as rectangles, and the rectangles are defined as a patch. An atlas is a set of patches generated from multiple videos by TMIV. The number of videos to transmit decreases by generating atlases because only the residuals are extracted from the additional views and merged into the atlases. The basic views are completed copied into the basic views atlases. Consequently, the TMIV encoder generates the atlases and their metadata, and the video encoder encodes them to stream the videos to the client side. Further details about the MIV standard are described in [18].



(a) Block diagram of group-based TMIV Encoder

(b) Example of group-based encoding atlases

FIGURE 2. **Block diagram of group-based TMIV encoder and example of atlases.**

B. GROUP-BASED TMIV

The TMIV supports group-based encoding to produce better rendering results with local coherent projections [19]. Group-based encoding has shown improvement in the subjective and objective results of the TMIV, particularly at high bitrate levels. The main feature of group-based encoding is dividing all input source views into groups to be processed separately, and this feature enables the TMIV to preserve important regions (for example, foreground objects and occluded scenes) in each group. Moreover, this feature can also lead to sub-bitstream accessibility across groups. This paper focuses on the fact that this method facilitates sub-bitstream separation, which enables adaptive rendering in MIV. Such a region of interest (ROI)-based approach can lead to the representation of viewport-dependent streaming in multiview, depending on the field-of-view (FoV) of the viewer.

This method simply separates all input source views into groups using camera parameters, and each group encodes the views independently. First, the view pool includes all the input source views, and then the camera parameters are listed. Then, the dominant axis is assigned as the axis with the largest valid range in the X, Y, and Z coordinates. The dominant axis is used to set the key direction, and the closest camera is selected in ascending order by distance from the key direction. The selected camera is labeled as the current group and removed from the pool of views. The second key position is assigned, and the process is repeated, covering all source views across the chosen number of groups. Using group-based coding, MIV produces better rendering results, especially for natural content sequences. [20].

Figure 2 illustrates an example of a group cluster using group-based encoding. The example sequence

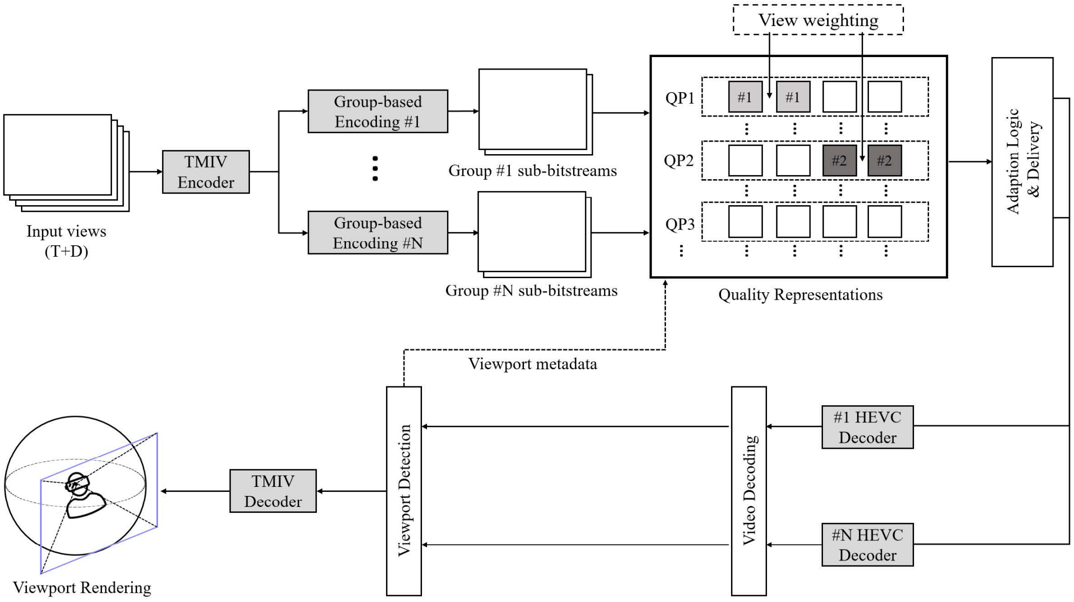


FIGURE 3. **Proposed adaptive immersive video rendering framework.**

is a 3 × 5 camera array of content. v6 and v8 indicate the basic view in each group. It is worth noting that all pixels in the basic view are preserved. In other words, as the number of groups increases, each important region is preserved, and the reconstruction of views becomes more accurate. However, limiting the pruning range has the disadvantage of increasing the overall pixel rate to be processed. The results will be discussed in further detail in Section 4.

III. PROPOSED SYSTEM

Traditional 360 video adaptive streaming allows tiles of quality to be mixed to trade off the quality between different regions of the videos and to utilize the bandwidth [21]. The rate between the tiles was efficiently allocated in the [22], and several useful conclusions were derived as a result of the experiments. The [23] suggested a technique for determining the ROI and allocating high quality tiles to those that correspond to the region. There are studies on quality allocation method for multi-view streaming in an environment with limited bandwidth. Priority-based adaptive multi-view streaming was proposed in [24] as a way to improve the quality of high-priority views in networks with limited capacity. Discussions on how to setup multi-view streaming have been explored from a system perspective, and streaming systems that consider buffer sharing and parallel processing have been presented [25]. An interactive multi-view adaptive streaming system was also developed in [26], which considers rate-distortion model-based quality allocation.

Although many multi-view streaming studies have been studied to render users' viewpoints with adaptive quality, these studies have limitations that require optimized equipment due to the large number of decoder instantiations required. This paper proposes a multiview streaming system based on MIV technology. To the best of our knowledge, this is the first work on the 6DoF adaptive streaming based on MIV standard technology. Figure 3 shows the proposed system for adaptive rendering. The system we propose includes the following considerations.

A. CONSIDERATIONS

1) DECODER FEASIBILITY

The MIV technology enables services to be provided on current-generation or near-next-generation hardware platforms. The pixel rate and simultaneous decoder instantiation restrictions become key considerations when considering actual implementation on real hardware. As the number of video decoder instantiations increases, it is clear that parallel processing performance decreases significantly. Therefore, minimizing the number of instantiations should be considered in multiview streaming. From these results, a practical maximum of two video decoders is assumed. If the lower limits of frame rates are stretched down to 30 fps, then up to four decoders can be instantiated. Based on this constraint, the TMIV tried to minimize the spatial resolution of the input videos to below 4096 × 2048, even though the process of inter-view redundancy removal causes information loss. Furthermore, beyond the HEVC decoder, the view must be rendered by a view synthesizer using the decoded streams. The larger the number of streams, the larger the memory and buffer capacity required.

Table 1 presents the pixel rate constraints for immersive videos. The maximum luma sample rate is the luma sample value per second across all decoders. The maximum luma picture size is the picture size value of each decoder instantiation. 'MP' means megapixel, which is a pixel rate per second

TABLE 1. Pixel rate constraint condition in MPEG-I.

Category	Max Sample Rate (Pixel rate per second)	Max Picture Size (Resolution)	Max Decoder Instance
Low pixel rate (HEVC Main 10 Profile Level 5.2)	1,069,547,520 (32 MP@30fps)	8,912,896 (4K, 4096×2048)	4
High pixel rate (HEVC Main 10 Profile Level 6.2)	4,278,190,080 (128 MP@30fps)	35,651,584 (8K, 8192×4096)	4

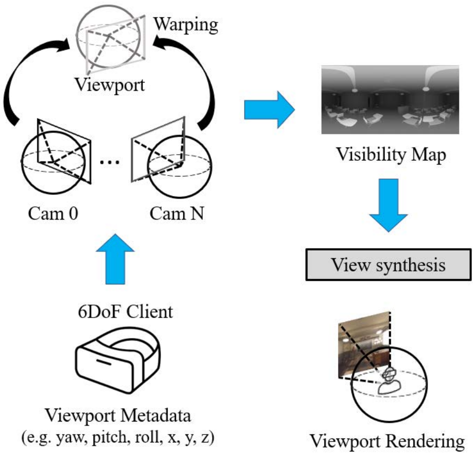


FIGURE 4. Rendering process in view weighting synthesizer (VWS).

TABLE 2. Results of group-based encoding and view weights (normalized, $N = 2$).

Class	view weight (p01)	view weight (p02)	view weight (p03)

S-1	**G1(64.95%)**
	G2(35.05%)
		G1(81.00%)
		G2(19.00%)
			G1(76.19%)
			G2(23.81%)

S-2	**G1(98.07%)**
	G2(1.93%)
		G1(10.64%)
		G2(89.36%)
			G1(7.18%)
			G2(92.82%)

S-3	**G1(80.99%)**
	G2(19.01%)
		G1(96.17%)
		G2(3.83%)
			G1(10.51%)
			G2(89.49%)

S-4	**G1(86.77%)**
	G2(13.23%)
		G1(20.99%)
		G2(79.01%)
			G1(90.44%)
			G2(9.56%)

S-5	**G1(92.55%)**
	G2(19.01%)
		G1(93.53%)
		G2(19.01%)
			G1(95.22%)
			G2(19.01%)

S-6	**G1(84.32%)**
	G2(15.68%)
		G1(90.97%)
		G2(9.03%)
			G1(10.99%)
			G2(89.01%)

in the table. The maximum number of simultaneous decoder instantiations is four. The MIV standard focuses on the low pixel rate condition because immersive video technology is already quite heavy for current devices. These considerations are critical for ideal multiview video streaming, such as real-time streaming. In this paper, the proposed method will be evaluated in both low ($N = 2, 3$) and high ($N = 4, 5$) conditions.

2) SPATIAL RANDOM ACCESS

The atlas domain presents content in a different form from the input videos, unlike conventional video streaming. Because the atlas representation ignores spatial information in the original representation, traditional adaptive streaming approaches are difficult to implement. This paper proposes an adaptive framework for multiview streaming based on group-based TMIV. In our framework, multiview processing is simplified using the MIV standard technology. Although the technical motivation for group-based encoding is to increase rendering quality, each group has no dependency on the others through the processing of TMIV. Accordingly, the sub-bitstream of each group can be independently transmitted and reconstructed. The discussions on transmission techniques through this sub-bitstream are underway in 360 video, including RoI-based and tile-based streaming [21], [27], [28]. As a result, discussions are also actively underway considering spatial random access, one of the main functions of next-generation media [29], but discussions have not yet been actively conducted on immersive video technology.

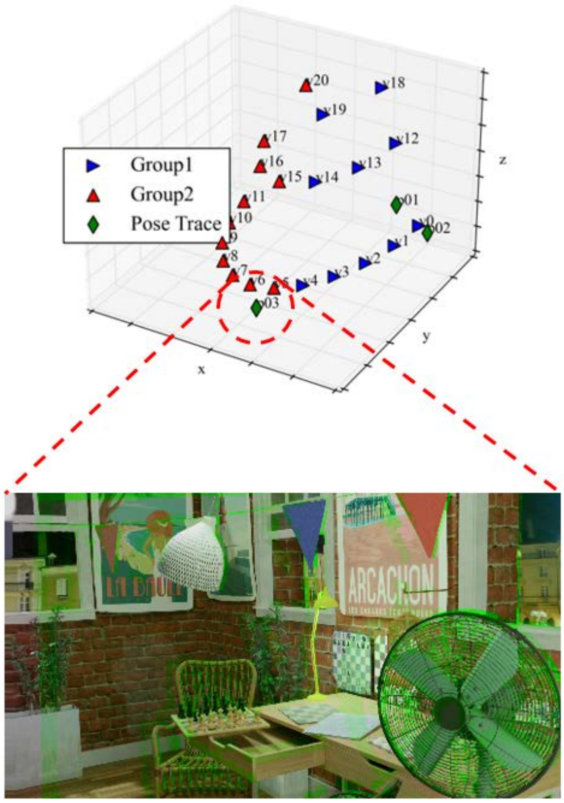
B. ADAPTIVE RENDERING

1) VIEWPORT WEIGHT CALCULATION

In the MIV standard, view weighting synthesizer (VWS) is adopted as a renderer software for synthesizing and rendering virtual viewpoints in TMIV [30]. The main feature of VWS is that when trying to synthesize a virtual viewpoint, the blending weight is calculated and reflected in the rendering process [31]. VWS generates a warped geometry map for each input view by unprojecting or reprojecting pixels from this view towards the target view. When several warped images are blended, a color value for each pixel inside the virtual view is computed using the weighted average of the warped pixels' color information. This calculation can be expressed as:

$$c_v(p) = \frac{\left(\sum_{i=0}^{N-1} w_i(p) \cdot c_i^{\text{warp}}(p)\right)}{\sum_{i=0}^{N-1} w_i(p)} \tag{1}$$

$c_v(p)$ is the color value of a pixel p in n a viewport image and $w_i(p)$ is a blending weight of a pixel p in the i-th warped image. Specifically, this paper use the blending weight factor $w_i(p)$ of a warped image, namely the visibility map. The rendering process with VWS is represented in Figure 4. Using this visibility map, the contribution of input views is calculated, and this information is used in the rendering process to synthesize more accurate and high-quality virtual viewpoints for view blending. To achieve efficient 6DoF streaming, this paper propose an adaptive rendering method for immersive video streaming. This paper implement the group weighting module with the visibility map. When transmitting 6DoF video to render a scene, we have to choose the quality of representation of bitstreams. Considering the rendering method of VWS, this paper proposes a method of calculating the contribution of each group and reflecting the contribution information when transmitting the bitstream of the priority group. A group containing many videos with a high contribution is defined as a priority viewpoint group in the adaptive rendering framework.



(a)

FIGURE 5. **(a): Camera coordinate information for each viewpoint ($N = 2$), (b): Rendering result of *p*03 view, group 1 (19.01%) sequence displays only luma colors for visualization.**

To render a virtual viewpoint, a bitstream with high quality applied to the priority viewpoint group is chosen. When selecting the bitstream quality, the contribution from the visibility map is used, and the user's viewpoint is rendered. Table 2 shows the contribution and adoption of priority viewpoint groups for each pose trace under common experimental conditions. In each user viewport, the contribution of *p01, p02, p03* points in rendering was calculated for all views, and then summed up for each group. Figure 5 shows the camera coordinates in the group sequence and the virtual viewpoint coordinates of the first frame of the viewpoint using the user pose trace dataset. As a result of the group division algorithm mentioned above, each camera viewpoint is divided into groups and the virtual viewpoint is synthesized mostly by utilizing a viewpoint with high contribution. For example, *p02* mainly utilize the viewpoint of group 1 to synthesize the viewpoint, and in the case of *p01* and *p03*, the viewpoint is used in combination. Therefore, even with the same bandwidth, the viewpoint group with a high contribution to the user's viewpoint will be able to efficiently adaptive stream in immersive video by transmitting a high-quality bitstream.

2) BIT ALLOCATION ALGORITHM

To validate the method proposed in this paper, a target bitrate is determined based on a streaming scenario. When a user's viewpoint is given, a priority viewpoint group is determined by calculating the viewpoint contribution of each view. Using the weight information, a bitstream of high quality is selected first for the viewpoint group to allocate a bitrate, and a group other than the viewpoint group selects and transmits a quality bitstream corresponding to the remaining bitrates. Atlases are pre-encoded with different quality representations that can be transmitted independently using group-based encoding in the proposed system. When the user's gaze information is given, a renderer calculates the contribution of each view to the virtual viewpoint. Then, the view weight module derives the bitrate allocation using these contributions. Once the target bitrates have been determined, a basic bitrate allocation strategy is implemented. The target bitrates are considered to be provided at high quality to the priority group first in the bit allocation stage. All groups are assigned to the lowest quality bitstream, and then a high quality bitstream is selected so that the priority group occupies as much view weight as possible. As a result, the bitstream of a group with a high contribution is assigned to the high bitrate to enable adaptive rendering at a given target bitrate. The total target bitrate is specified as R_t, the bitrate of the i-th group bitstream is specified as g_{bit_i}. Algorithm 2 details the operation of the proposed method of allocating bitrates to each group based on computed view weighting information.

Algorithm 2 Proposed Bitrate Allocation Algorithm Based on View Weighting

```
$R_t$ : target bitrate $\{5, 9, 16, 28, 50\}$ [Mbps]
$L$ : quality level QP=$\{22, 27, 32, 37, 42, 47\}$ representations
$g_i^L$ : each i-th group atlas bitstream with a bitrate corresponding to $L$
$g_{bit_i}^L$ : bitrate of $g_i^L$ of $L$ representation
$\mathcal{G}$ : list of groups sorted by view weight in each group, $\forall_i \in \mathcal{G}$
$Budget \leftarrow R_t - \sum g_{bit}^{min}$ % initialization for minimum bitrate
while $\mathcal{G} \neq \emptyset$ do
    for $L' \leq L$ do
        if $g_{bit_i}^{L'} - g_{bit_i}^{L} \leq Budget$ then
            $g_i^L \leftarrow g_i^{L'}$ % higher quality level $L'$ assignment
            $Budget \leftarrow Budget - (g_{bit_i}^{L'} - g_{bit_i}^{L})$
        else
            $\mathcal{G} \leftarrow \mathcal{G} - \{g_i\}$
        end if
    end for
    $i \leftarrow i + 1$ % next priority group
end while
```

IV. EXPERIMENTAL RESULTS

In this section, the efficiency of the proposed method is verified through video quality measurement for the virtual

TABLE 3. Characteristics of the immersive video test sequence.

Sequence name	Class	No. of source views	Resolution	View FoV (Field of view)
Painter	S-1	16 (4×4)	2048×1088	50° × 37°
Frog	S-2	13 (13×1)	1920×1080	63.65° × 38.47°
Fencing	S-3	10 (10×1)	1920×1080	63° × 48°
Carpark	S-4	9 (9×1)	1920×1088	63° × 48°
Hall	S-5	9 (9×1)	1920×1088	63° × 48°
Street	S-6	9 (9×1)	1920×1088	63° × 48°

TABLE 4. Experimental setting.

Items	Experimental values
Target bitrate	$\{5, 9, 16, 28, 50\}$ Mbps
Depth QP	Linear mapping (2)
GOP, framerate	GOP : 16, framerate : 30 fps
Evaluation frames	97 frames
Pose trace	*p01, p02, p03*
Quality switching frame	17,33,49,65,81 frame

TABLE 5. Comparison of pixel rate (atlas resolution) for the number of groups *N*. 100% is 4096 × 2048 × 4, the constraint on pixel rate specified in MPEG-I CTCs [33].

Sequence	S-1	S-2	S-3	S-4	S-5	S-6
Anchor	68%	72%	66%	61%	61%	55%
$N = 2$	64%	72%	60%	60%	60%	54%
$N = 3$	89%	88%	79%	80%	79%	76%
$N = 4$	**112%**	**109%**	**102%**	**105%**	**102%**	98%
$N = 5$	**134%**	**135%**	**132%**	**121%**	**118%**	**115%**

TABLE 6. Average BDBR and IV-BDBR performance comparison with TMIV anchor (*N* = 2,3).

Sequence	BDBR (%)		IV-BDBR (%)	
	Group-based	Proposed	Group-based	Proposed
Painter	-10.30%	-15.14%	-9.77%	-17.22%
Frog	7.07%	-13.35%	9.55%	-8.75%
Fencing	-47.99%	-61.77%	-42.25%	-52.22%
Carpark	-18.08%	-29.12%	-13.24%	-22.56%
Hall	-38.16%	-45.55%	-32.22%	-38.81%
Street	-40.65%	-52.51%	-33.42%	-45.56%
Average	**-24.68%**	**-36.24%**	**-20.22%**	**-30.85%**

TABLE 7. Average BDBR and IV-BDBR performance comparison with TMIV anchor (*N* = 4,5).

Sequence	BDBR (%)		IV-BDBR (%)	
	Group-based	Proposed	Group-based	Proposed
Painter	8.65%	-5.14%	12.24%	-7.88%
Frog	-20.17%	-30.19%	-22.13%	-28.89%
Fencing	-43.99%	-50.91%	-38.87%	-42.13%
Carpark	-26.32%	-35.12%	-18.87%	-25.57%
Hall	-41.16%	-52.47%	-35.52%	-45.24%
Street	-45.65%	-55.96%	-40.22%	-43.55%
Average	**-28.10%**	**-38.29%**	**-23.89%**	**-32.21%**

viewpoint rendered through the corresponding pose trace data set.

A. EXPERIMENTAL CONDITIONS

The proposed experimental method was conducted in compliance with CTCs, along with TMIV 6.0, a test model of MPEG-I with HM 16.16 [16], [32]. Table 3 illustrates the characteristics of the immersive video test sequence. The experiments used *Painter*, *Frog*, *Fencing*, *Carpark*, *Hall*, and *Street* as 6 MIV test sequences in the MPEG-I CTCs [33]. *Frog* and *Fencing* are linear camera arrays, with 13 (13 × 1) and 10 (10 × 1) views, respectively. *Carpark*, *Hall*, and *Street* contain 9 (9 × 1) views with a linear camera array, respectively. *Painter* has 16 views with a planar camera array, which is 16 (4 × 4) array. Because the proposed approach is based on group-based encoding, performance may differ considerably on the camera arrangement. Preprocessing and post-processing of multiple videos were conducted by TMIV, and the pre-processed videos were encoded by HM. The target bitrate has different bitrates depending on the video content, and the bitrate of the depth map complies with the common experimental conditions in which a linear transformation equation of the texture quantization parameter (QP) is presented [33]. According to the CTCs, the test sequences were encoded for target bitrate. For the proposed method, the experiments used immersive video depth estimation software (IVDE) 3.0 [34] to generate the geometry of the test sequences. To synthesize the virtual view, the view weighting synthesizer (VWS) 3.5 [30] was used. The group of picture (GOP) size is set to 16, and the total number of encoded frames in each view is 97. The framerate is set to 30 fps. Each video sequence's QP points are matched to the target bitrates under MPEG-I CTCs, which are $R_t = \{5, 9, 16, 28, 50\}$ Mbps. In the MIV anchor, the texture QPs are sequence-dependent, and the geometry QPs are simply linear mappings to the texture QPs. where Q_t is QP value of texture atlases and Q_g is QP value of geometry atlases. The mapping formulation with rounding operation is expressed as:

$$Q_g = \max\left(1, [-14.2 + 0.8\, Q_t]\right) \qquad (2)$$

B. RESULTS ANALYSIS

The proposed method has been evaluated using an immersive video quality assessment for the virtual viewpoint rendered by the pose trace dataset. This paper demonstrate the effectiveness of the proposed method by presenting objective and subjective evaluations for all three pose trace datasets. Through the proposed adaptive rendering framework, priority groups with a high contribution to rendering virtual viewpoints are transmitted with high quality, enabling more efficient adaptive streaming than when transmitting each group's quality uniformly. Although the group-based encoding is an advantage to preserving a more complete view, the pixel rate of each atlas increases, consuming more resources to process them. The group-based encoding has the disadvantage of increasing redundancy between views compared to TMIV, The results of pixel rate are summarized in Table 5.

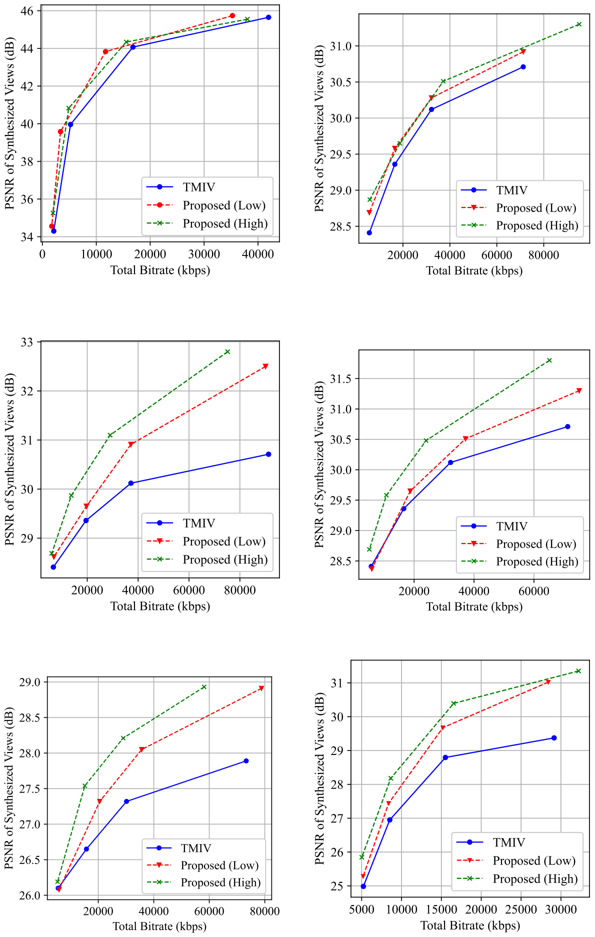


FIGURE 6. Objective performance comparison with synthesized pose trace views average BD-rate graph (TMIV, proposed (low, $N = 2, 3$), proposed (high, $N = 4, 5$).

TABLE 8. Configurations in the TMIV decoding time measurements.

Server Device	Intel Xeon E5-2687w v4 CPUs (24 cores, 48 threads)
Software Version	TMIV v6.0 / VWS v3.5
Render Resolution	1920x1080

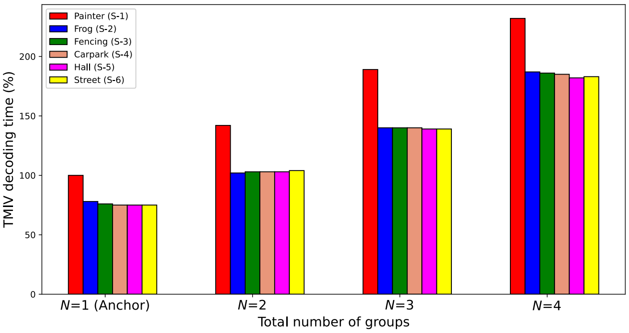


FIGURE 7. Comparison of relative TMIV decoding time for the number of groups N. (100% (S-1) is 49.64 seconds).

The group-based encoding has a negative impact on the pixel rate owing to the constraint of inter-view redundancy. To ensure compatibility with existing video codecs, TMIV considers the pixel rate as well as the coding efficiency, and it is important to understand these aspects.

With the proposed bit allocation algorithm, the group with a high contribution to view rendering is transmitted with high quality representation. This approach is related to prior tile-based 360 video streaming research, but the proposed method is more easily adaptable to group-based multiview streaming. The performance of the proposed method is evaluated by comparing the RD performance of other methods, such as the TMIV method from the convetional MPEG-I CTCs, and the group-based encoding, and the proposed method with adaptive rendering denoted by 'TMIV', 'Group-based', and 'Proposed', respectively. Table 5 and 6 demonstrates the Bjonteggrad delta bitrate (BDBR) and the Bjonteggrad delta PSNR (BDPSNR) performances of the group-based encoding and the proposed method, which uses the TMIV method as the anchor and the comparison basis [35]. Additionally, the experiments used the immersive video PSNR (IV-PSNR) for all sequences to evaluate the performance of the proposed method [36]. The results are measured by PSNR and IV-PSNR, respectively, according to the proposed method compared to the TMIV. For the low pixel rate condition ($N = 2, 3$), the proposed method shows an average BD-rate reduction of 36.24% in PSNR compared to TMIV anchor for six sequences, and 30.85% in IV-PSNR. Also the proposed method shows an average BD-rate reduction of 11.56% in PSNR and 10.63% in IV-PSNR compared to group-based TMIV. Table 8 shows the experimental conditions for the decoding time measurements.

Figure 7 presents a TMIV decoding time for the number of groups. As N increases, the overall decoding time also tends to increase. Even if parallel processing is currently possible in TMIV, the TMIV decoding time increases significantly according to the number of groups. The proposed method is capable of efficient transmission even though N is low. Therefore, the proposed method has the advantage of enabling more realistic deployment. The subjective quality comparison results for the proposed method with *p03* view are shown in Figure 8. For the all sequence, significant rendering artifacts are observed in TMIV anchor, but the proposed method achieves lower distortion in synthesized results with a total bitrate. A comparison result is shown when sequence is rendered at a target bitrate. The group-based encoding results show that viewpoint synthesis is improved during rendering, as reported in the proposal document [20]. Furthermore, the red region represents the part where quality improvement is identified in the result of applying the proposed adaptive quality allocation method. The proposed method allocates high quality to the priority group, so it can be observed that patches for texture parts are mainly replaced with high quality within the same bitrate, and subjective quality comparison is also improved.

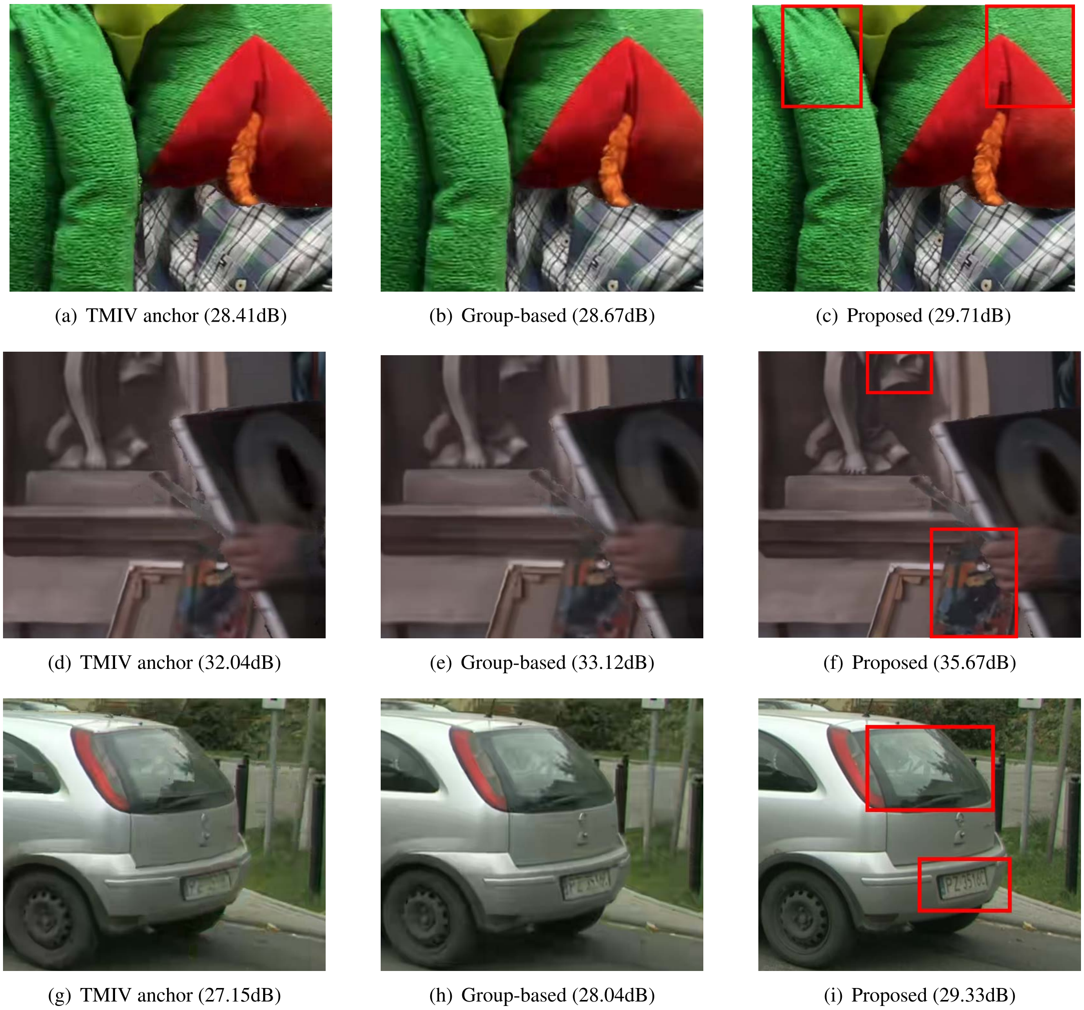

(a) TMIV anchor (28.41dB) (b) Group-based (28.67dB) (c) Proposed (29.71dB)
(d) TMIV anchor (32.04dB) (e) Group-based (33.12dB) (f) Proposed (35.67dB)
(g) TMIV anchor (27.15dB) (h) Group-based (28.04dB) (i) Proposed (29.33dB)

FIGURE 8. Subjective performance comparison with *p03* views (TMIV, group-based, proposed), $N = 2$; (a), (b), (c): Frog (S-2), target bitrate 7Mbps, (d), (e), (f): Painter (S-1), target bitrate 3Mbps, $N = 3$; (g), (h), (i): Street (S-6), target bitrate 1Mbps, $N = 4$.

V. CONCLUSION

This paper proposes a novel approach to adaptive streaming in terms of how group-based encoding can be transmitted independently. The view weighting calculation method is introduced to determine each group's adaptive bitrate allocation with visibility map. The priority groups with a high contribution to rendering virtual viewpoints are transmitted with high quality, enabling more efficient adaptive streaming. The experimental results demonstrate that the proposed technique is effective, especially on virtual viewports generated from pose trace datasets. Subsequently, efficient adaptive streaming is possible with low pixel rate conditions through the proposed method. The experimental results showed an average BD-rate reduction of 37.26% in PSNR and 31.53% in IV-PSNR with TMIV anchor.

REFERENCES

[1] *Proposed Architectures for Supporting Windowed 6DoF, Omnidirectional 6DoF 6DoF Media*, Standard ISO/IEC JTC1/SC29/WG11/M41555, Oct. 2017.

[2] *Call for Proposals on 3DoF+ Visual*, Standard ISO/IEC JTC1/SC29/WG11/M52994, Jan. 2019.

[3] *New Work Item Proposal on Coded Representation of Immersive Media*, 117th MPEG Meeting of ISO/IEC JTC 1/SC 29/WG 11, Standard MPEG/N16541, 2017.

[4] C. Fehn, "Depth-image-based rendering (DIBR), compression, and transmission for a new approach on 3D-TV," in *Stereoscopic Displays and Virtual Reality Systems XI*, vol. 5291, M. T. Bolas, A. J. Woods, J. O. Merritt, and S. A. Benton, Eds. Bellingham, WA, USA: SPIE, 2004, pp. 93–104.

[5] M. Domanski, O. Stankiewicz, K. Wegner, and T. Grajek, "Immersive visual media—MPEG-I: 360 video, virtual navigation and beyond," in *Proc. Int. Conf. Syst., Signals Image Process. (IWSSIP)*, May 2017, pp. 1–9.

[6] M. Wien, J. M. Boyce, T. Stockhammer, and W.-H. Peng, "Standardization status of immersive video coding," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 9, no. 1, pp. 5–17, Mar. 2019.

[7] *MPEG-I Project Plan*, Standard ISO/IEC JTC1/SC29/WG11/N17686, Apr. 2018.

[8] M. Domanski, A. Dziembowski, D. Mieloch, O. Stankiewicz, J. Stankowski, A. Grzelka, G. Lee, and J. Seo, *Call for Proposals on 3DoF+ Visual*, 129th MPEG Meeting of ISO/IEC JTC 1/SC 29/WG 11, Standard MPEG/N18145, 2019.

[9] *Technical Description of Proposal for Call for Proposals on 3DoF+ Visual prepared by Poznan University of Technology (PUT) and Electronics and Telecommunications Research Institute (ETRI)*, Standard ISO/IEC JTC1/SC29/WG11/M47407, Mar. 2019.

[10] *Description of Zhejiang University's Response to 3DoF+ Visual CfP*, Standard ISO/IEC JTC1/SC29/WG11/M47684, Mar. 2019.

[11] *Philips Response to CfP on 3DoF*, Standard ISO/IEC JTC1/SC29/WG11/M47179, Mar. 2019.

[12] J.-B. Jeong, S. Lee, D. Jang, and E.-S. Ryu, "Towards 3DoF+ 360 video streaming system for immersive media," *IEEE Access*, vol. 7, pp. 136399–136408, 2019.

[13] J.-B. Jeong, S. Lee, and E.-S. Ryu, "Sub-bitstream packing based lightweight tiled streaming for 6 degree of freedom immersive video," *Electron. Lett.*, vol. 57, no. 25, pp. 973–976, 2021.

[14] C.-F. Hsu, T.-H. Hung, and C.-H. Hsu, "Optimizing immersive video coding configurations using deep learning: A case study on TMIV," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 18, no. 1, pp. 1–25, Jan. 2022.

[15] J.-B. Jeong, S. Lee, and E.-S. Ryu, "Rethinking fatigue-aware 6DoF video streaming: Focusing on MPEG immersive video," in *Proc. Int. Conf. Inf. Netw. (ICOIN)*, Jan. 2022, pp. 304–309.

[16] B. Salahieh, J. Jung, A. Dziembowski, and C. Bachhuber, *Test Model 8 for MPEG Immersive Video*, 133rd MPEG Meeting of ISO/IEC JTC 1/SC 29/WG 4, Standard MPEG/WG4N0050, 2021.

[17] J. M. Boyce, M.-L. Chapel, Z. Deng, B. Kroon, and V. Malamal, *Proposed Draft Call for Proposals on 3DoF+*, 123rd MPEG Meeting of ISO/IEC JTC 1/SC 29/WG 11, Standard MPEG/M43973, 2018.

[18] M. Jill Boyce, R. Doré, A. Dziembowski, J. Fleureau, J. Jung, B. Kroon, B. Salahieh, V. K. M. Vadakital, and L. Yu, "MPEG immersive video coding standard," *Proc. IEEE*, vol. 109, no. 9, pp. 1521–1536, Sep. 2021.

[19] *Group-Based TMIV*, 127th MPEG meeting of ISO/IEC JTC 1/SC 29/WG 11, Standard MPEG/M49406, 2020.

[20] *Grouping and Anchor Study on MIV Content*, 131th MPEG Meeting of ISO/IEC JTC 1/SC 29/WG 11, Standard MPEG/M54151, 2020.

[21] X. Corbillon, G. Simon, A. Devlic, and J. Chakareski, "Viewport-adaptive navigable 360-degree video delivery," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2017, pp. 1–7.

[22] P. K. Yadav and W. T. Ooi, "Tile rate allocation for 360-degree tiled adaptive video streaming," in *Proc. 28th ACM Int. Conf. Multimedia*, New York, NY, USA, Oct. 2020, pp. 3724–3733.

[23] S. Lee, D. Jang, J. Jeong, and E.-S. Ryu, "Motion-constrained tile set based 360-degree video streaming using saliency map prediction," in *Proc. 29th ACM Workshop Netw. Operating Syst. Support Digit. Audio Video (NOSSDAV)*, New York, NY, USA, 2019, pp. 20–24.

[24] J. Chakareski, "Adaptive multiview video streaming: Challenges and opportunities," *IEEE Commun. Mag.*, vol. 51, no. 5, pp. 94–100, May 2013.

[25] X. Zhang, L. Toni, P. Frossard, and Y. Zhao, "Adaptive streaming in interactive multiview video systems," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 4, pp. 1130–1144, Apr. 2019.

[26] N. Carlsson, D. Eager, V. Krishnamoorthi, and T. Polishchuk, "Optimized adaptive streaming of multi-video stream bundles," *IEEE Trans. Multimedia*, vol. 19, no. 7, pp. 1637–1653, Jul. 2017.

[27] J. Son, D. Jang, and E.-S. Ryu, "Implementing motion-constrained tile and viewport extraction for VR streaming," in *Proc. 28th ACM SIGMM Workshop Netw. Operating Syst. Support Digit. Audio Video*, New York, NY, USA, Jun. 2018, pp. 61–66.

[28] K. M. A. M. Tourapis and J. Kim, *[V-PCC] A Tile Group Design for V-PCC*, 126th MPEG Meeting of ISO/IEC JTC 1/SC 29/WG 11, Standard MPEG/M47749, 2019.

[29] *Requirements for Immersive Media Access and Delivery*, 127th MPEG Meeting of ISO/IEC JTC 1/SC 29/WG 11, Standard MPEG/N18654, 2019.

[30] S. Kwak, J. Yun, J. Jeong, W.-S. Cheong, and J. Seo, *[MPEG-I Visual] Ray-Based Blending Weight for 6DoF View Synthesis*, 131st MPEG Meeting of ISO/IEC JTC 1/SC 29/WG 11, Standard MPEG/M54409, 2020.

[31] S. Kwak, J. Yun, J. Jeong, Y. Kim, I. Ihm, W. Cheong, and J. Seo, "View synthesis with sparse light field for 6DoF immersive video," *ETRI J.*, vol. 44, no. 1, pp. 24–37, Feb. 2022.

[32] Heinrich Hertz Institute. Fraunhofer Institute for Telecommunications. (2018). *High Efficiency Video Coding (HEVC) Reference Software HM*. [Online]. Available: https://hevc.hhi.fraunhofer.de/

[33] J. Jung and B. Kroon, *Common Test Conditions for MPEG Immersive Video*, 133rd MPEG Meeting of ISO/IEC JTC 1/SC 29/WG 4, Standard MPEG/WG4N0051, 2021.

[34] D. Mieloch, *Manual of Immersive Video Depth Estimation*, 130th MPEG Meeting of ISO/IEC JTC 1/SC 29/WG 11, Standard MPEG/W19224, 2020.

[35] G. Bjontegaard, *Calculation of Average PSNR Differences Between RD-curves*, Proceedings of the ITU-T Video Coding Experts Group (VCEG) Thirteenth Meeting, document VCEG-M33, 2001.

[36] A. Dziembowski, *Software Manual of IV-PSNR for Immersive Video*, 132nd MPEG Meeting of ISO/IEC JTC 1/SC 29/WG 4, Standard MPEG/WG4N00013, 2020.

SOONBIN LEE (Graduate Student Member, IEEE) received the B.S. degree from Gachon University, in February 2020, and the M.S. degree from Sungkyunkwan University (SKKU), in February 2022, where he is currently pursuing the Ph.D. degree with the Department of Computer Science Education. Since July 2022, he has been a Visiting Researcher at the Multimedia Communications Group, Fraunhofer Heinrich Hertz Institute (HHI). His current research interests include video compression standards, MPEG immersive video (MIV), and deep learning-based video coding.

JONG-BEOM JEONG (Graduate Student Member, IEEE) received the B.S. degree from Gachon University, in August 2018. He is currently pursuing the Ph.D. degree with the Department of Computer Science Education, Sungkyunkwan University (SKKU). From January 2020 to March 2020, he was a Visiting Student at the Department of Media Arts and Technology, University of California at Santa Barbara (UCSB), Santa Barbara, CA, USA. From August 2021 to January 2022, he was a Visiting Student at the Department of Computer and Information Technology, Purdue University, West Lafayette, IN, USA. His current research interests include video compression standards, MPEG immersive video (MIV), and deep learning-based video coding.

EUN-SEOK RYU (Senior Member, IEEE) received the B.S., M.S., and Ph.D. degrees in computer science from Korea University, Seoul, South Korea, in 1999, 2001, and 2008, respectively. He is currently an Associate Professor at the Department of Computer Education, Sungkyunkwan University (SKKU), Seoul. Prior to joining the university, in 2019, he was an Assistant Professor at the Department of Computer Engineering, Gachon University, Seongnam, South Korea, from 2015 to 2019. From 2014 to 2015, he was a Principal Engineer (Director) at Samsung Electronics, Suwon, South Korea, where he led a Multimedia Team. He was a Staff Engineer at InterDigital Labs, San Diego, CA, USA, from January 2011 to February 2014, where he researched and contributed to next generation video coding standards such as HEVC and SHVC. From September 2008 to December 2010, he was a full-time Visiting Research Scientist II at the Georgia Centers for Advanced Telecommunications Technology (GCATT), School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA. In 2008, he was a Research Professor at the Research Institute for Information and Communication Technology, Korea University. His research interests include multimedia computing systems that include video source coding and wireless mobile systems. He is a member of the ACM.