Efficient Group-Based Packing Strategy for 6DoF Immersive Video Streaming

Soonbin Lee, Jong-Beom Jeong, Eun-Seok Ryu
Department of Computer Science Education, Sungkyunkwan University, Seoul, Republic of South Korea
E-mails: {soonbinlee, uof4949, esryu}@skku.edu

***Abstract*—MPEG immersive video (MIV) standard technology is designed to support immersive video transmissions by removing the redundancy between multiview videos. Inter-view redundancy removal is achieved by packing the input videos into atlases, which are compact representations of multiview videos. Although this process can significantly reduce the pixel rates, it makes immersive video streaming difficult. The MPEG-I community has proposed group-based encoding as a way to improve the objective rendering performance. However, an efficient packing strategy for groups has yet to be investigated. This paper describes an efficient group-based packing strategy that enables immersive video streaming and parallel processing. The goal of this study was to evaluate the effectiveness of group-based encoding by comparing different packing strategies with MIV-encoding parameters. The experiment results show that the proposed packing method achieves better rendering results with a reasonable pixel rate.**

***Index Terms*—Virtual reality, MIV, 6DoF, immersive video**

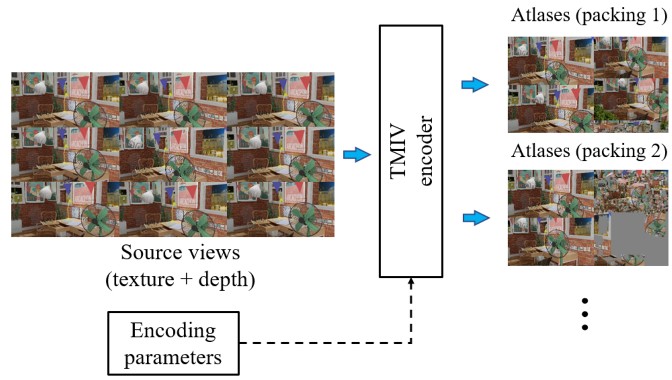


Fig. 1: Example of the proposed packing system using the TMIV encoder.

I. Introduction

With the current demand and interest in virtual reality (VR), the necessity for an efficient VR technology is critical because of the large number of data that must be processed in the systems [1]. A low latency and high resolution are significant factors that increase the user's quality of experience (QoE). In addition, the demand for graphical environments and realistic 360° scene representations on head-mounted displays (HMDs) is increasing, as is the need for technologies that provide users with a higher degrees of freedom (DoF). For example, a single 360-degree video does not allow the viewer to walk in a VR environment. In these media markets and technology movements, the Moving Picture Experts Group (ISO/IEC MPEG) has established immersive media standard projects to facilitate the compression, sharing, and distribution of immersive media between various devices and platforms.

In such an effort, the MPEG-Immersive (MPEG-I) subgroup defined three types of DoF for users. First, 3DoF supports only those experiences in which viewers are limited to rotational movements around the pitch, yaw, and roll. Second, 3DoF+ supports restricted movement of the user's head, which is an intermediate approach to 6DoF. Finally, the 6DoF supports a free viewpoint, which implies full movement of the user [2]–[4]. Motion parallax in 6DoF technologies can be achieved using the depth map-based image rendering (DIBR) technique with depth map information and associated camera parameters. Multiview video plus depth (MVD) representation technologies are still mainly studied for use in immersive videos. An immersive video includes several videos, including their corresponding depth maps [5]–[7]. Using the DIBR technique, 6DoF can generate the target views of a scene using a fixed number of input images and their corresponding depth maps. Given a target view to generate, the DIBR replaces the texture from the input videos with their new position corresponding to the depth maps. Immersive 6DoF videos should support many target views for user movements, and high computing resources and bandwidth are required to compress and transmit multiple videos acquired from different positions [8]. Although contributions to the compression performance improvement of immersive videos have been proposed in the MIV standard, MVD representation technologies must consider several representations of multiple views in a streaming scenario [9]–[11]. However, there have been only a few studies on immersive video streaming [12], [13]

The goal of this paper is to present a novel packing strategy for a group-based coding approach that can reduce the bandwidth and computing resources without significantly impacting the QoE. Although several techniques have been proposed for group-based MIVs, such techniques have not been investigated from an immersive video streaming perspective. The proposed packing strategy utilizes a group-based approach that separates all input views into groups, considering an efficient video representation. The evaluations included objective assessments based on common test conditions (CTCs), and the experiment results demonstrate the BD-rate gains and pixel rate savings achieved by the packing strategy.

ALGORITHM 1
MIV Interview Redundancy Removal Algorithm

Require: Multiple texture videos with corresponding depth maps,
camera parameters
Ensure: Pruned videos ($Atlases$)
V_S : set of all source views, e.g. $S = \{1, 2, 3, ...N\}$
// 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

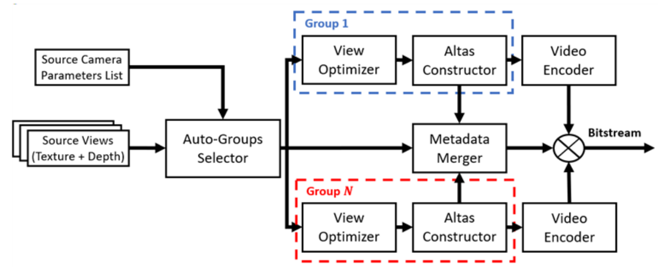


Fig. 2: Block diagram of the group-based TMIV encoding.

II. PRELIMINARIES

A. MPEG Immersive Video Standardization

The 3DoF+ and 6DoF technologies require the compression and processing of multiple videos to support the user's head and body movements, which is considerably challenging with high-efficiency video coding (HEVC) [14]. Because HEVC is designed for single video coding, it requires a large bandwidth and several computing limitations when handling multiple videos. MPEG-I proposed a test model for immersive video (TMIV) as a reference software for a 6DoF video compression [15], [16]. The TMIV supports pre-processing and post-processing for immersive video to more efficiently compress multiview videos [17].

Algorithm 1 illustrates the process of the inter-view redundancy removal in an MIV. The TMIV encoder divides the source views into basic and additional views, and the inter-view redundancy is removed using a pruning operation. The TMIV encoder collects the residual patches and merges them into atlases. Consequently, TMIV produces atlases that are pruned videos from all source views. This preprocessing significantly reduces the pixel rate, which also decreases the required decoder instantiations [8]. However, this patch-packing process has several hyperparameters, such as the ratio of basic views and the number of atlases [16], [18].

B. Group-based MPEG Immersive Video

The TMIV provides group-based encoding, which leads to more accurate rendering with locally coherent projections [19]. The subjective and objective results of the TMIV have been improved with group-based encoding, particularly for natural content or at high bitrate levels. The primary feature of group-based encoding is the division of all input source views into subgroups for processing, which enables the TMIV to preserve the relevant regions inside each group. This group division avoids an unnecessary depth estimation from all input source views, effectively avoiding view projection errors and generating better rendering results [20]. Figure 2 illustrates a block diagram of a group-based TMIV encoder. Although the technical motivation for group-based encoding is to increase the rendering quality, the subgroup has no dependence between groups through the processing of TMIV. This feature can lead to sub-bitstream accessibility across groups [21]–[25].

III. METHODOLOGY

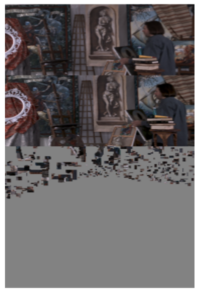
This section discusses the encoding parameters of the TMIV encoder. Because TMIV considers the pixel rate as well as the coding efficiency for compatibility with existing video codecs, it is important to investigate these factors. The experiments conducted in this study were aimed at evaluating the impact of the encoding parameters on the packing strategy for removing the inter-view redundancy algorithm.

TABLE I: Modified parameters in TMIV encoder configuration

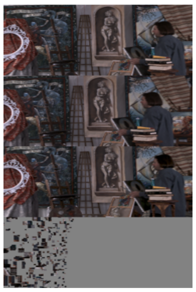
Encoding parameters	Experiment Configurations		
	Anchor	2G	3G
`numGroups`	1	2	3
`maxAtlases`	2	2	3
`minNonCodedViews`	3	2	1
`maxBasicViewFraction`	α = [0.25, 0.50, 0.75]		

Table I lists the specifications of the MIV-encoding parameters. The first parameter is the number of groups. With an increase in the number of groups, the scope of the inter-view redundancy removal was constrained, and the total bitrate increased. However, this grouping technique constrains the excessive depth estimation and view projection. As a result, MIV produces better rendering results, particularly when the quality of the depth map is low. The second parameter indicates the number of maximum atlases. The MPEG-I community defined the test condition constraints for the maximum number of simultaneous decoder instantiations in the CTCs. This is because MIVs are intended for use across a variety of platforms and devices. The number of decoder instantiations is a critical factor because parallel decoding is difficult to manage for many edge devices. Four decoder instantiations are recommended, two for texture atlases and two for depth map atlases. The `minNonCodedViews` parameter limits the number of basic views for each atlas. Determining the number of basic views is important to produce better packing, and this parameter preserves a meaningful objective evaluation of

(a) N=2, α=0.25



(b) N=2, α=0.5



(c) N=2, α=0.75

(d) N=2, α=1.0

Fig. 3: Example of TMIV encoded atlas (*Painter (NC-D)* sequence) for encoding parameters. N is `numGroup` and α is `maxBasicViewFraction`.

the source view positions. The `maxBasicViewFraction` parameter determines the basic view count. This parameter indicates the proportion of basic views among all views. This parameter allows for a flexible adjustment of the atlas representation of each group within the pixel rate constraints.

In this study, three experiment configurations were used to compare each strategy of the packing method for the encoding parameters. Figure 3 illustrates the results of the atlases for different values of the `maxBasicViewFraction` parameter. These examples demonstrate that the higher the parameter value is, the more basic the views that are included. Although there is an advantage to preserving a more complete view, the pixel rate of each atlas increases, consuming more resources to process them.

IV. Experimental Results

This section describes the experiment conditions, results, and analysis of various packing strategies. The presented experiment results are based on the use of group-based encoding, compared to a packing strategy in which immersive videos are encoded. The goal of the experiment was to assess the benefits of using a better rendering in a streaming scenario with 6DoF immersive videos compared to a baseline without the use of the modified parameters. This experiment was conducted under CTCs for immersive videos defined by MPEG-I [26]. Table II lists the characteristics of immersive video test sequences. The preprocessing and post-processing of multiview videos were conducted using TMIV 6.0, and the pre-processed videos

TABLE II: Characteristics of the immersive video test sequence

Sequence name	Class	No. of source views	Resolution	View FoV (Field of view)
Painter	NC-D	16 (4×4)	2048×1088	50° × 37°
Frog	NC-E	13 (13×1)	1920×1080	63.65° × 38.47°
Fencing	NC-L	10 (10×1)	1920×1080	63° × 48°
Carpark	NC-P	9 (9×1)	1920×1088	63° × 48°
Hall	NC-T	9 (9×1)	1920×1088	63° × 48°
Street	NC-U	9 (9×1)	1920×1088	63° × 48°

were encoded using HM [15], [27]. The group of pictures value was set to 16, and the frame rate was 30 fps.

TABLE III: BD-rate performances of the TMIV anchor compared to various packing methods (negative values indicate that the modified packing achieves a better performance)

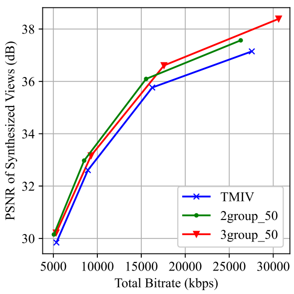
Class	2G_25 Y-PSNR BD-Rate	2G_50 Y-PSNR BD-Rate	2G_75 Y-PSNR BD-Rate	3G_25 Y-PSNR BD-Rate	3G_50 Y-PSNR BD-Rate	3G_75 Y-PSNR BD-Rate
NC-D	0.60%	-11.70%	-13.96%	3.53%	-8.64%	-4.46%
NC-E	19.73%	14.09%	-4.83%	2.51%	-7.83%	-11.75%
NC-L	-32.54%	-41.91%	-57.32%	-41.18%	-57.84%	-61.62%
NC-P	31.80%	-12.18%	-23.87%	0.55%	-24.01%	-24.01%
NC-T	-11.63%	-31.29%	-40.62%	12.12%	-44.94%	-44.94%
NC-U	58.68%	-23.50%	-52.94%	-22.66%	-56.72%	-56.72%
Average	**11.10%**	**-17.74%**	**-32.25%**	**-7.52%**	**-33.32%**	**-33.91%**

The inclusion of more basic views and fewer patches showed a better compression performance, and thus the TMIV anchor showed a slightly worse performance than the group-based encoding [20]. Group-based coding also increased the required pixel rate because each group selected their basic views individually and constrained the inter-view redundancy removal in the source views. The experiment results in Table III demonstrate the BD-rate gains of the various packing methods in comparison to the TMIV anchor.

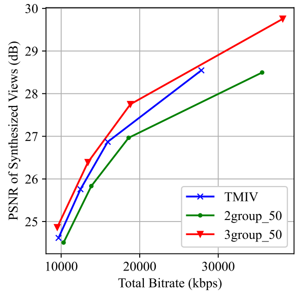
TABLE IV: Comparision of luma pixel samples with maximum luma pixel rate between anchor, 2G, 3G configurations (`maxBasicViewFraction` is set to 0.5).

Pixel ratio (%)	NC-D	NC-E	NC-L	NC-P	NC-T	NC-U
Anchor ($\alpha = 0.5$)	0.68	0.72	0.66	0.61	0.61	0.55
2G ($\alpha = 0.5$)	0.64	0.72	0.60	0.60	0.60	0.54
3G ($\alpha = 0.5$)	0.89	0.88	0.79	0.80	0.79	0.76

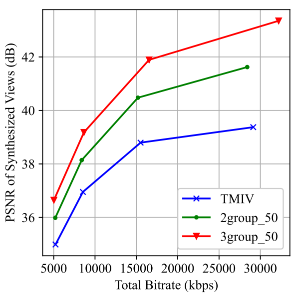
The pixel rate of packing strategies was also compared with maximum luma sample rate per frame (`4096×2048×4=33,554,432`), when the `maxBasicViewFraction` is set to 0.5. The results are summarized in Table IV. Although group-based encoding has a negative impact on the pixel rate owing to the constraint of inter-view redundancy, this method enables sub-bitstream extraction on the decoding side. To ensure compatibility with existing video codecs, TMIV considers the pixel rate as well as the coding efficiency, and it is therefore important to understand these aspects. As illustrated in Figure 4, different packing techniques can significantly reduce the bitrate required for transmission while preserving the objective video quality.



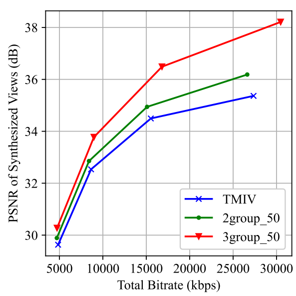
(a) Painter



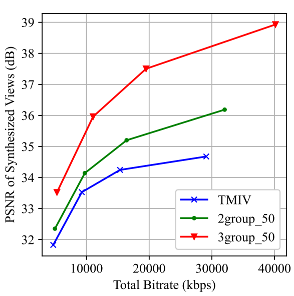
(b) Frog



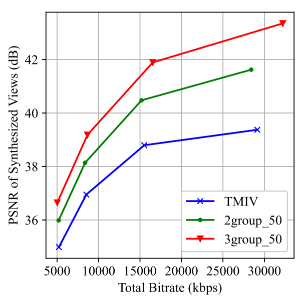
(c) Fencing



(d) Carpark



(e) Street



(f) Hall

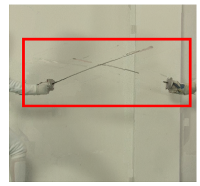
Fig. 4: Y-PSNR RD curves of the synthesized views for (a) *Painter (NC-D)*, (b) *Frog (NC-E)*, (c) *Fencing (NC-L)*, (d) *Carpark (NC-P)*, (e) *Street (NC-U)*, and (f) *Hall (NC-T)*. Here, `maxBasicViewFraction` is set to 0.5.

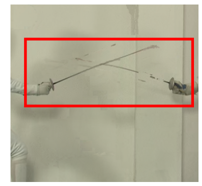
The experiment results show that preserving a more complete view is a significant factor in improving the rendering quality, particularly when using group-based encoding. A comparison of the subjective quality results of the packing method in the SE sequence is shown in Figure 5. The comparison results indicate when the synthesized view was rendered in the test sequences. As described above, it can be observed that the more basic views that are included in the subjective quality, the better the rendering result. As a reason for this, the more basic the views are, the smaller the number of errors occurring in the view synthesis of the patch. In particular, it is possible to observe a view synthesis error at the boundary of an object because the estimated depth map is inaccurate. These results showed the tendency according to the various packing strategies and encoding parameters.

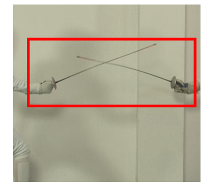

(a) Frog (TMIV)


(b) Frog (G2_50)


(c) Frog (G3_50)


(d) Fencing (TMIV)


(e) Fencing (G2_25)


(f) Fencing (G2_75)


(g) Street (TMIV)


(h) Street (G3_25)


(i) Street (G3_75)

Fig. 5: Synthesized view comparison with enlarged noticeable sections: (a), (b), (c) *Frog (NC-E) v4* rendered using TMIV anchor, G2_50, and G3_50 configurations. (d), (e), (f) *Fencing (NC-L) v4* rendered using TMIV anchor, G2_25, and G2_75 configurations. (g), (h), (i) *Street (NC-U) v6* rendered using TMIV anchor, G3_25, and G3_75 configurations.

V. Conclusion

In this paper, an efficient 6DoF immersive video packing strategy for the MIV coding standard was proposed. Specifically, the efficient representation of atlases was discussed in terms of the encoding parameters. It is important to explore these parameters because TMIV considers the pixel rate and coding efficiency as factors for compatibility with the video codecs. In particular, further study is required to develop efficient adaptive streaming based on the group-based packing algorithms discussed in this paper. Several packing strategies for MIV coding were explored in this study, and the experiment results indicate that the better the rendering, the more basic the views that are included. However, further study on the 6DoF streaming scenario in terms of pixel rate is required. Further research will be conducted on immersive video compression based on these observations.

Acknowledgment

This research was supported by the SungKyunKwan University and the BK21 FOUR(Graduate School Innovation) funded by the Ministry of Education(MOE, Korea) and National Research Foundation of Korea(NRF)

References

[1] M. Champel, T. Stockhammer, T. Fautier, E. Thomas, and R. Koenen, "Quality requirements for vr. 116th meeting of iso/iec jtc1/sc29/wg11, mpeg 116/m39532," 2016.

[2] M. Wien, J. M. Boyce, T. Stockhammer, and W.-H. Peng, "Standardization status of immersive video coding," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 9, no. 1, pp. 5–17, 2019.

[3] *Proposed architectures for supporting Windowed 6DoF, Omnidirectional 6DoF and 6DoF media*, ISO/IEC JTC1/SC29/WG11/M41555, Oct 2017.

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

[5] 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*, M. T. Bolas, A. J. Woods, J. O. Merritt, and S. A. Benton, Eds., vol. 5291, International Society for Optics and Photonics. SPIE, 2004, pp. 93 – 104. [Online]. Available: https://doi.org/10.1117/12.524762

[6] D. Mieloch, O. Stankiewicz, and M. Domański, "Depth map estimation for free-viewpoint television and virtual navigation," *IEEE Access*, vol. 8, pp. 5760–5776, 2020.

[7] A. Dziembowski, D. Mieloch, O. Stankiewicz, M. Domański, G. Lee, and J. Seo, "Virtual view synthesis for 3dof+ video," in *2019 Picture Coding Symposium (PCS)*, 2019, pp. 1–5.

[8] J. M. 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," *Proceedings of the IEEE*, pp. 1–16, 2021.

[9] J. Chakareski, "Adaptive multiview video streaming: challenges and opportunities," *IEEE Communications Magazine*, vol. 51, no. 5, pp. 94–100, 2013.

[10] X. Zhang, L. Toni, P. Frossard, and Y. Zhao, "Adaptive streaming in interactive multiview video systems," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. PP, 03 2018.

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

[12] J. Jeong, D. Jang, J.-W. Son, and E.-S. Ryu, "Bitrate efficient 3dof+ 360 video view synthesis for immersive vr video streaming," in *2018 International Conference on Information and Communication Technology Convergence (ICTC)*, 2018, pp. 581–586.

[13] J.-B. Jeong, S. Lee, and E.-S. Ryu, "Sub-bitstream packing based lightweight tiled streaming for 6 degree of freedom immersive video," *Electronics Letters*, vol. 57, no. 25, pp. 973–976, 2021. [Online]. Available: https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/ell2.12329

[14] G. J. Sullivan, J. Ohm, W. Han, and T. Wiegand, "Overview of the high efficiency video coding (hevc) standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1649–1668, 2012.

[15] 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, MPEG/WG4N0050*, 2021.

[16] 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, MPEG/N18145*, 2019.

[17] 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. 136 399–136 408, 2019.

[18] 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, MPEG/M43973*, 2018.

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

[20] "Grouping and anchor study on MIV content," *131th MPEG meeting of ISO/IEC JTC 1/SC 29/WG 11, MPEG/M54151*, 2020.

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

[22] J. Son, D. Jang, and E.-S. Ryu, "Implementing motion-constrained tile and viewport extraction for vr streaming," in *Proceedings of the 28th ACM SIGMM Workshop on Network and Operating Systems Support for Digital Audio and Video*, ser. NOSSDAV '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 61–66. [Online]. Available: https://doi.org/10.1145/3210445.3210455

[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 *Proceedings of the 29th ACM Workshop on Network and Operating Systems Support for Digital Audio and Video*, ser. NOSSDAV '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 20–24. [Online]. Available: https://doi.org/10.1145/3304112.3325614

[24] P. K. Yadav and W. T. Ooi, "Tile rate allocation for 360-degree tiled adaptive video streaming," in *Proceedings of the 28th ACM International Conference on Multimedia*, ser. MM '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 3724–3733. [Online]. Available: https://doi.org/10.1145/3394171.3413550

[25] J.-B. Jeong, S. Lee, I.-W. Ryu, T. T. Le, and E.-S. Ryu, "Towards viewport-dependent 6dof 360 video tiled streaming for virtual reality systems," in *Proceedings of the 28th ACM International Conference on Multimedia*, ser. MM '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 3687–3695. [Online]. Available: https://doi.org/10.1145/3394171.3413712

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

[27] H. H. I. Fraunhofer Institute for Telecommunications, "High efficiency video coding (hevc) reference software hm," https://hevc.hhi.fraunhofer.de/, 2018.