applied sciences

MDPI

Article

On the Selection of Transmitted Views for Decoder-Side Depth Estimation

Dominika Klóska, Adrian Dziembowski *, Adam Grzelka and Dawid Mieloch

Institute of Multimedia Telecommunications, Poznan University of Technology, 60-965 Poznań, Poland; dominika.kloska@put.poznan.pl (D.K.); adam.grzelka@put.poznan.pl (A.G.); dawid.mieloch@put.poznan.pl (D.M.)

* Correspondence: adrian.dziembowski@put.poznan.pl

Abstract

The selection of optimal views for transmission is critical for the coding efficiency of the MPEG Immersive Video (MIV) profile of Decoder-Side Depth Estimation (DSDE). Standard approaches, which favor a uniform camera distribution, often fail in scenes with complex geometry, leading to decreased quality of depth estimation, and thus, reduced quality of virtual views presented to a viewer. This paper proposes an adaptive view selection method that analyzes the scene's percentage of occluded regions. Based on this analysis, the encoder dynamically selects a transmission strategy: for scenes with a low occlusion ratio (smaller than 10%), a uniform layout is maintained to maximize spatial coverage; for scenes with a high occlusion ratio, the system switches to grouping cameras into stereo pairs, which are more robust for decreasing numbers of occlusions. Experiments conducted using the TMIV reference software demonstrated that this approach yields measurable quality gains (up to 2 dB BD-IVPSNR) for complex test sequences, such as MartialArts and Frog, without requiring any modifications to the decoder.

Keywords: immersive video; decoder-side depth estimation; video compression; depth maps

1. Introduction

Immersive video technologies [1] are continuously advancing, driven by the increasing demand for realistic and interactive multimedia experiences. These technologies implement various degrees of freedom, e.g., 3DoF, 3DoF+, and 6DoF [2,3]. The process of generating immersive video (Figure 1) requires using a multicamera system to register a 3D scene. Recent developments in immersive video have explored multiple configurations of multicamera systems, including linear, planar, and spherical camera arrangements.

The other component crucial for creating immersive video is the three-dimensional geometry of the registered scene, which can be obtained either from depth cameras or estimated using dedicated software. Since using depth cameras introduces a problem of interference between sensors [4], for the considerations presented in this paper, we will be assuming that depth information is obtained through the process of depth estimation.

Practical systems of immersive video [5] can enhance multicamera setups with a virtual camera placed among the real cameras, enabling dynamic and adaptive view synthesis to reflect the movement of the user in a 3D scene. In other words, a virtual camera is a viewport rendered by the system at the request of the viewer [6]. To represent all of this data, the multiview video plus depth (MVD [7]) format is usually used, as shown in Figure 2. It allows separate encoding of real views and their corresponding

Academic Editor: JungHwan Oh

Received: 21 November 2025
Revised: 12 December 2025
Accepted: 19 December 2025
Published: 20 December 2025

https://doi.org/10.3390/app16010072

depth maps using any available video encoder (HEVC, VVC [8,9]). Because traditional encoders were not developed for data such as depth maps (as they do not resemble naturally captured videos), alternative solutions have been explored (MV-HEVC, 3D-HEVC [10]). Unfortunately, experimental data show that their versatility is highly limited [11].

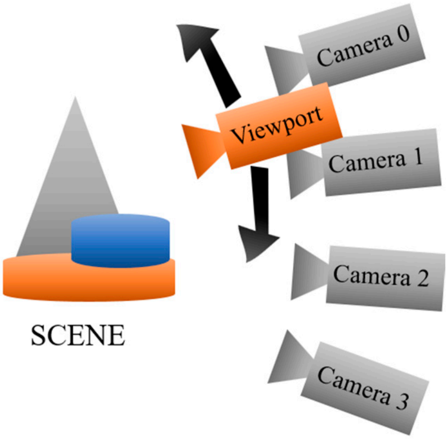

Figure 1. Scene acquisition in the immersive video system.

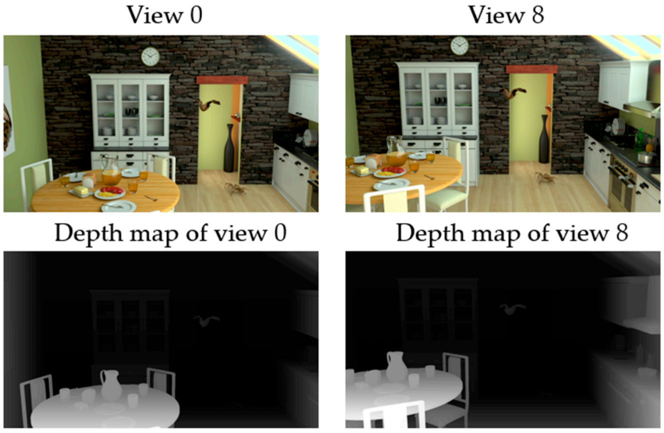

Figure 2. Multiview video plus depth format.

The latest approach to immersive video compression is the MPEG immersive video (MIV) standard [12], created by the ISO/IEC MPEG Video Coding. MIV utilizes multiview pre-processing and post-processing (Figure 3), making the resulting videos more efficient to compress with traditional video encoders. Additionally, MIV supports a range of profiles dedicated to different coding scenarios.

The MIV Main profile groups input views into a set of videos called atlases (usually four), which are shown in Figure 4. Each atlas is encoded independently. Input views, which are views captured by the multicamera system, are divided into basic and additional views. Basic views contain the most information about a scene and are usually packed into the first atlas. Additional views contain a large amount of redundant information, which is removed by the pruning process. This process creates small fragments (called patches) of the remaining view information. Patches are packed into the remaining space left in the first two atlases. The third and fourth atlases contain depth information corresponding to the first two atlases.

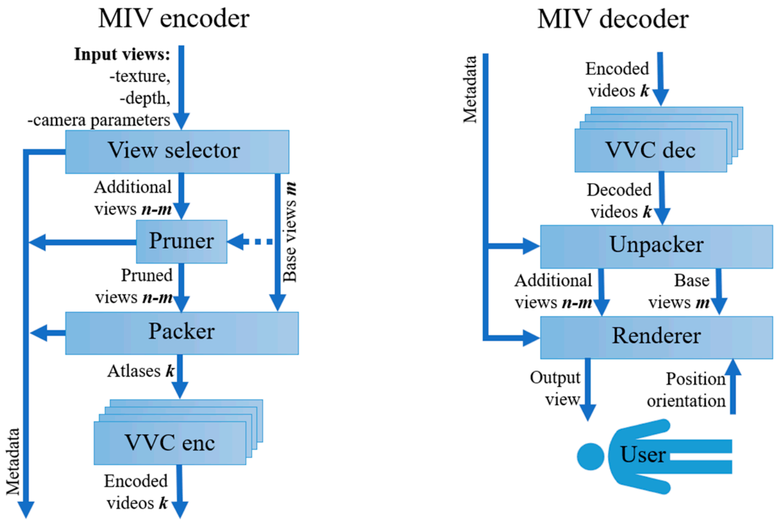

Figure 3. Simplified MIV scheme.

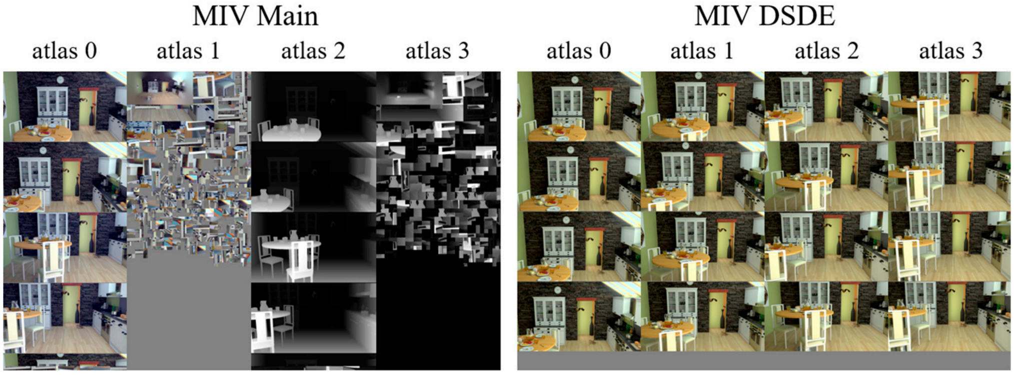

Figure 4. MIV Main and MIV DSDE atlases.

Another approach is called the MIV decoder-side depth estimation (DSDE) coding profile. Similarly to the MIV Main, this profile also utilizes atlases (Figure 4) consisting of multiple input views per atlas, but in this approach the atlases do not contain depth information. As depth is required to perform virtual view synthesis, it is estimated from the decoded views on the decoder side. In this scenario, the transmitted views should be chosen in such a way as to minimize mutual redundancy while maximizing coverage of scene content. The view selection process for the DSDE profile is further complicated by an intrinsic trade-off: increasing the number of transmitted views improves the accuracy of the depth estimates and, consequently, the quality of synthesized views, whereas a poorly chosen small subset can cause the depth estimator to fail, particularly in regions affected by occlusions. Hence, it is unclear whether the view selection method used for the MIV Main profile will remain effective for the DSDE profile, or whether selection criteria tailored to decoder-side depth estimation are required.

This paper addresses the challenge of optimal input view selection for the MIV DSDE profile, proposing a method that is adjusted to scene characteristics. We analyze how to ensure the highest possible depth-estimation quality when only a limited number of views can be transmitted, while simultaneously selecting them in a way that mitigates the negative impact of scene occlusions on the estimation process. Our proposed adaptive approach ensures the highest coding efficiency and visual quality across diverse immersive video content.

In Section 2, we review the state of the art in view selection, outline the issues that may arise during the view selection process, and provide a detailed analysis of how view selection operates in the MIV Main profile. In Section 3, we describe our proposed view selection method for the DSDE profile, which explicitly accounts for occlusions present in the scene. Section 4 presents an overview of the experiments, and Section 5 reports and analyzes the experimental results. Section 6 contains conclusions and future work.

2. View Selection for Immersive Video

View selection for immersive video is a complex process that requires careful consideration of the specific application scenario. A variety of factors influence the choice of input views in the DSDE scenario: the selected cameras determine not only the accuracy of estimated depth maps but also the quality of the synthesized virtual views [13], and therefore, ultimately, the overall quality of the immersive experience. Key challenges to consider include scene occlusions, finite image resolution, non-Lambertian surfaces, camera layout and type (e.g., linear, omnidirectional), and the decoding hardware's computational limits. In the following sections, we survey several view selection strategies designed to address these diverse problems and use cases.

2.1. View Selection Optimized for Virtual View Synthesis

This section describes the problem of selecting the best input views for the virtual view synthesis process in free viewpoint television (FTV) systems. As such systems require real-time synthesis, it is essential to limit the number of input views in order to maintain reasonable computational time. Adding more views to view synthesis [14–16] increases the quality of virtual views; nevertheless, it also increases computational time [6]. Therefore, an effective view selection method should aim to balance the quality of synthesized virtual views with the need to minimize computational complexity.

In the simplest implementations, a virtual view is synthesized from two manually selected real views [17]. The method described in [13] deals with the problem of choosing two views that will guarantee the highest quality of the synthesized virtual view. To achieve this, three main view synthesis challenges were taken into account:

1. Occlusions: Gaps occur where real views do not overlap; choosing cameras closest to the virtual viewpoint minimizes these holes.
2. Finite resolution: Objects appear at different scales across views; projecting from the views where each object is largest preserves geometric continuity.
3. Non-Lambertian reflectance: Surface brightness varies with angle; using the two cameras nearest the virtual position ensures more consistent lighting.

Through addressing the above challenges, it was shown that the best quality of the synthesized virtual view can be obtained when using two neighboring real views (nearest left and right), and this is true for any camera arrangement. However, this research considers view selection for virtual view synthesis in a scenario where all of the real views are available. In the scenario analyzed for the purpose of this article, we have only a subset of real views available at the decoder-side for depth estimation and virtual view synthesis. Consequently, it is not always possible to select the nearest-right and nearest-left views, and it is therefore necessary to investigate which selection strategy can provide a quality level closest to the scenario described in [13].

2.2. View Selection Optimized for Transmission

This section presents the problem of optimal selection of input views transmitted within the video bitstream [18] in immersive video systems for the MIV Main profile. In this scenario, only a subset of views available at the encoder side can be transmitted to the decoder. Selecting proper real views for transmission is crucial to achieving good quality on the decoder side. The method described in the previous section is unsuitable for

https://doi.org/10.3390/app16010072

this scenario, as it would require knowing which view was selected by the user prior to transmission, which is not possible.

The most straightforward approach to this scenario is using multiview simulcast coding. Unfortunately, it results in high bitrate and pixelrate [12], making this approach not valuable for a practical immersive video system.

Another approach is to extend the method described in the previous section. View selection had to be performed in such a way that would guarantee the highest average quality of the synthesized views independently of the user's viewpoint. For this purpose, Ref. [18] presents the results of a simulation of a simple practical immersive video system with the assumption of a reasonable pixelrate [12] and number of cameras [19]. The results showed that the best quality in the transmission scenario is achieved when cameras are distributed evenly; however, for omnidirectional content, there is a necessity to send input views from the horizontal axis rather than the vertical. The authors proved this through subjective tests. As noted in Section 1, the MIV DSDE profile differs materially from the MIV Main profile. While [18] describes a view selection strategy tailored to MIV Main, we contend that the DSDE scenario requires a different approach; our proposed method is presented in Section 3.

2.3. Impact of Camera Arrangement on Depth Estimation

The methods of view selection described in the previous sections did not consider one crucial process used in the creation of immersive video: depth estimation. When cameras are too far apart, fewer pixels are captured by at least two views, which negatively influences depth estimation, since any scene point must be visible in two or more images to have its depth reliably calculated, whereas occluded regions can only be interpolated or extrapolated. Moreover, larger baselines amplify lighting and reflectance discrepancies across views, complicating depth matching between views and degrading both depth maps and synthesized views.

There are other efficient camera setups [20–24], but these techniques require more input information, e.g., geometry of objects, which cannot be predicted in practical immersive video systems.

Stereopair grouping has been proposed to address these issues without increasing the number of cameras [25]. By organizing cameras into closely spaced pairs, each pair shares nearly identical viewpoints and lighting conditions, minimizing intra-pair occlusions and ensuring most scene points are visible to at least two cameras. However, the short baseline intrinsic to each pair limits depth precision. The solution is a hierarchical approach: Use long-baseline pairs (drawn from different stereo pairs) for precise depth estimation where possible, and rely on dense, short-baseline pairs to fill in occluded or poorly observed regions.

To present experimental results, the authors of [25] decided to use the PSNR metric because of its simplicity [26,27]. In the results, PSNR gains are split into baseline-adjustment and occlusion reduction. While long baselines improve depth, uniform layouts encounter a problem of a large number of occlusions in complex scenes [28,29], which causes horizontal displacements in virtual views. In typical two-step view-synthesis pipelines [30,31], occluded regions are inpainted and have lower quality. Empirical results indicated that when occlusions exceed roughly 20–25% of the scene, stereo-pair arrangements dramatically outperform uniform layouts, striking the optimal balance between depth accuracy and coverage in challenging immersive-video scenarios.

The experiments in [25] were conducted using the DERS [29] depth estimation algorithm, which is now outdated. In the present work, we employ IVDE [32], the current reference software for depth estimation, and perform experiments both on the sequences used in [25] and on new test sequences created specifically for research on immersive video [33].

3. The Proposal—Transmitted View Selection for DSDE

In immersive video systems with decoder-side depth estimation (DSDE), only a subset of available input views can be transmitted within the bitstream due to bitrate and pixelrate constraints [12]. Usually, a uniform selection of views is used [34] for limiting the number of transmitted views. However, such a strategy does not guarantee the highest quality on the decoder side.

The goal of the proposed method is to maximize the quality of synthesized virtual views while preserving the number of transmitted views and the overall bitrate and pixelrate.

The method relies on the analysis of occluded regions across input views. The occlusions are detected by a reprojection of 3D points between views, identifying pixels that are visible in only one view (thus, the pixels for which the depth is not estimable). In the proposed approach, the occlusion detection is implemented using the reprojection module already available in the TMIV v16.0 encoder [35]. At first, a set number of evenly spaced input views is set (in this paper, this number equals four), and each of them is reprojected to the remaining ones. For each pixel, the reprojected depth is compared with the original depth in the target view. A pixel is considered "visible" in the target view if the reprojected depth does not exceed the original depth (i.e., it is not occluded). The percentage of pixels visible in fewer than two views is accumulated. This provides a direct estimation of the occlusions ratio, i.e., the proportion of regions where depth cannot be reliably estimated.

To avoid fluctuations of the selected views within a single group of pictures (GOP)—which would significantly decrease video compression efficiency—the analysis is performed once per GOP (i.e., on the first frame of the GOP) and the resulting view selection remains fixed for the entire GOP. Because the underlying TMIV module is already highly optimized, the computation introduces only a negligible overhead. A detailed analysis of the processing time is provided in Section 5.3.

If a sequence contains a high ratio of occlusions (inestimable areas), the algorithm selects multiple camera stereopairs instead of uniformly arranged single cameras (see Figure 5). As proven in [18], such an approach reduces the number of non-estimated regions, thus increasing the overall quality of synthesized views. On the other hand, when the quantity of occlusions is lower, uniformly distributed views are chosen to maximize coverage and spatial diversity.

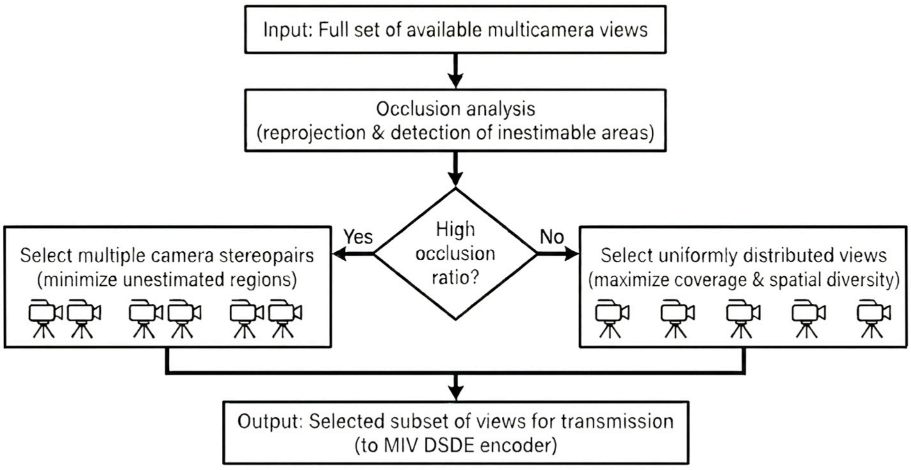
Input: Full set of available multicamera views
Occlusion analysis
(reprojection & detection of inestimable areas)
High
occlusion
ratio?
Yes
No
Select multiple camera stereopairs
(minimize unestimated regions)
Select uniformly distributed views
(maximize coverage & spatial diversity)
Output: Selected subset of views for transmission
(to MIV DSDE encoder)

Figure 5. Scheme of the proposal.

The proposed approach is straightforward to implement within an MIV encoder (e.g., in the Test Model for MPEG immersive video reference software, TMIV [35]) and does not require any changes on the decoder side.

Moreover, by selecting views captured by stereopairs instead of evenly distributed cameras, the method can also reduce the total bitrate of an immersive video, as similar neighboring views packed into a single atlas may be compressed more efficiently (especially when screen-content coding tools are used by a video encoder [34]).

4. Overview of Experiments

To comprehensively evaluate the effectiveness of the proposed view selection method, two complementary experiments were conducted:

1. "MIV experiment"—evaluation using modern immersive video content.
2. "Supplementary experiment"—evaluation using legacy, classical FTV multiview sequences.

Both experiments employed exactly the same processing pipeline (defined in MIV common test conditions, MIV CTC [33]), using the same software:

- TMIV reference software [35] for creating atlases, view selection, and view synthesis on the decoder side.
- VVenC + VvdeC [36] for VVC [9] atlas encoding and decoding.
- IVDE reference software [32] for decoder-side depth estimation.

All the parameter settings except one were aligned with MIV CTC. The only intentional change—applied consistently in both experiments—was restricting the system to transmit only four views in total (all packed into a single atlas). This design choice reflects the goal of evaluating the influence of camera pairing and spacing in the most controlled and interpretable setting. By focusing on a minimal, one-dimensional camera layout (cameras placed along a line or an arc), the experiment isolates the effect of horizontal baselines on video coding, depth estimation, and synthesized view quality, without confounding factors introduced by multidimensional camera rigs or additional transmitted views.

The four-view configuration also represents the simplest non-trivial case in which different pairing strategies meaningfully affect view synthesis. More complex scenarios (e.g., multi-atlas setups or two-dimensional camera arrangements) are natural extensions of this study, but addressing them would require analyzing several factors at once. Our intention was to start with a clean, analyzable scenario, establishing conclusions that can later be generalized to higher-dimensional arrangements and a higher number of transmitted views.

Both experiments differ solely in the test sequences, allowing for assessing whether the proposed methodology is consistent across datasets with different capture characteristics, resolutions, and scene geometries.

4.1. MIV Experiment

The main experiment was conducted on the modern immersive video content from the MIV CTC [33]. The MIV CTC test set contains 21 miscellaneous multiview sequences. However, only six of them satisfy the requirements of this study, i.e., multicamera setups with cameras arranged approximately along a line or along an arc. Such setups are essential for analyzing how different horizontal baselines influence immersive video processing in the DSDE scenario.

To systematically evaluate the influence of camera spacing, a uniformity coefficient U was introduced:

$$U = \frac{b}{d}$$

where b is the baseline of each stereopair, and d is the distance between two stereopairs (i.e., the distance between two middle cameras within the selected four), c.f. Figure 6.

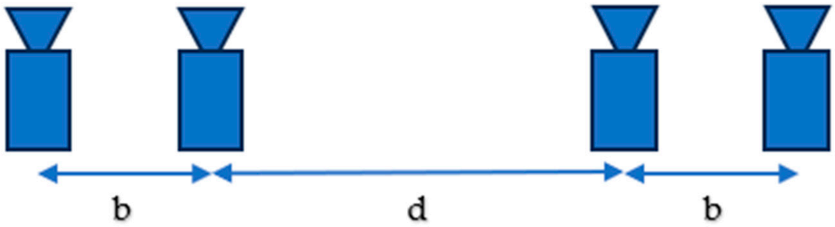

Figure 6. Analyzed camera arrangement; b—stereopair baseline (the same for both stereopairs), d—distance between two stereopairs; uniformity coefficient describes the ratio between b and d.

This parameter quantifies the uniformness of camera placement across the camera setup, where a coefficient equal to 1 corresponds to a perfectly uniform distribution (with equal distances between all neighboring cameras), and smaller values indicate non-uniform layouts with two stereopairs of cameras. A uniformity coefficient greater than 1 corresponds to a situation where the distance between two camera pairs is smaller than the baseline of each stereopair (i.e., the arrangement with a stereopair in the middle and two single cameras at each side).

The evaluation was based on the IV-PSNR metric, which measures the fidelity of the synthesized virtual views compared to reference ones, taking into account typical immersive video characteristics [37]. To assess the rate-distortion performance, BD-IVPSNR values were computed. The authors chose BD-IVPSNR instead of typical BD-rates because the RD-curves for different configurations did not always overlap. Therefore, the BD-IVPSNR metric ensured consistent comparison of the results.

Moreover, for each test sequence, occlusion maps were also generated in order to assess the proportion of scene areas not visible in multiple input views. This information was used to analyze how camera arrangement affects depth estimation and view synthesis quality for different levels of scene complexity.

The obtained results and their interpretation are discussed in Section 5, where the relationship between occlusion percentage, camera arrangement, and the quality of synthesized virtual views is analyzed in detail.

4.2. Supplementary Experiment

To validate the generality of the conclusions beyond modern immersive datasets, a supplementary experiment was performed using several classical FTV sequences: five BigBuckBunny sequences [38] (BBB Butterfly Arc, BBB Flowers Arc, BBB Rabbit Arc, BBB Butterfly Linear, and BBB Rabbit Linear), Bee [39], and three sequences from Nagoya University [40] (Champagne, Dog, and Pantomime). Although these sequences are of lower resolution and are no longer used in standardization activities, they provide diverse content characteristics and dozens of input views, allowing for analysis of multiple uniformity coefficients.

Moreover, the choice of sequences creates the possibility to partially compare the results of this research with the results presented in [25], where the authors analyzed the influence of camera pairing on depth estimation and view synthesis quality, but without the use of any video compression.

Crucially, the same experimental pipeline was used: TMIV + IVDE + VVenC/VVdeC, following MIV CTC guidelines and the same four-view, single-atlas constraint. This allowed a direct comparison of trends observed across datasets.

5. Experimental Results

5.1. MIV Experiment

Figure 7 presents a quality increase caused by camera pairing in comparison with uniform view distribution. Gain in quality is clearly visible for CBABasketball and MartialArts sequences, as well as the Frog and Fencing sequences, indicating that the camera pairing approach provides a noticeable advantage in scenes with complex geometry and frequent occlusions. The percentage of occlusions in each test sequence is presented in Table 1.

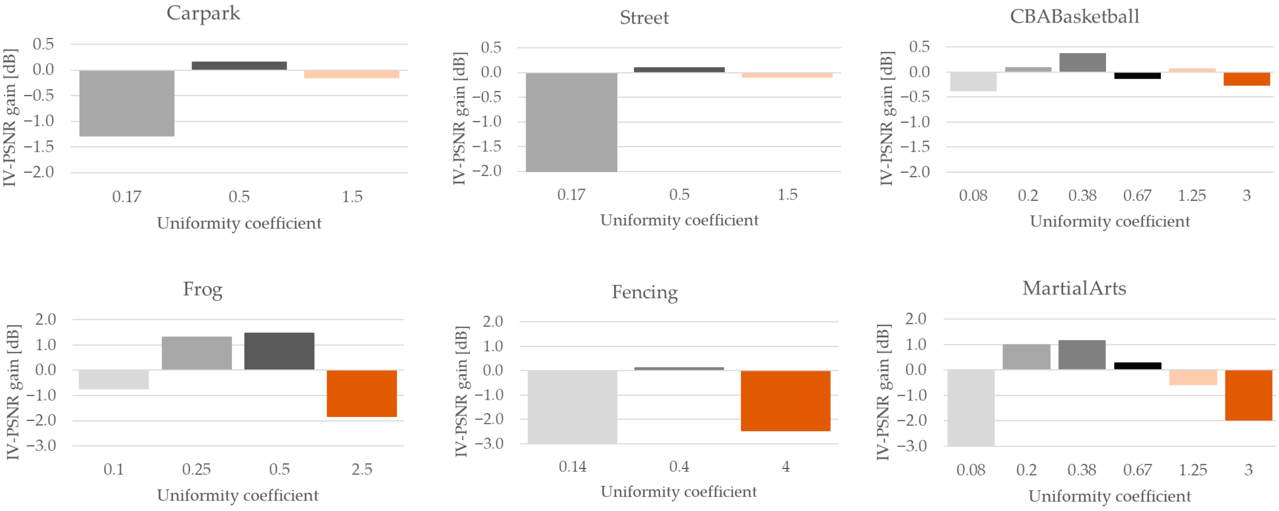

Figure 7. IV-PSNR of virtual views synthesized based on uncompressed atlases—quality increase compared to uniform view distribution (uniformity coefficient = 1); grey bars: two pairs of cameras (uniformity coefficient < 1), red bars: a pair of cameras in the middle and two single cameras at each side (uniformity coefficient > 1).

Table 1. Percentage of occluded areas in test sequences.

Sequence	Percentage of Occluded Areas
Frog	13.47%
Carpark	6.13%
Street	2.51%
Fencing	11.54%
CBABasketball	11.29%
MartialArts	21.47%

The existence of the relationship between the arrangement of selected real views and the amount of occlusions in the scene is further confirmed by the rate–distortion curves presented in Figure 8. For the sequences with more occlusions, the proposed view selection strategy consistently achieves higher IV-PSNR for a given bitrate.

Because the RD-curves for different configurations did not overlap, BD-rates could not be computed; therefore, in order to assess the average quality improvement across bitrates, the BD-IVPSNR values were used (Figure 9). Figure 9 summarizes the BD-IVPSNR gains as a function of the uniformity coefficient. The results confirm that pairing of the cameras (with a reasonable baseline—uniformity coefficient in the range [0.2, 0.7]) leads to a measurable quality increase for most of the tested sequences, especially for those with more occlusions.

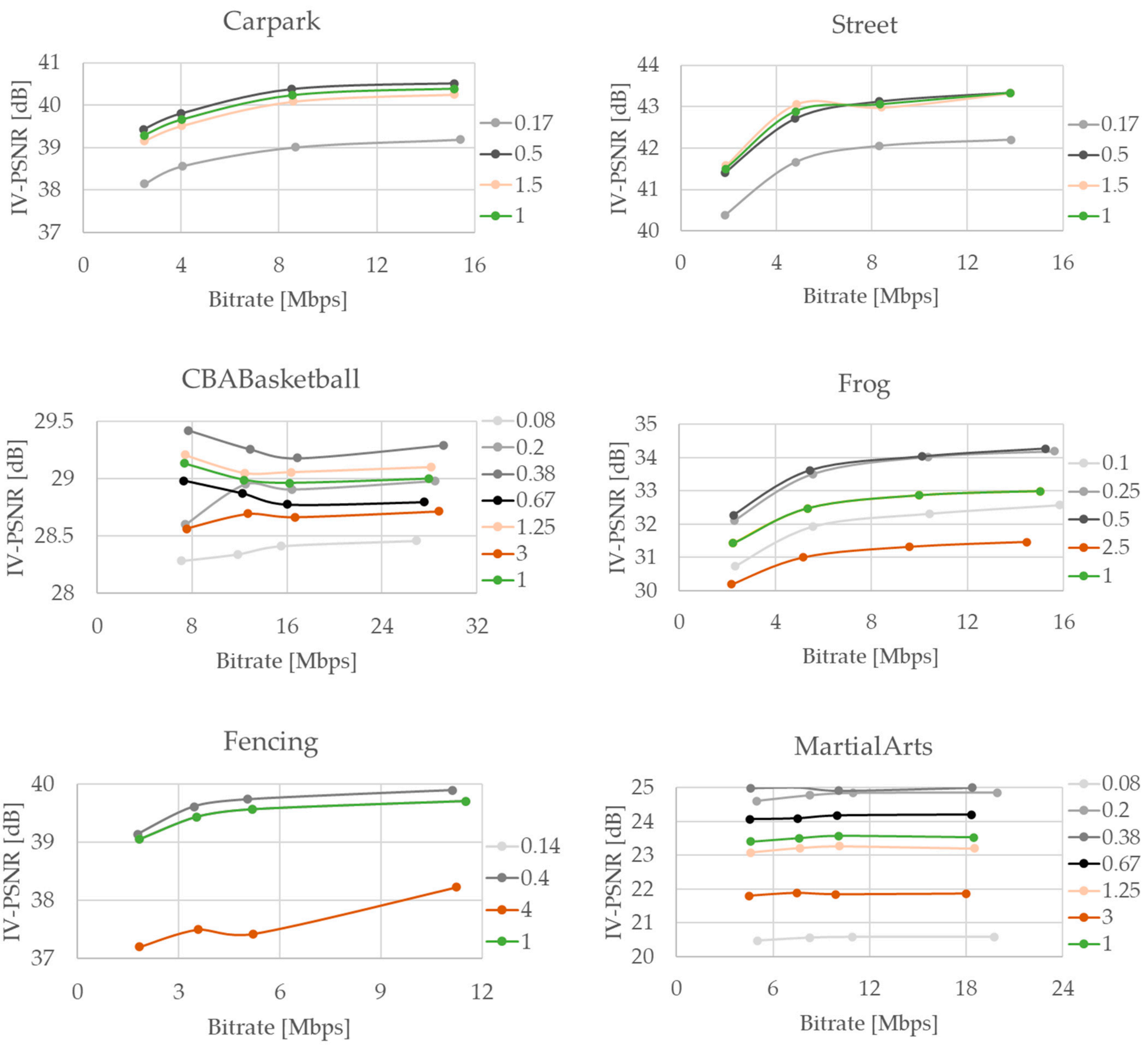

Figure 8. RD-curves of the immersive video system compared to uniform view distribution (uniformity coefficient = 1, green line); grey lines: two pairs of cameras (uniformity coefficient < 1), red lines: a pair of cameras in the middle and two single cameras at each side (uniformity coefficient > 1).

For scenes with very limited occlusions (e.g., Carpark, Street), both the IV-PSNR gains and BD-IVPSNR gains caused by camera pairing remain negligible. In such scenes, all the cameras observe nearly the same content, and an area with non-estimable depth is small for all camera arrangements. As a result, camera pairing does not introduce additional geometric cues that would noticeably improve depth estimation or view synthesis. On the other hand, uniform camera distribution—as described in Section 2.1 and [13]—minimizes problems with finite resolution of depth maps and existence of non-Lambertian reflections in the scene. Therefore, for scenes with limited occlusions, the proposed view selection method purposely selects views distributed uniformly.

The results demonstrate that the proposed view selection method brings clear benefits when the proportion of occluded regions exceeds approximately 10% (c.f., Table 1).

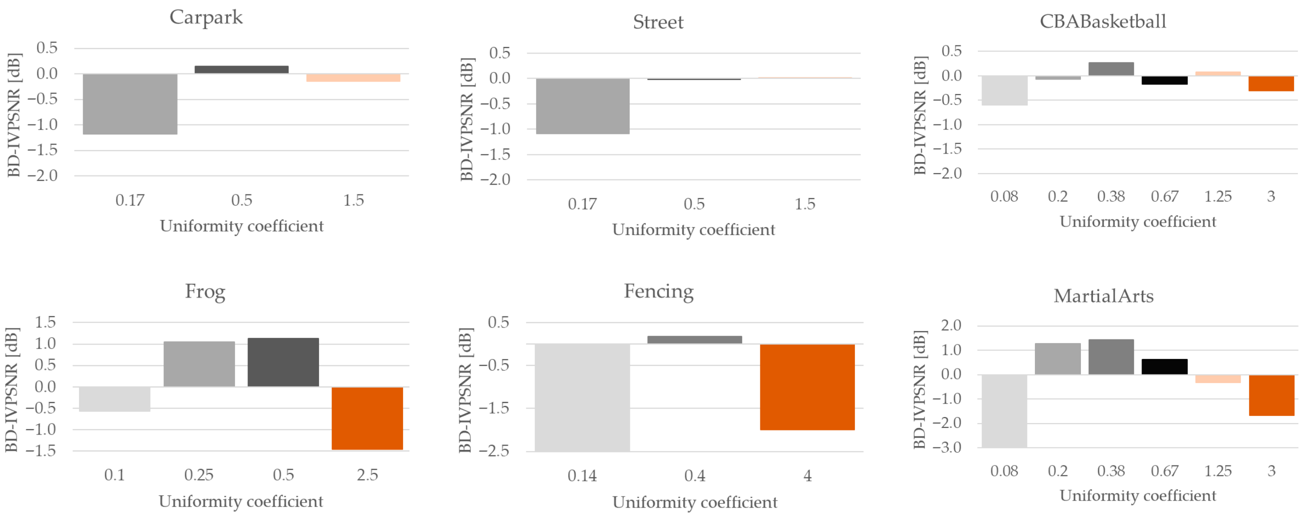

Figure 9. BD-IVPSNR of the immersive video system compared to uniform view distribution (uniformity coefficient = 1); grey bars: two pairs of cameras (uniformity coefficient < 1), red bars: a pair of cameras in the middle and two single cameras at each side (uniformity coefficient > 1).

In such cases, pair-wise camera grouping improves the reconstruction of disoccluded areas and reduces geometric distortions in synthesized views. On the other hand, for sequences with minimal occlusions (e.g., Carpark, Street), a uniform camera distribution remains sufficient, and camera pairing provides no significant quality gains (Table 2 and Figures 10 and 11).

Table 2. Gain of camera pairing (compared to the uniform camera arrangement).

Sequence	IV-PSNR Gain	BD-IVPSNR Gain
Frog	1.47 dB	1.14 dB
Carpark	0.13 dB	0.11 dB
Street	0.07 dB	−0.03 dB
Fencing	0.14 dB	0.20 dB
CBABasketball	0.37 dB	0.27 dB
MartialArts	1.15 dB	1.44 dB

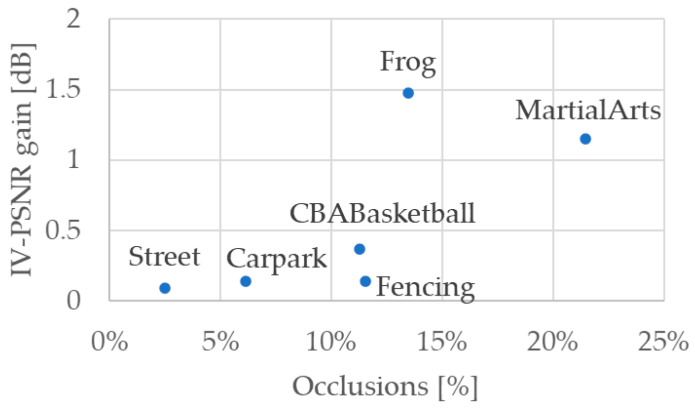

Figure 10. IV-PSNR gain from camera pairing (compared to the uniform camera arrangement) as a function of occlusions.

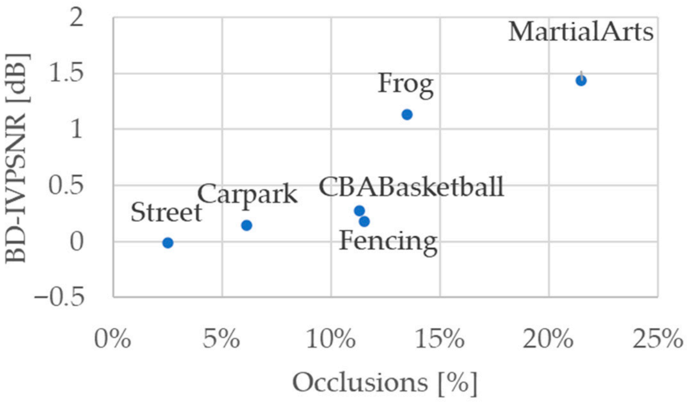

Figure 11. BD-IVPSNR gain from camera pairing (compared to the uniform camera arrangement) as a function of occlusions.

Overall, the experimental results confirm that the spatial arrangement of cameras has a measurable effect on the final quality of content presented to the user of an immersive video system. As presented, the pairwise view selection optimally balances occlusion handling and the precision of depth estimation, making it particularly suitable for complex scenes where occluded areas are a significant part of the visible content.

5.2. Supplementary Experiment

Figure 12 presents a quality increase caused by camera pairing in comparison with uniform view distribution. All sequences used in the supplementary experiment contain enough input views to provide results for ten different uniformity coefficients (including a coefficient equal to 1, representing uniform camera arrangement). Among all test sequences, there are three for which camera pairing always introduces quality loss: BBB Rabbit Arc, BBB Butterfly Arc, and Dog. As presented in Table 3, these three sequences are characterized by the lowest percentage of occlusions. For the remaining sequences, in which the occlusion ratio is higher, camera pairing with a sufficiently large baseline (uniformity coefficient > 0.5) provides gains in terms of the quality of synthesized views.

Table 3. Percentage of occluded areas in test sequences.

Sequence	Percentage of Occluded Areas
BBB Rabbit Arc	4.12%
BBB Butterfly Arc	8.71%
Dog	9.13%
BBB Rabbit Linear	15.30%
Pantomime	16.13%
BBB Butterfly Linear	16.22%
ChampagneTower	34.72%
Bee	37.11%
BBB Flowers Arc	39.34%

To more clearly illustrate the relationship between occlusions (Table 3) and the effectiveness of camera pairing (Figure 12), Figure 13 presents a scatterplot combining occlusion ratio and IV-PSNR gain. For all sequences, the results obtained for the arrangement with a uniformity coefficient of 0.7 are shown. The results confirm the outcomes from the MIV experiment: camera pairing yields consistent quality benefits when the proportion of occluded regions exceeds approximately 10%. In opposite cases, a uniform camera arrangement outperforms the pair-wise camera layout in terms of virtual view quality.

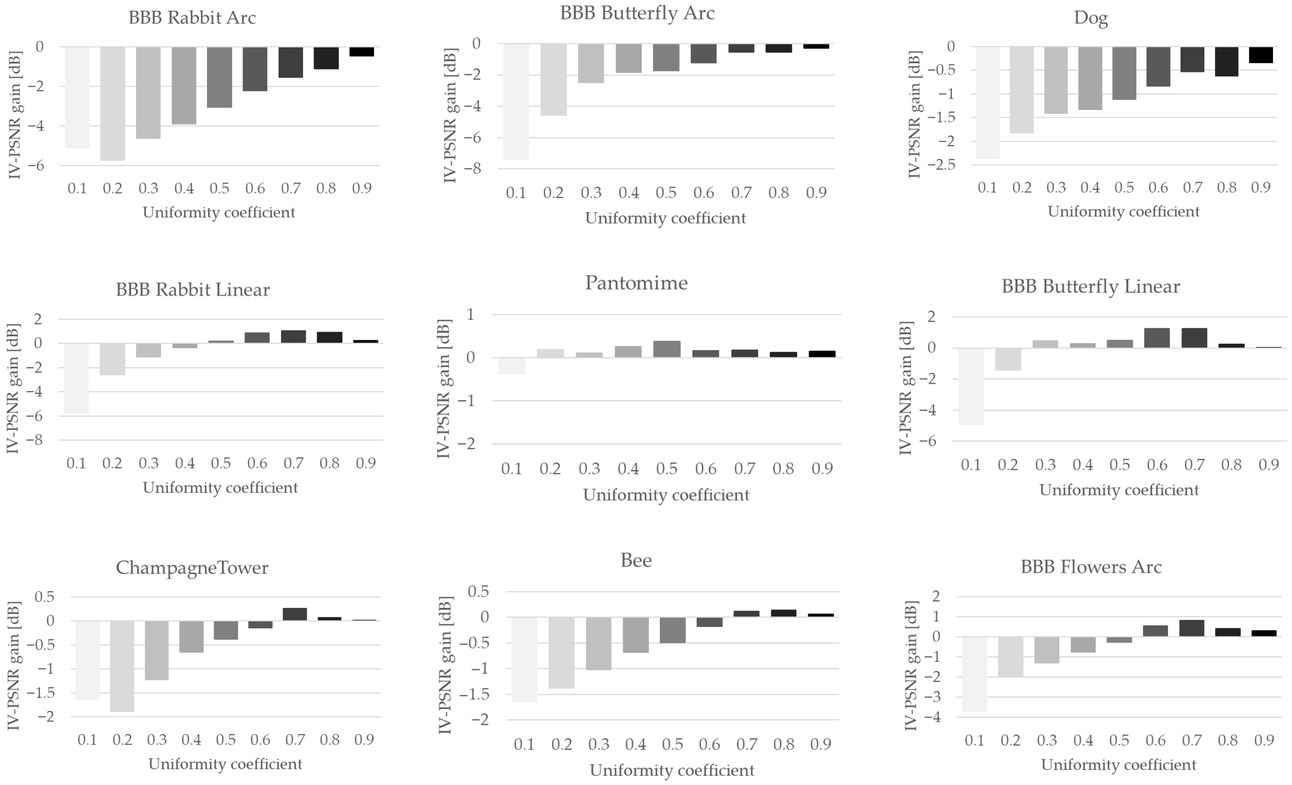

Figure 12. IV-PSNR of virtual views synthesized based on uncompressed atlases—quality increase compared to uniform view distribution (uniformity coefficient = 1); sequences sorted based on ascending ratio of occluded areas (Table 3).

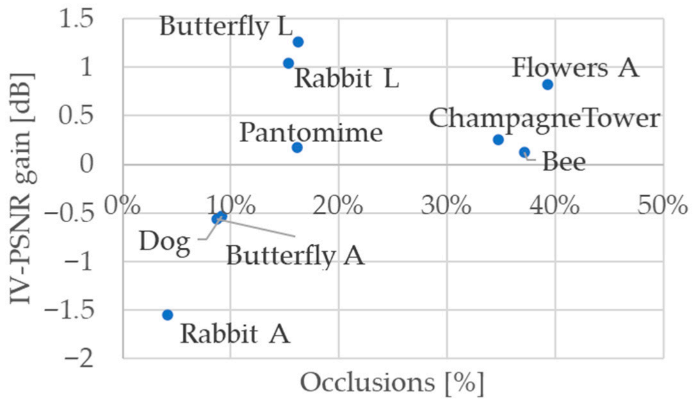

Figure 13. IV-PSNR gain from camera pairing (compared to the uniform camera arrangement) as a function of occlusions.

An additional observation concerns the three sequences with the highest occlusion ratios (above 30%), for which the quality gains from camera pairing are present but smaller than could be expected given the strong advantage of pair-wise camera setups. This behavior is consistent with the limitations of current immersive video pipelines—when the scene becomes overly complicated (the proportion of occluded areas becomes very large), the accuracy of depth estimation and view synthesis is constrained by the lack of information. In such cases, camera pairing cannot fully compensate for the severe geometric ambiguity, and the quality gains are restricted by incomplete scene information. Nevertheless, even in such challenging scenarios, the pair-wise camera layout still outperforms the uniform arrangement, proving that the proposed approach is beneficial for all tested content difficulty levels.

It is important to highlight that the 10% threshold—however consistent in both presented experiments—is smaller than a similar threshold reported in [25], where it exceeded 20%. The difference in the occlusion threshold used in this article, when compared with research conducted in [25], originates from the evolution of processing tools used for depth estimation and view synthesis. The study conducted in [25] relied on the DERS [29] algorithm for depth estimation and VSRS [41] software for virtual view synthesis. For the purpose of this article, the authors used IVDE v7.0 [32] depth estimation software and TMIV's view weighting synthesizer (VWS) [35] for virtual view synthesis. Both IVDE and VWS were developed as successors to DERS and VSRS, and they contain improvements such as inter-view and temporal consistency (IVDE) and the ability to synthesize views based on more than two input views (VWS). As a result, the newer software provides more robust depth estimation, improved handling of challenging or weakly textured regions, and significantly higher-quality virtual view synthesis. These improvements reduce the sensitivity of the system to missing information and therefore shift the effective threshold at which stereo-pair grouping becomes advantageous.

In order to present the evolution of immersive video processing pipeline efficiency, we have estimated PSNR values for all tested sequences and uniformity coefficients, keeping the same methodology as in [25] (PSNR for luma component only, averaged over all virtual views). Obtained results are presented in Table 4.

Table 4. Average PSNR of luma component for different camera arrangements.

Sequence	Uniformity Coefficient									
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
	PSNR of Luma Component [dB]									
BBB Rabbit Arc	26.73	27.26	28.39	29.27	30.17	31.00	31.60	32.06	32.41	32.65
BBB Butterfly Arc	27.37	29.85	31.87	32.56	32.85	33.18	33.85	33.83	33.93	34.27
Dog	22.28	22.74	23.17	23.36	23.83	24.12	24.48	24.61	24.72	25.30
BBB Rabbit Linear	23.19	24.92	26.07	26.97	27.55	28.11	28.47	28.51	28.31	28.22
Pantomime	24.07	24.57	24.69	24.96	25.72	25.93	26.04	26.32	26.27	26.47
BBB Butterfly Linear	26.29	29.11	30.65	31.25	31.67	32.86	33.17	32.46	32.17	32.25
ChampagneTower	17.70	17.73	18.47	19.01	19.04	19.20	19.37	19.66	19.48	19.44
Bee	16.30	16.63	16.92	17.17	17.31	17.51	17.62	17.70	17.57	17.58
BBB Flowers Arc	20.45	22.14	22.85	23.41	23.85	24.54	24.77	24.27	24.10	23.86

Table 5 contains differences between results gathered in Table 4 and the results reported in [25]. As presented, for most of the content, the quality is significantly higher when using the modern immersive video processing pipeline (TMIV + IVDE) than with the use of legacy depth estimation and view synthesis software (DERS + VSRS). The only exceptions are the Bee and Dog sequences (and Pantomime for several uniformity coefficients), where the multiview-based synthesis used in VWS performed worse than the simple two-view synthesis used in the VSRS algorithm.

Table 5. Difference in PSNR of the luma component between current research and research presented in [25]; positive value means that views synthesized using the current TMIV pipeline have better quality than those reported in [25].

Sequence	**Uniformity Coefficient**									
	0.1	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1**
	PSNR of Luma Component [dB]									
BBB Rabbit Arc	−0.08	0.48	1.46	2.17	2.73	3.26	3.70	3.97	4.13	4.22
BBB Butterfly Arc	0.68	2.33	1.75	1.46	1.90	1.98	2.30	1.89	2.10	2.12
Dog	−2.16	−1.84	−1.84	−2.12	−1.69	−1.55	−1.64	−0.99	−0.65	−0.47
BBB Rabbit Linear	1.64	2.26	2.78	3.51	3.94	4.36	4.16	4.94	4.39	4.57
Pantomime	4.22	2.44	0.52	−1.07	−0.81	−0.42	−0.54	−0.09	−0.52	−0.27
BBB Butterfly Linear	3.78	3.4	2.97	3.33	3.31	4.05	3.89	3.06	2.69	2.94
ChampagneTower	0.27	−1.18	−0.38	−0.20	0.33	0.18	0.97	1.86	1.34	1.79
Bee	−2.86	−4.05	−4.03	−3.98	−3.83	−3.50	−3.20	−2.97	−2.86	−2.46
BBB Flowers Arc	0.42	1.11	1.52	2.18	2.90	4.21	4.22	4.81	4.77	4.75
Average	0.66	0.55	0.53	0.59	0.98	1.40	1.54	1.83	1.71	1.91

Taken together, the results in Tables 4 and 5, and Table 2 from [25] illustrate how the substantial progress in depth estimation and view synthesis achieved over the past decade fundamentally changed the system's sensitivity to occluded content, fully justifying the lower empirical occupancy threshold observed in this work.

5.3. Computational Overhead

The coding pipeline of the proposal follows the standard TMIV workflow, with all modules and settings unchanged, except for the addition of the occlusion analysis step. The proposed occlusion analysis is executed using the effective and fast reprojection module already implemented within the TMIV v16.0 software [35]. To compute the occlusion ratio, four evenly spaced input views are selected, and each view is reprojected onto the remaining ones to determine the proportion of pixels that are visible by fewer than two cameras (i.e., for which the depth is not estimable).

To maintain encoding efficiency, the occlusion analysis is performed once per GOP (only for the first frame of GOP), and the resulting view selection persists for the entire GOP. Such an approach additionally decreases the computational overhead introduced by the proposed approach.

As presented in Table 6, the time required for performing the introduced occlusion analysis is negligible when compared to the total TMIV encoding time. In both experiments, the proposal increased the computational time by less than 0.5%. Moreover, the proposal does not change the decoding time, which is crucial in any practical video system.

Table 6. Computational time of the proposed approach compared to the unmodified TMIV encoder in DSDE configuration; results averaged over all sequences.

	MIV Experiment	**Supplementary Experiment**
Original encoding time (per GOP)	58.25 s	27.13 s
Occlusion analysis time (per GOP)	0.255 s	0.117 s
Computational time increase	0.44%	0.43%

https://doi.org/10.3390/app16010072

The presented results demonstrate that the proposed method is computationally efficient and practical for real-world immersive video encoding, providing an efficient and reliable view selection with minimal impact on processing time.

6. Conclusions

In this paper, we propose an adaptive view selection method for the MIV DSDE profile that dynamically switches between uniform camera placement and grouping them into stereo pairs. This decision is made based on an analysis of the occlusion level in the scene, performed at the encoder.

Our experimental results, obtained within the TMIV reference software, quantitatively validate this approach. We demonstrate a clear correlation between the percentage of occluded areas and the optimal camera layout. A key finding is the identification of a decision threshold: for scenes with occlusion levels exceeding approximately 10%, the stereo pair grouping strategy yields a significant and measurable gain in quality (up to 2 dB BD-IVPSNR). Below this threshold, the traditional uniform layout is sufficient or even more effective; thus, the uniform layout is automatically chosen by the proposed method.

However, it should be emphasized that this threshold is not a universal constant, but rather a consequence of the performance of the depth estimation, view synthesis, and video encoding algorithms used in an immersive video pipeline. In the modern MIV-based system, where TMIV with its efficient view weighting synthesizer [35], together with IVDE [32] for decoder-side depth estimation, is used, the threshold is equal to approximately 10%. Historically, when legacy state-of-the-art depth estimation and view synthesis were used (DERS [29] and VSRS [41], respectively), the effective threshold was significantly higher, reaching 20–25% for comparable multiview setups [25]. This evolution reflects the continuous improvement of geometric reconstruction of 3D scenes using modern immersive video processing techniques. Therefore, the empirical threshold reported in this paper should be interpreted as typical for modern MIV DSDE pipelines.

Overall, the presented results confirm that the spatial arrangement of transmitted views has a measurable impact on the quality of video watched by a viewer in a DSDE immersive video system. The proposed adaptive method effectively balances occlusion handling and geometric accuracy, making it suitable for practical use.

For future work, while this study confirmed the benefit of switching to stereo pairs, further investigation could focus on automating the selection of the optimal baseline (represented by the "uniformity coefficient") for those pairs, potentially adapting it dynamically based on scene geometry. Furthermore, exploring alternative or combined scene analysis metrics beyond a simple occlusion percentage could lead to an even more robust and fine-grained decision model for view selection.

Author Contributions: Conceptualization, D.K. and A.D.; methodology, D.K. and A.G.; software, D.K., A.G., and D.M.; validation, D.K. and A.D.; writing, D.K., A.D., and D.M. All authors have read and agreed to the published version of the manuscript.

Funding: The research was supported by the Ministry of Science and Higher Education of the Republic of Poland.

Data Availability Statement: The experiments described in Section 4 were conducted using the TMIV v16.0, https://gitlab.com/mpeg-i-visual/tmiv, accessed on 11 August 2025, and IVDE v7.0, https://gitlab.com/mpeg-i-visual/ivde, accessed on 11 August 2025, using MPEG test sequences [33,38–40].

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Wien, M.; Boyce, J.M.; Stockhammer, T.; Peng, W.-H. Standardization Status of Immersive Video Coding. *IEEE J. Emerg. Sel. Top. Circuits Syst.* **2019**, *9*, 5–17. [CrossRef]
2. *ISO/IEC JTC1/SC29/WG11MPEG/N18145*; MPEG Call for Proposals on 3DoF+Visual. ISO: Brussels, Belgium, 2020.
3. *ISO/IECJTC1/SC29/WG11 MPEG N17073*; Requirements on 6DoF (v1). ISO: Torino, Italy, 2017.
4. Xiang, S.; Yu, L.; Liu, Q.; Xiong, Z. A gradient-based approach for interference cancelation in systems with multiple Kinect cameras. In Proceedings of the 2013 IEEE International Symposium on Circuits and Systems (ISCAS), Beijing, China, 19–23 May 2013; pp. 13–16.
5. Stankiewicz, O.; Domanski, M.; Dziembowski, A.; Grzelka, A.; Mieloch, D.; Samelak, J. A free-viewpoint television system for horizontal virtual navigation. *IEEE Trans. Multimed.* **2018**, *20*, 2182–2195. [CrossRef]
6. Fachada, S.; Bonatto, D.; Schenkel, A.; Lafruit, G. Depth image based view synthesis with multiple reference views for virtual reality. In Proceedings of the 3DTV Conference 2018, Stockholm, Sweden, Helsinki, Finland, 3–5 June 2018.
7. Muller, K.; Merkle, P.; Wiegand, T. 3-D Video Representation Using Depth Maps. *Proc. IEEE* **2010**, *99*, 643–656. [CrossRef]
8. Sullivan, G.; Ohm, J.-R.; Han, W.-J.; Wiegand, T. Overview of the High Efficiency Video Coding (HEVC) standard. *IEEE Trans. Circuits Syst. Video Technol.* **2012**, *22*, 1649–1668. [CrossRef]
9. Bross, B.; Wang, Y.-K.; Ye, Y.; Liu, S.; Chen, J.; Sullivan, G.J.; Ohm, J.-R. Overview of the Versatile Video Coding (VVC) standard and its applications. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *31*, 3736–3764. [CrossRef]
10. Tech, G.; Chen, Y.; Müller, K.; Ohm, J.-R.; Vetro, A.; Wang, Y.-K. Overview of the Multiview and 3D Extensions of High Efficiency Video Coding. *IEEE Trans. Circuits Syst. Video Technol.* **2016**, *26*, 35–49. [CrossRef]
11. *ISO/IEC JTC1/SC29/WG4 MPEG2023/N0341*; Verification Test Report of MPEG Immersive Video. ISO: Antalya, Turkey, 2023.
12. Boyce, J.; Dore, R.; Dziembowski, A.; Fleureau, J.; Jung, J.; Kroon, B.; Salahieh, B.; Vadakital, V.K.M.; Yu, L. MPEG Immersive Video coding standard. *Proc. IEEE* **2021**, *109*, 1521–1536. [CrossRef]
13. Dziembowski, A.; Samelak, J.; Domanski, M. View selection for virtual view synthesis in free navigation systems. In Proceedings of the International Conference on Signals and Electronic Systems, ICSES, Kraków, Poland, 10–12 September 2018.
14. Dziembowski, A.; Grzelka, A.; Mieloch, D.; Stankiewicz, O.; Wegner, K.; Domanski, M. Multiview Synthesis–improved view synthesis for virtual navigation. In Proceedings of the 32nd Picture Coding Symposium (PCS), Nuremberg, Germany, 4–7 December 2016.
15. Ceulemans, B.; Lu, S.-P.; Lafruit, G.; Munteanu, A. Robust multiview synthesis for wide-baseline camera arrays. *IEEE Trans. Multimed.* **2018**, *20*, 2235–2248. [CrossRef]
16. Li, S.; Zhu, C.; Sun, M.-T. Hole filling with multiple reference views in DIBR view synthesis. *IEEE Trans. Multimed.* **2018**, *20*, 1948–1959. [CrossRef]
17. *ISO/IEC JTC1/SC29/WG11 MPEG2013/M40657*; View Synthesis Reference Software (VSRS) 4.2 with Improved Inpainting and Hole Filling. ISO: Geneva, Switzerland, 2017; pp. 3–7.
18. Klóska, D.; Dziembowski, A.; Samelak, J. Versatile input view selection for efficient immersive video transmission. In Proceedings of the 31st International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision, WSCG, Prague, Czechia, 15 May 2023.
19. *ISO/IEC JTC1/SC29/WG11 MPEG2018/M43748*; Kermit Test Sequence for Windowed 6DoF Activities. ISO: Geneva, Switzerland, 2018.
20. Rahimian, P.; Kearney, J.K. Optimal camera placement for motion capture systems. *IEEE Trans. Vis. Comput. Graph.* **2017**, *23*, 1209–1221. [CrossRef] [PubMed]
21. Chen, X.; Davis, J. An occlusion metric for selecting robust camera configurations. *Mach. Vis. Appl.* **2008**, *19*, 217–222. [CrossRef]
22. Olague, G.; Mohr, R. Optimal camera placement for accurate reconstruction. *Pattern Recognit.* **2002**, *35*, 927–944. [CrossRef]
23. Qian, N.; Lo, C.-Y. Optimizing camera positions for multiview 3D reconstruction. In Proceedings of the International Conference on 3D Imaging (IC3D), Liege, Belgium, 14–15 December 2015; pp. 1–8.
24. Gargallo, P.; Prados, E.; Sturm, P. Minimizing the reprojection error in surface reconstruction from images. In Proceedings of the IEEE 11th International Conference on Computer Vision, Rio de Janeiro, Brazil, 14–21 October 2007.
25. Domański, M.; Dziembowski, A.; Grzelka, A.; Mieloch, D. Optimization of camera positions for free-navigation applications. In Proceedings of the International Conference on Signals and Electronic Systems, ICSES 2016, Kraków, Poland, 5–7 September 2016.
26. Shao, F.; Lin, W.; Jiang, G.; Yu, M.; Dai, Q. Depth map coding for view synthesis based on distortion analyses. *IEEE J. Emerg. Sel. Top. Circuits Syst.* **2014**, *4*, 106–117. [CrossRef]
27. Wang, R.; Luo, J.; Jiang, X.; Wang, Z.; Wang, W.; Li, G.; Gao, W. Accelerating image-domain-warping virtual view synthesis on GPGPU. *IEEE Trans. Multimed.* **2017**, *19*, 1392–1400. [CrossRef]
28. Tanimoto, M.; Panahpour, M.; Fujii, T.; Yendo, T. FTV for 3-D spatial communication. *Proc. IEEE* **2012**, *100*, 905–917. [CrossRef]
29. *ISO/IEC JTC1/SC29/WG11 MPEG/M31518*; Enhanced Depth Estimation Reference Software (DERS) for Freeviewpoint Television. ISO: Geneva, Switzerland, 2013.

30. *ISO/IEC JTC1/SC29/WG11 MPEG/M37232*; Depth Based View Blending in View Synthesis Reference Software (VSRS). ISO: Geneva, Switzerlnad, 2015.
31. Sun, Z.; Jung, C. Real-time depth-image-based rendering on GPU. In Proceedings of the International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery, Xi'an, China, 17–19 September 2015; pp. 324–328.
32. *ISO/IEC JTC1/SC29/WG04 MPEG VC/N0058*; Manual of Immersive Video Depth Estimation. ISO: Geneva, Switzerland, 2021.
33. *ISO/IEC JTC1/SC29/WG04 MPEG VC/N0539*; Common Test Conditions for MPEG Immersive Video. ISO: Geneva, Switzerland, 2024.
34. Samelak, J.; Dziembowski, A.; Mieloch, D. Advanced HEVC Screen Content Coding for MPEG Immersive Video. *Electronics* **2022**, *11*, 23. [CrossRef]
35. *ISO/IEC JTC1/SC29/WG4 MPEG2021/N0084*; Test Model 9 for MPEG Immersive Video. ISO: Geneva, Switzerland, 2021.
36. Wieckowski, A.; Brandenburg, J.; Hinz, T.; Bartnik, C.; George, V.; Hege, G.; Helmrich, C.; Henkel, A.; Lehmann, C.; Stoffers, C.; et al. VVenC: An Open and Optimized VVC Encoder Implementation. In Proceedings of the IEEE International Conference on Multimedia Expo Workshops (ICMEW), Online, 5–9 July 2021.
37. Stankowski, J.; Dziembowski, A. Version [7.1]—[IV-PSNR: Software for immersive video objective quality evaluation]. *SoftwareX* **2024**, *28*, 101961. [CrossRef]
38. *ISO/IEC JTC1/SC29/WG11 M35721*; [FTV AHG] Big Buck Bunny Light-Field Test Sequences. ISO: Geneva, Switzerland, 2015.
39. *MPEG2013/M32995*; FTV AHG: EE1 and EE2 Results with Bee by NICT Document IEC JTC1/SC29/WG11. ISO: Geneva, Switzerland, 2014.
40. *ISO/IEC JTC1/SC29/WG11 M15378*; 1D Parallel Test Sequences for MPEG-FTV. ISO: Archamps, France, 2008.
41. *ISO/IEC JTC 1/SC 29/WG 11 M31520*; Enhanced View Synthesis Reference Software (VSRS) for Free-Viewpoint Television. ISO: Geneva, Switzerland, 2013.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

https://doi.org/10.3390/app16010072