AI-driven Enhancement of Depth Map Consistency

Jakub Kit[a], Dominika Klóska[b], Dawid Mieloch[c], and Adrian Dziembowski[d]
Institute of Multimedia Telecommunications, Poznan University of Technology, Poznań, Poland
{jakub.kit, dominika.kloska, dawid.mieloch, adrian.dziembowski}@put.poznan.pl

Keywords: Immersive Video, Depth Map Estimation, Virtual Reality

Abstract: Virtual Reality (VR) and Free-Viewpoint Television (FTV) systems rely on accurate depth maps for high-quality virtual view synthesis. However, standard depth estimation methods often suffer from temporal inconsistencies, particularly around moving objects, leading to flickering and visual artifacts. This paper proposes a novel, three-step process to enhance the consistency of depth maps. The approach involves extracting a static background using temporal median filtering, followed by AI-driven segmentation and tracking of moving objects using Detectron2, and finally, fusing independently estimated depth layers. An experimental evaluation of Common Test Conditions (CTC) for MPEG Immersive Video (MIV) sequences demonstrates that the proposed method significantly improves temporal stability and visual quality. While objective quality gains (IV-PSNR) are modest, the method achieves substantial coding efficiency improvements, with BD-rate savings of up to 15.6% compared to the anchor. Notably, the superior quality of the generated depth maps led to their adoption by the ISO/IEC MPEG group as the new reference for the Fencing sequence.

1. INTRODUCTION

Free-Viewpoint Television (FTV) (Tanimoto, 2012) and immersive video systems allow users to freely navigate within a three-dimensional scene captured by multiple cameras (grey cameras in Figure 1). Virtual navigation relies on depth maps and camera parameters to reconstruct intermediate viewpoints and synthesize virtual views (orange camera in Figure 1), (Stankowski and Dziembowski, 2022), (Yang et al., 2011). Depth maps are essential for accurate 3D reconstruction and providing a fully immersive navigation experience.

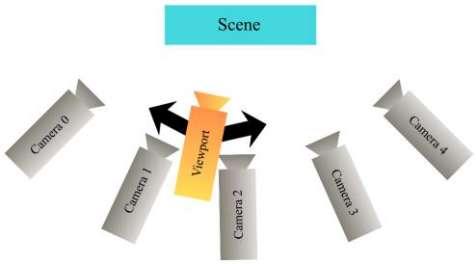

Figure 1: Idea of an immersive video system.

Depth estimation based on inter-view matching often introduces errors, such as artifacts or stretching around moving objects (Stankiewicz et al., 2013), (Mieloch et al., 2022). These inconsistencies result in temporal flickering and deformations in synthesized views, particularly affecting dynamic elements, as illustrated in Figure 2.

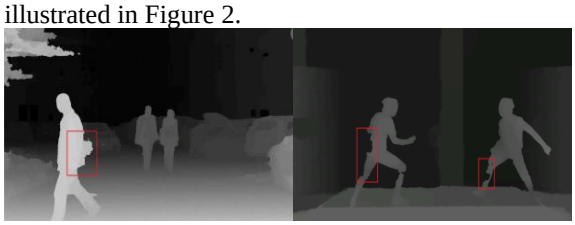

Figure 2: Typical depth artifacts on dynamic objects.

Additionally, compression of depth maps for natural scenes can produce blocking artifacts or blurring near object boundaries. Real-time immersive systems further face computational and bandwidth limitations, making efficient and accurate depth estimation essential for high-quality virtual view synthesis.

[a] https://orcid.org/0009-0005-0906-980X
[b] https://orcid.org/0000-0002-6867-3821
[c] https://orcid.org/0000-0003-0709-812X
[d] https://orcid.org/0000-0001-7426-3362

Recent foundation models like Depth Anything v2 (Yang et al., 2024) offer impressive zero-shot depth estimation and high boundary precision. However, these general-purpose frameworks prioritize per-pixel accuracy over temporal redundancy removal, which is critical for high-efficiency video coding. Therefore, their use can be not very efficient in systems which assume that video data will be compressed.

There are methods increasing spatial and temporal consistency of depth maps described in literature, for instance works based on input views denoising (Stankiewicz et al., 2015), modifications of the GraphCut algorithm (Mieloch and Grzelka, 2018), or depth voting strategies (Mieloch et al., 2021). While these methods provoide noticeable consistency improvements, they do not resolve all the problems, particularly for highly dynamic scenes, with fast-moving objects and frequent occlusions.

Since the solutions described above did not address all the issues with the temporal consistency of depth maps, the authors proposed a three-step depth map estimation method consisting of: static background extraction, AI-based moving object segmentation and tracking, and depth fusion. The proposed method is detailed in Section 2. Section 3 describes the implementation process and the tools employed, Section 4 is dedicated to the experimental results, and Section 5 provides conclusions and directions for future work.

2. PROPOSED AI-DRIVEN SEGMENTATION-BASED TECHNIQUE

2.1 Static background extraction

The first step of the proposed method utilizes the assumption that most objects in the scene remain relatively static or change position negligibly throughout the sequence. Although these elements can include objects both in any part of the scene (close to or far from camera), to simplify further considerations, we define these parts as a static background.

The authors decided to determine the static background for a given video sequence by calculating the median of all frames over time. The median was chosen over the arithmetic mean due to its robustness against artifacts caused by slow-moving objects. The mathematical representation of this process for each pixel (x, y) is as follows:

$$B(x,y) = \text{median}\{F_1(x,y), F_2(x,y), \ldots, F_N(x,y)\} \quad (1)$$

where $B(x,y)$ represents the resulting background pixel, $F_i(x,y)$ denotes the pixel value in the i-th frame, and N is the total number of frames in the video. An example of the resulting background is presented in Figure 3.

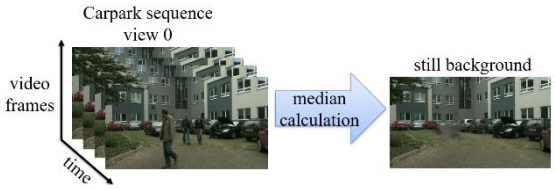

Figure 3: Still background calculation by median over time.

The static background extraction process is repeated for each view of the sequence. Subsequently, depth estimation is performed on the extracted background frames, as illustrated in Figure 4.

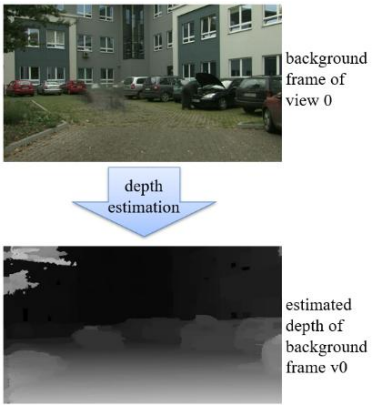

Figure 4: Background frame depth estimation.

During the calculation of the median over time, some cases can occur in which moving objects changed their position so slightly that the resulting background frame contained pixels belonging to the moving object. Consequently, depth artifacts can appear in the background frame (areas where the calculated depth was not accurate). To address this issue, a modification was introduced to the depth map estimation software. After calculating background depth, the whole frame is analyzed; if the algorithm detects elements with depth values significantly higher than their surroundings, they are identified as remnants of foreground objects. Subsequently, an inpainting operation is performed, assigning these mismatched areas depth values consistent with the surrounding background, as shown in Figure 5.

Following this modification, an effective removal of moving objects and noise was achieved, resulting in a single clean background frame.

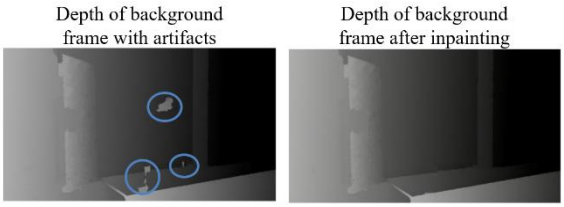

Figure 5: Depth inpainting.

The primary advantage of the static background estimation stage is that depth estimation is performed only once for the static elements of the given video. As a result, there is no noise in the static parts of the scene. This approach guarantees temporal consistency for said static parts.

2.2 AI-based moving objects tracking and segmentation

The second step of the proposed method involves AI-based tracking and segmentation of moving objects. The software implemented by the authors utilizes Detectron2 (Merz et al., 2023) (incorporating the PointRend module) for high-precision instance segmentation of each video frame. During implementation, a significant challenge arose: Detectron2 processes frames independently, resulting in a lack of temporal consistency in object IDs. The solution required adding the following elements:

- *Geometric matching*: use of the Jaccard index (intersection over union) to match object masks between consecutive frames.
- *Identity verification*: calculation of color difference coefficients to confirm object identity across frames.
- *Motion classification*: a decision logic that classifies an object as 'moving' based on established thresholds for color and position changes over time. Static objects are discarded from this path and treated as part of the background. Background areas are then excluded from further processing in this stage.

Figure 6 illustrates the output of the implemented software. After the segmentation stage is completed, the depth of the moving objects is estimated using the same depth estimation software and identical configuration parameters as those used for the background frame. An example of the resulting depth map is presented in Figure 7.

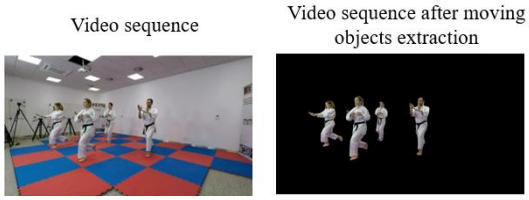

Figure 6: Video sequence after segmentation.

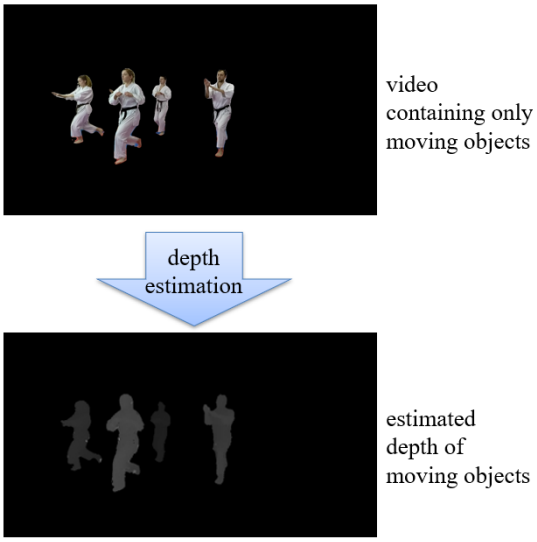

Figure 7: Depth map calculated for moving objects.

The proposed approach offers significant advantages for the depth estimation process. By maintaining object consistency through inter-frame analysis and inter-view projection, the workload of the depth estimation software is substantially optimized compared to processing full, non-segmented frames. Furthermore, this independent processing of foreground and background layers effectively eliminates the common problem of depth artifacts occurring at the boundaries between moving objects and the static background, as the depth values for each component are calculated independently.

2.3 Depth fusion

The final stage of the proposed solution involves merging a single static background depth map with the sequence of dynamic objects depth map to produce the final, temporally consistent output. To generate the complete depth map, the authors implemented a compositing algorithm, where dynamic moving objects were superimposed onto the static background using a pixel-wise maximum operation:

$$D_{merged}(x,y) = \max\left(D_{bg}(x,y), D_{obj}(x,y,t)\right) \quad (2)$$

where D_{bg} represents the static background depth, and D_{obj} represents the moving object depth at time t. This operation effectively overlays the moving objects onto the scene geometry.

Figure 8 depicts an example of depth map generated by the algorithm.

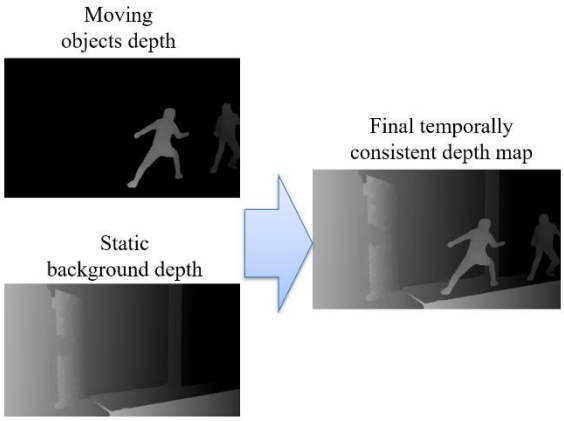

Figure 8: Depth map merging proccess.

The depth map created by the proposed method is free from frame-to-frame background noise, resulting in high temporal consistency.

3. IMPLEMENTATION AND TOOL SELECTION

The Segment Anything Model (SAM, (Kirillov et al., 2023)) is a significantly more advanced tool in terms of segmentation granularity and scene coverage when compared to Detectron2 (Merz et al., 2023). Its primary strength lies in its ability to produce dense, fine-grained segmentations of an image. However, in the context of the considered application, this feature becomes a notable limitation. SAM does not perform explicit object detection or classification, which are key components of the pipeline based on Detectron2. As a result, semantically coherent objects are not represented as single instances but rather as collections of independent segments. For example, a human figure may be decomposed into multiple regions corresponding to facial features, hair, or individual parts of clothing.

In contrast, Detectron2 models the same entity as a single, coherent object instance with an assigned class label. This representation significantly complicates tasks that require object identification and comparison across consecutive video frames, particularly object tracking and selective object filtering. The absence of stable instance identifiers and semantic class information in SAM necessitates the use of additional, often complex aggregation heuristics, thereby increasing the overall system complexity. Our reliance on a dedicated tracking and motion classification logic ensures that the estimated depth maps are not only spatially consistent but also optimized for inter-frame prediction in video codecs

An additional advantage of Detectron2 is the its relative consistency and reproducibility of detection results. Multiple executions of this model on identical input data typically yield highly similar instance masks, class assignments, and object identifiers. In contrast, SAM may produce more variable outputs, where the same object can be segmented differently across runs (e.g., clothing elements may be identified as separate segments in one execution and merged with the human body region in another).

Parameter Tuning

The detection score threshold was set to 0.9 in order to reduce the number of false positive detections. The optimal minimum confidence threshold, however, was found to be sequence-dependent. In simpler sequences containing a limited number of objects and featuring common categories (e.g., people and vehicles), a lower threshold of approximately 0.5 was sufficient to achieve correct object extraction in the majority of frames. Excessively high classification thresholds led to the rejection of valid detections that were required for subsequent stages of the analysis. Therefore, the threshold value of 0.9 was selected as a practical compromise between the number of detected objects and the overall classification reliability, ensuring stable performance across a wide range of video sequences.

Similarity Thresholds

Spatial Similarity (Jaccard Index): Spatial similarity between object masks was evaluated using the Jaccard index. A relatively low similarity threshold of 0.4 was adopted, as empirical observations and prior analysis indicated that the Jaccard score exhibits a highly non-linear behavior: masks corresponding to the same or strongly overlapping objects tend to produce rapidly increasing similarity values, whereas dissimilar or unrelated masks yield scores close to zero. Consequently, a threshold of 0.4 was sufficient to distinguish matching object instances while avoiding overly restrictive filtering.

Appearance Similarity (Color-Based Correlation): In addition to spatial overlap, appearance similarity was assessed using a color-based

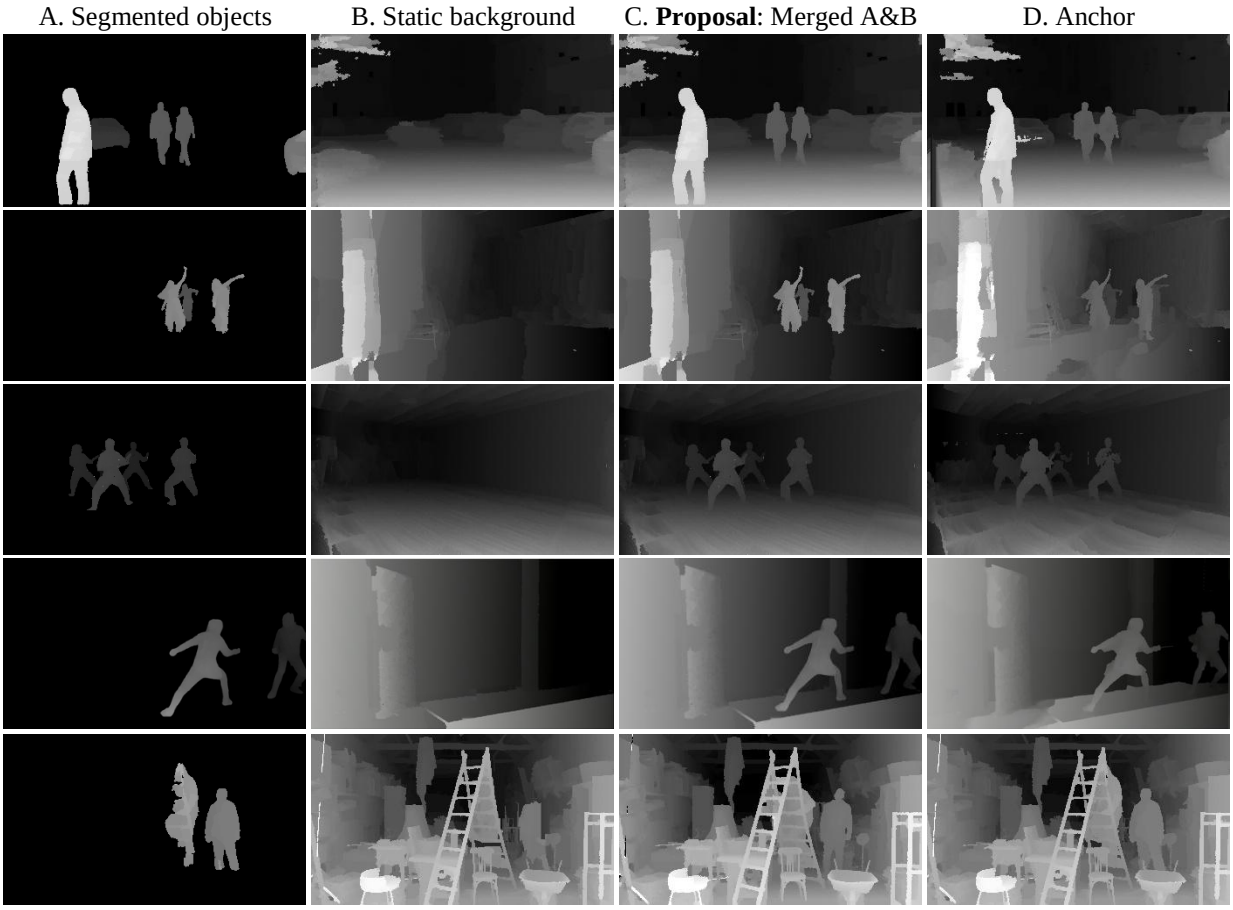

Figure 9: Single frame of depth maps for five tested sequences (from top): Carpark, Choreo, MartialArts, Fencing, and Barn. A: segmented moving objects, B: static background, C: combined background and moving objects, D: baseline anchor – depth calculated for full input views.

correlation measure. This approach was selected primarily for computational efficiency, serving as a lightweight alternative to more expensive similarity metrics such as Pearson correlation. The color-based similarity provided a fast approximation of appearance consistency across frames, enabling effective pruning of unlikely object correspondences without introducing significant computational overhead.

4. EXPERIMENTAL EVALUATION

The proposed method was evaluated using selected sequences from the MIV common test conditions (MIV CTC, (ISO/IEC, 2024)), focusing on natural scenes captured by real multicamera systems, containing dynamic foreground objects and different motion patterns. In particular, Fencing (Domański et al., 2016) and MartialArts (Mieloch et al., 2023) sequences were selected for quantitative evaluation, as they represent challenging scenarios with fast motion, frequent occlusions, and strong depth discontinuities.

For the remaining tested CTC sequences (Carpark (Mieloch et al., 2020), Choreo (Dziembowski et al., 2024), and Barn (Tapie et at., 2021)), the evaluation was limited to qualitative analysis, including visual comparison of estimated depth maps. This choice was motivated by the fact that the main contribution of this work lies in improving temporal consistency and visual stability, which are not always fully captured by objective metrics.

Depth estimation in both tested variants was performed using the same immersive video depth estimation (IVDE, (ISO/IEC, 2021)) software to ensure a fair comparison.

In the proposed approach, IVDE was run twice per sequence – once for estimating the depth of the static background (single frame) and once for estimating the depth of moving objects (for all 65 frames). Then, both depth maps were merged.

In the baseline approach (anchor), IVDE was used to estimate the depth of full input frames.

4.1 Qualitative results

Figure 9 illustrates a comparison of depth maps estimated using the proposed approach with those estimated based on full input views (anchor). As presented, the proposal enables a substantial increase in the spatial consistency of the depth maps. The objects' edges are sharper and more consistent (cf., Figure 10), and the background is smoother (cf., Figure 11). Moreover, fine details in the scene geometry are better preserved (cf., Figure 12).

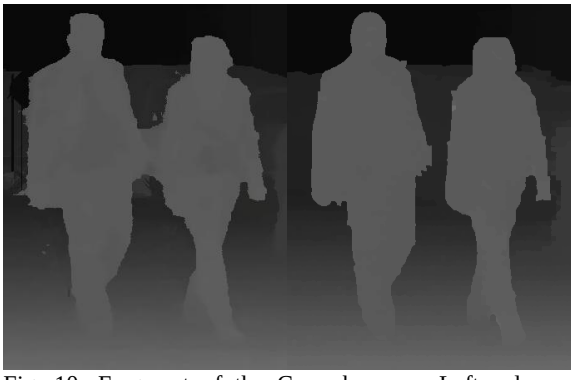

Fig. 10: Fragment of the Carpark scene. Left column: anchor, right column: proposal.

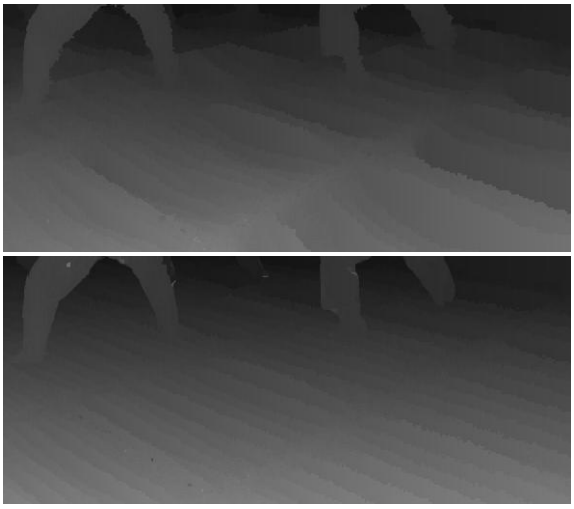

Fig. 11: Fragment of the floor in MartialArts sequence. Top row: anchor, bottom row: proposal.

Moreover, the proposed approach significantly increases the temporal consistency of the depth maps (Figure 13). The edges of the objects do not flicker, and there are fewer artifacts visible as background areas having foreground depth in proximity to moving objects (e.g., the part of the wall under the fencer's arm in Figure 13).

These improvements directly translate into more stable synthesized views, reducing visual discomfort and temporal artifacts during virtual navigation within the immersive video scene.

Fig. 12: Fragment of the Barn scene. Left column: anchor, right column: proposal.

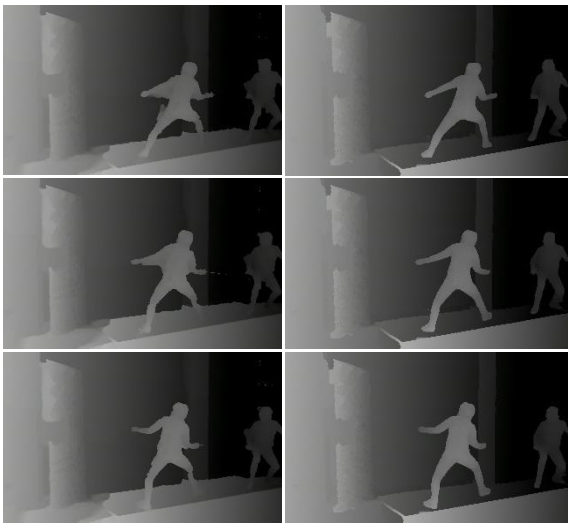

Figure 13: Three consecutive frames in Fencing sequence. Left: anchor, right: proposed approach.

4.2 Quantitative evaluation

Quantitative evaluation was conducted using the methodology defined in MIV CTC (ISO/IEC, 2024) on Fencing and MartialArts sequences. In particular, IV-PSNR (Dziembowski et al., 2022) was used for

objective quality evaluation, and bitrates, along with BD-rates (Bjøntegaard, 2001), were analyzed to assess the performance of immersive video coding when depth maps estimated using the proposed approach were used as input (in comparison to the anchor).

As presented in Tables 1 and 2, the average objective quality of a single frame is higher when using the depth estimated in the proposed approach. The increase, however, is slight.

Table 1: Objective quality of synthesized views, averaged over all views and all 65 frames; Fencing sequence.

Fencing	Average IV-PSNR [dB]	
	Anchor	Proposal
No compression	46.76	46.96
Rate point 1	43.79	44.00
Rate point 2	43.49	43.70
Rate point 3	42.31	42.47
Rate point 4	37.72	37.75

Table 2: Objective quality of synthesized views, averaged over all views and all 65 frames; MartialArts sequence.

MartialArts	Average IV-PSNR [dB]	
	Anchor	Proposal
No compression	35.59	36.02
Rate point 1	35.57	36.00
Rate point 2	35.52	35.93
Rate point 3	35.43	35.80
Rate point 4	34.67	34.95

The main advantage of the proposed method comes not from the quality of a single frame of a sequence, but from increased spatial and temporal consistency of the depth maps, what substantially increases depth map coding efficiency. It can be measured by the total bitrate required for transmission of the immersive video sequence. These results are presented in Tables 3 and 4.

Table 3: Total bitrate for the Fencing sequence; the bitrate includes texture, depth, and metadata subbitstreams.

Fencing	Total bitrate [Mbps]	
	Anchor	Proposal
Rate point 1	36.172	34.110
Rate point 2	16.843	15.855
Rate point 3	6.226	6.082
Rate point 4	2.048	1.921

Table 4: Total bitrate for the MartialArts sequence; the bitrate includes texture, depth, and metadata subbitstreams.

MartialArts	Total bitrate [Mbps]	
	Anchor	Proposal
Rate point 1	33.639	29.971
Rate point 2	23.917	21.110
Rate point 3	15.395	13.736
Rate point 4	5.050	4.586

As presented, the required bitrate is significantly lower than for the anchor, both for Fencing and MartialArts sequence. In terms of BD-rate (which combines the influence of quality and bitrate), the proposal outperforms the anchor by 10.4% for Fencing and by 15.6%, representing a significant coding efficiency gain.

5. CONCLUSIONS AND FUTURE WORKS

The proposed pipeline successfully improves the temporal consistency of depth maps by eliminating background flickering and preserving object edges. Independent processing of background and foreground objects further optimizes depth estimation, reducing computational time. Estimation is particularly efficient when applied to a single background frame obtained via a temporal median operation, and less complex sequences containing only moving objects against a uniform background also benefit from faster processing.

Unlike generic video depth frameworks that may introduce temporal drifts, the “static background” assumption is a deliberate design choice for systems which assume that video data will be compressed. By estimating the background depth only once, we completely eliminate temporal flickering in static regions. This drastically reduces the bitrate required for transmission, as the encoder can effectively exploit the lack of noise in the depth stream.

The processed depth maps demonstrate significant quality improvements, making them suitable for applications such as virtual view synthesis in multi-view sequences. Depth maps generated with the developed software were proposed to the ISO/IEC MPEG Video Coding group and have been adopted as the new depth maps for the Fencing sequence. Results obtained for the other sequences highlight potential directions for further research, including the development of improved motion detection algorithms leveraging instance segmentation produced by Detectron2. Some inaccuracies observed in various sequences can be attributed to limitations in Detectron2’s handling of

partially occluded objects, which may result in imprecise masks and, consequently, errors in matching similar sets and computing color differences.

Additionally, the developed software provides a flexible framework for evaluating alternative object detectors, enabling further exploration of instance segmentation and motion-based depth estimation strategies. Future work should focus on enhancing the motion detection algorithm (current bottleneck), handling occlusions more robustly, and exploring more sophisticated similarity measures to improve object association across frames.

ACKNOWLEDGEMENTS

This work was supported by the Ministry of Science and Higher Education of Republic of Poland.

REFERENCES

Bjøntegaard, G. (2001). Calculation of Average PSNR Differences between RD-curves. *Standardizaton document: ITU – T SG16, Doc. VCEG – M33.*

Domański, M. et al. (2016). Multiview test video sequences for free navigation exploration obtained using pairs of cameras. *Document ISO/IEC JTC1/SC29/WG11 MPEG2016, M38247*.

Dziembowski, A., et al. (2024). "Choreo" natural content for INVR applications. *Document ISO/IEC JTC1/SC29/WG4 MPEG 147, M68221*.

ISO/IEC (2021). Manual of Immersive Video Depth Estimation. *Document ISO/IEC JTC1/SC29/WG04 MPEG VC N0058*.

ISO/IEC, (2024). Common test conditions for MPEG immersive video. *Document ISO/IEC JTC1/SC29/WG04 MPEG VC N0539, 2024*.

Kirillov, A. et al. (2023). Segment anything, arXiv:2304.02643

Merz, G.M. et al. (2023). Detection, Instance Segmentation, and Classification for Astronomical Surveys with Deep Learning (DeepDISC): Detectron2 Implementation and Demonstration with Hyper Suprime-Cam Data. arXiv:2307.05826.

Mieloch, D., Grzelka, A. (2018). Segmentation-based Method of Increasing The Depth Maps Temporal Consistency. In *International Journal of Electronics and Telecommunications*, Vol. 64, No. 3, Electronics and Telecommunications Committee of Polish Academy of Sciences, Warsaw.

Mieloch, D. et al. (2020). Natural outdoor test sequences. *Document MPEG129/m51598.*

Mieloch, D., Dziembowski, A., Domański, M. (2021). Depth Map Refinement for Immersive Video. In *IEEE Access*, Vol. 9, 2021, pp. 10778 – 10788.

Mieloch, D., Garus, P., Milovanović, M., Jung, J., Jeong, J., Lingadahalli, R., S., Salahieh, B. (2022). Overview and Efficiency of Decoder-Side Depth Estimation in MPEG Immersive Video. In *IEEE Transactions on Circuits and Systems for Video Technology.*

Mieloch, D. et al. (2023). [MIV] New natural content – MartialArts. *Document ISO/IEC JTC1/SC29/WG4 MPEG 141, m61949*.

Stankiewicz, O., Wegner, K., Tanimoto, M., Domański, M. (2013). "Enhanced Depth Estimation Reference Software (DERS) for Free-viewpoint Television", *Document ISO/IEC JTC1/SC29/WG11 Doc. MPEG M31518*, Geneva.

Stankiewicz, O., Domański, M., Wegner, K. (2015). Estimation of Temporally-Consistent Depth Maps from Video with Reduced Noise. *3DTV Conference: The True Vision-Capture, Transmission and Display of 3D Video*, 3DTV- Con 2015, Lisbon, Portugal.

Stankowski, J., Dziembowski, A. (2022). Real-time CPU-based view synthesis for omnidirectional video. *30th International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision : WSCG 2022, Pilsen, Czech Republic*.

Tanimoto, M. (2012). FTV (free-viewpoint television) *Published online by Cambridge University Press*.

Tapie, T., Schubert, A., Gendrot, R., Briand, G., Thudor, F., Doré, R. (2021). [MIV] Barn new natural content proposal for MIV. *Document ISO/IECJTC1/SC29/ WG4, m56632.*

Yang, L., Teratani, M., Fujii, T., Tanimoto, M. (2011). High-quality virtual view synthesis in 3DTV and FTV. *3D Research. 2. 10.1007/3DRes.04(2011)*5.

Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H. (2024). Depth Anything V2. In *Advances in Neural Information Processing Systems*, Vol. 37, pp. 21875-21911.