Katsenou, A., Zhang, F., & Bull, D. (2022). A CNN-based Post-Processor for Perceptually-Optimized Immersive Media Compression. In *2022 IEEE International Conference on Image Processing (ICIP)* IEEE Computer Society. https://doi.org/10.1109/ICIP46576.2022.9897759

Peer reviewed version

Link to published version (if available):
10.1109/ICIP46576.2022.9897759

Link to publication record in Explore Bristol Research
PDF-document

University of Bristol - Explore Bristol Research
General rights

A CNN-BASED POST-PROCESSOR FOR PERCEPTUALLY-OPTIMIZED IMMERSIVE MEDIA COMPRESSION

Angeliki Katsenou, Fan Zhang, and David Bull

Visual Information Laboratory, University of Bristol, UK

ABSTRACT

In recent years, resolution adaptation based on deep neural networks has enabled significant performance gains for conventional (2D) video codecs. This paper investigates the effectiveness of spatial resolution resampling in the context of immersive content. The proposed approach reduces the spatial resolution of input multi-view videos before encoding, and reconstructs their original resolution after decoding. During the up-sampling process, an advanced CNN model is used to reduce potential re-sampling, compression and synthesis artefacts. This work has been fully tested with the TMIV coding standard using a Versatile Video Coding (VVC) codec. The results demonstrate that the proposed method achieves significant rate quality performance improvement for the majority of the test sequences, with an average BD-VMAF improvement of 3.07 over all sequences.

Index Terms— Immersive Media, Video Compression, TMIV, Spatial Resolution Adaptation, Perceptual Quality, CNN.

1. INTRODUCTION

In recent years the market for volumetric video has grown significantly, with investments supporting Virtual Reality (VR), Augmented Reality (AR), and extended reality (XR) products. In response to this trend and to support the streaming of such content, new standards have emerged for coding immersive media, in particular MPEG-I, "Coded Representation of Immersive Media" [1]. A major challenge for all types of volumetric video, besides the creation of the content, is its storage and distribution over existing and future networks. To this end, standardization bodies have developed MPEG Immersive Video (MIV) [2]. MIV provides 6 Degrees of Freedom (DoF) rendering capability, enabling the viewer to explore the scene from different angles.

Immersive video is defined as content synchronously captured by an array of cameras pointing towards the same scene. Cameras may be converging, parallel or diverging and the scene may be naturally captured or computer generated. Each camera captures a video sequence from a different viewpoint and this is accompanied by another sequence containing estimated depth information. The MIV coding workflow is based on 2D video coding and uses the concept of video compression of atlases, which contain pruned views of the input content [3, 2]. Since its initial versions, the MIV test model (TMIV) [4], has been refined and improved in terms of its compression performance. However, for certain content types (especially natural scenes with high levels of motion) it still produces noticeable artifacts (e.g. bad blending and disocclusions) that often degrade the quality of experience [5].

Efforts to improve the performance of TMIV have primarily focused on improving the depth estimation and on the synthesis at the decoder side [6, 7]. In [6], a patch-based estimation of the depth at the decoder side is proposed. In [7], an iterative depth refinement process is presented that improves inter-view consistency. A further approach, that claims to reduce noticeable artifacts as well as provide bitrate savings, was proposed in [8]. While all the aforementioned works succeeded in improving average media quality over the benchmark, significant scope for further perceptual rate-quality optimization remains.

Since TMIV is based on a 2D video coding framework, we hypothesise that it would benefit from resolution adaptation methods [9, 10, 11, 12, 13] and post-processing methodologies [14, 15] that have proved successful in conventional coding. In this context we propose a novel approach that exploits resolution adaptation and post-processing of synthesized views, realised using a Convolutional Neural Network (CNN). This allows the fixed budget for luma samples to be used to construct more accurate atlases, especially for those sequences with higher motion. The results confirm that the compression gains are significant for the majority of sequences evaluated, compared to the benchmark method, TMIV11.1. To the best of our knowledge, this is the first time that this approach has been applied to immersive content.

The remainder of this paper is organised as follows. Section 2 describes the proposed method. Section 3 presents the evaluation setup and test configurations. Section 4 provides the comparison results between the proposed method and the anchors on the test content. Finally, conclusions and future work are outlined in Section 5.

2. PROPOSED METHOD

The proposed approach is illustrated in Fig. 1. Compared to the benchmark method, the new approach downsamples (by a factor of two) the view sequences (texture and depth) before feeding these into the TMIV pipeline. After pre-processing at the TMIV Encoder and creation of the resulting texture and geometry atlases, a fast implementation of VVC, VVenC [16], is deployed for their compression (as recommended by TMIV). The TMIV Decoder is employed to generate the synthesized views. The final step is to spatially up-sample these sequences and feed them into a post-processing (PP) CNN-based block, a Multi-level Feature review Residual Network (MFRNet) [15].

2.1. Coding

As mentioned above, TMIV version 11.1, is employed for coding [4]. The outputs of the TMIV encoding process comprise texture atlases, depth map atlases (geometry video data), and the MIV bitstream [2]. The texture and depth atlases are fed into VVenC for
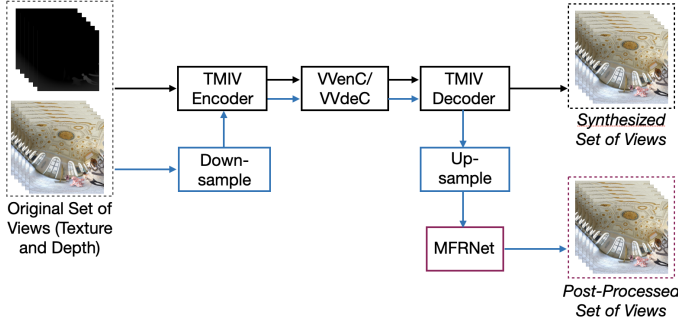
Fig. 1: Outline of proposed methodology. The workflow of the benchmark method is indicated by black arrows, while the proposed method is depicted by blue ones.

encoding and then to VVdeC for decoding [16]. The TMIV decoder consists of the MIV normative decoder and metadata parser, and a block to patch map decoder. Besides the normative part, the TMIV decoder comprises the geometry (depth) upscaler, a culler, and a renderer. The TMIV decoder takes both the decoded texture and depth atlases as well as the MIV bitstream as inputs to produce the final synthesized views.

2.2. Resolution Resampling

Spatial resolution adaptation has been demonstrated to deliver significant bitrate savings [10] and it has been used for optimizing the video streaming in large scale video streaming services, such as Netflix [9, 11]. Downsampling of the original sequences is achieved using a Lanczos filter [17], as implemented in ffmpeg [18]. This filter is commonly used for adaptive streaming purposes as it preserves high frequency content with reduced blurring compared to other methods [9, 11]. For upsampling the synthesized view sequence, after the TMIV Decoder, a Nearest Neighbors (NN) filter is first employed to restore the original resolution. This has proved effective, when combined with a CNN post-processor, in [19, 15].

2.3. CNN-based Post-Processing

To further improve video quality, and reduce re-sampling, compression and synthesis artefacts, an additional CNN is employed after NN-upsampling. We use the MFRNet architecture [15], which was proposed for PP and in-loop filtering (ILF) in the context of 2D video compression. This network consists of four Multi-level Feature review Residual dense Blocks, which are connected using a cascading structure. Each of these blocks extracts features from multiple convolutional layers using dense connections and a multi-level residual learning structure. In order to further improve information flow between these blocks, each of them also reuses high dimensional features from previous blocks. In this work, we employ MFRNet models that were trained [15] on the BVI-DVC database [20] with VVC, to post-process the upsampled rendered views.

3. EXPERIMENTAL CONFIGURATION

This section discusses the experimental setup: the test sequences used, the TMIV configuration, the compared methods, and the training of the CNN models for post-processing.

3.1. Test Conditions

For our experiments and the anchor generation, we adopt the MPEG methodology as defined in common test conditions (CTC) [21]. The CTC define seven mandatory test sequences used for the evaluation of proposed algorithms of various resolutions that are either computer generated (CG) or natural content (NC). The parameters of the test sequences are summarized in Table 1. A, B, and C are in Equirectangular Projection (ERP), while the others are in Perspective Projection (PTP). Each sequence (97 frames) was encoded in the setup of three groups, where each group has one atlas containing a basic view and one patch-atlas. Furthermore, in addition to the four given quantization parameter pairs (QP_T, QP_D) for video compression with VVenC/VVdeC [16], another quantization parameter pair is added to show the performance at low bit-rate. The set of (QP_T, QP_D) pairs is: $\{(22, 4), (27, 7), (32, 11), (37, 15)\}$. For the sequences generated by the proposed method, we adjusted the QP_T quantization parameter due to the spatial resolution downsampling to the following: $\{16, 21, 26, 31\}$ in order to result to a similar bitrate range as that of the anchors generated with benchmark method. This QP shift was based on a study from our previous work on 2D video resolution adaptation [10].

We evaluated the bitrate and synthesized view quality performance provided by the proposed method compared to the anchor, using the Bjøntegaard delta rate metrics [22], in terms of PSNR-Y, IV-PSNR [23, 24], and VMAF [25].

Id	Sequence	Type	Resolution	no Views
A	Classroom	CG	4096x2048	14
B	Museum	CG	2048x2048	18
C	Hijack	CG	4096x2048	9
D	Painter	NC	2048x1088	16
E	Frog	NC	1920x1080	13
J	Kitchen	CG	1920x1080	24
L	Fencing	NC	1920x1080	9

Table 1: Test Sequences and Parameters

3.2. Compared Methods

To the best of our knowledge, no other methods have performed either resolution adaptation and/or post-processing on immersive content. We thus compare against the methods described below:

- *Anchor*: This method represents the benchmark, and represents the results produced by applying the TMIV pipeline on the native resolution test sequences.

- *Re-scaled*: This method extends the benchmark by including the spatial resolution re-sampling (using the same filters as the proposed method), before and after the coding pipeline. The resulting sequences are not post-processed.

- *MFRNet*: This is the proposed method that uses a trained CNN, MFRNet, to post-process the sequences generated with the Re-scaled method.

3.3. Network Training

Datasets are key to the effective training of any CNN model. Since no large volumetric datasets exist, we employed a standard video format dataset, BVI-DVC [20] for training. BVI-DVC is one of

the largest video datasets for training deep video compression algorithms; it contains 800 video sequences, each of 64 frames with 10 bit depth in YCbCr 4:2:0 format at four different resolutions from 270p to 2160p. These sequences were compressed with the VVC VTM 7.0 codec using the JVET-CTC Random Access (RA) configuration with four QP values: 22, 27, 32 and 37. The CNN implementation and training was based on TensorFlow (version 1.8.0) framework [26]. Based on the generated training content, four CNN models aligned with the four QP groups were trained, which are subsequently used in the evaluation stage for different base QP values. More details on the MFRNet implementation and training can be found in [15].

4. RESULTS AND DISCUSSION

This section reports and discusses the results generated by the compared methods.

4.1. Objective Quality Assessment

IV-PSNR and VMAF have been reported to align better with subjective scores for volumetric video content [2]. In Figs. 2, the Rate-Quality (RQ) curves for the test sequences based on VMAF are illustrated. As can be seen from both figures, for the majority of sequences, the proposed method outperforms the benchmark. According to the VMAF metric, which has been proved to exhibit the highest correlation with perceptual quality for volumetric data [27], the proposed method demonstrates superior perceptual quality for the majority of the sequences compared to the Re-scaled and Anchor methods. This improvement in performance can be attributed as follows. Firstly resolution adaptation, while maintaining `maxLumaPictureSize`, results in very detailed atlases that can provide a better depth estimation and thus improve synthesizer perfomance. Secondly, and more importantly, MFRNet post-processing succeeds in "correcting" compression errors and resolution upsampling artifacts.

The exception, where the Anchor method delivers superior perceptual quality, is for sequences A and D. For sequence A, this is attributed to the fact that this static sequence has a high level of film grain noise, which is smoothed after the spatial resolution downsampling. The reason why our approach does not perform as well for sequence D can be attributed to the high density of textures in both background and foreground. These make segmentation and depth estimation challenging. Moreover, resolution adaptation is known to underperform for highly complex static textures.

From the RQ curves in Figs. 2, it is evident that the BD-Rate savings for sequences {B, C, E, J, L} are significantly higher than 100%, and similarly the Anchor method delivers results of similar significance for sequences A and D. In Table 2, we report the BD statistics on VMAF, IV-PSNR, and PSNR-Y[1]. Based on these figures, although the Re-scaled method reports quality losses compared to the Anchor, the proposed method delivers significant gains of BD–VMAF 3.07, BD–IV-PSNR 0.76dB, and BD–PSNR-Y 1.66dB on average compared to the Anchor.

4.2. Visual Inspection

To confirm the results presented above, we performed a visual inspection of the results. Due to page length limitations, we only in-

Fig. 2: Rate–VMAF curves of all compared methods.

clude in Fig. 3 a selection of patches sampled from the generated sequences for specific views using all three methods and including the original view. As can be seen from the patches 3(e)-(l), the proposed method produces sequences without dissocclusion or blocking artifacts. It is notable that, for sequence L, both Re-scaled and MFRNet methods successfully avoid the dissocclusion results of the benchmark method. On the other hand, the benchmark method successfully restructures the grain noise of sequence A (see 3(a)-(d)). These results confirm the need for an extended dataset of volumetric content to be able to properly investigate the relation of content characteristics to the effectiveness of the benchmark and the proposed approach.

4.3. Computational complexity

The training and evaluation processes were executed on a shared cluster provided by the University of Bristol, BlueCrystal Phase 4 (BC4) [28]. Each node contains two 14 core 2.4 GHz Intel E5-2680

---

[1]We did not include the BD-Rate results because differences in quality scales in most cases are so big that the extrapolation across the bitrate axis to compute the integral differences will not be accurate
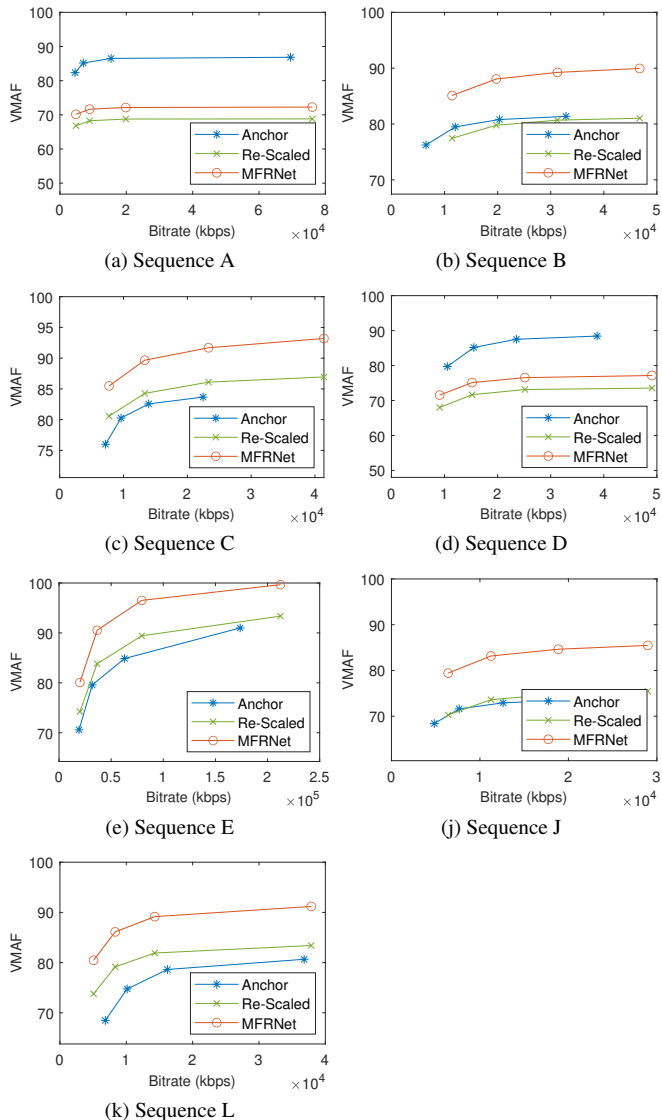
	Id	BD–VMAF	BD–IV-PSNR	BD–PSNR-Y
Re-scaled	A	-17.68	-2.76dB	-2.43dB
	B	-1.08	-3.22dB	-2.06dB
	C	2.11	0.49dB	-0.44dB
	D	-13.75	-2.37dB	-2.60dB
	E	2.73	-1.88dB	-1.41dB
	J	0.75	-2.61dB	-1.73dB
	L	4.42	-0.19dB	1.03dB
Total		-3.22	-1.45dB	-1.37dB

	Id	BD–VMAF	BD–IV-PSNR	BD–PSNR-Y
Proposed	A	-14.28	-1.75dB	-0.05dB
	B	7.12	-0.48dB	0.42dB
	C	7.43	3.30dB	3.07dB
	D	-10.29	-0.84dB	-0.65dB
	E	9.48	1.01dB	1.98dB
	J	10.28	0.67dB	2.25dB
	L	11.77	1.63dB	4.57dB
Total		3.07	0.76dB	1.66dB

Table 2: BD statistics of quality metrics of the Re-scaled and the Proposed Method over the Anchor.

V4 (Broadwell) CPUs, 128 GB of RAM, and NVIDIA P100 GPU devices. Regarding the computational complexity of the proposed method, MFRNet requires on average 18.85% more time per frame (roughly 4 times more than decoding one frame with VVdeC) compared to the total time required for the VVdeC and the TMIVdec for the Re-scaled method. Taking into account that the Re-scaled method compared to the Anchor is also 55.22% faster in terms of Decoding on average, the proposed method is around 36% faster than the Anchor method. The resolution adaptation execution time is negligible compared to the rest of the processes, as it accounts (on average) for just a few seconds per sequence. TMIV and VVC encoding times between the compared methods are very similar because, by using the same fixed `maxLumaPictureSize` as defined by CTC, the produced atlases have almost the same number of pixels.

5. CONCLUSION

In this paper, we presented a novel approach for perceptually enhancing immersive video coding. The proposed approach is based on spatial re-sampling and post-processing the TMIV decoded sequences with a CNN, MFRNet. The evaluation of the method is based on the mandatory sequences of the TMIV CTCs, while the training of the MFRNet was performed on a conventional video dataset, BVI-DVC. Our approach significantly reduces the impact of artifacts introduced by the TMIV Decoder, and particularly by the Synthesizer. It is notable that the proposed method performs particularly well for sequences with higher amounts of motion, such as Frog (E) and Fencing (L). Finally, despite the additional post-processing, the proposed method requires on average only 36% of execution time compared to the Anchor method.

The results presented here inspire further investigations including the integration of MFRNet in-loop, at the TMIV Decoder. Furthermore, we have demonstrated the need to expand the set of training and test sequences and complement the evaluation of the proposed method with subjective evaluation. Finally, as derived from
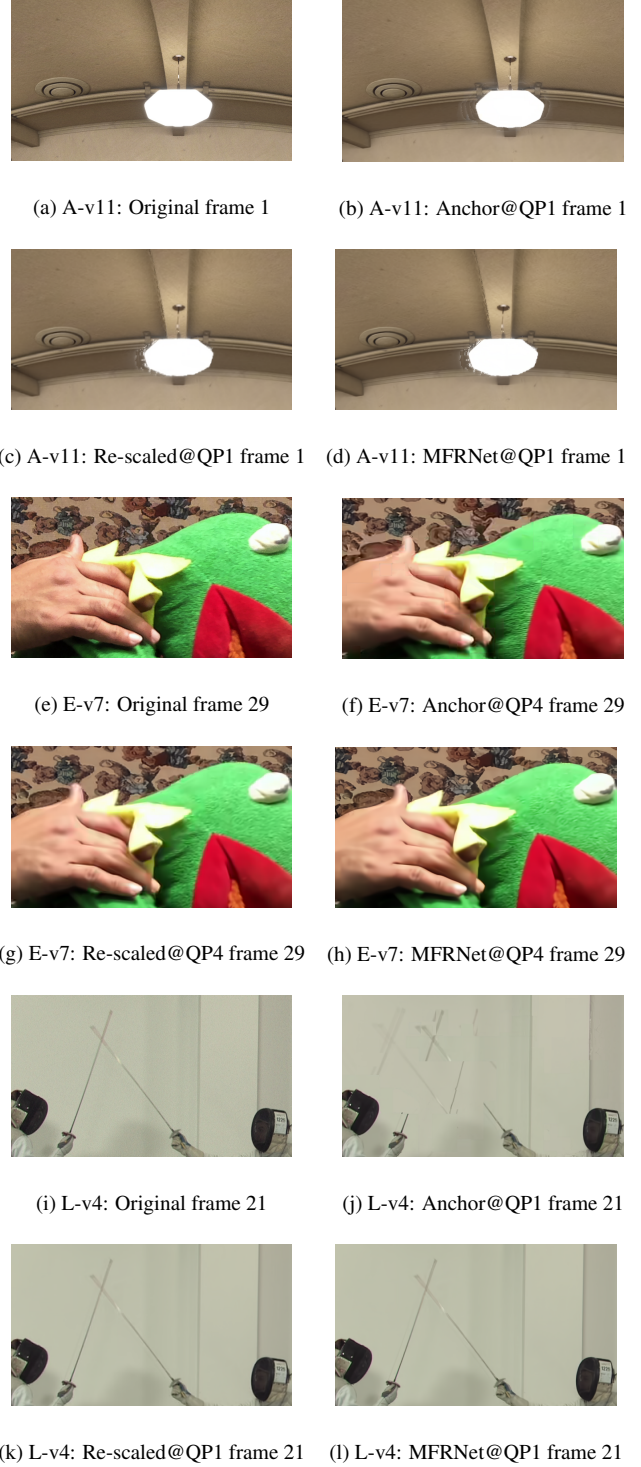

(a) A-v11: Original frame 1   (b) A-v11: Anchor@QP1 frame 1

(c) A-v11: Re-scaled@QP1 frame 1   (d) A-v11: MFRNet@QP1 frame 1

(e) E-v7: Original frame 29   (f) E-v7: Anchor@QP4 frame 29

(g) E-v7: Re-scaled@QP4 frame 29   (h) E-v7: MFRNet@QP4 frame 29

(i) L-v4: Original frame 21   (j) L-v4: Anchor@QP1 frame 21

(k) L-v4: Re-scaled@QP1 frame 21   (l) L-v4: MFRNet@QP1 frame 21

Fig. 3: Examples of patches for visual comparison using the original view and the compared methods.

the quality assessment results, the development of improved perceptual quality metrics would benefit this field of research.

6. REFERENCES

[1] 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.

[2] J. M. Boyce, R. Doré, A. Dziembowski, J. Fleureau, J. Jung, B. Kroon, B.and Salahieh, V. K. M. Vadakital, and L. Yu, "MPEG Immersive Video Coding Standard," *Proceedings of the IEEE*, vol. 109, no. 9, pp. 1521–1536, 2021.

[3] D. B. Graziosi and B. Kroon, "Video-based coding of volumetric data," in *2020 IEEE International Conference on Image Processing (ICIP)*, 2020, pp. 2706–2710.

[4] "MIV reference software," https://gitlab.com/mpeg-i-visual/tmiv, 2021.

[5] D. Bonatto, S. Fachada, and G. Lafruit, "RaViS: Real-time accelerated View Synthesizer for immersive video 6DoF VR," *Electronic Imaging*, pp. 382–1–382–9, 2020.

[6] M. Milovanović, F. Henry, M. Cagnazzo, and J. Jung, "Patch Decoder-Side Depth Estimation In MPEG Immersive Video," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2021, pp. 1945–1949.

[7] D. Mieloch, A. Dziembowski, and M. Domański, "Depth map refinement for immersive video," *IEEE Access*, vol. 9, pp. 10 778–10 788, 2021.

[8] S. Lee, J.-B. Jeong, and E.-S. Ryu, "Efficient Group-Based Packing Strategy for 6DoF Immersive Video Streaming," in *International Conference on Information Networking*, 2022, pp. 310–314.

[9] I. Katsavounidis, "Dynamic optimizer - a perceptual video encoding optimization framework," https://medium.com/netflix-techblog/dynamic-optimizer-a-perceptual-video-encoding-optimization-framework-e19f1e3a277f.

[10] M. Afonso, F. Zhang, A. Katsenou, D. Agrafiotis, and D. Bull, "Low complexity video coding based on spatial resolution adaptation," in *IEEE International Conference on Image Processing*, 2017, pp. 3011–3015.

[11] A. V. Katsenou, J. Sole, and D. R. Bull, "Efficient Bitrate Ladder Construction for Content-Optimized Adaptive Video Streaming," *IEEE Open Journal of Signal Processing*, vol. 2, pp. 496–511, 2021.

[12] M. Afonso, F. Zhang, and D. R. Bull, "Video compression based on spatio-temporal resolution adaptation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 1, pp. 275–280, 2018.

[13] D. Ma, F. Zhang, and D. R. Bull, "Cvegan: a perceptually-inspired gan for compressed video enhancement," *arXiv preprint arXiv:2011.09190*, 2020.

[14] F. Zhang, D. Ma, C. Feng, and D. R. Bull, "Video Compression With CNN-based Postprocessing," *IEEE MultiMedia*, vol. 28, no. 4, pp. 74–83, 2021.

[15] D. Ma, F. Zhang, and D. R. Bull, "MFRNet: A New CNN Architecture for Post-Processing and In-loop Filtering," *IEEE Journal of Selected Topics in Signal Processing*, vol. 15, no. 2, pp. 378–387, 2021.

[16] J. Brandenburg, A. Wieckowski, T. Hinz, A. Henkel, V. George, I. Zupancic, C. Stoffers, B. Bross, H. Schwarz, and D. Marpe, "Towards fast and efficient vvc encoding," in *2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP)*, 2020, pp. 1–6.

[17] C. E. Duchon, "Lanczos filtering in one and two dimensions," *Journal of Applied Meteorology*, vol. 18, no. 8, pp. 1016–1022, 1979.

[18] "FFMPEG," https://www.ffmpeg.org.

[19] F. Zhang, M. Afonso, and D. R. Bull, "Vistra2: Video coding using spatial resolution and effective bit depth adaptation," *Signal Processing: Image Communication*, vol. 97, p. 116355, 2021.

[20] D. Ma, F. Zhang, and D. Bull, "BVI-DVC: A Training Database for Deep Video Compression," *IEEE Transactions on Multimedia*, 2021.

[21] J. Jung, B. Kroon, and J. Boyce, "Common test conditions for immersive video," ISO/IEC JTC 1/SC 29/WG 11 N18997, Tech. Rep., 2020.

[22] G. Bjøntegaard, "Calculation of average PSNR differences between RD-curves," in *13th VCEG Meeting*, no. VCEG-M33. Austin, Texas, USA: ITU-T, April 2001.

[23] A. Dziembowski and M. Domański, "[MPEG-I Visual] Objective quality metric for immersive video," ISO/IEC JTC1/SC29/WG11 MPEG/M48093, Tech. Rep., 2019.

[24] J. Stankowski and A. Dziembowski, "Slightly faster IVP-SNR," ISO/IEC JTC1/SC29/WG04 MPEG/M55752, Tech. Rep., 2021.

[25] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, and M. Manohara, "Toward a practical perceptual video quality metric," *The Netflix Tech Blog*, 2016.

[26] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "Tensor-Flow: Large-Scale Machine Learning on Heterogeneous Systems," https://www.tensorflow.org/, 2015, software available from tensorflow.org.

[27] S. Van Damme, M. T. Vega, and F. De Turck, "A Full- and No-Reference Metrics Accuracy Analysis for Volumetric Media Streaming," in *13th International Conference on Quality of Multimedia Experience (QoMEX)*, 2021, pp. 225–230.

[28] "BlueCrystal Phase 4," https://www.acrc.bris.ac.uk/protected/bc4-docs/.