3D-STDF: Compressed Video Quality Enhancement with 3D Spatio-Temporal Fusion and Deformable Convolution

Garibaldi da Silveira Júnior
garibaldi.dsj@inf.ufpel.edu.br
ViTech - PPGC - UFPel
Pelotas, Brazil

Bruno Zatt
zatt@inf.ufpel.edu.br
ViTech - PPGC - UFPel
Pelotas, Brazil

Daniel Palomino
dpalomino@inf.ufpel.edu.br
ViTech - PPGC - UFPel
Pelotas, Brazil

Guilherme Correa
gcorrea@inf.ufpel.edu.br
ViTech - PPGC - UFPel
Pelotas, Brazil

ABSTRACT

Compressed videos frequently present artifacts that compromise visual quality. Deep learning models have demonstrated significant effectiveness in mitigating such distortions. In this study, we introduce 3D-STDF, an architecture based on the well-known Spatio-Temporal Deformable Fusion (STDF) and augmented with 3D convolutions to more effectively model temporal dependencies across video frames. Furthermore, we refine the Quality Enhancement (QE) module by integrating residual blocks, thereby enabling the extraction and representation of more intricate spatial features. Experimental results indicate that models based on the 3D-STDF architecture achieved an overall average improvement of up to 0.607 dB in PSNR, clearly outperforming previous STDF-based solutions.

KEYWORDS

Video Quality Enhancement, Video Coding, Deep Learning, Spatio-Temporal Deformable Fusion

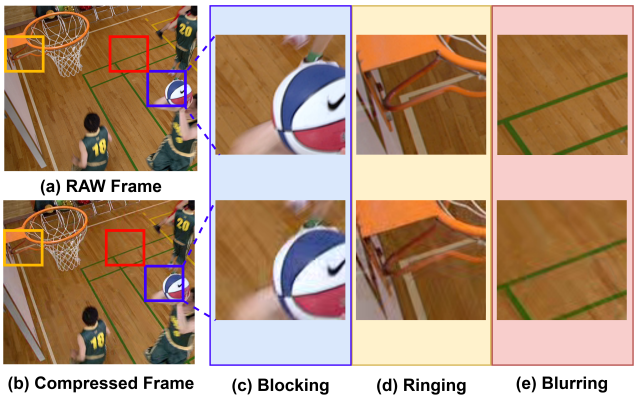

Figure 1: Compression Artifacts: (a) Original Frame (RAW); (b) Compressed Frame; (c) Blocking Artifact; (d) Ringing Artifact; (e) Blurring Artifact.

1 INTRODUCTION

The transmission of high-resolution videos, such as 4K and 8K, has expanded significantly in recent years, becoming one of the main drivers of global internet traffic growth. According to the 2024 Global Internet Phenomena Report published by Sandvine, video on demand remains the largest contributor to internet data traffic, surpassing categories such as online gaming and cloud services. The report indicates that in 2024, global traffic exceeded 33 exabytes per day, with the average daily consumption per user reaching 4.2 GB. Video content accounted for approximately 54% of the total downstream traffic volume, solidifying its position as the primary driver of broadband usage worldwide [19].

To meet this demand, researchers and industry professionals have developed compression techniques essential for video storage and transmission over limited bandwidth. These techniques are categorized into lossless, which enables exact reconstruction of the original video, and lossy, which achieves higher compression rates at the cost of some quality loss [20]. During compression, lossy methods introduce visual artifacts such as blocking, ringing, and blurring, which degrade the perceived quality [7]. Figure 1 illustrates such effects, with blocking in the blue area of the ball (c), ringing with wave-like patterns along the orange rim (d), and blurring with lost details like nails and floor separations (e). These artifacts often overlap, making them hard to isolate individually.

Aiming at these issues, modern video coding standards like High Efficiency Video Coding (HEVC), Versatile Video Coding (VVC), and AOMedia Video 1 (AV1) employ standardized in-loop filtering algorithms, including the Deblocking Filter (DF) [16] to reduce blocking artifacts, the Adaptive Loop Filter (ALF) [22] to minimize distortion between original and decoded samples, and the Sample Adaptive Offset (SAO) [8] to mitigate banding artifacts, all applied after frame reconstruction to enhance visual quality. However, compression artifacts are still visible in compressed videos, especially those encoded for transmission using low bitrates. These limitations indicate that traditional in-loop filters are not sufficient to fully restore perceptual quality, motivating the development of advanced post-processing techniques to further reduce artifacts and enhance the visual experience.

Unlike traditional in-loop filters, Deep Neural Networks (DNNs), especially those based on Convolutional Neural Networks (CNNs),

In: Proceedings of the Brazilian Symposium on Multimedia and the Web (WebMedia'2025). Rio de Janeiro, Brazil. Porto Alegre: Brazilian Computer Society, 2025.

ISSN 2966-2753

can analyze surrounding pixels to extract contextual and structural patterns, thereby minimizing the introduction of new artifacts during enhancement. In the context of Video Quality Enhancement (VQE), such models can be used either in-loop, where enhanced frames influence the encoding of future frames, or as post-processing modules, which refine the decoded frames independently of the video codec [13].

This work introduces a novel post-processing VQE model named 3D-STDF, built upon the established Spatio-Temporal Deformable Fusion (STDF) architecture [6]. The proposed model introduces temporal fusion blocks that utilize 3D convolutions to capture temporal features across frames during fusion. Additionally, the baseline QE module is improved by replacing flat CNN layers with residual blocks, enabling deeper feature extraction without losing critical spatial information. Experiments show that the proposed model achieves consistent objective improvements, with overall average improvement of up to 0.607 dB in Peak Signal-to-Noise Ratio (PSNR), surpassing both the original STDF and other related approaches.

2 RELATED WORK

Early VQE methods such as [3, 15] applied uniform linear heuristics across all pixels, disregarding spatial variability and often degrading unaffected regions. These were replaced by nonlinear machine learning models capable of analyzing local regions and applying more effective corrections [18]. With the advancement of DNNs, more robust architectures emerged, capable of learning complex patterns and significantly improving artifact reduction. A key milestone was the introduction of the Artifact Reduction Convolutional Neural Network (ARCNN) [7], which inspired further works such as VRCNN [5], an in-loop filter for HEVC, and MDCNN [12], a post-processing model applied at the decoder side.

Later studies began leveraging temporal redundancy through multi-frame architectures [2, 6, 9, 27], which use sliding windows to enhance the central frame by fusing information from adjacent ones. This exploits the GOP structure in video codecs, using high-quality frames to enhance low-quality ones [21]. Temporal alignment is often achieved via optical flow estimation, preserving motion continuity across frames [21, 25]. Alternatively, deformable convolutions [6, 23] offer a more adaptive solution, replacing fixed grids with learnable offsets (Fig. 2) that adjust sampling positions based on input features. Despite increased computational cost [4], they provide robustness to motion and spatial transformations, avoiding the limitations and errors of optical flow [14].

However, some models show limited generalization across different codecs. For instance, Multi-Codec STDF [11] found that the best performance was achieved when videos were compressed using the same codecs employed during training. To improve cross-codec performance, Multi-Domain STDF (MD-STDF) [10] was introduced, building on STDF [6] and training on videos encoded with various video codecs. Expanding on this idea, the present work proposes a novel VQE architecture that enhances frame alignment using 3D convolutions, which process spatial and temporal information jointly, unlike 2D convolutions or optical flow methods that treat frames separately. This approach improves motion modeling and temporal coherence, resulting in more robust feature alignment and fusion.

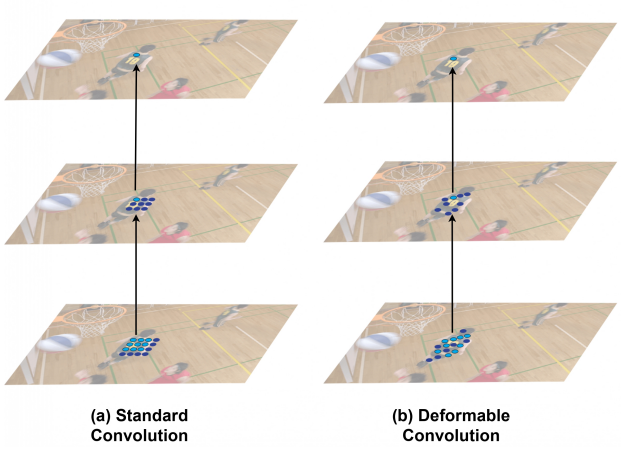

Figure 2: Types of convolution: (a) Fixed 3×3 pixel grid used in standard convolution, where the kernel samples values from a regular and uniform grid across the input feature map; (b) Grid with offset sampling points used in deformable convolution, which adaptively adjusts the sampling locations to better align with the geometric structure of the input.

Additionally, the proposed EnhancedQE module replaces standard flat CNNs with residual blocks, enabling the learning of deeper and more complex representations while preserving spatial and temporal details. The adoption of residual learning helps mitigate the vanishing gradient problem in deeper networks, facilitates more stable and efficient training, and allows the model to focus on learning meaningful residual mappings that refine the decoded frame quality. Furthermore, 3D convolutions, which have proven effective in tasks such as video super-resolution [28] and gesture recognition [29], apply filters jointly across height, width, and time to capture dynamic spatio-temporal patterns. This joint modeling strategy enhances the network's ability to mitigate compression artifacts and ultimately improves the perceptual quality of the reconstructed video.

3 PROPOSED 3D-STDF ARCHITECTURE

The proposed 3D-STDF architecture is based on the STDF architecture introduced by [6], which employs a multi-frame approach to enhance a central frame. The architecture is divided into two main modules: the first is responsible for frame alignment, fusion, and shallow feature extraction, while the second is dedicated to quality enhancement. The model receives as input the central frame to be enhanced (t_0), concatenated with its neighboring frames (t_i for future frames and t_{-i} for past frames). The number of neighboring frames concatenated with the central frame is determined by the Radius parameter (R), which defines how many adjacent frames will be considered. Thus, if R=1, the total number of frames concatenated and input to the model is 3.

As illustrated in Fig. 3, the alignment and fusion module in the 3D-STDF architecture is designed for feature extraction, beginning with a shallow U-Net-based network [17] that predicts the offset field. This U-Net employs stride-1 convolutions to preserve spatial

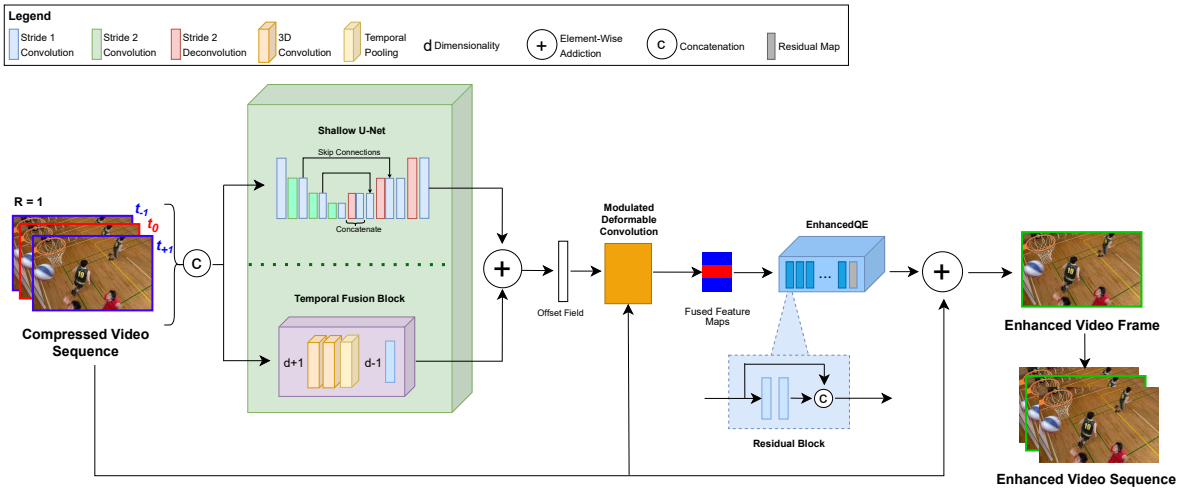

Figure 3: Structure of the proposed 3D-STDF architecture.

resolution, stride-2 convolutions for downsampling, and stride-2 deconvolutions for upsampling. A novel addition to the architecture is the temporal fusion block, which introduces 3D convolutions to explicitly model spatiotemporal dependencies. The input is reshaped to include a temporal dimension and processed by two sequential 3D convolutional layers (3×3×3 kernels, ReLU activations) that jointly capture spatial and temporal features more effectively than 2D convolutions. A subsequent Temporal Pooling layer computes statistics over consecutive frames to reduce the temporal dimension while retaining spatial structure and encoding dynamic patterns.

After temporal pooling, a 2D convolution refines the features spatially. The outputs of the U-Net and temporal fusion block are combined via element-wise summation to produce an offset field mask, which, together with the input frames, is passed to a single modulated deformable convolution layer. This layer adaptively learns offsets and modulation weights, enabling flexible sampling across varying object shapes, sizes, and positions. The resulting fused feature map is processed by the EnhancedQE module, which replaces the standard QE in the original STDF with a more expressive architecture based on residual blocks—each comprising two convolutional layers with ReLU activations and skip connections. EnhancedQE outputs a residual, which is added to the central input frame, enhancing each frame individually and generating a refined video sequence. The network uses a temporal radius of 3 (i.e., 7 input frames: one central, three past, and three future), with each convolutional layer in the U-Net and temporal fusion block using 38 filters, the modulated deformable convolution layer producing 64 output channels, and the EnhancedQE module composed of 8 residual blocks with 42 filters per layer.

4 EXPERIMENTAL RESULTS

This section presents both objective and perceptual evaluations of the proposed VQE method, offering a comprehensive assessment of its effectiveness. The model was trained on a system equipped with an AMD Ryzen 7 5700X CPU, 32 GB RAM, and an Nvidia GeForce RTX 3080 GPU with 12 GB of VRAM. Training was performed using a single GPU with a batch size of 32 and 300.000 iterations, totaling 10 epochs. The MFQE dataset [26], which includes 126 uncompressed videos (108 for training and 18 for testing) with resolutions from 352×240 to 1920×1080, was chosen due to its wide variety of content—including sports, faces, animals, and screen content—as well as diverse lighting conditions, camera angles, and environments. We chose to use only the first 300 frames of each video sequence that has more than 300 frames. Video compression was carried out using the HEVC reference software (HM 16.5) with the *low_delay_p* configuration and a Quantization Parameter (QP) of 37. The training utilized the Adam optimizer (learning rate 0.0001) and the Charbonnier loss function to minimize pixel-level reconstruction errors.

This base configuration required approximately 45 hours of training. A second, more robust variant of the model, named 3D-STDF-L, was also implemented, featuring 16 residual blocks. The training time for this extended model was 58 hours.

4.1 Objective Quality Results

Table 1 shows VQE results for the proposed 3D-STDF model across 18 test video sequences, grouped by resolution classes according to [1]: Class A (2560×1600), B (1920×1080), C (832×480), D (416×240), and E (1280×720). The objective VQE outcomes are expressed as differences (Δ) between the decoded video quality and the quality after processing by the 3D-STDF model. Quality metrics used are Peak Signal to Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) [24], where PSNR quantifies objective error and SSIM evaluates structural similarity, with values ranging from 0 to 1. Positive Δ values indicate quality improvement, while negative values denote degradation.

Tests were performed on videos compressed at various quality levels controlled by the QP, a key parameter in standards like HEVC. QP regulates the quantization of transform coefficients, with lower values yielding higher quality and larger file sizes, and higher values

Table 1: Results of the 3D-STDF model for compressed videos with different QP settings

Configuration	Video Class	Video	ΔPSNR(dB) and ΔSSIM(dB)							
			STDF-R3		MD-STDF		3D-STDF		3D-STDF-L	
			PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
QP 37	Class A	Traffic	0.667	0.012	0.663	0.011	0.722	0.013	0.729	0.013
		PeopleOnStreet	1.108	0.019	1.126	0.020	1.191	0.021	1.242	0.021
	Class B	Kimono	0.801	0.016	0.828	0.016	0.883	0.017	0.906	0.017
		ParkScene	0.538	0.014	0.554	0.013	0.593	0.014	0.599	0.015
		Cactus	0.679	0.013	0.696	0.013	0.754	0.014	0.781	0.014
		BQTerrace	0.558	0.010	0.541	0.009	0.584	0.010	0.624	0.011
		BasketballDrive	0.673	0.012	0.689	0.012	0.727	0.012	0.770	0.013
	Class C	RaceHorses	0.706	0.013	0.694	0.018	0.781	0.020	0.808	0.021
		BQMall	0.831	0.017	0.843	0.017	0.893	0.018	0.920	0.019
		PartyScene	0.617	0.020	0.637	0.019	0.686	0.022	0.679	0.021
		BasketballDrill	0.720	0.014	0.721	0.014	0.793	0.015	0.843	0.016
	Class D	RaceHorses	0.483	0.013	0.454	0.011	0.554	0.014	0.556	0.015
		BQSquare	0.899	0.014	1.023	0.015	0.979	0.016	1.088	0.016
		BlowingBubbles	0.620	0.020	0.632	0.019	0.682	0.021	0.692	0.022
		BasketballPass	0.947	0.019	0.968	0.019	1.038	0.020	1.081	0.021
	Class E	FourPeople	0.940	0.011	0.907	0.011	1.002	0.012	1.022	0.012
		Johnny	0.832	0.007	0.797	0.007	0.875	0.008	0.941	0.009
		KristenAndSara	0.984	0.009	0.970	0.009	1.054	0.010	1.067	0.010
	Average		0.756	0.014	0.764	0.014	0.827	0.015	0.853	0.016
QP 32	**Average**		0.525	0.007	0.399	0.004	0.477	0.006	0.446	0.007
QP 42	**Average**		0.595	0.017	0.662	0.019	0.679	0.019	0.710	0.020
QP 47	**Average**		0.351	0.015	0.416	0.018	0.406	0.017	0.420	0.017
	Total Average		**0.556**	**0.013**	**0.560**	**0.013**	**0.597**	**0.014**	**0.607**	**0.015**

producing stronger compression at the cost of quality. For benchmarking, besides the proposed 3D-STDF and 3D-STDF-L models, the STDF-R3 model from [6] and the MD-STDF model were also evaluated.

The proposed 3D-STDF and 3D-STDF-L models consistently outperform both the STDF-R3 and MD-STDF models across different compression levels, achieving the highest average improvements in ΔPSNR and ΔSSIM. At the training compression level (QP 37), 3D-STDF reached 0.827 ΔPSNR and 0.015 ΔSSIM, while 3D-STDF-L slightly improved these results to 0.853 ΔPSNR and 0.016 ΔSSIM. For higher QP values (42 and 47), both models maintained superior performance, with 3D-STDF-L achieving up to 0.710 ΔPSNR and 0.020 ΔSSIM at QP 42, and 0.420 ΔPSNR and 0.017 ΔSSIM at QP 47. The overall average results confirm that 3D-STDF and 3D-STDF-L surpass the STDF-R3 model, attaining up to 0.607 ΔPSNR and 0.015 ΔSSIM compared to STDF-R3's 0.556 ΔPSNR and 0.013 ΔSSIM.

4.2 Visual Perception Analysis

Figure 4 presents a comparative visual analysis of patches extracted from two test sequences, BasketballPass and FourPeople, to provide qualitative insights into the performance of models trained using the 3D-STDF architecture. The rows display patches from the cited sequences: first in RAW format, then compressed using HEVC (QP 37), and subsequently enhanced by the STDF-R3, MD-STDF, 3D-STDF, and 3D-STDF-L models.

The STDF-R3 model appears to introduce noticeable smoothing, which may lead to some loss of fine details—such as the court lines near a player's knee in the second column and the clothing folds in the first column. In contrast, the 3D-STDF and 3D-STDF-L models seem to retain finer structural information, with the latter, shown in the third column, suggesting a sharper restoration. Blocking artifacts visible in the STDF output appear less prominent in MD-STDF and are further reduced in the 3D-STDF variants. Additionally, edge definition and curvature, particularly in text regions, seem better preserved in the 3D-STDF and 3D-STDF-L results. Overall, these visual tendencies suggest that the 3D-STDF and 3D-STDF-L models may better preserve perceptual details compared to STDF-R3 and MD-STDF, potentially benefiting from a more effective exploitation of temporal information from higher-quality neighboring frames.

5 CONCLUSION

This paper introduced 3D-STDF, a novel video quality enhancement model based on the STDF architecture, which incorporates 3D convolutions to more effectively capture temporal features across video frames and replaces the standard convolutional structure with a residual block-based enhancement network. Experimental results showed that both 3D-STDF and its variant 3D-STDF-L significantly improved the visual quality of compressed videos, surpassing previous methods in most evaluations by achieving up to 0.607 dB gain in ΔPSNR and 0.015 in ΔSSIM at QP 37, while maintaining strong performance at higher compression levels (QP 42 and 47) with effective detail preservation. As this research is ongoing, future work will involve further analysis using perceptual metrics like VMAF and LPIPS for a more comprehensive assessment of visual quality, computational cost evaluations, and ablation studies to explore the solution's feasibility and adaptability. Additional refinements will target the number and structure of residual blocks, network depth and width, and temporal fusion strategy.

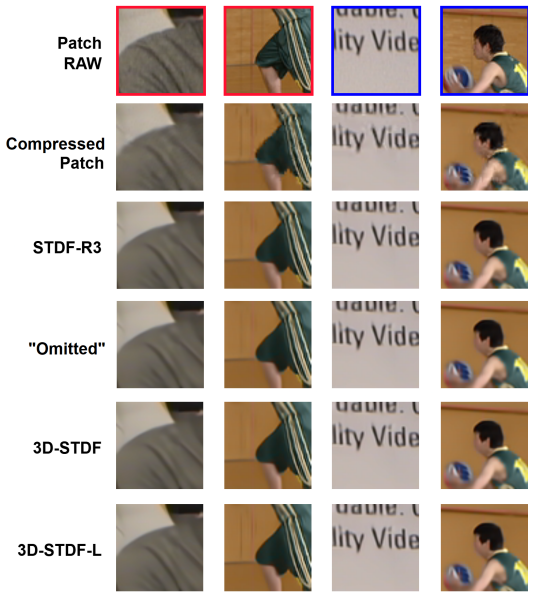

Figure 4: Comparison of results through visual quality perception of images.

6 ACKNOWLEDGEMENTS

This study was financed in part by the *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior* – Brasil (CAPES) – Finance Code 001, Foundation for Research Support of the State of Rio Grande do Sul (FAPERGS), and National Council for Scientific and Technological Development (CNPq).

REFERENCES

[1] J. Boyce, K. Suehring, and X. Li. 2018. JVET-J1010: JVET common test conditions and software reference configurations. *JVET-J1010* (2018).

[2] Jose Caballero, Christian Ledig, Andrew Aitken, Alejandro Acosta, Johannes Totz, Zehan Wang, and Wenzhe Shi. 2017. Real-time video super-resolution with spatio-temporal networks and motion compensation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4778–4787.

[3] H-Y Cheong, Alexis M Tourapis, Joan Llach, and Jill Boyce. 2004. Adaptive spatio-temporal filtering for video denoising. In *2004 International Conference on Image Processing, 2004. ICIP'04.*, Vol. 2. IEEE, 965–968.

[4] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. 2017. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*. 764–773.

[5] Yuanying Dai, Dong Liu, and Feng Wu. 2017. A convolutional neural network approach for post-processing in HEVC intra coding. In *MultiMedia Modeling: 23rd International Conference, MMM 2017, Reykjavik, Iceland, January 4-6, 2017, Proceedings, Part I 23*. Springer, 28–39.

[6] Jianing Deng, Li Wang, Shiliang Pu, and Cheng Zhuo. 2020. Spatio-temporal deformable convolution for compressed video quality enhancement. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 10696–10703.

[7] Chao Dong, Yubin Deng, Chen Change Loy, and Xiaoou Tang. 2015. Compression artifacts reduction by a deep convolutional network. In *Proceedings of the IEEE International Conference on Computer Vision*. 576–584.

[8] Chih-Ming Fu, Elena Alshina, Alexander Alshin, Yu-Wen Huang, Ching-Yeh Chen, Chia-Yang Tsai, Chih-Wei Hsu, Shaw-Min Lei, Jeong-Hoon Park, and Woo-Jin Han. 2012. Sample adaptive offset in the HEVC standard. *IEEE Transactions on Circuits and Systems for Video technology* 22, 12 (2012), 1755–1764.

[9] Zhenyu Guan, Qunliang Xing, Mai Xu, Ren Yang, Tie Liu, and Zulin Wang. 2019. MFQE 2.0: A new approach for multi-frame quality enhancement on compressed video. *IEEE transactions on pattern analysis and machine intelligence* 43, 3 (2019), 949–963.

[10] Garibaldi Silveira Júnior, Gilberto Kreisler, Bruno Zatt, Daniel Palomino, and Guilherme Correa. 2024. Multi-Domain Spatio-Temporal Deformable Fusion model for video quality enhancement. In *Proceedings of the 30th Brazilian Symposium on Multimedia and the Web* (Juiz de Fora/MG). SBC, Porto Alegre, RS, Brasil, 223–230. https://doi.org/10.5753/webmedia.2024.241618

[11] Gilberto Kreisler, Garibaldi da Silveira Junior, Bruno Zatt, Daniel Palomino, and Guilherme Correa. 2023. Modelo Multi-Codec Baseado em Spatio-Temporal Deformable Fusion para Melhoria de Qualidade de Vídeos Comprimidos. In *Anais do L Seminário Integrado de Software e Hardware*. SBC, 143–154.

[12] Shiba Kuanar, Christopher Conly, and KR Rao. 2018. Deep learning based HEVC in-loop filtering for decoder quality enhancement. In *2018 Picture Coding Symposium (PCS)*. IEEE, 164–168.

[13] Tianyi Li, Mai Xu, Ce Zhu, Ren Yang, Zulin Wang, and Zhenyu Guan. 2019. A deep learning approach for multi-frame in-loop filter of HEVC. *IEEE Transactions on Image Processing* 28, 11 (2019), 5663–5678.

[14] Khoi-Nguyen C Mac, Dhiraj Joshi, Raymond A Yeh, Jinjun Xiong, Rogerio S Feris, and Minh N Do. 2019. Learning motion in feature space: Locally-consistent deformable convolution networks for fine-grained action detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 6282–6291.

[15] Mona Mahmoudi and Guillermo Sapiro. 2005. Fast image and video denoising via nonlocal means of similar neighborhoods. *IEEE signal processing letters* 12, 12 (2005), 839–842.

[16] Andrey Norkin, Gisle Bjontegaard, Arild Fuldseth, Matthias Narroschke, Masaru Ikeda, Kenneth Andersson, Minhua Zhou, and Geert Van der Auwera. 2012. HEVC deblocking filter. *IEEE Transactions on Circuits and Systems for Video Technology* 22, 12 (2012), 1746–1754.

[17] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 234–241.

[18] Claudio Rota, Marco Buzzelli, Simone Bianco, and Raimondo Schettini. 2023. Video restoration based on deep learning: a comprehensive survey. *Artificial Intelligence Review* 56, 6 (2023), 5317–5364.

[19] I Sandvine. 2024. Global internet phenomena report.

[20] Gary J Sullivan, Jens-Rainer Ohm, Woo-Jin Han, and Thomas Wiegand. 2012. Overview of the high efficiency video coding (HEVC) standard. *IEEE Transactions on circuits and systems for video technology* 22, 12 (2012), 1649–1668.

[21] Junchao Tong, Xilin Wu, Dandan Ding, Zheng Zhu, and Zoe Liu. 2019. Learning-based multi-frame video quality enhancement. In *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 929–933.

[22] Chia-Yang Tsai, Ching-Yeh Chen, Tomoo Yamakage, In Suk Chong, Yu-Wen Huang, Chih-Ming Fu, Takayuki Itoh, Takashi Watanabe, Takeshi Chujoh, Marta Karczewicz, et al. 2013. Adaptive loop filtering for video coding. *IEEE Journal of Selected Topics in Signal Processing* 7, 6 (2013), 934–945.

[23] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. 2019. Edvr: Video restoration with enhanced deformable convolutional networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*. 0–0.

[24] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* 13, 4 (2004), 600–612.

[25] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. 2019. Video enhancement with task-oriented flow. *International Journal of Computer Vision* 127 (2019), 1106–1125.

[26] Ren Yang, Mai Xu, Tie Liu, Zulin Wang, and Zhenyu Guan. 2018. Enhancing quality for HEVC compressed videos. *IEEE Transactions on Circuits and Systems for Video Technology* 29, 7 (2018), 2039–2054.

[27] Ren Yang, Mai Xu, Zulin Wang, and Tianyi Li. 2018. Multi-frame quality enhancement for compressed video. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6664–6673.

[28] Xinyi Ying, Longguang Wang, Yingqian Wang, Weidong Sheng, Wei An, and Yulan Guo. 2020. Deformable 3d convolution for video super-resolution. *IEEE Signal Processing Letters* 27 (2020), 1500–1504.

[29] Yifan Zhang, Lei Shi, Yi Wu, Ke Cheng, Jian Cheng, and Hanqing Lu. 2020. Gesture recognition based on deep deformable 3D convolutional neural networks. *Pattern Recognition* 107 (2020), 107416.