Multi-Domain Spatio-Temporal Deformable Fusion model for video quality enhancement

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

Gilberto Kreisler
gkfneto@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

Lossy video compression introduces artifacts that can degrade the perceived visual quality of the video. Improving the quality of compressed videos involves mitigating these artifacts through filtering techniques. Deep neural network (DNN) models have emerged as powerful tools for this task, demonstrating effectiveness in artifact reduction. However, traditional approaches typically evaluate these models using videos compressed by a single coding standard, limiting their applicability across diverse codecs. To address this limitation, this study proposes a novel multi-domain architecture built upon the Spatio-Temporal Deformable Fusion technique. This innovative approach enables the development of models capable of enhancing videos compressed by various codecs, ensuring consistent performance across different standards. Experimental results showcase the efficacy of the proposed method, yielding significant improvements in average Peak Signal-to-Noise Ratio (PSNR) for videos compressed with HEVC, VVC, VP9, and AV1, with enhancements of 0.764 dB, 0.448 dB, 0.736 dB, and 0.228 dB, respectively. The code of our MD-STDF approach is available at *https://github.com/Espeto/md-stdf*

KEYWORDS

Redes neurais profundas, Melhoria de qualidade de vídeo, Codificação de vídeo, Aprendizado multi-domínio

1 INTRODUCTION

Video compression plays a crucial role in services dealing with the distribution and storage of audiovisual content, becoming essential for the operation of companies like Netflix, TikTok, and YouTube. Due to the high demand for this type of service, digital video represented the highest volume of data transmitted over the Internet in recent years. It has been predicted that 4K videos represented 66% of Internet consumption on television devices by the end of 2023, surpassing the 2018 estimate [8]. Consequently, research efforts by both academia and industry are dedicated to improving not only compression efficiency but also reducing undesired visual effects caused by this process. As video content continues to proliferate across various platforms and devices, maintaining high visual quality while efficiently managing data transmission and storage becomes increasingly paramount.

Compressed videos suffer from visual effects such as blocking, ringing and blurring artifacts [12], which compromise the perceived video quality for users. In Figure 1, the visual effects of these artifacts can be observed. The patches (c), (d) and (e) are separated by the artifact that predominantly affects the selected part of the image. Generally, artifacts can blend with or appear next to others, making it difficult to generate a patch that isolate only one compression artifact. In (c), the division between blocks used in the compression process is perceptible in the middle of the blue part of the ball, evidencing a blocking effect. In (d), we can see the ringing effect, noticeable as wave-like patterns along the edges of the orange rim. In (e), details from the original image, such as the nails on the ground and the separations in the wooden floor, are lost during compression, leading to a blurring effect.

Filtering algorithms like the Deblocking Filter (DF) [24], addressing blocking effects, Adaptive Loop Filter (ALF) [29], that minimize the distortion between the original and decoded samples, and the Sample Adaptive Offset (SAO) [13], focused on reducing banding effects, are standardized processes in formats such as High Efficiency Video Coding (HEVC) and Versatile Video Coding (VVC).

Both DF and SAO are heuristic-based methods, devised based upon statistical observations for reducing compression artifacts. These models are applied as filters that traverse all pixels in each frame, aiming to enhance visual quality. While these heuristic-based approaches have demonstrated effectiveness in mitigating certain artifacts, they have inherent limitations. For example, DF may introduce a blurring effect on the image as it attempts to reduce blockiness, potentially sacrificing fine details and sharpness. Additionally, both DF and SAO may inadvertently amplify the presence of other compression artifacts, such as ringing or mosquito noise, particularly in regions with high contrast or intricate textures [17]. Despite that, heuristic-based methods remain valuable tools of video compression techniques, especially when used in conjunction with more sophisticated algorithms and Deep Neural Networks (DNN) to achieve comprehensive artifact reduction and enhance overall visual quality.

Currently, a significant amount of studies exploring the Video Quality Enhancement (VQE) problem employ DNN models based on

In: Proceedings of the Brazilian Symposium on Multimedia and the Web (WebMedia'2024). Juiz de Fora, Brazil. Porto Alegre: Brazilian Computer Society, 2024.

ISSN 2966-2753

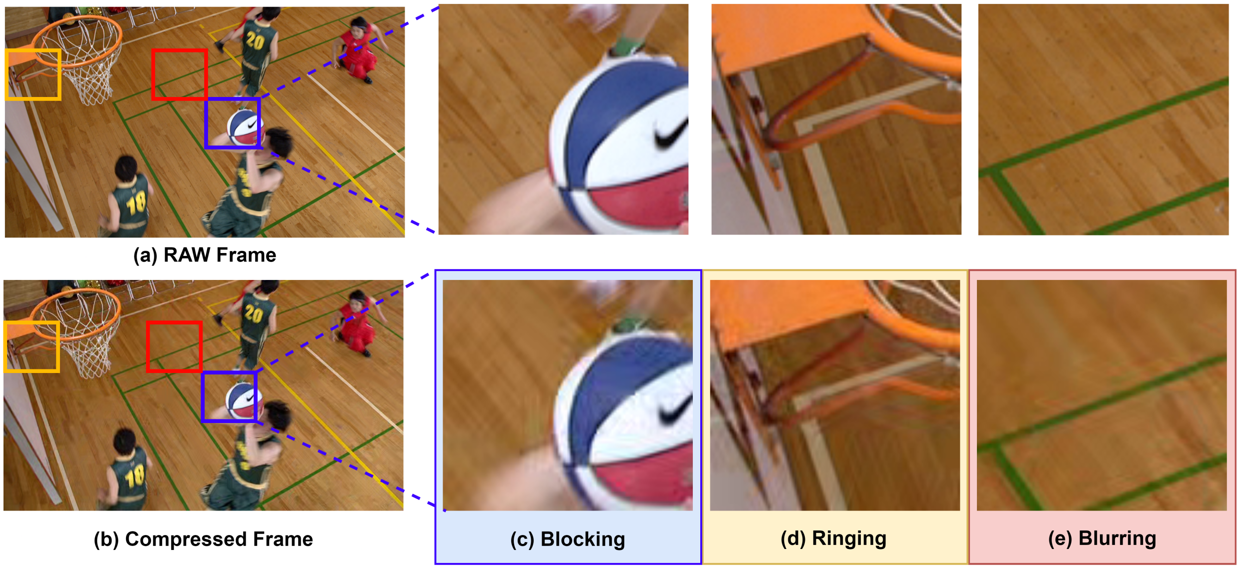

Figure 1: Compression Artifacts: (a) Original frame (RAW); (b) Compressed Frame; (c) Blocking Artifact; (d) Ringing Artifact; (e) Blurring Artifact

Convolutional Neural Networks (CNN) [9-27]. CNNs have emerged as a powerful tool in image and video processing tasks due to their ability to automatically learn hierarchical features from data. Unlike traditional image processing techniques that operate on individual pixels, CNNs operate by convolving learned filters across the input image, enabling them to capture spatial hierarchies and dependencies. By processing local image patches and learning from their correlations, CNNs can effectively extract meaningful features that represent various visual patterns, such as edges, textures, and object shapes. This enables them to understand the contextual connection between neighboring pixels and learn complex patterns in the data [18]. Therefore, CNN-based models are well-suited for VQE tasks as they can identify and address overall image quality degradation caused by compression artifacts rather than focusing solely on specific types of artifacts.

It has been observed that many DNN models for VQE are tested using videos compressed with the same codec and configuration as the training videos. The authors in [15] show that the VQE models tend to produce better results for videos compressed with the same codec and quantization parameter as those used for training. Oppositely, for videos compressed with different codecs and configurations the VQE models lead to little improvement in quality or even quality degradation. Thus, considering the large number of video codecs and encoding configurations available nowadays, it is desirable that the VQE model at the decoder side is generic enough to be used for enhancing videos compressed in different scenarios.

To address this issue, this work proposes the use of a multi-domain training method [6] that allows identifying the video encoding scenario, generating a model that is adaptive to the video codec and its associated compression artifacts. The proposed Multi-Domain Spatio-Temporal Deformable Fusion (MD-STDF) architecture explores multi-domain training for quality improvement of compressed videos, ensuring that a single model is efficient in enhancing the quality of videos originating from any of the domains involved in training. In this work, the videos used in both the training and testing phases have been categorized into domains based on the video codec used for compression: HEVC, VVC, VP9, and AV1.

Experimental results demonstrate that the proposed architecture generates a model that achieves a consistent increase in objective quality for videos compressed with multiple codecs, reaching an average ΔPSNR value of up to 0.764 dB. To the best of the authors' knowledge, this is the first DNN architecture to employ multi-domain training for VQE and that has been trained and tested for multiple video codecs and formats.

This paper is organized as follows. Section 2 presents previous VQE works that are based on deep learning and other multi-domain solutions for other video-related tasks. Section 3 presents the proposed Multi-Domain Spatio-Temporal Deformable Fusion (MD-STDF) architecture. Section 4 presents and discusses the obtained results. Finally, Section 5 concludes this work.

2 RELATED WORK

In recent years, different architectures have been proposed to investigate the VQE problem. In this section, we present the main studies that have contributed to the evolution of learning-based models for VQE. Also, related works focusing on multi-domain learning for video-related problems are discussed.

2.1 Video Quality Enhancement

The problem of VQE emerged with the application of frame-by-frame image processing algorithms in videos. These methods originated from linear algorithms based on heuristics, processing all pixels in the frame, disregarding spatial differences within the same frame, and applying the same equation to the entire image. While this approach effectively addressed degradation issues concentrated in specific parts of the image, it often detrimentally affected other areas that did not share the same problem.

These algorithms have gradually given way to nonlinear machine learning models, which aim to comprehend a zone of pixels and detect their characteristics, thereby determining which heuristic yields superior results in a given scenario. Consequently, the side effects caused by these processes tend to diminish. Moreover, the advent of DNN has accompany in more robust architectures for nonlinear learning, enabling the detection of increasingly intricate characteristics as network depth increases. This evolution has significantly contributed to the development of more effective methods for Video Quality Enhancement.

One of the first solutions for improving video quality based on deep learning is [12], which proposes an Artifact Reduction CNN (ARCNN) that processes each frame individually, exploring only the spatial information within the image. Building on [12], other applications for the model were explored, such as an in-loop filter that replaces DF and SAO filters [10] or as a post-processing filter [16], which performs the filtering after the frames are fully decoded.

Some studies emerged with the proposal to explore the existing temporal correlation between frames. Initially, models based on multiple frames [5, 11, 14, 34] proved effective for the VQE problem. These models define a temporal sliding window that processes a fixed number of frames to improve the central one. This way, the Group of Pictures (GOP) structure present in most of the video coding standards can be explored, allowing information from high-quality frames to be used to improve low-quality frames [28].

models based on multiple frames aim to synchronize past and future frames with the currently processed frame. This alignment process is typically achieved through motion compensation techniques, such as optical flow estimation [28, 31]. Optical flow helps in determining the motion vectors between consecutive frames, facilitating the alignment process and enabling the model to understand the temporal relationships between frames.

Following alignment, the fusion of processed frames occurs, with the objective of incorporating the best quality characteristics from each frame. Two common fusion approaches include direct fusion and slow fusion. In direct fusion, all frames are fused simultaneously, leveraging the information from each frame to enhance the overall quality. On the other hand, slow fusion involves iteratively fusing pairs of frames until only one frame remains, gradually synthesizing the final enhanced frame [20].

This alignment and fusion strategy allows models to exploit temporal information across multiple frames, leading to more comprehensive and effective video quality enhancement. By aligning frames and fusing their features judiciously, these models can better capture and preserve the temporal coherence and visual details present in the video sequence.

An alternative to the optical flow-based fusion is the use of deformable convolutions [11, 30]. This mechanism replaces the traditional convolution used by CNN layers with a deformable convolution, where, instead of using a fixed conventional matrix as a filter, a matrix with variably displaced points is used, learning the process of displacing neighboring pixels relative to the processed pixel through information obtained from previous convolutions.

Another strategy employed to harness temporal information in video processing involves the utilization of recurrent models, particularly Recurrent Neural Networks (RNNs). RNNs are designed to handle sequential data by processing input sequences one step at a time, maintaining a hidden state that captures information from previous steps [19]. In the context of video quality enhancement, RNNs operate by progressively analyzing each frame of the video sequence. As each frame is processed, its characteristics are extracted and incorporated into the hidden state of the network. This hidden state serves as a condensed representation of the temporal information extracted from the video sequence so far, enabling the network to understand the temporal dependencies and patterns present in the data. Subsequently, this aggregated temporal information can be utilized to enhance the quality of future frames through filtering or other enhancement techniques. Studies explore recurrent networks in different ways, either in a unidirectional manner [22], where the characteristics are propagated from past frames to the currently processed frame, or using a bidirectional improvement [35], where both features from past and future frames are used to improve the quality of the current frame.

2.2 Multi-Domain Learning

The advancement of machine learning algorithms has been made possible due to the abundance of available data. However, the current training paradigms are limited in the variety of data they can handle. Most methods operate on data from specific domains, causing the model to learn the inherent bias of the dataset. As a result, the models perform well on tasks specific to the domains they were trained on, severely limiting their ability to generalize when processing data from new and previously unseen domains. These challenges become more pronounced when dealing with data from highly variable domains, especially when there is a need to develop a single model capable of handling multiple distinct datasets. Training a single model to encompass this diversity of domains prevents the capture of specific nuances from each one. The common approach to address this issue involves recreating a model for each domain and applying each model to the corresponding data. However, this methodology proves to be highly inefficient

Videos exhibit characteristics with a high variability, making it challenging for a model to generate a unique representation that can capture them all. Therefore, the Multi-Domain Network (MDNet) [23] emerged with the proposal of separating videos into annotated domains, so that each domain follows a distinct path in the final layers of the network. In this way, common features among all domains are extracted in the initial layers of the network, while features specific to the domain are extracted in the final layers. Following the same proposal as the previous study, the Branch-Activated Multi-Domain Convolutional Neural Network for Visual Tracking (BAMCNN) [6] was developed. It leverages the concept of multi-domain training to create an architecture for visual tracking, separating videos into different domains based on their similarities. This involves creating branches in the network for each domain, thus detecting their specific characteristics. During testing, the branch with the highest level of similarity to the processed video sequence is identified and activated.

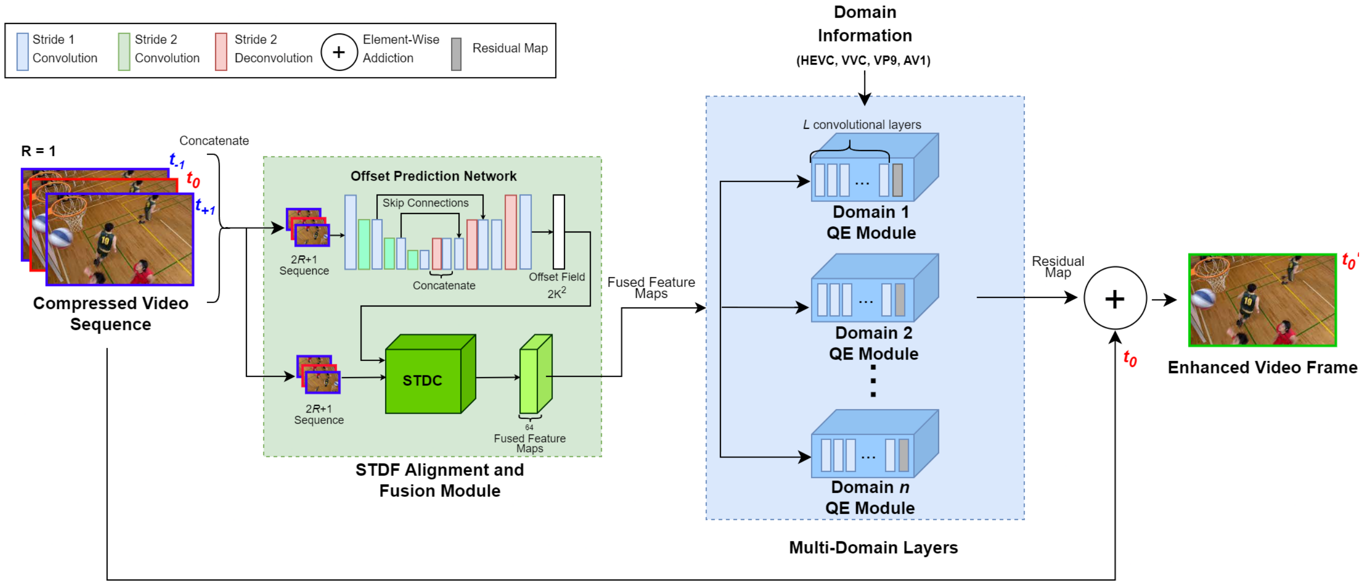

Figure 2: The MD-STDF architecture consists of the STDF Alignment and Fusion Module, that is shared by all domains and *n* branches of QE multi-domain. The red-border frame represents the central one, under processing, whereas blue-border neighboring frames are aligned and used in the enhancement of the central frame.

Recently, other studies have employed multi-domain learning for different video-related problems. In [25], the authors addressed the quality enhancement of multiview video coding. The work [1] explores the detection of deepfake videos by combining features from two distinct domains, spatial and frequency, in a discriminative learning model.

This work aims to explore VQE under a generalized approach, seeking to create a model that is not limited to improving the quality of videos in a specific domain but is generic enough to enhance videos belonging to other domains. No studies were found that addressed the multi-domain training method from such a perspective.

3 MULTI-DOMAIN STDF

The proposed Multi-Domain Spatio-Temporal Deformable Fusion (MD-STDF) architecture is based on the STDF architecture [11], which employs the approach of using multiple frames to enhance a central frame. Additionally, the architecture applies the multi-domain training strategy, with training data originated from different domains. The domain information is also incorporated in the dataset to allow the model to learn the domain-specific characteristics along the training.

The STDF architecture [11] is divided into two modules, with the first one focusing on frame alignment and fusion, as well as shallow feature extraction. The second module is dedicated to quality enhancement. The model receives as input the central frame to be improved (t_0), concatenated with temporally neighboring frames (t_1 for the future frames and t_{-1} for past frames). The number of neighboring frames to be concatenated with the central frame is determined by the Radius (R) parameter, which represents the number of neighboring frames. Thus, with R defined as 1, the total number of frames to be concatenated and introduced into the model is 3.

As depicted in Figure 2, in the development of MD-STDF, the alignment and fusion module is dedicated to obtaining general features — i.e., those that are common to all videos regardless of the domain. The network starts with a shallow U-Net Model [26] that extracts the features for offset prediction. This model is based on Stride 1 convolutions, which mantain the sample dimensionality, Stride 2 convolutions, which reduce sample dimensionality, and Stride 2 deconvolutions, which perform an upsampling. This part of the network captures the temporal characteristics of the sequence and generates an offset field mask. The kernel size of the offset mask is $2K^2$. This mask is used with the input 2R+1 sequence in the Spatio-Temporal Deformable Convolution (STDC) network, generating a Fused Feature Map with 64 channels.

Instead of calculate the diference between two frames, that is the method used in optical flow, in studies like [14], the STDC modules uses a modulated deformable convolution layer [36], this layer realize the motion compensation of the entire input sequence at once. As an alternative of using a fixed grid of positions to apply the convolutional kernels (as in conventional convolutions), deformable convolution introduces additional offsets that are learned by the network during training. These offsets allow the convolutional kernel to 'deform' to better align with the input patterns, resulting in improved capture of spatial and temporal variations [9].

The multi-domain training is based on a label linking the video to a specific domain (codec), which is employed in the transition between the alignment and fusion module and the quality enhancement (QE) module. The branches created for each domain ensure that the QE module is updated with parameters specific to the codec.

The activated branch delivers the Fused Feature Map to the corresponding QE module according to the video label. The QE Modules are composed by L convolutional layers with Stride 1, maintain the dimensionality of the samples and extract domain-related features. At last, the QE module generates a Residual Map, which is added to the central frame (t_0), generating an enhanced frame (t_0'). The procedure is repeated for each frame of the compressed video, generating an enhanced video sequence.

3.1 Data Preparation

A survey of the main datasets used for VQE was conducted. This survey did not include image datasets, only videos. Additionally, datasets with a specific purpose, such as screen content videos or sports videos only, were not included, but rather those that cover a wide category range. Among the possible choices for datasets are the MFQE dataset [14], the Large-scale Diverse Video (LDV) dataset used in the challenge proposed at the New Trends in Image Restoration and Enhancement workshop and challenges on image and video processing (NTIRE) in 2021 [32], and the Vimeo-90K dataset [31].

Since this work was not restricted to a specific video category, the videos from the chosen dataset encompass the broadest spectrum of possible categories (sports, screen content, face videos, animals, etc.), as well as differences related to brightness, video environment, and camera angle. The MFQE dataset [33] was chosen, which contains 126 uncompressed videos (108 for training and 18 for testing) at different resolutions ranging from 352 × 240 to 1920 × 1080.

The video sequences were organized according to the standards used for compression, in order to split the dataset for the multi-domain training. Four versions of the training dataset were generated, each corresponding to the addressed domains. Thus, the 108 videos were encoded and decoded using the reference software for the four standards/formats, resulting in a total of 432 videos. The four distinct domains defined correspond to the High Efficiency Video Coding (HEVC) standard [27], the Versatile Video Coding (VVC) standard [4], the AOMedia Video 1 (AV1) format [7], and the VP9 [21] format.

For HEVC encoding, the reference software HEVC Model (HM) version 16.5 was used, and for VVC, the VVC Test Model (VTM) version 13.0 was employed, both configured with *Low Delay P* temporal setting. For AV1 encoding, the reference software *libaom*, hashcode3.3 was used, whereas for VP9 the *libvpx*, hashcode1.12.0, was used. For HEVC and VVC, the quantization parameter (QP) was set to 37, while for VP9 and AV1, constant quality (CQ) parameter was set to 55.According to the [2] analysis, there is no correlation between the QP and CP values that indicates the same quality loss is occurring in both encoders. Thus, CQ and QP values were chosen as those that lead to the highest level of degradation according to the recommended test conditions of HEVC/VVC and VP9/AV1. The QP/CQ controls the level of quantization applied to the transforms of video data blocks. A high QP/CQ value during compression causes information to be discarded during the quantization process, leading to a loss of fine details and, consequently, lower image quality. A lower CQ value means that less quantization is applied, resulting in superior quality.

3.2 Training Process

For training, a computer with the following configuration was used: AMD Ryzen 7 5700X processor, 32 GB RAM, Nvidia Geforce RTX 3070 GPU with 8 GB VRAM. The batch size and the number of iterations were adjusted to achieve 10 epochs with one GPU (i.e., a batch size of 32 and 1,200,000 iterations over the dataset). The training was conducted using Adam as optimizer and a learning rate of 0.0001.

The algorithm employed was Stochastic Gradient Descent (SGD) [23], where each training iteration was executed under a specific domain. In other words, the batch of videos used belongs to a single domain, activating only one branch of the network.

After a certain number of iterations, the model is updated based solely on the processed batch. Subsequently, a new batch from another domain is processed, causing the shared layers of the network to be updated while keeping the previously updated branch unchanged. This process is repeated until the predefined number of iterations is reached. Following this approach, the generic features common to all processed videos are obtained in the shared layers of the network, while for each specific branch of each domain, modeling is done to acquire domain-specific characteristics.

4 EXPERIMENTAL RESULTS

The obtained results offer a comprehensive view of the conducted study and aid in assessing the effectiveness of the adopted multi-domain training methodology. The results are presented in terms of Delta Peak Signal-to-Noise Ratio (ΔPSNR), which measures the objective difference between the enhanced and the low-quality decoded video. Positive numbers indicate and increase in objective quality, whereas negative numbers indicate a quality decrease.

For comparison purposes, three single-codec STDF models were also trained following the methodology presented in [11]: the first with a dataset containing only videos compressed with the HEVC codec; the second with videos compressed with the VVC codec; and the third with videos compressed with the AV1 codec. Additionally, the multi-codec, single-domain approach proposed in [15] is also presented. The same encoder configurations and training setup mentioned in the previous section were used.

Table 1 shows the comparison of different models trained for comparison purposes. The *Training Dataset* column represents the dataset used to train each model. The subsequent columns represent the test dataset used to obtain the ΔPSNR values. The first three rows (HEVC, VVC and AV1) present the VQE results obtained from training models using the single-codec approach. In the fourth row, the results obtained from training using the multi-codec approach [15] are shown. Finally, the last row presents the results obtained with the model trained using the proposed multi-domain method.

As observed, the model trained with videos compressed with HEVC achieves the best results (0.755 dB) when tested with videos encoded with the same standard. However, for videos encoded with AV1 standard, the model yields a negative result for VQE (-0.506 dB). A similar trend is observed for the model trained with videos compressed with VVC, which performs poorly for AV1-compressed videos. The multi-codec model proposed in [15] presents more constant results (varying between 0.210 dB and 0.375 dB), but does not perform as well as the single-codec models. Finally, the proposed

Table 1: Comparison between single-codec, multi-codec and multi-domain approaches.

STDF Model	ΔPSNR (dB)			
	HEVC	VVC	VP9	AV1
	QP 37	QP 37	CQ 55	CQ 55
HEVC QP 37 [11]	0.755	0.250	0.357	-0.506
VVC QP 37	0.529	0.371	0.385	-0.016
AV1 CQ 55	0.285	0.144	0.389	0.286
Multi-Codec [15]	0.335	0.210	0.375	0.229
Multi-Domain	**0.764**	**0.448**	**0.736**	**0.228**

Table 2: VQE results for the MD-STDF model (ΔPSNR and ΔSSIM).

Test Dataset		ΔPSNR (dB) and ΔSSIM							
		HEVC QP 37		VVC QP 37		VP9 CQ 55		AV1 CQ 55	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Class A	*Traffic*	0.662	0.011	0.420	0.006	0.727	0.009	0.120	0.003
	People on Street	1.126	0.020	0.582	0.009	0.911	0.014	0.200	0.004
Class B	*Kimono*	0.831	0.016	0.547	0.008	0.462	0.007	0.184	0.004
	ParkScene	0.551	0.013	0.426	0.011	0.487	0.008	0.152	0.005
	Cactus	0.698	0.013	0.404	0.007	0.641	0.010	0.130	0.003
	BQTerrace	0.543	0.009	0.217	0.005	0.402	0.006	-0.024	0.001
	BasketballDrive	0.686	0.012	0.292	0.005	0.552	0.008	0.156	0.004
Class C	*RaceHorses*	0.447	0.011	0.220	0.005	0.473	0.011	0.153	0.003
	BQMall	0.838	0.017	0.529	0.009	0.828	0.012	0.231	0.004
	PartyScene	0.640	0.019	0.380	0.011	0.731	0.014	0.197	0.005
	BasketballDrill	0.718	0.015	0.290	0.005	0.733	0.014	0.197	0.004
Class D	*RaceHorses*	0.691	0.018	0.436	0.011	0.661	0.015	0.344	0.008
	BQSquare	1.020	0.015	0.667	0.009	1,234	0.010	0.488	0.004
	BlowingBubbles	0.635	0.020	0.453	0.015	0.702	0.015	0.328	0.008
	BasketballPass	0.971	0.019	0.716	0.015	0.860	0.015	0.496	0.008
Class E	*FourPeople*	0.913	0.012	0.548	0.006	1.046	0.008	0.325	0.002
	Johnny	0.804	0.008	0.414	0.003	0.843	0.006	0.164	0.001
	KristenAndSara	0.969	0.009	0.515	0.004	0.954	0.006	0.270	0.002
Average		**0.764**	**0.014**	**0.448**	**0.008**	**0.736**	**0.010**	**0.228**	**0.004**

multi-domain model is the one that presents the best results in all cases (varying between 0.228 dB for AV1 and 0.764 dB for HEVC), performing even better than the single-codec models trained specifically for each standard.

Table 2 presents the results of objective quality variation for each video sequence. The results are presented in terms of ΔPSNR and ΔSSIM for each video sequence, with the last row showing the average results. The videos are grouped according to their Classs [3]: Class A (2560x1600), Class B (1920x1080), Class C (832x480), Class D (416x240), Class E (1280x720). Due to the good results presented in terms of ΔPSNR, it was decided to add the SSIM metric to complement the analysis of the results.

As can be observed in the table, most of the results are positive, with only one specific case showing a negative ΔPSNR value. However, in terms of ΔSSIM, the result remained positive. On average, all the results were positive. The worst result achieved, in terms of ΔPSNR, was -0.024 dB for AV1 in the *BQTerrace* sequence, but in terms of ΔSSIM, the video showed a slight improvement. This is also the worst result in terms of ΔSSIM. The best result in ΔPSNR was 1.234 dB for VP9 in the *BQSquare* sequence; in ΔSSIM, it was 0.020 for HEVC in the *People on Street* sequence. For average results of ΔPSNR, the worst improvement achieved was 0.228 dB for videos encoded with AV1, and the best improvement was 0.764 dB for videos encoded with HEVC. In terms of average ΔSSIM, the worst improvement was 0.004 for videos encoded with AV1, and the best was 0.014 for videos encoded with HEVC. Some specific cases, in addition to the best case, showed results in terms of ΔPSNR above 1 dB, such as the *People on Street* sequence encoded by HEVC with QP 37, which achieved a result of 1.126 dB; the *BQSquare* sequence encoded by HEVC with QP 37 achieved a result of 1.020 dB; and the *FourPeople* sequence encoded by VP9 with CQ 55 achieved a result of 1.046 dB.

5 VISUAL QUALITY PERCEPTION

This section presents an analysis of the perceived visual quality improvement of the multi-domain STDF solution. For this section, the selected frame of video sequence was the one in which the

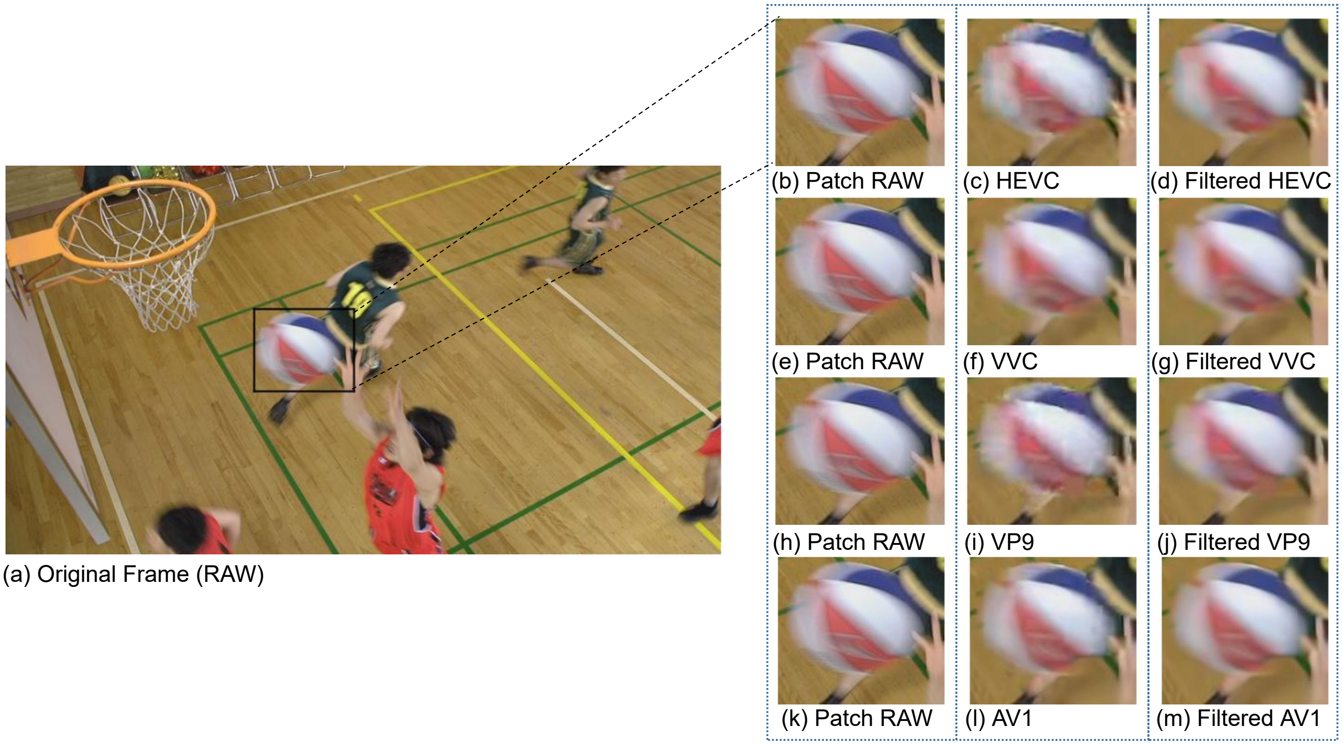

Figure 3: Frame number 14 of the BasketballDrill sequence: (a) Original frame (RAW); (b)(e)(h)(k) Cropped section of the original frame; (c) HEVC compressed version; (d) Improved HEVC version; (f) VVC compressed version; (g) Improved VVC version; (i) VP9 compressed version; (j) Improved VP9 version; (l) AV1 compressed version; (m) Improved AV1 version.

greatest difference in visual quality was observed after a series of visual analyses.

The Figure 3 presents a composition based on frame number 14 of the *BasketballDrill* sequence. Observing the images in the second column, i.e., the images that went through the compression process, most of them exhibit a significant number of artifacts. The most deteriorated images are (c) and (i), which correspond to the HEVC and VP9 codecs, respectively. In image (l), which corresponds to the AV1 codec, some artifacts are still noticeable, although in smaller quantities. On the other hand, image (f), corresponding to the VVC codec, has the fewest artifacts.

In all the compressed images, the blurring effect is noticeable, especially in the background. The blocking effect can be observed in images (c), (i), and (l). In image (c), this artifact creates a stair-step effect on the edges of the ball. In image (i), it creates a mosaic effect. In image (l), the stair-step effect is also noticeable on the upper right edge of the ball.

Comparing the low-quality versions in the second column with their respective enhanced versions in the third column, it is evident that images (d) and (j) show a significant improvement in visual quality. Image (m) also demonstrates a smoothing of artifacts. Image (g), on the other hand, is the one that most resembles its compressed version before the application of the enhancement filter, since image (f) exhibits an extremely low incidence of compression artifacts.

6 CONCLUSION

This work proposed a novel architecture named Multi-Domain Spatio-Temporal Deformable Fusion (MD-STDF), which employs multi-domain learning to enhance the quality of videos compressed with different codecs, overcoming limitations of previous approaches. The model was trained with videos generated by multiple video codecs, thus learning characteristics of different types and levels of compression artifacts more effectively. In the conducted experiments, MD-STDF achieved promising results by providing significant improvements in VQE for videos compressed with HEVC, VVC, AV1, and VP9. The multi-domain approach proved superior to single-codec and single-domain techniques, consistently yielding gains across all tests. On average, the quality improvement in terms of ΔPSNR ranged between 0.228 dB (AV1) and 0.764 dB (HEVC),indicating that the model achieves a strong generalization capability for different video compression scenarios. This can be explained by the larger number of training samples, which may have led to a greater refinement of the alignment and fusion network, which is the network shared among all domains. The good results from the objective analysis of the multi-domain model are also reflected in the perception of subjective quality improvement. As shown in the visual quality analysis, the multi-domain model was able to satisfactorily remove the artifacts present in the frames, smoothing the images. Some details cannot be restored due to the lossy nature of the compression process; however, overall, the quality improvement in the frames is noticeable. Knowing that PSNR does not have a strong correlation with subjective video quality, it is intended that for future work, other evaluation metrics such as VMAF and LPIPS will be adopted. Furthermore, the authors intend to explore training models with an even more diverse set of codecs and QP/CQ configurations, aiming for increased generalization and effectiveness across different scenarios. Additionally, it is also intended to conduct an ablation study and explore cost reduction techniques.

7 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] Aayushi Agarwal, Akshay Agarwal, Sayan Sinha, Mayank Vatsa, and Richa Singh. 2021. MD-CSDNetwork: Multi-domain cross stitched network for deepfake detection. In *2021 16th IEEE international conference on automatic face and gesture recognition (FG 2021)*. IEEE, 1–8.

[2] Isis Bender, Daniel Palomino, Luciano Agostini, Guilherme Correa, and Marcelo Porto. 2019. Compression efficiency and computational cost comparison between AV1 and HEVC encoders. In *2019 27th European Signal Processing Conference (EUSIPCO)*. IEEE, 1–5.

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

[4] Benjamin Bross, Ye-Kui Wang, Yan Ye, Shan Liu, Jianle Chen, Gary J Sullivan, and Jens-Rainer Ohm. 2021. Overview of the versatile video coding (VVC) standard and its applications. *IEEE Transactions on Circuits and Systems for Video Technology* 31, 10 (2021), 3736–3764.

[5] 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.

[6] Yimin Chen, Rongrong Lu, Yibo Zou, and Yanhui Zhang. 2018. Branch-Activated Multi-Domain Convolutional Neural Network for Visual Tracking. *Journal of Shanghai Jiaotong University (Science)* 23 (2018), 360–367.

[7] Yue Chen, Debargha Murherjee, Jingning Han, Adrian Grange, Yaowu Xu, Zoe Liu, Sarah Parker, Cheng Chen, Hui Su, Urvang Joshi, et al. 2018. An overview of core coding tools in the AV1 video codec. In *2018 picture coding symposium (PCS)*. IEEE, 41–45.

[8] V Cisco. 2020. Cisco visual networking index: Forecast and trends, 2018–2023. *White Paper* 1 (2020).

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] Zewen Li, Fan Liu, Wenjie Yang, Shouheng Peng, and Jun Zhou. 2021. A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE transactions on neural networks and learning systems* 33, 12 (2021), 6999–7019.

[19] Ming Liang and Xiaolin Hu. 2015. Recurrent convolutional neural network for object recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3367–3375.

[20] Xiandong Meng, Xuan Deng, Shuyuan Zhu, and Bing Zeng. 2019. Enhancing quality for VVC compressed videos by jointly exploiting spatial details and temporal structure. In *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 1193–1197.

[21] Debargha Mukherjee, Jim Bankoski, Adrian Grange, Jingning Han, John Koleszar, Paul Wilkins, Yaowu Xu, and Ronald Bultje. 2013. The latest open-source video codec VP9-an overview and preliminary results. In *2013 Picture Coding Symposium (PCS)*. IEEE, 390–393.

[22] Seungjun Nah, Sanghyun Son, and Kyoung Mu Lee. 2019. Recurrent neural networks with intra-frame iterations for video deblurring. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8102–8111.

[23] Hyeonseob Nam and Bohyung Han. 2016. Learning multi-domain convolutional neural networks for visual tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4293–4302.

[24] 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.

[25] Bo Peng, Renjie Chang, Zhaoqing Pan, Ge Li, Nam Ling, and Jianjun Lei. 2022. Deep in-loop filtering via multi-domain correlation learning and partition constraint for multiview video coding. *IEEE Transactions on Circuits and Systems for Video Technology* 33, 4 (2022), 1911–1921.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] Ren Yang and Radu Timofte. 2021. NTIRE 2021 Challenge on Quality Enhancement of Compressed Video: Dataset and Study. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*.

[33] 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.

[34] 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.

[35] Chao Zhu, Hang Dong, Jinshan Pan, Boyang Liang, Yuhao Huang, Lean Fu, and Fei Wang. 2022. Deep recurrent neural network with multi-scale bi-directional propagation for video deblurring. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 36. 3598–3607.

[36] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. 2019. Deformable convnets v2: More deformable, better results. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9308–9316.