Modelo Multi-Codec Baseado em *Spatio-Temporal Deformable Fusion* para Melhoria de Qualidade de Vídeos Comprimidos

Gilberto Kreisler, Garibaldi da Silveira Junior, Bruno Zatt, Daniel Palomino, Guilherme Correa

[1]Video Technology Research Group (ViTech) - Universidade Federal de Pelotas (UFPel)
Rua Gomes Carneiro, 1 - Pelotas/RS - Brasil - 96010-610

{gkfneto,garibaldi.dsj,zatt,dpalomino,gcorrea}@inf.ufpel.edu.br

Abstract. *Compressed videos often suffer from visual effects that decrease the quality perceived by the user. Currently, different deep learning architectures have been shown to be efficient for the problem of quality improvement in videos. However, most of them are trained and validated using videos generated by a single video encoding standard. This paper proposes a new model based on the Spatio-Temporal Deformable Fusion (STDF) architecture, providing quality gains for videos compressed by different standards. The results demonstrate that when considering different standards and video encoding settings in model training, a significant increase in quality improvement is achieved, with an average PSNR increment of up to 0.382 dB.*

Resumo. *Vídeos comprimidos geralmente sofrem com efeitos visuais que prejudicam a qualidade percebida pelo usuário. Atualmente, diferentes arquiteturas de aprendizado profundo têm se mostrado eficientes para o problema de melhoria de qualidade em vídeos. No entanto, a maioria delas é treinada e validada usando vídeos gerados por um único padrão de codificação de vídeo. Este artigo propõe um novo modelo baseado na arquitetura Spatio-Temporal Deformable Fusion (STDF), proporcionando ganhos de qualidade para vídeos comprimidos por diferentes padrões. Os resultados demonstram que ao considerar diferentes padrões e configurações de codificação de vídeo no treinamento do modelo, um aumento significativo na melhoria de qualidade é alcançado, com um incremento médio de PSNR de até 0,382 dB.*

1. Introdução

O volume de dados provenientes de vídeos digitais tem crescido cada vez mais na internet. Com a pandemia de COVID-19, empresas com atividades relacionadas a vídeos digitais como TikTok, NetFlix e YouTube tiveram um grande aumento de demanda. Só durante o primeiro mês da pandemia, o volume de dados proveniente da transmissão de vídeos pela internet aumentou em 32,6% [Statista 2022]. Além disso, vídeos de alta resolução, como *Ultra-High Definition* (UHD) 4K e 8K, têm se tornado cada vez mais comuns. De acordo com [Cisco 2020], até o fim do ano de 2023, vídeos em 4K representarão 66% do consumo de internet por aparelhos de televisão, sendo este percentual maior do que o previsto em 2018 (33%).

Desta forma, tendo em vista o alto consumo deste tipo de mídia pela população, profissionais da indústria e academia constantemente investem recursos e pesquisa no desenvolvimento de soluções de compressão. Estas soluções são essenciais para que vídeos

em alta qualidade possam ser armazenados e transmitidos em uma largura de banda limitada. Para isso, existem dois grupos de métodos de compressão: os métodos sem perda e os métodos com perda. Nos métodos sem perda, é possível recuperar o arquivo exatamente como era antes da compressão. Já nos métodos com perda, não é possível que o vídeo seja recuperado conforme sua representação original; porém, estes métodos atingem maiores taxas de compressão quando comparados aos métodos sem perdas. Portanto, se a perda for aceitável para o processo, métodos com perda são os mais indicados.

Em processos de compressão com perda, a redução de qualidade ocorre devido à etapa de quantização. Nesta etapa, os coeficientes transformados passam por uma divisão inteira, de forma que os coeficientes de mais alta frequência acabam por ser zerados e não podem ser reconstruídos no processo de decodificação. Este processo acaba por introduzir artefatos indesejados ao usuário, também conhecidos como artefatos de compressão (AC) [Dong et al. 2015].

Padrões de codificação de vídeo como o *High Efficiency Video Coding* (HEVC) e o *Versatile Video Coding* (VVC), bem como o formato livre de royalties *AOMedia Video 1* (AV1), adotam em seu processo de codificação filtros que reduzem esses efeitos visuais causados por artefatos de compressão específicos, como o filtro de deblocagem, que alivia os efeitos de bloco, e o *Sample Adaptive Offset* (SAO), focado em reduzir os efeitos de banda. Neste processo, segundo [Li et al. 2019], são utilizadas estratégias baseadas em heurísticas. Este tipo de estratégia utiliza um conhecimento prévio do domínio de codificação, onde um modelo estatístico é determinado para a criação de um filtro para a melhoria de qualidade visual. O problema deste tipo de filtro é que mesmo algoritmos considerados estado da arte, como o de [Foi et al. 2007], tendem a produzir resultados com excesso de borrão (*blur*) e bordas pouco definidas, impactando negativamente na qualidade visual percebida pelo usuário.

Diferentemente de filtros tradicionais, que focam em AC específicos, arquiteturas de Redes Neurais Profundas (*Deep Neural Networks* – DNN) baseadas em Redes Neurais Convolucionais (*Convolutional Neural Networks* – CNN) ao invés de processar uma imagem pixel a pixel, correlacionam pixels vizinhos para compreender o contexto de ligação entre eles, aprendendo formas. Quando este tipo de rede é utilizado no contexto de melhoria de qualidade visual, percebe-se a perda de qualidade da imagem como um todo ao invés de focar em um AC específico, o que permite que não ocorra a incidência de novos AC com o processamento da imagem.

Os modelos de DNN para melhoria de qualidade de vídeo (*Video Quality Enhancement* – VQE) podem ser aplicados de duas diferentes formas em relação ao padrão de codificação de vídeo. São considerados *in-loop* aqueles que estão vinculados a arquitetura do padrão. Este tipo de modelo promove a melhoria do quadro antes de ser armazenado como referência futura para os próximos quadros. Quando o modelo não está vinculado a um padrão de codificação de vídeo, o filtro é considerado de pós-processamento, em que a melhoria ocorre com a imagem já decodificada, sendo o quadro melhorado exibido diretamente ao usuário sem retroalimentar o codificador [Li et al. 2019]. Dessa forma, este tipo de filtro não tem vínculo com nenhuma ferramenta de codificação específica e tecnicamente poderia ser utilizada como ferramenta de pós-processamento de qualquer padrão/formato.

Grande parte das arquiteturas propostas para o problema de VQE utiliza apenas vídeos comprimidos conforme o padrão HEVC, tanto para treinamento como teste, não analisando a efetividade dos modelos gerados para a melhoria de vídeos comprimidos por outros padrões de codificação. Este é o caso da arquitetura *Spatio-Temporal Deformable Fusion* (STDF), proposta por [Deng et al. 2020], que adota um alinhamento de características utilizando convoluções deformáveis aplicadas a múltiplos quadros ao invés do típico processo de estimação e compensação de movimento adotado por outras arquiteturas, como a *Multi Frame Quality Enhancement* (MFQE) de [Yang et al. 2018a].

Neste artigo, apresentamos um novo modelo baseado na arquitetura STDF [Deng et al. 2020]. Esta proposta, denominada multi-codec, é realizada usando um *dataset* misto composto por vídeos comprimidos pelos codecs VVC e AV1. Resultados experimentais mostram que o modelo proposto alcança um aumento de qualidade objetiva consistente para vídeos comprimidos com múltiplos codecs, atingindo um valor de ΔPSNR até 0,382 dB. No conhecimento dos autores, este é o primeiro modelo de melhoria de qualidade de vídeos treinado e testado para múltiplos padrões e formatos de codificação de vídeo.

2. Trabalhos Relacionados

A primeira arquitetura baseada em CNN específica para a redução de AC em imagens foi a *Artifact Reduction Convolutional Neural Network* (ARCNN) [Dong et al. 2015]. Alguns estudos posteriores utilizaram essa arquitetura como base, como o de [Dai et al. 2017], um dos primeiros estudos a abordar CNN para o problema de VQE, em que foi desenvolvida a arquitetura *Variable-Filter-Size Residue-Learning* (VRCNN), composta de um modelo de CNN de quatro camadas. Na proposta, o modelo é utilizado como um filtro *in-loop* e substitui os filtros SAO e de deblocagem do padrão HEVC. Tomando como base a ARCNN, [Kuanar et al. 2018] desenvolveu a *Multiview Deep Convolutional Neural Network* (MDCNN), uma arquitetura de 9 camadas de CNN, aplicada somente no lado do decodificador, como um filtro de pós-processamento.

Outros tipos de rede que vão além da arquitetura básica de uma CNN também foram explorados em arquiteturas que usam informações espaciais para a melhoria, nas quais a utilização de blocos residuais acabou se tornando tendência. O estudo de [Wang et al. 2018] utiliza uma arquitetura baseada em redes residuais denominada *Dense Residual Convolutional Neural Network* (DRN), que utiliza um novo tipo de bloco residual denso, que permite que as informações multinível consigam se propagar pela rede sem a atenuação causada pelo problema de dissipação de gradiente. Da mesma forma, [Zhang et al. 2018] também desenvolveu uma arquitetura que utiliza unidades residuais denominada *Residual Highway Convolutional Neural Network* (RHCNN).

De acordo com [Deng et al. 2020], arquiteturas para a melhoria de qualidade visual baseadas em um único quadro foram aderindo ao uso de camadas de CNN cada vez mais profundas, para obter melhores resultados. No entanto, essa estratégia acaba elevando o número de parâmetros envolvidos no processo de treinamento, consequentemente gerando um aumento no custo computacional. Desta forma, alguns autores optam por investigar informações além do domínio espacial para o treinamento de modelos.

As primeiras soluções baseadas em CNN que foram propostas para *Video Quality Enhancement* (VQE) realizavam o processamento de cada quadro do vídeo de forma

individual. Como esses modelos evoluíram de algoritmos utilizados no processamento de imagem, somente eram observadas características espaciais durante o treinamento, não analisando a correlação temporal existente entre quadros. Estudos mais recentes têm como proposta o processamento de múltiplos quadros, utilizando informações de quadros vizinhos para melhorar a qualidade de um quadro central, desta forma englobando informações espaço-temporais.

O primeiro estudo que utiliza informações espaço-temporais para a melhoria de qualidade visual de um único quadro foi o de [Yang et al. 2018b], que propõe a arquitetura *Multi-Frame Quality Enhancement* (MFQE) tomando como base estudos semelhantes para tarefas de super-resolução. O MFQE utiliza uma abordagem de detecção de quadros vizinhos com alta qualidade (*Peak Quality Frames – PQF*) para serem utilizados como referência na melhoria de qualidade visual de um único quadro, sendo utilizado como um filtro de pós-processamento. A arquitetura MFQE possui três sub-redes, que fazem a detecção de *Peak Quality Frames* (PQF), compensação de movimento dos quadros vizinhos e melhoria de qualidade visual de todos os quadros. No mesmo ano em que a arquitetura MFQE foi lançada, [Soh et al. 2018] também propôs a utilização de informação temporal entre o quadro processado $(X^{(t)})$ e dois quadros vizinhos imediatos $(X^{(t-1)}$ e $X^{(t+1)})$ no treinamento do modelo.

A arquitetura MFQE aplica os conceitos de alinhamento e fusão no processamento da imagem. Neste tipo de arquitetura, são utilizados múltiplos quadros como entrada, os quais são alinhados através da abordagem de *optical flow* em nível de *pixel* para compensar os movimentos naturais em um vídeo. Esta abordagem provê o deslocamento de objetos entre dois quadros consecutivos, gerando vetores de movimento que apontam a coordenada de um pixel no primeiro quadro para a coordenada do mesmo pixel no segundo quadro, indicando a direção do movimento.

De acordo com [Xue et al. 2019], a maioria dos algoritmos baseados em movimento utilizam um processo de duas etapas, em que é primeiramente realizada a estimativa do movimento entre os quadros (por exemplo, através da abordagem *optical flow*), para só então realizar a fusão destes quadros, gerando um novo quadro como resultado. Em [Xue et al. 2019], o autor propõe a arquitetura *Task-Oriented Flow* (TOFlow), que, diferentemente dos algoritmos tradicionais usados na época de sua publicação, utiliza uma rede baseada em CNN que possibilita, além do processo de *optical flow*, a redução de ruído e artefatos de compressão nas bordas dos objetos.

A etapa de fusão é tipicamente realizada de duas formas principais. Na primeira, denominada *slow fusion*, os quadros de entrada são divididos em duplas e fundidos, gerando um novo quadro para cada dupla. São novamente definidas duplas com os quadros resultantes da fusão anterior, para que as mesmas sejam novamente fundidas. Esse processo ocorre até que seja formado apenas um quadro. A segunda forma é a *direct fusion*, em que todos os quadros são fundidos de uma só vez, gerando um único quadro como resultado [Tong et al. 2019]. Já o método *direct fusion* é mais direto, colapsando todos os dados temporais na primeira camada [Meng et al. 2019], diferentemente do método *slow fusion*, que mescla as camadas parcialmente durante o progresso na rede.

Conforme citam os autores de [Deng et al. 2020], apesar de muitos estudos utilizarem abordagens baseadas em *optical flow*, este esquema se demonstra sub-ótimo para

problemas de VQE, já que a distorção da imagem pode afetar a correlação de pixels entre os quadros, fazendo com que, por exemplo, um pixel presente no primeiro quadro e que deveria estar presente também no quadro seguinte seja afetado por um AC, impactando em uma perda de eficácia para a estimação de movimento. No estudo [Deng et al. 2020], foi adotada a estratégia de *Deformable Convolution* em substituição ao *optical flow*. Este mecanismo substitui a convolução tradicional utilizada pelas camadas de CNN por um tipo de convolução deformável, em que, ao invés de utilizar uma matriz fixa convencional como filtro, é utilizada uma matriz com pontos de deslocamento variável, que aprende o processo de deslocamento de pixels vizinhos em relação ao pixel processado por informações obtidas de convoluções anteriores.

Apesar dos recentes avanços nos estudos de VQE baseados em CNN, a grande maioria das soluções considera, tanto para treinamento quanto para teste dos modelos, apenas vídeos gerados em cenários específicos, geralmente compactados seguindo um único padrão de compressão em uma única configuração. Embora a maioria das soluções se concentre em vídeos compactados seguindo o padrão HEVC, alguns estudos também consideram o padrão VVC, como é o caso de [Nasiri et al. 2021] e [HoangVan and Nguyen 2020]. Diferentemente de trabalhos anteriores, este artigo apresenta uma solução que leva em consideração os diferentes tipos de artefatos que diferentes padrões/formatos de compressão e diferentes configurações de codecs introduzem aos vídeos comprimidos. Portanto, a principal contribuição deste trabalho está no fato de apresentar um modelo de VQE genérico que funcione como método de pós-processamento de mais de um codec.

3. Avaliação da Arquitetura STDF

A eficiência da arquitetura STDF [Deng et al. 2020] em termos de melhoria de qualidade foi analisada em diferentes cenários de teste com vídeos comprimidos por diferentes codecs e configurações de QP/CQ [1]. A avaliação centrou-se na observação dos resultados alcançados em termos de melhoria objetiva da qualidade de vídeos. A sessão de teste foi realizada usando o modelo treinado por [Deng et al. 2020], que está disponível publicamente[2]. O conjunto de dados de vídeo utilizado na análise é o mesmo empregado em [Yang et al. 2018a], que compreende 126 vídeos sem compressão (RAW) de diferentes resoluções, dos quais 108 foram utilizados para treinamento do modelo e 18 para teste.

O *dataset* de teste, composto por 18 vídeos, foi totalmente comprimido usando os codecs HEVC, VVC, VP9 e AV1. Para codificação HEVC, foi utilizado o software de referência HEVC Model (HM), versão 16.5, com a configuração *Low Delay P* e QP 37. Para codificação VP9, foi usado o software de referência *libvpx*, *hashcode* 1.12.0. Para HEVC e VVC, o QP foi definido como 32 e 37, enquanto que para VP9 e AV1 o CQ foi definido como 43 e 55. Assim, foram geradas 8 versões dos 18 vídeos de teste, totalizando 144 vídeos. Todos os 144 vídeos foram melhorados usando o modelo treinado e disponibilizado em [Yang et al. 2018a] e o PSNR foi calculado tomando como base as sequências originais (RAW) não compactadas.

[1] O termo QP (*Quantization Parameter*), utilizado nos padrões VVC e HEVC, é o parâmetro que indica o nível de quantização que será utilizado para a compressão do vídeo. O termo CQ (*Constant Quality*) é o equivalente ao QP nos formatos AV1 e VP9. O valor de QP/CQ é diretamente proporcional à taxa de compressão e inversamente proporcional à qualidade da imagem resultante.

[2] https://github.com/ryanxingql/mfqev2.0/wiki/MFQEv2-Dataset

Tabela 1. Melhoria de qualidade obtida com STDF [Deng et al. 2020] em diferentes cenários. Todos os valores são apresentados em PSNR (dB).

Codec	Quantização	Decodificado	Filtrado	ΔPSNR
HEVC	QP 32	34,192	34,659	0,466
	QP 37	31,608	32,327	0,719
VVC	QP 32	34,713	34,607	-0,106
	QP 37	32,186	32,457	0,271
VP9	CQ 43	36,271	35,987	-0,230
	CQ 55	33,696	34,148	0,452
AV1	CQ 43	37,701	36,431	-1,270
	CQ 55	35,227	34,867	-0,360

A Tabela 1 mostra os resultados médios obtidos para todas as sequências de teste. A coluna *Decodificado* apresenta o PSNR médio dos vídeos decodificados (ou seja, antes de ser processado pelo modelo STDF). A coluna *Filtrado* apresenta o PSNR médio dos vídeos decodificados após filtragem pelo STDF. A coluna $\Delta PSNR$ apresenta a diferença de PSNR entre os dois casos. A tabela mostra que o ganho de qualidade é maior para vídeos comprimidos com HEVC do que para vídeos codificados com outros codecs. Além disso, o maior ΔPSNR (0,719 dB) ocorre para vídeos comprimidos com HEVC e QP 37. Isso pode ser explicado porque esta é a mesma configuração usada para compactar todos os vídeos usados no treinamento do modelo STDF proposto em [Deng et al. 2020].

Também podem ser observados valores negativos na Tabela 1 para algumas configurações de codificação, o que significa que o STDF acaba por diminuir a qualidade dos vídeos comprimidos em alguns casos. Além disso, mesmo quando mantido o parâmetro de quantização, mas alterando-se o codec, percebe-se uma diferença significativa em ΔPSNR (por exemplo, HEVC e VVC com QP 32). Assim, a análise revela que existe uma correlação entre o codec empregado para gerar as sequências de treinamento e a melhoria de qualidade alcançada pelo modelo VQE.

4. Modelo STDF Multi-codec Proposto

Após a avaliação apresentada, esta seção propõe um novo modelo baseado em STDF com base na hipótese de que um modelo para VQE treinado com vídeos comprimidos por diferentes codecs é capaz de atingir melhoria significativa de qualidade em diferentes cenários. Para comprovar a hipótese, o modelo STDF foi treinado com vídeos comprimidos pelos padrões VVC e AV1, os quais diferem significativamente entre si em termos de processo de codificação, parâmetros e artefatos de compressão típicos.

O processo de treinamento do modelo STDF multi-codec começa com a divisão do *dataset* de vídeo em duas partes iguais: 54 vídeos comprimidos usando VVC com QP 37 e 54 vídeos comprimidos usando AV1 com CQ 55. O *VVC Test Model* (VTM) (versão 13.0) foi utilizado como software de referência para todas as codificações de VVC, seguindo a configuração temporal *Low Delay*. Por outro lado, o software de referência libaom (*hashcode* 3.3) foi utilizado para todas as codificações AV1. A divisão dos vídeos foi feita de acordo com o tipo de resolução, como mostrado na Tabela 2. No entanto, algumas resoluções possuem um número ímpar de vídeos, o que é resolvido adicionando um vídeo extra para um dos padrões (VVC QP 37), e a compensação é feita na próxima resolução

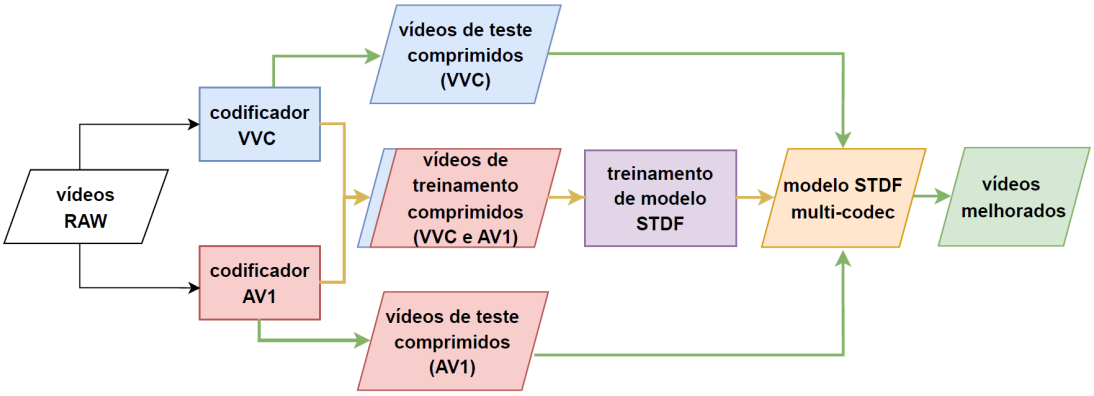

Figura 1. Metodologia de treinamento e teste do modelo STDF multi-codec.

que possui um número ímpar de vídeos, sendo esta compactada com o outro padrão (AV1 CQ 55).

A partir dessa divisão, foi gerado o *dataset* multi-codec, utilizado para treinamento o modelo STDF [Deng et al. 2020]. Durante o processo de treinamento, o modelo é alimentado com pares de imagens, sendo uma imagem de referência não compactada e outra imagem compactada pelo VVC ou AV1. O objetivo do treinamento é fazer com que o modelo aprenda a mapear as imagens compactadas de volta para as imagens originais de alta qualidade. O processo de treinamento é repetido várias vezes, ajustando os pesos do modelo para reduzir a diferença entre as imagens geradas pelo modelo e as imagens originais de alta qualidade. O modelo treinado é então avaliado usando um conjunto de dados de validação e testado em vídeos comprimidos com outros codecs e configurações de quantização.

O processo de treinamento é representado pelo caminho com setas amarelas na Fig. 1, enquanto a etapa de testes é representada pelo caminho com setas verdes. O treinamento foi realizado usando a implementação de referência STDF [Xing and Deng 2020]. Como os dois módulos do STDF são baseados em CNN, a arquitetura é unificada e pode ser treinada de ponta a ponta. Para o treinamento, foi utilizado um computador com a seguinte configuração: processador *AMD Ryzen 7 5700X*; 32 GB de memória RAM; GPU *Nvidia Geforce RTX 3070* com 8 GB de VRAM. Os parâmetros referentes ao tamanho do lote (*batch size*) e número de iterações foram adaptados para atingir o mesmo número de épocas usado em [Deng et al. 2020] com uma única GPU (ou seja, um tamanho do lote de 8 e 300.000 iterações ao longo do conjunto de dados).

5. Resultados Experimentais

A seguir, são apresentados os resultados experimentais obtidos de forma objetiva e através de análise de percepção visual das imagens após o processo de VQE. Esses resultados fornecem uma visão abrangente do estudo realizado e ajudam a avaliar a eficácia da metodologia adotada para treinamento.

5.1. Resultados Objetivos de Qualidade

A Tabela 3 apresenta os resultados de VQE obtidos para o modelo STDF multi-codec proposto considerando as 18 sequências de vídeo de teste. Os resultados são apresentados

Tabela 2. Divisão dos vídeos do *dataset* utilizado.

Resolução	Quantidade	
	VVC QP 37	AV1 CQ 55
2048×1080	3	3
1920×1080	18	18
704×576	5	4
640×360	18	19
704×576	5	4
352×288	9	9
352×240	1	1
Total	54	54

para as 8 versões dos vídeos de teste, ou seja, considerando os quatro codecs e duas configurações de quantização. Os vídeos de teste são agrupados por dimensão de acordo com a sua classe [Boyce et al. 2018]: Classe A: 2560×1600, Classe B: 1920×1080, Classe C: 832×480, Classe D: 416×240, Classe E: 1280×720. Os resultados são apresentados em PSNR, pois esta é a métrica mais comumente utilizada em trabalhos relacionados, permitindo comparações.

Tabela 3. Resultados do modelo STDF multi-codec para vídeos comprimidos com diferentes codecs.

Dataset de treino		ΔPSNR(dB)							
		HEVC		VVC		VP9		AV1	
		QP 32	QP 37	QP 32	QP 37	CQ 43	CQ 55	CQ 43	CQ 55
Classe A	*Traffic*	0,314	0,291	0,193	0,190	0,277	0,358	-0,012	0,166
	PeopleOnStreet	0,37	0,368	0,130	0,149	0,340	0,391	0,055	0,19
Classe B	*Kimono*	0,249	0,193	0,114	0,112	0,217	0,195	0,093	0,12
	ParkScene	0,15	0,105	0,115	0,078	0,183	0,164	0,123	0,155
	Cactus	0,244	0,239	0,123	0,163	0,199	0,230	0,011	0,095
	BQTerrace	0,14	0,198	-0,009	0,052	0,074	0,192	-0,127	0,032
	BasketballDrive	0,313	0,316	0,088	0,152	0,261	0,317	-0,001	0,16
Classe C	*RaceHorses*	0,254	0,246	0,084	0,113	0,244	0,283	0,66	0,176
	BQMall	0,388	0,342	0,211	0,248	0,314	0,375	0,013	0,221
	PartyScene	0,492	0,372	0,265	0,258	0,436	0,428	0,203	0,279
	BasketballDrill	0,447	0,443	0,008	0,149	0,424	0,442	0,031	0,234
Classe D	*RaceHorses*	0,348	0,298	0,177	0,175	0,352	0,328	0,22	0,261
	BQSquare	0,803	0,643	0,431	0,375	0,752	0,775	0,571	0,689
	BlowingBubbles	0,46	0,352	0,328	0,316	0,403	0,377	0,246	0,311
	BaskeetballPass	0,573	0,463	0,380	0,392	0,549	0,515	0,257	0,43
Classe E	*FourPeople*	0,514	0,431	0,316	0,342	0,450	0,514	-0,064	0,209
	Johnny	0,398	0,349	0,221	0,256	0,352	0,410	0,032	0,21
	KristenAndSara	0,428	0,384	0,226	0,260	0,345	0,461	-0,075	0,153
Média		**0,382**	**0,335**	**0,189**	**0,210**	**0,343**	**0,375**	**0,091**	**0,229**

Pode ser observado que o modelo multi-codec proposto obtém resultados médios positivos em todos os cenários, diferentemente dos testes apresentados anteriormente na Tabela 1, que levam a uma queda de qualidade visual em metade dos casos. Em média, o modelo proposto é capaz de aumentar a qualidade objetiva da imagem entre 0,091 dB (AV1, CQ 43) e 0,382 dB (HEVC, QP 32). Além disso, também pode-se observar que

Tabela 4. Comparação entre o STDF multi-codec e o STDF single-codec.

Dataset de treino	ΔPSNR (dB)							
	HEVC		VVC		VP9		AV1	
	QP 32	QP 37	QP 32	QP 37	CQ 43	CQ 55	CQ 43	CQ 55
HEVC QP 37	0,362	0,755	-0,217	0,250	-0,465	0,357	-1,479	-0,506
VVC QP 37	0,446	0,529	0,216	0,371	0,050	0,385	-0,530	-0,016
AV1 CQ 55	0,346	0,285	0,137	0,144	0,368	0,389	0,109	0,286
Multi-codec	**0,382**	**0,335**	**0,189**	**0,210**	**0,343**	**0,375**	**0,091**	**0,229**

o modelo foi capaz de melhorar a qualidade de vídeos comprimidos com outras configurações de quantização além daquelas usadas para comprimir os vídeos de treinamento, como é o caso do HEVC QP 32 (0,382 dB), VVC QP 32 (0,189 dB), VP9 CQ 43 (0,343) e AV1 CQ 43 (0,091).

Para fins de comparação, três modelos STDF single-codec também foram treinados seguindo a mesma metodologia apresentada em [Deng et al. 2020]: o primeiro com um *dataset* contendo apenas vídeos comprimidos pelo codec HEVC, o segundo com vídeos comprimidos pelo codec VVC, e o terceiro com vídeos comprimidos pelo codec AV1. Para isso, um novo conjunto de dados de treinamento foi criado com vídeos comprimidos com o codec HEVC. Foi utilizada a mesma configuração no software de referência HM mencionada na seção anterior.

A Tabela 4 apresenta resultados médios obtidos com os modelos STDF single-codec e, na última linha, os resultados obtidos com o modelo STDF multi-codec replicados. Em dois terços dos modelos single-codec, podem ser percebidos resultados negativos no teste, indicando que os modelos não são eficazes para todos os quatro padrões/formatos de codificação testados. De fato, em 6 dos 24 testes (25%) realizados com modelos single-codec, valores negativos (isto é, redução na qualidade visual) foram percebidos. O único modelo single-codec que gera resultados positivos em todos os casos de teste é aquele treinado com vídeos gerados pelo codec AV1. Ainda assim, o modelo multi-codec alcança melhores resultados do que este em praticamente todos os casos de teste, exceto para vídeos comprimidos usando o AV1, o que é esperado.

Em trabalhos futuros, é importante destacar a necessidade de realizar a avaliação objetiva da qualidade de imagem por meio de outras métricas, que permitam medi-la com mais precisão. Algumas dessas métricas são o *Structural Similarity Index* (SSIM) e o *Video Multi-Method Assessment Fusion* (VMAF), que permitem realizar comparações mais precisas do que o PSNR e chegar mais perto da qualidade visual observada em avaliações subjetivas. Essas métricas fornecem informações mais detalhadas sobre a qualidade da imagem produzida e devem ser consideradas na avaliação da eficácia de modelos e codecs.

5.2. Percepção de Qualidade Visual

As imagens e recortes apresentados na Figura 2 permitem a percepção de qualidade visual a partir do modelo treinado com vídeos comprimidos pelo codec VP9 utilizando o CQ 43. Foram selecionados aqueles quadros em que os autores perceberam a maior diferença em nível de qualidade visual. De maneira geral, observa-se que o modelo STDF suavizou em excesso as imagens nas regiões destacadas, o que resultou na perda de alguns detalhes importantes.

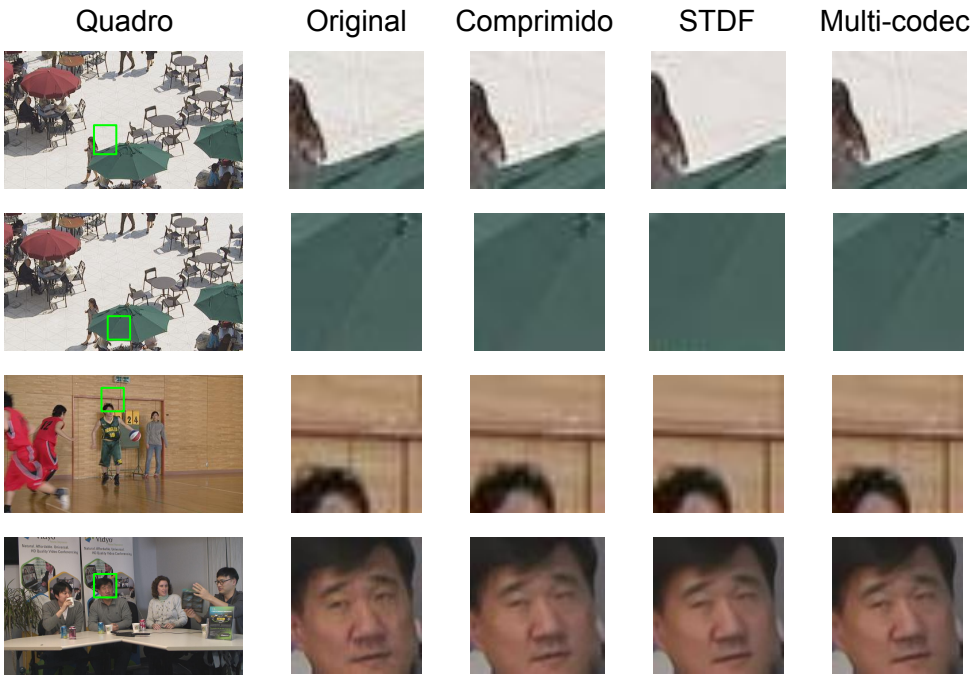

Figura 2. Comparação de resultados através da percepção de qualidade de imagens.

As duas primeiras linhas de exemplos na Figura foram retiradas da sequência de vídeo *BQSquare*. Na primeira linha, é perceptível que o modelo STDF suavizou as linhas dos ladrilhos no fundo da imagem, enquanto o modelo multi-codec conseguiu mantê-las. O mesmo ocorreu com o vinco do guarda-sol apresentado na segunda linha: o modelo STDF suavizou o vinco, tornando-o praticamente imperceptível, enquanto o modelo multi-codec manteve o detalhe. A terceira linha apresenta um quadro da sequência *BasketballDrive*, onde é possível observar que o modelo STDF suavizou a linha da porta acima da cabeça do jogador, praticamente removendo-a. Por fim, a última linha mostra um quadro da sequência *FourPeople*, no qual o modelo STDF aplicou um borramento na imagem, deixando-a visualmente mais próxima da imagem comprimida e sem melhorias.

Essa análise de percepção visual indica que o modelo multi-codec apresenta um desempenho superior em relação ao modelo STDF single-codec, no que diz respeito à preservação dos detalhes das imagens. Esse resultado não é surpreendente, já que o modelo multi-codec utiliza um dataset mais complexo e representativo, que permite uma melhor conservação dos detalhes da imagem, ao contrário do modelo STDF original.

6. Conclusão

Este artigo apresentou um novo modelo para a arquitetura *Spatio-Temporal Deformable Fusion* (STDF) de *Video Quality Enhancement* (VQE) capaz de melhorar a qualidade visual de vídeos comprimidos com diferentes codecs de vídeo. O modelo apresentado propõe uma abordagem inovadora, que leva em consideração que diferentes padrões e formatos de codificação de vídeo introduzem diferentes tipos e níveis de artefatos de compressão na imagem. Dessa forma, o modelo foi treinado com vídeos gerados por

múltiplos padrões de codificação de vídeo, para que pudesse aprender a reduzir esses artefatos de forma mais eficaz. Ao contrário dos trabalhos estado da arte, que geralmente são treinados com apenas um codec de vídeo, o modelo proposto neste artigo foi treinado com vídeos comprimidos pelos codecs HEVC, VVC, VP9 e AV1. Os resultados obtidos foram promissores, mostrando que o modelo foi capaz de melhorar a qualidade visual dos vídeos comprimidos em todos os casos testados.

O artigo apresentou uma análise sobre a eficácia do modelo STDF treinado com codec único, indicando que este é ineficaz para fins gerais. Ao propor um modelo baseado em STDF com vídeos comprimidos por dois padrões de codificação de vídeo (VVC e AV1), foram percebidos resultados de melhoria de qualidade positivos para todos os casos testados, superando os resultados obtidos por todos os modelos STDF treinados da maneira tradicional. Em média, os resultados de melhoria de qualidade variaram entre 0,091 dB (teste com AV1 e CQ 43) e 0,382 dB (teste com HEVC QP 32), indicando que o modelo apresenta uma boa capacidade de generalização para diferentes cenários de compressão de vídeo. Em trabalhos futuros, pretende-se incluir um conjunto ainda mais diversificado de codecs e configurações de QP/CQ para construir o conjunto de dados de treinamento, atingindo assim uma maior generalização e eficácia para diferentes cenários. Além disso, também é planejada a utilização de outras informações relevantes no processo de treinamento do modelo, como o nível de quantização empregado em cada quadro do vídeo e em cada bloco do quadro e outros parâmetros que possam ser obtidos em tempo de decodificação.

Referências

Boyce, J., Suehring, K., and Li, X. (2018). Jvet-j1010: Jvet common test conditions and software reference configurations. *JVET-J1010*.

Cisco (2020). Cisco annual internet report (2018–2023) white paper. `https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html`. Acessado em 15/02/2023.

Dai, Y., Liu, D., and Wu, F. (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*, pages 28–39. Springer.

Deng, J., Wang, L., Pu, S., and Zhuo, C. (2020). Spatio-temporal deformable convolution for compressed video quality enhancement. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 10696–10703.

Dong, C., Deng, Y., Loy, C. C., and Tang, X. (2015). Compression artifacts reduction by a deep convolutional network. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 576–584.

Foi, A., Katkovnik, V., and Egiazarian, K. (2007). Pointwise shape-adaptive dct for high-quality denoising and deblocking of grayscale and color images. *IEEE transactions on image processing*, 16(5):1395–1411.

HoangVan, X. and Nguyen, H.-H. (2020). Enhancing quality for vvc compressed videos with multi-frame quality enhancement model. In *2020 International Conference on Advanced Technologies for Communications (ATC)*, pages 172–176. IEEE.

Kuanar, S., Conly, C., and Rao, K. (2018). Deep learning based hevc in-loop filtering for decoder quality enhancement. In *2018 Picture Coding Symposium (PCS)*, pages 164–168. IEEE.

Li, T., Xu, M., Zhu, C., Yang, R., Wang, Z., and Guan, Z. (2019). A deep learning approach for multi-frame in-loop filter of hevc. *IEEE Transactions on Image Processing*, 28(11):5663–5678.

Meng, X., Deng, X., Zhu, S., and Zeng, B. (2019). Enhancing quality for vvc compressed videos by jointly exploiting spatial details and temporal structure. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 1193–1197. IEEE.

Nasiri, F., Hamidouche, W., Morin, L., Dhollande, N., and Cocherel, G. (2021). A cnn-based prediction-aware quality enhancement framework for vvc. *IEEE Open Journal of Signal Processing*, 2:466–483.

Soh, J. W., Park, J., Kim, Y., Ahn, B., Lee, H.-S., Moon, Y.-S., and Cho, N. I. (2018). Reduction of video compression artifacts based on deep temporal networks. *IEEE Access*, 6:63094–63106.

Statista (2022). Semiconductor market size worldwide from 1987 to 2020. `https://www.statista.com/statistics/266973/global-semiconductor-sales-since-1988/`. Acessado em 10/01/2022.

Tong, J., Wu, X., Ding, D., Zhu, Z., and Liu, Z. (2019). Learning-based multi-frame video quality enhancement. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 929–933. IEEE.

Wang, Y., Zhu, H., Li, Y., Chen, Z., and Liu, S. (2018). Dense residual convolutional neural network based in-loop filter for hevc. In *2018 IEEE Visual Communications and Image Processing (VCIP)*, pages 1–4. IEEE.

Xing, Q. and Deng, J. (2020). PyTorch implementation of STDF. https://github.com/ryanxingql/stdf-pytorch, version 1.0.0, 2020.

Xue, T., Chen, B., Wu, J., Wei, D., and Freeman, W. T. (2019). Video enhancement with task-oriented flow. *International Journal of Computer Vision*, 127(8):1106–1125.

Yang, R., Xu, M., Liu, T., Wang, Z., and Guan, Z. (2018a). Enhancing quality for hevc compressed videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(7):2039–2054.

Yang, R., Xu, M., Wang, Z., and Li, T. (2018b). Multi-frame quality enhancement for compressed video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6664–6673.

Zhang, Y., Shen, T., Ji, X., Zhang, Y., Xiong, R., and Dai, Q. (2018). Residual highway convolutional neural networks for in-loop filtering in hevc. *IEEE Transactions on image processing*, 27(8):3827–3841.