# MD-STDF: UMA ABORDAGEM PARA MELHORIA DE QUALIDADE DE VÍDEOS CODIFICADOS POR MÚLTIPLOS CODECS

GILBERTO KREISLER[1]; GARIBALDI DA SILVEIRA JUNIOR[2]; BRUNO ZATT[3]; DANIEL PALOMINO[4]; GUILHERME CORRÊA[5]

[1]*Universidade Federal de Pelotas – gkfneto@inf.ufpel.edu.br*
[2]*Universidade Federal de Pelotas – garibaldi.dsj@inf.ufpel.edu.br*
[3]*Universidade Federal de Pelotas – zatt@inf.ufpel.edu.br*
[4]*Universidade Federal de Pelotas – dpalomino@inf.ufpel.edu.br*
[5]*Universidade Federal de Pelotas – gcorrea@inf.ufpel.edu.br*

## 1. INTRODUÇÃO

O tráfego de vídeo domina a internet, representando 60% do fluxo global de dados em 2018 (EFOUI-HESS, M. et al., 2019), com projeções de aumento de 79% no tráfego de vídeo entre 2021 e 2027 (JONSSON, P.; CARSON, S.; DAVIS, S. et al., 2021). Para gerenciar a enorme quantidade de dados, a compressão de vídeo é essencial, pois vídeos não comprimidos exigem grandes quantidades de espaço e largura de banda. No entanto, a compressão introduz artefatos como blocagem, desfoque e zumbido, que afetam a qualidade da experiência do espectador. Diversas técnicas de filtragem, como o Filtro de Deblocagem (*Deblocking Filter* - DF) e o Filtro de Loop Adaptativo (*Adaptive Loop Filter* - ALF), foram desenvolvidas para reduzir esses artefatos, mas possuem limitações, como introduzir novos problemas ao tentar eliminar outros.

Modelos baseados em Redes Neurais Convolucionais (*Convolutional Neural Networks* - CNNs) ganharam destaque na melhoria da qualidade do vídeo (*Video Quality Enhancement* - VQE) devido à sua capacidade de capturar padrões visuais complexos e lidar com diferentes tipos de degradação causados pela compressão. Entretanto, muitos modelos de redes neurais profundas (*Deep Neural Network* - DNN) funcionam bem apenas em vídeos comprimidos com os mesmos codecs usados no treinamento. A arquitetura MD-STDF (*Multi-Domain Spatio Temporal Deformable Fusion*) resolve essa limitação usando um treinamento multi-domínio, permitindo que o modelo se adapte a diferentes codecs e artefatos de compressão. Os experimentos mostram que o MD-STDF melhora a qualidade dos vídeos em diversos codecs, alcançando ganhos significativos no *Peak Signal-to-Noise Ratio* (PSNR), estabelecendo um novo marco na melhoria da qualidade de vídeos comprimidos.

## 2. METODOLOGIA

A arquitetura MD-STDF baseia-se na STDF (DENG, J. et al., 2020), que utiliza múltiplos quadros de vídeo para melhorar um quadro central. Incorporando uma estratégia de treinamento multi-domínio, o modelo é treinado com dados de diferentes domínios, permitindo que aprenda características específicas de cada um. A STDF possui dois módulos principais: um para o alinhamento, extração e fusão de características dos quadros, e outro para a melhoria da qualidade do quadro central com base nas características fundidas, que utiliza quadros passados e futuros. A quantidade de quadros vizinhos utilizados depende do parâmetro de Raio (R), e quando R=1, três quadros são usados no total.

O modelo MD-STDF implementa um módulo de alinhamento e fusão que extrai características gerais dos vídeos, independentemente do domínio, utilizando uma abordagem de convoluções deformáveis para capturar variações espaciais e temporais. O treinamento multi-domínio usa rótulos específicos para vincular cada lote de treinamento a um domínio particular, ajustando o módulo de melhoria de qualidade (*Quality Enhancement* - QE) de acordo com essas características. Cada ramo do QE é otimizado para parâmetros específicos de cada domínio, produzindo um Mapa Residual que é somado ao quadro central para gerar um quadro aprimorado. Esse processo é repetido para todos os quadros de um vídeo comprimido, resultando em uma sequência de vídeo aprimorada.

O conjunto de dados utilizado neste trabalho é o MFQE (YANG, R. et al., 2018), composto por 126 vídeos não comprimidos (108 para treinamento e 18 para teste) em resoluções variando de 352×240 a 1920×1080. As sequências de vídeo foram organizadas com base no padrão e no parâmetro de qualidade usados para a compressão, permitindo a divisão do dataset para o treinamento multi-domínio. Oito versões do conjunto de treinamento foram geradas, cada uma correspondendo a um domínio específico, resultando em 864 vídeos, codificados e decodificados com quatro padrões de codificação de vídeo *High Efficiency Video Coding* (HEVC), *Versatile Video Coding* (VVC), *AOMedia Video 1* (AV1) e VP9; e dois valores de parâmetro de quantização. Para o HEVC e VVC, o parâmetro de quantização (*Quantization Parameter* - QP) foi ajustado para 32 e 37, enquanto para o VP9 e AV1, o parâmetro de qualidade constante (*Constant Quality* - CQ) foi definido como 43 e 55. Esses parâmetros controlam o nível de quantização aplicado, afetando diretamente a qualidade do vídeo, onde valores mais altos causam maior perda de detalhes.

Para o treinamento do modelo, foi utilizado um computador com a seguinte configuração: processador AMD Ryzen 7 5700X, 32 GB de RAM, GPU Nvidia Geforce RTX 3070 com 8 GB de VRAM. O tamanho do lote (*batch size*) e o número de iterações foram ajustados para alcançar 10 épocas com uma GPU (ou seja, um batch size de 32 e 1.200.000 iterações sobre o conjunto de dados). O treinamento foi realizado utilizando o otimizador Adam com $\beta_1 = 0{,}9$, $\beta_2 = 0{,}999$ e $\varepsilon = 10^{-8}$, e uma taxa de aprendizado de 0,0001.

## 3. RESULTADOS E DISCUSSÃO

A Tabela 1 apresenta a variação objetiva de qualidade, medida em ΔPSNR, para cada sequência de vídeo. Essa métrica mostra a diferença de qualidade entre as sequências de vídeo comprimidas e aprimoradas, com valores positivos indicando melhorias e valores negativos indicando uma queda na qualidade. As sequências de vídeo estão organizadas em categorias com base nas Condições Comuns de Teste do JVET (BOYCE, J.; SUEHRING, K.; LI, X., 2018): Classe A (2560x1600), Classe B (1920x1080), Classe C (832x480), Classe D (416x240) e Classe E (1280x720).

Como observado na Tabela 1, a maioria dos resultados é positiva, com apenas um caso específico mostrando um valor negativo de ΔPSNR. Em média, todos os resultados foram positivos. O pior resultado foi -0,024 dB para AV1 na sequência *BQTerrace*. O melhor resultado foi 1,437 dB para HEVC QP 32 na

sequência *BQSquare*. Para os resultados médios, a menor melhoria foi 0,228 dB para vídeos codificados com AV1 CQ 55, enquanto a maior foi 0,787 dB para vídeos codificados com HEVC QP 32. Alguns casos específicos, além do melhor caso, apresentaram resultados superiores a 1 dB, como a sequência *People on Street* codificada com HEVC com QP 37, que atingiu 1,126 dB, e a sequência *BQSquare*, que alcançou 1,020 dB, 1,379 dB e 1,234 dB quando codificada com HEVC (QP 37), VP9 (CQ 43) e VP9 (CQ 55), respectivamente.

Tabela 1: Resultados de VQE para o modelo Multi-Domínio em ΔPSNR

| *Dataset* de Teste | | 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* | 0,730 | 0,662 | 0,480 | 0,420 | 0,697 | 0,727 | 0,290 | 0,120 |
| | *People on Street* | 0,919 | 1,126 | 0,465 | 0,582 | 0,741 | 0,911 | 0,298 | 0,200 |
| Classe B | *Kimono* | 0,777 | 0,831 | 0,499 | 0,547 | 0,464 | 0,462 | 0,237 | 0,184 |
| | *ParkScene* | 0,628 | 0,551 | 0,509 | 0,426 | 0,530 | 0,487 | 0,255 | 0,152 |
| | *Cactus* | 0,642 | 0,698 | 0,391 | 0,404 | 0,554 | 0,641 | 0,168 | 0,130 |
| | *BQTerrace* | 0,498 | 0,543 | 0,244 | 0,217 | 0,333 | 0,402 | 0,102 | -0,024 |
| | *BasketballDrive* | 0,586 | 0,686 | 0,266 | 0,292 | 0,488 | 0,552 | 0,206 | 0,156 |
| Classe C | *RaceHorses* | 0,456 | 0,447 | 0,263 | 0,220 | 0,486 | 0,473 | 0,228 | 0,153 |
| | *BQMall* | 0,880 | 0,838 | 0,587 | 0,529 | 0,807 | 0,828 | 0,369 | 0,231 |
| | *PartyScene* | 0,854 | 0,640 | 0,579 | 0,380 | 0,857 | 0,731 | 0,557 | 0,197 |
| | *BasketballDrill* | 0,622 | 0,718 | 0,241 | 0,290 | 0,753 | 0,733 | 0,423 | 0,197 |
| Classe D | *RaceHorses* | 0,698 | 0,691 | 0,491 | 0,436 | 0,731 | 0,661 | 0,420 | 0,344 |
| | *BQSquare* | 1,437 | 1,020 | 0,966 | 0,667 | 1,379 | 1,234 | 1,147 | 0,488 |
| | *BlowingBubbles* | 0,834 | 0,635 | 0,622 | 0,453 | 0,833 | 0,702 | 0,475 | 0,328 |
| | *BasketballPass* | 1,013 | 0,971 | 0,763 | 0,716 | 0,947 | 0,860 | 0,585 | 0,496 |
| Classe E | *FourPeople* | 0,936 | 0,913 | 0,590 | 0,548 | 0,967 | 1,046 | 0,356 | 0,325 |
| | *Johnny* | 0,757 | 0,804 | 0,406 | 0,414 | 0,757 | 0,843 | 0,222 | 0,164 |
| | *KristenAndSara* | 0,905 | 0,969 | 0,508 | 0,515 | 0,829 | 0,954 | 0,299 | 0,270 |
| **Média** | | **0,787** | **0,764** | **0,493** | **0,448** | **0,731** | **0,736** | **0,369** | **0,228** |

A Tabela 2 mostra uma comparação de diferentes abordagens. A coluna *Modelo STDF* representa o conjunto de dados usado para treinar cada modelo. As colunas subsequentes representam o conjunto de dados de teste usado para obter os valores de ΔPSNR. As três primeiras linhas (HEVC, VVC e AV1) apresentam resultados de VQE obtidos a partir de modelos treinados usando uma abordagem de codec único, ou seja, com um conjunto de dados composto apenas por vídeos comprimidos com um codec específico. A quarta linha apresenta resultados de uma abordagem multi-codec (KREISLER, G. et al., 2024), na qual o conjunto de dados é composto por vídeos comprimidos com diferentes codecs. Finalmente, a última linha apresenta resultados obtidos com o método multi-domínio proposto.

Tabela 2: Comparação entre Single-Codec, Multi-Codec e Multi-Domínio

| Modelo STDF | 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 (DENG, J. et al., 2020) | 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 (KREISLER, G. et al., 2024) | 0,382 | 0,335 | 0,189 | 0,210 | 0,343 | 0,375 | 0,091 | 0,229 |
| **Multi-Domínio** | **0,787** | **0,764** | **0,493** | **0,448** | **0,731** | **0,736** | **0,369** | **0,228** |

O modelo treinado com vídeos comprimidos com HEVC apresenta melhor desempenho em vídeos HEVC (0,755 dB), mas um desempenho fraco em vídeos AV1 (-0,506 dB). Da mesma forma, o modelo treinado com VVC tem dificuldades

com vídeos AV1. O modelo multi-codec de KREISLER, G. et al. (2024) fornece resultados mais consistentes entre os codecs (0,210 dB a 0,375 dB), mas não supera os modelos de codec único. O modelo multi-domínio proposto supera tanto os modelos de codec único quanto os modelos multi-codec, alcançando os melhores resultados na maioria dos codecs, com melhorias de qualidade variando de 0,228 dB para AV1 CQ 55 a 0,787 dB para HEVC QP 32.

## 4. CONCLUSÕES

Este estudo introduziu uma nova arquitetura de aprimoramento de qualidade de vídeo chamada *Multi-Domain Spatio-Temporal Deformable Fusion* (MD-STDF), que utiliza aprendizado multi-domínio para melhorar a qualidade de vídeos comprimidos com vários codecs. O modelo foi treinado em vídeos de múltiplos codecs, permitindo lidar melhor com diferentes tipos e níveis de artefatos de compressão. Os experimentos mostraram que o MD-STDF melhorou significativamente a qualidade do vídeo para HEVC, VVC, AV1 e VP9, superando abordagens de codec único e domínio único. A melhoria média no ΔPSNR variou de 0,228 dB para AV1 CQ 55 a 0,787 dB para HEVC QP 32, demonstrando a forte capacidade de generalização do modelo em diferentes cenários de compressão. Essa eficácia pode ser atribuída ao extenso conjunto de dados de treinamento, que provavelmente aprimorou as capacidades de alinhamento e fusão da rede compartilhada entre os domínios. Trabalhos futuros envolverão o uso de métricas adicionais, como VMAF e LPIPS, treinamento com uma gama mais ampla de codecs e configurações, além da realização de um estudo de ablação e análise de redução de custos.

## 5. REFERÊNCIAS BIBLIOGRÁFICAS

JONSSON, P.; CARSON, S.; DAVIS, S. et al. **Ericsson mobility report (2021)**. Stockholm: Ericsson, 2021. Tech. Rep.

EFOUI-HESS, M. et al. **Climate crisis: The unsustainable use of online video**. Paris: The Shift Project, 2019.

DENG, J. et al. **Spatio-temporal deformable convolution for compressed video quality enhancement**. In: PROCEEDINGS OF THE AAAI CONFERENCE ON ARTIFICIAL INTELLIGENCE, v. 34, 2020. p. 10 696–10 703.

KREISLER, G. et al. **Multi-codec video quality enhancement model based on spatio-temporal deformable fusion**. In: 2024 IEEE 15th LATIN AMERICA SYMPOSIUM ON CIRCUITS AND SYSTEMS (LASCAS), 2024.

YANG, R. et al. **Enhancing quality for hevc compressed videos**. IEEE Transactions on Circuits and Systems for Video Technology, v. 29, n. 7, p. 2039–2054, 2018.

BOYCE, J.; SUEHRING, K.; LI, X. **JVET-J1010: JVET common test conditions and software reference configurations**. JVET-J1010, 2018.