Avaliação da rede Coarse-to-Fine Spatio-Temporal Information Fusion (CF-STIF)

LUIS CARLOS RODRIGUES LINARES[1]; GILBERTO KREISLER[2]; DANIEL PALOMINO[3]; GUILHERME CORRÊA[4]; BRUNO ZATT[5]

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

1. INTRODUÇÃO

O aumento do volume de dados provenientes de vídeos digitais na internet continua crescendo exponencialmente, com uma grande aceleração desde a pandemia de COVID-19 (STATISTA, 2022). Em 2021, o tráfego global de vídeos online representou 60% do fluxo de dados global, com um crescimento estimado de 79% entre 2021 e 2027 (ZHANG, 2018).

Com o consumo massivo de vídeos digitais e a crescente demanda por vídeos de alta qualidade, pesquisadores têm se esforçado para desenvolver soluções de compressão mais eficientes. Essas soluções são fundamentais para viabilizar o armazenamento e a transmissão de vídeos em ambientes com limitação de largura de banda como redes móveis — as aplicações de vídeo representaram 66,2% do uso global de dados móveis em janeiro de 2021 (STATISTA, 2022).

Oferecer altas taxas de compressão para atender às restrições impostas por ambientes com limitação de largura de banda pode levar à degradação da qualidade do vídeo e à inserção de artefatos visuais. Como medidas de prevenção, os padrões atuais de codificação de vídeo empregam estratégias baseadas em conhecimento prévio para criar filtros com o objetivo de melhorar a qualidade visual (LI, 2022). No entanto, questões como borrões e bordas serrilhadas são comuns mesmo nos padrões de última geração, impactando negativamente a qualidade visual percebida pelos usuários.

Em contraste com os filtros tradicionais, que se concentram em artefatos de compressão específicos, arquiteturas baseadas em aprendizado, como redes neurais profundas, são capazes de correlacionar *pixels* vizinhos e entender o contexto visual e os padrões de imagem/vídeo. Quando utilizadas para aprimorar a qualidade visual, essas redes tendem a melhorar a qualidade da imagem de forma integral, em vez de se concentrarem em artefatos de compressão específicos, o que ajuda a evitar a introdução de novos artefatos durante o processamento da imagem (LI, 2022).

Várias soluções de aprimoramento da qualidade de vídeo (VQE) baseadas em aprendizado foram propostas nos últimos anos. O *Coarse-to-Fine Spatio-Temporal Information Fusion* (CF-STIF) (LUO, 2022) destaca-se como uma solução eficaz que emprega redes neurais profundas para fornecer aprimoramento de qualidade de última geração para vídeos codificados. O CF-STIF atinge um aumento médio de ~0,9 dB na métrica *Peak Signal-to-Noise Ratio (*PSNR) considerando vídeos codificados no padrão HEVC. No entanto, os artefatos visuais/degradação variam entre diferentes padrões de codificação

devido ao uso de um conjunto diferente de ferramentas de codificação. Como resultado, não é possível atestar o desempenho do CF-STIF quando aplicado a vídeos codificados usando outros padrões/*codecs*, como VVC ou AV1.

Neste artigo, apresentamos uma avaliação da arquitetura CF-STIF para avaliar seu potencial como uma ferramenta de aprimoramento de qualidade de vídeo independente de *codec*. Medimos o aprimoramento da qualidade do CF-STIF considerando diferentes *codecs* (HEVC, VCC, VP9, AV1) e níveis de qualidade, para múltiplas métricas de qualidade (PSNR, SSIM e LPIPS).

2. METODOLOGIA

Para o treinamento da rede CF-STIF, utilizamos a estrutura de treinamento desenvolvida por XING, 2020. O treinamento foi realizado em uma GPU Titan V com 12 GB de memória. Neste experimento, o conjunto de dados de treinamento foi o MFQEv2Dataset, onde os vídeos de baixa qualidade (LQ) foram codificados no formato HEVC com QP 37 usando a versão 16.5 do HM.

Para o processamento, a GPU contou com 12 *workers* e um *batch size* de 32. A arquitetura da rede neural foi configurada com um raio de entrada de 3, resultando no uso de 7 quadros (3 anteriores, 1 atual e 3 subsequentes). O Módulo *Multi-Level Residual Fusion* (MLRF) foi configurado com 1 canal de entrada (canal Y), 64 canais de saída e 32 mapas de características, utilizando 3 camadas de convolução com um tamanho de *kernel* de 3x3. O Módulo de Reconstrução recebeu os 64 canais de saída do módulo anterior, gerando 1 canal de saída (canal Y) e trabalhando com 48 mapas de características distribuídos em 8 camadas de convolução, também com um tamanho de *kernel* de 3x3.

As configurações de treinamento consistiram em um total de 300.000 iterações, com um intervalo de validação a cada 5.000 iterações. O otimizador escolhido foi o Adam, com uma taxa de aprendizado inicial de 1×10^{-4}, $\beta1=0,9$, $\beta2=0,999$ e $\epsilon=1\times10^{-8}$. A função de perda aplicada foi a *CharbonnierLoss*, com $\epsilon=1\times10^{-6}$, enquanto o critério de avaliação utilizado foi o PSNR.

Para uma avaliação completa dos resultados, o conjunto de testes composto por 18 vídeos foi medido por três métricas diferentes: PSNR**,** *Structural similarity index measure* (SSIM) e *Learned Perceptual Image Patch Similarity* (LPIPS).

Os testes foram realizados utilizando vídeos comprimidos em diferentes formatos, incluindo HEVC (QP 37)**,** VVC (QP 37)**,** VP9 (CQ 55) e AV1 (CQ 55), permitindo uma análise comparativa do desempenho do modelo entre vários tipos de compressão. Dessa forma, o modelo é avaliado em cenários diversos, refletindo a capacidade de generalização do processo de aprimoramento da qualidade de vídeo em vários formatos de compressão.

4. RESULTADOS E DISCUSSÃO

Na Tabela 1, observa-se a superioridade do CF-STIF ao aprimorar vídeos comprimidos em HEVC, em comparação com vídeos gerados por outros *codecs* (VVC, VP9 e AV1). A análise das métricas PSNR e SSIM, que medem a qualidade de reconstrução e a similaridade estrutural, respectivamente, revela um desempenho consistentemente melhor do CF-STIF ao aprimorar vídeos comprimidos em HEVC.

Em média, os vídeos comprimidos em HEVC alcançaram uma melhoria de 0,813 dB, um valor consideravelmente superior ao de VVC (0,164 dB), VP9 (0,371 dB) e vídeos comprimidos em AV1 (0,0337 dB). Esses valores representam o Δ

Tabela 1 – Resultados de melhoria de qualidade de vídeo para HEVC, VVC, VP9 e AV1 em termos de ΔPSNR (dB), ΔSSIM (×10−4) e ΔLPIPS (×10−4)

Resolução	Vídeo	HEVC			VVC			VP9			AV1		
		Δ PSNR	Δ SSIM	Δ LPIPS	Δ PSNR	Δ SSIM	Δ LPIPS	Δ PSNR	Δ SSIM	Δ LPIPS	Δ PSNR	Δ SSIM	Δ LPIPS
1920x1080	Kimono	0.912	70	-151	0.177	52	18	0.322	48	-130	0.068	-19	-121
1920x1080	ParkScene	0.584	71	-186	0.300	62	-46	0.236	40	-242	0.254	-46	-179
1920x1080	Cactus	0.750	154	-65	0.128	29	-39	0.430	57	-200	0.056	-47	-159
1920x1080	BQTerrace	0.622	127	-72	0.088	17	-22	0.201	35	-131	-0.001	-36	-113
1920x1080	BasketballDrive	0.675	153	-4	0.093	21	17	0.488	65	-114	0.019	-21	-90
1280x720	FourPeople	0.999	116	-32	0.232	27	-9	0.298	43	-71	0.100	-40	-86
1280x720	Johnny	0.860	79	-6	0.038	14	-47	0.382	37	-101	-0.011	-25	-93
1280x720	KristenAndSara	1.016	96	-64	0.080	17	-37	0.218	34	-123	-0.002	-26	-101
832x480	RaceHorses	0.501	235	-102	0.052	6	-178	0.388	74	-276	-0.009	-45	-280
832x480	BQMall	0.999	200	-31	0.250	56	13	0.460	68	-124	0.051	-51	-134
832x480	PartyScene	0.674	260	42	0.207	53	-6	0.276	55	-95	-0.015	-99	-116
832x480	BasketballDrill	0.847	137	-76	0.088	5	-73	0.354	75	-145	0.005	-91	-221
416x240	RaceHorses	0.799	308	-107	0.221	68	-77	0.557	113	-129	-0.011	-24	-224
416x240	BQSquare	0.994	184	-2	0.243	51	3	0.453	21	-12	-0.027	-101	-128
416x240	BlowingBubbles	0.708	280	-92	0.194	99	-36	0.226	65	-49	0.037	-63	-154
416x240	BasketballPass	1.070	260	-51	0.241	83	4	0.650	93	-74	0.026	-38	-173
1920x1080	AVG	0.709	115	-95.6	0.157	36.2	-14.4	0.335	49	-163.4	0.079	-33.8	-132.4
1280x720	AVG	0.958	97	-34	0.116	19.333	-31	0.299	38	-98.333	0.029	-30.333	-93.333
832x480	AVG	0.755	208	-41.75	0.149	30	-61	0.369	68	-160	0.008	-71.5	-187.75
416x240	AVG	0.892	258	-63	0.224	75.25	-26.5	0.471	73	-66	0.006	-56.5	-169.75
–	**AVG**	**0.813**	**170.6**	**-62.43**	**0.164**	**41.25**	**-32.18**	**0.371**	**57.69**	**-126**	**0.0337**	**-48.25**	**-148.25**

de melhoria de na qualidade de reconstrução, demonstrando que o CF-STIF é mais eficaz para vídeos codificados em HEVC. Para resoluções como 1920x1080 e 1280x720, os vídeos comprimidos em HEVC também mostraram ganhos expressivos, como no vídeo *KristenAndSara* (1,016 dB) e no vídeo *BasketballPass* (1,070 dB), enquanto vídeos comprimidos por outros *codecs* apresentaram melhorias menores ou até perda de qualidade, como foi o caso do AV1 em vários vídeos.

Considerando a métrica SSIM, o CF-STIF também demonstra resultados superiores para vídeos comprimidos em HEVC, com um SSIM médio de 170,6 (multiplicado por $\times 10^{-4}$). Para VVC, VP9 e AV1, os resultados médios de SSIM foram 41,25, 57,69 e -48,25, respectivamente. Essa métrica, que reflete a percepção humana da qualidade, é essencial para garantir que as melhorias sejam claramente percebidas pelos usuários. No caso do AV1, todos os vídeos apresentaram valores negativos de SSIM, indicando degradação visual quando o CF-STIF é utilizado.

Para a métrica LPIPS, utilizando a rede “Alex”, os vídeos comprimidos em VVC alcançaram um LPIPS médio de -32,43 (multiplicado por $\times 10^{-4}$), superando HEVC (-62,43), VP9 (-126) e AV1 (-148). O LPIPS identificou quase todos os vídeos como piores após o aprimoramento e classificou o AV1 como o pior. A métrica LPIPS usa uma rede neural para estimar a similaridade perceptual, focando mais na percepção visual humana do que em métricas tradicionais como PSNR ou SSIM. Embora o codec HEVC apresente melhores resultados em PSNR e SSIM, o LPIPS pode perceber a qualidade como inferior, pois avalia diferenças em texturas, bordas e detalhes mais finos que o PSNR e o SSIM podem não captar. Isso pode ocorrer devido à forma como o CF-STIF modifica o conteúdo do vídeo, possivelmente introduzindo mudanças que afetam esses aspectos perceptuais de maneira que o LPIPS identifica como degradação, apesar das melhorias na fidelidade do sinal.

Esses resultados demonstram que a rede CF-STIF é otimizada para melhorar a qualidade de vídeos codificados em HEVC, o que não ocorre com os outros codificadores. O desempenho inferior nos codificadores AV1, VVC e VP9 pode ser atribuído às diferenças na maneira como esses codificadores aplicam a compressão e aos tipos de artefatos introduzidos. Isso sugere que, embora o CF-

STIF seja altamente eficaz para HEVC, ajustes específicos podem ser necessários para melhorar o desempenho em outros codificadores.

4. CONCLUSÕES

Os resultados deste artigo evidenciam a eficácia do CF-STIF em melhorar a qualidade de vídeos codificados em HEVC, superando outros codecs analisados (VVC, VP9 e AV1) em termos de PSNR e SSIM. A arquitetura do CF-STIF proporcionou um ganho médio significativo de 0,813 dB no PSNR, assim como uma percepção visual consistentemente superior, medida pelo SSIM. No entanto, ao avaliar outros codecs como VVC, VP9 e AV1, o desempenho do CF-STIF foi inferior, sugerindo que, apesar de sua eficácia no HEVC, o modelo enfrenta limitações em cenários mais diversos. O AV1, especificamente, apresentou valores negativos de SSIM em todos os vídeos, indicando degradação de qualidade. Isso ressalta a necessidade de uma adaptação mais robusta do CF-STIF para lidar com os diferentes tipos de artefatos gerados por esses padrões de compressão.

É importante destacar que o modelo CF-STIF foi treinado utilizando o PSNR como métrica de critério, o que pode explicar por que o modelo melhora a qualidade em termos de PSNR e SSIM, mas a diminui ao considerar o LPIPS. Trabalhos futuros podem explorar o uso de métricas alternativas, como o LPIPS, durante o treinamento. Além disso, pesquisas futuras poderiam se concentrar em ajustar a arquitetura do CF-STIF para torná-la mais generalizável a diferentes codecs, com o desenvolvimento de módulos especializados para artefatos específicos e o treinamento com dados mais diversificados. Com esses ajustes, o CF-STIF pode se tornar uma solução ainda mais eficaz para aprimorar a qualidade de vídeo em uma ampla variedade de padrões de compressão, não se limitando apenas ao HEVC.

5. REFERÊNCIAS BIBLIOGRÁFICAS

STATISTA. **Semiconductor market size worldwide from 1987 to 2020**. Disponível em: https://www.statista.com/statistics/266973/global-semiconductor-sales-since-1988/. Acesso em: 1 out. 2024.

ZHANG, R.; ISOLA, P.; EFROS, A.; SHECHTMAN, E.; WANG, O. **The unreasonable effectiveness of deep features as a perceptual metric**. 2018.

LI, Z. et al. A survey of convolutional neural networks: analysis, applications, and prospects. **IEEE Transactions on Neural Networks and Learning Systems**, v. 33, n. 12, p. 6999–7019, 2022.

LUO, D.; YE, M.; LI, S.; LI, X. Coarse-to-fine spatiotemporal information fusion for compressed video quality enhancement. **IEEE Signal Processing Letters**, v. 29, p. 543–547, 2022.

XING, Q.; DENG, J. **PyTorch implementation of STDF (Version 1.0.0)**. Disponível em: https://github.com/ryanxingql/stdf-pytorch. Acesso em: 1 out. 2024.