Cross-Frame Attention を用いた映像補間モデルの一検討

A Method for Video Frame Interpolation Using Cross-Frame Attention

渡部泰樹† 進藤嵩紘† 巽優衣† 渡辺裕†
Taiju Watanabe† Takahiro Shindo† Yui Tatsumi† Hiroshi Watanabe†

†早稲田大学
†Waseda University

Abstract Recently, diffusion models are used for video frame interpolation. However, for videos with a lot of movement, the texture collapses when frame interpolation is performed. In this study, we extend the existing research using Cross-Frame Attention to improve the accuracy of frame interpolation.

1. はじめに

近年のデータ量の増大に伴い，映像伝送技術において，通信帯域を圧迫させないためにも伝送する映像量を減らす必要がある．その一つの手法として AI を用いたフレーム補間が考えられる．これは映像全てを伝送するのではなく，最初と最後のフレームを伝送し，その２枚のフレームから間のフレームを AI を用いて補間する手法である．この手法によって，大幅に伝送量を削減できる一方で，AI を用いた映像補間技術は発展途上であり，HEVC や VVC といった映像符号化技術と組み合わせるまでには至っていない．というのも，既存手法では，物体の動きやカメラの動きが激しいような映像に対してはフレーム補間精度が著しく劣化する．最近では AI 技術の中で映像生成によく用いられるのが，拡散モデルである．一方で，既存研究の拡散モデルを用いたフレーム補間モデルでは，フレーム補間をした際にテクスチャが崩壊するなど精度があまり良くない．本研究では，既存のフレーム補間モデルを拡張し，フレーム補間精度の向上を目指す．

2. 関連研究

2.1. フレーム補間技術

フレーム補間とは，既存のフレーム間に新しいフレームを生成する手法である．この技術は，映像のフレームレートを高めることで映像を滑らかにするなど，映像生成技術において重要な役割を果たしている．古典的なフレーム補間技術は主にモーション推定に基づいている．モーション推定とは，映像内のオブジェクトの動きを解析し，次のフレームを予測する技術である．この技術は，映像伝送技術にも使用されており，フレーム間の動きを正確に捉えることで圧縮効率を向上させている．最近では，ディープラーニングベースのフレーム補間モデルも提案されている．RIFE[1]では，CNN によってオプティカルフローを計算し，中間フレームを補間している．

2.2. 拡散モデル

拡散モデルは，元の画像データにノイズを加えていく拡散過程と，ノイズからノイズを除去することで画像データを作成する逆拡散仮定の２種類のプロセスで構成されている．この拡散モデルは最近では動画生成にも使用されており，画素を生成するものと，潜在表現を生成するものに分けられている．画素を生成するものとして，1 フレームずつ条件付き画像生成問題として解く LDMVFI[2]などがある．また，潜在表現を生成するものとして，画像用の潜在表現を使用する Latent-Shift[3]などがある．また，VDT[4]ではマスクする潜在表現を選択することによって，動画予測，フレーム補間，映像修復を可能にしている．本研究では，VDT を参考にし，フレーム補間精度を向上させることを目指す．

3. 提案手法

3.1. Cross-Frame Attention

VDT の補間精度を向上させるために，VideoBooth[5]で提案された Cross-Frame Attention を参考にして，フレーム補間用の Cross-Frame Attention に改良する．また，これを VDT の Attention 層に導入することで，VDT のフレーム補間精度を向上させる．図 1 に提案手法の Cross-Frame Attention の概要を示した．Cross-Frame Attention は，各フレームの潜在変数の時間的な整合性を改善するために用いられる．

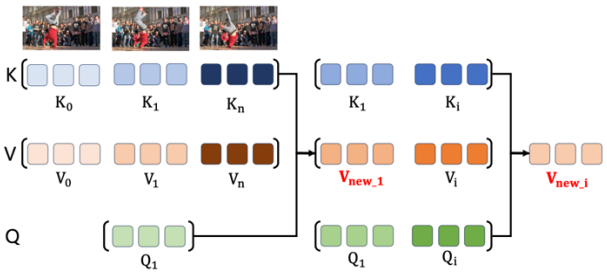

図 1. Cross-Frame Attention の概要

表１．フレーム補間結果(Vimeo90K)

モデル	PSNR↑	SSIM↑	LPIPS↓
VDT	19.1	0.670	0.080
提案手法	19.8	0.694	0.070

表２．フレーム補間結果(DAVIS2017)

モデル	PSNR↑	SSIM↑	LPIPS↓
VDT	17.6	0.445	0.169
提案手法	17.8	0.450	0.153

図１のように，t=1 のフレームに該当するトークンは，t=0 と t=n の最初と最後のフレームに該当するトークンによって条件づけられ，それが新しい Value となって，他のフレームにも伝搬される．この Attention によって，最初と最後のフレームの情報が他のフレームにも伝搬し，生成したフレーム全体として整合性の取れたものとなることが期待できる．

4. 実験

提案手法の有効性を検証するために実験を行った．Vimeo90K データセット[6]の訓練用データを用いて VDT と提案手法の学習を行った．いずれも 400Kstep 学習させた．学習したモデルを用いて Vimeo90K データセットの検証用データを用いてフレーム補間精度を比較した．いずれも t=0 と t=6 のフレームを受け取り，t=1~5 の 5 枚のフレームを生成し，それらの補間精度を検証した．評価指標としては，PSNR，SSIM，LPIPS を用いた．表１にその結果を示した．また，同様にして DAVIS2017 データセット[7]に対しても補間精度を比較した．表２にその結果を示した．いずれのデータセットに対しても，提案手法の方が補間精度が向上していることがわかる．これは Cross-Frame Attention を導入したことにより，時間的な整合性が改善したためだと考えられる．また，図２に DAVIS2017 データセットに対する定性的な結果を示した．図２からわかるように提案手法では，馬や人といった物体の形状が保たれているのに対して，従来手法の VDT では，形状が崩壊してしまっていることがわかる．

5. まとめ

本研究では，拡散モデルを用いたフレーム補間に対する有効な手法である Cross-Frame Attention を提案した．実験より，定性的にも定量的にも提案手法によってフレーム補間精度が向上したことがわかった．今後の展望としては，依然として物体が少々崩壊してしまっているので，より最初と最後のフレームに存在する物体を反映させるようなフレーム補間手法について検討していく．

図２．定性的な結果
（上：VDT，中央：提案手法，下：正解）

謝　辞

本研究成果は，国立研究開発法人情報通信研究機構の委託研究 (No. 05101) により得られたものである．

文　献

[1] Z. Huang, T. Zhang, W. Heng, B. Shi, S. Zhou, “Real-Time Intermediate Flow Estimation for Video Frame Interpolation, ” ECCV 2022. Lecture Notes in Computer Sciences, vol 13674. 2022, pp 624-642.

[2] Duolikun Danier, Fan Zhang, David Bull, “LDMVFI: Video Frame Interpolation with Latent Diffusion Models, ” AAAI 2024. Proceedings of the AAAI Conference on Artificial Intelligence, vol 38. 2024, pp 1472-1480.

[3] J. An, S. Zhang, H. Yang, S. Gupta, J-B. Huang, J. Luo, X. Yin, “Latent-Shift: Latent Diffusion with Temporal Shift for Efficient Text-to-Video Generation,” arXiv preprint, arXiv : 2304.08477.

[4] H. Lu, G. Yang, N. Fei, Y. Huo, Z. Lu, P. Luo, M. Ding, “ VDT: General-purpose Video Diffusion Transformers via Mask Modeling,” ICLR 2024. The International Conference on Learning Representations, 2024.

[5] Y. Jiang, T. Wu, S. Yang, C. Si, D. Lin, Y. Qiao, C. C. Loy, Z. Liu, “VideoBooth: Diffusion-based Video Generation with Image Prompts, ” CVPR 2024. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2024, pp 6689-6700.

[6] T. Xue, B. Chen, J. Wu, D. Wei, W. T. Freeman,“Video Enhancement with Task-Oriented Flow,” IJCV 2019. International Journal of Computer Vision, vol 127. 2019, pp 1106-1125.

[7] J. Pont-Tuset, F. Perazzi, S. Caelles, P. Arbeláez, A. Sorkine-Hornung, L. V. Gool, “The 2017 DAVIS Challenge on Video Object Segmentation,” arXiv preprint, arXiv : 1704.00675.

†早稲田大学大学院　基幹理工学研究科　情報理工・情報通信専攻

〒140-0001 東京都新宿区大久保 3-14-9 早大シルマンホール 401 号室

TEL.080-1337-2708　E-mail: lvpurin@fuji.waseda.jp