6X-03

事前学習済みの拡散モデルを使用したフレーム補間

渡部泰樹† 進藤嵩紘† 巽優衣† 渡辺裕†
早稲田大学大学院 基幹理工学研究科†

1. まえがき

フレーム補間とは，2 つの時系列で離れたフレームからその間の 1 フレーム，もしくは複数のフレームを補間する技術である．フレーム補間は，スローモーション生成，高フレームレートの動画生成，映像圧縮に応用されている．従来は，CNN を活用したフローベース，もしくはカーネルベースのフレーム補間手法が主流であった．最近では，画像，動画や音声生成において革新的な成果を出している拡散モデルを応用したフレーム補間手法も提案されている．本研究では，事前学習済みの動画生成用拡散モデルである Stable Video Diffusion (SVD) [1]を使用して先行研究よりも高速かつ高品質にフレーム補間を行う手法を提案する．さらに，実験より本手法の有効性を検証する．

2. 関連研究

2.1. 拡散モデル

拡散モデルは，拡散過程と逆拡散過程という 2 種類のプロセスを踏むような深層生成モデルである．拡散過程では，入力データに徐々にガウシアンノイズを加えていき，逆拡散過程では，ノイズが加えられたデータから，そのノイズをニューラルネットワークによって予測することで，そのノイズを除去していく．この学習によって，推論時にはランダムノイズから真のデータ分布に近い画像や動画を生成することができる．

2.2. Stable Video Diffusion (SVD) を使用したフレーム補間手法

Stable Video Diffusion (SVD) は，テキストから動画を生成するような拡散モデルである．SVD は，テキストから画像を生成する Stable Diffusion を元に設計されており，Stable Diffusion に時間的な畳み込み層とアテンション層を挿入し，高品質な動画のデータセットを用いてファインチューニングすることで，画像から動画の生成を可能にしている．SVDKFI[2]では，最初のフレームからその先のフレームを予測する通常の SVD に加えて，最後のフレームからその前のフレームを予測する SVD の 2 種類を利用して複数枚のフレーム補間を実現している．後者は通常の SVD のアテンションマップを 180 度回転させ，時系列処理に関わる一部の層をファインチューニングすることで，逆方向の SVD を作成している．推論時では，各タイムステップで，2 種類の SVD で生成した順方向と逆方向のノイズを合算することで，時間的に一貫したフレームの生成を可能にしている．

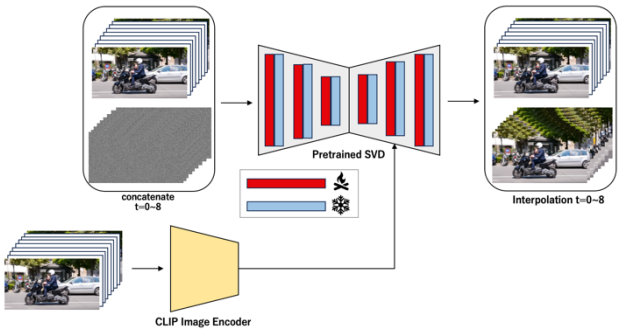

図 1 提案手法の概略図

3. 提案手法

既存手法である SVDKFI と同様に SVD を使用して，最初と最後の入力フレームから複数枚のフレームの補間を行う．SVDKFI は，2 種類の SVD を推論時に使用しているため，推論に時間がかかる．SVDKFI で使用している逆方向のフレームを生成する SVD の学習では，豊富なデータで事前に学習されている順方向の SVD の潤沢な知識を失わないために，一部の層のみをファインチューニングしている．しかしながら，学習が安定せず，順方向のものよりも生成されたフレームの品質が悪い．提案手法では，事前に学習されている SVD の潤沢な知識を効率的に使用するために，順方向の SVD に対して，最後のフレームも条件付けることで，フレーム補間を実現する．SVDKFI と同様，時間的なアテンション層のみをファインチューニングする．推論時にはこの一つの SVD のみを使用しているため，SVDKFI に比べ，推論時間が短縮される．単に最後のフレームを条件付けることで通常の SVD に

Video Frame Interpolation Using Pretrained Diffusion Model
† Taiju Watanabe, Shindo Takahiro, Yui Tatsumi and Hiroshi Watanabe

表１　DAVIS データセットの評価結果

Method	PSNR↑	SSIM↑	LPIPS↓	FID↓	FVD↓
FILM	19.68	0.6458	**0.2768**	**11.53**	341.1
AMT	**20.16**	**0.6729**	0.3160	25.06	391.5
SVDKFI	16.25	0.5265	0.3781	27.14	300.03
Ours	17.51	0.5790	0.3368	24.00	**198.5**

制約を加えており，SVD の豊富な知識を損なわないと考える．提案手法の概略について，図１に示す．推論時は，ランダムノイズと最初と最後のフレームを結合し，それがモデルである 3D-UNet の入力となる．3D-UNet はその入力に加えて，CLIP Image Encoder で得られた最初と最後のフレームの Embedding も中間の層で利用している．図 1 のようにランダムノイズから全てのフレームを再構成するように推論が進められる．

４．　実験

4.1 実験条件

訓練データセットとして，OpenVid-1M データセット[3]を使用して，100k step 学習を行う．最初と最後の 2 フレームを条件付け，9 フレームを再構成するような学習を行う．時間的なアテンション層（temporal attention block）のみファインチューニングする．テストデータセットには，DAVIS データセット[4]（448x832 ピクセル）を使用する．学習時と同様，最初と最後の 2 フレームを条件付け，９フレームを生成する．評価指標としては，PSNR, SSIM, LPIPS, FID, FVD を採用する．比較のために，CNN ベースのフレーム補間手法である FILM[5]，AMT[6]と，SVD ベースの SVDKFI についても評価をする．

4.2 実験結果

表１に DAVIS データセットの評価結果を示す．表１から，PSNR，SSIM，LPIPS，FID においては CNN ベースの FILM や AMT が優れている．これは，CNN ベースの手法では，ピクセルベースでフレームの予測を行っているが，SVD ベースでは VAE を通し，潜在空間上で予測をおこなっているため，PSNR などピクセル間の差分で評価する指標において CNN ベースが秀でる．しかし，フレーム間の時間的な一貫性においては，SVD ベースの方が優れており，それを考慮する FVD では我々の提案手法が比較手法よりも優れていることがわかる．図２に DAVIS データセットの生成結果（間の 3 フレーム）について示す．図から，CNN ベースでは物体が崩壊しており，SVDKFI では物体の動きが少ないことがわかる．

図２　DAVIS データセットの生成結果
（t=3～5）

提案手法では一部物体が崩壊している部分もあるが，比較手法よりも視覚的にみても良いことがわかる．

５．　むすび

本研究では，SVD を使用したフレーム補間について検討した．順方向の SVD をファインチューニングすることで，SVD の豊富な知識を損なわずにフレーム補間を可能にした．実験から提案手法が時間的な一貫性を保ったフレームの生成が可能であることを確認した．

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（JPJ012368C05101）により得られたものです．

参考文献

[1] A. Blattmann et al., “Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets,” arXiv preprint arXiv:2311.15127, 2023.
[2] X. Wang et al., “Generative Inbetweening: Adapting Image-to-Video Models for Keyframe Interpolation,” arXiv preprint arXiv:2408.15239, 2024.
[3] K. Nan et al., “OpenVid-1M: A Large-Scale High-Quality Dataset for Text-to-video Generation,” arXiv preprint arXiv:2407.02371, 2024.
[4] J. Pont-Tuset et al., “The 2017 DAVIS Challenge on Video Object Segmentation,” arXiv preprint arXiv:1704.00675, 2017.
[5] F. Reda et al., “FILM: Frame Interpolation for Large Motion,” ECCV, pp. 4614-4618, 2022.
[6] Z. Li et al., “AMT: All-Pairs Multi-Field Transforms for Efficient Frame Interpolation,” CVPR, pp. 9801-9810, 2023.