IS1-123

Dual Attention機構と動画拡散モデルを活用した長尺動画のカラー化

木村倫太郎, 柳井啓司（電気通信大学）

概要

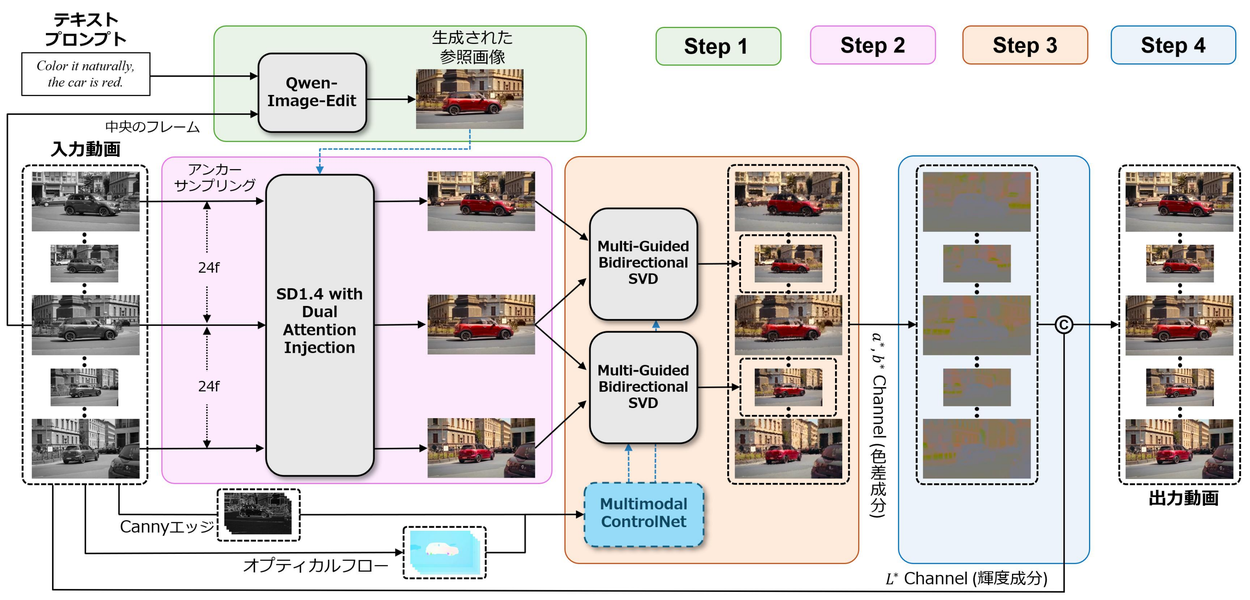
背景・課題
拡散モデルによるカラー化は，自然な発色に優れるが，全フレーム一括生成ではメモリ使用量が動画長に比例し，またフリッカーや色ドリフトも課題.

研究目的
テキストで色を指定でき，長尺の白黒動画でも，一貫性のある自然なカラー化を，一定のピークメモリ使用量で実現する.

アプローチ
中央フレームから参照画像を生成 → 24フレーム間隔のアンカーへ色転写 → 双方向SVDで区間補完 → Lab色空間で入力輝度を復元.

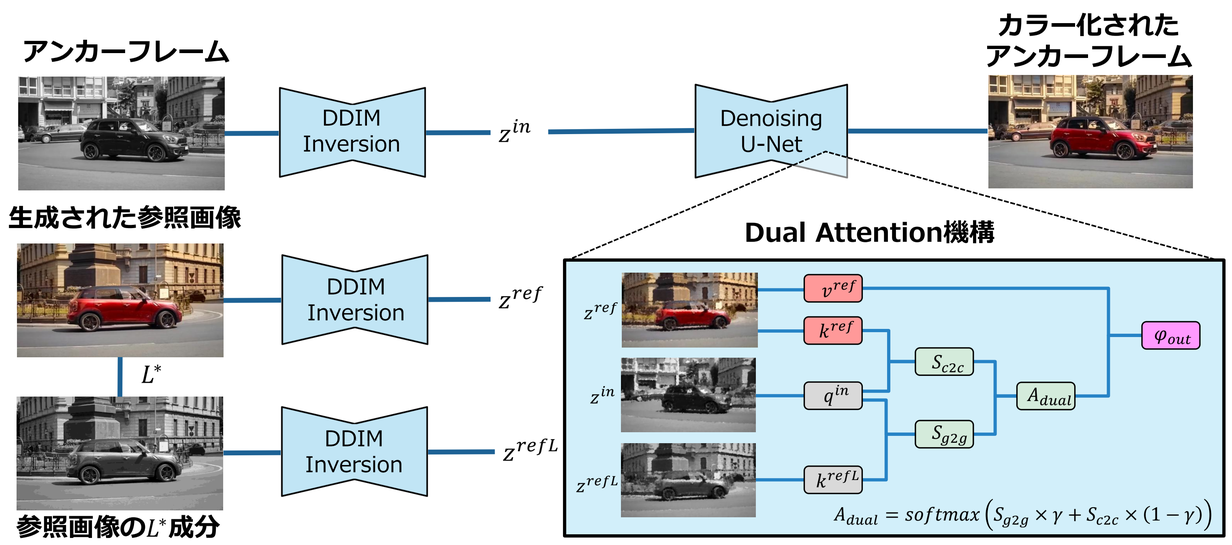
結果
DAVIS30：PSNR 26.28 / LPIPS 0.117，Videvo：27.38 / 0.107.
動画長Tに対して，ピークメモリ$O(1)$，計算量$O(T)$.

提案手法



Step 1｜テキストから参照画像を生成
Qwen-Image-Edit-2509で中央フレームと入力テキストプロンプトから後続の処理で用いる参照画像となるカラー画像を生成する.

Step 2｜アンカーフレームへ色を転写
疎なアンカーフレームを個別にカラー化する.
入力動画から24フレームごとにアンカーフレームを抽出し，アンカーフレーム，Step 1で生成した参照画像，参照画像のL^*成分画像をDDIM Inversionする.
構造情報の対応を見るS_{g2g}と，色情報の対応を見るS_{c2c}のAttentionを重み付き融合しながらデノイズし，カラー化されたアンカーフレームを得る.

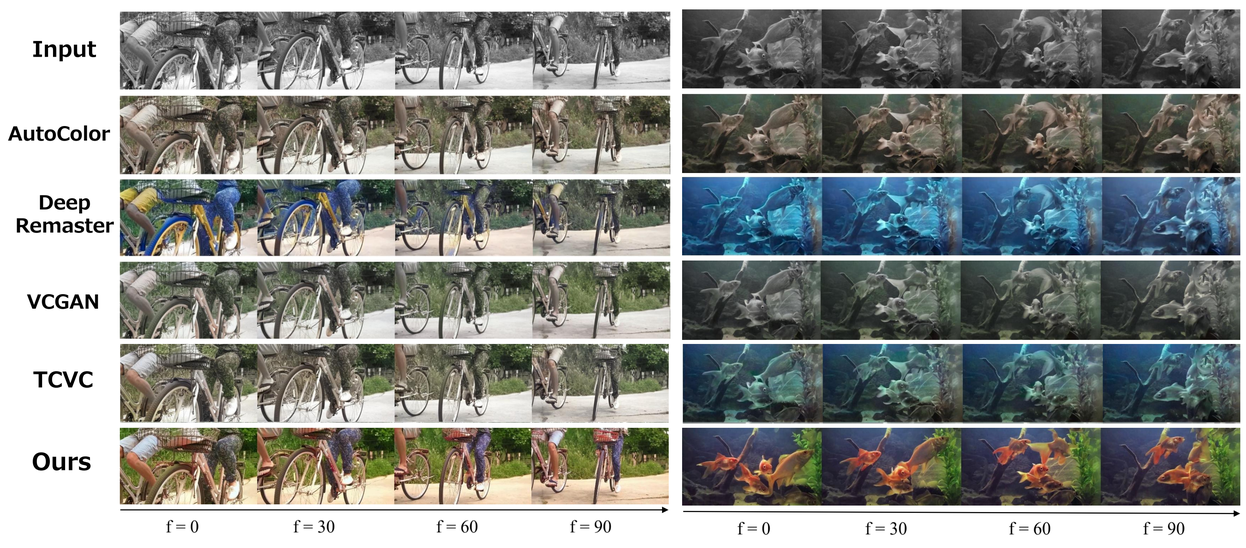


Step 3｜双方向SVDで中間フレームを補完
隣り合う2つのアンカーフレームを入力とし，その間のフレームを補間.
順方向だけでなく，逆方向からも推論を行うStable Video Diffusionで，前後のフレームと齟齬のない補間が可能.
条件入力として，入力動画のCannyエッジとオプティカルフローの情報を統合したマルチモーダルControlNetを用い，元動画の構造を維持する.

Step 4｜Lab空間で入力の細部を復元
Step 3までで生成した動画をLab色空間へ変換し，その色差情報だけを含むa^*, b^*チャンネルを抽出する．元動画から輝度成分だけを含むL^*チャンネルを抽出し，これらをLab色空間上で統合する．これにより，生成動画の豊かな色合いと，元動画の詳細な構造を両立させる.

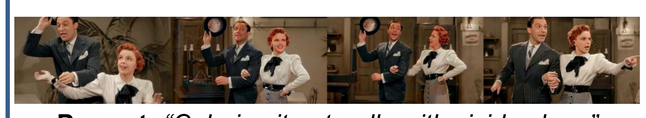
結果

従来手法との比較



提案手法は終始一貫した自然なカラー化を実現.

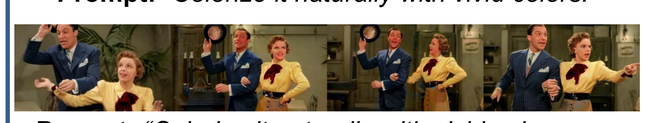
プロンプトによる色の制御
テキストプロンプトにより，任意のカラー化が可能.



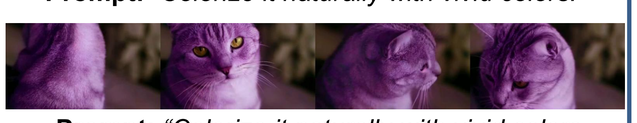
Prompt: *“Colorize it naturally with vivid colors.”*



Prompt: *“Colorize it naturally with vivid colors.”*



Prompt: *“Colorize it naturally with vivid colors. The man’s suit is **blue**, and the woman’s shirt is **yellow**.”*



Prompt: *“Colorize it naturally with vivid colors. The cat has **purple** fur and **yellow** eyes.”*

長尺動画への適応
数百フレームを超える動画でも一定のピークメモリで安定して処理可能.



定量評価
DAVIS dataset (一般的な評価データセット)**:**

Method	PSNR ↑	LPIPS ↓	FID ↓	Colorfulness↑
AutoColor (2019)	24.41	0.264	83.05	14.14
DeepRemaster (2019)	21.95	0.354	97.54	25.66
VCGAN (2021)	23.90	0.247	70.29	15.89
ColorDiffuser (2025)	23.73	0.213	<u>69.51</u>	29.13
L-C4 (2026)	<u>25.69</u>	0.209	-	29.33
VanGogh (2025)	23.20	<u>0.191</u>	-	**60.09**
Ours	**26.28**	**0.117**	**61.50**	<u>43.51</u>

Videvo dataset (長尺動画を含むデータセット)**:**

Method	PSNR ↑	LPIPS ↓	FID ↓	Colorfulness↑
AutoColor (2019)	<u>25.90</u>	0.277	76.28	13.23
DeepRemaster (2019)	21.88	0.358	86.23	28.72
VCGAN (2021)	24.67	0.276	**63.83**	14.90
ColorDiffuser (2025)	25.27	0.205	66.11	20.73
L-C4 (2026)	25.17	<u>0.198</u>	-	<u>32.59</u>
VanGogh (2025)	-	-	-	-
Ours	**27.38**	**0.107**	<u>65.53</u>	**38.98**

両データセットでPSNRとLPIPSが最良となり，正解動画に対して高い画素・知覚的類似度を達成.
FIDと色の鮮やかさを表すColorfulnessも良好であり，提案手法が視覚的な品質と豊かな発色を両立できることを示した.