CONFIDENTIAL EXTENDED ABSTRACT.
DO NOT DISTRIBUTE ANYWHERE.

AniDepth: 深度画像に基づく線画変形を用いた条件付き拡散モデルによるアニメ中割手法

古賀 荘翠[1,a)]　舩冨 卓哉[1]　品川 政太朗[1]　藤村 友貴[1]
北野 和哉[1]　久保 尋之[2]　前島 謙宣[3]　向川 康博[1]

概要

アニメ制作の中割り作業は高い専門性と多大な作業コストを要するため，その自動化が望まれている．本稿では，動画拡散モデルを用いたアニメの中割りを支援する補間フレーム生成手法 “AniDepth” を提案する．AniDepth では，彩色後のキーフレーム補間に線画を用いることで実際の制作工程に近い条件設定で補間結果の品質の向上を行う．キーフレームを深度画像に変換し，アニメと実写のドメインギャップを埋めるモダリティに変換することで，動画拡散モデルの事前知識の活用を図る．実験により，事前学習済みの動画拡散モデルが明示的な深度画像による学習を介さずとも補間を実現し，定量的指標で既存手法を上回り視覚的破綻の低減が得られることを確認した．

1. はじめに

アニメ制作においては，動作の大枠を定義する原画を作成した後，原画同士を滑らかにつなぐ**中割り**が作成される．しかし，中割り作業は高度な専門技術を要し，多量のフレームを限られたスケジュール内で描く必要があるため，制作現場にとって大きな負担となっている．そこで本研究では，アニメ制作における中割作業の補助を実現できるよう，与えられた 2 枚の原画に対する滑らかな補間を実現することを目的とする．

フレーム補間技術として，実写動画を対象としたオプティカルフロー推定に基づく手法 [4], [9] や，動画拡散モデルを用いる手法 [5], [6], [12] が提案されており，高い補間性能を示している．これらの技術を中割作成に用いる際には，**原画**と呼ばれる線画のキーフレームを描いた後にこれらを彩色し，彩色済みの**アニメ画像**を補間する流れを取ることが多い．しかし実制作において使用されるアニメの画像では，領域ごとの平坦な彩色や単一の輪郭線が多用されるなど，実写画像とは異なる特徴がある．そのため，既存の補間手法を直接適用すると，輪郭線の崩れや色の滲みなど，アニメ画像として許容できない破綻が発生しやすい．アニメ画像で調整された動画拡散モデル [11] を用いたとしても，キャラクタの大きな変形や急激な動きが伴うシーンでは滑らかに補間されず，結果としてキーフレームとほとんど変わらない中間フレームや，輪郭線・彩色の破綻が目立つ結果が得られる．これらの問題により，実際のアニメ制作に利用可能な中割り画像が制限される．

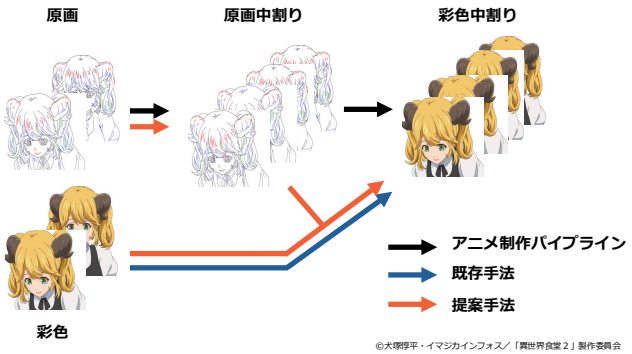


図 1　アニメ制作工程とフレーム補間手法の比較概要．

図 1 にアニメ制作工程と既存のフレーム補間手法と提案手法の概要の比較を示す．一般に，通常のアニメ制作では，原画の間を補間する**線画の中割**を行い，最後に**全線画の彩色**が行われる．この工程は，中割フレームでアニメ画像としての特徴を維持するのに重要であるが，既存のフレーム補間技術の多くはこの工程を踏んでいない．そこで本研究では，近年注目されている動画拡散モデルを用いる手法をベースに，線画に対する中割工程をフレーム補間処理に組み込み，中割フレームがアニメ画像として特徴を維持するよう条件付けることで，自動生成された中割フレームの品質向上を試みる．

2. 提案手法：AniDepth

本研究では，アニメ画像を深度画像に変換することで，実写画像ドメインで学習された動画拡散モデルが獲得した空間的事前知識を有効に活用可能であるという仮説を立

1　奈良先端科学技術大学院大学
2　千葉大学
3　OLM Digital, Inc. IMAGICA GROUP, Inc.
a)　koga.sosui.ko9@naist.ac.jp

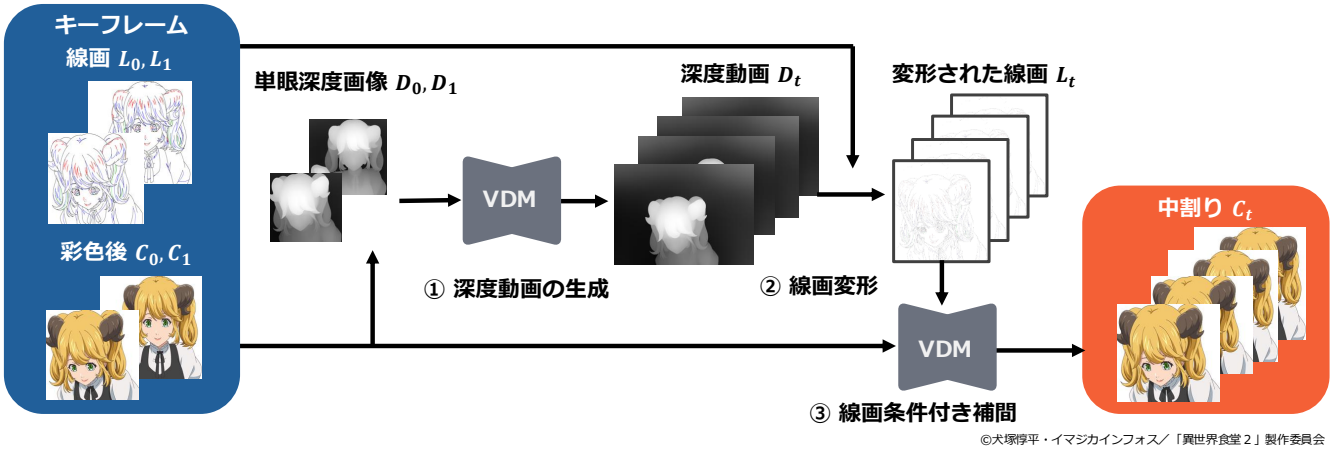


図 2　提案手法パイプラインの概要

てる．

以上の仮説を踏まえ，提案手法では基盤モデルを用いてアニメ画像を深度化することで，動画拡散モデルを用いてアニメ特有の表現かつ大きな変形があるケースでも滑らかな補間を実現する手法を提案する．図 2 に提案手法のパイプラインを示す．提案手法は以下で述べる 3 つのステップから構成される．

2.1　深度動画の生成

彩色後キーフレーム C_0, C_1 からそれぞれ深度画像 D_0, D_1 を推定する．

$$D_0 = \mathcal{F}_{\text{Depth}}(C_0), \quad D_1 = \mathcal{F}_{\text{Depth}}(C_1) \tag{1}$$

ここで，$\mathcal{F}_{\text{Depth}}$ は単眼深度推定モデルを表す．その後，動画拡散モデル $\mathcal{V}_{\text{Diff}}$ を用いて，深度化したキーフレームの補間を行い深度動画を生成する．時刻 t における補間深度画像 D_t は以下のようにして得らえる．

$$D_t = \mathcal{V}_{\text{Diff}}(D_0, D_1, t) \tag{2}$$

2.2　線画変形

深度画像にすることで滑らかな補間結果が得られるが，輪郭線やその内部の色は失われる．これらを維持した補間を行うため，生成された深度動画に対してオプティカルフローを推定し，これに基づいて線画を変形する．

得られた深度動画 D_t に対してオプティカルフロー推定器を適用する．例えば，深度キーフレーム D_0 から中割フレームへのオプティカルフローは以下のように表す．

$$\omega_{0\to t} = \mathcal{F}_{\text{Flow}}(D_0, D_t) \tag{3}$$

このフロー $\omega_{0\to t}$ を用いて，線画キーフレーム L_0 を中割フレームに変形する．

$$L_t = \mathcal{W}(L_0, \omega_{0\to t}) \tag{4}$$

ここで，$\mathcal{W}$ は変形演算子である．彩色後のキーフレームを直接変形するのではなく，線画を変形することにより，消失や混色を防ぎ線画の単一色のままにすることが可能である．

2.3　線画条件付き彩色補間

彩色後キーフレーム C_0, C_1 と，前段で変形した線画 L_t を条件として，拡散モデルにより中割フレームの彩色画像 C_t を生成する．

$$C_t = \mathcal{V}^{\mathcal{S}}_{\text{Diff}}(C_0, C_1, L_t) \tag{5}$$

ここで，$\mathcal{V}^{\mathcal{S}}_{\text{Diff}}$ は線画を条件とした動画拡散モデルである．これにより，補間した線画を考慮した線画の詳細な構造維持と彩色後のキーフレームへの高忠実度の両立を実現する．

2.4　実装

深度化に用いる基盤モデルには Depth AnythingV2[13]，オプティカルフロー推定器には DOT[8] を採用した．事前学習済みの条件付き動画拡散モデルは，以下に示す目的関数に基づいて学習されている．

$$\min_{\theta} \mathbb{E}_{\mathcal{E}(\mathbf{x}),\,\mathbf{s},\,t,\epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[\|\epsilon - \epsilon^{\mathcal{S}}_{\theta}(\mathbf{z}_t;\,\mathbf{c}_{\text{img}},\,\mathbf{c}_{\text{txt}},\,\mathbf{s}',\,t,\,f_{\text{ps}})\|_2^2\right] \tag{6}$$

ここで，$\epsilon^{\mathcal{S}}_{\theta}$ は動画拡散モデルと線画エンコーダ $\mathcal{S}$ を組み合わせたモデル出力である．$\mathbf{s}$ は線画に対応する．

また，動画拡散モデルを最大限活用するため，彩色後キーフレームと変形された線画は事前学習時と同じ解像度（512×320）にリサイズして入力する．しかし，一般的なアニメ画像をこの解像度に縮小すると，輪郭線が途切れたり平滑化によって色にグラデーションが生じることがある．そこで，縮小後に改めて線画を重ね合わせるなどの工夫によって品質の向上を図っている．

表 1　各手法の定量評価結果

手法	PSNR(↑)	LPIPS(↓)
(A) 線画ガイド無し拡散モデル [11]	18.08	0.1705
(B) 提案手法 AniDepth	**18.79**	**0.1595**
(C) 彩色後画像の変形	15.12	0.2656

3. 実験

3.1 既存手法との比較

提案手法 “AniDepth” の有効性を検証するため，3 体の異なるキャラクタや多様な動作シーンを含むアニメ画像合計 130 サンプルを用いて評価した．各サンプルは中割り作業済み動画と背景との合成前の段階で取得しており，隣接する 3 フレームをスライディングウィンドウ処理し，両端のフレームをキーフレーム，中央のフレームを参照フレームとして抽出している．動画拡散モデルでは 16 フレームを生成するが，そのうち真値と対応すると仮定した 7 フレーム目を参照フレームとして比較した．比較対象として，以下の 3 つの手法を用いる．

(A) 線画ガイド無し動画拡散モデル [11]

(B) 提案手法（AniDepth）

(C) オプティカルフローによる彩色済み画像の変形

$$D'_t = \mathcal{W}(D_0, \omega_{0\to t})$$

(A) と (B) の比較においては，テキスト条件づけや乱数シードなどのハイパーパラメータは共通とし，線画ガイドの有無だけを変化させている．定量評価として Peak Signal-to-Noise Ratio (PSNR) と Learned Perceptual Image Patch Similarity (LPIPS) などを算出した結果を表 1 に示す．

表 1 より，提案手法 (B) は (A) に比べて PSNR が高く，LPIPS が低い．これは誤差の低減と視覚的類似度の向上を意味する．一方，(C) の単純な変形法は大きく差をつけられており，線画条件付き補間を用いる有効性を示している．定性的な結果を図 3 に示す．既存手法 [11] では，多くのフレームに動きがほとんど見られず，中間フレーム付近で急激な破綻が生じている．一方，AniDepth は輪郭線の連続性を維持し色の変化・消失が抑制されている．

3.2 シーンダイナミクスに対する頑健性の検証

提案手法のキャラクタ変形の大きさに対する頑健性を調べるために，Chen ら [3] で導入された動作スケール指標 RRLD を用いて詳細な評価を行った．

$$\mathrm{RRLD}(\omega_{0\to t}, \omega_{1\to t}) = \frac{1}{|\Omega|}\sum_{(i,j)\in\Omega}\frac{\frac{1}{2}\|\omega_{0\to t}[i,j] + \omega_{1\to t}[i,j]\|}{\|\omega_{0\to t}[i,j] - \omega_{1\to t}[i,j]\|} \tag{6}$$

Ω は画像座標の集合を表し，$\|\cdot\|$ は L2 ノルムである．ここで，$\omega_{0\to t}$ および $\omega_{1\to t}$ はそれぞれキーフレーム 0 とキーフレーム 1 から中割フレーム t へ向かうオプティカルフローである．

図 4 に各指標の傾向を示す．各散布図は，AniDepth で算出された評価値（x 軸）と ToonCrafter で算出された評価値（y 軸）を比較している．プロット点は，各サンプルの RRLD 値を対数スケールのカラーマッピングで表現しており，シーンにおける変化が大きいほど RRLD は低い値をとる．キャラクタの変形が大きい場合でも，AniDepth は大幅な指標の悪化が見られず，動作スケールが大きいシーンにおいても補間性能を維持していることが確認できる．

3.3 深度化によるドメインへの影響の検討

本章ではアニメ画像を深度画像に変換すること（深度化）の効果を検証する．Tang ら [10] は，拡散モデルの U-Net が持つ中間表現が，異なる画像間の対応関係を暗黙的に捉える高い表現力を持つと指摘している．そこで，動画拡散モデルの中間特徴を用いて，深度化によって実写ドメインの知識をどの程度活用できるようになるかを検証した．

データセットとして (a) アニメ画像，(b) a を深度化したもの，(c) 実写画像，(d) c を深度化したもの，(e) 事前学習に利用された動画データセット WebVido-10M[1] を検証に使用した．このデータセットは Web で収集された多様な動画を含んでいるが，そのほとんどは実写動画である．また比較対象として，(c) は異なる実写画像のデータセット DAVIS2017[2] からサンプリングした．

実写動画で事前学習された後，アニメで調整された動画拡散モデル [11] に (a)–(e) の画像を入力し，逆拡散過程を担う U-Net のアップサンプリング第 1 層の特徴量を取得した．特徴量の可視化のため，(a)–(d) の特徴量を対象として低次元化する写像を UMAP[7] を用いて取得した．各ドメイン (a)–(e) を同じ写像を用いて 2 次元で可視化した結果を図 5 に示す．

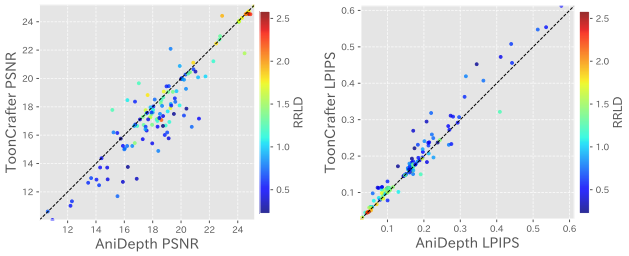
図 5 から，実写画像と実写深度が非常に近接した位置に分布していることが確認された．一方で，実写，アニメ，アニメ深度はそれぞれ異なるクラスに分類されることから，アニメ画像の深度化が実写ドメインの知識を十分に活用しているとは言い難い．さらに，アニメ深度のデータドメインが実写およびアニメの双方から離れた領域へシフトしていることも明らかとなった．以上の結果を踏まえると，現段階ではアニメ画像の深度化が動画拡散モデルの事前知識を有効に活用できるという明確な根拠は得られなかった

4. 結論

本稿では，動画拡散モデルを用いる手法をベースに，線画に対する中割工程をフレーム補間処理に組み込み，中割フレームがアニメ画像として特徴を維持するよう条件付ける手法 “AniDepth” を提案した．また，アニメ画像を深度画像へ変換することで，大きな変形があっても動画拡散モ

©犬塚惇平・イマジカインフォス／「異世界食堂 2」製作委員会

図 3　提案手法によるアニメ中割生成例.



(1) PSNR(↑)
右下領域で提案手法が優位

(2) LPIPS(↓)
左上領域で提案手法が優位

図 4　RRLD 値と補間性能の優劣との関係性.

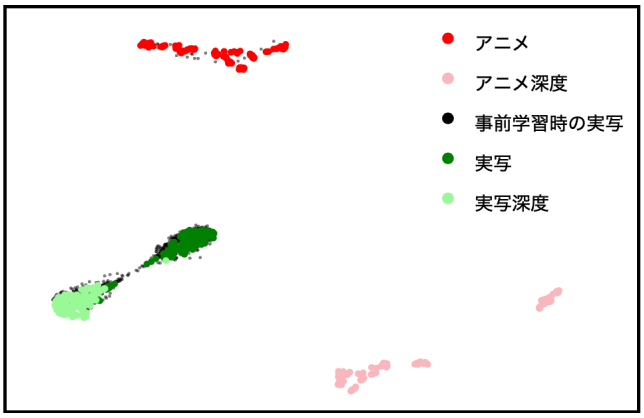


図 5　次元削減による動画拡散モデル中間表現の可視化

デルを用いて滑らかに補間できることを示した．本手法を用いて，実際のアニメ制作現場で採用可能な中割りの選択肢を増やすことが可能である．また，本手法は追加学習を必要とせず，既存のアニメ制作パイプラインに導入できる利点がある．今後は，より精密なオプティカルフロー推定手法や線画変形手法との組み合わせによるさらなる品質向上に取り組む．また，深度変換による補間精度向上の要因および動画拡散モデルの事前知識の活用について，さらなる検証が必要である．

謝辞

研究目的にイラストの使用許諾をいただいた犬塚惇平氏に感謝の意を表する．本研究は，経済産業省と国立研究開発法人新エネルギー・産業技術総合開発機構（NEDO）が実施する，国内の生成 AI の開発力強化を目的としたプロジェクト「GENIAC（Generative AI Accelerator Challenge）」(課題番号 [JPNP20017]) の成果をもとに作成されました．

参考文献

[1] Bain, M., Nagrani, A., Varol, G. and Zisserman, A.: Frozen in Time: A Joint Video and Image Encoder for End-to-End Retrieval, *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1708–1718 (2021).

[2] Caelles, S., Pont-Tuset, J., Perazzi, F., Montes, A., Maninis, K.-K. and Van Gool, L.: The 2019 DAVIS Challenge on VOS: Unsupervised Multi-Object Segmentation, *arXiv:1905.00737* (2019).

[3] Chen, S. and Zwicker, M.: Improving the Perceptual Quality of 2D Animation Interpolation, *Proceedings of the European Conference on Computer Vision* (2022).

[4] Huang, Z., Zhang, T., Heng, W., Shi, B. and Zhou, S.: Real-Time Intermediate Flow Estimation for Video Frame Interpolation, *Proceedings of the European Conference on Computer Vision (ECCV)* (2022).

[5] Huang, Z., Yu, Y., Yang, L., Qin, C., Zheng, B., Zheng, X., Zhou, Z., Wang, Y. and Yang, W.: Motion-aware Latent Diffusion Models for Video Frame Interpolation, *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, New York, NY, USA, Association for Computing Machinery, p. 1043–1052 (2024).

[6] Liang, J., Fan, Y., Zhang, K., Timofte, R., Van Gool, L. and Ranjan, R.: MoVideo: Motion-Aware Video Generation with Diffusion Model, *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part XLIV*, Berlin, Heidelberg, Springer-Verlag, p. 56–74 (2024).

[7] McInnes, L., Healy, J., Saul, N. and Großberger, L.: UMAP: Uniform Manifold Approximation and Projection, *Journal of Open Source Software*, Vol. 3, No. 29, p. 861 (2018).

[8] Moing, G. L., Ponce, J. and Schmid, C.: Dense Optical Tracking: Connecting the Dots, *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE/CVF, pp. 19187–19197 (2024).

[9] Reda, F., Kontkanen, J., Tabellion, E., Sun, D., Pantofaru, C. and Curless, B.: FILM: Frame Interpolation for Large Motion, *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VII*, Berlin, Heidelberg, Springer-Verlag, p. 250–266 (2022).

[10] Tang, L., Jia, M., Wang, Q., Phoo, C. P. and Hariharan, B.: Emergent correspondence from image diffusion, *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, Curran Associates Inc. (2023).

[11] Xing, J., Liu, H., Xia, M., Zhang, Y., Wang, X., Shan, Y. and Wong, T.-T.: ToonCrafter: Generative Cartoon Interpolation, *ACM Trans. Graph.*, Vol. 43, No. 6 (2024).

[12] Xing, J., Xia, M., Zhang, Y., Chen, H., Yu, W., Liu, H., Liu, G., Wang, X., Shan, Y. and Wong, T.-T.: DynamiCrafter: Animating Open-Domain Images with Video Diffusion Priors, *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part XLVI*, Berlin, Heidelberg, Springer-Verlag, p. 399–417 (2024).

[13] Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J. and Zhao, H.: Depth Anything V2, *Advances in Neural Information Processing Systems* (Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J. and Zhang, C., eds.), Vol. 37, Curran Associates, Inc., pp. 21875–21911 (2024).