I-024

流れ場に基づくワーピングによる深層生成モデルを用いた人物動画生成
Generating Human Animation Using Deep Generative Model with Flow-based Warping

鬼塚 慎吾[1)]　遠藤 結城[1)]　栗山 繁[1)]
Shingo Onizuka　Yuki Endo　Shigeru Kuriyama

1　はじめに

深層学習を用いた人物の動画生成技術は，アニメーションコンテンツや機械学習用データセットの作成など様々な用途が期待されている．代表的な手法として，Chan ら [1] は対象の人物動画とポーズ動画のペアを学習することで，その人物の任意の姿勢を高品質に再現している．しかし，再現したい人物の動画を学習データとして十分に用意できるとは限らない．

そこで本研究では，Si らの敵対的生成ネットワーク (GAN) [2] をベースに，1 枚の人物画像から異なる姿勢の画像を高品質に生成する手法を検討する．ベースとなるモデルは任意の人物画像を扱えるように設計されているものの，結果画像を直接生成する学習をしているため，学習していない人物の顔や服装を忠実に再現するには限界がある．この問題に対処するため，画像を直接生成するのではなく，入力画像のワーピングを流れ場として推定する間接的な手法を検討する．流れ場を基に入力画像のみを使って画像を再構成すれば，その人物のディテールを損なうことなく，異なるポーズが再現されることが期待できる．実データを用いた既存手法との比較実験を通して，その有効性を検証する．

2　関連研究

Chan ら [1] の手法は人物の高品質な動画の生成を達成している．しかし学習に使用した特定の人物しか生成できず，人物ごとに動画データを学習させる必要がある．生成したい人物の動画が手に入る状況は限られているため，本研究では 1 枚の画像からの人物動画生成を目的とする．

単一画像から人物動画を生成する既存研究もいくつか存在する．Esser ら [3] は U-net に Variational Autoencoder (VAE) を適用したモデルを提案している．入力にはターゲットの姿勢およびソースの人物画像と姿勢を与え，これらの特徴量を中間層で結合することで結果を生成する．また，Si らの GAN を用いた手法 [2] では，ソースの姿勢画像とターゲットの姿勢画像の特徴量の差分と人物の画像の特徴量から結果画像を生成している．しかしこれらの手法は，未知の人物画像に対して，顔や服装を鮮明に再現できない．一方，人物画像生成に流れ場を用いているものとして，Dong ら [4] や Siarohin ら [5] の手法がある．これらは生成モデルの中間層に推定した流れ場を組み込むことで結果を生成している．本研究では流れ場を中間層の特徴量として使うのではなく，入力画像をワーピングするために利用する点で異なる．

3　提案手法

最初にベースとする Si らの手法 [2] について説明する．この手法はまずソースの姿勢画像，ターゲットの姿勢画像，そしてソースの人物画像から，畳み込み層のエンコーダによってそれぞれの特徴量を抽出する．ターゲットとソースの姿勢画像については，各々の特徴量の差分を姿勢変換の特徴量として計算する．姿勢変換の特徴量とソースの人物画像の特徴量を結合し，畳み込みのデコーダで処理することで結果画像を生成する．

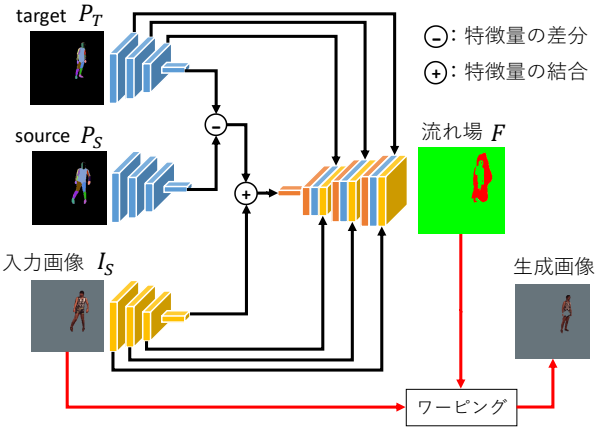


図 1: 提案モデル．赤色の矢印部分は既存手法からの変更部分であり，流れ場 F を推定し，入力画像 I_S をワーピングさせることで画像を生成する．

ベースとなる手法を変更した提案モデルを図 1 に示す．モデルが生成する画像を人物画像から流れ場 F へと変更し，F を用いてソース画像 I_s をワーピングしている．推定する F の各画素は 2 チャンネルの値が $[-1,1]$ で定義されており，活性化関数の tanh により最終層で正規化される．目的関数として，正解の画像とワーピング後の画像との L_1 損失と，推定した F を平滑化する total variation (TV) 損失を使用する．平滑化の導入は，隣り合うピクセルの移動量が極端に変化する状況はさほど多くないと考えられるためである．また推定する流れ場は相対的な移動量を表現しているが，安定した学習をするために極端に大きな移動はないと仮定し，移動範囲を調節するパラメータ m を導入した．ワーピングの際，m を推定した F に掛け合わせたものを最終的な流れ場として使用する．

入力で扱う 2 次元の姿勢画像には，DensePose [6] によって推定した結果を使用する．従来研究では，姿勢画像として OpenPose [7] 等で推定された骨格情報を用いることが多かった．一方，DensePose の場合，骨格情報だけでなく人物の領域を得られるため，今回のタスクにより有用な情報になると考えられる．図 2 に DensePose と OpenPose で推定した姿勢画像の比較を示す．また DensePose の利点として，推定した姿勢画像から人物領域のみを抽出するマスクを作成できる．そこでそのマスクを利用することで，ワーピングによる人物の生成に影響する背景を除き，人物領域のみを対象として学習できるようにする．

1)　豊橋技術科学大学，Toyohashi University of Technology

(a) DensePose による推定　(b) OpenPose による推定

図 2: 姿勢画像の推定手法の比較．入力画像には図 3(b) をマスクする前の画像を用いた．本手法では，骨格だけでなく人物領域も取得できる DensePose の結果を入力の姿勢画像として利用する．

4　画像生成実験

4.1　データセットと実行環境

本研究では Human3.6M [8] の動画データをデータセットとして使用する．Human3.6M は人物が 3D データとしてキャプチャされているデータセットであるが，キャプチャ時の動画も提供されているため，その動画を使用する．画像サイズは 224×224 に縮小した．

学習には一人の人物の動画から 1000 枚の画像を使用した．計算リソースとして，NVIDIA GeForce RTX 2080 Ti を使用した．バッチサイズは 6 とし，4000 エポック学習を行った．学習に要した時間はおよそ 185 時間である．パラメータ m は経験則から 0.1 と設定した．

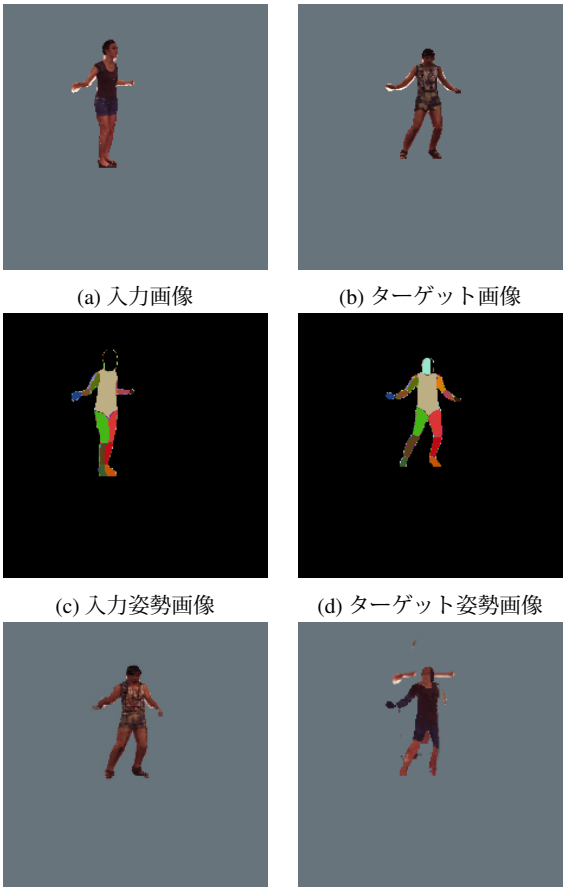
4.2　生成結果

図 3 はデータセットに含まれていない人物画像 (a) とその姿勢画像 (c)，及びターゲットの姿勢画像 (d) から，別姿勢の画像を生成した結果である．ターゲットとなる姿勢画像 (d) は人物画像 (b) から抽出した画像である．提案手法によって推定された流れ場を用いて，(a) をワーピングさせた結果 (f) が得られた．服装や脚などをはじめ，全体的にはソースの人物画像の外見を保持したまま，ターゲットの姿勢に変換されている．しかし，腕の位置や顔の鮮明さなど再現が不十分な箇所も多くみられた．

比較手法として，直接画像を生成する Si らの手法による実験も行った．学習時の条件は提案手法と同じく単一の人物の画像を 1000 枚学習した．なおエポック数は結果画像が十分に再構成されるのを確認したため，100 エポックまでとした．図 3(e) に示す既存手法による結果では，顔や服装，手足の部分が生成されているようにみえる．しかし，一人の人物のみを学習しているため，ソース画像の人物は再現されずに，ターゲットと同じような外見となっている．提案手法では，入力画像からワーピングによって出力画像を再構成するため，図 3(f) の結果のように学習に含まれていない人物でも扱える可能性があるのは，大きなメリットである．

5　まとめと今後の課題

本研究では異なる姿勢の人物画像生成の手法をベースに，直接画像を生成するのではなく画像の変形を流れ場として推定し，ワーピングによって間接的に人物画像を生成する方法を検討した．実験結果から，服装や脚などいくつかの部位の移動を適切に推定できることを確認した．しかし現状の方法では，入力の人物画像の外見を忠実に再現できているとは言えない．一方で，既存手法との比較を通してある程度の改善はみられたことから，本アプローチによる今後の可能性を示すことができた．



(a) 入力画像　(b) ターゲット画像

(c) 入力姿勢画像　(d) ターゲット姿勢画像

(e) 既存手法の生成画像　(f) 提案手法の生成画像

図 3: 提案手法と既存手法による生成結果の比較．

現在の問題点として，ワーピング後の画像同士を比較する損失関数のみから，流れ場を推定するモデルを更新しているため，流れ場の解が多数存在することが学習を困難にしていると考えられる．また，生成画像は入力画像の画素に依存するため，横顔から正面顔を生成するなど，存在しない情報の復元が困難である．これらの問題点を今後の課題とする．

参考文献

[1] C. Chan *et al*. Everybody dance now. *CoRR*, Vol. abs/1808.07371, , 2018.

[2] C. Si *et al*. Multistage adversarial losses for pose-based human image synthesis. In *CVPR*, pp. 118–126, 2018.

[3] P. Esser *et al*. Towards learning a realistic rendering of human behavior. *ECCV*, pp. 409–425, 2018.

[4] H. Dong *et al*. Soft-gated warping-gan for pose-guided person image synthesis. In *NeurIPS*, pp. 472–482, 2018.

[5] A. Siarohin *et al*. Animating arbitrary objects via deep motion transfer. In *CVPR*, pp. 2377–2386, 2019.

[6] R. A. Güler, *et al*. Densepose: Dense human pose estimation in the wild. In *CVPR*, pp. 7297–7306, 2018.

[7] Z. Cao *et al*. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, pp. 1302–1310, 2017.

[8] C. Ionescu *et al*. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Trans. on PAMI*, Vol. 36, No. 7, pp. 1325–1339, 2014.