G-23

エッジ特徴と Optical Flow を用いた動画像表情変換の精度向上
Improving Facial Video Conversion Using Edge Feature and Optical Flow

上野 遥平†　　瀬尾 昌孝†
Youhei Ueno　Masataka Seo

1. はじめに

深層生成モデルによる動画像生成において、動く物体などのフレーム間で変化のある部分がぼやけがちに生成される一方、フレーム間で変化のない部分のエッジやテクスチャがフレームごとに変化して生成されるという問題が存在する。そこで本研究では、生成目標となる本物データと生成データの各フレームのエッジ特徴の誤差をモデルに与えることで、より本物らしいエッジが明瞭な動画像を生成することができるようにモデルを学習させる。それに加えて、本物データと生成データの両方で、各隣接フレームから Optical Flow を算出し、その誤差をモデルに与えることで、フレーム間の各物体の動きがより本物らしい動画像が生成されるようにモデルを学習させる。また、Generator に Self-Attention 層を追加することで、畳み込みによって得られた各特徴マップの重みを調整する。さらに、Discriminator に自己教師あり学習を導入することで、Discriminator が画像特徴をより詳細に捉えられるようにする。

2. 関連研究

2.1　Pix2Pix

Pix2Pix は画像変換を行う生成モデルであり、Generator と Discriminator の 2 つのニューラルネットワークで構成される GAN 構造である[1]。これら 2 つのネットワークを互いに競合するように学習させることで、Generator は入力画像に基づいて変換された画像を生成できる。

2.2　Recurrent U-Net

Recurrent U-Net は U-Net 内の畳み込み層を Recurrent 畳み込み層に変更したネットワークである[2]。Recurrent 構造を畳み込み層に追加することで、時系列データを入力とすることができる。各層の計算では、現時刻の入力に前時刻での畳み込みによって得られた特徴マップを加える。これにより、過去の時刻の入力から得た特徴を利用することができる。

2.3　Self-Supervised Discriminator

Self-Supervised Discriminator では GAN の Discriminator をエンコーダとみなし、中間層の出力を分岐させ、入力画像を再構成するデコーダに接続する[3]。デコーダから出力された再構成画像と入力画像の誤差を Discriminator の損失関数に加えることで Discriminator は自己教師あり学習を行う。学習では敵対的損失と再構成誤差の両方を最小化する。Discriminator は良好な再構成を行うために、入力画像の特徴をよりよく捉えるように学習するため、より正確な真偽判別を行えるようになる。

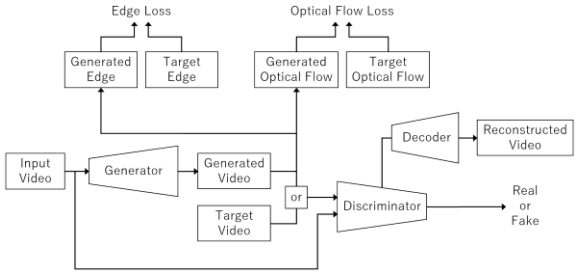

図 1. 提案モデルの構造

3. 提案手法

提案するモデルは、Pix2Pix の Generator の U-Net 構造を Recurrent U-Net 構造に変更することで動画像生成を行う。動画像データの各フレームを Generator に入力することによって、表情変換されたフレームが順番に生成される。モデルの構造を図 1 に示す。

3.1　Edge and Optical Flow Loss

Generator には、新たに 2 つの損失を追加して学習を行う。1 つ目の Edge Loss は、まず目標データと生成データのエッジ特徴を Sobel フィルタによる 1 次微分によって取得する。これら 2 つのエッジ特徴の誤差を Generator に損失として与えることで、目標動画像のような明瞭なエッジを持った動画像を生成するように Generator は学習する。2 つ目の Optical Flow Loss では、まず目標データと生成データの Optical Flow を取得する。Optical Flow とは、隣接フレーム間の物体の動きを 2 次元ベクトルとして表したものである。目標データと生成データの Optical Flow の誤差を Generator に損失として与えることで、各フレーム間での物体の動きがより目標動画像のように生成されるように Generator は学習する。

3.2　Self-Attention

Generator のエンコーダ、ボトルネック層、デコーダにそれぞれ 1 つずつ Self-Attention 層を追加する。Self-Attention 層を追加することによって、下層の畳み込み層で得られた特徴マップの中で、目標の画像生成に重要な特徴量が増幅され、重要度の低い特徴量は減衰される。また、近傍の画素のみから計算する畳み込みに比べ、Self-Attention は入力された特徴マップ全域から計算するため、より大域的な相関関係を捉えることができる。そのため、小さな畳み込みカーネルでは捉えられない、画像の離れた位置にある物体同士の依存関係を捉えることができる。

3.3　自己教師あり学習

Discriminator の中間層の出力から再構成デコーダに接続し、入力画像との再構成誤差を計算する。その値を損失に加えることで、Discriminator は自己教師あり学習を行う。

† 大阪工業大学, Osaka Institute of Technology

Discriminator はエンコーダとして入力画像の再構成を高精度に行えるように入力画像の特徴をよく捉えられるように訓練されるため、画像の細かい特徴をより捉えた上で真偽判別を行うことができるようになる。Generator にはこの強化された Discriminator を騙すために、より本物らしい画像を生成するようになることを期待する。

4. 実験

4.1 データセット

顔動画像の表情変換実験に用いるデータセットとして、動画像データセットを使用する。各データは入力データと生成目標データのペアであり、各データの解像度は 256×256、フレーム数は 15 である。実験に使用する訓練データにはデータ数を 30 倍にするように Data Augmentation を行った。訓練データには 1800 組のデータを使用し、テストデータには 18 組のデータを使用した。

4.2 実験結果

行った実験では、真顔の発話データを入力とし、笑顔の発話データを目標データとした。ベースモデルは Recurrent Pix2Pix とし、動画像変換は Generator に Recurrent 構造を追加することで実現した。そして、ベースモデルを Model 1 として、そのほかに Model 2 として Edge loss と Optical Flow Loss を追加した Recurrent Pix2Pix、Model 3 として Generator に Self-Attention 層と Discriminator に自己教師あり学習を導入した Recurrent Pix2Pix、Model 4 として Edge Loss、Optical Flow Loss、Self-Attention、自己教師あり学習のすべてを追加した Recurrent Pix2Pix の合計 3 つのモデルを用意し、それぞれで生成した表情変換された動画像を比較した。図 2 に入力データと 4 つのモデルの生成結果の全体画像と局所拡大画像を示す。損失の追加によって顔の輪郭が明瞭に生成され、self-attention 層、自己教師あり学習の追加によって肌が滑らかに生成されていることがわかる。

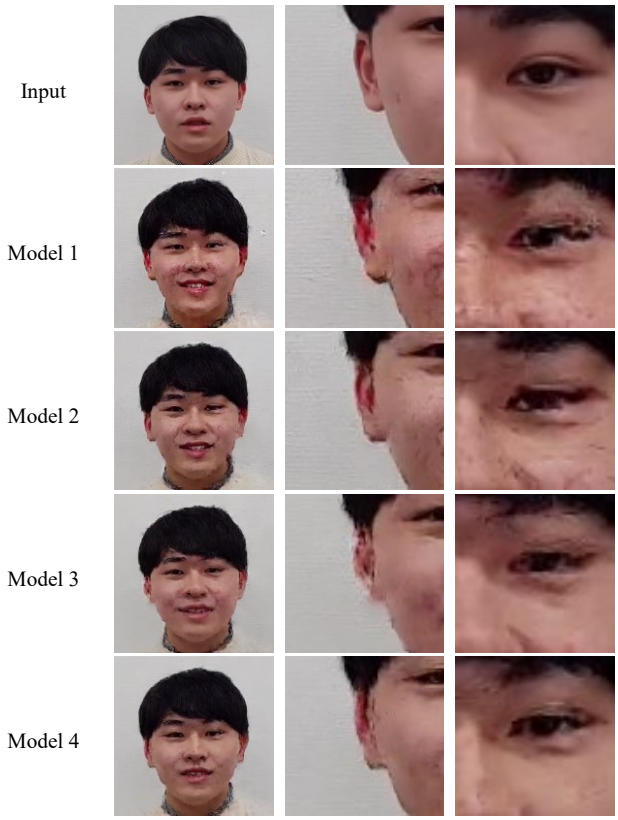

図 2. 表情変換結果

各フレームにおける目標動画像と 4 つのモデルによる生成動画像の PSNR、SSIM をそれぞれ算出した。結果を図 3 に示す。ベースモデルである Model 1 と提案手法をすべて追加した Model 4 を比較すると、Model 4 のほうが全フレームにおいて PSNR、SSIM の値が高かった。

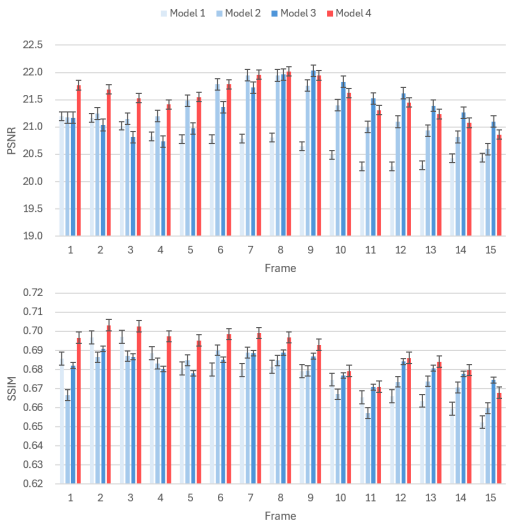

図 3. 目標データと生成結果の PSNR と SSIM

5. 結論

以上の実験より、Generator に Edge Loss と Optical Flow Loss を追加して学習することで、各フレームにおける物体のエッジのぼやけが低減され、フレーム間で変化しない部分のエッジやテクスチャの変化が低減されていることがわかった。目標動画像と各モデルの生成動画像から PSNR、SSIM を算出した結果、ベースモデルに比べ提案手法を追加したモデルでの生成結果がより本物動画像に近い結果であることを確認できた。また、Generator に Self-Attention を追加し、Discriminator に自己教師あり学習を導入することで、生成精度が向上することが確認できた。

今後の展望として、空間方向と時間方向の両方に Attention を導入することが考えられる。そうすることで、直前のフレームだけでなく、未来のフレームの情報も使って動画を生成することができ、フレームによる対象物の変化をより正確に捉えることが可能になる。

参考文献

[1] P. Isola, J.-Y. Zhu, T. Zhou and A. A. Efros, “Image-to-Image Translation with Conditional Adversarial Networks,” In CVPR, 2017.

[2] M. Z. Alom, M. Hasan, C. Yakopcic, T. M. Taha and V. K. Asari, “Recurrent Residual Convolutional Neural Network based on U-Net (R2U-Net) for Medical Image Segmentation,” arXiv:1802.06955, 2018.

[3] B. Liu, Y. Zhu, K. Song and A. Elgammal, “Towards Faster and Stabilized GAN Training for High-Fidelity Few-Shot Iimage Synthesis,” In ICLR, 2021.