7S-05

Transformerを活用したGANによる手振れ補正モデルの検討

大口瑞妃† 佐藤裕二‡
法政大学情報科学部ディジタルメディア学科‡

1. はじめに

本研究では, GAN を利用した手振れ補正モデルの精度向上を目的として, Transformer を活用した手法を提案する. GAN はデータの分布を捉えデータを生成するモデルと, 入力データが, 生成データの実データかを識別するモデルが相互に学習し, 精度を高める. 先行研究では, GAN を発展させ, 手振れ画像から鮮明な画像を生成するモデルが提案された. CNN を利用したモデルには, 画像全体を加味した特徴を捉えることが不得手という課題がある. そこで大域的に特徴を捉えることが可能な Transformer を生成モデルに活用することで, より効果的に特徴を捉えることができる手振れ補正モデルを提案する. 画像の劣化度を示す評価指標により, 提案手法が効率的に学習を行えることを示す.

2. 従来手法の課題

先行研究のモデルの特徴抽出器は, CNN のみであったが, CNN には局所的な情報にしか注目できないという課題がある. そこで, マルチスケール特徴を取り入れることが精度向上に繋がると考える. 大域的な特徴を捉えることが可能なネットワークを CNN とともに Generator に挿入することで, 解決するモデルの提案を考える.

3.手振れタスクへの Transformer の応用

Transformer が大域的な特徴抽出が可能であることから, 本研究では, 手ぶれ補正を目的とした条件付き画像生成モデルの Generator の一部に Transformer を利用したモデルを提案する. Transformer 以外のアーキテクチャとして, 先行研究からスキップ接続があることで精度が向上することが報告されているため, スキップ接続が有用であると考えられる. Transformer とスキップ接続の両方を Generator に持つモデルとして 8 種類の猫の写真を線画から作成する Transpix2pix[1]というモデルが存在する. 本提案手法では Transpix2pix をベースとして, 手振れ補正のために以下に示す拡張を提案する.

- バッチ学習からミニバッチ学習への変更
- 客観的評価指標の実装
- embedding の変更
- Henge loss の導入
- TTUR の導入

提案する Generator ネットワークの構成を図 1 に示す. Generator 全体に Transformer を使うのではなく, 層の学習の過程に挿入することで CNN の利点と Transformer の利点の両方を取りいれ, マルチスケールな特徴の抽出を狙う.

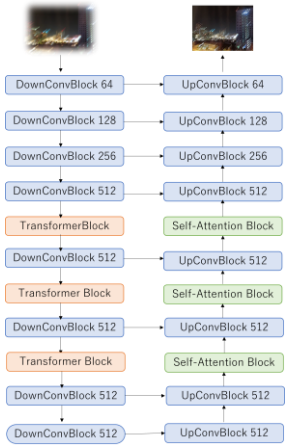

図 1. 提案モデル A の Generator 図

4.1 評価方法

GoPro, DVD, NFS 3 種類のデータセットを合わせて 2 万枚にした手振れ画像と鮮明な画像のペアデータセットを対象に実験を行う. 実験は DeblurGANv2, 初期提案モデル A に加えて, 図 6 の提案モデル B, C の 4 つのモデルに対して行う. テストデータは, 学習で使用したものとパターンの異なる画像を使用する. 評価は画像の劣化度を示す客観的評価指標である PSNR, SSIM と, 実行時間の平均で行う. 実験に関するパラメータ設定は表 2 に示す.

表 2. ハイパーパラメータ

Generator 学習率	0.0003
Discriminator 学習率	0.00001
バッチサイズ	4
epoch 数	120

Image stabilization model using GAN with Transformer
† Mizuki Oguchi†, Yuji Sato
† Faculty of Information Science, Hosei University

図 6. 初期提案モデルの変形
Transformer, Self-Attention の数を 2 つにし, B はデータサイズが大きい側に, C は小さい側に配置した.

4.2 実験結果と考察

表 3 に実験の評価結果を示す. 3 つの提案モデル全てにおいて DeblurGANv2 以上に精度が向上することはなかったが, B が DeblurGANv2 に迫るスコアを記録し, 実行時間に関しては DeblurGANv2 よりも早いことがわかる. A-B の結果を比較すると, B, C, A の順に精度が良い. Transformer, Self-Attention ブロックの数がそれぞれ 2 つの B, C において C の精度が低いことから, データのサイズが小さいダウンサンプリングの終盤やアップサンプリングの初期に, Transformer ブロックや Self-Attention ブロックがあると精度が下がることがわかる.

表 3. 評価結果

	DeblurGAN-v2	A	B	C
PSNR	26.9	22.8	**25.3**	**25.3**
SSIM	0.79	0.72	**0.76**	0.75
TIME	0.063	0.049	0.045	**0.043**

図 7 に DeblurGAN-v2, A-C の各モデルが 3 枚のぶれの程度が異なる画像に対して手振れ補正を行った結果を示す. 上段のぶれの程度が小さい画像に関しては, どのモデルでもブレが軽減されていることが画像内の看板の文字を比較することでわかる. 中段の中程度の画像に関しては, A-C では, 光が当たり薄っすらと明るくなっているような箇所の境界線がはっきりとし, のっぺりとした印象となっている. 下段のぶれが大きい画像に関しては, 中段の画像よりも A-C において輪郭の協調が強くなり, 光が当たっている箇所のグラデーションが失われている.

上記で挙げたぶれが大きい場合にのっぺりとした表現が生じる原因として, 細かな特徴の違いをモデルが認識することができず, 一つのまとまった特徴として抽出し, 過度な表現変換を行ったために情報の欠落が起きていることが考えられる. Transformer, Self-Attention ブロックでは, 画像全体における各セルの特徴の類似度からグループ化するようにして特徴を抽出している. この方法が原因となり, 本研究で使用した手振れデータセットのように細かな特徴をもつデータに対して, 過度な類似点の抽出を行い情報の欠損が起きているのではないかと推測する. 上記の考察と実験結果から, 特にデータサイズが小さいとき, 抽象的に類似度を計算するため, 情報の欠損が起こりやすく, 精度が低下すると考えられる.

下段の画像に関しては DeblurGAN-v2 と比べて大きなぶれのなかで特徴をつかむことが出来ていることがわかる. B, C のモデルは評価結果が DeblurGAN-v2 に近く, 実行速度も速いため, 上記で述べた過度な表現変換を抑える工夫を行うことで, DeblurGAN-v2 よりも優れたモデルを作成できる可能性が高い.

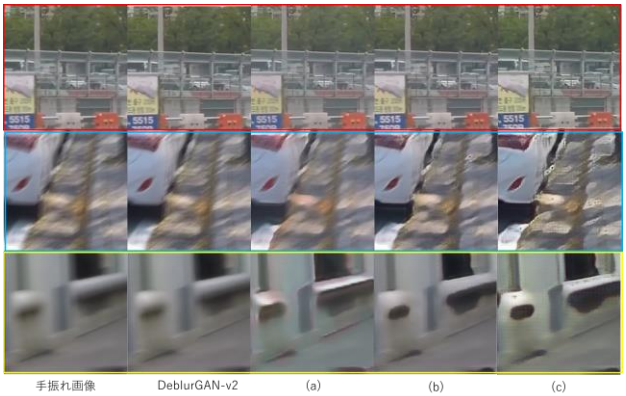

図 7. 各モデルにおける 3 種類の手振れ補正結果

5. むすび

本研究では, GAN に Transformer を利用した手振れ補正モデルの提案を行い, 手振れ精度の向上を目指した. 実験では, DeblurGAN-v2 と比べて提案モデル全てで精度が低くかった. しかし, Transformer の類似度から特徴を抽出する機構が, 過度な表現変換を行っていると推測されることから, この現象を軽減するアーキテクチャを考察することで優れたモデルとなる可能性を示した. 今後の課題として, GAN に Transformer を挿入する際に, 過度な表現変換を抑えるアーキテクチャを調査する必要がある.

文献

[1] Artem-gorodetskii,TransPix2Pix,https://github.com/artemgorodetskii/TransPix2Pix, 2022. Accesed. 13. Jan. 2023.