アニメキャラクターの顔画像から全身画像への画像翻訳手法の検討

斎藤健三郎† 清 雄一† 田原 康之†† 大須賀昭彦††

† 電気通信大学情報理工学域 〒 182-8585 東京都調布市調布ヶ丘 1-5-1
†† 大須賀・田原・清研究室 〒 182-8585 東京都調布市調布ヶ丘 1-5-1 電気通信大学 西 10 号館 728 号室
E-mail: †s1910271@edu.cc.uec.ac.jp, ††{seiuny,tahara,ohsuga}@uec.ac.jp

あらまし 近年, イラスト自動生成手法が劇的に進化している. イラストレーターの仕事が奪われるという声がある反面, アイデア捻出の助けになっているという意見もある. 本研究では, イラスト初心者にありがちな「顔しか描けない」という状態から, 全身のイラストを描くための手助けになるよう, キャラクターの肩から上の画像から全身画像への画像翻訳を検討する. 既存の高精度な画像翻訳モデルである Council GAN をベースに, 大きな形状変化を伴う画像翻訳に対応するためのパラメータ変更に着目し, 新規モデルを提案した. キャラクターの全身画像のデータセットを独自に作成して画像翻訳を試み, 既存手法で生成した場合と比較した. 客観的評価値については, 既存手法に比べて提案手法の方が良い値が得られた. しかし, どちらの手法についても, キャラクターの全身画像であると明確に言えるものを生成することはできなかった.

キーワード 深層学習, GAN, 画像スタイル変換, イラスト

1 はじめに

近年, イラスト自動生成手法が劇的に進化している. 中国産の画像生成 AI「Stable Diffusion」[3] などは, 画像のイメージをテキストで入力するとそれにマッチしたイラストが生成される. このような技術の進歩によって, イラストレーターの仕事が奪われるという声がある反面, アイデア捻出の助けになっているという意見もある. よって, クリエイターの手助けになるような応用手法を発展させる必要がある. 本研究では, イラスト初心者にありがちな「顔しか描けない」という状態から, 全身のイラストを描くための手助けになるよう, キャラクターの肩から上の画像から全身画像への画像翻訳を検討する. (図 1) その際, ただ顔に体を付けるだけでなく, 肩から上のイラストの特徴を汲み取った形で全身画像を生成することを目指す.

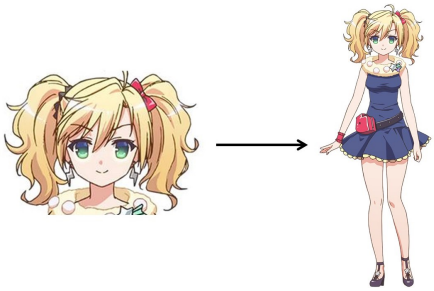

図 1 画像翻訳のイメージ [9]

2 関連研究

2.1 GAN

画像翻訳タスクの研究では、様々な種類の GAN (敵対的生成ネットワーク) が用いられている. これは, 画像を生成する Generator と画像が本物か判別する Discriminator を競い合わせて, 高精度な画像を生成できるように学習する.

GAN の目的関数は式 1 で表される.

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{data}(x)}[logD(x)] + \mathbb{E}_{x\sim p_z(x)}[log(1-D(G(z)))] \quad (1)$$

$\mathbb{E}_{x\sim p_{data}(x)}$ は Discriminator に本物の画像が与えられた時の期待値, $\mathbb{E}_{x\sim p_z(x)}$ は Discriminator に生成された画像が与えられた時の期待値, $G(z)$ は Generator がノイズ z から生成した画像であり, $D(x)$ は Discriminator に画像 x が与えられた時のスコアを $0 \leqq D(x) \leqq 1$ で表す. Discriminator が理想的な動作をすれば右辺は最大になり, Generator が理想的な動作をすれば右辺は最小になる.

2.2 Council GAN

画像翻訳に用いられる GAN においては, Council GAN [16] が大きな成果を上げている. この GAN は, 1 つの Generator と 2 つの Discriminator を 1 つのグループにした Council というユニットを複数用いる. この Council では, Generator が入力画像を受け取って画像を生成し Discriminator が入力画像と生成された画像の識別を行うが, もう 1 つの Discriminator では生成された画像が自分の Council の Generator で生成されたのか, 別の Council の Genarator で生成された画像なのかを識別するように学習する. つまり, それぞれの Council の生成画像は他の Council が同意するような特徴を持つ必要があるので, 生成される画像はソース画像の重要な特徴を維持することが可能となる. この手法では, アニメキャラクターの顔画像から実際の人物写真への変換などを高精度に行うことに成功している. しかし, 形状が大きく異なる画像間での翻訳については既存のモデルをそのまま適応できないケースが多い. Council GAN の目的関数は式 2 で表さ

れる.

$$\min_{G_i}\max_{D_i}\max_{\hat{D}_i} V(G_i, D_i, \hat{D}_i)$$
$$= GAN\_loss_i + \lambda_1 Council\_loss_i + \lambda_2 Focus\_loss_i \quad (2)$$

i は Council の番号, GAN_loss_i は LSGAN [?] で用いられている損失関数であり, 式 1 の対数の代わりに平均二乗誤差を用いている. Counsil が N 個存在する時, $Council\_loss_i$ の定義は式 3 のようになる.

$$Council\_loss_i(G_i, \hat{D}_i, \{G_j\}_{j\neq i}, X_s, z_i, \{E_i\}_{1\leqq j\leqq N})$$
$$= \mathbb{E}_{x\sim p(X_s)} \sum_{j\neq i} [log(1 - \hat{D}_i(G_i(E_i(x), z_i), x))$$
$$+ log(\hat{D}_i(G_j(E_j(x), \alpha z_j), x))] \quad (3)$$

E_i はエンコーダされた i 番目のソース画像, z_i は $1 \leqq i \leqq N$ の council に関連付けられたランダムなベクトル, α は他の Council の Generator のサブドメインの大きさを表すスカラー, $G_i(E_i(x), z_i)$ はソース画像 x とランダムベクトル z_i によって生成された画像である.

$Focus\_loss_i$ は, 学習した対象と背景を分離し, 背景をそのままに学習対象を変換する作用を与える. 定義は式 4 のようになる.

$$Focus\_loss_i = \delta(\sum_k mask_i[k])^2 + \sum_k \frac{1}{|mask_i[k] - 0.5| + \epsilon} \quad (4)$$

$mask_i[k]$ は画素 k の第 4 チャンネルの値であり, ϵ は学習対象と背景を分離する作用の強さを示すパラメータである.

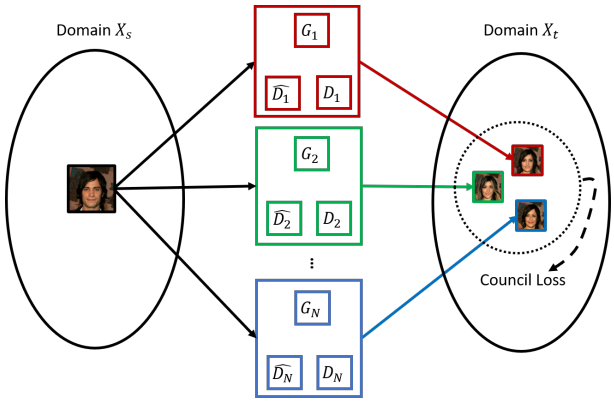

図 2　Council GAN の概要 [16]

2.3　キャラクター画像の生成

キャラクターの全身画像の生成については様々な手法が提案されているが、入出力で大きな形状変化を伴う画像翻訳は既存モデルをそのまま当てはめるのが依然困難である. NovelAI [4] など, 既存の image-to-image サービスでも現状対応できていない. また、現在一般利用可能なアニメキャラクターの全身画像のデータセットは存在しないため, 学習データの収集も容易ではない.

3　提 案 手 法

本研究では, GAN を用いてキャラクターの顔画像から全身画像への翻訳手法を検討する.

3.1　モ　デ　ル

Council GAN を用いた anime-to-face の手法 [16] をベースラインに, パラメータを以下のように調整した.

（1） 画像の入出力サイズを 512 × 512 に変更した.

（2） メモリ使用量削減のため, バッチサイズを 1 に変更した.

（3） Instance Normalization [17] による正規化を導入した. これにより, 画像とチャネルごとに正規化を行うため, バッチサイズが小さくても問題なく正規化を行うことが出来る.

（4） council の数を 4 から 3 に変更した. council が複数あることでソース画像側の特徴を残す作用を生むので, council の数を減らすことでターゲット画像への大きな形状変化に対応することを期待できる.

提案モデルは図 3 のようになる.

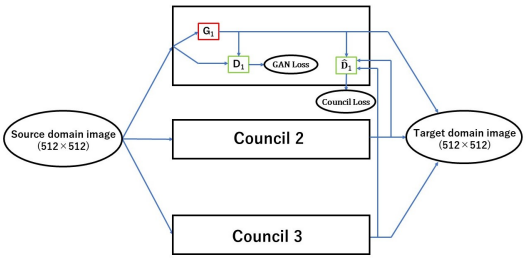

図 3　提案モデルの概要

3.2　データセット

一般利用可能な 2 次元のアニメキャラクターの全身画像のデータセットは現在存在していないので, 独自に作成した. まず, 2013 2023 年放送のアニメの公式ホームページのキャラクターページや, イラストコミュニケーションサービス「pixiv [6]」から以下の基準を満たす, 図 4 のようなキャラクターの全身画像を手動で収集した.

（1） キャラクターの頭から足先までの全身が描かれていること.

（2） キャラクターが正面或いは斜め方向を向いていること.

（3） 身体と衣装以外の情報が多く描かれていないこと.

次に, 画像のキャラクター部分のみを切り抜き背景を黒色に統一した. 最後に, 画像の縦横比を保ったまま余白部分を黒に塗りつぶし, モデルの入力サイズに合わせて画像をリサイズした. このようにして, 合計 2050 枚のアニメキャラクターの全身画像によるデータセットを作成した.

4　実　　験

4.1　概　　要

既存手法で 400,000iteration, 3 章での提案手法で 313,000iteration で学習を行い, 画像翻訳の結果を比較することで生成の精

図 4　収集した全身画像の具体例 [7] [8] [9] [10] [11] [12] [13] [14]

度を検証した. 既存手法として, 画像の入出力のサイズを 256 × 256 にし他のパラメータを anime-to-face の手法と同様にした Council GAN を用いた. 評価指標には FID(フレシェ開始距離) [15] を用いた. この値は生成された画像とターゲット画像の類似度を評価するものであり, 小さい程類似度が高い.

4.2　データセット

ソース画像として,Kaggle の selfie2anime [5] から 2050 枚のアニメの顔画像を使用し, ターゲット画像として独自作成したデータセットを使用した. それぞれのデータセットのうち 2000 枚を学習データ, 50 枚をテストデータとした. また, 背景処理の効果を確認するため, 既存手法に対しては背景を黒に統一していない状態の全身画像を使用した.

4.3　結　　果

それぞれの手法で生成された画像は図 5, 6 のようになった. また, それぞれの手法で算出された学習終了時の FID は表 1 のようになり, iteration ごとの FID の変化は図 7, 8 のようになった. それぞれのグラフに係数 0.9 で指数平滑移動平均による平滑化を行った.

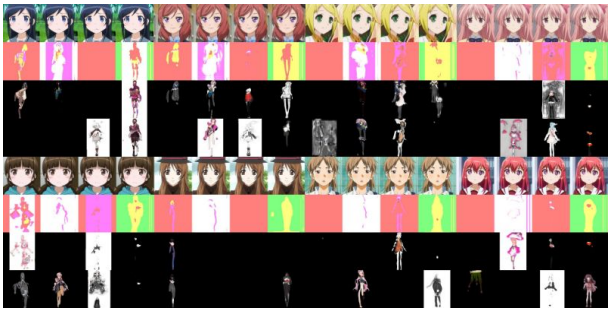

図 5　既存手法での生成結果

図 6　提案手法での生成結果

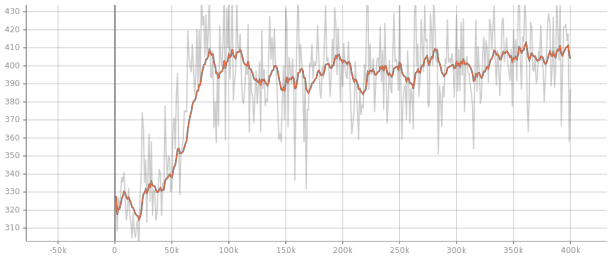

図 7　既存手法での FID の変化

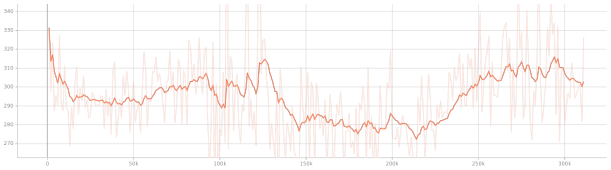

図 8　提案手法での FID の変化

表 1　各手法の FID スコア

	既存手法	提案手法
FID	387.21	281.59

5　考　　察

5.1　客観的評価値の比較

表 1 より, 実験から得られた FID は提案手法の方が大幅に小さくなった. これは, 提案モデルによる精度改善と考えられるが, データセットの前処理で背景色を統一したことでよって類似性を得やすくなったことも要因であると思われる.

5.2　生成画像の比較

図 5, 6 より, 提案手法の方が人体の輪郭線に近いものが生成されており, 背景色も黒に統一されている. これは, データセット作成の段階で背景を同一色に統一し, 画像変換の対象を見分けやすくした成果だと考えられる.

しかし, 既存手法と提案手法の両方において, キャラクターの

全身画像であると明確に言えるものを生成することはできなかった. これは, モデルとデータセットの両方に問題があったためだと考えられる.

モデルについては Council GAN をベースライン手法としたが, このモデルは元々非常にサイズが大きいため, 画像の入出力のサイズを大きくした時点で, 本研究で使用可能であった GPU のメモリの殆どを使用してしまうサイズになってしまった. これにより, 既存のモデルにあまり大きな変更・追加を行うことが出来なかった. また, 今回はベースラインとするモデルを選ぶ際に, アニメ画像への変換を高精度に行うことに成功していることを重視していたが, 本研究は大きな形状変化を伴う画像翻訳であるということを念頭に置いて, それに成功しているモデルを取り入れる必要があったと思われる.

データセットについては, 今回は手動で 2050 枚の画像を収集したが, GAN で高精度な画像翻訳を行ううえでは十分であったとは言い難い. 画像の収集及びノイズ画像の除去を自動で行うことが出来たならば, より多くの画像で安定して学習を行うことが出来たと考えられる. また, ターゲット画像については独自でデータセットの作成および前処理を行ったが, ソース画像については既存のデータセットを用いた. 顔画像と全身画像の対応を学習させるために, 全身画像の肩から上の部分をトリミングしたものを顔画像をとして, ペア化したデータセットを作成して使用することができればより生成精度を上げることが出来たと思われる.

6 おわりに

本研究では, キャラクターの肩から上の画像から全身画像への画像翻訳を試みた. 提案モデルと新規データセットを用いた画像翻訳は, 客観的評価値においては既存手法よりも良い精度を得ることが出来た. しかし, 本研究の目的であった「キャラクターの肩から上の画像から全身画像への画像翻訳」「肩から上のイラストの特徴を汲み取った全身画像を生成」は達成することが出来なかった.

今後の展望として, ペア化したデータセットの作成と, ペア化したデータセットを用いる画像翻訳のモデルをベースラインとした手法が提案できれば, 生成精度の向上を期待できる. また, 今回は正面を向いた女性キャラクターの画像のみを対象としたが, 将来的にはキャラクターの性別や顔の向きを問わず, 顔画像から適切な全身画像の生成ができるようにしたい.

謝　　辞

本研究は, 電気通信大学人工知能先端研究センター (AIX) の計算機を利用して実施したものです.

本研究は JSPS 科研費 JP21H03496, JP22K12157 の助成を受けたものです.

文　　献

[1] http://ongaku-shoujo.jp/
[2] pixiv. https://www.pixiv.net/.
[3] Stable diffusion. https://stablediffusionweb.com/.
[4] NovelAI - The GPT-powered AI Storyteller, https://novelai.net/
[5] selfie2anime, https://www.kaggle.com/datasets/arnaud58/selfie2anime
[6] pixiv, https://www.pixiv.net/
[7] https://nonnontv.com/tvanime/
[8] https://www.toei-anim.co.jp/tv/precure/
[9] http://ongaku-shoujo.jp/
[10] http://citrus-anime.com/
[11] https://www.pixiv.net/artworks/70043091
[12] https://www.pixiv.net/artworks/104391463
[13] https://www.pixiv.net/artworks/84264082
[14] https://www.pixiv.net/artworks/79847750
[15] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In Advances in Neural Information Processing Systems, pages 6626–6637, 2017
[16] Ori Nizan, Ayellet Tal. Breaking the Cycle - Colleagues Are All You Need. CVPR(2020)
[17] Dmitry Ulyanov, Andrea Vedaldi, Victor Lempitsky. Instance Normalization: The Missing Ingredient for Fast Stylization. CVPR. 2016.