情報処理学会論文誌 Vol.60 No.3 870–879 (Mar. 2019)

服の領域を考慮した写真上の人物の自動着せ替えに関する研究

久保 静真[1,a)] 岩澤 有祐[1,b)] 鈴木 雅大[1,c)] 松尾 豊[1,d)]

受付日 2018年6月7日, 採録日 2018年12月4日

概要：本稿では Generative adversarial networks（GAN）に基づく写真上の自動着せ替えについて研究する．GAN に基づく自動着せ替えの手法としては Conditional Analogy GAN（CAGAN）がすでに提案されているが，複雑なパターンの服の生成は難しい．本研究では，衣類の領域を考慮することで CAGAN よりも服のパターンをよりよく反映させることができる SwapGAN を提案する．この SwapGAN は，まず大規模なデータセットで訓練されたセグメンテーションのモデルを使用して，写真上の人物の衣服の領域を取得する．その後，取得した領域を用いて衣服部分を人物画像から除去し，空白領域に所望の衣服を描写する．実験によって，提案手法ではセグメンテーションが有効に機能し，既存手法に対して優位性があることを確認した．

キーワード：深層生成モデル，GAN，ファッション

SwapGAN: Cloth-Region Aware Generative Adversarial Networks toward Virtual Try-On System

Shizuma Kubo[1,a)] Yusuke Iwasawa[1,b)] Masahiro Suzuki[1,c)] Yutaka Matsuo[1,d)]

Received: June 7, 2018, Accepted: December 4, 2018

Abstract: We investigate a virtual try-on method based on generative adversarial networks (GAN). Conditional Analogy GAN (CAGAN) has already been proposed as a virtual try-on method based on GAN, though this method is not good at generating with complex patterns of clothing. In this study, we propose SwapGAN which can better reflect clothing pattern than CAGAN by considering clothing area. Our method first obtains the clothing region on a person by using a human parsing model trained with a large-scale dataset. Next, using the acquired region, the clothing part is removed from a human image. A desired clothing image is added to the blank area. Our experimental results showed that our proposed method has superiority over the existing method by a human parsing functioning effectively.

Keywords: deep generative model, GAN, fashion

1. はじめに

近年，利便性の高さや取り扱い品目の多さから E-commerce（EC）サイトの需要が増加している．そのなかでもファッション分野は需要の高い分野の 1 つである．ファッション EC サイト市場は国内市場での伸び率が高く[*1]，また，アメリカにおいても市場の拡大が予測されている[*2]ことから，国内外でその市場の重要性は増していることが伺える．

[1] 東京大学大学院工学系研究科
Graduate School of Engineering, The University of Tokyo, Bunkyo, Tokyo 113–8654, Japan
[a)] kubo@weblab.t.u-tokyo.ac.jp
[b)] iwasawa@weblab.t.u-tokyo.ac.jp
[c)] masa@weblab.t.u-tokyo.ac.jp
[d)] matsuo@weblab.t.u-tokyo.ac.jp

[*1] アパレル EC 市場の動向と事業戦略の方向性，最終閲覧日：2018 年 9 月 30 日，http://www.smbc.co.jp/hojin/report/investigationlecture/resources/pdf/3_00_CRSDReport053.pdf

[*2] Apparel, footwear and accessories retail e-commerce revenue in the United States from 2016 to 2022 (in million U.S. dollars)，最終閲覧日：2018 年 9 月 30 日，https://www.statista.com/statistics/278890/us-apparel-and-accessories-retail-e-commerce-revenue/

しかしながら，今後のファッション EC 市場拡大のためには課題もある．その課題とは EC サイト上において自分に似合う服の選択が難しいことである．実店舗の場合，鏡の前で試着して服が自分の身体に合っているか，色合いや模様が似合っているかどうかを確かめることが可能であるが，EC サイト上では難しい．この課題を改善するためには，実店舗の試着では可能だった服のサイズのフィット感の確認と服の色合いや模様が似合うかどうかの確認を EC サイト利用時にもできるようにする必要がある．

服のサイズのフィット感の確認を行う方法としては，スーツを用いたサイズ計測がすでに行われている．START TODAY 社は，顧客が身体のサイズを計測するためにスーツの無償提供を行った[*3]．このスーツを着用した状態で全身を撮影することで身体のサイズを 3D モデルとして再現することができるため，EC サイト利用時にも身体のサイズにあった服の提供が可能となった．

一方，服の色合いや模様が似合うかどうかの確認を行う方法としては，スマートフォンのカメラを用いたアプリケーションが考えられる[*4]．ユーザは自分自身の写った写真を撮影・アップロードし，その写真内のユーザが着用している服をサイト内の特定の服に自動的に着せ替えて，あたかもユーザ自身が着用しているような画像を生成することで着用のイメージを得るというものである．この手法は自動着せ替えとして技術の研究がなされているところである．

自動着せ替えの研究として有力なものに，VITON [1] と呼ばれる手法を提案した研究と CAGAN [2] と呼ばれる手法を提案した研究の 2 つがある．VITON は粗い画像を生成するネットワークとその出力を精錬した画像を出力するネットワークの 2 段階のモデルを構築して自動着せ替えを行う手法である．また, CAGAN は Generative Adversarial Networks（GAN）[3] と呼ばれる手法に，ネットワークの過程でセグメンテーション[*5]を生成するような学習の工夫を加えることで，自動着せ替えを行う手法である．GAN は本物らしい画像を生成できるという点でこのタスクにふさわしいと考え，本研究では GAN を用いた自動着せ替え手法に着目する．

CAGAN は単色の服のような模様の単純な服ではうまくいくことが分かっているが，実際にアプリケーションとして利用する際に問題となることがある．それは複雑な模様の服に対してはうまく適用できないということである．服の模様は，服が自分に似合うかどうかの判断には必要な要素であるため，これを改善することは重要である．

本研究では CAGAN の抱える問題点を改善する SwapGAN を提案する．CAGAN はネットワークの中で服の領域を示すセグメンテーションを生成するが，そのセグメンテーションは複雑な模様に対しては十分に機能していない可能性が考えられた．そのため，本研究では明示的に服の領域のセグメンテーションを行うことでその問題の解決を目指す．本研究で提案する SwapGAN は領域を検出する学習済みのセグメンテーションのネットワークを用いる．評価に関しては，定性的・定量的評価によって提案モデルの着せ替えの精度の優位性を検証する．

本研究の貢献は以下のとおりである．

- 写真上の自動着せ替えのタスクにおいて，学習済みのセグメンテーションのネットワークを活用することで既存のモデルで指摘されていた複雑な模様の服に対する課題を改善したモデルである SwapGAN を提案した．
- 定性的・定量的評価によって提案モデルの着せ替えの精度の優位性を検証した．

本稿の構成は次のとおりである．まず，2 章では本研究に関する前提知識を説明する．次に 3 章では既存手法と提案手法の概要について述べ，4 章では本研究に関連する研究を紹介する．5 章では実際に EC サイトから取得したデータを用いて実験し，既存手法との比較により提案手法の優位性を示す．そして，最後に 6 章でまとめを述べる．

2. 前提知識

本章では，本研究の前提知識について述べる．2.1 節では Generative Adversarial Networks（GAN）[3] について説明し，続く 2.2 節では GAN を応用した conditional GAN [4] について説明する．

2.1 Generative Adversarial Networks（GAN）

Goodfellow らの研究 [3] で Generative Adversarial Networks（GAN）が提案された．GAN は深層生成モデルと呼ばれる深層ニューラルネットワークを用いた生成モデルの中でも主流の手法の 1 つである．ここで，生成モデルとは訓練データを学習することで，その訓練データを生成する確率分布を推定し，訓練データと似たようなデータを生成するモデルのことをいう．

GAN のネットワークは敵対する 2 つのネットワークからなる．1 つが Discriminator と呼ばれる入力サンプルがモデルから生成されたものであるかデータセットから取り出されたものであるかを見分けるネットワークで，もう一方が Generator と呼ばれる Discriminator に生成されたものかどうか見分けがつかないようなデータセット内のデータに近いデータを生成できるように学習するネットワークである．

[*3] zozosuit，最終閲覧日：2018 年 9 月 30 日，http://zozo.jp/zozosuit/

[*4] ZOZO スーツなどのバーチャル試着が EC サイトにもたらすメリット，最終閲覧日：2018 年 9 月 30 日，https://ec-orange.jp/ec-media/?p=19246

[*5] 本稿内のセグメンテーションとは画像のセマンティックセグメンテーションのことを指し，ピクセル単位でラベル付けを行うことである．

Generator，Discriminator の両ネットワークは多層パーセプトロンによりなる．データ x に対する Generator の生成分布 p_g を学習するためにインプットとなるノイズのデータ分布 $p_z(z)$ を定義し，データ空間へのマッピングを $G(z;\theta_g)$ と表現する（以下，簡易的に G のように表す）．G はパラメータ θ_g で定義された多層パーセプトロンによって定義された微分可能な関数である．同様に Discriminator 側の多層パーセプトロンも $D(x;\theta_d)$ のように定義し単一スカラーを出力とする（以下，簡易的に D のように表す）．D は x がデータセット内のデータである確率を表す．学習段階では，D を x はデータセットからのデータであるか p_g から生成されるデータであるかの正しいラベル付けをできるような確率を最大化させるように学習させる．それと同時に G を $\log(1-D(G(z)))$ を最小化するように学習する．つまり，以下の式のように G と D の V(G, D) の関数の min-max ゲームによって学習を進める．

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

なお，$x\sim p_{data}(x)$ はデータセットのなかからミニバッチのためにデータをサンプリングしていることを，$z\sim p_z(z)$ はノイズ z のサンプリングをしていることをそれぞれ表しており，E は期待値を表している．

学習プロセスは，生成の質を評価する関数である損失関数を最小化することで行う．学習プロセス自体は自動化されるが効果的な損失関数を設計する必要がある．この損失関数をうまく設定必要があり，たとえばユークリッド距離を用いるとぼやけた画像を生成しがちであった．これに対して，GAN は生成画像が本物かどうかという学習を行うことで目的を満たす適切な損失関数を自動的に学習することができる．ぼやけた画像は本物には見えないため，生成されにくくなる．

画像の生成において，GAN は比較的くっきりとした画像が出力される傾向にあるが，学習は不安定となる．また，画像の生成については GAN のネットワークとして Convolutional Neural Networks（CNN）を使うことが効果的であることが分かっている [5]．

2.2 conditionalGAN

前項の GAN では画像を生成するときにラベル情報のコントロールまではできなかった．たとえば，手書き文字のデータセット（MNIST）を例にすると，GAN では数字のような画像を生成できても，どの数字の画像を生成するかまでは指定することができなかった．conditinalGAN [4] では，Generator と Discriminator の入力として y という新しい情報を以下の式のように条件付けることにより生成する画像をコントロールすることができるようにした．つまり，MNIST の例であればどの数字を出力するかをコントロールすることを可能にした．

$$\min_G \max_D V(D,G) = E_{x\sim p_{data}(x)}[\log D(x|y)] + E_{z\sim p_z(z)}[\log(1-D(G(z|y)))]. \quad (2)$$

3. 既存手法と提案手法

本章ではまず GAN を用いた自動着せ替えの従来手法である CAGAN について説明し，その課題まで述べる．その後に提案手法である SwapGAN について説明する．

3.1 既存手法とその課題

3.1.1 画像の生成

図 1 の上側が CAGAN の Generator の模式図である．CAGAN では Generator の入力として「人物画像」,「その人物が着用している服の画像」,「着せ替えたい服の画像」の 3 枚が必要となっており，着せ替わった人物の画像が出力される．なお，x_i，y_i をそれぞれ人物画像とその人物が着ている服の画像を表すとすると，学習のデータとしては $\{x_i, y_i\}_{i=1}^N$ のような N 組のペアの集合からなる．

CAGAN のネットワークには，画像の畳み込みによって入力画像の特徴量を圧縮する Encoder と畳込みの逆操

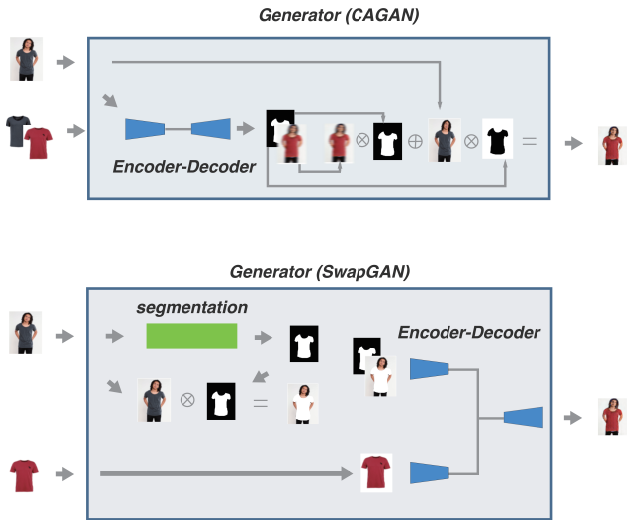

図 1 CAGAN と SwapGAN の Generator のネットワーク構成を表している．上側が CAGAN で下側が SwapGAN の Generator を表している．各 Generator への入力と出力の画像はテスト時の例を示している．SwapGAN はまず，入力の人物画像の服の領域を特定する．そして，その領域を取り除いた人物の画像を用意する．それと服の画像を入力として Encoder-Decoder のネットワークが着せ替わった画像を生成する

Fig. 1 The architecture of CAGAN's generator (on the upper side) and SwapGAN's generator (on the lower side). This figure shows an example at the time of testing. SwapGAN first identifies the area of clothing in the person image and removes the area. Then encoder-decoder network generates a image by using the removed image and a clothing item.

作である逆畳み込みによって画像に変換する Decoder からなる Encoder-Decoder ネットワークが含まれる．特に CAGAN は 3 種類の画像を入力とするトリプレットと呼ばれ形式である．また 4 チャンネルが出力となり，そのうち 1 チャンネルを α_i^j，3 チャンネルを $\tilde{x}_i^j$ とする．α_i^j は服の領域のマスクを表し，$\tilde{x}_i^j$ は着せ替え後の服の画像を表す．Encoder-Decoder の関数を F，着せ替えたい服の画像を y_j とすると，以下の式のような合成によって Generator から画像が出力される．

$$\begin{aligned}&[\tilde{x}_i^j, \alpha_i^j] = F(x_i, y_i, y_j).\\&G(x_i, y_i, y_j) = \alpha_i^j \tilde{x}_i^j + (1-\alpha_i^j)x_i.\end{aligned} \tag{3}$$

3.1.2 損失関数

既存手法である CAGAN が学習するための損失関数は以下の式のように $L_{cGAN}(G,D)$，$L_{id}(G)$，L_{cyc} の 3 つの項の和からなる．

$$\begin{aligned}\min_G \max_D V(D,G) = L_{cGAN}(G,D)\\ +\gamma_i L_{id}(G) + \gamma_c L_{cyc}(G).\end{aligned} \tag{4}$$

CAGAN の研究 [2] では $\gamma_i = 0.1, \gamma_c = 1$ として学習を行っている．

まず，以下の式に表す $L_{cGAN}(G,D)$ は GAN の損失関数である．この項が Generator と Discriminator に関わる最も重要な項である．

$$\begin{aligned}L_{cGAN}(D,G) = &E_{x_i,y_i\sim p_{data}}[\log D(x_i,y_i)]\\ &+E_{x_i,y_j\sim p_{data}}[\log(1-D(G(x_i,y_i,y_j),y_j))]\\ &+E_{x_i,y_j\sim p_{data}}\log(1-D(x_i,y_j))].\end{aligned} \tag{5}$$

以下の式に表す $L_{id}(G)$ は α_i^j に対して設ける制約である．この項によって，服に関係ない領域はできるだけ変換後も残すようになる．結果的に，α_i^j は服のセグメンテーションを意味することになる．

$$L_{id}(G) = E_{x_i,y_i,y_j\sim p_{data}}||\alpha_i^j||. \tag{6}$$

最後に，以下の式は服の着せ替えは着せ替えた後にもう一度はじめの服を着せ替えると元に戻ることが期待されるための関数である．元々の画像との L1 ロスをとる．

$$L_{cyc}(G) = E_{x_i,y_i,y_j\sim p_{data}}||x_i - G(G(x_i,y_i,y_j),y_j,y_i)||. \tag{7}$$

3.1.3 課題

CAGAN のセグメンテーションの出力は入力の服が複雑な模様の場合，その模様の影響を受けて崩れてしまう．模様がセグメンテーションに影響を与えてしまうのは，Generator が服の領域のセグメンテーションと服の画像の両方を同時に出力しているためである．次節で説明する提案手法の SwapGAN では，セグメンテーションの出力を服の出力とは別のネットワークで行うことにより，複雑な模様に影響を受けずにセグメンテーションを出力できる．これによって，複雑な模様の服の着せ替えもうまく行うことを期待している．

3.2 SwapGAN

3.2.1 画像の生成

本稿の提案手法である SwapGAN は，従来手法である CAGAN の持つ複雑な模様に対してはうまくいかない課題を改善するために，Generator のネットワークに服の領域を考慮してセグメンテーションのネットワークを明示的に組み込んだ．本提案手法の Generator のネットワークの模式図を図 1 の下側に示す．まず，入力した人物画像の服の領域を特定し，セグメンテーションを生成する．そして，領域のセグメンテーションと服の領域を取り除いた人物画像の 2 枚と入力の服の画像の計 3 枚を入力とした Encoder-Decoder のネットワークによって，自動着せ替えが行われる．なお，この Encoder-Decoder ネットワークは Pix2Pix [6] のネットワークを参考にしている．

Encoder-Decoder ネットワークを関数 F，セグメンテーションのネットワークを関数 M で表すと，SwapGAN の Generator は以下の式のように表される．

$$G_{swap}(x_i, y_i) = F(x_i \odot M(x_i), M(x_i), y_j). \tag{8}$$

3.2.2 損失関数

提案手法の損失関数は以下の式のように定義される．各項についても以下で説明する．

$$\begin{aligned}\min_{G_{swap}} \max_D V(D, G_{swap}) = L_{cGAN}(G_{swap}, D)\\ +L_{cyc}(G_{swap}) + L_{per}(G_{swap}).\end{aligned} \tag{9}$$

以下では $L_{cGAN}(G_{swap}, D)$，$L_{cyc}(G_{swap})$，$L_{per}(G_{swap})$ をそれぞれ L_{cGAN}，L_{cyc}，L_{per} のように省略して表記する．

まず，以下の式は L_{cGAN} を表す．

$$\begin{aligned}L_{cGAN} = &E_{x_i,y_i\sim p_{data}}[\log D(x_i,y_i)]\\ &+E_{x_i,y_j\sim p_{data}}[\log(1-D(G_{swap}(x_i,y_j),y_j))]\\ &+E_{x_i,y_j\sim p_{data}}[\log(1-D(x_i,y_j))].\end{aligned} \tag{10}$$

次に，以下の式が L_{cyc} を表す．

$$L_{cyc} = E_{x_i,y_i,y_j\sim p_{data}}||x_i - G_{swap}(G_{swap}(x_i,y_j),y_i)||. \tag{11}$$

また，Generator に対応する人物画像と服の画像のペア (x_i, y_i) を入力して得られる出力と元の人物画像 (x_i) は同じになることが期待される．それぞれを一般物体認識で高い性能を示した VGG19 [7] の学習済みモデルに入力して得られる各ブロックの特徴マップの差 l_ϕ の和を取ったもの

が Perceptual Loss である．以下の式にて Perceptual Loss と呼ばれる L_{per} を示す．

$$L_{per} = E_{x_i,y_i\sim p_{data}}\left[\sum_{i=1} \lambda_i l_{\phi,blocki\_conv2} \right]. \quad (12)$$

なお，λ は各層のパラメータ数の逆数である．また，$l_{\phi,blocki\_conv2}$ は各ブロックの 2 番目の Convolution から出力される特徴マップを表す．Perceptual Loss では，VGG19 のような一般物体認識のタスクで学習済みのネットワークの隠れ層を使用する．この隠れ層は模様のような高次の特徴をとらえることができる．Perceptual Loss を損失関数に加えることで生成画像と対象画像の高次の特徴を近づけることができ，生成画像に対象画像と同様の特徴を持たせることを可能にする．Johnson らの研究 [8] では，生成する画像と特定のスタイルを持った画像の Perceptual Loss を最小化するように学習を行うことで，同様のスタイルを持った画像を生成できるようにした．提案手法でも，画像間の Perceptual Loss を損失関数に加えることで，生成する画像の模様に着せ替えたい服の模様が反映させるようにした．

4. 関連研究

本章では，本研究の関連研究について述べる．まず，4.1 節では GAN を用いた画像のスタイル変換の説明する．続く 4.2 節では，ファッション分野における GAN による画像生成について述べる．

4.1 GAN による画像のスタイル変換

画像の変換において，対応するドメインに変換する研究が続けられてきた．conditionalGAN の仕組みを用いて一般化された画像の変換の仕組みを提案したのが Pix2Pix と呼ばれるモデルである [6]．

Pix2Pix の学習では Generator への入力として x にあたる画像を入力し，それに対応する画像を生成する．学習は conditionalGAN の損失関数を Generator と Discriminator の敵対的学習によって行われる．また，古典的な損失関数である $L1$ を加えると効果的であることも分かっており，Generator の学習は conditionalGAN の損失関数と $L1$ を混合させた損失関数を最小化することで学習を進める．

Generator のネットワークとしては U-Net と呼ばれる skip connection 付きの Encoder-Decoder ネットワーク[*6]を採用している．また，PAN [9] という研究によれば Pix2Pix の変換において，Perceptual Loss と呼ばれる画像分類のネットワークとして訓練済みのネットワークから視覚的な高次元の情報を抽出する要素を追加することが有効であることが示されており，この要素は本研究でも採用している．

Pix2Pix は変換前と変換後のペアからなるデータセットを期待している．しかしながら，たとえば馬とシマウマのペアのように，変換前と変換後で対になる画像のペアからなるデータセットを得ることができない場面も存在する．CycleGAN [10] はペアのデータセットがない場面でも異なるドメインのそれぞれ画像からなるデータセットがあれば学習が可能となる．CycleGAN の学習は両ドメイン方向の変換に対応する conditionalGAN の損失関数に，画像を一度変換してそれを逆変換したときの画像と元の画像との間に生じる差を損失として定義する Cycle Loss と呼ばれる損失関数を足し合わせた損失関数を最小化することで学習を進めることができる．この Cycle Loss は本研究でも採用している．

4.2 ファッション分野への GAN の利用

最近の動向として，ファッションの分野においても Generative Adversarial Networks（GAN）を利用した画像生成の技術を用いた研究がいくつか行われいてる．そのなかで，服を着用した人物画像の生成の研究も行われている．

ClothNet [11] は服を着た人物の全身画像を生成するモデルの研究である．それまで人物の全身画像は姿勢や服装のバリエーションの多さから難しいとされ，人物画像の生成は 3 次元情報を元にしたものであった．この研究では 2 つのモデルを組み合わせることで，画像のみを学習に利用して服を着用した人物画像の生成を可能にした．1 つのモデルは Variational Autoencoder（VAE）[12] と呼ばれる手法で人物の領域を示した画像を生成するモデルで，もう 1 つのモデルは生成された領域の画像を元にして本物らしい画像を生成する Pix2Pix のモデルである．この 2 つの組み合わせによって人物の全身画像を生成することはできる．しかしながら，服装を本物らしく生成するだけで意図した服を生成することはできない．

FashionGAN [13] の研究は人物の画像の服装を変換する研究である．この研究では人物や姿勢はそのままに与えたテキストに記述された内容の服装を再現する．提案されているモデルでは，まず人物画像のセグメンテーションのデータを利用し，与えた画像の人物や服装の領域を示すセグメンテーションを GAN のネットワークを用いて生成する．その後，生成したセグメンテーションと与えたテキスト情報を別の GAN のネットワークの入力とすることで，テキストの内容に応じた人物画像を生成することができる．この研究では，テキストの情報に応じて服を変えることはできるが着せ替えたい服の画像データを条件とすることはできない．

CAGAN [2] は人物の画像を着せ替えたい服の画像を条件に着せ替わった画像を生成する研究である．ネットワー

[*6] Encoder-Decoder ネットワークとは，入力を中間表現に変換する Encoder とその中間表現を別の形式へ変換する Decoder の 2 つのモデルを組み合わせたネットワークである．skip connection 付きの場合は中間層を軸に対称な畳み込み層と逆畳み込み層を直接つなぐリンクを加える．

ク内部にセグメンテーションを生成する機構を含みながら，Pix2Pix の U-Net と同じような構造の GAN のネットワークを利用して画像を生成することができる．ただし，この研究では複雑な模様の服の着せ替えがうまくいっておらず，本研究ではこれの改善を行う．なお，CAGAN の詳細については本研究の比較手法として 3 章で述べたとおりである．

5. 実験

5.1 データセット

学習に使用する人物画像とその人物の着用する服の画像のペアのデータセットは [1], [2] にならって，オンラインショッピングサイト Zaland [*7] の Website から取得した．**図 2** は人物画像とその人物の着用する服の画像のペアのデータの 1 つの組である．画像は 128 × 96 のサイズを用いた．

人物画像は正面の画像で，服の画像は服 1 着が写ったものを使用した．使用した画像は，男性画像と対応する服の画像の 5,447 組，女性画像と対応する服の画像の 3,839 組の計 9,286 組である．このうち，9,000 組を学習用のデータとして利用した．そして，残りの 286 組をテストデータとして利用した．評価にはこのテストデータを用いている．

5.2 ネットワーク構造

実装を行ったネットワークの詳細を説明する．**図 3** が提案手法の Generator の Encoder-Decoder ネットワークの模式図である．Encoder 側の一方はセグメンテーションの画像とその人物の服の部分を取り除いた画像の 4 チャンネルの入力で，もう一方は着せ替えたい服の画像の 3 チャンネルの入力となっている．ネットワークは Encoder 側は 3×3 のフィルタでストライド 2 の Convolution 層，BatchNorm 層，LeakyRelu 層のレイヤが続き，Decoder 側は 3 × 3 のフィルタでストライド 2 の Transposed convolution 層[*8]，BatchNorm 層，Relu 層のレイヤが続く．そして最後に 3 × 3 のフィルタでストライド 1 の Convolution 層が続く．また，**図 4** は提案手法の Discriminator のネットワーク構造の模式図である．なお，この Decoder のネットワーク構造は CAGAN で用いられる Decoder のネットワーク構造と同様である．

また，本提案手法では Gong らの研究 [14] で高い精度が示されたセグメンテーションのモデルを利用する．このモデルは Attention [15] モデルに Self-supervised Structure-sensitive Learning（SSL）という身体の構造を考慮した仕組みを取り入れた Attention+SSL と呼ばれるモデルである．このセグメンテーションのネットワークのパラメータは提案手法の学習では更新せずに，Cong らの研究 [14] 内で提案される LIP データセットで学習済みのパラメータを利用する．

既存研究に習って，最適化手法には Adam を利用し，パラメータを $\beta_1 = 0.5$，$\beta_2 = 0.999$，学習率を 0.0002 とした．また，実装は Tensorflow バックエンドの Keras で行い，バッチサイズを 16 で 13.5 万ステップの学習を行った．提案手法の学習は 12 GB メモリの TITAN X（Pascal）GPU 上で約 16 時間程度時間を要した．

5.3 生成画像の定性評価

5.3.1 生成画像の比較

図 5 では既存手法の CAGAN と提案手法の SwapGAN の生成画像の比較を行っている．(a) は着せ替える対象の人物画像で，(b) は (a) の人物に着せ替える服の画像である．CAGAN による生成結果が (c) で SwapGAN による生成結果が (d) である．

(A)，(B) の行は (a) の画像中の人物の服が単色の画像である場合の着せ替えである．両手法ともうまくいっているようにみえるが，既存手法の (c) では元の服の色が若干残っているのに対して，提案手法の (d) では比較的元の服

図 2 データセット例
Fig. 2 Dataset examples.

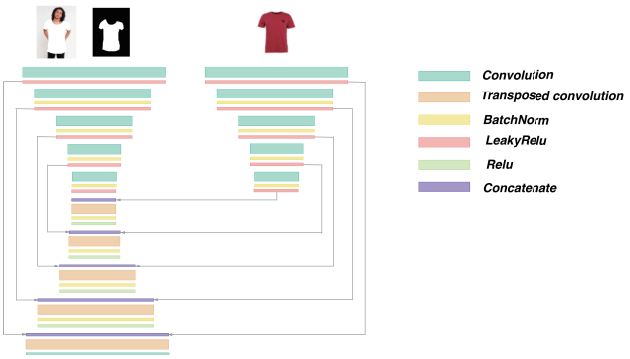

図 3 SwapGAN の Encoder-Decoder の模式図
Fig. 3 The encoder-decoder architecture of SwapGAN.

図 4 SwapGAN の Disciriminator の模式図
Fig. 4 The discriminator's architecture of SwapGAN.

*7 http://www.zalando.com/

*8 Transposed convolution 層とは逆畳み込み層のことを指す．

図 5 CAGAN および SwapGAN の画像の生成結果の例を示している．(a) と (b) はデータセット中の画像である．(a) の人物の服を (b) の服に着せ替える．(c) が CAGAN の生成結果を表しており，(d) が SwapGAN による生成の結果を表している

Fig. 5 (a) and (b) are images in the dataset. The human model (a) wears the item (b). (c) shows the result of CAGAN, and (d) shows the result of SwapGAN.

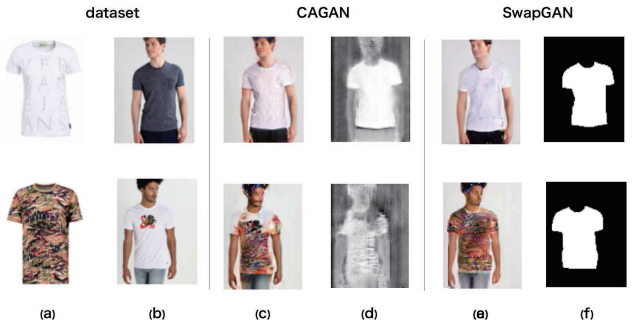

図 6 CAGAN と SwapGAN のセグメンテーション．(a)・(b) はデータセット中の画像，(c)・(d) がそれぞれ CAGAN による生成画像とセグメンテーション（論文中の α），(e)・(f) が SwapGAN による生成画像とセグメンテーションを示す

Fig. 6 The results of CAGAN and SwapGAN. (a) and (b) are images in the dataset. (c), (d) shows the generated image and segmentation by CAGAN respectively. (e), (f) shows the generated image and segmentation by SwapGAN respectively.

図 7 セグメンテーションの影響．(a) が着せ替える服で (b) が着せ替える人物の画像である．(d) がセグメンテーションの結果であり，(c) が着せ替えの結果の画像である

Fig. 7 The influence of segmentation. The human model (b) wears the item (a). (c) shows the generated image and (d) shows the result of segmentation.

の色の影響が小さい．また，(C)，(D) の行は画像中の人物の画像に模様がある場合の着せ替えである．既存手法の (c) は元の服の色や模様の影響が (A)，(B) のときよりも大きく残っていることがみられる．一方，提案手法の (d) は元の服の色や模様の影響が小さく，着せ替えのタスクが比較的よくできていることが確認できる．なお，生成画像の他の結果は付録として掲載している．

5.3.2 セグメンテーションの有効性の検証

本研究では CAGAN では複雑な模様を持つ服の模様に対して領域を十分に特定できていないことが課題になっていると考えた．それを比較により確認する．**図 6** に CAGAN と提案手法の生成画像とそのセグメンテーションの例を示す．図中上部の行の服は模様が単調で CAGAN による服の位置の特定も比較的うまくいっており SwapGAN との差は小さい．しかし，図中下部の服のように複雑な模様を持つ場合には CAGAN のセグメンテーション (d) がうまくいっておらず，結果的に生成画像 (c) が崩れたものになってしまっている．それに対して提案手法のセグメンテーション (f) は図中の上下ともうまくできており，生成画像が CAGAN と比較して複雑な模様に対しても機能していることが分かる．以上より，SwapGAN では CAGAN と比較して複雑な模様の服に対しても精度の高い服の領域の特定ができており，服の着せ替えに貢献していることが伺える．

5.3.3 SwapGAN の適用範囲

SwapGAN の画像生成のなかには必ずしもうまくいかない場合も存在したので，その例を示す．SwapGAN ではまず人物画像の服の領域を特定したセグメンテーションを生成するが，そのセグメンテーションがうまくいかない場合に着せ替えが失敗する例がみられた．以下の**図 7** の (d) は人物の上着の部分を表すセグメンテーションである．一部分しか上着として特定できなかったため (c) の着せ替えの結果の画像にうまくできていない領域が残ってしまっている．

また，本研究では上着の着せ替えに限定した．そのため他のファッションアイテムに関しては考慮していない．ただし，他のアイテムについても個々に同様のモデルを用いることで対処できると考えられる．そして，本研究では人物写真の正面画像に限定した．そのため，人物が横や後ろを向いている場合や何らかのポーズをとっている場合など人物の身体の状態の差異については考慮していない．今後はファッションアイテムや人物の身体の状態に対しても汎

表 1 CAGAN と SwapGAN の生成画像の定量評価
Table 1 Quantitative evaluation of images generated by CAGAN and SwapGAN.

モデル	IS [*9] [16]	FID [17]	アンケート
CAGAN [2]	2.71	98.30	17.8%
SwapGAN（提案手法）	2.88	126.37	82.2%
テストデータ	2.79	-	-

化させたモデルとなることが望まれる．

5.4 生成画像の定量評価

生成結果の比較として定量評価も行った．生成モデルに生成画像の評価として用いられることの多い Inception Score [16] と FID [17] の指標による評価とアンケートを利用した評価を行った．

5.4.1 Inception Score と FID

生成した画像の評価を行うために，Inception Score と FID による指標で評価を行う．Inception Score [16] は Inception モデルで識別されるラベルが多様で識別しやすい画像であるほど数値が高くなるように設計されている指標である．また，FID [17] は画像の集合間の距離を表す．今回では，テストデータの着せ替えを行う人物画像の集合と生成画像の集合の距離を表している．画像生成の品質を測るときに両指標は使用され，Inception Score は数値が高いほど，FID は数値が低いほどよいと判断される．Inception Score と FID は着せ替えがうまくできているかどうかを考慮しているわけではないが，生成した画像が本物らしい画像になっているかを確認する指標に用いることはできる．実験結果は**表 1** に示すとおりである．提案手法の Inception Score は CAGAN やテストデータと比較して十分な値になっている．また，FID の値も CAGAN の数値より大きくなってはいるものの極端に大きな値にはなっておらず，セグメンテーションを加えることで生成画像の品質が著しく落ちてはいないことを示している．

5.4.2 アンケート評価

Inception Score や FID で生成画像の品質を評価はできても，適切に着せ替えを行えているかという本研究での目的を十分に評価することはできない．本研究では，着せ替えの対象の人物の服以外の部分は維持して着せ替えたい服の色や模様などをうまく反映できているかを評価する必要がある．そのため，アンケートによる評価も行った．アンケートの目的は着せ替えとして既存手法と提案手法のどちらが適切に着せ替えができているかを確認することである．そのためにアンケートでは両手法の生成結果を提示してどちらのほうが着せ替えとして適切かどうかを尋ねた．アンケートの具体的なステップとしては以下のとおりである．

- 人物画像と着せ替える服の画像をみせる．
- 同時に，CAGAN，SwapGAN の両手法による生成した着せ替え画像を提示する．
- どちらの画像が着せ替えとして適切かを選択する．

このステップを 1 回として異なるサンプルで 1 人あたり合計 30 回のステップを行い，被験者 131 人から回答を集めた．実験結果は表 1 に示すとおりである．表の数値は適切であると回答した割合を示す．結果から着せ替えの生成画像は CAGAN よりも SwapGAN ほうがよい結果となることが示される．

5.5 考察

写真上の人物の服を替えるという処理は，以下の 2 点から成り立つと考えられる．

(1) 写真上の人物の元の服の位置をとらえること
(2) その形に変形した着せ替える対象の服を当てはめること

CAGAN では (1) を満たすため，人物画像と着用している服の画像から人物画像のどこが衣服であるかの特定を行っていると考えられる．そのために，入力画像として着用している服の画像が必要になっている．しかし，CAGAN では人物が複雑な模様の服を着ている場合，この (1) の服の位置をとらえるところがうまく機能していなかった．本研究の提案手法では，服の領域をとらえるセグメンテーションの役割を別のデータセットで学習させたネットワークを組み込むことで，5.3.2 項に示されるように精度の高い服の領域の特定ができ，モデルの不十分さを改善したと考えられる．セグメンテーションのネットワークの精度が高いのは人物画像と対応する服の領域の画像データセットですでに学習を行っているためであり，このようなモデルを適切に組み込めたことが SwapGAN の有効性の要因であると考えられる．

また，入力画像として人物の着用している服の画像はなくても服の位置の特定をセグメンテーションのモデルで行えるため，提案手法では人物の着ている服の画像の入力は不要となった．これにより，CAGAN では「人物画像」，「その人物が着用している服の画像」，「着せ替えたい服の画像」の計 3 枚が入力として必要だったが，SwapGAN では「人物画像」，「着せ替えたい服の画像」の 2 枚のみの入力で着せ替え画像の生成が可能となった．EC サイト上でユーザが着せ替えを実行するときに自分の服を脱いでその服の画像を撮影することは手間であるため，「その人物が着用している服の画像」が不要となることはアプリケーションとして利用するうえで利点となる．

6. おわりに

本稿では，ファッション EC 分野の発展を背景に需要の高まる写真上の自動着せ替えに関して，本物らしい画像を生成することができる GAN を用いた自動着せ替えの精度

[*9] IS は Inception Score のこと．

向上を目指した．既存手法である CAGAN の服の領域を特定するセグメンテーションに注目して，服の位置の特定を行うために学習済みのセグメンテーションのモデルを組み込んだ手法 SwapGAN を提案した．定性評価による生成画像の検証と定量評価による CAGAN との比較によって，提案手法の有効性を示した．

今後も引き続き，服の模様について改善に取り組む予定である．また，様々な姿勢やアイテムに適用範囲を広げることも予定している．

参考文献

[1] Han, X., Wu, Z., Wu, Z., Yu, R. and Davis, L.S.: VITON: An Image-based Virtual Try-on Network (2017).

[2] Jetchev, N. and Bergmann, U.: The Conditional Analogy GAN: Swapping Fashion Articles on People Images, *International Conference on Computer Vision* (*ICCV*) (2017).

[3] Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y.: Generative Adversarial Networks, *Neural Information Processing Systems* (*NIPS*) (2014).

[4] Mirza, M. and Osindero, S.: Conditional Generative Adversarial Nets, pp.1–7 (2014).

[5] Radford, A., Metz, L. and Chintala, S.: Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks, pp.1–16 (2015).

[6] Isola, P., Efros, A.A., Ai, B. and Berkeley, U.C.: Image-to-Image Translation with Conditional Adversarial Networks, *Computer Vision and Pattern Recognition* (*CVPR*) (2017).

[7] Simonyan, K. and Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition, *International Conference for Learning Representations* (*ICLR*) (2015).

[8] Johnson, J., Alahi, A. and Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution, *European Conference on Computer Vision* (*ECCV*) (2016).

[9] Wang, C., Xu, C., Wang, C. and Tao, D.: Perceptual Adversarial Networks for Image-to-Image Transformation, pp.1–20 (2017).

[10] Zhu, J.-Y., Park, T., Isola, P. and Efros, A.A.: Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks, *International Conference on Computer Vision* (*ICCV*) (2017).

[11] Lassner, C., Pons-Moll, G. and Gehler, P.V.: A Generative Model of People in Clothing (2017).

[12] Kingma, D.P. and Welling, M.: Auto-Encoding Variational Bayes, No.Ml, pp.1–14 (2013).

[13] Zhu, S., Fidler, S., Urtasun, R., Lin, D. and Loy, C.C.: Be Your Own Prada: Fashion Synthesis with Structural Coherence, No.Figure 1 (2017).

[14] Gong, K., Liang, X., Zhang, D., Shen, X. and Lin, L.: Look into Person: Self-supervised Structure-sensitive Learning and A New Benchmark for Human Parsing, *Computer Vision and Pattern Recognition* (*CVPR*) (2017).

[15] Chen, L.-C., Yang, Y., Wang, J., Xu, W. and Yuille, A.L.: Attention to Scale: Scale-aware Semantic Image Segmentation (2015).

[16] Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A. and Chen, X.: Improved Techniques for Training GANs, *Neural Information Processing Systems* (*NIPS*) (2016).

[17] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B. and Hochreiter, S.: GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium, *Neural Information Processing Systems* (*NIPS*) (2017).

付　　録

A.1 生成例

付録として生成例をいくつか示す．**図 A·1**，**図 A·2**，**図 A·3**，**図 A·4**，**図 A·5** に生成例を示す．1 行目が SwapGAN の結果を表しており，2 行目が既存手法である CAGAN

図 A·1 生成例 1
Fig. A·1 Examples of generated images 1.

図 A·2 生成例 2
Fig. A·2 Examples of generated images 2.

図 A·3 生成例 3
Fig. A·3 Examples of generated images 3.

図 A·4 生成例 4
Fig. A·4 Examples of generated images 4.

図 A·5 生成例 5
Fig. A·5 Examples of generated images 5.

の結果を示している．3 行目は着せ替えた服の画像である．一番左の列が着せ替える人物の画像と元々着用している服の画像である．

図 A·1，A·2，A·3 は着せ替えるモデルの服が単色の場合で，図 A·4，A·5 は着せ替えるモデルの服に模様がある場合を示している．

久保 静真 （学生会員）

2018 年東京大学工学部卒業．2018 年に同大学院工学系研究科技術経営学専攻修士課程在籍．専門は機械学習・深層学習応用．

岩澤 有祐

2012 年上智大学理工学部卒業，2014 年同大学院理工学系研究科修士課程，2017 年東京大学大学院工学系研究科博士課程修了．博士（工学）．2018 年東京大学大学院工学系研究科技術経営戦略学専攻松尾研究室特任研究員．同年より同大学特任助教．専門はウェアラブルセンシング，深層学習．

鈴木 雅大 （正会員）

2013 年北海道大学工学部卒業．2015 年同大学大学院修士課程修了．2018 年東京大学工学系研究科博士課程修了．博士（工学）．2018 年より東京大学大学院工学系研究科技術経営戦略学専攻特任研究員．人工知能，深層学習の研究に従事．

松尾 豊 （正会員）

1997 年東京大学工学部卒業．2002 年同大学院博士課程修了．博士（工学）．産業技術総合研究所，スタンフォード大学を経て，2007 年東京大学大学院工学系研究科技術経営戦略学専攻准教授．2012 年人工知能学会理事・編集委員長，2014 年倫理委員長．人工知能学会論文賞，情報処理学会長尾真記念特別賞，ドコモモバイルサイエンス賞等受賞．専門は，Web 工学，Deep Learning，人工知能，Deep Learning，スケーラビリティ，Web Mining．