

# 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

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

久保 静真<sup>\*1</sup> 岩澤 有祐<sup>\*1</sup> 松尾 豊<sup>\*1</sup>  
Shizuma Kubo Yusuke Iwasawa Yutaka Matsuo

<sup>\*1</sup>東京大学大学院工学系研究科 松尾研究室  
The University of Tokyo, Matsuo Laboratory

We propose a novel virtual try-on method based on generative adversarial networks (GANs), which we call SwapGAN. Conditional Analogy GAN (CAGAN) has already been proposed as a virtual try-on method based on GANs, though this method is not good at generating with complex patterns of clothing. By considering clothing regions, SwapGAN enables us to reflect the pattern of clothes better than CAGAN. 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. The network learns how to apply new clothing to the area of people's clothing. Furthermore, an image of the clothes that the person is originally wearing becomes unnecessary during testing.

## 1. はじめに

近年、ファッション分野の E-commerce(EC) サイトの需要が高まっている。今後も国内外ともファッション EC 市場は拡大することが予測されている。衣服の選択に際して、それが自分に似合うかどうかは重要であり、店舗では試着によってそれが確認出来る。一方、ファッション EC サイトではそのような確認は難しい。もしサイト上でも衣服の相性を判断する材料を提示出来れば、ファッション EC サイトにおける購買体験が改善される可能性は高い。また、技術動向として、深層生成モデルのひとつである Generative Adversarial Networks (GAN) [Goodfellow 14] と呼ばれる手法の研究が進み、画像の生成において有効であることが分かっている。特に、Conditional Analogy GAN (CAGAN) [Jetchev 17] は写真上の人物の着せ替えに関する GAN の研究として既に提案されている。写真上の人物の自動着せ替えとは図 1 のような処理である。つまり、人物画像と服の画像を入力として、着せ替えを行った後の人物の画像を生成することである。なお、本稿では上着の着せ替えを行う。この自動着せ替えは衣服の相性を判断する材料となりうるため、EC サイト上の購買体験向上が期待される。本研究では、CAGAN の持つ服の模様に関する課題点を改善することを目指した。複雑な模様の服に関しても有効となるように、服の領域のセグメンテーションを明示的に組み込んだネットワークを提案し、評価を行った。

## 2. 先行研究

自動着せ替えの技術としては、DRAPE(DResing Any Person) [Guan 12] がある。DRAPE は 3 次元の様々な姿勢のバーチャルアバターに対して、2 次元の服のデザインを着用させた。また、[Sekine 14] では、深度カメラで得られた情報を利用して 2 次元の服のデータを現実世界のユーザーにフィッティングするシステムを提案した。システムでは人物の身体を映した画像と深度カメラで得られた情報を入力として身体の形を推定し、その後身体に適するサイズの服をデータベースから取得し、服を身体に覆うように合成する。最近では、[Yang 16]

連絡先: 久保静真, 東京大学工学系研究科松尾研究室,  
08015475717, kubo@weblab.t.u-tokyo.ac.jp

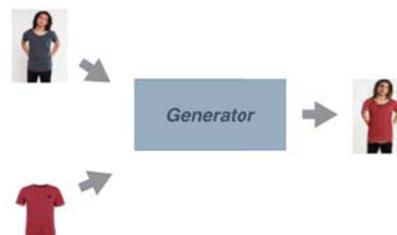


図 1: 本稿の行う着せ替えの模式図を表している。人物の画像と着せ替える対象の服を入力として、着せ替わった画像を出力することになる。

で 2 次元の画像から身体と服の 3 次元モデルへの復元を行った。服を他の身体に当てはめることで着せ替えが出来ることについても言及されている。また、[Pons-Moll 17] は人物をスキャンして、その人物のアバターに服を当てはめてフィット感を確認出来る。一方で、3 次元の計測を利用せずに 2 次元の画像から自動着せ替えを行う研究も進んでいる。CAGAN [Jetchev 17] では GAN を利用し end-to-end で自動着せ替え学習を行った。また、VITON [Han 17] は粗い画像を生成する Encoder-Decoder のステージとその出力を精錬した画像を出力するステージの 2 段階のモデルを構築し自動着せ替えを行った。いずれの研究でも、学習のデータセットは容易に集められるデータセットであり、さらに 3 次元計測を利用した手法の研究と比較して計算コストが小さく済み効率的であるところに利点がある。

本稿における提案手法もこの 2 次元の画像からの自動着せ替えにあたり、CAGAN の持つ服の模様に関する課題点を改善したモデルを提案する。

## 3. 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。その後提案手法である SwapGAN について説明する。図 2 の上

側が CAGAN の Generator の模式図, 下側が SwapGAN の Generator の模式図である。

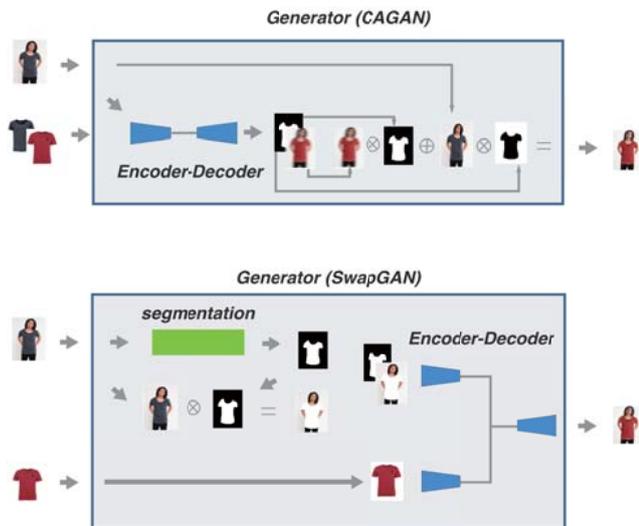


図 2: CAGAN と SwapGAN の Generator のネットワーク構成を表している。上側が CAGAN で下側が SwapGAN の Generator を表している。SwapGAN はまず、入力的人物画像の服の領域を特定する。そして、その領域を取り除いた人物の画像を用意する。それと服の画像を入力として Encoder-Decoder のネットワークが着せ替わった画像を生成する。

### 3.1 CAGAN とその課題

CAGAN が学習するためには損失関数が必要で、その項は以下の式 (1) のように  $L_{cGAN}(G, D)$ ,  $L_{id}(G)$ ,  $L_{cyc}$  の 3 つがある。[Jetchev 17] では  $\gamma_i = 0.1, \gamma_c = 1$  として学習を行っている。

$$\min_G \max_D V(D, G) = L_{cGAN}(G, D) + \gamma_i L_{id}(G) + \gamma_c L_{cyc}(G). \quad (1)$$

CAGAN の Encoder-Decoder はトリプレットと呼ばれ 3 種類の画像を入力とする。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  $\alpha_i^j$ , 3 チャンネルを  $\tilde{x}_i^j$  とする。Encoder-Decoder の関数を  $F$  とすると、式 (2) のような合成によって Generator の出力の画像となる。なお、 $x_i, y_i$  がそれぞれ人物画像とその人物が着ている服の画像を表しており、データセットは  $\{x_i, y_i\}_{i=1}^N$  のような  $N$  組のペアの集合からなる。

$$\begin{aligned} [\tilde{x}_i^j, \alpha_i^j] &= F(x_i, y_i, y_j). \\ G(x_i, y_i, y_j) &= \alpha_i^j \tilde{x}_i^j + (1 - \alpha_i^j) x_i. \end{aligned} \quad (2)$$

まず、 $L_{cGAN}(G, D)$  は conditional 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。

$$\begin{aligned} L_{cGAN}(D, G) &= E_{x_i, y_i \sim p_{data}} [\log D(x_i, y_i)] \\ &+ E_{x_i, y_j \sim p_{data}} [\log(1 - D(G(x_i, y_i, y_j), y_j))] \\ &+ E_{x_i, y_j \sim p_{data}} [\log(1 - D(x_i, y_j))]. \end{aligned} \quad (3)$$

$L_{id}(G)$  は  $\alpha_i^j$  に対して設ける損失関数である。この項によって、服に関係ない領域は出来るだけ変換後も残すようになる。結果的に、 $\alpha_i^j$  は服のセグメンテーションを意味することになる。

$$L_{id}(G) = E_{x_i, y_i, y_j \sim p_{data}} \|\alpha_i^j\|. \quad (4)$$

最後に、式 (5) は服の着せ替えは着せ替えた後にもう一度はじめての服を着せ替えると元に戻ることが期待されるための式である。元々の画像との L1 ロスをとる。

$$L_{cyc}(G) = E_{x_i, y_i, y_j \sim p_{data}} \|x_i - G(G(x_i, y_i, y_j), y_j, y_i)\|. \quad (5)$$

CAGAN の課題としては単色の服のような単純な模様の服の着せ替えについては比較的うまくいっているが、複雑な模様に対してはうまくいかない課題がある。また、入力として写真の人物が着ている服の画像が必要となるため、実用上不便となる。

### 3.2 SwapGAN

本稿の提案手法である SwapGAN のは、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 2 の下側に示す。まず、入力した人物画像の服の領域を特定するセグメンテーションを行う。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 のネットワークは [Isola 17] のネットワークを参考にしている。

CAGAN の Generator は、Encoder-Decoder の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 は式 (6) のように表される。なお、 $M$  は人物の服の部分のマスクとなる。

$$G_{swap}(x_i, y_i) = F(x_i \odot M(x_i), M(x_i), y_j). \quad (6)$$

提案手法の損失関数は式 (7) のように定義される。

$$\begin{aligned} \min_{G_{swap}} \max_D V(D, G_{swap}) &= L_{cGAN}(G_{swap}, D) \\ &+ L_{cyc}(G_{swap}) + L_{perceptual}(G_{swap}). \end{aligned} \quad (7)$$

以下の式 (8), (9) にそれぞれ、 $L_{cGAN}(G_{swap}, D)$  と  $L_{cyc}(G_{swap})$  を示す。 $L_{cGAN}(G_{swap}, D)$  と  $L_{cyc}(G_{swap})$  の項は CAGAN と類似の項であるが、Generator の中にセグメンテーションのネットワークが明示的に組み込まれていることに違いがある。なお、それによって入力に写真上の人物が着ている服の画像である  $y_i$  は不要となる。

$$\begin{aligned} L_{cGAN}(D, G_{swap}) &= E_{x_i, y_i \sim p_{data}} [\log D(x_i, y_i)] \\ &+ E_{x_i, y_j \sim p_{data}} [\log(1 - D(G_{swap}(x_i, y_j), y_j))] \\ &+ E_{x_i, y_j \sim p_{data}} [\log(1 - D(x_i, y_j))]. \end{aligned} \quad (8)$$

$$L_{cyc}(G_{swap}) = E_{x_i, y_i, y_j \sim p_{data}} \|x_i - G_{swap}(G_{swap}(x_i, y_j), y_i)\|. \quad (9)$$

また、以下の式 (10) に  $l_{\text{perceptual}}$  と呼ばれる  $L_{\text{perceptual}}(G_{\text{swap}})$  を示す。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[Simonyan 17] 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  $l_{\phi}$  の和を取ったものが  $l_{\text{perceptual}}$  である。  $\lambda$  は各層のパラメータ数の逆数である。[Hohson 16, Han 17] に習って、  $l_{\text{perceptual}}$  を追加することで服の模様の領域を考慮出来るようになることを期待している。

$$L_{\text{perceptual}}(G_{\text{swap}}) = E_{x_i, y_i \sim p_{\text{data}}} \left[ \sum_{i=1} \lambda_i l_{\phi, \text{block}_i \text{conv}_2} \right]. \quad (10)$$

## 4. 実験

### 4.1 実験内容

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ファッション EC サイト Zalando (<https://www.zalando.de>) の Website から取得した。また、画像のサイズは 128x96 とした。取得した人物画像は正面を向いたもの、服の画像は 1 着が写ったものに限定し、他はノイズとして取り除いた。計 9286 枚のうち、9000 枚を学習に、286 枚をテストに利用した。

実装は代表的な DeepLearning のフレームワークの一つである Keras を Tensorflow バックエンドで利用して行った。最適化手法には Adam[Kingma 15] を用いた。なお、ネットワークは畳み込み層と逆畳み込み層を多層に積み上げた形になっている。各層はバッチ正規化を行い、ReLU または LeakyReLU を活性化関数として利用している。同様に従来手法の CAGAN も実装し比較を行った。また、セグメンテーションのネットワークは他の研究 [Gong 17] 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[Chen 15] というモデルに 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本提案手法の学習では更新せずに、[Gong 17] 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。各手法の生成結果は図 3 に示す。

### 4.2 評価

画像を生成するモデルを評価するための指標は Inception Score [Salimans 16] や FID [Heusel 17] のような指標など、いくつか提案されているが、着せ替えの良し悪しを判断する指標にはならない。そのため、アンケートによって提案手法の有効性を評価した。

テストにはテストデータセット内の画像を用いて 30 件の着せ替えを本提案手法と従来手法の CAGAN でそれぞれ行い、どちらが着せ替えとして適切かのアンケートを行った。回答は 131 人から得た。質問はテストデータセットの中から人物画像と服の画像を取得し、CAGAN 及び SwapGAN でそれぞれ着せ替えの処理を行った生成画像を提示しどちらのほうが着せ替えとして適切かを質問した。全質問の回答を平均すると提案手法を適切と答えた割合は 82.2% であり、本提案手法の有効性が示された。

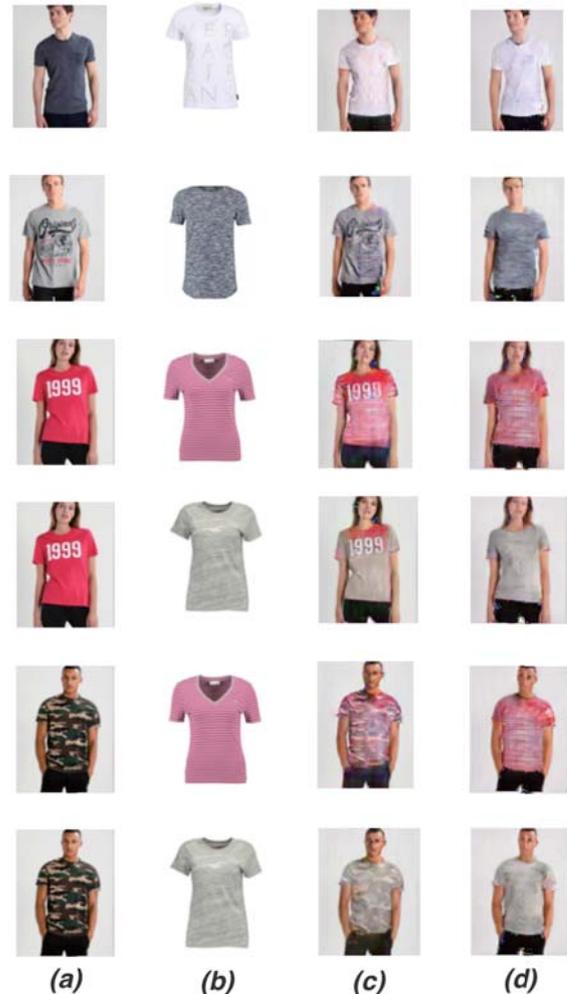


図 3: CAGAN 及び SwapGAN の画像の生成結果の例を示している。(a) と (b) はデータセット中の画像である。(a) の人物の服を (b) の服に着せ替える。(c) が CAGAN の生成結果を表しており、(d) が SwapGAN による生成の結果を表している。

## 5. 考察

図 4 では 2 つのサンプルに対して各手法の生成結果の比較を示している。1 行目を見てみると、単調な服の模様に対しては CAGAN と SwapGAN の生成結果はあまり差はないように見える。一方、2 行目を見てみると、複雑な模様の服に対しては CAGAN よりも SwapGAN のほうが生成結果が良いように見える。ここで、CAGAN の服の領域部分を示す (d) の列をみてみると、複雑な模様の服に対しては比較的うまくいっていないように見える。一方で、SwapGAN の服の領域部分を示す (f) の列をみてみると、どちらの服のパターンに対してもうまくいっているように見える。SwapGAN では、服の領域を慎重に考慮したことによって、元々の服の模様による悪影響を小さくすることが出来たと考えられる。結果として、着せ替える服の模様をより良く結果に反映出来たと思われる。

実験によって提案手法が従来手法と比較してより適切な着せ替えが行えたといえる。領域を明示的に示したことで、元の服

の模様の影響が小さくなり、着せ替える服の模様の反映がうまく出来ていると考えられる。



図 4: 生成の例を 2 つ示している. (a) と (b) はデータセット中の画像である. (b) の人物の服を (a) の服に着せ替える. (c) は CAGAN の結果を表しており, (d) は CAGAN が示す服の領域を意味する領域を表す. 一方で, (e) は SwapGAN が生成した結果を表しており, (f) はセグメンテーションのモデルが生成した領域を表している.

## 6. まとめ

本研究では, GAN による自動着せ替えにおいて, 服の位置のセグメンテーションを明示的に行う手法を提案し, その評価を行った. そして, 従来手法の CAGAN と比較して, その有効性が示された. また, 従来手法の入力では必要であった人物の着ている服の画像も不要となった.

本研究においても, 服の模様に関してまだ改善の余地があり, また, 人物の姿勢を限定したり, 服を上着に限定したりするなどの適用範囲の制限もある. この点においてさらなる向上が期待される.

## 参考文献

- [Goodfellow 14] Goodfellow, Ian J. and Pouget-Abadie, Jean and Mirza, Mehdi and Xu, Bing and Warde-Farley, David and Ozair, Sherjil and Courville, Aaron and Bengio, Yoshua: Generative Adversarial Networks, in Neural Information Processing Systems (NIPS) (2014)
- [Jetchev 17] Jetchev, Nikolay and Bergmann, Urs: The Conditional Analogy GAN: Swapping Fashion Articles on People Images, in International Conference on Computer Vision (ICCV) (2017)
- [Guan 12] Guan, Peng and Reiss, Loretta and Hirshberg, David A. and Weiss, Alexander and Black, Michael J.: The Conditional Analogy GAN: Drape, in ACM Transactions on Graphics (2012)
- [Sekine 14] Sekine, Masahiro and Sugita, Kaoru and Perbet, Frank and Stenger, Björn and Nishiyama, Masashi: Virtual Fitting by Single-Shot Body Shape Estimation (2014)
- [Yang 16] Yang, Shan and Amert, Tanya and Pan, Zherong and Wang, Ke and Yu, Lich and Eng and Lin, Tamara Berg and Ming C. and Hill, University of North Carolina at Chapel: Detailed Garment Recovery from a Single-View Image (2016)
- [Pons-Moll 17] Pons-Moll, Gerard and Pujades, Sergi and Hu, Sonny and Black, Michael J.: ClothCap: Seamless 4D Clothing Capture and Retargeting, in ACM Transactions on Graphics (2017)
- [Han 17] Han, Xintong and Wu, Zuxuan and Wu, Zhe and Yu, Ruichi and Davis, Larry S.: VITON: An Image-based Virtual Try-on Network (2017)
- [Gong 17] Gong, Ke and Liang, Xiaodan and Zhang, Dongyu and Shen, Xiaohui and Lin, Liang: Look into Person: Self-supervised Structure-sensitive Learning and A New Benchmark for Human Parsing, in Computer Vision and Pattern Recognition (CVPR) (2017)
- [Isola 17] Isola, Phillip and Efros, Alexei A and Ai, Berkeley and Berkeley, U C: Image-to-Image Translation with Conditional Adversarial Networks, in Computer Vision and Pattern Recognition (CVPR) (2017)
- [Simonyan 17] Simonyan, Karen and Zisserman, Andrew: Very Deep Convolutional Networks for Large-Scale Image Recognition, in International Conference for Learning Representations (ICLR) (2017)
- [Kingma 15] Kingma, Diederik P. and Ba, Jimmy: Adam: A Method for Stochastic Optimization, in International Conference for Learning Representations (ICLR) (2015)
- [Salimans 16] Salimans, Tim and Goodfellow, Ian and Zaremba, Wojciech and Cheung, Vicki and Radford, Alec and Chen, Xi: Improved Techniques for Training GANs, in Neural Information Processing Systems (NIPS) (2016)
- [Heusel 17] Heusel, Martin and Ramsauer, Hubert and Unterthiner, Thomas and Nessler, Bernhard and Hochreiter, Sepp: GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium, in Neural Information Processing Systems (NIPS) (2017)
- [Hohson 16] Johnson, Justin and Alahi, Alexandre and Fei-Fei, Li: Perceptual losses for real-time style transfer and super-resolution, in European Conference on Computer Vision (ECCV) (2016)
- [Chen 15] Chen, Liang-Chieh and Yang, Yi and Wang, Jiang and Xu, Wei and Yuille, Alan L.: Attention to Scale: Scale-aware Semantic Image Segmentation (2015)