The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

深層学習を用いたアパレルアイテム平置き画像から着装状態への変換

Image-to-image Translation from Apparel Item Image Placed Flat to Image Put on Using Deep Neural Networks

積際 早紀[*1] Saki Tsumugiwa　黒澤 義明[*2] Yoshiaki Kurosawa　目良 和也[*2] Kazuya Mera　竹澤 寿幸[*2] Toshiyuki Takezawa

[*1] 広島市立大学 情報科学部 School of Information Sciences Hiroshima City University

[*2] 広島市立大学大学院 情報科学研究科 Graduate School of Information Sciences Hiroshima City University

This paper deals with image-to-image translation of apparel items. The images are difficult to be translated because the items are variously set, when they are took photos: being placed flat, being put on the mannequin and so on. We try to investigate and improve the previous work also known as 'pix2pix' based on deep neural networks, especially deep convolutional generative adversarial network (DCGAN). We propose a new two-stage procedure. Some experimentation revealed that our proposed method was superior to the previous work, evaluated using structural similarity index. Moreover, we confirmed it generated item details (zipper, button) and patterns (dot) as the result of visual confirmation. This knowledge is very important because the fault image of the item without buttons should be completely different from the original item image.

1. はじめに

インターネットの発展に伴い，各種通販サイトが開設されてきた．アパレル分野も例外ではなく，各メーカー/ブランドのサイトに加え，ZOZOTOWN，MAGASEEK 等の大規模複合サイトが開設され，今に至っている．

これら複合サイトの利点は，様々なブランドの商品を一度に閲覧し，また購入できることである．その一方で，ブランド毎に用意される画像が統一しておらず，ユーザのイメージ想起を困難にすることがある．

例えば，サーキュラースカート（図 1(a)）のように広がりが大きいスカートは，着装した時に「フレア感」と呼ばれる印象を形成する． この点，図 1(a)左よりも図 1(a)右の方が印象を伝えやすい．しかし，短いサイクルでデザインから縫製，販売まで手掛ける現代のファストファッションでは，写真撮影に時間を割いてばかりもいられない．このため，『基本的な写真（例えば，平置き画像）からの，自動的なバリエーション画像（例えば，着用画像）変換』が望ましいと考えられる．

そこで本研究は，このような異なるバリエーション画像を，深層学習を用いて自動で変換し，ユーザの印象形成を容易にすることを目的とする．特に，図 1(a)のように変形の大きい対象についても変換できる手法の確立を目指す．

(a) 変形大　(b) 変形小

図 1　アパレルサイト内バリエーションの例

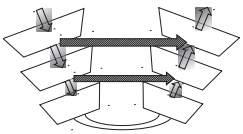

図 2　U-Net の模式図

連絡先：黒澤 義明，広島市立大学大学院 情報科学研究科，広島市安佐南区大塚東 3-4-1，082-830-1500（代），kurosawa@hiroshima-cu.ac.jp

2. 関連研究

本研究はアパレルアイテム画像から，様々なバリエーション画像の変換生成を目的とする．深層学習を用いた画像変換の研究として，いわゆる pix2pix が挙げられる[Isola 17]．

本研究はこの pix2pix を利用して，大きな変形を伴う画像変換を実現する．そこで，まず本研究にとって重要な pix2pix の特徴と検討項目を挙げた上で，関連研究を 2 つ挙げる．

2.1 pix2pix

pix2pix は conditional Generative Adversarial Networks（GAN）の一種に基づく画像変換器である[Isola 17]．線画への着色やセグメンテーション等，様々な対象に利用されている．

次に pix2pix の特徴を 2 つ挙げる．

(1) Generative Adversarial Networks（GAN）

第一の特徴は，2 つの畳み込みニューラルネットワーク（生成器 G(Generator)と判別器 D(Discriminator)）を使用する点である．

まず生成器 G は，訓練画像に似た画像(fake)を生成するように訓練される．次に判別器 D は，D に入力された画像が訓練画像か，G が生成した fake かを判定する．そして，G は D を騙すように訓練され，D は G に騙されないように訓練される．このように，G と D が互いに争い合うことで画像の変換精度を高める．

(2) U-Net

次の特徴は，畳み込みネットワークに U-Net を使用している点である．このネットワークは，医療分野での画像セグメンテーション（細胞の着色 etc）を目的として開発された[Ronneberger 15]．重要な点は，U-Net という名前が示す構造にある（図 2）．

U 字の左右をつなぐ結合（skip connection）がある（右向き車線矢印）．この結合により，U 字の右側では下層から出力された特徴マップに，左側の対応する層からの特徴マップが加わる．これがお手本のようになり，結合のない単なる Encoder-Decoder モデルよりも，速く正確に学習する．一般的に，GAN による画像生成は細部が破綻することも多く，U-Net を使う利点は大きい．

2.2 pix2pix における要検討項目

[Isola 17]は車や人のセグメンテーション，線画に対する着色等，様々な変換事例を紹介している．しかし，pixel 同士の対応を前提としており，変形を伴う対象を検討していない．よく似た CycleGAN [Zhu 17]では，変形の問題が指摘されているため，吟味する必要があると言える．

本研究が対象とするアパレルアイテムは，図 1(a)に挙げたような変形を含む．こうしたアイテムの変形が可能であれば，背景からのアイテム抽出や，アイテムの変形加工等の複数の処理を，モデルを変えるだけで実行できるという利点がある．

以上の観点から，pix2pix が変形に対応可能かどうか議論することを本研究の課題の 1 つとする．問題があるとすればどのような点か，そしてどのように解決するかに検討する．

次に，アパレルアイテムに関する関連研究について，以下に補足しておく．

2.3 Pixel Level Domain Transfer

pix2pix 以外にも変換を試みる研究がある．[Yoo 16]は，本研究同様，アパレルアイテム画像の様態変換を行っている．この研究も GAN を利用している．

異なる点は，U-Net ではなく単純な Encoder-Decoder モデルを使用していること，G と D に加えて，Domain 判別用の別のネットワークが用意されている点，の 2 点である．先に述べたように，Encoder-Decoder モデルより U-Net が優れる．

2.4 Detailed Garment Recovery

単なる着用画像の変換だけではなく，動画化を試みた研究もある[Yang 16]．フレアスカートのフレアを揺らすことができる等，ユーザのイメージ想起に有効に働く手法である．

ただ，彼らの手法の問題は，型紙から衣服を作るかのような詳細なアイテム記述が必要なことであり，コストがかかる点である．短いサイクルで様々なアイテムが発売されるファストファッションの場合には，このようなコストをかけることは困難である．したがって，シンプルな画像変換による方法の方が，コストが少なく，より有効であると考えられる．

3. 提案手法

本研究の提案は，pix2pix を用いた変形を伴う画像変換に対する解決法である．以下に，その手法について述べる．

先述の通り，pix2pix の手法は，画像間のピクセルについて対応が取れているとき，または対応のずれが小さいときには有効である．一方で，変形が大きく，対応のずれが大きくなると，学習が困難になると考えられる．

この理由としては，pix2pix の利点となっている skip connection に原因がある．すなわち，お手本として振る舞うべきであるのに，対応のずれが災いする．つまり，オリジナルの画像情報が反映されない学習結果となるわけである．

そこで本研究は，お手本となるべき入力を新たに導入するため，2 段階の処理を採用し，つなぎ目に処理を加える方法を提案する(図 3)．

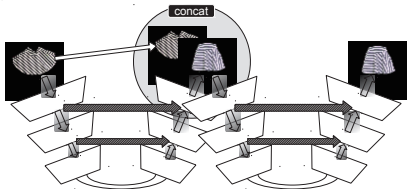

図 3　本研究の提案手法

本研究の手法はシンプルであり，2 つの pix2pix をつなげ，そのつなぎ目で，1 つめの pix2pix の出力と，最初の入力(白抜き矢印)を重ね合わせて(concat)，2 つめの pix2pix への入力とする．これにより，オリジナル画像の情報を有した 6 次元の画像による学習が可能となる．これにより，変形への対応を目指す．

4. 実験と考察

本節では，pix2pix の変形にの可否に関する調査を行い，その問題点を解消する実験について報告する．

4.1 本研究で使用するデータ

本稿では対象をスカート，また，平置き画像(図 1(a, b)左)から着装状態(図 1(a, b)右)への変換に限定する．に限定した理由は，変形量が大きく，検証に有効と考えたからである．図 1(a)が変形の大きい例，(b)が小さいタイトスカートの例である．

以下に，本研究で使用したデータの詳細について述べる．

(1) 撮影

1 枚のスカートに付き，平置き画像(床に広げ，上方から撮影)と着装状態画像(マネキンに着用)の 2 枚ずつを撮影した．その後，それぞれ背景を削除した．アイテム数は 283 であった．

(2) Data Augmentation

上記アイテム数は，深層学習に十分な量とは言えないため，data augmentation を行った．以下の表 1 の命令からランダムに 4 種類を選び，色等を変更する．なお，オプションランダムに選ぶ．

上の手続きを 1 枚のアイテムにつき 21 回行った上で，さらに左右反転を行った．これにより，オリジナルの画像も含め，22 (21+1) x 2(反転)倍の画像となった(283x44=12,452 画像)．

(3) 画像サイズ調整

平置きと着装状態は別々に撮影しているため，サイズの調整を行う．それぞれ，輪郭抽出を行った後，外接四角形を決定する．そして，着装画像の縦サイズに合わせ，縦横比を維持したまま，平置き画像を縮小/拡大し，縦横 256 pixel の画像を得た．

4.2 実験 1 ~基礎的な変形に関する検討~

本研究は pix2pix の実装として，原著者らが公開している実装を用いた．なお，特に断らない限り，デフォルトのパラメータを使用する．GPU は GeforceGTX 1080Ti を使用した．

まず基本的にな変形に関する情報収集のため，判別器 D の層数を 5 に変更(デフォルト:3)し，バッチサイズを 10 で 1,000 epoch(同:200)繰り返す実験を行った．学習時の loss の推移を図 5 に示す．学習開始直後の L1 の値が大きく見づらいため，補正した図になっている．

表 1　data augmentation に使用した命令とオプション

命令	オプション		
画像の伸縮	[0.9], [1.05], [1.1]		
回転	[3], [5], [7]（度）		
透視変換	[0.015]（画像幅 x0.015）, [0.03], [0.05], [0.07],		
色交換	[0, 2, 1, 3]（BGR→BRG）, [1, 0, 2, 3], [1, 2, 0, 3], [2, 0, 1, 3], [2, 1, 0, 3]		
色軽減	[0.95]（Bx0.95，Gx0.95，Rx0.95）, [0.9]		
特定色軽減	[0, 0.95]（Bx0.95）, [1, 0.95], [2, 0.95]		
色除去/半減	[0, 0]（Bx0）, [0, 0.5], [1, 0], [1, 0.5], [2, 0], [2, 0.5]		
HSV 円環上の移動	[-40], [-30], [-20], [-10], [10], [20], [30], [40]		
平滑化/先鋭化	[0], [1], [2]		
	[0]: Gaussian Filter Kernel Size(11x11)	[1]: [-1, -1, -1] [-1, 9, -1] [-1, -1, -1]	[2]: [0, -1, 0] [-1, 5, -1] [0, -1, 0]

ここで，「G_GAN」は『生成器 G が判別器 D を騙せているか』についての指標，「G_L1」は『生成器 G が生成した fake(predict)と正解(ground truth)との L1 距離』に関する指標である．今回の学習では収束条件が L1 となっている．また，「D_real」は『正解に対する，判別器 D の判定』，「D_fake」は『fake に対する，判別器 D の判定』について表している．

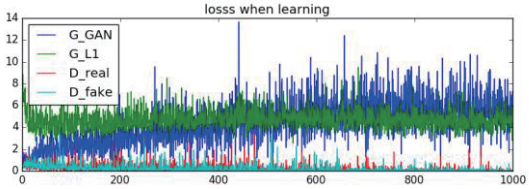

図 5　loss の推移（補正後）

G_L1，G_GAN 共に変動が大きいことがわかる．また，L1 が横這いである点も学習が不十分である可能性を示唆している．このため今回の検討は，以下の実験を含め，全てクローズドテストで行うこととする．学習時のアイテムをモデルに入力し，生成した画像を挙げる（図 6～図 8）．

変形の必要なサーキュラースカート（図 6）やフレアスカート（図 7）では，アイテム領域の内外に破綻が見られる（ex. 600 epoch）．ただ，変形はおおむね成功しているように見える．また，シンプルなタイトスカート（図 8）でも変形はできていても，テクスチャーの変換には成功していない．仮説としては，ネットワークの重み学習が変形を中心に行われて，余裕がなかったのではないか，つまり学習資源に関するパラメータ設定が必要と考えられる．

4.3 実験 2 ~パラメータ変更~

前項の仮説を確認するため，生成器 G が持つカーネルの数を増加させる．デフォルトが 512 の層 1024 にする等，各層 2 倍にする[1]．また，性能向上のため，判別器 D の性能向上を目指しカーネルの大きさを 1 にした実験を行う[2]．アイテムを学習モデルに入力した例を，次の図 9～図 11 に示す．

実験 1 では成功していなかったストライプ（図 6）の復元に成功している（図 9）．また図 10 では，フレアスカートのフレア感も再現している．一方で，ドット柄の再現については不十分であり，さらに図 11 からわかるように，ディテール（ボタンやジッパー）がどの epoch でも再現されていない．

つまり，重み学習のためにパラメータの変更～特に，カーネル数等の増加～を行うことは重要である．しかし，それでも細かい模様やディテールを再現できないという問題が残る．

4.4 実験 3　提案手法 ~2 段階処理への拡張 ~

3 節に述べた提案手法に基づく実験を行う．第一段階には前節で学習した中で，200 epoch のモデルを選んだ．200 はデフォルトの値である．

このモデルにより得られた変換画像を 6 次元に合成した上で，2 段階目の学習を行った．結果を図に示す（図 12，図 13）．

図 10 に比べた図 12 では，細かいドット柄を表現できていることがわかる．また，図 11 で不十分だったディテールが，図 13 では表現できていることもわかる．

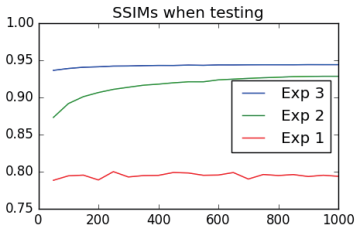

図 4　3 実験比較を目的とした SSIM 値の推移

4.5 考察

変換がどの程度向上したか，以下の Structural Similarity（SSIM）を用いて評価する．画素中の小領域の平均・分散・共分散を用いて，局所的な破綻に敏感であり，最良値は 1 である．

$$\mathrm{SSIM} = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}$$

撮影画像 283 枚それぞれについて，50 epoch ごとに平置き-着装間の平均 SSIM を求めた（図 4）．

実験 1 と比べ，実験 2 と実験 3（提案手法）との値が高くなっていることがわかる．また実験 3 が優れていることも明らかである．

ただ，実験 2 との差はあまりなく，pix2pix のパラメータを変更しただけであるので，「提案手法のようなコストをかけなくてもよいのでは」という指摘もあろう．しかし，前節で述べたように，見た目の変化は大きい（図 10 vs. 図 12，図 11 vs. 図 13）．特に，ボタン等のディテールが再現できなければ，全く別のデザインになることは指摘しておきたい．その意味で，数値以上に提案手法が優れていると言える．

5. おわりに

アパレルアイテム画像の様態変換を目的に，pix2pix を利用した 2 段階の学習方法を提案した．

実験の結果，SSIM による数値による比較で，本提案が優れていることが明らかになった．pix2pix は細かい模様やディテールを再現できない一方，提案手法では再現できたからである．

今後の課題としては，アイテム数を増やした上で，さらなる検討を試みることが挙げられる．また SSIM 値は，ディテールの有無等を反映していないこともわかった．さらなる改善のため，適切な指標を検討することも必要であろう．

謝辞

この研究の一部は，国立研究開発法人科学技術振興機構(JST)の研究成果展開事業「センター・オブ・イノベーション(COI)プログラム」．広島市立大学特定研究費（先端学術研究費 H27～29，30 年度科研費獲得支援費）の補助を得ている．

参考文献

[Isola 17] P. Isola, J.-Y. Zhu, T. Zhou, A. A. Efros : Image-to-Image Translation with Conditional Adversarial Networks, In IEEE Conference on Computer Vision and Pattern Recognition, 2017.

[Ronneberger 15] O. Ronneberger, P. Fischer, T. Brox: U-Net: Convolutional Networks for Biomedical Image Segmentation, arXiv:1505.04597, 2015.

[1] 関数 UnetGenerator の ngf を 64 から 128 に変更した．

[2] 関数 PixelDiscriminator．なお，実験 1 と 2 の間にはカーネルのサイズを変更した複数の実験がある．ページの都合により今回は割愛する．

[Yang 16] S. Yang, T. Ambert, Z. Pan, K. Wang, L. Yu, T. Berg, M. C. Lin：Detailed Garment Recovery from a Single-View Image, arXiv:1608.01250v4, 2016.

[Yoo 16] D. Yoo, N. Kim, S. Park, A. S. Paek, I. S. Kweon: Pixel-Level Domain Transfer, arXiv:1603.07442, 2016.

[Zhu 17] J.-Y. Zhu, T. Park, P. Isola, A. A. Efros.: Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks, In IEEE International Conference on Computer Vision, 2017.

図 6　実験 1 item002: epoch 毎生成画像（左から 50, 100, 200, 400, 600, 800, 1000 epoch, ground truth）

図 7　実験 1 item215: epoch 毎生成画像（左から 50, 100, 200, 400, 600, 800, 1000 epoch, ground truth）

図 8　実験 1 item200: epoch 毎生成画像（左から 50, 100, 200, 400, 600, 800, 1000 epoch, ground truth）

図 9　実験 2 item002: epoch 毎生成画像（左から 50, 100, 200, 400, 600, 800, 1000 epoch, ground truth）

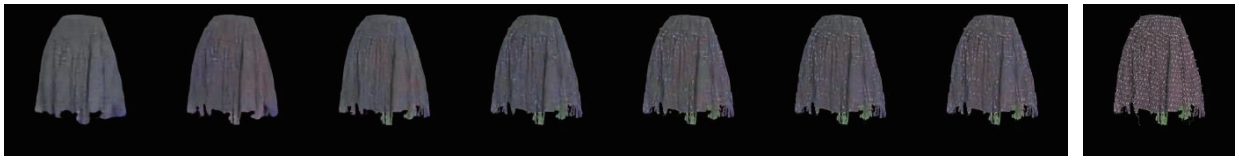

図 10　実験 2 item226: epoch 毎生成画像（左から 50, 100, 200, 400, 600, 800, 1000 epoch, ground truth）

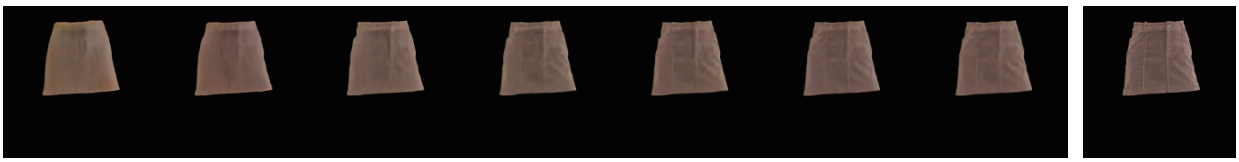

図 11　実験 2 item200: epoch 毎生成画像（左から 50, 100, 200, 400, 600, 800, 1000 epoch, ground truth）

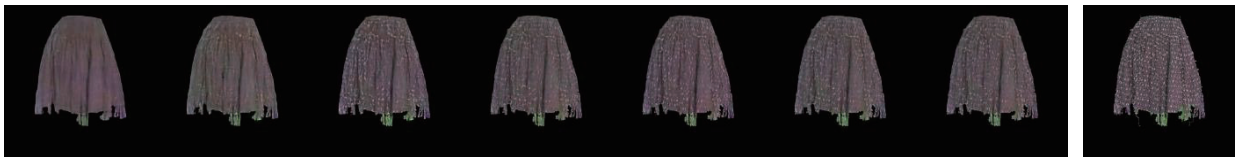

図 12　実験 3 item226: epoch 毎生成画像（左から 50, 100, 200, 500, 550, 600, 800 epoch, ground truth）

図 13　実験 3 item200: epoch 毎生成画像（左から 50, 100, 200, 500, 550, 600, 800 epoch, ground truth）