

身体の3次元構造を考慮したニューラル仮想試着

Neural Virtual Try-On System considering 3D human model

久保 静真^{*1}

Shizuma Kubo

岩澤 有祐^{*1}

Yusuke Iwasawa

鈴木 雅大^{*1}

Masahiro Suzuki

松尾 豊^{*1}

Yutaka Matsuo

^{*1}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 (GAN), which uses 3D surface model of body. In existing GAN-based methods (CAGAN, SwapGAN) sometimes do not work on a human image of rare posture. In our proposed method, by using DensePose to estimate a point corresponding to 3D surface model for each pixel point of 2D image, 3D surface based information is incorporated into our model. Therefore, it is possible to change clothes of people in various postures. Our proposed method uses a coarse-to-fine strategy. First, *Parts Generation Network* generates parts and they are mapped to 2D image to produce coarse dressing image. After that, *Refine Network* refines the coarse dressing image. In our experiment, we show the result of the proposed method and our method has effect on rare postures by comparison with existing methods.

1. はじめに

オンラインショッピングサイトの需要の増加に伴い、仮想試着は注目を集めている。既存の仮想試着の研究では服の模様や特徴を対象の人物の服の領域に遷移させることに注力しているが、身体の向きが変わると自然な画像が生成できないという課題がある。例えば、既存のモデルは身体が直立正面となるような姿勢ではうまくいくが、腕が身体の正面にくるような姿勢ではうまくいかないことがある。既存の研究における身体情報の利用はセグメンテーションや2次元の姿勢推定に限定されているが、身体の向きが変わっても自然な画像を生成するためには、身体構造を考慮するのに必要な新たな情報を補完する必要があると考えられる。

本研究では、様々な姿勢に対しても有効な着せ替えを行なうために、身体の3次元構造を考慮した仮想試着の手法を提案する。仮想試着では、身体の3次元構造の推定とその推定された3次元構造に服をマッピングする2つのプロセスが存在すると考えられるが、既存の手法ではこのプロセスを同時に行っている。提案手法ではこのプロセスを明示的に分離し、身体の3次元構造の推定に既存研究のDensePose[Güler 18]の出力であるIUV^{*1}データを利用することで、既存手法よりも身体構造として正確な服の着せ替え画像を生成できることを示す。提案モデルは2段階のネットワークを用いる。1段目のネットワークでは服の画像から身体の3D表面のUV座標に対応するパーツを生成する。その生成したパーツをIUVデータに沿って、対象の人物に貼り付けることで対象の服に着せ替わった画像を生成する。生成するパーツの各ピクセルは身体の3次元構造に即してマッピングされるため、様々な姿勢であっても対応できる。このマッピングの段階では粗さが残るため、2段目のネットワークでは精錬してより本物らしい画像を生成する。

2. 先行研究

近年では、人物画像に任意の服を着せた画像を生成する仮想試着の研究が複数行われている。画像生成において、Generative Adversarial Networks (GAN) [Goodfellow 14] がよく使われるのと同時に、仮想試着においてもGANを使った手法は提案されている。[Jetchev 17] は [Jun-Yan 17] を応用したGANによる仮想試着の手法を提案した。[Kubo 18] は [Jetchev 17] の研究において、服の領域を考慮することで服の模様の遷移がうまくいくことを示した。また、その他の仮想試着の研究としては、[Han 17] や [Wang 18] がある。[Han 17] では粗い画像を生成するEncoder-Decoderのステージとその出力とthin plate spline (TPS) によって変形した服を入力に画像を精錬するステージの2段階のモデルを構築し、仮想試着を行った。[Wang 18] は [Han 17] を応用して、Geometric Matching Module (GMM) を使って服の変形を行なう機構を取り入れ、服の特徴をより生成画像に反映できることを示した。

最近では、身体の領域を検出し、各ピクセルに対して3Dサーフェスモデルに対応する点を推定する[Güler 18]を利用した研究も行われている。[Neverova 18] は[Güler 18]を姿勢推定のタスクに応用した。また、[Wu 18] は[Güler 18]を用いて2つの画像人物間の服の交換を行った。本稿では、この[Güler 18]によって得られる身体の3次元構造の情報を仮想試着のモデルに組み込むことで身体構造をより考慮した仮想試着のモデルを提案する。

3. 提案手法

提案手法では、2段階のネットワークを学習する。全体像は図1に示す通りである。1段目のネットワークであるパーツ生成ネットワーク(Parts Generation Network)によって、着せ替え後の人物の3次元表面を表す身体各パーツを生成し、そのパーツを利用し、テクスチャマッピング(Texture Mapping)によって2次元の着せ替え画像を生成する。そして、その着せ替え画像を2段目のネットワークである精錬ネットワーク(Refine Network)によって精錬させたものを最終的な結果とする。

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

^{*1} 3Dモデルにテクスチャをマッピングするときに使うUV座標系に各UV座標が身体の中のどのパーツに属するかの情報も加えたものである。UV座標系は2次元の直交座標系で横方向をU、縦方向をVとする。

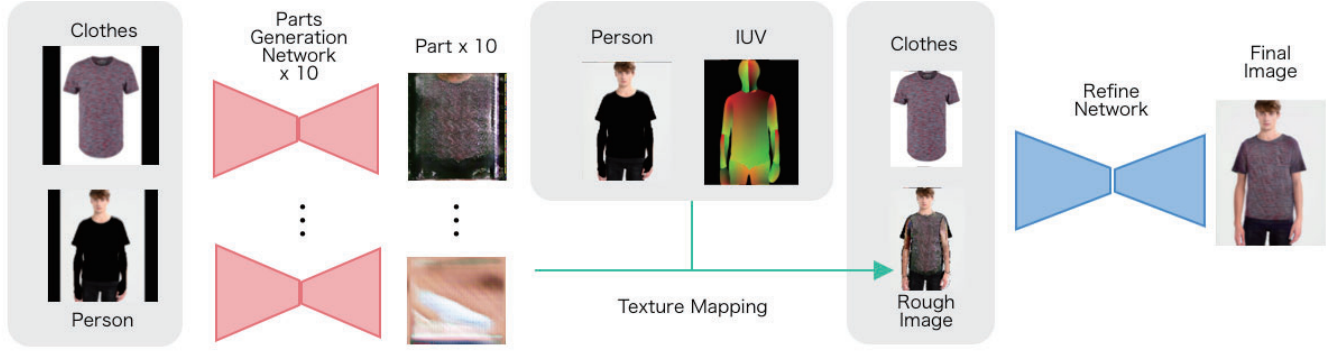


図 1: 全体像

赤のネットワークがパーツ生成ネットワーク（Parts Generation Network）を、緑の線がテクスチャマッピング（Texture Mapping）を表しており、これらによって粗い着せ替え画像（Rough Image）が生成される。その後、図中の青で表された精錬ネットワーク（Refine Network）によってより本物らしい着せ替え画像へと精錬される。なお、入力の服の領域を切り取られた「Person」と DensePose の出力の「IUV」は着せ替える対象の人物画像から得られる。

3.1 パーツ生成ネットワーク

1 つ目のネットワークであるパーツ生成ネットワーク（Parts Generation Network）では対象の服の画像から着せ替え後の人物の 3 次元表面を表す身体各パーツを生成するように学習する。本稿では既存手法と同様に、着せ替える対象を上着に限定するため、生成するパーツは上半身に含まれる合計 10 個のパーツとなる（ただし、両手は除く）。生成するパーツそれぞれに対してネットワークを用意する。それぞれのネットワークは服の画像と人物の服部分を切り抜いた画像の 2 枚の画像を入力とする。人物の服部分を切り抜いた画像を入力に含めたのは人物の肌の色のような身体情報を加味するためである。入力の画像は出力のサイズに合わせ、2 枚とも正方形となるように補完し、128x128 にリサイズして入力とした。10 個のネットワーク全てにおいて入力は同じであり、出力の画像のサイズも 128x128 である。教師データとして、入力の服を着用した人物画像から DensePose [Güler 18] によって得られる IUV データを用いてパーツにしたものを用いる。この人物画像の元のサイズは 1100x762 または 1650x1143 のサイズであり、128x128 のパーツの画像を教師データとして用意した。パーツの例を図 2 に示す。

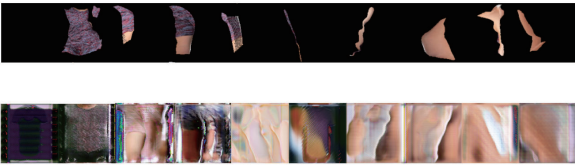


図 2: パーツの生成例

上の段の画像が教師データとなるパーツの画像となる。下の段の画像がネットワークによって出力されるパーツの画像例である。

学習は、各パーツごとに下記の式 (1) の損失関数を最小にするように行われる。G と D はそれぞれ多層パーセプトロンからなる Generator, Discriminator の関数を表したものであり、式 (1) のように G と D の関数 $V(G, D)$ の min-max ゲームによって学習を進める。各項については以下で説明する。

$$\min_{\mathbf{G}} \max_{\mathbf{D}} V(\mathbf{D}, \mathbf{G}) = L_{GAN_{parts}}(\mathbf{G}, \mathbf{D}) + L_1(\mathbf{G}) + L_{perceptual}(\mathbf{G}). \quad (1)$$

用いるデータは、入力の服の画像を c_i 、人物の服部分を切り抜いた画像を r_i 、対応するパーツの画像を y_i 、対応する教師データのパーツの存在する領域のマスクを m_i とし、 $\{c_i, r_i, y_i, m_i\}_{i=1}^N$ のような N 組のペアの集合からなる。以下の式 (2), (3) にそれぞれ、 $L_{GAN}(\mathbf{G}, \mathbf{D})$ と $L_1(\mathbf{G})$ を示す。Generator の出力をパーツの存在領域でマスクして学習に利用する。

$$L_{GAN_{parts}}(\mathbf{D}, \mathbf{G}) = E[\log D(y_i, c_i, r_i)] + E[\log(1 - D(G(c_i, r_i) \odot m_i, c_i, r_i))] \quad (2)$$

$$L_1(\mathbf{G}) = E\|y_i - G(c_i, r_i) \odot m_i\|. \quad (3)$$

なお、 E は期待値を表している。

また、以下の式 (4) では perceptual loss と呼ばれる $L_{perceptual}(\mathbf{G})$ を示す。 $G(c_i, r_i) \odot m_i$ と対応するパーツ画像 y_i を一般物体認識で高い性能を示した VGG19 [Simonyan 15] の学習済みモデルにそれぞれ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。 λ は各層のパラメータ数の逆数である。 [Johnson 16, Han 17] に習って、perceptual loss を追加することで服の模様を考慮出来るようになることを期待している。

$$L_{perceptual}(\mathbf{G}) = E \left[\sum_{i=1} \lambda_i l_{\phi, block_{i-conv2}} \right]. \quad (4)$$

3.2 テクスチャマッピング

パーツ生成ネットワークで生成したパーツを IUV データに基づいて、対象の人物の身体表面にマッピングする。対象の人物の画像は 256x192 のサイズで、貼り付けるパーツの部分は [Chen 15] によって得られる上半身のセグメンテーションデータを用いて取り除いた。なお、両手は IUV データにより元画像から再現した。図 3 に例を示す。

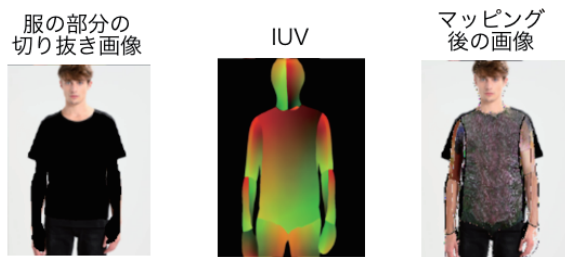


図 3: テクスチャマッピングの例
服の部分を切り抜いた画像に IUV データを元に生成したパーツを貼り付けることでマッピングを行なう。

3.3 精錬ネットワーク

テクスチャマッピング後の着せ替え画像は粗く、補完しきれない部分が残るため、精錬ネットワークにより本物らしい画像に精錬する。入力としてはテクスチャマッピング後の着せ替え画像と着せ替える服の画像を入力とする。全ての入力のサイズは 256×192 である。学習は式 (5) の損失関数を最小にするように行われる。ここでの G , D はパーツ生成ネットワークとは別のネットワークである。 $L_1(G)$, $L_{\text{perceptual}}(G)$ では着用している服と着せ替える服を同じ服にして着せ替え処理を行って、着せ替え前後の差をパーツ生成ネットワークと同様に損失とする。 $L_{\text{GAN}_{\text{refine}}}(G, D)$ については以下で説明する。

$$\min_G \max_D V(D, G) = L_{\text{GAN}_{\text{refine}}}(G, D) + L_1(G) + L_{\text{perceptual}}(G). \quad (5)$$

入力の人物の画像を h_i , 元々着用している服を c_i , 着せ替える服の画像 c_j , テクスチャマッピングによって生成された粗い着せ替え画像を r_{ij} として、以下の式 (6) に $L_{\text{GAN}_{\text{refine}}}(G, D)$ を示す。なお、この項は既存手法である CAGAN[Jetchev 17] や SwapGAN[Kubo 18] と同様の項である。

$$L_{\text{GAN}_{\text{refine}}}(D, G) = E[\log D(h_i, c_i)] + E[\log(1 - D(G(r_{ij}, c_j), c_j))] + E[\log(1 - D(h_i, c_j))]. \quad (6)$$

4. 実験

4.1 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ファッション EC サイト Zalando (<https://www.zalando.de>) の Website から取得したものである。また、画像のサイズは 256×192 とし、1 枚に複数の服が写っているようなノイズとなる画像は取り除いた。用意したデータは合計で 9286 組であり、9000 組を学習に、286 組をテストに利用した。実装は DeepLearning のフレームワークの Pytorch を利用し、最適化手法には Adam を用いた。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ており、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。比較として、従来手法の CAGAN[Jetchev 17], SwapGAN[Kubo 18] も実装し比較を行ったが、公平性のため、各手法とも論文内の 128×96 のサイズの画像ではなく、本提案手法に合わせて 256×192 のサイズ

の画像で学習を行った。また、セグメンテーションの生成には [Gong 17] のデータセットで学習した [Chen 15] を利用し、また、[Güler 18] によって得られる IUV データを利用する。

4.2 生成結果

図 4, 図 5 は提案手法の生成画像を示したものである。図 4 は特に特徴の模様のある服の着せ替えを行った例である。図 5 は着用している服と着せ替える服の丈の長さの違う場合の着せ替え及び学習データに少ない正面向きの姿勢ではない人物の着せ替えの例である。



図 4: 提案手法の生成画像 1
Rough Image はテクスチャマッピング後の生成結果、Final Image は精錬ネットワークの生成結果である。

4.3 既存手法との比較

図 6 は GAN ベースの既存手法である CAGAN と SwapGAN との比較を示している。既存手法である CAGAN 及び SwapGAN は直立正面の姿勢の人物の着せ替えであれば着せ替えを行なうことができるが、図の人物のような他の姿勢の画像に対してはうまく着せ替えを行なうことが難しい。これは、学習データに多く存在する直立正面以外の姿勢までネットワークの学習が及んでいないためであると考えられる。一方、提案手法では身体構造をネットワーク内で学習しようとするのではなく、Densepose によって得られる IUV データを使ったマッピング (パーツの貼り付け) に担わせることにより、直立正面の特定の姿勢だけではなく様々な姿勢に対応できるようになったと考えられる。

5. まとめ

本研究では、ファッション分野におけるオンラインショッピングサイトの需要増加によって利用が期待される仮想試着に関して 3 次元表面を利用した着せ替え手法の提案を行った。既存の GAN を用いた自動着せ替え手法である CAGAN, SwapGAN

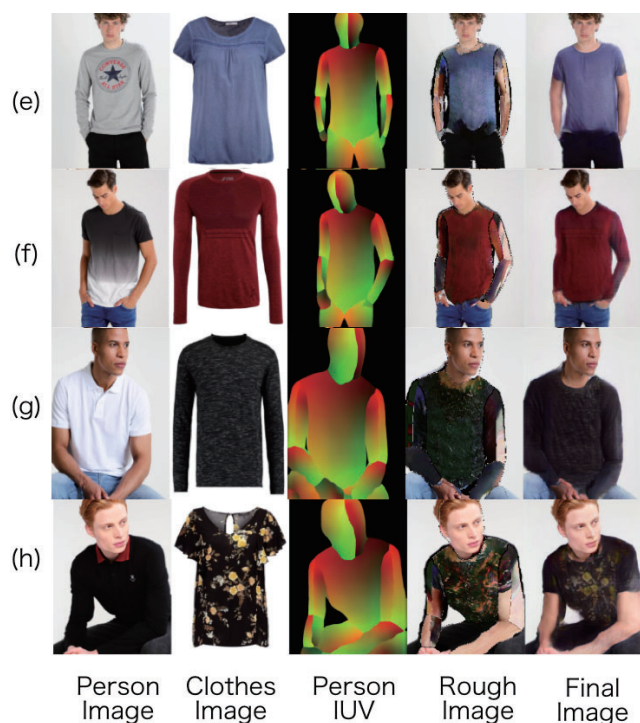


図 5: 提案手法の生成画像 2

Rough Image はテクスチャマッピング後の生成結果, Final Image は精錬ネットワークの生成結果である. (e), (f) はそれぞれ長袖から半袖, 半袖から長袖の服に着せ替えた例である. (g), (h) は正面向きでない姿勢の画像を着せ替えた例である.

との比較により 3 次元構造を利用した着せ替え手法の有効性を示した. 今後も引き続き, 改善に取り組む予定である.

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)
- [Jun-Yan 17] hu, Jun-Yan and Park, Taesung and Isola, Phillip and Efros, Alexei A.: Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks, in International Conference on Computer Vision (ICCV) (2017)
- [Kubo 18] Shizuma, Kubo and Yusuke, Iwasawa and Masahiro, Suzuki and Yutaka, Matsuo: SwapGAN: Cloth-Region Aware Generative Adversarial Networks toward Virtual Try-On System (2018)
- [Han 17] Han, Xintong and Wu, Zuxuan and Wu, Zhe and Yu, Ruichi and Davis, Larry S.: VITON: An Image-



図 6: 既存手法との比較

based Virtual Try-on Network (2017), in the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)

- [Wang 18] Wang, Bochao and Zheng, Huabin and Liang, Xiaodan and Chen, Yimin and Lin, Liang and Yang, Meng.: Toward characteristic-preserving image-based virtual try-on network, in European Conference on Computer Vision (ECCV) (2018)
- [Güler 18] Güler, Rıza Alp and Neverova, Natalia and Kokkinos, Iasonas: DensePose: Dense Human Pose Estimation In The Wild, in the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
- [Neverova 18] Neverova, Natalia and Alp Güler, Rıza and Kokkinos, Iasonas: Dense Pose Transfer, in European Conference on Computer Vision (ECCV) (2018)
- [Wu 18] Wu, Zhonghua and Lin, Guosheng and Tao, Qingyi and Cai, Jianfei: M2E-Try On Net: Fashion from Model to Everyone (2018)
- [Simonyan 15] Simonyan, Karen and Zisserman, Andrew: Very Deep Convolutional Networks for Large-Scale Image Recognition, in International Conference for Learning Representations (ICLR) (2015)
- [Johnson 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)
- [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)
- [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)