トレンド検出に基づくファッション画像生成

Trend Analysis-Based Fashion Image Generation

益川 良藏 [*1] Ryozo Masukawa
青葉 紗矢香 [*1] Sayaka Aoba
佐藤 勇元 [*1] Yugen Sato
櫨 翔佑 [*1] Shosuke Haji
佐藤 真 [*1] Makoto Sato
松井 太我 [*2] Taiga Matsui
石川 桂太 [*2] Keita Ishikawa
高木 友博 [*1] Tomohiro Takagi

[*1] 明治大学理工学部情報科学科
Meiji University, School of Science and Technology, Department of Computer Science

[*2] 株式会社エアークローゼット
airCloset, Inc.

One of the most essential tasks in fashion retail business is to predict the inventory by analyzing the future fashion trend. Generally speaking, fashion designers design a novel clothing by converting iconic designs shown in catwalk show into a general design that consumers use in their daily lives. As a result, fashion experts have a duty to detect novel fashion trend by aggregating external information sources such as fashion web magazines. However, external fashion information sources are too diverse for experts to analyze all of them objectively, constantly, and equally. To deal with this problem, we propose a system that generates images of novel clothing designs that matches existing fashion trends by combining generative models and information retrieval models. Experiments show that our proposed system can generate high resolution clothing images that reflects iconic fashion trend information.

1. はじめに

敵対的生成ネットワークや変分オートエンコーダー, 拡散モデルを中心として, 近年は画像生成に関する研究が盛んである. その応用例の一つとして, ファッション分野における画像生成に関する研究が行われており, 仮想試着やデザイン創出など, 多岐に渡る分野で発展している.

ファッション販売業界においては, 商品の発注をする際, 未来の流行を予測することが不可欠であり, そのデザインはファッションショー等で示される衣服の情報から, 一般の消費者が実生活で着用するものへトップダウン的に変換される.

本研究では, 上記のような衣服の特徴を抽出し, 小売店で販売されるような一般的な衣服のデザインへ変換するという, 衣服のデザイン創出を行う上で必須となる業務を画像生成モデルで代用するシステムを提案する. ファッションの専門家による定量的, 定性的な評価により, 提案システムの有効性を示した.

2. 関連研究

2.1 敵対的生成ネットワークのファッションへの応用

敵対的生成ネットワーク (Generative Adversarial Networks:GAN)[1] 技術の進歩は現実的な画像生成を可能にした. 特に, PGGAN[2] を改良して潜在変数の各要素を独立して扱うことで, 高解像度な画像生成が可能な StyleGAN[3] をファッション分野へ応用する研究は, 例えば仮想試着 [4] やコーディネート生成 [5] など多様な分野で行われている. 本研究においても, 高解像度ファッションデザイン生成のために, StyleGAN[3] の発展形である StyleGAN2-ADA[6] をベースとした.

連絡先: 益川 良藏, 明治大学理工学部情報科学科, 〒 214-8571 神奈川県川崎市多摩区東三田 1-1-1, ryozo0209@cs.meiji.ac.jp

2.2 画像処理と時系列予測を利用したファッショントレンド予測

ファッショントレンドを予測する研究としては, 画像情報とテキスト情報を読み込んでファッションの時系列予測を行う Grauman[7] らのシステムや, Ma[8] らのファッション要素の階層的な構造を捉え, LSTM を用いて時系列予測を行うシステム等が存在する. しかし, これらは時系列予測に主眼を置いており, 実際のファッション専門家のように, 多様な情報源を集約して最新のトレンド情報を把握するものではない. 発注実務における予測では, 時系列のトレンド要因よりも, ファッション誌等の外的要因による影響の方が大きい. そのため, 本研究における提案手法は, ファッション専門家がトレンドを検出する手順を再現することを目的として設計した.

2.3 CLIP

Radford らにより提案された CLIP[9] は, 与えられた画像に対して最適なテキストを推測する目的で画像情報とテキスト情報をそれぞれをベクトルへ変換して, そのコサイン類似度が最大になるように学習するため, 画像情報と言語情報を同一のベクトル空間上で結びつけることが可能である. CLIP は 1 億枚以上の多様な画像から訓練されており, ゼロショットであっても高い精度で画像分類を行うことができる. 本研究で用いたデータセットは衣服の部分画像のみで構成された, 特徴が均一なデータセットであり, 実際のファッションショー等で見られる周りの状況も含めて撮影された多様な画像とはドメインが異なる. これらの異なるドメインに属するデータを同一に扱うため, 我々は膨大な数の画像で事前学習済みである CLIP の画像エンコーダーを利用した.

3. 手法

提案システムは大きく分けて二つのモジュールからなる.

一つ目のモジュールは, CLIP を用いてファッション誌に登場する象徴的な衣服の画像の傾向を, 一般的な商品の特徴へ変換す

る機能を持つ CLIP 変換モジュール (**3.2** に後述) である.

一方, 二つ目のモジュールは CLIP 変換モジュールの出力を基に, 衣服の傾向を示すタグ表現から画像生成を行う敵対的生成ネットワークベースの画像生成モジュールである (**3.3** に後述).

3.1 データセット

提案手法の学習には, エアークローゼットのデータの一部である約 3 万枚の衣服の商品画像からなるデータセットを用いた. このデータセットは, 衣服毎に一意の ID と衣服の画像, そのカテゴリ (7 種類), シルエット (10 種類), 袖丈 (12 種類), 色 (16 種類), 柄 (20 種類) のタグ情報, そして, 衣服画像の特徴を実際のスタイリストが記述した言語情報によって構成されている. 更に, これらに加えて, 素材や質感等の詳細なタグが付与されており, その数は合計約 200 種類程である.

3.2 CLIP 変換モジュール

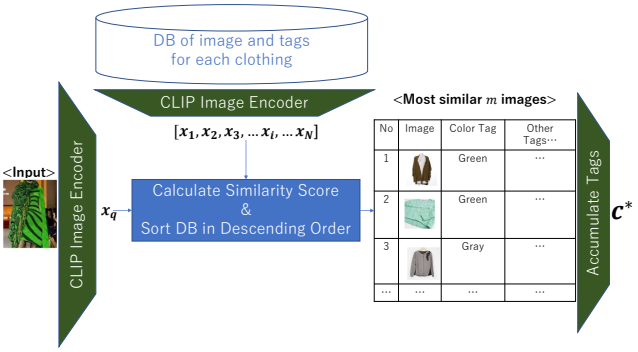

図 1　CLIP 変換モジュールの構造

我々はまず, 一般的な衣服の画像情報と, その特徴を示している言語情報を同一のベクトル空間へ変換するモジュールを CLIP を用いて開発した. その構造を図 1 に示す.

まず, エアークローゼットのデータセット内の衣服の画像情報と対応するコメントのテキストデータによって CLIP のファインチューニングを行った. 次に, この学習済みの CLIP を用い, 式 (1) のようにファッション Web マガジン等から得られた象徴的なデザインの衣服の画像 I_q を CLIP の画像エンコーダー E によって検索ベクトル $\boldsymbol{x}_q$ へ変換する.

$$\boldsymbol{x}_q = E(I_q) \tag{1}$$

同様に, N 枚のエアークローゼットデータセット内の一般的な衣服の商品画像の集合 $D : \{I_i | i \in [1, N]\}$ を式 (2) のようにベクトルへ変換し, 集合 $X : \{\boldsymbol{x}_i | i \in [1, N]\}$ を得る.

$$\boldsymbol{x}_i = E(I_i)(I_i \in D, i \in [1, N]) \tag{2}$$

そして, 検索対象の画像ベクトル $\boldsymbol{x}_q$ と集合 X 内の全ての画像ベクトル $\boldsymbol{x}_i \in X$ との類似度 s_i を, 式 (3) の内積によって算出し, 類似度の集合 $S : \{s_i | i \in [1, N]\}$ を得る.

$$s_i = \boldsymbol{x}_q \cdot \boldsymbol{x}_i \tag{3}$$

この集合 S 中, 類似度上位 $m (0 < m < N)$ 枚の衣服画像のタグ情報 (カテゴリ, シルエット, 袖丈, 色, 柄) の出現頻度を計算し, $\boldsymbol{x}_q$ の類似画像に現れる最頻出のタグの組 $\boldsymbol{c}_q$ を式 (4) のように得る.

$$\boldsymbol{c}_q = [c_{\text{category}}, c_{\text{silhouette}}, c_{\text{length}}, c_{\text{color}}, c_{\text{pattern}}] \tag{4}$$

なお, $c_{\text{category}}, c_{\text{silhouette}}, c_{\text{length}}, c_{\text{color}}, c_{\text{pattern}}$ はそれぞれ, 最頻出のカテゴリ, シルエット, 袖丈, 色, 柄のタグ情報である.

我々は式 (1)〜(4) の手順をファッションウェブマガジン等から集約した検索対象の M 枚の画像の集合 $Q : \{I_q : q \in [1, M]\}$ 毎に繰り返し, 各画像の尤もらしい特徴として得られたタグの組の集合 $C : \{\boldsymbol{c}_q : q \in [1, M]\}$ を得る. そして, C 内の全てのタグの組み合わせについて, 再度カテゴリ, シルエット, 袖丈, 色, 柄の最頻出のタグの組み合わせ $\boldsymbol{c}^*$ を式 (5) のように得る.

$$\boldsymbol{c}^* = [c^*_{\text{category}}, c^*_{\text{silhouette}}, c^*_{\text{length}}, c^*_{\text{color}}, c^*_{\text{pattern}}] \tag{5}$$

この $\boldsymbol{c}^*$ は, 全ての外部情報を集約して得られた衣服のタグ情報の組み合わせとなり, トレンド情報を示している.

3.3 画像生成モジュール

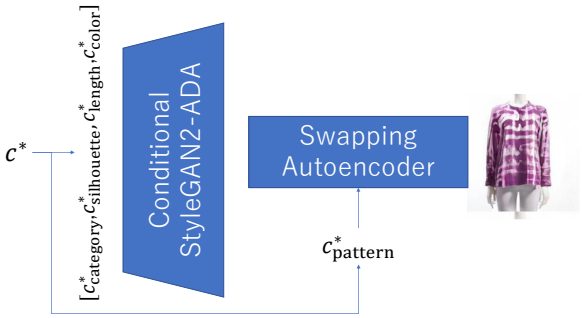

図 2　画像生成モジュールの構造

本モジュールは, 式 (5) に示す CLIP 変換モジュールの出力によって得られた条件の組み合わせ $\boldsymbol{c}^*$ を基に, StyleGAN2-ADA[6] をベースとする, 図 2 に示す機構で画像を生成する.

StyleGAN2-ADA[6] の生成器は, 潜在空間 $\mathcal{Z}$ 内の潜在変数 $\boldsymbol{z}$ を, マッピングネットワーク f により中間潜在変数 $\boldsymbol{w} \in \mathcal{W}$ へ写像する ($f : \mathcal{Z} \to \mathcal{W}$). f には $\boldsymbol{z}$ を式 (6) のように条件ラベルの埋め込み表現 $\boldsymbol{y}$ を含めて写像するための機構がある.

$$\boldsymbol{w} = f(\boldsymbol{z}, \boldsymbol{y}), \quad \boldsymbol{y} = g(c) \tag{6}$$

ここで, g と c はそれぞれ, 埋め込み層と条件ラベルである. 本モジュールにおける StyleGAN2-ADA ではカテゴリ, シルエット, 袖丈, 色の 4 条件を反映する必要がある. そのため, 複数の条件ラベルを入力できるよう, マッピングネットワーク f 内部の埋め込み層を, 任意の数のの条件ラベル各々に用意できるよう変更した. 特に, 前述した 4 条件を組み合わせた条件ラベルを入力するため, 条件ラベル $\boldsymbol{c} = \{c_{\text{cate}}, c_{\text{sil}}, c_{\text{len}}, c_{\text{color}}\}$ と, それぞれの埋め込み層 $g_c(c \in \boldsymbol{c})$ を式 (7) のように定めた.

$$\boldsymbol{w} = f(\boldsymbol{z}, \boldsymbol{y}), \boldsymbol{y} = [g_{\text{cate}}(c_{\text{cate}}), g_{\text{sil}}(c_{\text{sil}}), g_{\text{len}}(c_{\text{len}}), g_{\text{color}}(c_{\text{color}})] \tag{7}$$

なお, $\{c_{\text{cate}}, c_{\text{sil}}, c_{\text{len}}, c_{\text{color}}\}$ はカテゴリ, シルエット, 袖丈, 色のそれぞれの条件ラベルである. そして, これらは $\boldsymbol{c}^*$ の $\{c^*_{\text{category}}, c^*_{\text{silhouette}}, c^*_{\text{length}}, c^*_{\text{color}}\}$ にそれぞれ対応している.

ここで, StyleGAN2-ADA の生成器 (Synthesis Network) を G_s とすると, 複数条件付き StyleGAN2-ADA の出力画像 I は, 式 (8) のように表せる.

$$I = G_s(\boldsymbol{w}) \tag{8}$$

なお, 実験により, 柄を指定した生成は条件付き StyleGAN2-ADA のみでは困難であると判明したため, Swapping Autoencoder[10] を用いて柄 c^*_{pattern} の適用を行った. 本生成モデルの機構や詳細な実験については,Masukawa[11] らの論文に従った.

4. 実験・評価

本システムの有効性を示すため, その構成要素である CLIP 変換モジュールと画像生成モジュールに対して以下の 3 通りの実験を行った.

4.1 実験 1: CLIP 変換モジュールの性能評価

CLIP 変換モジュールの出力として得られるタグ情報が, 人間の感性を反映しているかを検証するため, ファッションレンタルサービスを展開している, エアークローゼットのスタイリストと, 提案手法の有効性を定量的に評価した. 具体的には, ファッションウェブマガジンから得た 75 枚の象徴的な衣服を着用したファッションモデルの画像に対し, 提案システムのデータセットに付与されている 200 種類のタグの中から合致するものをスタイリストが選択することで, 本システムの学習に利用したデータセットと同様のタグ付けを実施し, これをテストデータとした.

定量評価は, CLIP 変換結果のタグと, 人間の選択したタグの一致度を Top-n Accuracy に類似した方式で行った. 具体的には, CLIP の変換結果から得られた上位 n タグの集合の中に含まれる単語と, スタイリストが選択したタグの集合の積集合をとり, この積集合が空集合でなければ正解, 空集合であれば不正解として, 全てのテストデータ中の正解率を計算した. これは, 提案手法の式 (4) で得られる変換結果が人間の専門家の判断と一致するかを評価している. この結果を表 1 に示す. 表 1 のように,Top-5 Accuracy は 0.75 を超えており, 実際のスタイリストの判断と CLIP 変換モジュールの出力結果が 7 割以上一致することを示している.

表 1 CLIP 変換モジュールの出力とスタイリストの判断の間の Top-n Accuracy

n	Top-n Accuracy
5	0.75
10	0.77
15	0.80
20	0.84

4.2 実験 2: 画像生成モジュールの性能評価

画像生成モジュールの定量評価には, 実画像の分布と生成画像の分布との距離を計測する, フレチェのインセプション距離 (FID) を用いた. 同時に, 入力条件を反映しているか定性的に評価を行った.

表 2 画像生成モジュールの FID スコア

	FID score
Conditional StyleGAN2-ADA	4.53

表 2 の FID と図 3 が示すように, 画像生成モジュールは入力された 5 条件 (カテゴリ, シルエット, 袖丈, 色, 柄) を忠実に反映した上で, 多様かつ高解像度の画像生成が可能である. 画像生成モジュールの性能評価の詳細については Masukawa[11] に従った.

図 3 画像生成モジュールの生成例

4.3 実験 3: システム全体の定性評価

ここでは象徴的な入力画像の特徴を CLIP 変換によって実際の商品の情報へと変換し, それを入力条件として画像生成を行った結果によって定性的な評価を行う. 本システム全体の流れによって出力された画像の例を図 4 に示す.

図 4 右上のワンピースとスカートの中間のようなデザインの衣服は, 黄色いロングフレアスカートという形に変換され, 画像生成が行われた. また, 図 4 右下のファッションモデルが着用しているボタンが特徴的な白いオーバーサイズの衣服は, 白くルーズなシルエットのシャツとして一般的な商品画像へ変換された.

このように, CLIP 変換モジュールが入力画像の象徴的な衣服の色やシルエット, カテゴリ等の特徴を精確に捉えており, 画像生成モジュールが変換結果を反映した商品画像を生成できていることが分かる. 以上から, 提案システムは, 外部情報を集約して現状のトレンドと合致する正確な画像生成を行っていることがわかる.

5. おわりに

本論文では, 外部情報からファッショントレンドを検出し, そのトレンドと合致する衣服の画像生成を行うシステムを提案した. 実験により, 提案システムの有効性が示された.

一方で, CLIP 変換モジュールに関しては, 例えばネイビーの服のみで構成されているデータセットである場合, データセット内の類似画像の情報によって大きくバイアスがかけられてしまい, 変換結果ではネイビーの服が最頻出のタグであると判定されてしまう不具合が生じる可能性がある. ゆえに, 提案システムを有効活用するには, CLIP 変換モジュールに利用する衣服画像のデータベースの属性情報が均一かつ多様であることが不可欠であると考えられる.

また, 提案システムでは学習に利用したデータセットのタグ情報のうち, 実際に生成に用いているのはカテゴリ, シルエット, 袖丈, 色, 柄といった衣服における基本的な属性情報である. しかし, 素材や質感などの情報も衣服のデザイン創出において非常に重要な要素である. Stable Diffusion[12] 等により, このような詳細な情報を反映した衣服画像の生成を行うことが将来的には考えられる.

参考文献

[1] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes,

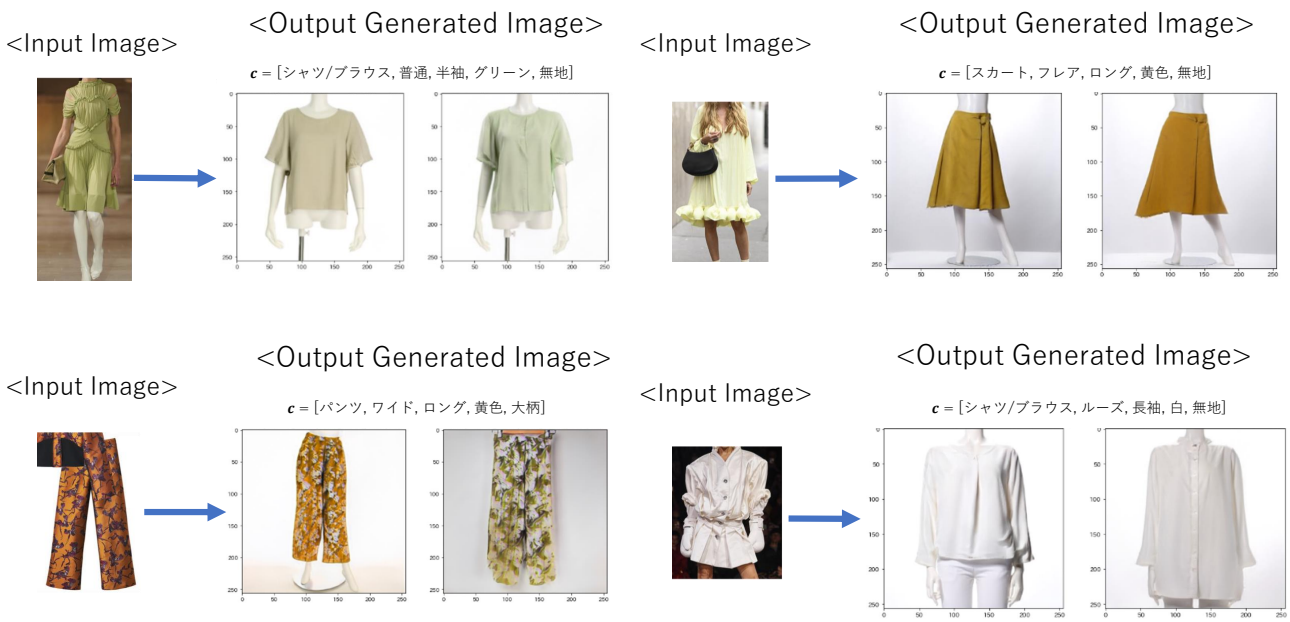

図 4　提案システムによる商品画像の生成例

N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

[2] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations (ICLR)*, 2018.

[3] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[4] Kathleen M. Lewis, Srivatsan Varadharajan, and Ira Kemelmacher-Shlizerman. VOGUE: try-on by StyleGAN: interpolation optimization. *CoRR*, abs/2101.02285, 2021.

[5] Gökhan Yildirim, Nikolay Jetchev, Roland Vollgraf, and Urs Bergmann. Generating high-resolution fashion model images wearing custom outfits. *CoRR*, abs/1908.08847, 2019.

[6] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12104–12114. Curran Associates, Inc., 2020.

[7] Wei-Lin Hsiao and Kristen Grauman. From culture to clothing: Discovering the world events behind a century of fashion images, 2021.

[8] Yunshan Ma, Yujuan Ding, Xun Yang, Lizi Liao, Wai Keung Wong, and Tat-Seng Chua. Knowledge enhanced neural fashion trend forecasting, 2020.

[9] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[10] Taesung Park, Jun-Yan Zhu, Oliver Wang, Jingwan Lu, Eli Shechtman, Alexei Efros, and Richard Zhang. Swapping autoencoder for deep image manipulation. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 7198–7211. Curran Associates, Inc., 2020.

[11] Ryozo Masukawa, Shosuke Haji, Masane Fuchi, Kazuki Yamaji, Tomohiro Takagi, Taiga Matsui, and Keita Ishikawa. Gan-based detailed clothing generation system. In *The 17th International Conference on Knowledge, Information and Creativity Support Systems*, 2022.

[12] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022.