DEIM Forum 2022 C31-2

気象センサ情報を用いた屋外画像における人物の服装変換

岡田　　渓† 新田　直子† 中村　和晃†† 馬場口　登†

† 大阪大学大学院工学研究科　〒 565–0871 大阪府吹田市山田丘 2-1
†† 東京理科大学工学部　〒 125–8585 東京都葛飾区新宿 6-3-1
E-mail: †{okada,naoko,babaguchi}@nanase.comm.eng.osaka-u.ac.jp, ††nakamura.kazuaki@rs.tus.ac.jp

あらまし　近年発展を遂げている条件に応じた画像変換技術の適用により，ある時間に撮影された屋外画像を，同地点の異なる時間の気象センサの観測値に応じて変換し，疑似的に該当時間の気象状況に応じた空の様子などを表す屋外画像を生成できる．気象センサの観測値を条件とした画像変換器は一般に，同じ環境における気象センサの観測値と屋外画像の対から学習されるが，画像中の人物の服装などは気象状況に応じた視覚的変化が多様であり，その変化パターンを学習するのに十分なデータ対の収集は困難である．そこで本研究では，気象状況に対し，人物の服装の種類が概ね一様に変化することに着目し，まず，比較的収集が容易な服装の種類がラベル付けされた人物画像データセットを用いて，画像中の人物に対し，与えられた服装の種類を条件として服装を変換する画像変換器を学習する．さらに，人物画像に対する服装の種類の推定器を学習し，気象センサの観測値と対として収集した人物を含む屋外画像に対し，人物の服装の種類を推定し，気象センサの観測値と服装の種類の対に変換する．このデータセットを用いて，気象センサの観測値に対する服装の種類の推定器を学習し，屋外画像と気象センサの観測値が与えられたとき，まず気象状況に応じた服装の種類を推定した後，服装の種類に応じて画像を変換するといった 2 段階の処理を通して気象センサの観測値に応じた人物の服装変換を実現する．

キーワード　気象センサ, 画像変換, 服装変換

1 はじめに

近年，定点カメラの映像やソーシャルネットワーキングサービス (SNS) の投稿画像などが Web 上に公開されており，多様な場所の状況を視覚的に観測できる．しかし，カメラの設置コストやプライバシ面における問題，ユーザの投稿は自発的であることから観測できる時間，場所は限定的である．一方で，非視覚的ではあるが時空間的に比較的密に観測値を取得できるものとして気象センサや GPS などのセンサ値が挙げられる．例えば，WorldWeatherOnline [1] では，世界各地に設置された気象センサから数時間ごとに気象状況を観測して公開しているため，時空間的に密に観測値を取得できる．よって，ある場所の画像を取得した際に，その地点の異なる時間のセンサ値を基に画像変換を行うことにより，本来取得できない時間の視覚的情報である画像を疑似的に生成できると考えられる．

画像変換の技術として，学習のために大量に画像を用意し，本物の画像と区別がつかない偽物の画像を生成する敵対的生成ネットワーク (Generative Adversarial Network:GAN) [2] が注目されている．また，顔画像と性別や髪色のように画像とその画像に対応する条件の対を学習用データとして用いることにより，与えられた条件に応じて偽物の画像を生成する条件付き GAN [3,4] が存在する．この技術を適用し，ある時間に撮影された屋外画像を，同地点の異なる時間の気象センサの観測値に応じて変換し，疑似的に該当時間の気象状況に応じた空の様子などを表す屋外画像を生成する研究 [5,6] が行われている．しかし，空の様子以外にも気象状況と共に変化すると考えられる人物の服装などについては，まだ対応していない．

画像中の人物の服装を変換する研究としては，人物画像と共に服装画像を与え，人物の姿勢に応じて変換した服装画像を基に，与えられた服装の人物画像を生成する技術 [7] が主流である．この手法を気象状況に応じた服装変換に適用するためには，気象状況に応じた服装画像を入力として与える必要がある．一方で，条件付き GAN を用いた気象センサの観測値を条件とした画像変換器は，同じ環境における気象センサの観測値と屋外画像の対が学習データとして必要となるが，画像中の人物の服装は気象状況に応じた視覚的変化が多様であり，その変化パターンを学習するのに十分なデータ対の収集は困難である．

そこで本研究では，気象状況に対し，人物の服装の種類が概ね一様に変化することに着目し，まず，比較的収集が容易な服装の種類がラベル付けされた人物画像データセットを用いて，画像中の人物に対し，与えられた服装の種類を条件として服装を変換する画像変換器を学習する．さらに，人物画像に対して服装の種類を推定する推定器を学習することにより，気象センサの観測値の対として収集した人物を含む屋外画像に対し，人物の服装の種類を推定し，気象センサの観測値と服装の種類の対のデータセットに変換する．このデータセットを用いて，気象センサの観測値に対する服装の種類の推定器を学習することにより，屋外画像と気象センサの観測値が与えられたとき，まず気象状況に応じた服装の種類を推定した後，服装の種類に応じて画像を変換するといった 2 段階の処理を通して気象センサの観測値に応じた人物の服装変換を実現する．

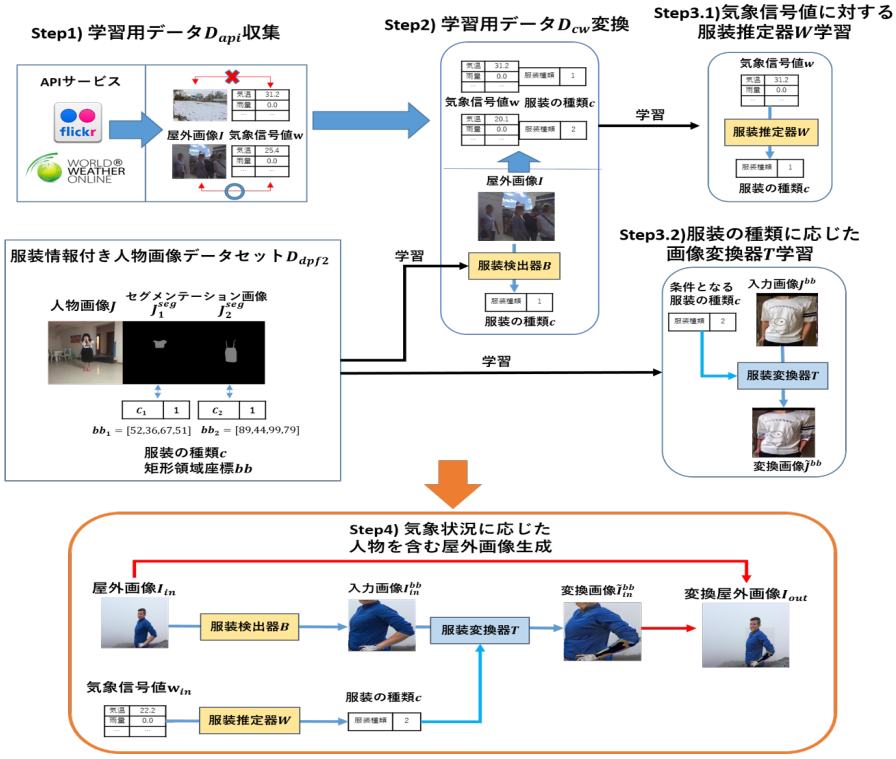

図 1　提案手法の概要

2　提 案 手 法

本研究では，任意の時間，場所における気象センサからの信号値 w_{in} が与えられた際に，同地点で撮影された人物を含む屋外画像 I_{in} に対し画像変換，特に服装部分の画像変換を行い，与えられた気象状況として適切な画像 I_{out} を生成することを目的とする．これを実現する画像変換器の学習には，同じ気象状況を観測した人物を含む屋外画像と気象センサの信号値の対が大量に必要となる．本研究では，SNS に世界各地の人々が投稿した人物を含む屋外画像が，その投稿時間，位置情報と共に公開されていることに着目し，SNS から投稿画像を収集し，投稿時間，位置と最も近い時間，位置の気象信号値をその投稿画像の対として収集する．

しかし，気象状況に応じた人物の服装の変化は多様であり，そのパターンを学習するために十分な人物を含む屋外画像の収集は困難であることから，比較的収集が容易な服装の種類がラベル付けされた人物画像データセットを利用する．まず，服装の種類に応じて画像を変換する画像変換器の学習を行う．また，人物画像からその人物の服装の種類を推定する服装推定器を学習し，収集した気象信号値の対となる人物画像に対し服装の種類を推定することにより，気象信号値と服装の種類の対に変換する．そして作成したデータ対を用いて気象信号値から服装の種類を推定する推定器を学習することにより，任意の時間場所の気象信号値 w_{in} と同地点の人物を含む屋外画像 I_{in} が与えられた際にまず，気象信号値 w_{in} から服装の種類を推定した後，推定された服装の種類を条件として画像 I_{out} に変換するという 2 段階の処理を通して，気象状況に応じた人物画像の服装変換を実現する．ただし，画像中には人物が複数含まれることがあり，画像全体の服装変換の学習は困難であるため，服装の矩形領域における画像変換器を学習し，画像生成時には，各矩形領域ごとに画像変換し，元画像に埋め込むことにより画像全体を変換する．

提案手法は，図 1 に示すように，以下の 5 ステップからなる．

Step 1) 学習用データ収集

SNS から人物を含む屋外画像 I_n を収集し，投稿時間，位置情報と最も近い気象信号値 w_n を対とし，データセット $\mathcal{D}_{api} = \{(I_n, w_n)|n = 1, ..\}$ を作成する．

Step 2) 学習用データ変換

人物画像 J_m に対し画像中に写る a_m 個の服装の種類 $c_m = \{c_{m0}, c_{m1}, ..c_{ma_m}\}$ や矩形領域 $bb_m = \{bb_{m0}, bb_{m1}, .., bb_{ma_m}\}$，セグメンテーション画像 $J_m^{seg} = \{J_{m0}^{seg}, J_{m1}^{seg}, .., J_{ma_m}^{seg}\}$ のメタデータが付与された服装データセット $\mathcal{D}_{dpf2} = \{(J_m, c_m, bb_m, J_m^{seg})|m = 1, ..\}$ を用いて，$\{(c_{m0}, bb_{m0}), (c_{m1}, bb_{m1}), .., (c_{ma_m}, bb_{ma_m})\} = B(J_m)$ となる服装の種類と矩形領域推定器 B(以後検出器 B と呼ぶ) を学習する．ただし服装の種類は，t 種類のラベルからなり，矩形領域は，服装矩形の左上の座標 (x_0, y_0)，右下の座標 (x_1, y_1) からなる 4 次元の値 (x_0, y_0, x_1, y_1) で表される．その後，Step 1) で収集したデータセット $\mathcal{D}_{api}$ に含まれる画像 I_n 中の人物の服装の種類と矩形領域 $\{(c_{n0}, bb_{n0}), (c_{n1}, bb_{n1}), ..\} = B(I_n)$ を検出し，検出結果から画像 I_n に対応する服装の種類 c_n を決定することにより，データ対 $\mathcal{D}_{cw} = \{(c_n, w_n)|n = 1, ..\}$ に変換する．

Step 3.1) 気象信号値に対する服装推定器学習

Step 2) で作成したデータ対 $\mathcal{D}_{cw}$ の気象信号値 w_n と服装の種類 c_n の対を用いて $c_n = W(w_n)$ となる推定器 W を学習する．

Step 3.2) 服装の種類に応じた画像変換器学習

服装情報付きデータセット $\mathcal{D}_{dpf2}$ を基に，画像 J_m と矩形領域 bb_m から作成される矩形領域の画像 J_m^{bb} と矩形に対応する服装の種類 c_m の対から $\tilde{J}_m^{bb} = T(J_m^{bb}, c_m)$ となる画像変換器 T を学習する．

Step 4) 気象状況に応じた人物を含む屋外画像生成

ある場所で撮影された人物を含む屋外画像 I_{in} と，同地点の異なる時間における気象信号値 w_{in} を取得する．まず Step 3.1) で学習した推定器 W を用いて取得した気象信号値 w_{in} から服装の種類 $c = W(w_{in})$ を推定する．その後，Step 2) で学習した検出器 B を用いて画像 I_{in} から服装の矩形領域画像 I_{in}^{bb} を検出し，Step 3.2) で学習した変換器 T を用い，推定した服装の種類 c に応じて服装変換する．最後に，変換された服装画像 $\tilde{I}_{in}^{bb} = T(I_{in}^{bb}, c)$ を画像 I_{in} の矩形領域に埋め込むことにより，気象状況 w_{in} に応じた人物を含む屋外画像 I_{out} を生成する．

以降の節で各ステップの詳細について述べる．

2.1 学習用データ収集

まず SNS に投稿された画像の中から，気象状況を観測できる屋外で撮影された画像を収集する．屋外画像の判定方法としては Zhou ら [8] の提案する 365 種類のシーンラベル付きデータセットを基に学習されたシーン推定器を用いる．365 種類のシーンにはそれぞれ屋内ラベル 0，屋外ラベル 1 が振り分けられているため，投稿画像 I に対する k 位の推定結果のシーンを $Pred_{scene}(I)_k = s$，$s \in \{0, 1\}$ とした時，上位 K 位の推定シーンが $\frac{1}{K}\sum_{k=1}^{K} Pred_{scene}(I)_k \geqq Th_{outdoor}$ を満たす画像 I を屋外画像と判定し収集する．

次に，収集した屋外画像 I に対して対となる気象信号値 w を取得する．SNS には画像と共に投稿時間，位置情報が公開されているため，その投稿位置と最も近い気象センサにおいて投稿時間と最も近い時間に観測された気象信号値 w を取得する．ただし，投稿画像に付属する時間，位置情報はあくまで投稿された位置，時間情報であり，撮影された時間，位置であるとは限らないため，対となる屋外画像 I と気象信号値 w が同じ時間，位置の状況を観測しているか判定する必要がある．そこで，同じ時間，位置を観測する屋外画像と気象信号値の対は，同じ気象状況を示すという前提のもと，収集した対からそれぞれ天候状況を推定し，その推定結果を基に各対の整合性を判定する．

ここでは，晴れ，雨のような天候状況がラベル付けされた屋外画像データセットが存在すること，気象センサの観測値には，温度，湿度といった気象信号値と共に天候ラベルが含まれることに着目し，屋外画像と気象信号値からそれぞれ天候状況を推定する推定器を学習する．屋外画像 I と気象信号値 w に対する天候推定結果を $\hat{\boldsymbol{g}}_I$，$\hat{\boldsymbol{g}}_w$ とした時，整合性を示す指標として乖離度 L を $\hat{\boldsymbol{g}}_I$ と $\hat{\boldsymbol{g}}_w$ 及び気象センサの観測値内に含まれる天候ラベルをワンホットベクトル化した $\boldsymbol{g}_w$ のそれぞれの交差エントロピーの和で定義する．

$$L = -\lambda \hat{\boldsymbol{g}}_w \log \hat{\boldsymbol{g}}_I \quad - (1-\lambda)\boldsymbol{g}_w \log \hat{\boldsymbol{g}}_I \tag{1}$$

第一項では，屋外画像 I と気象信号値 w の推定結果の類似度を，第二項では，屋外画像 I と気象センサの観測値内の天候ラベル $\boldsymbol{g}_w$ との類似度を測っている．いずれも，推定された天候状況が類似している場合，乖離度 L は小さい値になることから，$L < Th_{sim}$ を満たす人物を含む屋外画像と気象信号値の対を整合性の取れたデータ対と判定する．

ここで，整合性判定によるデータ数の減少を防ぐため，整合性の取れた屋外画像と気象信号値の対を基に，屋外画像に対する気象信号値の推定器を学習することにより，整合性判定の結果，対が作成されなかった屋外画像に対し気象信号値を推定し疑似的な対を作成する．

最後に，画像中の物体を検出するモデルである YOLO(You Look Only Once) [9] を用いて，収集した屋外画像のうち，人物を含む屋外画像を選別する．収集した屋外画像 I に対して人物検出を行い，画像中に人物が写っている信頼度 $Conf_{hum}(I)$ が $Conf_{hum}(I) > Th_{hum}$ を満たす屋外画像を人物を含む画像と判定し，人物を含む屋外画像と気象信号値の対 $\mathcal{D}_{api} = \{(I_n, w_n)|n = 1, ..\}$ として収集する．

2.2 学習用データの変換

2.1 節において収集したデータ対 $\mathcal{D}_{api}$ を服装の種類と気象信号値の対のデータセット $\mathcal{D}_{cw} = \{(c_n, w_n)|n = 1, ..\}$ に変換する．人物画像 J_m に対し画像に含まれる a_m 個の服装の種類 $c_m = \{c_{m0}, c_{m1}, ..c_{ma_m}\}$ や矩形領域 $bb_m = \{bb_{m0}, bb_{m1}, .., bb_{ma_m}\}$，セグメンテーション画像 $J_m^{seg} = \{J_{m0}^{seg}, J_{m1}^{seg}, .., J_{ma_m}^{seg}\}$ のメタデータが付与された既存のデータセット $\mathcal{D}_{dpf2} = \{(J_m, c_m, bb_m, J_m^{seg})|m = 1, ..\}$ が容易に入手できることから，データセット $\mathcal{D}_{api}$ 内の人物を含む屋外画像 I_n に対して服装の種類 c_n の推定を行い，服装の種類 c_n と気象信号値 w_n のデータ対に変換する．しかし，SNS から収集される画像 I は複数の人物を含む場合があり，同じ気象状況下であっても同じ種類の服装を着用するとは限らない．同じ気象信号値に対して各服装の種類を対応させ，複数の対を作成することも可能であるが，ここでは画像中で最も着用されている服装の種類 c が，気象信号値 w の表す気象状況において最も適切な服装であると考える．

そこで，$\mathcal{D}_{dpf2}$ 内の人物画像 J_m と，画像に対応する服装の種類 c_m と矩形領域 bb_m を用いて，画像中の服装の種類と矩形領域を検出する検出器 B を学習する．損失関数は，人物画像 J_m に対する検出結果を $B(J_m) = ((\hat{\boldsymbol{c}}_{m1}, \hat{bb}_{m1}), ..(\hat{\boldsymbol{c}}_{mi}, \hat{bb}_{mi}), ..)$，正解データを $((\boldsymbol{c}_{m1}, bb_{m1}), .., (\boldsymbol{c}_{mi}, bb_{mi}), ..)$ として

$$\mathcal{L}_B = \sum_m \sum_i -\boldsymbol{c}_{mi} \log \hat{\boldsymbol{c}}_{mi} + |bb_{mi} - \hat{bb}_{mi}| \tag{2}$$

と表される．ただし，$|x|$ は x の L1 ノルムを，$\boldsymbol{c}_{mi}$ はワンホットベクトルに変換された c_{mi} を示す．第一項は推定した服装の種類の推定誤差，第二項は推定した矩形領域の回帰誤差を表し，これらを最小化するよう，検出器 B を学習する．その後，2.1 節で収集したデータ対 $\mathcal{D}_{api}$ の人物画像 I_n に対し，検出器 B を用いて服装検出 $((c_{n1}, bb_{n1}), (c_{n2}, bb_{n2}), ...) = B(I_n)$ を行い，$(c_{n1}, c_{n2}, ...)$ の中で最も多い服装の種類 c_n を人物画像 I_n に対

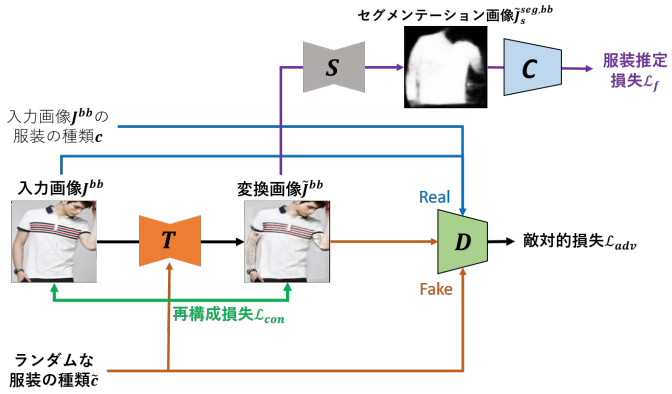

図 2　画像変換器ネットワーク構成

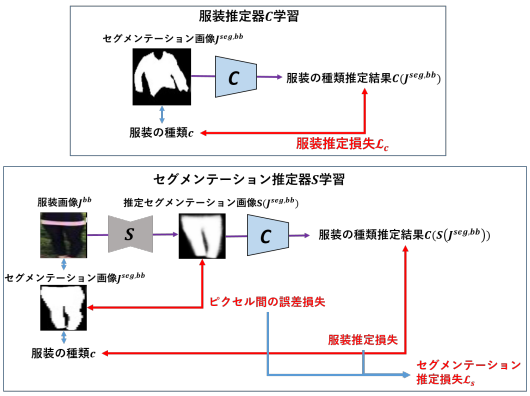

図 3　事前学習ネットワーク構成

する服装推定結果とする．以上の手順により 2.1 節で収集したデータ対 $\mathcal{D}_{api} = \{(I_n, w_n)|n = 1, ..\}$ を服装の種類 c_n と気象信号値 w_n の対の学習用データセット $\mathcal{D}_{cw} = \{(c_n, w_n)|n = 1, ..\}$ に変換する．

2.3　気象信号値に対する服装推定器学習

2.2 節において作成したデータ対 $\mathcal{D}_{cw} = \{(c_n, w_n)|n = 1, ..\}$ を用いて，気象信号値に対して服装の種類を推定する推定器 W を学習する．次式に示す，服装の種類 c_n をワンホットベクトルに変換した $\boldsymbol{c}_n$ と気象信号値 w_n に対する推定結果 $W(w_n)$ の交差エントロピーを損失関数とする．

$$\mathcal{L}_W = \sum_n -\boldsymbol{c}_n \log(W(w_n)) \tag{3}$$

2.4　服装の種類に応じた画像変換器学習

服装データセット $\mathcal{D}_{dpf2} = \{(J_m, c_m, bb_m, J_m^{seg})|m = 1, ..\}$ 内の人物画像 J_m と矩形領域 bb_m から服装の矩形領域画像 J_m^{bb} を作成し，$\tilde{J}_m^{bb} = T(J_m^{bb}, c_m)$ となる服装領域における画像の服装変換器 T を学習する．

属性ベクトルを条件とする画像変換器の学習では一般に，判別器に偽画像と見抜かれない画像を生成するための敵対的損失，入力画像と類似した内容の画像を生成するための再構成損失，また変換画像に対して属性を推定し，条件として与えた属性に応じた画像を生成するための属性推定損失からなる．本手法も同様の損失を用いるが，$\mathcal{D}_{dpf2}$ にセグメンテーション画像 J^{seg} が含まれることに着目し，属性推定損失を算出する際に使用する．セグメンテーション画像を通して属性推定を行うことにより，服装の形状が明確に変換されることを狙う．図 2 にネットワークの構成を示す．ネットワークは，服装の種類 $\tilde{c}$ を条件として入力画像 J^{bb} を変換する画像変換器 T，変換された偽画像 $\tilde{J}^{bb}$ と実画像 J^{bb} を判別する判別器 D，変換画像 $\tilde{J}^{bb}$ に対して，服装部分のセグメンテーションを推定する推定器 S と矩形領域のセグメンテーション画像 $\tilde{J}^{seg,bb}$ から服装の種類を推定する推定器 C の 4 つから構成される．

画像変換器 T の学習にあたり，まず事前に矩形領域のセグメンテーション画像に対する服装推定器 C とセグメンテーション推定器 S を学習する．それぞれのネットワークの構成を図 3 に示す．セグメンテーション画像に対する服装推定器 C は，畳み込みニューラルネットワークを用いて，セグメンテーション画像 J_{mi}^{seg} と矩形領域情報 bb_{mi} を基に作られる矩形領域のセグメンテーション画像 $J_{mi}^{seg,bb}$ に対応するワンホットベクトル化された服装の種類 $\boldsymbol{c}_{mi}$ と推定された服装の種類 $C(J_{mi}^{seg,bb})$ の間の交差エントロピーを損失とする．

$$\mathcal{L}_C = \sum_m \sum_i -\boldsymbol{c}_{mi} \log(C(J_{mi}^{seg,bb})) \tag{4}$$

セグメンテーション推定器 S は，完全畳み込みネットワーク (Fully Convolutional Network:FCN) を用い，入力となる服装の矩形領域画像 J_{mi}^{bb} に対するセグメンテーション画像 $J_{mi}^{seg,bb}$ と推定したセグメンテーション画像 $S(J_{mi}^{bb})$ の対応するピクセル毎の交差エントロピーを損失関数とする．更に，推定したセグメンテーション画像 $S(J_{mi}^{bb})$ に対し服装推定器 C を用いて服装推定を行い，推定された服装 $C(S(J_{mi}^{bb}))$ と入力画像 J_{mi}^{bb} に対応するワンホットベクトル化された服装の種類 $\boldsymbol{c}_{mi}$ との交差エントロピーを損失関数に加える．これにより，服装推定器 C が判別可能な鮮明なセグメンテーション画像が推定されることを狙う．以下に損失関数 L_S を示す．ただし，画像の各画素を p，全画素数を P とし，それぞれの損失関数のバランスを λ_s で取る．

$$\begin{aligned}\mathcal{L}_S = \sum_m \sum_i & -\frac{1}{P}\sum_p (J_{mip}^{seg,bb} \log(S(J_{mip}^{bb}))) \\ & -\lambda_s \boldsymbol{c}_{mi} \log(C(S(J_{mi}^{bb})))\end{aligned} \tag{5}$$

以上の学習済みのセグメンテーション推定器 S と服装推定器 C を用いて，画像変換器 T と判別器 D を学習する．

画像変換器 T は，入力画像と共に条件となる服装の種類を与えられるように対応させた FCN を用い，判別器 D を用いた敵対的損失 $\mathcal{L}_{adv}$ と，変換画像 $\tilde{J}^{bb}$ の再構成損失 $\mathcal{L}_{con}$，変換画像 $\tilde{J}^{bb}$ からセグメンテーション画像を推定した $\tilde{J}^{seg,bb}$ に対する服装の種類の推定損失 $\mathcal{L}_f$ の 3 つの要素からなる損失関数 $\mathcal{L}_T$ を最小化するよう学習する．損失関数 $\mathcal{L}_T$ は，各要素のバランスをとる係数である λ_{con}，λ_f を用いて以下のように表される．

$$
\begin{aligned}
\mathcal{L}_T &= \mathcal{L}_{adv} + \lambda_{con}\mathcal{L}_{con} + \lambda_f\mathcal{L}_f \\
\mathcal{L}_{adv} &= \mathbb{E}_{J^{bb}\sim p_J,\tilde{\boldsymbol{c}}\sim p_{\boldsymbol{c}}}\left[Relu(1-D(T(J^{bb},\tilde{\boldsymbol{c}}),\tilde{\boldsymbol{c}}))\right] \\
\mathcal{L}_{con} &= \mathbb{E}_{J^{bb}\sim p_J,\tilde{\boldsymbol{c}}\sim p_{\boldsymbol{c}}}\left[\frac{\frac{1}{P}|J^{bb}-T(J^{bb},\tilde{\boldsymbol{c}})|}{|\boldsymbol{c}-\tilde{\boldsymbol{c}}|}\right] \\
\mathcal{L}_f &= \mathbb{E}_{J^{bb}\sim p_J,\tilde{\boldsymbol{c}}\sim p_{\boldsymbol{c}}}\left[-\tilde{\boldsymbol{c}}\log(C(S(T(J^{bb},\tilde{\boldsymbol{c}}))))\right]
\end{aligned}
\tag{6}
$$

$\mathcal{L}_{adv}$ は，実画像と見分けがつかないように画像を変換するための制約となる．$\mathcal{L}_{con}$ の分子は再構成損失であり，画像の内容を大きく変化させないための制約であるが，分母に変換前後の条件変化を用いることにより，条件の変化が大きい場合に制約を弱める．$\mathcal{L}_f$ は $\tilde{\boldsymbol{c}}$ を条件として変換した画像 $\tilde{J}^{bb}$ に対してセグメンテーション画像を推定し，セグメンテーション画像 $\tilde{J}^{seg,bb}$ から推定される服装の種類と条件となる $\tilde{\boldsymbol{c}}$ との交差エントロピーであり，条件に応じて適切な変換が行われるための制約となる．$\mathcal{L}_T$ を最小化することにより，変換器 T は，入力画像を，条件となる服装に応じ，かつ実画像と見分けがつかないように変換するよう学習される．

また，同時に学習を行う判別器 D は，画像と共に服装の種類を入力として真偽を判別する畳み込みニューラルネットワークを用い，損失関数 $\mathcal{L}_{\mathcal{D}}$ は，

$$
\begin{aligned}
\mathcal{L}_D = &\mathbb{E}_{J^{bb}\sim p_J,\tilde{\boldsymbol{c}}\sim p_{\boldsymbol{c}}}\left[Relu(1-D(J^{bb},\boldsymbol{c}))\right] \\
&+\mathbb{E}_{J^{bb}\sim p_J,\tilde{\boldsymbol{c}}\sim p_{\boldsymbol{c}}}\left[Relu(1+D(T(J^{bb},\tilde{\boldsymbol{c}}),\tilde{\boldsymbol{c}}))\right]
\end{aligned}
\tag{7}
$$

とする．$\mathcal{L}_{\mathcal{D}}$ は敵対的損失であり，これを最小化することにより，判別器 D は，実画像 J^{bb} と変換前の服装の種類 $\boldsymbol{c}$ が入力された際は正の値を，変換された偽画像 $\tilde{J}^{bb}$ と変換条件として与えられた服装の種類 $\tilde{\boldsymbol{c}}$ が入力された際には負の値を返し，実画像と偽画像を判別するように学習される．

3 実　　験

本章では，まず実験に用いるデータセットについて説明し，これらを用いて学習した気象信号値に対する服装推定器 W と服装の種類に応じた画像変換器 T をそれぞれ評価した後，W と T を用いた気象に応じた画像変換の結果を評価する．

3.1 実験に用いるデータセット

提案手法は，人物画像を含む屋外画像と気象信号値の対のデータセット $\mathcal{D}_{api}$ を収集し，服装情報付き人物画像データセット $\mathcal{D}_{dpf2}$ を用いて，服装の種類と気象信号値の対のデータセット $\mathcal{D}_{cw}$ に変換し，$\mathcal{D}_{dpf2}$ と $\mathcal{D}_{cw}$ を学習に用いる．実験のため用意した上記の 3 つのデータセットについて述べる．

[**屋外画像と気象信号値の対のデータセット** $\mathcal{D}_{api}$]

Flickr [10] から 2016，2017 年にアメリカで投稿された画像を，投稿時間，位置情報と共に収集し，$K=10$，$Th_{outdoor}=1.0$ として Zhou らの手法 [8] を用いて屋外画像判定を行い，495,669 枚の屋外画像を収集した．その後，WorldWeatherOnline [1] からアメリカ各地点における 2016，2017 年の気象センサの観測値を収集し，各屋外画像 I に対して，観測位置の誤差が 10km 以内かつ観測時間の誤差が 1 時間以内の最も近い気象信号値 w を対として収集した．

対となる屋外画像 I と気象信号値 w の整合性を測るため，屋外画像から晴れ，曇り，霧，雨，雪の 5 種類の天候状況を推定する天候推定器と，気象信号値から同様に 5 種類の天候状況を推定する推定器を学習した．屋外画像に対する天候推定器は，約 18 万枚の屋外画像に対し，人手で天候ラベルが付けられたデータセットである Image2weather [11] が取得可能であることから，晴れ，曇り，霧，雨，雪の画像それぞれ 3,000 枚，3,000 枚，350 枚，1,250 枚，1,250 枚を学習用とテスト用に 4:1 の比率で分け，事前学習済みの VGG16 [12] を転移学習した．また気象信号値に対する天候推定器は，WorldWeatherOnline [1] に信号値と共に天候状況も含まれていることから各天候 1,250,000 個ずつ合計 6,250,000 個の観測値を収集し学習用とテスト用に 4:1 に分けて使用した．そして，3 層からなるパーセプトロンを用い，データ内に含まれる信号値の中で天候と関連すると考えられる温度，UV インデックス，視程，風速，雲量，湿度，気圧，露点の 8 つの気象信号値に対する天候推定器を学習した．学習の結果，画像に対する天候推定精度は 81.3%，気象信号値に対する天候推定精度は 90.5%となり概ね天候が推定できることを確認した．

これらの天候推定器を用いて，収集した屋外画像と気象信号値の対に対して天候推定を行い，$\lambda=0.1$ として乖離度 L を算出し，$Th_{sim}=0.5$ を満たす 126,107 対のデータ対を整合性が高いと判定した．ここで，整合性の高い対 126,107 対を基に，屋外画像から上記の 8 種の気象信号値を推定する推定器を VGG16 [12] を用いて学習した．そして，369,562 個の整合性が低いと判定された対の屋外画像に対し気象信号値を推定し，疑似的な気象信号値対を作成した．最後に，YOLO [9] を用いて，$Th_{hum}=0$ を満たす画像 24,571 対を人物の写る屋外画像であると判定し，人物を含む屋外画像と気象信号値の対のデータセット $\mathcal{D}_{api}=\{(I_n,w_n)|n=1,..\}$ とした．

[**服装情報付きデータセット** $\mathcal{D}_{dpf2}$]

人物画像 J に対して服装の種類 c や矩形領域 bb，服装領域のセグメンテーション画像 J^{seg} からなるメタデータが付与されているデータセット DeepFahion2 [13] が容易に入手できる．DeepFashion2 には，224,114 枚の人物画像 J に対してメタデータが付与されており，服装の種類は，short sleeveed shirt，long sleeved shirt，vest や trousers など 13 種類からなる．

実験では画像変換対象を上衣に絞り，気象に応じて画像中に変化が見られると考えられる short sleeved shirt と long sleeved shirt の 2 種の服装を半袖（short），長袖（long）とする．検出器 B の学習の際，検出クラスを 2 種類に限定すると布の量などから安易に半袖長袖を判別されると考えられることから，trousers，shorts をまとめて bottom とした．さらに，袖部分が見切れていたり，体勢により半袖長袖が判別し辛い画像が含まれることを考慮し，short，long，bottom の 153,253 枚の服装画像 J^{bb} のうち，10,000 枚を目視で確認し，判別の困難な上衣の画像は unused_top，下衣の画像は unused_bottom とした．

目視で確認していない 143,253 枚の服装領域画像 J^{bb} に対し

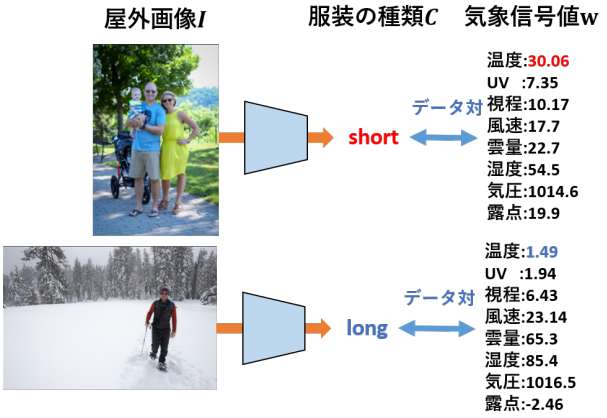

図 4　データセット $\mathcal{D}_{cw}$ データ例

て同様に 5 種類のラベルを付与するため，ResNet18 [14] を基に，服装領域画像を short，long，bottom，unused_ bottom，unused_top の 5 クラスに分類する推定器を学習した．その後，学習した推定器を用いて残りの 143,253 枚の服装領域画像 J^{bb} を分類した．その結果，服装の種類 short，long，bottom，unused_bottom，unused_top のラベルが付与された服装領域画像がそれぞれ 29,068 枚，12,788 枚，55,943 枚，32,897 枚，22,557 枚得られた．以上のように本研究の目的に合わせてラベルを付与した服装領域を含む画像を抽出し，111,317 枚の人物画像のデータセット $\mathcal{D}_{dpf2} = \{(J_m, c_m, bb_m, J_m^{seg})|m = 1, ..\}$ が得られた．

[**服装の種類と気象信号値の対のデータセット** $\mathcal{D}_{cw}$]

作成した 111,317 枚の服装情報付き人物画像データセット $\mathcal{D}_{dpf2} = \{(J_m, c_m, bb_m, J_m^{seg}|m = 1, ..\}$ を学習データとして 95,155 枚，テストデータとして 16,162 枚に分割し，resnet50 [14] をバックボーンモデルとした fasterrcnn [15] を基に，画像中の服装矩形領域と服装の種類を検出する検出器 B を学習した．検出する服装は c_m に含まれる short，long，bottom，unused_bottom，unused_top の 5 種類とした．short または long に限定し，正解となる矩形領域と検出領域が 50% 以上重なり，かつ服装の種類が正しく推定されている検出領域を正解とみなし，検出器 B を評価した結果，適合率は 91.7%，再現率は 94.1%となった．評価により半袖と長袖の服装領域は概ね検出できていることが確認できた．

学習した服装検出器 B を用いて，収集したデータ対 $\mathcal{D}_{api} = \{(I_n, w_n)|n = 1, ..\}$ の屋外画像 I_n から服装 c_n を推定した．その際，short，long のいずれかに判定された領域が検出されなかった画像は学習用データから除き，15,060 対の服装の種類と気象信号値の対のデータセット $\mathcal{D}_{cw} = \{(c_n, w_n)|n = 1, ..\}$ を得た．作成したデータ対の例を図 4 に示す．低温の気象信号値には long ラベル，高温の気象信号値には short ラベルが付与された画像が対となっており，気象信号値に応じた画像が収集されていることが確認できた．

3.2　気象信号値に対する服装推定器の評価

作成した服装の種類と気象信号値の 15,060 対からなる $\mathcal{D}_{cw} = \{(c_n, w_n)|n = 1, ..\}$ を学習用に 9,060 対，テスト用に 6,000 対に分割し，気象信号値から服装の種類を推定する推定器 W を学習した．3 層からなるニューラルネットワークを使用し，服装の変化と関連すると考えられる温度，UV インデックス，雲量，湿度，露点の 5 種の気象信号値から short，long の 2 種の服装を推定するよう学習した．

表 1　服装推定器 W 推定結果

		推定ラベル		
		short	long	再現率
正解ラベル	short	2364	636	0.788
	long	1911	1089	0.363
適合率		0.553	0.631	0.576

表 2　服装推定器 W 尤度 0.65 以上の推定結果

		推定ラベル		
		short	long	再現率
正解ラベル	short	139	109	0.560
	long	93	345	0.788
適合率		0.601	0.760	0.706

推定器 W は表 1 に示すように 57.6%と低精度であった．一般的に人物の服装は暑い時には半袖を着用し，寒い時には長袖を着用するが，その基準は人それぞれであることから中途半端な気象状況であるほど人物の服装は一意に定まらない．そのため，服装推定尤度は基本的に 0.5 付近の値が出力され，極端な気象状況であれば高い尤度を出力すると考えられる．そこで，推定器 W の推定尤度が 0.65 を越えている推定結果を表 2 に示す．高い尤度の推定における推定精度は全体的に高くなっており，推定尤度が高くなるような極端な気象状況であれば気象に応じた服装の分類が可能であることが確認できた．

3.3　服装の種類に応じた画像変換器の評価

本節では，作成した服装情報付きデータセット $\mathcal{D}_{dpf2} = \{(J_m, c_m, bb_m, J_m^{seg}|m = 1, ..\}$ を用いて服装の矩形領域における画像変換器 T を学習し，変換画像を評価する．事前に学習が必要な服装推定器 C，セグメンテーション推定器 S の評価を行った後，画像変換器 T を評価する．

3.3.1　ネットワーク事前学習

服装推定器 C は，服装データセット $\mathcal{D}_{dpf2}$ 内に含まれる矩形領域 bb_{mi} とセグメンテーション画像 J_{mi}^{seg} から作成できる矩形領域のセグメンテーション画像 $J_{mi}^{seg,bb}$ と服装の種類 c_{mi} を用いて入力を白黒画像に対応させた resnet18 [14] を基に学習した．画像変換器の学習の入力には short，long と判定された画像のみが与えられるため，short，long とそれ以外のラベルを判定できればよいものとし，推定する服装の種類は，short，long，bottom の 3 種類とした．3 種の服装がラベル付けされたデータ対 97,799 対を学習用に 83,407 対，テスト用を 14,392 対に分割して使用した．推定精度は 98.5%となり，高精度の推定が学習されていることを確認した．

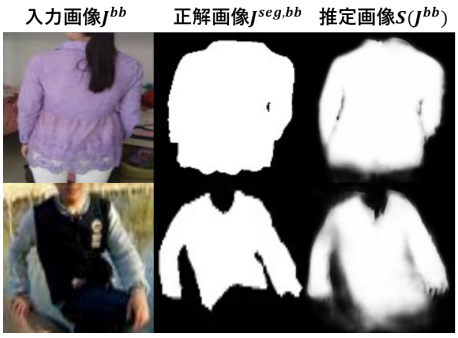

図 5　セグメンテーション推定器 S 推定結果

次に，画像 J_m と矩形領域 bb_{mi} から作成できる矩形領域画像 J_{mi}^{bb} と矩形領域におけるセグメンテーション画像 $J_{mi}^{seg,bb}$，服装の種類 c_{mi} のデータ対を用いて，U-net [16] を基にセグメンテーション推定器 S を学習した．その際，推定器 S から推定されるセグメンテーション画像が服装推定器 C が判別可能な自然な画像になることを狙い，学習済みの服装推定器 C を損失計算に使用した．推定器 S により推定されたセグメンテーション画像例を図 5 に示す．図のように正解となるセグメンテーション画像 $J^{seg,bb}$ に類似した形状のセグメンテーション画像が推定された．

3.3.2　画像変換器の評価

学習したセグメンテーション推定器 S とセグメンテーション画像に対する服装推定器 C を用いて，画像変換器 T と判別器 D を同時に学習した．本実験では上衣における半袖長袖の変換を学習するため，服装データセット $\mathcal{D}_{dpf2}$ 内の服装の種類が short または long のデータを学習用データ 35,580 対，テストデータ 6,276 対に分割して使用した．また，変換の際に条件として与える服装の種類は，画像に付与されたラベルと異なるラベルが与えられた際には入力画像を変換し，同一ラベルが与えられた際には入力画像を保つことが望まれる．前者の方が学習が困難なため，条件として，異なるラベルと同一ラベルを 4:1 の割合で与えるように設定した．

画像変換器 T には，U-net [16] をベースに，画像と条件が入力できるよう変更した条件付き U-net，判別器 D は，同様に画像と条件が入力できる SNGAN [17] の Projection Discriminator を用いて学習した．Adam のパラメータとして $\beta_1 = 0.5, \beta_2 = 0.9$，学習率は判別器 D に対して $lr = 1e-4$，画像変換器 T に対して $lr = 1e-3$，損失関数に対するパラメータは $\lambda_{con} = 1, \lambda_f = 1$ とし，変換器 T を 3 回更新する毎に判別器 D を 1 回更新する学習頻度で，バッチサイズ 8，エポック数 300 エポックとして学習した．学習後の変換画像例を図 6 に示す．図の左側に short の画像に対して long を条件として与えた例，右側に long の画像に対して short を条件として与えた例を示す．変換が薄くノイズのように変換されている画像もあるが，short を条件とした際は，腕部分に肌が生成された一方，long を条件とした際は，元の服装の色を保持して，腕部分に袖が生成されることを確認した．

図 6　画像変換器 T 変換結果例

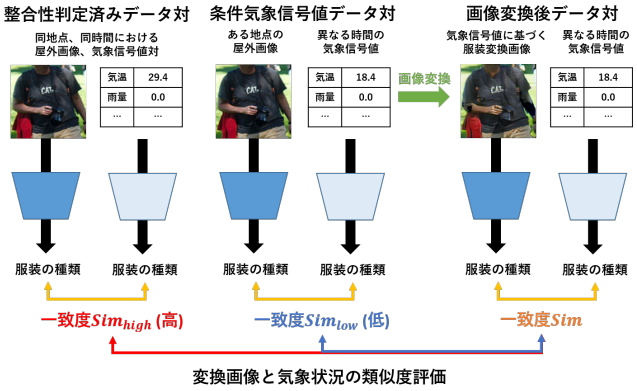

図 7　変換画像の評価方法

3.4　気象状況に応じた人物画像生成と評価

最後に，学習した気象信号値に対する服装推定器 W，画像変換器 T を用い，気象状況に応じて画像中の服装を変換した．図 7 に示すように，変換画像と気象信号値に対して服装を推定し，推定されたラベルの一致度 Sim により評価した．その際，整合性の取れた屋外画像と気象信号値対に対する一致度 Sim_{high}，変換前画像と気象信号値対に対する一致度 Sim_{low} を算出して比較した．

評価データとして，SNS から収集した 15,060 対の $\mathcal{D}_{cw}$ のうち，極端な気象状況の対，690 対を用いた．また，画像変換器 T が SNS 画像に対しても自然な変換が行われるように，評価に用いない残りの SNS 画像 14,370 枚を用いて 10 エポック再学習した．その結果，$Sim = 0.716, Sim_{high} = 0.730, Sim_{low} = 0.484$ となり画像変換により気象状況に応じた画像が増加していることから気象状況に応じた人物の服装変換が実現されたことを確認した．

画像変換例を図 8 に示す．変換画像例より，概ね袖や肌が生成されたことが確認できた．しかし，服装データセットに対する画像変換結果で見られた変換前の服装の色を保持した変換は見られなかった．服装データセットと比べて SNS の画像は多様な姿勢や多数の人物が写るなどの画像の違いがあることから変換結果にも差異が生じたと考察し，SNS 画像など多様な画像に対しても自然に変換できる画像変換器の学習に取り組む必要がある．

図 8　SNS 画像変換例

4　ま　と　め

本研究は任意地点の屋外画像中の人物の服装を，任意時間の気象信号値に応じて変換し，疑似的に該当時間の気象状況に応じた服装の人物画像を生成することを目的とした．特に，気象状況に応じた人物の服装の変換パターンを学習するために十分な量の人物画像と気象信号値の対データの収集が困難という問題を解決するため，大量の画像が含まれる既存の服装データセットが収集可能であること，気象信号値に対する服装の推定は少量のデータで学習可能であることに着目し，ある気象信号値と屋外画像が与えられた際に，まず信号値に応じた服装を推定し，推定された服装に応じて画像変換するという 2 段階の処理により気象状況に応じて人物の服装を変換する手法を提案した．

実際に，SNS の投稿画像と気象センサから収集した 15,060 対の学習データ対と服装に関するメタデータが付与された 111,317 枚の服装画像データセットを用い，気象状況に応じて画像を変換した．変換画像と気象信号値に対する服装推定結果の一致度により変換画像を評価した結果，同じ環境における屋外画像と気象信号値の対と同等の一致度となり気象に応じた画像に変換されていることを確認した．今後の課題として，SNS 画像と服装データセット内の画像の変換結果に差異が生まれたことが挙げられ，多様な画像に対しても自然に変換できる画像変換器を学習する必要がある．

本研究の一部は，JST CREST JPMJCR20D3，科学研究費補助金基盤（C）19K12019 の支援を受けたものである．

文　　献

[1] WorldWeatherOnline. `https://www.worldweatheronline.com/`.

[2] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Neural Information Processing Systems*, pp. 2672–2680, 2014.

[3] Guim Perarnau, Joost Van De Weijer, Bogdan Raducanu, and Jose M Álvarez. Invertible conditional gans for image editing. *arXiv preprint arXiv:1611.06355*, 2016.

[4] Guillaume Lample, Neil Zeghidour, Nicolas Usunier, Antoine Bordes, Ludovic Denoyer, and Marc'Aurelio Ranzato. Fader networks: Manipulating images by sliding attributes. *Neural Information Processing Systems*, pp. 5969–5978, 2017.

[5] 松崎大輔, 新田直子, 中村和晃, 馬場口登. 半教師あり学習による気象条件に応じた屋外画像の生成. データ工学と情報マネジメントに関するフォーラム,7*pages*, 2020.

[6] Sota Kawakami, Kei Okada, Naoko Nitta, Kazuaki Nakamura, and Noboru Babaguchi. Semi-supervised outdoor image generation conditioned on weather signals. *International Conference on Pattern Recognition*, pp. 4268–4275, 2021.

[7] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S Davis. Viton: An image-based virtual try-on network. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7543–7552, 2018.

[8] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 40, No. 6, pp. 1452–1464, 2017.

[9] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 779–788, 2016.

[10] Flickr. `https://www.flickr.com/`.

[11] Wei-Ta Chu, Xiang-You Zheng, and Ding-Shiuan Ding. Image2weather: A large-scale image dataset for weather property estimation. *IEEE International Conference on Multimedia Big Data*, pp. 137–144, 2016.

[12] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[13] Yuying Ge, Ruimao Zhang, Xiaogang Wang, Xiaoou Tang, and Ping Luo. Deepfashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5337–5345, 2019.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

[15] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 39, No. 6, pp. 1137–1149, 2016.

[16] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241, 2015.

[17] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.