Vol.2021-CG-184 No.20
Vol.2021-DCC-29 No.20
Vol.2021-CVIM-227 No.20
2021/11/5

詳細とスタイルを制御可能にした スケッチからの顔画像生成手法

吉川 天斗[†1,a)]　遠藤 結城[†1,b)]　金森 由博[†1,c)]　三谷 純[†1,d)]

概要：
深層学習の発展に伴い、スケッチから画像を生成するニューラルネットワークモデルが数多く研究されてきた。これら多くの研究は、写実的な出力画像を得ることに注力してきたが、クリエイティブなコンテンツ制作のためには多様な出力を得ることも重要になる。そこで本研究では実用性の高さから顔画像を対象に、1 枚のスケッチから多様な画像を生成可能な、一対多のマッピングを学習する深層生成モデルを提案する。提案手法の主なアイデアは、顔画像をしわなどの「詳細」と、肌の色などの「スタイル」の 2 要素に分解して学習する点にある。各要素はそれぞれ別のネットワークによって学習され、その際出力に確率的なブレを与えることで、要素ごとの操作を可能にする。また、従来のスケッチ補正ネットワークを組み合わせることで、粗いスケッチに対しても写実的な画像が生成できる。実験結果を通して、提案手法は写実的な画像を生成しながら、多様かつ柔軟な編集ができることを示す。

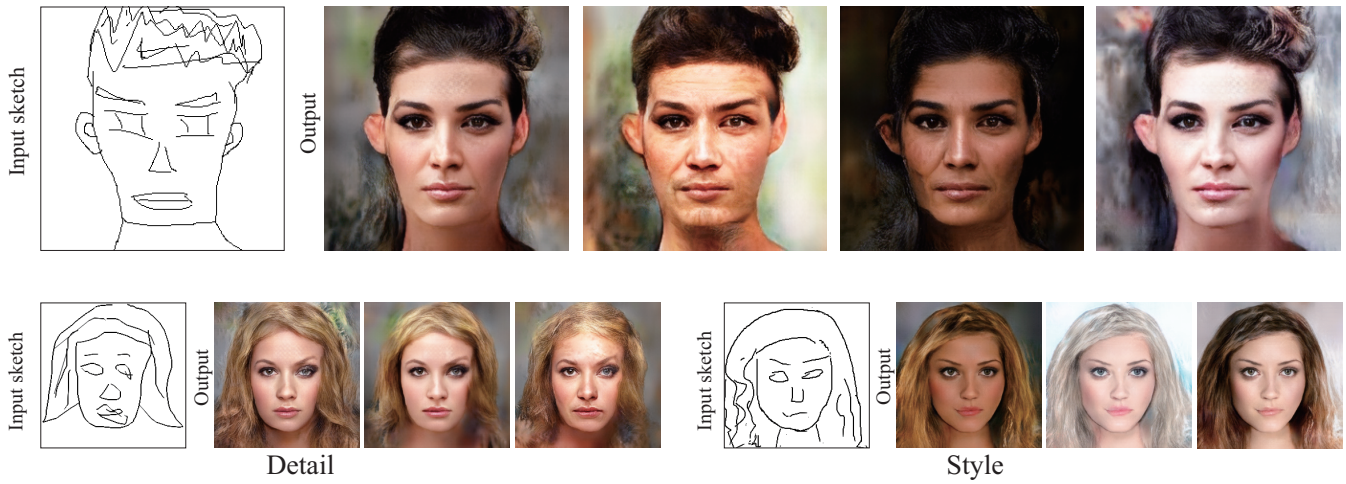


図 1　提案する深層生成モデルでは、1 枚の顔スケッチから多様で写実的な顔画像を生成できる (上段)。また、「詳細」と「スタイル」の要素で別々に制御可能である (下段)。

1. はじめに

スケッチは物の概形などを簡単に表現することができる。スケッチであれば素人のユーザでも、専門的な画像編集ツールなどを使わずに簡単に写真の加工が可能になる。高性能なタッチスクリーンデバイスの普及によりその利点はさらに高まっており、こうした背景からスケッチからの対話的な画像生成・編集手法が盛んに研究されている。このような技術が実現すれば、アイデアをすぐに写実的な画像にして見せたり、犯罪捜査を円滑に進めるために犯人の似顔絵から写真を生成したりする応用も期待できる。そこで本研究では、特に実用性の高い顔画像の生成を目的とする。

スケッチからの画像生成において、重要なのは**写実性**と**多様性**である。写実性とは生成画像中の顔が人物の顔として自然な見た目かどうかということである。写実性の向上は先行研究でも多く取り組まれている。例えば深層学習を用いた最先端の手法では、不自然なスケッチの形状を補正するようにネットワークモデルを設計することで、より自然な画像を生成できる [32]。一方、多様性とは 1 枚のスケッチに対して生成される画像のバリエーションの多さであ

†1　現在，筑波大学
Presently with University of Tsukuba
a)　tenten0727@icloud.com
b)c)d)　{endo,kanamori,mitani}@cs.tsukuba.ac.jp

り、コンテンツ制作においてユーザの創造性を高めるために重要である。スケッチは実写画像よりも情報量が少ないため、生成画像は様々な候補が考えられるはずである。例えば顔画像において、肌の色や髪の毛の色、髪の毛の流れ、肌のしわなど、スケッチだけで表現しきれない要素は、多様な出力が想定される。既存手法の多くは 1 枚のスケッチから一意に定まった生成画像しか得られず、多様性に関して十分に考慮できていない [3], [4], [7], [22], [26], [32]。

そこで本研究では、写実性と多様性の両方を考慮したスケッチからの顔画像生成手法を提案する。本手法は、敵対的生成ネットワーク (GAN) に基づいた画像対画像変換ネットワークを用いる。写実性については、*Deep Plastic Surgery* (DPS) [32] における形状補正ネットワークを組み合わせることで、粗いスケッチに対しても自然な画像を生成する。多様性については、顔画像を「詳細」と「スタイル」の 2 つの要素に分けて、それぞれ独立に制御可能とすることで多様な出力を実現する。顔画像の「詳細」とは髪の毛の流れや肌のしわなどで、「スタイル」とは肌や髪の色などのことを指す。この 2 つの要素を独立に制御するための具体的な方法として、「詳細」を表現するネットワークと「スタイル」を表現するネットワークを分けて訓練する。「詳細」のネットワークでは疎なスケッチから密なスケッチへの変換を学習する。この訓練のために密なエッジマップを学習データとして使用するが、様々なエッジマップで試し、最もよく「詳細」を表現しているエッジマップを用いた。「スタイル」のネットワークでは、「詳細」のネットワークで生成された密なエッジマップからカラー画像への変換を学習する。また、それぞれのネットワークに *Wasserstein Auto-Encoder* (WAE) [28] という手法を組み込むことで確率的なブレを与え、多様な出力を可能としている。**図 1** に示す通り、本手法では 1 枚のスケッチから多様な顔画像を生成できる。また、「詳細」と「スタイル」のネットワークを分けて訓練しているため、「詳細」のみの多様化や「スタイル」のみの多様化が可能である。既存手法との比較を行い、最新の手法と遜色ない写実性を保ちながら多様な出力が可能であることを実証する。

2. 関連研究

2.1 画像対画像変換

入力画像を目的のドメインに変換するために、深層学習を用いた様々な画像対画像変換手法が提案されてきた。代表例として、Isola らは pix2pix [12] という画像変換フレームワークを設計した。この手法ではセマンティックラベルマップやエッジマップを写真に変換できる。生成手法として Conditional GAN [23] を用いており、識別器と生成器による敵対的学習によって写実的な画像変換を実現している。その後、より高解像度の画像 [13], [30] やペアでない画像 [6], [19], [20], [35] 同士の変換が可能になった。

2.2 スケッチからの画像生成

pix2pix [12] をはじめとしたいくつかのモデル [7], [26] は、スケッチよりも簡単に取得できるエッジマップから写真への変換を学習している。しかし、人が描いたスケッチとエッジマップでは構造が異なるため、前述のモデルをスケッチに一般化することは難しい。その後、いくつかのスケッチデータセット [25], [34] が公開されたものの、汎化性能の高いモデルを学習するには十分なデータ量とは言えない。Chen らはこのデータ不足を補うために、スケッチに近づくようにデータ拡張を施したエッジマップを GAN の学習に用いることで、多様なクラスに対する画像生成を実現した [4]。しかし、依然スケッチに一般化したモデルにはなっておらず、生成画像の写実性はそれほど高くない。シンプルにスケッチの学習データを増やせば生成画像の品質向上も見込めるが、アノテーションには多大なコストがかかる。

この問題に対して、エッジマップベースのモデルをスケッチに適応させるアプローチもある。ContextualGAN [22] ではエッジマップと写真を結合した画像の分布を学習し、入力したスケッチに近いものを検索するというアプローチを取った。さらにこのようなアプローチで顔画像の生成に特化した手法 [3] も提案された。また、Yang らはスケッチの忠実度を制御可能にしたスケッチベースの画像編集フレームワークである DPS [32] を提案した。学習の際、前処理としてエッジマップを変形・膨張させることで擬似的なスケッチを生成し、これを処理前のエッジマップに戻すようにネットワークを学習させる。これにより推論時には手描きスケッチをエッジ風に補正し、最終結果としてより写実的な画像を得られる。この際ユーザはパラメータによって、どの程度スケッチを補正するかを指定できる。これらの研究では生成画像の写実性を向上させたが、多様性について十分な考慮ができていない。一方本手法は、写実性と多様性の両方を考慮したスケッチ対画像変換を実現できる。

2.3 生成画像の多様化

GAN における生成画像の多様化はチャレンジングな問題である。例えば pix2pix [12] ではモード崩壊という問題が見られる。出力を制御するために追加の入力としてノイズを使用しても、もっともらしい単一の結果しか得られない。この問題を解決するために、様々な手法が提案された。Ghosh らはモードが異なるように学習した複数の生成器を用いてマルチモーダル画像合成を実現した [10]。しかし、この手法では 1 枚の入力画像から固定数の画像しか生成できない。モード崩壊の問題を解決しつつ、可変のモードを扱うために、Larsen らは *Variational Auto-Encoder* (VAE) [15] を GAN と共に用いた VAE-GAN [16] を提案した。また、BicycleGAN [36] では潜在空間への回帰器を用いて、VAE-GAN を条件付き GAN で行えるようにし、

Vol.2021-CG-184 No.20
Vol.2021-DCC-29 No.20
Vol.2021-CVIM-227 No.20
2021/11/5

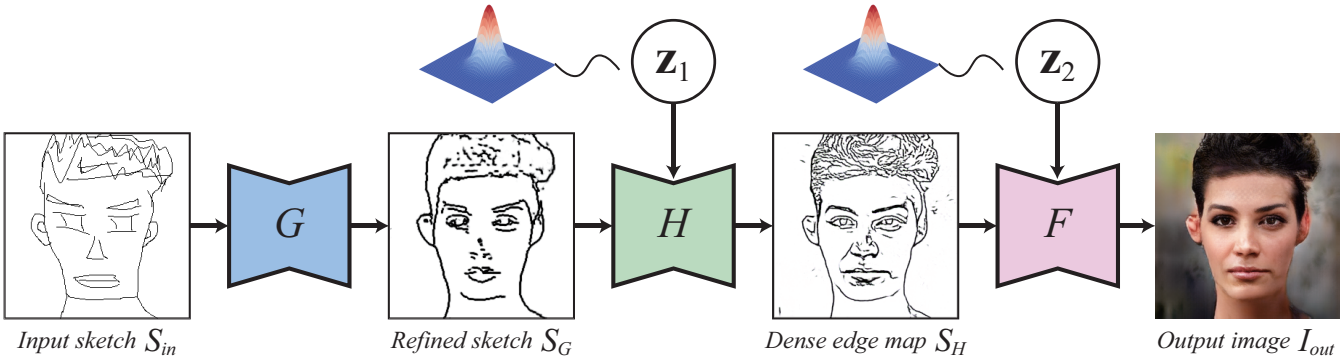


図 2 本手法の推論時の概要図。G はスケッチを補正するネットワークであり DPS [32] の学習済みネットワークを用いる。H と F がそれぞれ「詳細」と「スタイル」を表現するネットワークであり、潜在変数 $\mathbf{z}_1$、$\mathbf{z}_2$ によってそれぞれの要素を制御できる。

pix2pix で多様な生成画像を出力できるようにした。これらのような汎用的な画像変換だけでなく、ラベルマップを入力としたセマンティック画像合成においても出力のマルチモーダル化が盛んに研究されている [8], [18], [21], [24]。上記の手法では十分な量と品質の学習データがあることが前提であり、既存のスケッチと画像のペアのデータセットはその条件を満たさないため、そのまま適用しても良好な結果は得られない。

そのため、スケッチからの画像生成でも多様な出力が可能な手法がいくつか提案された。Yang らは属性で制御可能な手法 [33] を提案した。属性での制御では意味的な編集を簡単に行うことができるが、スケッチの形状を無視した編集をする可能性がある。Lee らの手法 [17] では参照画像を入力として与えることで、「スタイル」を指定できるようにした。これらの手法に対し、本手法では「スタイル」に加えて「詳細」についても制御でき、スケッチの形状をどれくらい忠実に反映するかも指定できる。さらに、参照画像を用意しなくても、潜在変数に事前分布を仮定することで、ランダムで出力を多様化できる。

最近では Wang らが学習済みの生成器を、与えられたスケッチの形状やポーズだけ合わせるようにファインチューニングする手法 [29] を提案した。この手法では写実性、多様性共に高いレベルの生成画像を得ることができる。一方で、スケッチごとにモデルをファインチューニングしなければならないため、インタラクティブな画像生成はできない。それに比べて、本手法では高速なフィードフォワード方式で多様な画像を出力するので、ユーザはインタラクティブに画像を生成できる。

3. 提案手法

本研究の目的は「詳細」と「スタイル」に関して、多様な顔画像をスケッチから生成することである。**図 2** に提案手法の推論時の概略を示す。まず、手描きのスケッチ S_{in} がスケッチ補正ネットワーク G に入力され、補正されたスケッチ S_G に変換される。次に S_G が、「詳細」を表現するネットワーク H に入力され、さらにエッジが細かく描かれた S_H に変換される。最終的に「スタイル」を表現するネットワーク F に S_H を入力して顔画像の写真 I_{out} が生成される。推論時、H と F では潜在変数 $\mathbf{z}_1$、$\mathbf{z}_2$ による制御ができる。$\mathbf{z}_1$、$\mathbf{z}_2$ は正規分布 $\mathcal{N}(\mathbf{0}, \mathbf{I})$ からサンプリングした値である。また、それぞれのネットワークは教師あり学習によって訓練される。3.1 節でネットワーク構造、3.2 節で学習の方法について説明する。

3.1 ネットワーク構造

ネットワーク G には DPS [32] と同じ構造を用いる。また、**図 3**(a) にネットワーク H の学習時の概要図を、図 3(b) に点線内部の詳細を示す。図 3 には H の学習のみ示しており、F の学習については示していない。F については用いる損失関数が異なるがネットワーク構造は H と同様であるため、これ以降は H を例に説明する。H はエンコーダ・デコーダ型の画像変換ネットワークである。図 3(b) の青で囲まれたブロックは活性化関数 (ReLU)、Reflection Padding、フィルタサイズ 3×3 の畳み込み層、*Adaptive Instance Normalization* (AdaIN) [11] による正規化から構成される。このブロックがエンコーダに 6 つ含まれる。また、デコーダでは Reflection Padding と畳み込み層の部分を転置畳み込み層にしたブロックが、エンコーダと同様 6 つ含まれる。ネットワーク D_H は H で生成された S_H と正解画像 S_H^{gt} の真偽を判別する識別器である。D_H は DPS [32] の識別器と同様のものを使用した。ネットワーク E_W と D_W はそれぞれ WAE のエンコーダと識別器である。これらのエンコーダは Endo らの研究 [9] で用いられているエンコーダを使用した。また、D_W には 2 層の全

結合層からなるネットワークを使用した。

WAE による「詳細」と「スタイル」の多様化

多様性を実現するためによく使われている手法として VAE [15] が挙げられる [8], [16], [24]。VAE では、Kullback-Leibler divergence を用いた正則化によって、ある特定の事前分布に従った潜在空間を学習でき、推論時にその分布から潜在変数をサンプリングすることで多様な出力が得られる。しかし、VAE では学習の際 Reparameterization Trick という手法を用いており、学習の度に確率分布 (ここでは正規分布) からサンプリングしてデコーダに渡すため、同じ潜在変数で異なる画像を再構成する訓練をしてしまう可能性がある。このため、VAE では生成画像がぼやけるという問題があった。これを解決するために本手法では WAE [28] を導入する。図 3(b) の右側のように、S_H^{gt} をエンコーダ E_W によって潜在変数 $\mathbf{z}_f$ にエンコードし、仮定した事前分布からサンプリングした潜在変数 $\mathbf{z}_r$ と共に識別器 D_W で損失を測る。このように GAN ベースの学習で分布を近づけることで VAE における問題を解決する。また、$\mathbf{z}_f$ は H でデコードされ、S_H^{gt} が再構成される。WAE を本手法に組み込むことで、ある事前分布に従った「詳細」や「スタイル」を表現する潜在空間を学習し、多様な出力を実現できる。

AdaIN による「詳細」と「スタイル」の情報の挿入

本手法では E_W によりエンコードした潜在変数 $\mathbf{z}_f$ の情報を、ネットワーク H や F に反映させるために AdaIN を用いた。AdaIN [11] はスタイルベースの画像生成 [14] で用いられている正規化手法である。AdaIN は Instance Normalization で正規化した特徴量を、別の特徴量の分散でスケーリング、平均でシフトすることで情報を挿入している。図 3(b) に示す通り 2 つの線形変換層を用いて、$\mathbf{z}_f$ を平均と分散にマッピングしている。H や F のエンコーダでは最初以外の各畳み込み層の後に、デコーダでは最後以外の各逆畳み込み層の後に AdaIN で正規化している。

3.2 学習方法

ネットワーク G には DPS [32] で事前学習されたものを使用する。具体的には、まずエッジマップの線に対して変形、削除、膨張の前処理を施す。入力を前処理済みのエッジマップとし、元のエッジマップが出力となるように訓練することで、G はスケッチの補正を学習する。

ネットワーク H と F は個別に学習する。それぞれの学習方法は次の通りである。図 3 に示す通り、H は粗いエッジから詳細なエッジへの変換を学習する。粗いエッジ S_G^{gt} としては、写真から HED [31] という手法で抽出したエッジマップを用いた。詳細なエッジマップ S_H^{gt} としてはいくつかの候補が考えられるが、実験 (4.4 節参照) を通して最も「詳細」をよく表せていた DFE [2] によるエッジマップを採用した。一方、F は H で出力されたエッジマップから最終的な顔画像への変換を学習する。そのため、F は入力画像に詳細なエッジ S_H^{gt}、正解画像にそのエッジに対応する顔画像の写真 I_{gt} を用いて学習する。

損失関数

以下はネットワーク H の学習時の損失関数である。正解画像と生成画像を画素単位で近づけるため、L_1 損失 $\mathcal{L}_{rec}$ を用いる。

$$\mathcal{L}_{rec} = \mathbb{E}\left[\left\|S_H - S_H^{gt}\right\|_1\right]. \tag{1}$$

さらに正解画像と生成画像の意味上の類似性を評価するために、知覚損失 $\mathcal{L}_{prec}$ を以下のように計算する。

$$\mathcal{L}_{perc} = \mathbb{E}[\sum_i \lambda_i \left\|\Phi_i(S_H) - \Phi_i(S_H^{gt})\right\|_2^2], \tag{2}$$

ここで、$\Phi_i(x)$ は、VGG19 [27] の i 番目の層における x の特徴マップであり、 λ_i は各層における重みである。また、敵対的損失 $\mathcal{L}_G$ と $\mathcal{L}_D$ にはヒンジ損失を用いる。

$$\mathcal{L}_G = -\mathbb{E}\left[D_H(S_H)\right], \tag{3}$$

$$\mathcal{L}_D = \mathbb{E}\left[\mathbf{ReLU}(\tau + D_H(S_H))\right] + \mathbb{E}\left[\mathbf{ReLU}(\tau - D_H(S_H^{gt}))\right], \tag{4}$$

ここで、τ は定数である。さらに、エッジ同士の特徴を一致させるために Feature Matching Loss [16] を導入する。

$$\mathcal{L}_{FM} = \mathbb{E}[\sum_j \left\|D_H^{(j)}(S_H) - D_H^{(j)}(S_H^{gt})\right\|_1], \tag{5}$$

ここで、$D_H^{(j)}(x)$ は D_H の j 番目の層における x の特徴マップである。最後に、WAE の敵対的損失を計算する。

$$\mathcal{L}_{GW} = -\mathbb{E}\left[D_W(E_W(S_H^{gt}))\right], \tag{6}$$

$$\mathcal{L}_{DW} = -(\mathbb{E}\left[D_W(\mathbf{z})\right] - \mathbb{E}\left[D_W(E_W(S_H^{gt}))\right]), \tag{7}$$

ここで、$\mathbf{z} \in \mathbb{R}^n$ は $E_W(S_H^{gt})$ と同じ次元数 n の正規分布 $\mathcal{N}(\mathbf{0}, \mathbf{I})$ からランダムにサンプリングした値である。なお、ネットワーク F は $\mathcal{L}_{FM}$ 以外の損失関数で学習する。

4. 実験

4.1 データセット

モデルの学習には CelebA-HQ データセット [13] を使用した。このデータセットの顔画像から HED [31] と DFE [2], [3] でエッジマップを抽出した。これらの画像は全て 256×256 にリサイズされている。最初の 28,000 枚を学習データに、残りの 2,000 枚をテストに用いた。

4.2 学習時のパラメータ

最適化手法は Adam を採用し、学習率は 0.0002 に固定し、モーメント推定値の指数減衰率を {0.5, 0.999} とした。バッチサイズは 4 とした。すべての実験において、$\mathcal{L}_{rec}$、$\mathcal{L}_G$、$\mathcal{L}_D$、$\mathcal{L}_{FM}$、$\mathcal{L}_{GW}$、$\mathcal{L}_{DW}$ の重みは 100、1、1、10、10000、

Vol.2021-CG-184 No.20
Vol.2021-DCC-29 No.20
Vol.2021-CVIM-227 No.20
2021/11/5

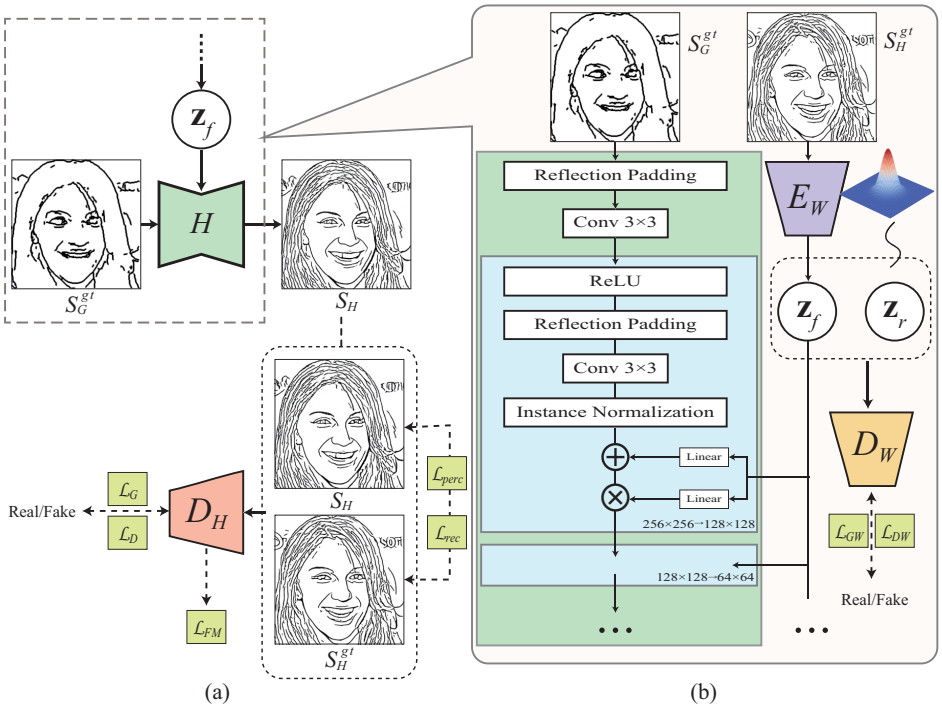


図 3 本手法のネットワーク H の訓練時の概要図。ネットワーク F については $\mathcal{L}_{FM}$ を損失関数として用いないこと以外は H の訓練と同様である。(a) は学習時の全体像を表しており、(a) の一部の詳細を (b) に示す。

1 とした。$\mathcal{L}_{perc}$ の計算には VGG19 [27] の conv2 の第 1 層と conv3 の第 1 層を、それぞれ 1 と 0.5 で重み付けして使用した。ヒンジ損失について、τ を 10 として計算した。潜在変数 $\mathbf{z}$ の次元数 n は 8 とした。本手法で用いたモデルは、ほぼ学習が収束したとみられる 20 エポックまで学習したものを使用している。

4.3 既存手法との比較

図 4 に示す通り、既存手法との定性的な比較を行った。Park らの手法 (SPADE) [24] では潜在変数による多様化が可能である。手描きスケッチをそのまま SPADE のネットワークに入力した場合、スケッチの線に忠実過ぎてしまい、写実性の低い画像が生成される。そこで、本手法でも用いているスケッチ補正ネットワーク G によって補正されたスケッチを SPADE の入力として用いた (SPADE+G)。その結果、写実性の向上は見られるが、本手法の方がより良好な結果を得られている。また、どちらの手法も出力は多様化できているが、本手法の方がより多様な顔画像を得られている。DPS [32] はスケッチの補正ネットワークを用いた本手法のベースラインとなる手法だが、最終結果としてぼやけた画像が生成されやすい。本手法ではネットワークを増やし段階的に生成を行うことで、図 4 に示すように鮮明な結果が得られる。このように写実性を向上させながら、DPS ではできなかった多様な画像生成が可能となったことが本手法の利点といえる。

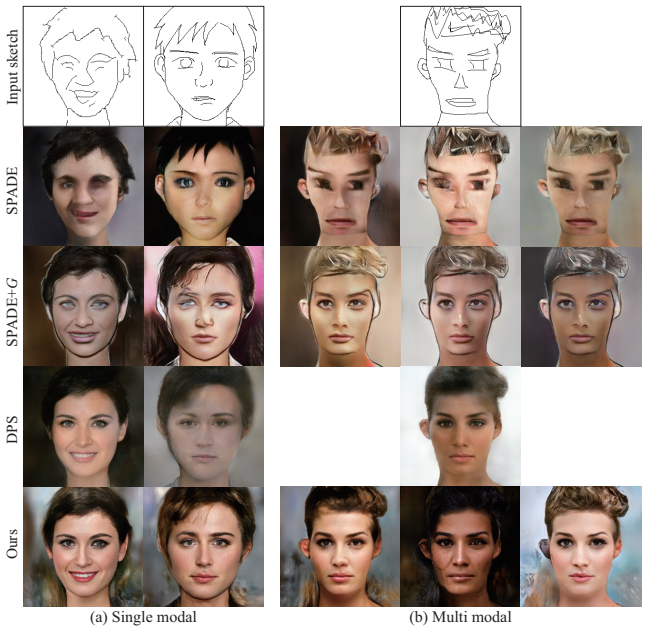


図 4 既存手法との比較。SPADE [24]、スケッチ補正ネットワーク G を追加した SPADE、DPS [32]、本手法の 4 つの手法での生成画像を比較した。本手法は既存手法よりも写実性と多様性において優れた結果となっている。

4.4 H の学習に用いるエッジマップの比較

本節では、「詳細」を表現するネットワーク H の学習に、様々な正解エッジマップを用いた場合の結果の違いを議論する。HED [31] によるエッジマップは対象の概形を良く表せているが、対象の「詳細」は表現できていないことが

Vol.2021-CG-184 No.20
Vol.2021-DCC-29 No.20
Vol.2021-CVIM-227 No.20
2021/11/5

図 5 H の学習に用いるエッジマップの種類による S_H、I_{out} の比較。Canny エッジマップ (a) と Sobel エッジマップ (b) では「詳細」を表現できず、最終的な生成画像の写実性も低下している。これに対し、DFE で抽出したエッジマップでは髪の流れなどの顔の「詳細」を表現できており、その情報が適切に反映された生成画像が得られる。

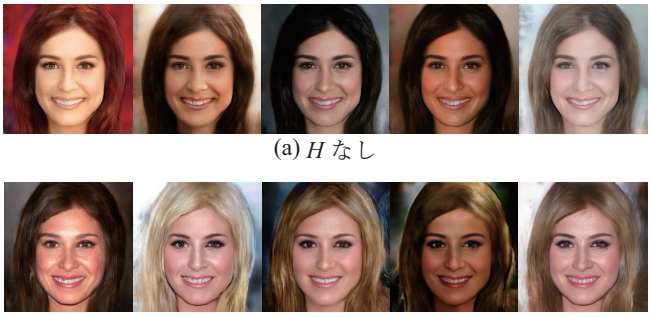


図 6 ネットワーク H の有無による生成画像の比較。H がある方がない場合よりも、写実性と多様性の両方が向上している。

表 1 多様化を行う手法として VAE と WAE を使った場合の評価指標の値の比較。太字は最良値を表す。

	LPIPS ↑	FID ↓
VAE	0.327	128.8
WAE	**0.377**	**103.3**

多い。そこで、**図 5** に示すように HED よりも密に線が描画される Canny エッジ [1]、Sobel エッジ、そして DFE で抽出されたエッジの 3 つのエッジマップを正解データとして用い、結果を比較した。Canny エッジや Sobel エッジで H を学習した結果を見ると、エッジの線は密になっているが、肌のしわや髪の流れのような「詳細」を表した線にはなっていない。その結果、最終的な生成画像も写実性に欠ける結果となっている。それに対し、DFE で学習した場合は「詳細」を良く表現できており、最終的な生成画像ではその「詳細」が反映され写実的な画像となっている。

4.5 Ablation Study

「詳細」を表現するネットワーク H

本節では、「詳細」を表現するネットワーク H の有無が、多様性や写実性に与える影響を検証する。図 6 は本手法の H の有無による生成画像の比較である。H なしのモデルは、S_G^{gt} から I_{gt} への変換を行うように F の学習データを変更し、G と F のネットワークだけで構成されている。結果に示す通り、H を追加したことにより要素の制御が可能になっただけでなく、H なしのモデルよりも鮮明な画像が

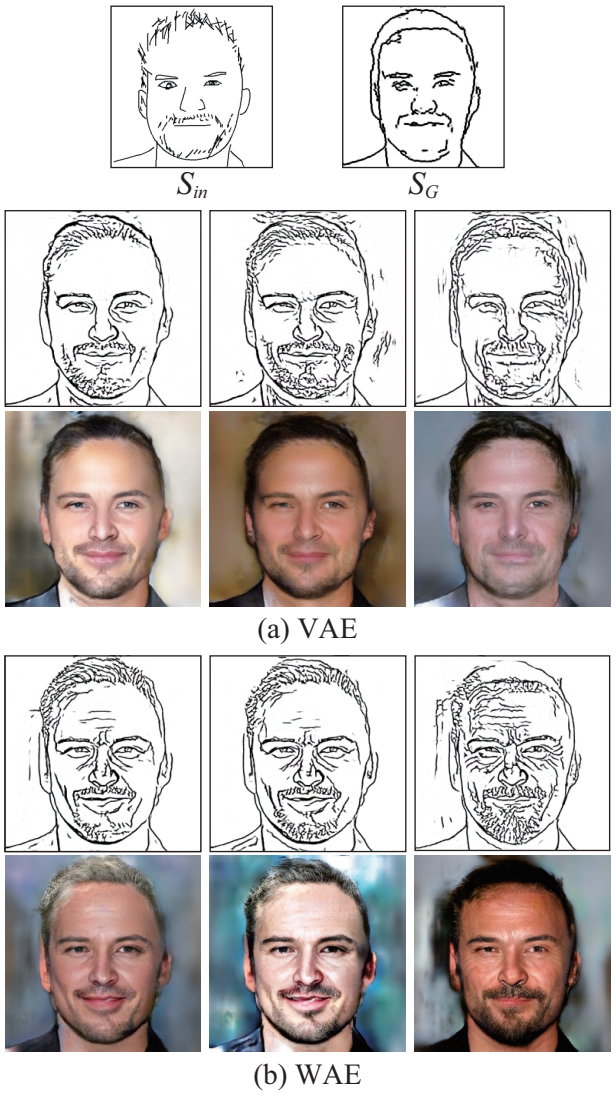
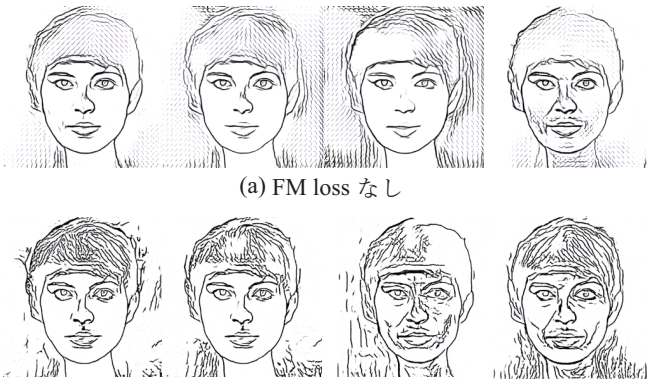


図 7 多様化を行う手法として VAE を使った場合と WAE を使った場合の生成画像の比較。それぞれ上の段が S_H 下の段が I_{out}。WAE を使った場合の方が S_H のエッジに多様性があり、I_{out} も肌の色や髪の色に関して多様な画像となっている。

生成されている。また、肌の質感や髪の色の変化が H ありのモデルの方が大きいことから、「詳細」と「スタイル」の多様性を向上させる効果もあることがわかる。これは要素を分離して学習することにより、モデルがそれぞれの特徴を捉えやすくなったためだと考えられる。

Vol.2021-CG-184 No.20
Vol.2021-DCC-29 No.20
Vol.2021-CVIM-227 No.20
2021/11/5



(a) FM loss なし

(b) FM loss あり

図 8 H の学習時の *Feature Matching Loss* (FM loss) の有無による S_H の比較。FM loss なしの場合は格子状のアーティファクトが生じる。FM loss を導入するとそのアーティファクトは見られなくなり、「詳細」を表現した S_H が得られる。

VAE と WAE の比較

図 7 は H、F ともに VAE [15] を用いたモデルと WAE [28] を用いたモデルの S_H と I_{out} を比較している。VAE を用いた画像では、「詳細」と「スタイル」の多様性が低く、髭の質感などがあまり鮮明でない。これは Reparameterization Trick によって、同じ潜在変数で異なる画像を再構成する訓練をしてしまう可能性があるという VAE の欠点が原因で、「詳細」と「スタイル」の潜在空間がうまく学習できていないと考えられる。これに対し WAE を用いた画像では、それぞれの要素の多様性が向上しており、鮮明な結果が得られている。**表 1** では、VAE と WAE を定量的に比較している。文献 [5] に倣い、多様性を LPIPS、写実性を FID によって評価した。具体的には図 7 のスケッチ S_{in} からランダムに生成された 100 枚の生成画像 I_{out} に対して、生成画像間の LPIPS 、およびテストセットとの FID を計算した。どちらのスコアも WAE が VAE を上回っていることから、本手法では多様化に用いる手法として VAE よりも WAE が適していることがわかる。

***Feature Matching Loss* (FM loss)**

図 8 では、ネットワーク H に FM loss を導入した場合としていない場合の出力 S_H を比較している。FM loss なしの場合、エッジの特徴を上手く学習できておらず、格子状のアーティファクトが発生している。FM loss を導入することで H は正解のエッジと生成されたエッジの特徴が一致するように学習が進み、図 8(b) に示すようなアーティファクトの目立たない髪の流れや肌のしわを表現できる。

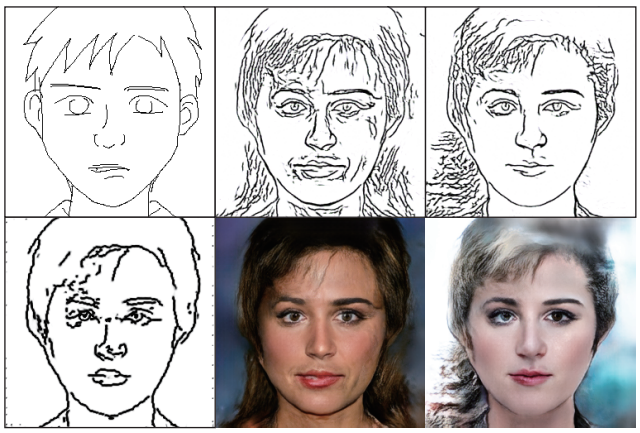
4.6 アプリケーション

本手法では**図 9** に示す通り、参照画像を用いた顔画像の「詳細」や「スタイル」の転写も可能である。「詳細」や「スタイル」の潜在変数をランダムに抽出するのではなく、参照画像をエンコーダ E_W に入力して得られた潜在変数を使用することで、参照画像の「詳細」または「スタイル」を生成画像に反映できる。

(a) Detail transfer

(b) Style transfer

図 9 提案手法の応用例。本手法では参照画像を用いた「詳細」変換 (a) や「スタイル」変換 (b) が可能である。



(a) (b) (c)

図 10 提案手法の失敗例。(a) 入力と S_G。(b) 顔の周りに余計な線が追加される例。(c)「詳細」と「スタイル」の潜在変数の組み合わせによって写実性が落ちる例。

5. まとめと今後の課題

本研究では 1 枚のスケッチから多様な画像を生成可能な、一対多のマッピングを学習する深層生成モデルを提案した。「詳細」と「スタイル」の 2 つの要素に分けて、別々のモジュールで学習することでそれぞれの要素を制御可能としている。また、多様化の手法として WAE を用いており、AdaIN によって「詳細」と「スタイル」の情報をネットワークに挿入することで多様性を実現している。

本手法は 1 枚のスケッチから写実的で多様な出力を得ることができたが、いくつかの問題が存在する。**図 10** に示す通り、H のネットワークでエッジを詳細化する際に顔の周りに余計な線が追加されてしまうことがある。これは顔

のスケッチに沿ったマスクを作り、詳細化する範囲を限定することによって解決できる可能性がある。次に、「詳細」と「スタイル」の潜在変数の組み合わせによっては画像の写実性が落ちてしまう場合がある。今後の課題として、どの組み合わせでも写実的な画像が生成されるようなモデルの設計を検討したい。

参考文献

[1] John F. Canny. A computational approach to edge detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 8, No. 6, pp. 679–698, 1986.
[2] Shu-Yu Chen, Feng-Lin Liu, Yu-Kun Lai, Paul L. Rosin, Chunpeng Li, Hongbo Fu, and Lin Gao. DeepFaceEditing: deep face generation and editing with disentangled geometry and appearance control. *ACM Trans. Graph.*, Vol. 40, No. 4, pp. 90:1–90:15, 2021.
[3] Shu-Yu Chen, Wanchao Su, Lin Gao, Shihong Xia, and Hongbo Fu. DeepFaceDrawing: deep generation of face images from sketches. *ACM Trans. Graph.*, Vol. 39, No. 4, p. 72, 2020.
[4] Wengling Chen and James Hays. SketchyGAN: Towards diverse and realistic sketch to image synthesis. In *CVPR 2018*, pp. 9416–9425, 2018.
[5] Yen-Chi Cheng, Hsin-Ying Lee, Min Sun, and Ming-Hsuan Yang. Controllable image synthesis via SegVAE. In *ECCV 2020*, Vol. 12352 of *Lecture Notes in Computer Science*, pp. 159–174. Springer, 2020.
[6] Yunjey Choi, Min-Je Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation. In *CVPR 2018*, pp. 8789–8797, 2018.
[7] Tali Dekel, Chuang Gan, Dilip Krishnan, Ce Liu, and William T. Freeman. Sparse, smart contours to represent and edit images. In *CVPR 2018*, pp. 3511–3520, 2018.
[8] Yuki Endo and Yoshihiro Kanamori. Diversifying semantic image synthesis and editing via class- and layer-wise VAEs. *Comput. Graph. Forum*, Vol. 39, No. 7, pp. 519–530, 2020.
[9] Yuki Endo, Yoshihiro Kanamori, and Shigeru Kuriyama. Animating landscape: self-supervised learning of decoupled motion and appearance for single-image video synthesis. *ACM Trans. Graph.*, Vol. 38, No. 6, pp. 175:1–175:19, 2019.
[10] Arnab Ghosh, Viveka Kulharia, Vinay P. Namboodiri, Philip H. S. Torr, and Puneet Kumar Dokania. Multi-agent diverse generative adversarial networks. In *CVPR 2018*, pp. 8513–8521, 2018.
[11] Xun Huang and Serge J. Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV 2017*, pp. 1510–1519, 2017.
[12] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR 2017*, pp. 5967–5976, 2017.
[13] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *ICLR 2018*, 2018.
[14] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR 2019*, pp. 4401–4410, 2019.
[15] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR 2014*, 2014.
[16] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. Autoencoding beyond pixels using a learned similarity metric. In *ICML 2016*, Vol. 48 of *JMLR Workshop and Conference Proceedings*, pp. 1558–1566, 2016.
[17] Junsoo Lee, Eungyeup Kim, Yunsung Lee, Dongjun Kim, Jaehyuk Chang, and Jaegul Choo. Reference-based sketch image colorization using augmented-self reference and dense semantic correspondence. In *CVPR 2020*, pp. 5800–5809, 2020.
[18] Ke Li, Tianhao Zhang, and Jitendra Malik. Diverse image synthesis from semantic layouts via conditional IMLE. In *ICCV 2019*, pp. 4219–4228, 2019.
[19] Ming-Yu Liu, Thomas M. Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *NeurIPS 2017*, pp. 700–708, 2017.
[20] Ming-Yu Liu, Xun Huang, Arun Mallya, Tero Karras, Timo Aila, Jaakko Lehtinen, and Jan Kautz. Few-shot unsupervised image-to-image translation. In *ICCV 2019*, pp. 10550–10559, 2019.
[21] Xihui Liu, Guojun Yin, Jing Shao, Xiaogang Wang, and Hongsheng Li. Learning to predict layout-to-image conditional convolutions for semantic image synthesis. In *NeurIPS*.
[22] Yongyi Lu, Shangzhe Wu, Yu-Wing Tai, and Chi-Keung Tang. Image generation from sketch constraint using contextual GAN. In *ECCV 2018*, Vol. 11220 of *Lecture Notes in Computer Science*, pp. 213–228, 2018.
[23] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *CoRR*, Vol. abs/1411.1784, , 2014.
[24] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *CVPR 2019*, pp. 2337–2346, 2019.
[25] Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. The sketchy database: learning to retrieve badly drawn bunnies. *ACM Trans. Graph.*, Vol. 35, No. 4, pp. 119:1–119:12, 2016.
[26] Patsorn Sangkloy, Jingwan Lu, Chen Fang, Fisher Yu, and James Hays. Scribbler: Controlling deep image synthesis with sketch and color. In *CVPR 2017*, pp. 6836–6845, 2017.
[27] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR 2015*, 2015.
[28] Ilya O. Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schölkopf. Wasserstein auto-encoders. In *ICLR 2018*, 2018.
[29] Sheng-Yu Wang, David Bau, and Jun-Yan Zhu. Sketch your own GAN. *CoRR*, Vol. abs/2108.02774, , 2021.
[30] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional GANs. In *CVPR 2018*, pp. 8798–8807, 2018.
[31] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *ICCV 2015*, pp. 1395–1403, 2015.
[32] Shuai Yang, Zhangyang Wang, Jiaying Liu, and Zongming Guo. Deep plastic surgery: Robust and controllable image editing with human-drawn sketches. In *ECCV 2020*, Vol. 12360 of *Lecture Notes in Computer Science*, pp. 601–617, 2020.
[33] Yan Yang, Md. Zakir Hossain, Tom Gedeon, and Shafin Rahman. S2FGAN: semantically aware interactive sketch-to-face translation. *CoRR*, Vol. abs/2011.14785, , 2020.
[34] Qian Yu, Feng Liu, Yi-Zhe Song, Tao Xiang, Timothy M. Hospedales, and Chen Change Loy. Sketch me that shoe. In *CVPR 2016*, pp. 799–807, 2016.
[35] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV 2017*, pp. 2242–2251, 2017.
[36] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A. Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In *NeurIPS 2017*, pp. 465–476, 2017.