社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

StyleGANによるCLIP-Guidedな画像形状特徴編集

銭　　雨晨†　柳井　啓司†

† 電気通信大学 大学院情報理工学研究 情報学専攻
E-mail: †qian-y@mm.inf.uec.ac.jp, ††yanai@cs.uec.ac.jp

あらまし　近年，画像と自然言語のマルチモーダルモデルを利用し，自然言語から画像特徴を編集する研究が注目されている．テキストを用いた画像特徴編集タスクにおいて，既存の手法では，画像内のオブジェクトの外観特徴（色やテクスチャなどの特徴）に対する編集が主流で，オブジェクトの形状特徴（一部の形状やサイズなどの特徴）に対する編集の研究は少ない．そこで本研究は画像生成モデルのStyleGAN2とimage-textマッチングモデルのCLIPを利用し，事前学習済みStyleGAN2生成器のパラメータを調整することにより，画像形状特徴の編集を実現する手法を提案する．定性評価と定量評価の実験を行い，提案モデルが目標特徴の変換を達成でき，編集後の画像品質を維持できることを示した．

キーワード　GAN，画像編集

1. はじめに

近年，深層学習に基づいた画像変換や画像特徴編集の研究が盛んに行われており，自然でリアルな画像変換や編集が可能になっている．その成果は，写真編集やコンテンツ創造などの領域に応用されている．深層学習モデルの力をなるべく完全に発揮するため，人と機械の間に，より使いやすいインタフェースが必要になる．それに応じて，深層学習技術を用いたマルチモーダルモデルに関する研究が発展しており，画像と自然言語，動画と自然言語など，複数種類のデータを処理してそれらの関係性を把握することにより，より人間に近い感覚の操作や判定をできるようになる．そのため，マルチモーダルモデルのimage-textマッチングモデルを利用し，自然言語をインタフェースとして，画像特徴を編集する研究が注目されている．

深層学習モデルの学習には，計算資源，大量なデータと時間が必要である．そのため，ゼロからモデルを学習する代わりに，事前学習済みモデルを活用する方法が一般的である．大規模な画像データで事前学習済み生成モデルには大量な画像特徴情報が含まれ，画像変換や編集タスクに直接的に応用できる．大規模な画像とテキストのペアデータで事前学習済みのimage-textマッチングモデルには，豊富な画像特徴とテキスト特徴が埋め込まれ，それらの特徴を用い，広い範囲の画像変換や編集を実現できる．

自然言語を用いた画像編集において，既存研究では，画像内のオブジェクトの外観特徴（色やテクスチャなどの特徴）に対する編集が主流で，オブジェクトの形状特徴（一部の形状やサイズなどの特徴）に対する編集の研究が少ない，という問題点が残っている．その不足点を踏まえ，本研究の目的は，事前学習済み画像生成モデルとマルチモーダルモデルを利用し，入力テキストに基づいた画像の形状特徴の編集を実現することである．SOTAな画像生成モデルのStyleGAN2 [1] とSOTAなimage-textマッチングモデルのCLIP [2] を結合し，事前学習済みStyleGAN2 [1] モデルのパラメータを調整することにより，画像形状特徴の編集を実現する．

2. 関連研究

2.1 潜在空間と画像編集

学習済みのGANモデルの潜在空間には，豊富な分離されたかつ解釈可能な画像特徴が埋め込まれている．初期GANベースモデルのDCGAN [3] の研究では，潜在空間と生成画像の関係を調査した．入力ベクトルを少しずつ線形的に変えることで，生成画像には段階的な変化を起こすことができることが発見された．近年，StyleGAN [4] などの高性能GANモデルの事前学習済みモデルの潜在空間を利用した画像変換や編集をする研究が多い．その方法は，教師なしと教師ありに分けられる．教師なし手法 [5], [6] は，分析的な手法で潜在空間のもつれを解くかつ解釈可能な変換方向を探すことに注目する．探した変換方向の具体的な変換種類を確認するため，人工的なチェックやアノテーションか必要である．一部の教師あり手法 [7], [8] は，例画像や事前学習済み特徴分類器などの外部ガイドを用い，目標特徴を起こす変換方向を探す．その原因で，外部ガイドは変換方向探索の精度や効率への影響が大きい．最近のテキストガイドを用いた手法 [9], [10] は，CLIPを用いて変換をガイドする．大規模image-textマッチングモデルの外部ガイドを使用することで，範囲が広いかつ柔軟性がある画像編集を実現する．

StyleGANの潜在空間を利用する画像編集の研究について，多くの研究は潜在空間 W または $W+$ に変換方向を探す．他の潜在空間を利用する研究もある．StyleSpace [11] は，潜在空間 W と $W+$ より特徴がもつれを解くのスタイル空間 S に目標特徴を起こす変換方向を探す．StyleCLIPにもその潜在空間 S を利用した手法が提案される．それ以外，一部の研究 [12], [13] は潜在空間の代わりに事前学習済みモデルのパラメータを操作することで目標特徴の編集を実現する．NaviGAN [12] は事前学習済みモデルの重みにシフトを加えることで，出力画像に意味ある

特徴変換を起こす．教師なし手法を用いて意味ある変換を起こすシフトを探し，人工チェックにより具体的な変換種類を確認する．潜在空間でなくモデルのパラメータを操作する原因で，探した変換方向には形状特徴の編集に関する変換が多い．モデルパラメータを調整することにより，従来手法に利用した潜在空間 W などの空間では難しい形状変換を実現できる．しかも，一度最適化でシフトを取得すれば，別の画像もこのシフトで調整されるモデルで変換できるので，大量な画像の高速変換が可能になる．本研究は NaviGAN のアイデアを画像とテキストのマルチモーダルモデルと結合し，テキストからの形状特徴編集を実現する．

2.2 テキストを用いた画像編集

自然言語は GAN に基づいた画像生成や画像編集モデルのインタフェースの一つとして使用される．比較的早期なモデル [14], [15] には，ゼロから学習されるモデルが多い．これらのモデルはゼロから学習されるため，大量な時間とテキスト付け画像データが必要である．かつ，処理できる編集種類も学習データのアノテーションに含まれるものに限られる．近年，大規模データで学習された高性能モデルが現れると共に，事前学習済みモデルを活用する研究も増加している．テキストを用いた画像編集タスクにおいて，事前学習済み画像生成モデルを用いて高品質な画像合成機能を提供し，事前学習済みマルチモーダルモデルで画像編集のガイドを提供する方法がいくつかある．Paint by Word [10] は事前学習済み StyleGAN2 と CLIP を結合し，CLIP でガイドを提供し，潜在空間 W にある入力画像の潜在コード w の編集により，ユーザーが指定するマスク内の部分を編集をする．StyleCLIP [9] も事前学習済み StyleGAN2 と CLIP を利用したモデルである．この研究は三つの手法を提案し，二つの手法は潜在空間 W を利用する手法で，もう一つの手法は StyleSpace [11] に提案された潜在空間 S で適切な変換方向を探す．ゼロからモデルを学習する方法より，事前学習済みモデルを活用する方法は学習コストを減少するだけでなく，大規模データセットで事前学習されたマルチモーダルモデルのおかげで，広い範囲の変換が可能となっている．しかし，今までのテキストを用いた画像編集の研究には，画像の全体または指定された部分の外観特徴を編集する研究が多かったが，形状特徴の編集に力を入れる研究は少ない．本研究は，画像形状特徴の編集に重点を置く．

2.3 GAN 逆マッピング

GAN モデルは潜在ベクトルから画像を合成するため，GAN モデルを用いてリアル画像の入力を画像編集などの処理をするのは，入力画像を GAN モデルの潜在空間に逆マッピングするモジュールが必要である．GAN 逆マッピングの目標は，入力画像を GAN 潜在空間の潜在コードにエンコードすることで，その潜在コードを用いて入力画像になるべく近い画像を再構築できる．

GAN 逆マッピングの手法は最適化，エンコーダー，ハイブリッドに分けられる．Image2StyleGAN [16] は最適化のフレームワークを使い，入力画像を StyleGAN の潜在空間 W にエンコードする．そのロス関数は逆マッピングで得た潜在コードからの合成画像と入力画像の perceptual loss と pixel-wise MSE loss を使用する．Encoder4Editing (e4e) [17] は一つのエンコーダーを提案し，そのエンコーダーで入力画像を潜在空間 W にエンコードする．画像編集タスクに相応しいかつ画像品質を保つ潜在コードを得るため，このエンコーダーは潜在空間 W の分布に近い潜在コードを得ることを目標とする．

本研究はエンコーダーに基づいた方法の e4e [17] を使用し，入力画像を StyleGAN2 生成器の潜在空間に逆マッピングする．

3. 手　　法

NaviGAN [12] の「生成器のパラメータを調整する」アイデアに基づいて，事前学習済み StyleGAN2 生成器を利用し，提案モデルを構築する．一つのシフト x で，事前学習済み生成器のパラメータを調整する．CLIP により，生成画像と入力テキストの類似程度を測定し，最適化のためのロスを求める．モデルの概要を図 1 に示す．

3.1 利用するモデル

StyleGAN [4] とその拡張版の StyleGAN2 [1] は，異なるスタイルベクトルを生成器の各畳み込み層に入力することにより，高い品質の画像を生成できる．StyleGAN はマッピングネットワークと合成ネットワークの二つの部分で構成される．StyleGAN で画像を合成する全体像は，正規分布に従う入力ベクトル z をマッピングネットワークで潜在空間 W の潜在ベクトル w にマッピングし，潜在ベクトル w を生成器の各畳み込み層に送って画像を合成する．合成ネットワークの各レイヤーは異なる特徴を制御するので，一部の手法は入力ベクトル z または潜在コード w を調整することにより，画像のセマンティック編集をする．

しかし，画像の形状特徴の編集について，先行研究 [12] では潜在コード w に対する調整では形状特徴の編集を達成できないことが示されている．そこで本研究では，NaviGAN [12] の方法に従って，StyleGAN2 の畳み込みレイヤーのパラメータを調整することで，画像の形状特徴の編集を実現する．StyleGAN を使用した画像編集タスクでは，入力画像を StyleGAN の潜在空間にマッピングできるエンコーダモジュールが必要である．本研究は事前学習済み StyleGAN2 モデルを画像生成モデルとして利用し，入力画像のエンコーダには encoder4editing(e4e) [17] を使用する．

最近リリースされた Contrastive Language-Image Pre-training (CLIP) [2] は画像とテキストの共有空間を学習し，画像とテキストのマッチング度を測定するマルチモーダルモデルである．CLIP は一つのテキストエンコーダーと一つのイメージエンコーダーで構成され，入力画像と入力テキストをそれぞれ共有空間にエンコードし，それらの類似程度を計算する．インターネットから収集された 4 億の画像とテキストのペアデータで学習され，CLIP 空間には大量の画像特徴とテキスト特徴が埋め込まれるため，CLIP は SOTA な image-text マッチング性能を持つ．

テキストを利用して画像合成や画像編集をするタスクにおいて，StyleCLIP など，その image-text マッチング性能を用いて画像合成や編集をガイドするモデルはいくつかある．従来の StyleGAN と CLIP を利用した画像編集モデルは，StyleGAN の

図 1　モデル概要

潜在空間に潜在コードを調整することにより画像編集を実現する．それに対して，本研究は潜在空間の潜在コードではなく，事前学習済み StyleGAN のパラメータを調整することで，画像編集を実現する．本研究は CLIP で生成画像と入力テキストの類似程度を測定し，変換をガイドする．

3.2　モデルアーキテクチャ

提案モデルの入力は入力画像 I（または潜在コード w）と入力テキスト T，出力画像は I' とする．入力が画像の場合，その画像 I を事前学習された逆マッピングモジュール e4e [17] に入力し，対応する潜在コード w を取得する．生成器 G は事前学習済みの StyleGAN2 の生成器を使う．目標は，テキスト T のセマンティック意味に合うように I を編集し，編集された画像 I' を出力することである．シフト x は正規分布に従ってランダムに初期化され，生成器のパラメータを操作する．シフトされた生成器から出力された画像 I' は入力テキストと一緒に CLIP に送られ，出力画像と入力テキストの類似程度を測定し，最適化のためのフィードバックを提供する．CLIP のロスとシフト x の L2 ロスを用い，シフト x を最適化する．

一回の最適化で，シフト x は生成器 G の一つの畳み込みレイヤーのパラメータを調整し，一つの目標特徴の編集するシフトを得る．レイヤーの選定について，一定範囲のレイヤーを選び，それらのレイヤーの生成結果を観察し，目標レイヤーを決める．また，シフト x だけに対して最適化し，ほかの部分のパラメータは固定される．

画像特徴編集の従来研究に比べて，この研究の特徴は次の通りとなる．

- 多くの従来研究は画像外観特徴の編集に注目するが，この研究は画像形状特徴の編集に重点を置く．
- 事前学習済み GAN モデルを利用した画像特徴編集において，従来研究には GAN 潜在空間の操作により画像編集を実現することが多い．この研究は，事前学習済みモデルのパラメータを調整することにより，画像特徴を編集する．

3.3　損 失 関 数

提案モデルに用いた損失関数には，CLIP matching loss の L_{clip} と L2 loss の L_2 を使い，シフト x を最適化する．損失関数は式 1 で，D_{clip} は CLIP により計算される二つの入力の cosine similarity である．

CLIP matching loss について，生成画像 I' と入力テキスト T のセマンティック意味に合うように，画像 I' とテキスト T を CLIP モデルに入力し，それらの cosine similarity を計算する．L2 loss について，目標特徴を編集しながら元画像の関係なし特徴を最大限に維持するため，目標特徴の編集を満足する上に，編集を最小限にする．ロス関数には，シフト x の L_2 ノルムを最小化するロスを使用する．

目標特徴の編集と関係なし特徴の維持の間にはトレードオフがあり，λ_2 を大きくすると特徴編集の効果が抑えられ，λ_1 を大きくすると画像が歪んでしまうことがある．実験に用いたハイパーパラメータ値は $\lambda_1 = 10$ で，λ_2 の値は変換種類により 0.01-0.1 の範囲に最適な数値を選択する．

$$\begin{aligned} L &= \lambda_1 L_{clip} + \lambda_2 L_2 \\ &= \lambda_1 D_{clip}(I', T) + \lambda_2 ||x||_2 \end{aligned} \tag{1}$$

4.　実　　験

4.1　実 験 配 置

StyleGAN2 の事前学習に用いたデータセットは，FFHQ [4]，LSUN-Car [18]，LSUN-Horse [18] である．それぞれの解像度は，1024x1024，512x384，256x256 である．画像逆マッピングモジュールの e4e [17] も対応なデータセットで事前学習されたモデルを利用する．一枚の画像と目標テイストのペアに対して，一つの変換のシフトを Adam アルゴリズム [19] で 100 ステップの最適化するのはおよそ 1 分 30 秒かかる．使用する GPU は一つの NVIDIA GeForce GTX 1080 Ti である．損失関数のハイパーパラメータ値は $\lambda_1 = 10$ で，λ_2 の値は変換種類により 0.01-0.1 の範囲に複数の値で変換し，最適な数値を選択する．一般的には，顔の幅さなど全体的な特徴を編集するのは λ_2 の値は 0.02-0.03 で，目の編集など細かい部分を編集するのは λ_2 の値は 0.07-0.08 とする．

4.2　ビジュアル効果

画像内のオブジェクトの一部のサイズを強度を調整しながら変換する形状特徴編集について，いくつかの変換結果を図 2 に示す．変換程度による変化を示すために，最適化で得たシフト x に-3 から 3 までの係数をかけ，編集された画像を生成する．

外観特徴の編集において，画像の一部の色やテクスチャの変換は，他の部分の外観や形状に影響を及ばない．それに対して，形状特徴の編集では，一つの部分に対する編集は他の部分に影

響を与える可能性がある．鼻の長さの変換には，鼻が長くなると口の位置も下に移動する．車輪のサイズが大きくすると，他の部分も変化し，車全体の比例に変化を起こす．提案手法はそのような全体的な変化を起こさずに，目的特徴のみの編集を達成できる．その一方で，一部の変換には不自然なパターンが現れる，または関係なし特徴が編集されてしまうこともある．

一部の変換（a wide car など）には，変換の強さがマイナスになると目標特徴の効果が現れるという現象がある．同じ部分に関する変換，図 3 の示すように，a small-wheel car と a big-wheel car という形容詞だけが異なるの二つの変換は，同じレイヤーでシフトを最適化すれば，それらの目標が逆であるが，結果の傾向が同じになった．

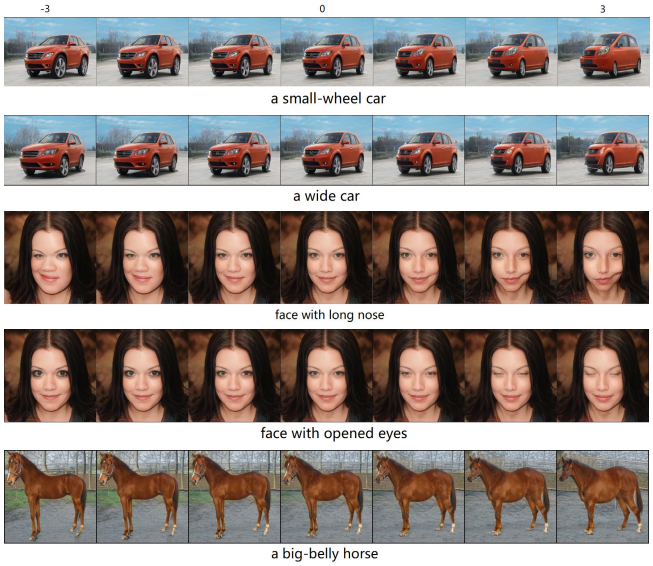

図 2　変換効果の例

図 3　形容詞だけが異なるの二つの変換

4.3　比 較 実 験

テキストを用いた画像の形状特徴編集のタスクにおいて，提案モデルと同じように事前学習済み StyleGAN2 と CLIP を利用した手法の StyleCLIP [9] と比較実験をした．StyleCLIP は潜在空間での潜在コードに対して調整をすることにより画像編集をするモデルである．StyleCLIP は三つの手法を提案する．第一の手法 latent optimization は潜在空間 W に潜在コードを直接的に最適化する手法である．第二の手法 latent mapper も潜在空間 W に潜在コードを調整するが，それは一つのマッピングモジュールを学習して潜在コードを変換する手法である．第三の手法 global directions は StyleGAN2 の潜在空間 S（StyleSpace [11] で提案された潜在空間で，位置は StyleGAN2 モデルの affine transformation と modulation の間）を利用して編集をする手法である．潜在空間 S を利用した手法は，より精度の高い変換が可能である．

StyleCLIP と提案手法の主な区別は，StyleCLIP は StyleGAN2 の潜在空間に対する調整で，提案手法は StyleGAN2 のパラメータに対する調整，という点である．各データセットに二つの変換を選択し，二つのモデルを用いて比較実験を行った．結果を図 4 に示す．

図 4　比較実験の結果

実験結果によると，StyleCLIP の潜在空間 W で調整をする二つの方法は形状特徴の編集をほぼ達成できなかった．Global directions の方法は一部の形状特徴編集を達成できるが，目標特徴の編集と関係ないスタイルやテクスチャの変換を起こしてしまうことがある．顔画像の鼻の長さの変換では，StyleCLIP より提案手法が目標特徴の変換を達成できた．車のマーカーのような小さい部分の変換にも，提案手法が目標特徴の変換を達成できた．車輪のサイズ変換と馬の腹の変換では，StyleCLIP による変換された画像は，スタイルやテクスチャを変えてしまうことがあった．しかし，提案手法も同様に，鼻の長さを変換しながら口の形状も変えてしまうように，目標とする特徴とは関係なく大きな変換をすることがあった．

4.4　定 量 評 価

提案モデルにより得られる画像の品質を示すため，定量評価を行った．使用する指標は Fréchet Inception Distance (FID) [20] である．

まずは，元データセットに比べて提案手法による生成画像の品質は大幅に下がることがないことを示すため，大量な結果画像を生成し，FID でベースラインセットとテストセットの距離を測定し，定量評価をした．ベースラインセットは 3000 枚のリアル画像を用いる．テストセットについて，まずは一つの画像に対して最適化をして，一つの変換のシフト x を得る．そしてこのシフトを事前学習済みモデルに適用し，3000 枚の真実画

像を逆マッピングして変換し，テストセットの 3000 枚画像を得る．

使用する変換種類は wheel size と cheek size である．それぞれの変換に対してシフトを最適化し，事前学習済みモデルに適用してテストセットを得る．しかも，シフトに係数をかけて，異なる変換の幅に対してそれぞれの FID 値を測定する．実験に使用した変換の幅は ±3，±5，±10 で，それらの幅により得た画像の例は図 5 に示す．実験結果は表 1 に載せる．逆マッピングで得られた画像は完全にリアル画像と一致することができないため，逆マッピングで再構築された画像セットの FID 値も載せ，変換幅は 0 とする．

次は，StyleCLIP に対して定量評価をし，提案モデルと比較する．前と同じように，3000 枚のベースラインセットのリアル画像，と StyleCLIP で変換された 3000 枚の画像の FID 値を測定した．ベースラインセットは前と同じセットを使い，テストセットは StyleCLIP の global direction 方法で，なるべく提案手法の対応な幅と同じ効果になるようにパラメータを選択して得られる．ここでは，前の変換幅に対応する実験を行った．StyleCLIP による生成画像の例は図 5 に示す．

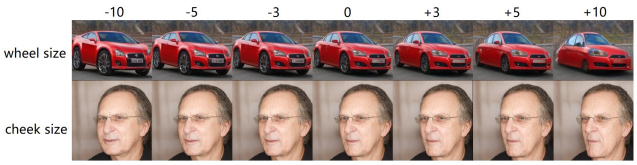

図 5　提案モデルによる異なる幅の生成画像例

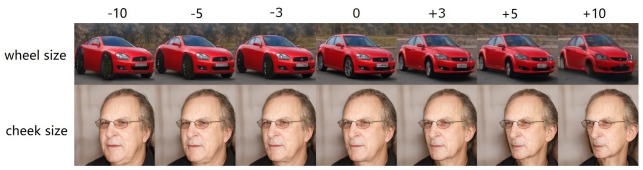

図 6　StyleCLIP による異なる幅の生成画像例

表 1 に，提案モデルと StyleCLIP による定量評価の結果を載せる．多くの場合には，提案手法がより良い FID 値を取得した．タイヤのサイズ変換には，提案手法がタイヤのテクスチャを保つことができた．チークのサイズ変換には，提案手法が StyleCLIP と同様の結果を得た．StyleCLIP より，提案モデルの変換結果が画像品質を保つことができると証明された．

4.5　複数種類の編集

提案手法には，一回の最適化に，一種類の変換だけのシフトを得られる．しかし，複数の最適化されたシフトを同時にモデルに適用することにより，複数の変換をする画像を得られる．つまり，複数のシフトで起こす変換効果を一つの画像に適用する．

複数種類の編集の実験を行った．二つの変換のシフトをそれぞれ最適化して取得し，同時にモデルに加える．二つのシフトに対応するレイヤーが違うとき，それぞれのシフトを対応するレイヤーに加える．二つのシフトが同じレイヤーに対応する場合，それらを線形的に加えてモデルに適用する．違うレイヤーの変換の結果画像は図 7，同じレイヤーの変換の結果画像は図 8 に示す．

表 1　定量評価の結果

		幅	FID	幅	FID
Wheel Size	StyleCLIP	-3	33.36	+3	18.16
		-5	42.35	+5	23.33
		-10	54.34	+10	67.57
	提案手法	-3	**15.34**	+3	**15.22**
		-5	**17.96**	+5	**21.30**
		-10	**26.27**	+10	**62.39**
	逆マッピング	0	12.54		
Cheek Size	StyleCLIP	-3	28.90	+3	28.40
		-5	**29.16**	+5	29.20
		-10	30.55	+10	30.12
	提案手法	-3	**28.89**	+3	**27.96**
		-5	29.34	+5	**28.41**
		-10	**30.21**	+10	**29.86**
	逆マッピング	0	25.6		

提案モデルにより最適化で得たシフトは，完全にもつれを解く変換を得られないことがある．シフトを用いて目標特徴を変換しながら，関係なし特徴を変えてしまうことがある．そこで，二つの種類の変換のシフトを同時に適用するとき，それぞれのシフトによる変換効果が互いに干渉することがある．顔画像に対して opened eyes と short nose の二つ変換をするとき，この二つのシフトによる効果が互いに関係ないため，結果画像には二つの目標変換を同時にできる．車画像に対して small marker と small wheel の変換では，small marker の変換のシフトがマーカーを小さくすると同時に，車輪を大きくしてしまう．それは small wheel の変換効果と中和し，車輪のサイズが中間的なサイズになった．

5.　お わ り に

本研究では，事前学習済み StyleGAN2 の生成器を利用し，生成器のパラメータを調整する方法により，マルチモーダルモデル CLIP のガイドによりテキストからの画像形状特徴の編集を実現した．実験結果から見ると，従来の潜在空間を調整をする方法に比べて，形状特徴の編集において提案手法の表現が上回った．提案手法は，従来の潜在コード調整の方法より目標特徴の変換を達成でき，定量評価により変換後の画像品質を保つことをできると証明された．また，複数の最適化されたシフトを同時にモデルに適用することにより，複数特徴の編集を実現できた．

提案モデルでは，複数の事前最適化されたシフトを同時にモデルに適用することで複数特徴の編集を実現できるが，一回の最適化には複数の特徴を編集できるシフトを学習することができない．そして，提案手法は一部のサイズ変換による他の部分に対する影響を処理することができるが，最適化で得たシフトは完全にもつれを解くものではないため，目標特徴を編集するとき，目標特徴以外の部分を変えてしまうことがある．今後の

図 7　違うレイヤーの二つの変換を同時に応用する

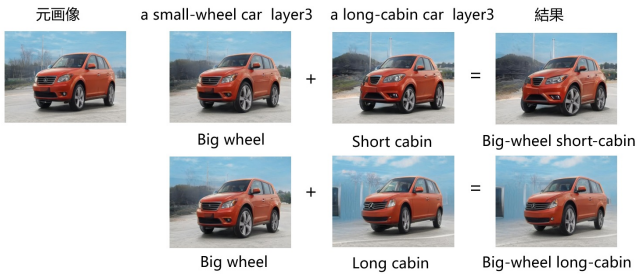

図 8　同じレイヤーの二つの変換を同時に応用する

課題として，複数の特徴変換を処理でき，かつ関係の無い特徴への影響を最小化する方法を考える．

文　献

[1] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[2] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

[3] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *International Conference on Learning Representations*, 2016.

[4] 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*, 2019.

[5] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. In *Advances in Neural Information Processing Systems*, 2020.

[6] Yujun Shen and Bolei Zhou. Closed-form factorization of latent semantics in gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1532–1540, 2021.

[7] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9243–9252, 2020.

[8] Rameen Abdal, Peihao Zhu, Niloy J Mitra, and Peter Wonka. Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. *ACM Transactions on Graphics (TOG)*, Vol. 40, No. 3, pp. 1–21, 2021.

[9] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2085–2094, 2021.

[10] David Bau, Alex Andonian, Audrey Cui, YeonHwan Park, Ali Jahanian, Aude Oliva, and Antonio Torralba. Paint by word. *arXiv preprint arXiv:2103.10951*, 2021.

[11] Zongze Wu, Dani Lischinski, and Eli Shechtman. Stylespace analysis: Disentangled controls for stylegan image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12863–12872, 2021.

[12] Anton Cherepkov, Andrey Voynov, and Artem Babenko. Navigating the gan parameter space for semantic image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[13] David Bau, Steven Liu, Tongzhou Wang, Jun-Yan Zhu, and Antonio Torralba. Rewriting a deep generative model. In *Proceedings of the European Conference on Computer Vision*, 2020.

[14] Hao Dong, Simiao Yu, Chao Wu, and Yike Guo. Semantic image synthesis via adversarial learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5706–5714, 2017.

[15] Seonghyeon Nam, Yunji Kim, and Seon Joo Kim. Text-adaptive generative adversarial networks: Manipulating images with natural language. *arXiv preprint arXiv:1810.11919*, 2018.

[16] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4432–4441, 2019.

[17] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. *ACM Transactions on Graphics (TOG)*, Vol. 40, No. 4, pp. 1–14, 2021.

[18] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

[19] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[20] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems*, Vol. 30, , 2017.