H-029

StyleCLIP とセマンティックセグメンテーションを利用した背景画像の編集についての検討
Study of editing background images using StyleCLIP and semantic segmentation

石幡 柊介[1)]　折原 良平[1)]　清 雄一[1)]　田原 康之[1)]　大須賀 昭彦[1)]
Syuusuke Ishihata　Ryohei Orihara　Yuichi Sei　Yasuyuki Tahara　Akihiko Ohsuga

1　はじめに

StyleGAN[7] という教師なし学習の GAN モデルは，高品質な画像の生成ができ，画像間の補間性能に優れている．この補間性能を利用して StyleGAN を画像編集タスクに利用する研究事例が増えている．例えば，StyleCLIP は，自然言語テキストを入力して学習済み StyleGAN の潜在変数をテキストの内容に即した画像に編集する研究である．この StyleCLIP は StyleGAN の研究でよく用いられる顔画像だけでなく背景の画像でも応用可能である．しかし，背景には山や森のような屋外のものもあれば，部屋やビルの中，スタジアムなどの屋内も存在する．このように背景画像は顔画像のデータよりも多様性が高いため，事前学習済み StyleGAN の生成画像の品質が落ちてしまう．また，StyleCLIP によって画像のスタイルを編集することが可能であるが，画像のコンテントの指定をテキストのみで行う場合，直感的な編集をすることが難しい．これは，画像のコンテントをテキストで細かく指定したとしても，わずかに位置がずれている，物体間の前後関係が違っているなど編集者の意図とは異なる編集結果になってしまう場合があるためである．そのため，StyleCLIP では任意のテキストだけでは自由な背景の編集が困難である．それに対し，セマンティックセグメンテーションは編集者の意図するコンテントを視覚的に表現できる．

GAN Inversion とは，入力画像を Generator が再構成するような潜在変数を推定するタスクであり，推定した潜在変数を変更することによって画像を編集することが可能である．GAN Inversion タスクの Encoder ベースの手法の中には HyperStyle や HyperInverter [15][16] のように，Generator のパラメータを修正して画像の再構成品質と編集性能の両立を目指した手法がある．

本研究では，HyperStyle と StyleTransformer[17] という GAN Inversion の Encoder を組み合わせ，入力にセマンティックセグメンテーションを追加する手法を提案する．この手法で画像編集を試みたところ，背景画像のスタイル，コンテントを分離して編集することができたことを報告する．また，StyleCLIP を応用して新たに Text Encoder を追加し，その出力の特徴量を潜在変数に加えるテキストによる画像編集手法でテキストによる背景画像のスタイル編集ができることを確認した．

2　関連研究

2.1　StyleCLIP

StyleCLIP は StyleGAN の表現力を活かして，テキストで画像を編集する手法の一つである．これは，StyleGAN に加えて OpenAI が提案した CLIP[3] を損失関数に用いることが特徴である．CLIP は事前学習済みの Text Encoder と Image Encoder を用意し，これらを用いて自然言語のテキストと画像との関係を学習するマルチモーダルの画像分類モデルである．これはテキストと画像，それぞれの特徴量間のコサイン類似度に基づいており，その値が大きいほど画像がテキストの内容に適していることを意味する．CLIP を損失関数に用いることで，テキストの内容に即した画像編集を行うことができる．

本研究では，StyleCLIP の 1 アプローチである Mapping Network の学習法 (Latent Mapper) を使用しており，そのモデルのアーキテクチャを図 1 に記す． この Mapper は

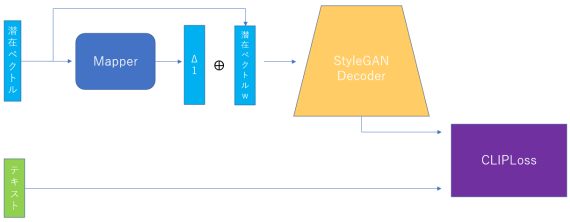

図 1　StyleCLIP のモデル

レイヤーを 3 つのグループに分割しており，それぞれが画像全体の粗いスタイル，中間レベルのスタイル，画像の細かいスタイルの 3 つになっている．この各グループの出力を結合したものと元の潜在変数を足し合わせて潜在変数を編集する．

編集された潜在変数 w' は次の式で定式化される．

$$w' = w + \lambda \Delta_1 \tag{1}$$

Δ_1 は Mapper の出力であり，λ はその重みである．

2.2　GAN Inversion

実画像から，GAN の Generator でその画像を再現できるような潜在変数を推定することを GAN Inversion という．このような手法には潜在変数を直接最適化する手法 [12] や，画像を直接潜在ベクトルに Encode する手法などがある [13][17]．直接最適化する手法では，再構成品質は高いが，最適化に時間がかかってしまう．それに対して，Encoder を使う手法は推定時間が早いが，再構成品質は低い傾向にある．この GAN Inversion，特に Encoder ベースの手法に対して，Generator のパラメータを Hypernetwork[11] で更新することで生成画像の再構成品質を向上させる手法も存在する [15][16]．本手法では HyperStyle を使用し，そのモデルは図 2 のようになる．

GAN Inversion の Encoder モデルによって再構成した画像と元の入力画像を HyperNetwork H に入力する．その出力 Δ から式 (2) によってパラメータ $\hat{\theta}$ を求める．その $\hat{\theta}$ で Generator を修正する．$\hat{\theta}_l^{i,j}$ は 1 層目の Generator

1) 電気通信大学大学院情報理工学研究科　情報学専攻　University of Electro-Communications Graduate School of Informatics and Engineering Department of Informatics

に関する畳み込み層の i 番目のフィルタの j チャネルに関する重みである．

$$\hat{\theta}_l^{i,j} = \theta_l^{i,j}(1 + \Delta_l^{i,j}) \tag{2}$$

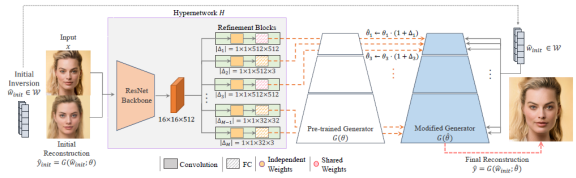

図 2　HyperStyle のモデルアーキテクチャ ([15]p3，図 2 より引用)

3　手法

3.1　スタイル・コンテントの編集モデル

画像のスタイル，コンテントの編集をするための GAN Inversion モデルの概要は図 3 のようになる．

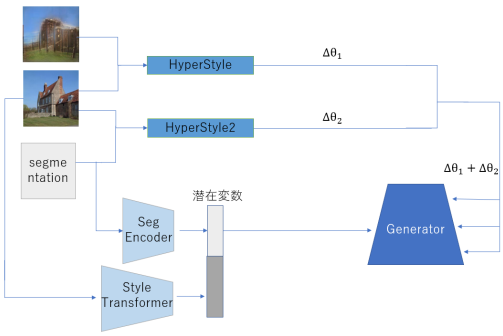

図 3　スタイル・コンテントの編集のモデル

2 つの HyperStyle を用意し，それぞれの出力である残差パラメータを学習済みの Generator のパラメータに加える．HyperStyle による各パラメータの更新の式は (2) と同様である．Generator の畳み込み層が低い部分にセマンティックセグメンテーションの Encoder(Seg Encoder) の出力という，画像のコンテントにかかわる情報を用いることで，画像の形を制御できるのではないかと考えた．そこで画像の Encoder の出力となる潜在変数の解像度が低〜中となる部分を Seg Encoder の出力に置換した．画像の Encoder である StyleTransformer[17] の出力は本来，各層に入力する潜在変数がそれぞれ異なる w+空間 ($\mathbb{R}^{n \times 512}$) の点である．この n は Generator の層の数である．しかし，HyperStyle では編集性能と再構成品質の両立を図るために，潜在変数については再構成品質は w+空間よりも低いが編集性能が高いとされている w 空間 ($\mathbb{R}^{512}$) を前提にしている [15]．この w 空間は各層に入力する潜在変数が同一のものとなる．今回は，StyleTransformer の出力の次元を調整して $w \in \mathbb{R}^{1 \times 512}$ とした．

3.2　テキストによる潜在変数の編集

テキストによるスタイル編集部分のモデルは StyleCLIP をベースにしており，潜在変数 w を Mapper に入力し，その出力 Δ_1 を w との演算に用いる．この Δ_1 は式 (1) と同じである．既存の StyleCLIP との大きな違いは，どのような編集をするのかを指定するテキストを Encode する Text Encoder の存在である．StyleCLIP では潜在変数の編集でテキストの特徴をそのまま使用せず，損失関数で使用していた．今回の目的のように背景そのものを提案する手法では Text Encoder を用いて，その出力，μ，σ から正規分布に沿ってランダムベクトル Δ_2 をサンプリングする．Text Encoder と StyleCLIP の Mapper を利用した潜在変数の編集は，StyleGAN の画像編集では潜在変数にある値を足し合わせる方法を主に用いていることから Δ_1，Δ_2 を線形補間によって足し合わせる．この手法で編集された潜在空間 w' は次の式で定式化される．

$$w' = w + \alpha\Delta_1 + (1 - \alpha)\Delta_2 \tag{3}$$

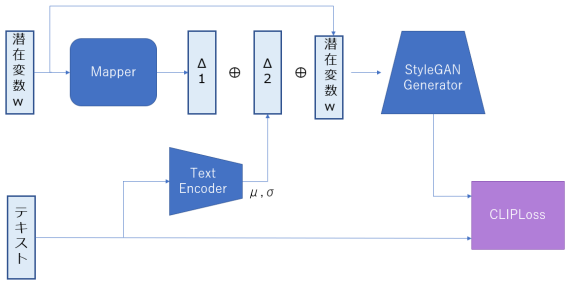

図 4　テキストによるスタイル編集のモデル

3.3　損失関数

コンテント，スタイル編集モデルの損失関数は，HyperStyle のものと同様である．テキストによるスタイル編集モデルでは，Mapper と Text Encoder のパラメータを更新させている．StyleGAN の再構成性能を活かすために Generator は事前学習済みのモデルを扱うので，Generator のパラメータは更新しない．

データにはない背景の画像の編集において Discriminator を使用しないため，通常の GAN で用いられる adversarial loss を使用しない．全体のモデルの損失関数 L は以下の式 (4) で表す．そして，この L を最小化するように Mapper と Text Encoder を学習する．

$$L = \lambda_{clip}L_{clip}(t, \hat{y}) + \lambda_{sim}L_{sim}(x, y, \hat{y}) \tag{4}$$

L_{clip} は潜在空間で生成した画像がテキストに即するように，1 からコサイン類似度を引いたものとなる．

L_{sim} としては，顔画像を生成・編集するタスクの場合，アイデンティティベースの顔認識モデルを利用した類似損失を使用するケースが多い．しかし本研究では背景画像の編集を中心に行うため，MoCo ベースの類似損失を用いる [13][14].

4　実験

4.1　概要

事前学習済みの StyleGAN2[8] を用いて，HyperStyle を利用した画像の編集がどのようにできるのかの実験とテキストを利用したスタイルの編集を行った．

本論文では，StyleTransformer を GAN Inversion の Encoder として使用し，Generator のパラメータの更新を HyperStyle で行っている．また，定量的な評価では HyperInverter を使用した場合との比較を行った．

図 5　再構成画像の結果

本研究では，ADE20K データセット [2] で学習したものを使用した．式 (4) については $\lambda_{clip} = 0.9$，$\lambda_{sim} = 0.1$ とした．元の画像と編集画像はどちらとも 256 × 256 の解像度で学習を行った．

4.2　GAN Inversion の定性評価

定性的な評価では背景画像における StyleTransformer を Encoder とした HyperStyle の再構成品質の結果 (図 5)，元画像のコンテントを残しつつスタイルのみをミキシングした時の結果 (図 6)，元画像のスタイルを残しつつコンテントのみをミキシングした時の結果 (図 7) の 3 点によって評価する．

本実験では，スタイルのみをミキシングするときは潜在変数の高解像度部分を入れ替える．また，元画像のスタイルを残しつつコンテントのみをミキシングする場合は，Seg Encoder の出力と残差パラメータを置換する．

図 5 の再構成品質の結果について，2，3 列目で比較してみると，提案手法のほうが StyleTransformer のみよりも元の入力画像により近い Inversion ができていた．図 6 では，一番左の列の画像をベースに一番上の行の画像のスタイルをミキシングしている．元の画像のコンテントを残しつつ，スタイルだけが変化していることがわかる．一方で，図 7 では，元の画像のスタイルを残しつつ，コンテントだけが変化していることがわかる．

4.3　テキストによるスタイル編集の結果

StyleCLIP を利用したテキストによる再構成画像に対するスタイル編集の結果と，モデルの構成要素である Text Encoder に対する，Ablation Study を図 8 に記す．Text Encoder の有無による編集結果の画像の変化はあまりみられなかった．

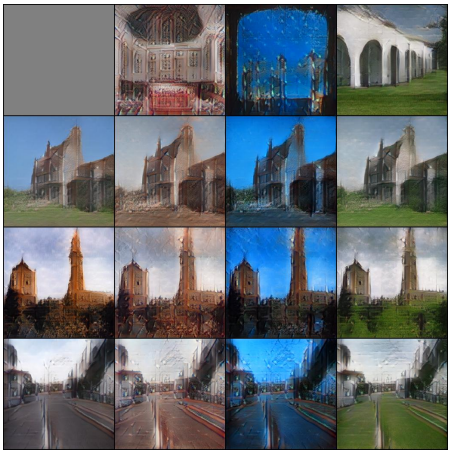

図 6　スタイルの編集結果 (1 行目，1 列目が入力画像．各列の画像のスタイルを 1 行目の画像のスタイルに編集)

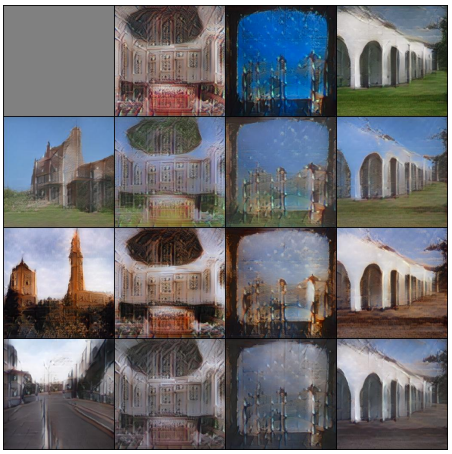

図 7　コンテントの編集結果 (1 行目，1 列目が入力画像．各列の画像のコンテントを 1 行目の画像のコンテントに編集)

4.4　定量的評価

本実験では，HyperInverter と同様の指標を用いて定量的な評価を行った [16]．これは，損失関数の L2 と LPIPS[18] に加え，FID[5] や KID[20], PSNR, MS-SSIM[19] による画像の再構成品質の評価である．その結果を表 1 に記す．

5　考察

従来の研究では，背景画像のように多様性が高いドメインに対して画像のスタイルとコンテント，それぞれを分離して編集することが困難であった．本手法では図 6，7 から，画像のスタイルとコンテントそれぞれが独立して編集ができていることがわかる．スタイルとコンテントどちらか一方を編集することができ，従来の手法よりも自由な編集ができるというメリットがあると考える．

しかし 4.4 節から，セマンティックセグメンテーションを導入せず，StyleTransformer と HyperStyle を組み合わせた手法のほうが定量的に良くなる傾向になった．これは，HyperStyle2 の出力 $\Delta\theta_2$ が影響していると考える．パラメータの変化量が増えたために Generator のパラメータの細かい調整ができなかったと考える．

テキストによるスタイルの編集では，Text Encoder

表 1　セマンティックセグメンテーションの有無によるの定量的な比較結果

手法	L2(↓)	LPIPS(↓)	FID(↓)	KID(× 10^{-3})(↓)	PSNR(↑)	MS-SSIM(↑)
提案手法	0.06412	0.27514	**44.97**	**18.36959**	18.25542	0.57358
StyleTransformer+HyperStyle	**0.05276**	**0.22515**	47.50	19.98976	**19.11787**	**0.64569**

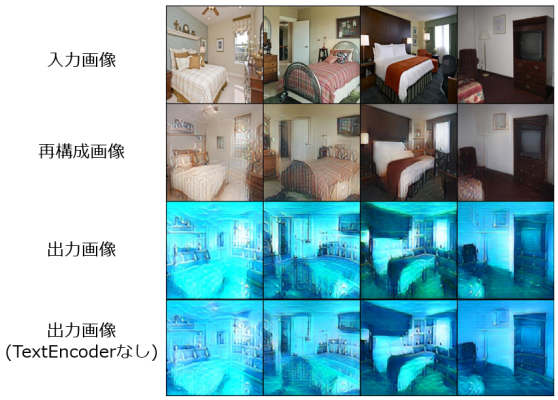

図 8　テキスト編集の結果 (word:”ocean”)

の有無で大きな変化が見られなかった．これは，Text Encoder で出力される値が小さく，潜在変数の変化が少なかったためにこのような結果になったと考える．

6　まとめと今後の展望

HyperNetwork を用いた Generator のパラメータを更新する手法を応用し，新たにセマンティックセグメンテーションを利用した GAN Inversion を提案した．また，この手法を利用することによって背景画像のスタイル，コンテントを分離した編集ができたことを確認し，テキストを用いて，画像のスタイルを編集することができた．

今後の展望として，第一に GAN Inversion の Encoder 自体の再構成品質の向上が挙げられる．HyperStyle を利用することによって品質の向上はできているが，画像の再構成品質の向上は限定的である (図 5 の 4 行目)．これは Encoder による再構成品質が十分ではないためであると考える．そのため Encoder の品質を向上する手法を模索したいと考えている．第二に，さらなるコンテント編集性能の向上である．画像のコンテントは Seg Encoder と Hyperstyle の出力である残差パラメータによってミキシングが実現できた．しかし，編集したい画像とは別に画像が必要なため，編集の制約がある．今後は，セマンティックセグメンテーションのみでコンテントの編集を制御できる手法を模索し，より直感的で自由な背景画像の編集に着手したいと考えている．

謝辞

本研究は JSPS 科研費 JP21H03496, JP22K12157 の助成を受けたものです．

参考文献

[1] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, DavidWarde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. Neural Information Processing Systems (NIPS), 2014.

[2] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba. Scene parsing through ade20k dataset. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[3] 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.

[4] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. arXiv preprint arXiv:2103.17249, 2021.

[5] Martin Heusel and Hubert Ramsauer and Thomas Unterthiner and Bernhard Nessler and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. Neural Information Processing Systems (NIPS), 2017

[6] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. CoRR, abs/1710.10196, 2017.

[7] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In Proc. CVPR, pages 4401〜4410, 2019.

[8] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten,Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In Proc. CVPR, pages 8110〜8119, 2020.

[9] Xun Huang and Serge J. Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. CoRR, abs/1703.06868, 2017

[10] Hongwei Yong, Jianqiang Huang, Xiansheng Hua, and Lei Zhang. Gradient-centralization: A new optimization technique for deep neural networks. 2020.

[11] David Ha, Andrew Dai, and Quoc V Le. Hypernetworks. arXiv preprint arXiv:1609.09106, 2016.

[12] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In Proceedings of the IEEE international conference on computer vision, 2019.

[13] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021.

[14] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation, 2021.

[15] Yuval Alaluf, Omer Tov, Ron Mokady, Rinon Gal, and Amit H. Bermano. Hyperstyle:Stylegan inversion with hypernetworks for real image editing,2021.

[16] Tan M. Dinh, Anh Tuan Tran, Rang Nguyen, and Binh-Son Hua. Hyperinverter: Improving stylegan inversion via hypernetwork. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022.

[17] Xueqi Hu, Qiusheng Huang, Zhengyi Shi, Siyuan Li, Changxin Gao, Li Sun, and Qingli Li. Style transformer for image inversion and editing. arXiv preprint arXiv:2203.07932, 2022.

[18] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In CVPR, 2018.

[19] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003.

[20] Mikołaj Binkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. arXiv preprint arXiv:1801.01401, 2018.