H-030

StyleMap を用いた GAN 反転による画像編集と評価
Image Editing and Evaluation by GAN Inversion using StyleMap

本田 爽[1)] 折原 良平[1)] 清 雄一[1)] 田原 康之[1)] 大須賀 昭彦[1)]
So Honda Ryohei Orihara Yuichi Sei Yasuyuki Tahara Akihiko Ohsuga

1 はじめに

Generative Adversarial Networks(GAN)[1] は画像を生成する Generator と，入力された画像が真の画像か生成画像かを判別する Discriminator からなる生成モデルであり，これら 2 つのネットワークを交互に学習することで，品質の高い画像を生成することができる．ノイズを入力に画像を生成するモデルの最先端である StyleGAN[2, 3] は，高解像度かつ高品質な画像を生成でき，ノイズ $\mathbf{z}$ を潜在変数 $\mathbf{w}$ に変換してから用いることで画像の属性の制御を可能にした．これらの性質を画像編集に利用したのが GAN 反転である．GAN 反転は入力画像を再現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編集することで画像を編集することができる．GAN 反転において入力画像の再現度である再構成品質は高いほうが望ましいが，画像全体の性質を 512 次元の潜在変数で表すため，正確な再構成はいまだ困難である．

本研究では，StyleGAN の潜在変数を空間方向に拡張した StyleMap を用いて GAN 反転を行うことで，再構成品質の向上を図った．

2 関連研究

2.1 StyleGAN

Karras ら [2] は Style 変換の分野で用いられていた AdaIN を画像生成に取り入れた StyleGAN を提案した．AdaIN は中間特徴マップ $\mathbf{x}_i$ を正規化したあと潜在変数 $\mathbf{w}$ によってスケーリングすることで所望の平均・分散に変換する操作である．

StyleGAN2[3] では，StyleGAN に特有のアーティファクトの原因が AdaIN であることから，中間特徴マップではなく畳み込み重みを正規化・スケーリングする Weight Demodulation が用いられている．Weight Demodulation は式 (1) に示す方法で重みを変換し中間特徴マップを畳み込む．ただし w_{ijk}, w'_{ijk} はそれぞれ Weight Demodulation 前後の畳み込み重み，s_i は潜在変数 $\mathbf{w}$ を 1 層の FC 層で変換したものである．

$$w'_{ijk} = s_i \cdot w_{ijk} \Big/ \sqrt{\sum_{i,k} (s_i \cdot w_{ijk})^2 + \epsilon} \tag{1}$$

画像生成時は，512 次元の標準正規分布からサンプリングしたノイズ $\mathbf{z}$ を 8 層の FC 層からなるマッピングネットワークによって潜在変数 $\mathbf{w}$ に変換し，各 Weight Demodulation 付畳み込み層に入力する．StyleGAN2 のアーキテクチャを図 1 に示す．

2.2 GAN 反転

StyleGAN は，潜在変数 $\mathbf{w}$ の従う分布である $\mathcal{W}$ 空間内で，画像の属性同士が分離されていることが知られており，この性質を利用することで画像の属性を制御可能である．例えば，顔画像で学習された StyleGAN の潜在空間には顔の向きを制御するベクトル $\mathbf{w}_{dir}$ が存在するため，ある潜在変数 $\mathbf{w}$ に対して $\mathbf{w} + \alpha\mathbf{w}_{dir}$ を用いて画像を生成すると顔の向きのみ異なる画像が得られる．$\mathbf{w}$ および $\mathbf{w} + \alpha\mathbf{w}_{dir}$ を用いて生成した画像を図 2 に示す．

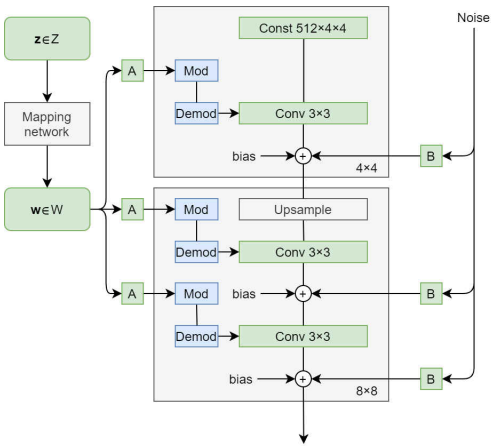

図 1 StyleGAN2 のアーキテクチャ

図 2 上段:$\mathbf{w}_{org}$ で生成した画像
下段:$\mathbf{w}_{org} + \alpha\mathbf{w}_{dir}$ で生成した画像

GAN 反転は前述の性質を利用して，入力画像を再現するように推定した潜在変数を編集することで画像編集を行う．実際は，$\mathbf{w} \in \mathbb{R}^{512}$ のみでは十分な再構成が行えないため，Weight Demodulation 付き畳み込み層の数 (1024^2 解像度では通常 18 個) だけ潜在変数を用いる $\mathcal{W}+$ 空間で推定することが多い．

2.3 pSp エンコーダ

Richardson ら [4] は，エンコーダを用いた GAN 反転である pixel2style2pixel(pSp) を提案した．pSp エンコーダは ResNet[5] ベースの Feature Pyramid[6] によって 3 段階の中間特徴マップを生成し，Map2Style と呼ばれる FCN によって変換することで潜在変数を推定する．pSp エンコーダのアーキテクチャを図 3 に示す．

2.4 空間方向に拡張した Style

StyleGAN において特徴マップの各チャンネルは，Style を制御する潜在変数の対応する要素でスケーリングされてから畳み込まれる．これに対して，いくつかの

1) 電気通信大学 大学院情報理工学研究科
Graduate School of Informatics and Engineering, The University of Electro-Communications

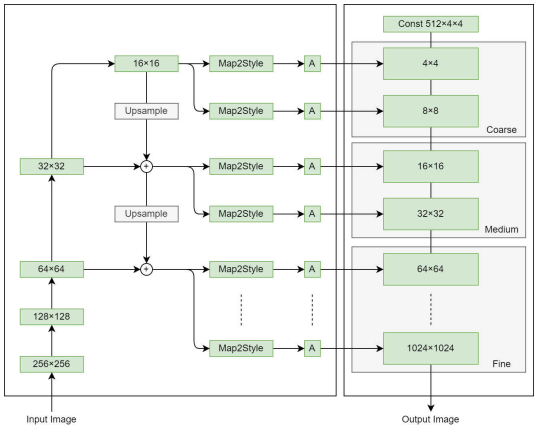

図 3　pSp エンコーダのアーキテクチャ

研究では，特徴マップと同じ形状の Style で AdaIn を行うことを提案している．

2.4.1　SPADE

Park ら [7] は正規化した特徴マップに対し，セマンティックセグメンテーションマップを用いて AdaIN 操作を行う SPADE を提案した．SPADE は，セグメンテーションマップを基に画像を生成可能である．

2.4.2　StyleMapGAN

Kim ら [8] は Style を空間方向に拡張した StyleMap を用いて AdaIN を行う StyleMapGAN を提案した．StyleMap はノイズ z をマッピングネットワークによって変換することで生成される．

StyleMapGAN では，編集箇所を指定するマスクを用いて異なる StyleMap 同士を合成することで画像の局所編集が可能である．

2.4.3　LDBR

Hong ら [9] は，StyleGAN2 で生成した画像の局所編集のため，画像の一部の潜在変数のみリサンプリングする Low Distortion Block-Resampling(LDBR) を提案した．LDBR では StyleGAN2 を用いているため，Weight Demodulation を一般化した SpatiallyModulatedConv と呼ばれる操作を用いて画像を生成する．SpatiallyModulatedConv の定義を式 (2) に示す．

$$\mathrm{SpatiallyModulatedConv}_w(\mathbf{x},\mathbf{s}) = \frac{w * (\mathbf{s} \odot \mathbf{x})}{\sqrt{\sum_i \sum_j (w^2 * s^2)_{i,j}}} \quad (2)$$

3　空間方向に拡張した Style による GAN 反転

StyleMapGAN は Style を空間方向に拡張にすることで局所編集を可能にした．一方で (1) StyleGAN2 ではなく StyleGAN をベースにしているため特有のアーティファクトが依然として発生する，(2) StyleGAN のアーキテクチャを改変しているため GAN 反転にあたって既存の学習済みモデルを活用することができない，という問題がある．

LDBR は StyleGAN2 において局所的な画像編集を可能にした．しかしながら，GAN 反転に焦点を当てていないため，ランダムノイズから生成した画像の編集はできても実画像の編集はできない．

4　提案手法

4.1　ネットワークの概要

本研究では pSp エンコーダを基に，Map2Style ネットワークのダウンサンプリングの回数を減らすことで StyleMap を推定する Map2Map を定義し，アーキテクチャを構成した．提案手法のアーキテクチャを図 4 に示す．

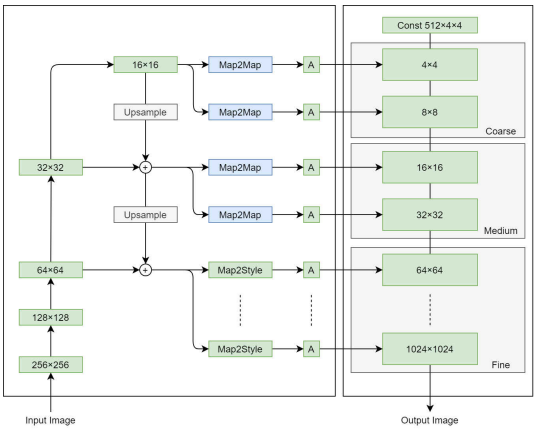

図 4　StyleMap に対応させた pSp エンコーダ

それぞれの Map2Map ネットワークでは 3 回のダウンサンプリングを行う．ダウンサンプリングごとに特徴マップの解像度は 1/2 になるため，Coarse スケールでは 2^2，Medium スケールでは 4^2 の StyleMap が推定される．推定した StyleMap は畳み込み時に中間特徴マップと同じサイズに拡大されて用いられる．Fine スケールでは，省メモリのために通常の Map2Style ネットワークを使用している．

4.2　損失関数

pSp に用いられている損失関数を用いた．すなわち，入力画像と再構成画像について，L2 損失，LPIPS 損失 [10]，ID 損失である．ID 損失は，顔の類似度を出力する事前学習済みの ArcFace[11] ネットワーク R について $\mathcal{L}_{ID} = 1 - R(\mathbf{x}) \cdot R(G(E(\mathbf{x})))$ で定義される．

5　実験・評価

5.1　データセット

顔画像のデータセットである FFHQ[2] で学習された StyleGAN2 に対して，同データセットを用いてエンコーダを学習し，CelebA-HQ[12] によって評価した．

5.2　実験結果

CelebA-HQ のランダムな画像に対して GAN 反転を行った．結果を図 5 に示す．3 列目の画像における髪の明部や，目元のアイシャドウなど，提案手法は既存手法と比べてより適切に再構成ができている．

5.3　定量評価

5.3.1　再構成品質

pSp と同様に，LPIPS，MSE，Similarity による評価を行った．Similarity は CurricularFace[13] による顔の類似度である．

表 1　再構成品質の定量評価

手法	Similarity↑	LPIPS↓	MSE↓
pSp	0.56	0.17	0.04
提案手法	0.67	0.10	0.02

5.3.2　編集品質

Tov ら [14] は GAN 反転における編集品質の指標として，式 (3) に示す Latent Editing Consistency(LEC) を提案

図 5　再構成画像

した．

$$LEC(f_\theta) = \mathop{\mathbb{E}}_{\mathbf{x}}[\|E(\mathbf{x}) - (f_\theta^{-1} \circ E \circ G \circ f_\theta \circ E)(\mathbf{x})\|_2] \quad (3)$$

ここで，$f_\theta(\mathbf{w}) = \mathbf{w} + \alpha \mathbf{w}_{dir}$ であり，編集ベクトル $\mathbf{w}_{dir}$ は InterfaceGAN[15] によって出力される．LEC の StyleMap に対する拡張としては，avg・min・max の三つを定義した．すなわち，StyleMap の各ピクセルを Style とみなして差の二乗和を計算し，これらの平均値，最大値，最小値を用いて LEC を算出する拡張である．Age ベクトルについて $\alpha = 3, -3$ をそれぞれ Old,Young として，Smile ベクトルについて $\alpha = 3, -3$ をそれぞれ Smile,No Smile として LEC を評価した．評価結果を表 2 に示す．

表 2　LEC↓ による評価結果

手法	Old	Young	Smile	No Smile
pSp	63.54	59.14	53.95	54.00
提案手法 (avg)	297.13	288.23	278.93	279.97
提案手法 (min)	202.39	195.53	189.32	189.28
提案手法 (max)	427.38	415.82	402.14	404.25

5.4　定性評価

5.4.1　Age ベクトルによる年齢編集

pSp，提案手法の推定した潜在変数について，InterfaceGAN により出力された Age ベクトルを用いて画像編集を行った結果を図 6 に示す．pSp，提案手法ともに年齢編集はできているが，提案手法は pSp と比べて，編集の影響が薄いことが観察できる．

5.5　Style 補間

二枚の実画像から推定した潜在変数を補間した結果を図 7 に示す．2 枚の入力画像から潜在変数 $\mathbf{w_1}, \mathbf{w_2}$ を推定し，それぞれの手法について左から $\mathbf{w_1}, 0.75\mathbf{w_1} + 0.25\mathbf{w_2}, 0.25\mathbf{w_1} + 0.75\mathbf{w_2}, \mathbf{w_2}$ を用いて画像を生成した．pSp，提案手法ともに，2 枚の入力画像に対して中間的な画像を出力できている．

5.5.1　StyleMap の空間的補間

提案手法は StyleMapGAN と同様に，StyleMap 同士をマスクで合成することによって局所編集が可能である．垂直方向になだらかな補間を行うマスクを用いて画像編集を行った結果を図 8 に示す．

1 行目の 2〜4 列および，1 列目の 2〜4 行は入力画像である．それ以外の r 行 c 列目の画像については，画像上部では 1 行目 c 列の入力画像から推定した StyleMap，画像下部では r 行 1 列目の入力画像から推定した StyleMap を用いて画像が生成される．提案手法の Fine スケールでは StyleMap ではなく通常の Style を用いているため，補間にあたっては Style を 1 ピクセルの StyleMap とみなし，拡大してから補間した．

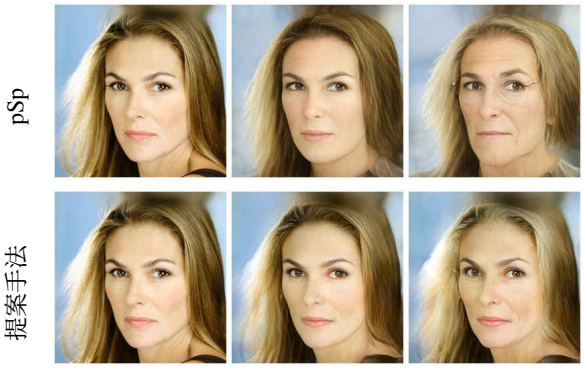

図 6　年齢編集の比較．
左から入力画像，再構成画像，
Age ベクトルを加算した編集画像．

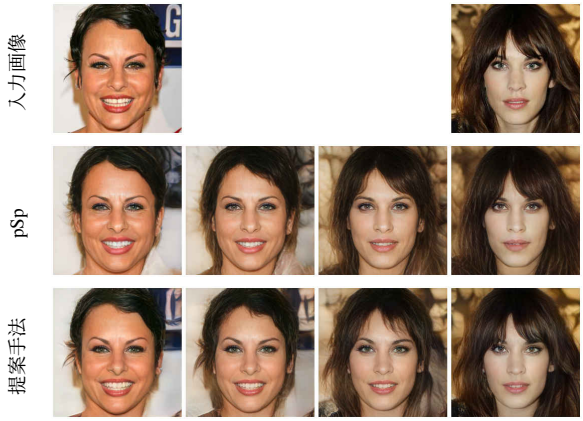

図 7　Style 補間の比較．
pSp, 提案手法ともに両端は入力画像の再構成．

図 8　StyleMap の空間的補間

6　考察

再構成品質は定量的・定性的に pSp より良い結果となった．これは潜在変数を空間方向に拡張したことで，

単一の Style を用いると画像の細部をコントロールできないのに対して，局所的に異なる Style を推定できるからであると考える．

また，編集品質は定量評価において pSp と比べて大幅に悪い結果となったが，一方で定性的には編集ベクトルによる影響が pSp と比較して小さいという軽微な悪化に収まった．この現象の原因は次の 2 つが考えられる，(1) 顔画像のデータセットは多様性が低い，(2) LEC が編集品質を正確に評価できていない．前者は，顔画像のデータセットは多様性が低く，編集品質の悪さが顕在化しなかったという可能性である．この仮説を検証するためには，車や馬のようなより多様性の高いデータセットで実験する必要がある．後者については，より定性評価に整合した評価指標の調査をする必要があると考える．

以上のように編集品質の評価に関しては課題が残るが，編集対象を顔画像に限定すれば，定性的には Style 補間ができている点，通常の GAN 反転ではできない Style の空間的な補間ができるという点から提案手法は画像編集に十分有用である．

7 まとめおよび今後の展望

StyleGAN において画像の性質を制御する潜在変数を空間方向に拡張した StyleMap を用いて，GAN 反転を行った．定量的・定性的に既存手法より良い再構成が行えることを確認し，編集についても定量的には劣るものの，定性的に十分な編集が行えることを確認した．

今後の展望として第一に，顔画像以外のデータセットでの実験が挙げられる．前述のように，編集品質の定量評価および定性評価の差異の原因として，多様性の低いデータセットでは性能が十分に評価できていない可能性がある．そのため，様々なデータセットで同様の実験を行い，性能の評価を行いたい．第二に，推定潜在変数についての正則化が挙げられる．GAN 反転の編集品質は，推定潜在変数の分布と StyleGAN のマッピングネットワークの分布の距離に依存するという指摘がある [14]．StyleMap による高い再構成品質を活かしつつ，適切な制約を与えることによって，再構成品質と編集品質を両立した GAN 反転を行いたいと考える．

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, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, Vol. 27, pp. 2672–2680. Curran Associates, Inc., 2014.

[2] 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.

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

[4] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015.

[6] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 936–944, 2017.

[7] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[8] Hyunsu Kim, Yunjey Choi, Junho Kim, Sungjoo Yoo, and Youngjung Uh. Exploiting spatial dimensions of latent in gan for real-time image editing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021.

[9] Sarah Hong, Martin Arjovsky, Darryl Barnhart, and Ian Thompson. Low distortion block-resampling with spatially stochastic networks. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, Vol. 33, pp. 4441–4452. Curran Associates, Inc., 2020.

[10] 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.

[11] Jiankang Deng, Jia Guo, Xue Niannan, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, 2019.

[12] 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*, 2018.

[13] Yuge Huang, Yuhan Wang, Ying Tai, Xiaoming Liu, Pengcheng Shen, Shaoxin Li, and Feiyue Huang Jilin Li. Curricularface: Adaptive curriculum learning loss for deep face recognition. pp. 1–8, 2020.

[14] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. *arXiv preprint arXiv:2102.02766*, 2021.

[15] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *CVPR*, 2020.