特徴空間における位置情報を条件付けに用いた拡散モデルの生成制御手法

早田 啓介[1,a)]

概要

拡散モデルを用いた画像生成によるデータ拡張は、多様な学習用データを大量に生成しモデルの精度向上が期待できる一方、生成の詳細な条件付けを行えないため、データの効率的な選択的生成という点で課題を持つ。例えば、識別精度に大きく寄与するクラス境界付近のデータを指定して生成することは難しい。本研究ではこの課題を解決するため、画像の特徴空間における位置情報を新たな条件付け情報として導入し、ピンポイントな条件指定による画像生成を可能とする手法を提案する。2 種類のデータセットを用いた実験で生成画像の画質と生成制御性の評価を行った結果、提案手法は従来法に比べてより高品質かつ条件に合致した画像を生成可能であることを確認した。

1. はじめに

近年、拡散モデルを活用したデータ拡張技術は、クラス分類やセグメンテーションをはじめ各種識別タスクにおけるモデルの汎化性能向上に革新的な進展をもたらしている。従来の幾何学的変換や色調整といった基本的なデータ変換や GAN[1] など Deep Learning ベースの類似技術に比べ、拡散モデルは高品質かつ多様な合成画像を生成可能な点が特徴的である。特に医療画像や不均衡データセットへの適用事例では、精度が大幅に向上したとする研究結果が報告されている [2][3][4]。さらにその後、テキストや画像など多様な条件付け生成を可能にする手法 [5][6] や、潜在空間上で拡散モデルを適用することで計算効率と生成品質を向上し、より実用的な高品質画像を生成する手法 [7] が提案されている。本研究では、画像データの特徴空間における位置ベクトルを新たな条件情報として用いた手法を提案し、より厳密な条件付けによる画像生成を試みた。本稿ではその手法と検証結果について報告する。

[1] コニカミノルタ株式会社
[a)] keisuke.hayata@konicaminolta.com

2. 関連研究と課題

2.1 拡散モデル

拡散モデルの基本概念は、2015 年 Sohl-Dickstein らが提案した非平衡熱力学に基づく深層学習フレームワークに端を発する [8]。この手法は、データ分布を段階的にノイズ化する「拡散過程」と、その逆プロセスである「逆拡散過程」から構成される。具体的には、元画像にガウシアンノイズを徐々に付加していき、最終的に純粋なノイズに変換する過程を学習する。2020 年に発表された Denoising Diffusion Probabilistic Models (DDPM) [9] と Denoising Diffusion Implicit Models (DDIM) [10] は、Sohl-Dickstein らの拡散確率モデル [8] や Song らのスコアベースモデル [11] の理論的・実践的な限界を克服しより高品質な画像を高速かつ決定的に生成可能とした。

2.2 生成画像をクラス識別タスクに用いる際の課題

拡散モデルによる生成画像をクラス識別タスクのモデル学習のための拡張データとして利用しようとする場合、生成される画像の特徴を指定して制御することが難しい。特に、高精度な識別に寄与するクラス境界付近のデータを分布の中からピンポイントで指定して生成することが困難という課題がある。拡散モデルは、基本的に学習データから抽出された特徴分布から均等な確率で表現を生成しようとする。そのため、クラスラベルを与えてもそのクラスの分布からランダムにデータを生成してしまう。複数のクラスの中間的な画像を生成する方法として、例えば複数のクラスラベルの尤度を比率で設定してソフトラベルとして与える方法 [12] があるが、その場合単にそれぞれのクラスの特徴を「混ぜ合わせた」ような、平均的な画像が生成されやすく、意図する「境界」の特徴が強く現れないことがある。「クラス境界」というのは多くの場合、高次元の特徴空間における曖昧で複雑な領域であり、単純なラベルの補間だけでは、この複雑な領域を正確に表現することは困難なためだと考えられる。

また Latent Diffusion Model (LDM) [7] などでのテキス

トの埋め込みベクトルや、ControlNet [5] で画像を条件付けに用いる場合でも、条件付けの自由度は向上するが、同様にクラス境界などをピンポイントに条件指定することは困難である。本研究ではこれらの課題の解決方法として、クラスラベルに加えて生成画像を詳細に制御するための情報を追加した手法を提案する。

3. 提案手法

提案手法では、画像データの特徴空間における位置ベクトル情報を条件付けの情報に追加することで、「クラス境界」のような限定された条件を指定した画像生成を行えるようにする。

モデルの構成を図 1 に示す。U-Net ベースの conditional DDPM モデルにおいては、各層に対し時刻 t と条件付けラベル情報 y を入力するが、提案手法ではこれに加えて画像の特徴空間上の位置ベクトル情報を条件付けに追加する。ラベル情報 *label* と、Uniform Manifold Approximation and Projection (UMAP) [13] で次元削減を行った UMAP 位置ベクトル情報 *umap_loc* を入力する構造をとることで、クラス情報に加えて画像の特徴空間上の位置ベクトル情報を同時に学習させる。生成時にはクラス情報と UMAP 位置ベクトル情報を条件付けとして入力することで、入力に該当する条件の画像を生成することができる。

今回、UMAP 位置ベクトル情報は、元画像データセットに対し ImageNet データセットで学習した Swin Transformer v2 [14] を用い、事前学習モデルとして `swinv2_base_window12to16_192to256` を使用した [15]。モデルのバックボーンで特徴ベクトルを抽出し、これを UMAP で 20 次元に次元削減を行ったものを用いる。UMAP を用いた理由は、UMAP は高次元データを低次元に削減して可視化するための強力な手法であり、データセット内のクラスタが低次元空間でより明確に集合し、かつ各クラスタを適切に分離して可視化する性能に優れているためである。こうした特長から、UMAP で「クラス境界」のような情報を表現しやすくすると考えた。

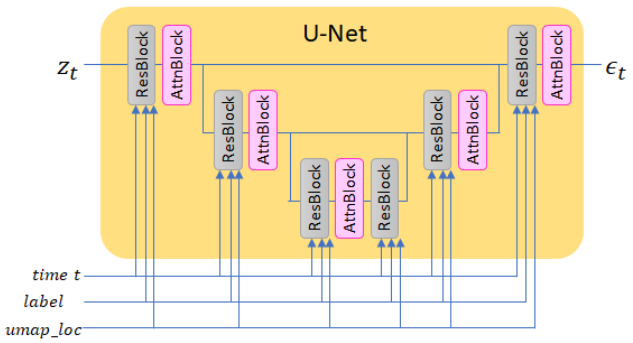

図 1 提案手法のモデル構造

UMAP 位置ベクトル情報を条件付けとして行う本手法は、画像から抽出した特徴量を何らかの「制御信号」として Diffusion モデルに与えるという考え方に基づくものであり、概念的には ControlNet でポーズやエッジ情報を条件付けとして利用することに近い。ControlNet は、より具体的な「構造情報」を条件としているが、本提案手法では、抽象的な「特徴量」を条件として与えることで、例えば「クラス境界」のようなの曖昧な特徴を表現することができると考える。

4. 評価実験

提案した拡散モデルを用いて、画像生成の評価実験を実施した。データセットは 2 種類で、外観検査評価用の社内データおよび MNIST データセットを用いる。

4.1 実験データ

(1) 外観検査データ：9 クラス、計 5427 枚。
自動車塗装の外観検査評価用の社内データ（非公開データ）。自動車外装の塗装後、表面についた傷や汚れを検出するための評価データセット。
内訳は以下の通り。

class	0	1	2	3	4	5	6	7	8
num	1552	414	119	1845	568	160	249	290	230

表 1 外観検査データ

(2) 手書き数字画像データ MNIST：0〜9 の数字で構成させる 10 クラス。train セットより各クラス 1000 枚、計 10000 枚を使用。

4.2 学習条件

モデルの学習条件として、optimizer は AdamW、学習率（lr）は 1×10^{-5}、学習 epoch 数は 500 に設定した。

4.3 画質評価

4.3.1 評価条件

提案手法の評価として、拡散モデルへ条件付けする情報を変えて生成画像の比較を行う。各クラスごとに 1 枚ずつ選択した参考画像から UMAP 位置ベクトルを計算し、これを生成の条件付けに用いる。条件付け情報は以下の 3 パターンを用いる。

- 従来法：クラスラベル（label）のみ
- 提案手法：クラスラベル + UMAP 位置ベクトル（label + umap）
- 比較手法：UMAP 位置ベクトルのみ（umap）

UMAP 情報のみを用いた条件付けは既存手法として一般的ではないが、本手法の効果検証のため比較手法として設定した。

評価指標は、以下を用いた。

- 生成画像の目視による主観評価
- Frechet Inception Distance（FID）スコア
- Inception Score（IS）スコア

FID スコアは、「生成画像の分布が実画像の分布にどれだけ近いか」を、IS スコアは「生成画像の多様性」を評価する指標であり生成画像の評価指標の 1 つとして用いられる。

学習時に用いる UMAP 位置ベクトルは、学習データから 20 次元の UMAP 空間を構築し射影されたベクトルを使用する。生成時に用いる UMAP 位置ベクトルは、学習データから各クラスごとに無作為に選択した参考画像に対し、学習時に構築した UMAP 空間を用いて算出する。

4.3.2 結果

従来法（label）、提案手法（label+umap）、比較手法（umap）の結果を図 2〜図 4 に示す。ここでは、MNIST の結果についてのみ掲載する。

従来法 (図 2) や比較手法 (図 4) では、外観が似たクラスに時折混同して生成されているが、提案手法 (図 3) ではほぼ正しいクラスの画像が生成されており、UMAP 情報を導入することでより正確な条件付けによる生成が行われていることが定性的にわかる。

次に定量評価として FID および IS スコアの結果を見てみる。FID スコアでは、表 2、表 4 に示すように両データセットにおいて表従来法や比較手法に比べて提案手法で FID スコアが小さくなり、複数の情報を条件付けに利用することでより実画像の分布を崩さずに精度よく生成が行えていることを示している。

IS スコアは MNIST では実画像と同等の値を保っている (表 5) 一方で、外観検査データでは生成画像は実画像よりも低い値となっている (表 3)。これは、MNIST は比較的クラス判別がしやすい分布であるのに対し、外観検査データではクラスインバランスやクラス内の外観が多様なものがありクラス判別が比較的難しいデータであることや、それにより拡散モデルが十分に分布を学習し生成できていない可能性を示している。

外観検査データ

表 2 外観検査データにおける FID スコア

条件付け	label	label + umap	umap
FID score	79.71	53.03	77.93

表 3 外観検査データデータにおける IS スコア

条件付け	実画像	label	label + umap	umap
IS score	3.48	2.73	2.83	2.49

MNIST データ

表 4 MNIST データにおける FID スコア

条件付け	label	label + umap	umap
FID score	92.09	71.29	92.18

表 5 MNIST データにおける IS スコア

条件付け	実画像	label	label + umap	umap
IS score	2.13	2.11	2.19	2.11

図 2 生成画像 (従来法：label 情報のみ使用）

図 3 生成画像 (提案手法：label+UMAP 情報を使用）

図 4 生成画像 (比較手法：UMAP 情報のみ使用）

4.4 生成制御評価

4.4.1 実験内容

各クラスごとに選択した 2 種類の参考画像から UMAP 位置ベクトルを計算し、これを生成の条件付けに用いることで、生成画像の制御が行われているかを評価する。参考画像と生成画像が類似しているかどうかを目視で評価する。

また両データセットにおいて、異なる 2 種の参考画像から求めた UMAP 位置ベクトルを用いて、それぞれに類似した画像が生成されるかの評価を行った。

各クラスで 2 種類の参考画像を選定し、参考画像から算出した UMAP 位置ベクトルで条件付けした画像を各 10 枚ずつ生成する。

4.4.2 期待される結果

提案手法によって画像の生成制御ができている場合は、以下のような結果が期待される

- 各参考画像から 10 枚ずつ生成した画像は、同じクラスラベル＋ UMAP 位置ベクトルを使っているので全部似た画像になる
- 2 つの参考画像 1, 2 では UMAP 位置ベクトルが異なるので、参考画像 1 の場合と 2 の場合ではそれぞれ違う外観の画像が生成される (それぞれの参考画像に似

た画像が生成される)

4.4.3 結果

外観検査評価データにおける生成結果を図 5 に示す。外観検査データは非公開データのため参考画像は提示せず生成データのみを提示する。

各行が各クラスごとの生成画像で、9 クラス × 各クラス 10 枚の生成結果を表示する。左右 2 つのセットは、2 種それぞれの参考画像に対する生成結果である。

外観検査評価データでは各参考画像から計算した UMAP 位置ベクトルを条件付け入力することで、生成画像の制御が行われていることを確認した。各行の同クラス内の生成画像が一貫性した類似性をもつ画像でであり、かつ左右 2 セットを比較すると異なる外観の画像が生成されている。このことから条件付けに用いた UMAP 位置ベクトルによって生成が制御できていることがわかる。

続いて MNIST データにおける生成結果を図 6 に示す。MNIST データでは各行の同クラス内の生成結果を見てわかるように外観の一貫性が乏しく、また 2 つの参考画像のどちらでも参考画像に類似した外観が再現されておらず、制御がうまく行われていなかった。

2 つのデータセットで制御性が分かれた原因としては、UMAP 座標空間におけるクラス内の分布の密度が影響すると考えられる。図 7 に、外観検査データおよび MNIST データを 2 次元の UMAP 座標にプロットした図を示す。MNIST データではクラス毎にデータがまとまっている一方、外観検査データではクラスごとのデータが明確に分離されず、特徴空間上で分散的に広がっている。

MNIST データでは同クラスの位置ベクトルが近くにまとまっているため位置ベクトルがクラス内の変動を表現できるだけの違いのある情報を表現できておらず、外観検査データでは同クラス内での位置ベクトルの変動が大きいため、位置ベクトルがクラス内の変動を表現する情報として有効に働いていると考えられる。

5. まとめ

本研究では、新たな条件付けの情報を定義しそれを学習に組み込むことで、より厳密な条件付けによる画像生成が行える手法を提案した。本手法を用いて、社内保有の外観検査用データと MNIST データセットの生成評価を行い、従来のラベルデータのみを用いる手法に比べてより正確な画像が生成されることを示した。生成制御については外観検査データについては UMAP 位置ベクトルを条件付けに利用することで生成が制御できることを確認したが、MNIST データでは制御の確認はできなかった。原因について UMAP 空間上のデータ密度の観点から考察を実施した。今後は生成データをクラス識別の追加学習データに用いたデータ拡張を行い、より効率的に識別精度が改善できるかの評価実験を行う予定である。

今回は提案手法の原理検証のためデータ数が十分にある条件下で評価を行ったが、識別モデルのデータ拡張は検査や医療など入手できるデータが少数の環境でより必要とされるため、将来的には少数データでも精度よく生成が行える技術を目指す。

6. 謝辞

本研究において有益な技術的アイディアをご提供くださった指田岳彦氏、ならびに建設的な技術ディスカッションを通じて多くの示唆を与えてくださった池田信氏、田中朋陽氏に深く感謝する。また、論文執筆に際し、多大な支援をいただいた筒川和樹氏、長野紘士朗氏、池田大志氏、Quan Hoangdanh 氏にも深く感謝の意を表する。

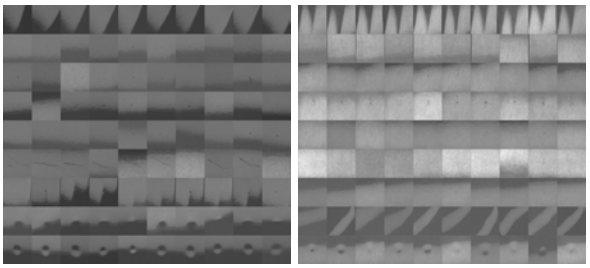

図 5 外観検査データ：各参考画像に対する生成画像
(左: 参考画像 1 に対する結果、右: 参考画像 2 に対する結果)

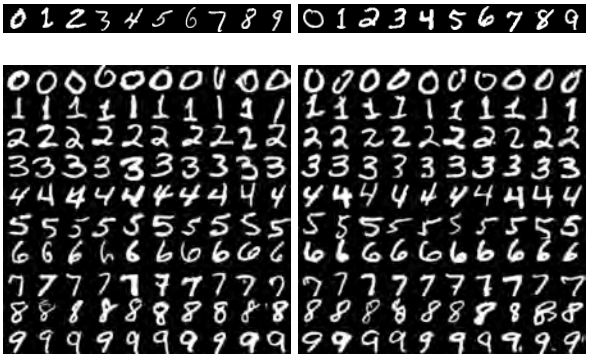

図 6 MNIST：参考画像 (上段) と参考画像に対する生成画像 (下段)
(左: 参考画像 1 に対する結果、右: 参考画像 2 に対する結果)

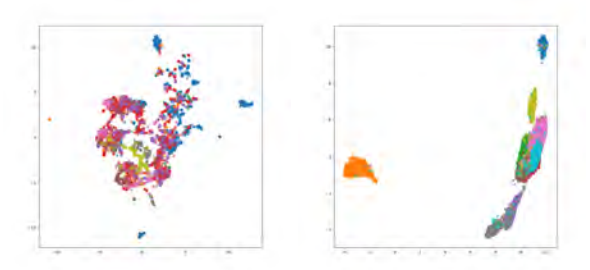

図 7 UMAP 座標プロット (左：外観検査データ、右：MNIST データ)

参考文献

[1] Ian J. Goodfellow, et al., "Generative Adversarial Networks", Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2, 2014

[2] Ira Ktena, Olivia Wiles, et al., "Generative models improve fairness of medical classifiers under distribution shifts", Nat Med. 2024 Apr;30(4):1166-1173

[3] Firas Khader, et al., "Denoising diffusion probabilistic models for 3D medical image generation", Scientific Reports volume 13, Article number: 7303 (2023)

[4] Kulikov et al., "Diffusion Model-Based Data Augmentation for Lung Ultrasound Classification with Limited Data", Proceedings of the 3rd Machine Learning for Health Symposium, PMLR 225:664-676, 2023.

[5] Lvmin Zhang, et al., "Adding Conditional Control to Text-to-Image Diffusion Models", International Conference on Computer Vision（ICCV 2023）

[6] Alex Nichol, et al., "GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models", Proceedings of the 39th International Conference on Machine Learning, PMLR 162:16784-16804, 2022.

[7] Robin Rombach, et al., "High-Resolution Image Synthesis with Latent Diffusion Models", Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10684-10695

[8] Jascha Sohl-Dickstein, et al., "Deep Unsupervised Learning using Nonequilibrium Thermodynamics", ICML'15: Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume37, Pages 2256-226

[9] Jonathan Ho, Ajay Jain, Pieter Abbeel, "Denoising Diffusion Probabilistic Models", Advances in Neural Information Processing Systems 33 (NeurIPS 2020)

[10] Jiaming Song, Chenlin Meng & Stefano Ermon, "DENOISING DIFFUSION IMPLICIT MODELS", International Conference on Learning Representations (ICLR) 2021

[11] Yang Song, Stefano Ermon, "Generative Modeling by Estimating Gradients of the Data Distribution", Neural Information Processing Systems（NeurIPS）2019

[12] Prafulla Dhariwal and Alex Nichol, "Diffusion Models Beat GANs on Image Synthesis", NeurIPS 2021

[13] Leland McInnes,et al.,"UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction",Journal of Open Source Software（JOSS）Vol. 3, No. 29, p. 861, 2018

[14] Ze Liu(Microsoft Research Asia) et al., "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows", International Conference on Computer Vision (ICCV), 2021, pp. 10012–10022

[15] Hugging Face, swinv2_base_window12to16_192to256, Available at: `https://huggingface.co/timm/swinv2_base_window12to16_192to256.ms_in22k_ft_in1k` [Accessed: 2025-06-06].

[16] Xin Ding, et al., "CCDM: Continuous Conditional Diffusion Models for Image Generation", arXiv preprint