

論文

StableDiffusion による 作家性が再現された画像の生成

若月 玲之 平綿 素望 大谷 紀子

小説や絵画、楽曲などの生成 AI が質の高い作品を生成するようになり、危機感を覚える作家がいる一方で、自らの「作家性」が反映された作品を永続的に発表し続けるための道具として、生成 AI の使用可能性に期待を寄せる作家もいる。本研究では、作家の創作活動における生成 AI の活用可能性の示唆を目的として、既存の画像生成 AI を用いて作家性を反映したイラストを生成する手法を提案する。デジタルツールで描かれたカラーイラストとアナログツールで描かれたモノクロイラストの2種類を処理対象とし、画像生成 AI には StableDiffusion を採用する。モノクロイラストに関しては、複数のシーンが含まれている画像をシーンごとに切り分ける手法も合わせて提案する。デジタルアーティスト・wataboku 氏のカラーイラスト、および漫画家・窪之内英策氏のモノクロイラストを教師データとして、提案手法によりイラストを作成した結果、作家の代理を任せられるレベルには達していないが、ある程度の作家性の再現は可能であることが示された。wataboku 氏に関しては、オリジナルアイコンである SAI に加え、背景も高精度に再現された。窪之内英策氏については、人物の全身画像を教師データとすると目の特徴を再現できないという問題があったが、顔周りのみをトリミングして教師データとすることで、まつげなどの細部まで再現することに成功した。モノクロイラストのシーン切り分けについても、提案手法は既存手法より適切にシーンを切り分けられることが確認された。

キーワード：画像生成 AI, StableDiffusion, LoRa, 線画, クラスタリング

1. はじめに

第3次 AI ブームのなかで「AI は人間の仕事を奪う存在」という声が多く聞かれるが、小説や絵画、楽曲などの生成 AI がある程度質の高い作品を生成するようになると、作品の提供側、被提供側の両者から「AI は作家の仕事すら奪う」と言われるようになった。2023年、生成 AI の活用を巡って全米脚本家組合が大規模なストライキを実施し、ロサンゼルスやニューヨークでは俳優たちが一斉デモを行ったことから、生成 AI に対して危機感を覚える人がいることがわかる。一方、自らの「作家性」が反映された作品を永続的に発表し続けるための道具として、生成 AI の使用可能性に期待を寄せる作家もいる。

本研究では、作家の創作活動における生成 AI の活用可能性の示唆を目的として、既存の画像生成 AI を用いて作家性を反映したイラストを生成する手法を提案する。デジタルツールで描かれたカラーイラストとアナログツールで描かれたモノクロイラストの2種類を処理対象とし、

画像生成 AI には少量の VRAM でパラメータ調整が可能な StableDiffusion ^[1] を使用する。

2. 提案手法

2.1 イラスト生成手順

提案手法におけるイラスト生成手順を以下に記す。

1. 特定の作家のイラストから教師データを作成する。
2. Dataset Tag Editor¹において、インターロゲーターとして WD 1.4 SwinV2 Tagger V2モデル²を使用し、各教師データのキャプションを生成する。
3. 目的の特徴が学習されるようにキャプションを編集する。
4. 教師データと編集後のキャプションを使用し、Stable Diffusion V1-5 Model Card³をチェックポイントとする追加学習を実行して、LoRA モデル^[2]を作成する。

WAKATSUKI Reiji

東京都市大学メディア情報学部情報システム学科2年生

HIRAWATA So

東京都市大学メディア情報学部情報システム学科4年生

OTANI Noriko

東京都市大学メディア情報学部情報システム学科教授

¹ <https://github.com/toshiaki1729/stable-diffusion-webui-dataset-tag-editor>

² <https://huggingface.co/SmilingWolf/wd-v1-4-swinv2-tagger-v2>

³ <https://huggingface.co/runwayml/stable-diffusion-v1-5>

5. StableDiffusionでプロンプトを入力し、作成したLoRAモデルを使用してイラストを生成する。

StableDiffusionでは入力された画像を正方形にリサイズして学習に使用する⁴。縦横比の変更による画像の歪みを回避し、限定した対象を適切に学習させるため、1では教師データが学習対象でないものが可能な限り含まれないような正方形になるよう、教師データをトリミングする。2で生成されるキャプションは、学習対象を識別するための文字列であるトリガーワードに続いて、画像の内容を説明する複数の単語をカンマで区切って列挙した文字列である。3では、重複する単語および学習対象でない特徴の説明単語のみからなるキャプションを追加学習時に使用することで、教師データの特徴をより詳細に獲得することができる。

2.2 画像切り分け

マンガやアニメーション制作において、キャラクターや背景、アイテムなどのビジュアルコンセプトを固定するための参考資料を設定資料と呼ぶ。一般に、1枚の設定資料に複数のシーンが描かれていることが多い。設定資料の例を図1に示す。複数シーンが混在した画像を追加学習に用いると、細部が学習されないため、生成されるイラストの品質が低下する。品質向上のためには、画像

⁴ 本研究遂行時点（2023年9月）

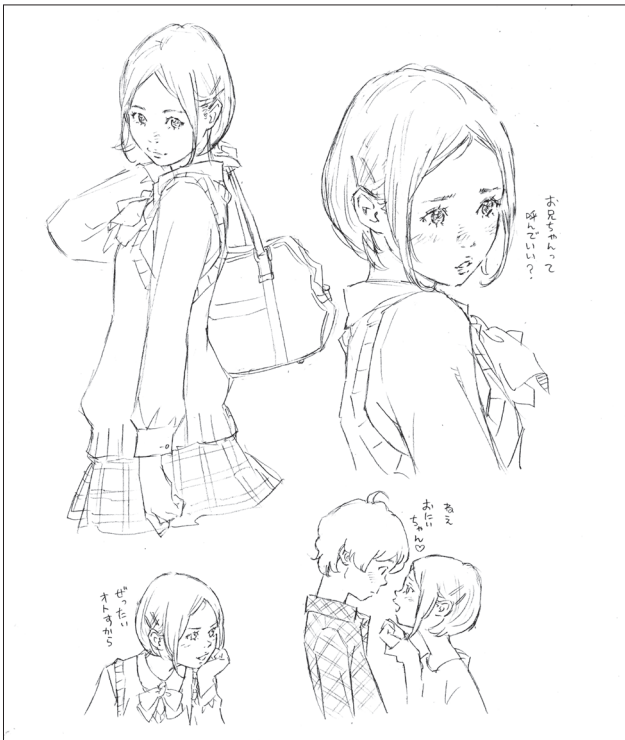


図1 設定資料の例

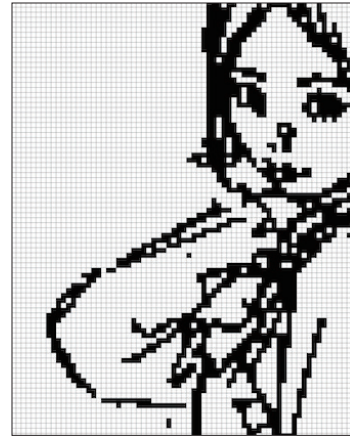


図2 モルフォロジー変換後の画像

をシーンごとに切り分ける必要がある。本研究では、モノクロイラストとして提供された設定資料をシーンごとに自動で切り分ける手法を提案する。シーン数を k とするときの処理手順を以下に示す。

1. 画像を適切な大きさの方眼で覆い、モルフォロジー変換を適用する。モルフォロジー変換とは、主に二値画像を対象として、特徴抽出やノイズ除去を行うための処理である。膨張と収縮という2つの基本操作のうちの膨張を適用して、1つの方眼の中に1つ以上白でないドットが含まれる場合に方眼内を黒く塗りつぶし、絵の内部に連結領域が出現しやすいようにする。図1のイラストに膨張を適用した画像の一部分を図2に示す。



図3 $k = 4$ としたときの確定済み連結領域

2. すべての連結領域を列挙する。連結領域とは、隣接する方眼が塗りつぶされている領域を指す。
3. 含まれる方眼の数が多い k 個の連結領域を確定済み連結領域とする。図1のイラストで $k=4$ としたときの確定済み連結領域を図3に示す。赤、緑、水色、紫の部分が確定済み連結領域、黒の部分が未確定連結領域である。
4. 確定済み連結領域に近いものから順に未確定連結領域を確定済み連結領域に統合し、方眼にラベリングする。
5. 方眼のラベリング結果を用いて元の画像のすべての点にラベリングする。
6. ラベルごとに別の画像として保存する。

高速で実用的なクラスタリング手法として、k-means法がある。k-means法は、クラスタ内の点とクラスタ中心との距離が最小になるようにクラスタを作成する。しかし、k-means法は円で囲めるようなクラスタを想定しており、設定資料においてはシーンを分断するため不適である。

また、データマイニングや機械学習の分野で広く活用されているクラスタリング手法に、DBSCAN (Density-Based Spatial Clustering of Applications with Noise) [3] がある。DBSCANの結果は、半径 eps と最小点数 $MinPts$ という2つのパラメタの設定に大きく依存する。DBSCANを設定資料の切り分けに適用する場合、距離の近いシーンを切り分けるために eps を小さくすると、少し離れたところにあるオブジェクトや顔のパーツなどが別のシーンとして判定される。データの質やクラスタリングの目的に応じてパラメタを設定する必要があるが、適切な値を導くことは難しい。

一方、提案手法のパラメタは方眼の細かさとシーン数の2つであり、DBSCANのパラメタより直感的でわかりやすいため、設定が容易である。また、セリフや飛沫などのオブジェクトや途切れている線は、離れていても他のシーンより正しいシーンに近いという仮定のもと、方眼数の多い連結領域からの距離を基準としてクラスタを生成するため、DBSCANよりシーンの切り分けに適した方法といえる。

3. 適用事例

デジタルアーティスト・wataboku氏のカラーイラスト、および漫画家・窪之内英策氏のモノクロイラストを教師データとして、本研究で提案したイラスト生成手法、および画像切り分け手法の有用性を確認する。

3.1 カラーイラストの場合

wataboku氏はオリジナルアイコンである制服の少女SAIをモデルにした作品を多く発表していることから、



図4 wataboku氏が描いたSAI



(a) 全体を使用

(b) 顔部分のみを使用

図5 提案手法により生成されたSAI

SAIのイラストを教師データとして使用し、SAIが再現されるようなキャプションとプロンプトを使用してイラストを生成した。wataboku氏が描いたSAIのイラストを図4に示す。

提案手法により生成されたイラストを図5に示す。図5 (a) はSAIの身体を描いた部分も含むイラストを教師データとして使用し、図5 (b) はSAIの顔部分のみを切り取って教師データとした場合のイラストである。いずれにおいても目つきや髪型にSAIの特徴が現れているが、図5 (b) のほうがより上目遣いで、口の開き方にもSAIの雰囲気を感じられる。表情をより反映させるためには、顔部分を切り出すことが有効であるといえる。図5 (a) では、制服姿というSAIのいつもの服装が反映されていることに加え、背景ではデジタルアート特有の色使いや光の表現が再現されている。

3.2 モノクロイラストの場合

3.2.1 切り分け手法の評価

窪之内英策氏が描いた図1のイラストをk-means法、DBSCAN、および提案手法により切り分けた。k = 4としたk-means法の切り分け結果を図6、DBSCANによる切り分け結果を図7に示す。また、方眼の1辺を20px、k = 4とした提案手法による結果を図8、方眼の1辺を10px、k = 5とした提案手法による結果を図9に示す。k-means法では上部の2人の人物が分断されている。重心からの距離で直線的に切り分けるため、過不足なくシ

ーンを切り分けることが難しいことが示されている。DBSCANでは、鼻や腕のライン、セリフ部分がそれぞれ別のシーンとして判別されており、適切に切り分けられているとはいえない。一方、k = 4とする提案手法では、人物が分断されることなく、セリフ部分も話し手と同じシーンとして判別された。さらに、方眼サイズを細かくし、k = 5とすることで、右下の2人の人物を別のシーンと判別することができた。セリフ部分は両者ともに近い位置にあるが、話し手と同じシーンに割り当てられている。以上より、提案手法では方眼サイズとkを調整することで、意図を反映した切り分けが可能であるといえる。

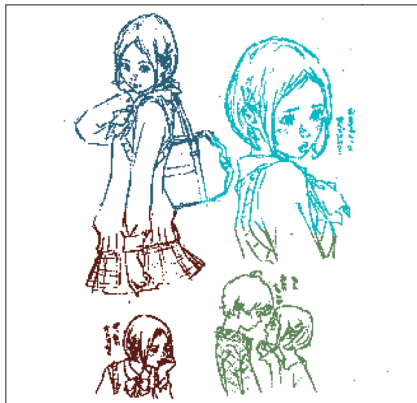


図6 k-means法による切り分け結果



図9 提案手法による切り分け結果 (10px, k = 5)



図7 DBSCANによる切り分け結果



図10 離れたオブジェクトを含むイラスト



図8 提案手法による切り分け結果 (20px, k = 4)



図11 離れたオブジェクトを含むイラストの切り分け結果

32枚の設定資料をDBSCANと提案手法により切り分けた後のシーン数を正解のシーン数とともに表1に示す。網掛け部分では、提案手法によって正解となるシーンに切り分けられている。作家の細部へのこだわりを失わないために、飛沫などの微小なオブジェクトが含まれているかまで厳密にみて、シーン数を計測した。DBSCANでは5枚しか完全に切り分けることができなかったのに対して、提案手法では21枚を完全に切り分けることができ、8枚は1箇所間違いにとどまった。

表1 切り分け後のシーン数

No.	正解数	提案手法	DBSCAN
1	10	10	9
2	9	9	5
3	8	8	5
4	8	8	2
5	8	6	4
6	8	6	0
7	8	4	3
8	7	7	5
9	7	7	3
10	7	7	2
11	7	7	2
12	7	7	2
13	7	5	4
14	7	5	4
15	7	5	3
16	6	6	4
17	6	6	3
18	6	4	1
19	6	4	0
20	6	2	1
21	5	5	5
22	5	5	5
23	5	5	3
24	5	5	1
25	5	5	1
26	5	3	1
27	5	0	0
28	4	4	4
29	4	4	4
30	4	4	4
31	4	4	0
32	3	3	1



図12 シーン同士の距離が近いイラストの切り分け結果



図13 窪之内英策氏が描いた女性

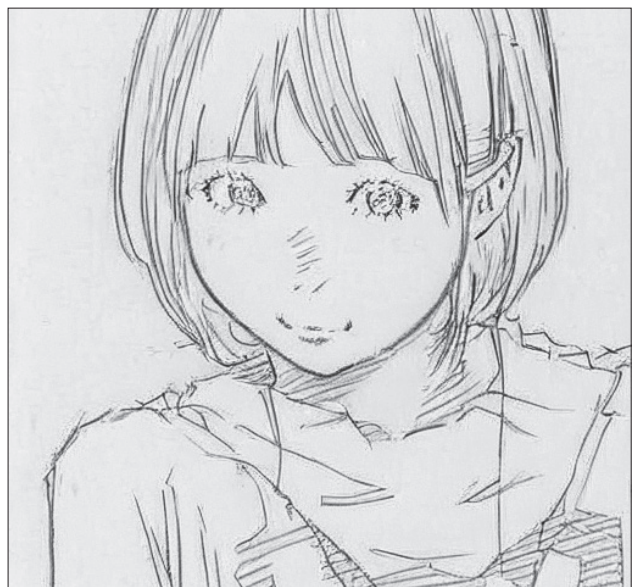


図14 提案手法により生成された窪之内英策氏風の少女

メインの人物から離れたオブジェクトが多く含まれているイラストの例を図10に示す。単純に大きな連結領域からの距離だけで判断すると、飛んでいる虫は右足に近いが、「ぶぶぶ」という文字や飛行の軌跡などが本来のシーンとの間には存在する。以上ような手がかりを順次辿っていくことによって、図11に示すように正しくラベリングができる。以上のように、離れた位置にある他のシーンに近いオブジェクトであっても、本手法であれば意図通りに切り分けられるケースが多く見られた。

適切にラベリングできない例を図12に示す。シーン内部が疎でシーン同士の距離が極めて近いものや、シーン同士が完全に連結しているものは正確に切り分けられない。

3.2.2 イラスト生成

窪之内英策氏のイラストのうち、女性が描かれている線画を教師データとして、女性のイラストを生成した。窪之内英策氏が描いた女性のイラストを図13、提案手法により生成されたイラストを図14に示す。教師データが線画の場合、キャプション生成において目の特徴を検出しにくく、生成結果においても目が潰れるが、顔周辺のみをトリミングした画像を用いて集中的に学習させたことで、窪之内氏のイラストに特有のまつげや唇の質感などの細部を再現することができた。

4. おわりに

本研究では、生成AIを活用して作家性が反映されたイラストを生成する方法を模索した。デジタルとアナログの両方のツールで描かれたイラストにおいて、StableDiffusionを用いて、作家の代理を任せられるレベルには達していないが、ある程度の作家性を再現したイラストを生成することが可能であることを示した。切り分け手法も提案し、既存手法より正確に切り分けられることを示した。

謝辞

本研究の遂行にあたり、多大なご協力をいただきました株式会社カルトン東京の皆様に深く感謝申し上げます。

参考文献

- [1] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, “High-Resolution Image Synthesis with Latent Diffusion Models” , CoRR, abs/2112.10752, 2021.
- [2] E.J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, W. Chen, “LoRA: Low-Rank Adaptation of Large Language Models” , CoRR, abs/2106.09685, 2021.

- [3] M. Ester, H.P. Kriegel, J. Sander, X. Xu. “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise” , Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining, 1996.