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.

特定の作家の作風に酷似した顔アイコンを創作する拡散モデル (仮題)

創作するイラスト深層生成モデルを活用する受容性の調査

尾崎安範†

† 個人
住所は個人情報であるため、公開は控えます。メールにて回答します。
E-mail: †ozaki.yasunori@outlook.com

あらまし　本研究の目的は、「もし、有名なイラストレーターの画像数百枚とゲーム用パソコンを持ってさえいれば、そのイラストレーターの新作を個人で無数に作ることはできるのか」との問いに答えることである。実験の結果、答えは技術的にも倫理的にも「はい」であった。本研究会で発表する目的は、この答えに対する技術的詳細と倫理的影響の顛末を報告し、この答えの取り扱いについて議論することである。なお、個人サイトにある場合の本報告書はプレプリント版である。

キーワード　拡散モデル、倫理

Diffusion Model Specializing In An Illustrator (Preprint)

Yasunori OZAKI†

† Individual
Please ask me my address via a e-mail because the address is private.
E-mail: †ozaki.yasunori@outlook.com

1. はじめに

“「序曲」は、５分でできました。そういうパッとメロディーが浮かんだ曲のほうが、こねくり回して作った曲よりも、素直で出来がいいものです。ただ、この曲は５分プラス、僕がそれまで生きてきた５５年分が詰まっている。” [1] と語るのは作曲家のすぎやまこういち氏である。この「序曲」はやがて 2020 東京大会の選手団入場曲として使われ、日本が世界に誇る曲となった。すぎやまこういち氏の例のように、人間の優れた創作能力は数十年の努力を得て習得するものである。しかし、もし、深層生成モデルによって、この 55 年間の努力をパソコンにより**半日**で習得できるとしたら、5 分の創作をスマートフォンにより **5 秒**で創作できるとしたら、どのような世界になるのだろうか。

本研究の目的は、「もし、有名なイラストレーターの画像数百枚とゲーム用パソコンを持ってさえいれば、そのイラストレーターの新作を個人で無数に作ることはできるのか」との問いに技術的側面と倫理的側面から答えることである。具体的には、個人が他人の新作を制作することは技術的に可能であり、著作権法第四十七条の四にある “著作権者の利益を不当に害すること” に該当しない限り、著作権法の目的である “著作物並びに実演、レコード、放送及び有線放送に関し著作者の権利及びこ

(a) 顔アイコン生成例　　**(b)** デモサイトへの二次元コード

図 1: 図 1a は、拡散モデル [2] が特定の作家からイラスト画像から画風を学習し、ノイズから生成した画像である。この現象を問題視した著者は、このイラストの学習画像を描いた、Web サイト「いらすとや」を運営する、みふねたかし氏と相談の上、この現象が再現できることを技術デモとしてインターネット上に報告した。この技術デモは図 1b の二次元コードをスマートフォンなどで読み取ることで体験できる。この技術デモはスマートフォン程度の性能の仮想マシンで運用されており、およそ **5 秒**でイラストを生成できる。

れに隣接する権利を定め、これらの文化的所産の公正な利用に留意しつつ、著作者等の権利の保護を図り、もつて文化の発展に寄与すること” を倫理的に反さない行為であるか社会に問うことである。この問いに至った理由は、図 1 のとおり、拡散モデル [2] が特定の作家の作風に酷似した顔アイコンをノイズから生成する現象を引き起こしたためである。

本研究のうち、本報告書の学術的貢献は以下の通りである。

（1） 特定のイラストレーターの新作さを定量的に定義し、従来研究よりも特定のイラストレーターの新作さが定量的かつ定性的に高くなる現象を発見したこと

2. 関 連 研 究

本セクションでは生成技術と著作権法を紹介し、残された課題を明らかにする。

2.1 画像生成技術

絵画の画風を学習し、その画風を他の画像へ転移させる技術として Style Transfer [3] がある。Style Transfer は風景を撮影した画像を入力とし、その画像がゴッホなどの画風になって出力される技術である。Style Transfer と拡散モデルが異なる点は生成に用いる元の画像が、意味のある画像であるか、意味のないノイズであるかの違いである。このため、意味のないノイズを入力とする拡散モデルは、意味のある画像を入力とする Style Transfer と異なり、入力画像を用意する手間がないという特徴がある。

拡散モデルと同じく意味のないノイズを入力をもとに画像を生成する技術として GAN がある。GAN は識別機と生成器をお互いに騙し合うように学習させることで、限りなく学習画像に近い偽物の画像を生成する生成器を学習させる技術である。DCGAN [4] はこの GAN の先駆け的存在である。さらに、この GAN に Style Transfer を応用したのが StyleGAN2 [5] である。StyleGAN2 とほぼ同時期に異なる生成原理を用いた拡散モデルが開発された。拡散モデルはノイズからノイズを取り除くことで画像を生成する技術である。GAN と拡散モデルは画像の生成原理が根本的に異なっており、どちらが優れていると一概には言えない。このため、各研究ごとに各技術を比較することが好ましいと考えられる。

画像生成技術を通して、イラストを生成する試みは数多くある。その中でも最も問題設定が似ているものは、いらすとやにある顔を含むイラストを DGGAN で生成した試み(注1)や StyleGAN2 で生成した試み(注2)である。

しかしながら、特定のイラストレーターの新作さを定量的に定義し、それを目的とする研究は筆者の調べる限り見当たらなかった。このため、従来の手法と今回の現象を起こす手法を用いて、特定のイラストレーターの新作さを定量的に追加調査する必要があると判断した。

(注1)：https://mickey24.hatenablog.com/entry/irasutoya_deep_learning

(注2)：https://www.gwern.net/Faces

2.2 著作権法とその法的解釈

他人のイラストを画像生成できる場合は、文化庁によると “コンピュータ等を用いて情報解析（※）を行うことを目的とする場合には，必要と認められる限度において記録媒体に著作物を複製・翻案することができる。” とあり、情報解析とは “大量の情報から言語，音，映像等を抽出し，比較，分類等の統計的な解析を行うこと” であるとしている [6]。つまり、インターネット上から必要な範囲で画像を収集し、その画像をもとにパソコン内で画像生成すること自体は翻案に相当することであり、合法的な行為であると解釈できる。

ただし、学術論文などに公開することは、“著作権者の利益を不当に害すること” に抵触するおそれがある。これは画像生成のみならず、翻案全般に関わる問題である。本研究では、みふねたかし氏から「技術の紹介に限り、透かしを入れ、素材の再配布をできない状態にするならば、生成結果を公開しても良い」とをメールにて許可を得ているため、この問題に抵触する恐れはない。

しかし、イラストを生成し、許可を得た上で公開することが合法だからといって、必ずしも倫理的に好ましいとはわからず、広く調査が必要であると考えられる。

3. 拡散モデルによる現象の再現方法

デフォルトの拡散モデルをイラストに適用するだけで本現象は起きる。さらに、拡散モデルに用いる U-Net の構造を極端な砂時計型ニューラルネットワークにすることで本現象を高頻度に起こすことができる。この極端な砂時計型ニューラルネットワークを PyTorch による拡散モデルの実装(注3)で具体的に説明する。通常の拡散モデルにおける U-Net のソースコード 1 の通りである。

ソースコード 1: 拡散モデルに用いる U-Net の標準実装

```
model = Unet(
    dim = 64,
    dim_mults = (1, 2, 4, 8)
)
```

このデフォルトの拡散モデルでも本現象は起きる。ソースコード 1 の `dim_mults` を `(1,4,16)` などとすることでさらに本現象を高頻度に起こすことができる。また、パラメータだけでなくとも、U-Net の構造をソースコードレベルで変更しても本現象を起こすことができる。

4. 提 案 手 法

（特許や実験の都合、プレプリント版では、結果を公知公用するが手法そのものは公知にはしない。研究会の原稿には記載する予定である。）

5. イラストレーターの新作さの比較実験

従来の手法と今回の現象を起こす手法を用いて、特定のイラ

(注3)：https://github.com/lucidrains/denoising-diffusion-pytorch

ストレーターの新作さを定量的に追加調査する必要があると判断したため、定量的、さらには定性的な比較実験を行った。本セクションでは、その実験を方法とともに結果を報告する。

5.1 比較方法

本現象を定量的に比較するにはまずイラストレーションの新作さを操作化して、定量化する必要がある。今回の問題を定量化した指標であるイラストレーションの新作さ NWI (New Work Index) を式 1 のように定義する。

$$\mathrm{NWI} = \mathrm{MAE} - \mathrm{FID} \tag{1}$$

FID (Fréchet Inception Distance) は画像生成の品質を評価する際によく用いられる指標である。Inception V3 の神経活動をもとにフレシェ距離を求めることで FID が得られる。今回この FID を画風の類似度として操作化する。MAE (Mean Absolute Error) は画像がピクセル単位で類似しているかを表す際によく用いられる。今回は学習データセットとのピクセル単位での類似度として操作化する。以上より、NWI は、特定のイラストレーターの新作さを「生成された画像群は、学習データセットと画風が似ているが、学習データセットとはピクセル単位での類似度は低い」として操作化した指標であると考えられる。本実験では、この NWI の妥当性をまず検証する。

次に、NWI を用いて具体的な問題に本研究を絞り、次のとおりにした。

問題設定

Web サイトいらすとやに掲載されている、人間の顔アイコン 158 枚があり、ゲーム用パソコンの GPU として、GeForce RTX 3080 があるとする。この顔アイコンとゲーム用パソコンから FID を可能な限り低くしつつ NWI が高くなるような 128 ピクセル x128 ピクセルの画像を任意の z 次元の疑似乱数から生成できる生成モデル $G : \mathbb{R}^z \to \mathbb{R}^{128\times128}$ を求めよ。また、十分に収束した FID を計算するため、その生成モデルを使い、1000 枚以上生成せよ。なお、事前学習モデルは使えないものとし、学習時間や生成時間は**半日 (12 時間)** 以内でそれぞれ終わるものとする。

この問題を解く方法、すなわち比較条件は従来手法である DCGAN と StyleGAN2、拡散モデル、提案手法の 4 種類とした。

5.2 NWI の妥当性検証

NWI が本研究の定量的指標として妥当なのかを調べるために、簡易な検証を行った。まず、いらすとやにて 2013 年に公開されていた顔アイコンを「いらすとや（初期)」とし、2015 年に公開された顔アイコンを「いらすとや（新)」とした。次に、ゲーム会社 C が運営しているゲーム U における初期リリース時点では存在したキャラの顔アイコンを「ゲーム U（初期)」と 2022 年 7 月時点までに新しく追加したキャラの顔アイコンを「ゲーム U（新)」とした。同じく、ゲーム会社 C が運営しているゲーム P における初期リリース時点では存在したキャラの顔アイコンを「ゲーム P（初期)」と 2022 年 7 月時点までに新しく追加したキャラの顔アイコンを「ゲーム P（新)」とした。以上の顔アイコン同士の組み合わせで NWI を計算した結果を表 1 に示す。表 1 の FID は pytorch-fid(注4)で計算した。表 1 を見ると、新しい顔アイコンが増えたときは NWI が高くなることがわかる。また、データセットのドメインがあまりに異なると低くなることもわかる。

表 1 の結果から、NWI が本研究の定量的指標として妥当だと推測される。

表 1: 特定のイラストレーターの新作さを表す指標 NWI の妥当性検証結果。確かに新しい顔アイコンが増えたときは NWI が高くなることがわかる。また、データセットがあまりに異なると低くなる。

データセット 1	データセット 2	FID↓	NWI↑
いらすとや (初期)	いらすとや (新)	57.4	3.91
ゲーム U(初期)	ゲーム U(新)	96.2	35.6
ゲーム P(初期)	ゲーム P(新)	67.5	26.1
ゲーム U(初期)	いらすとや (新)	352	-165
ゲーム P(初期)	いらすとや (新)	398	-201

表 2: いらすとやデータセットにおいて、従来手法とイラストの新作さを比較した結果。NWI が高いほど新作さが高いことを示している。通常の拡散モデルでも従来手法よりも FID が低く、NWI が高いことがわかる。提案手法では通常よりもさらに NWI は高くなる。

手法	FID↓	NWI↑
DCGAN [4]	293	-202
StyleGAN2 [5]	67.8	-35.5
拡散モデル [2]	14.3	9.99
提案手法	27.4	29.4

5.3 比較結果

比較実験は、可能な限り再現性を重視し、画像を学習し、生成した。乱数のシードを 3407 [7] に固定した。また、ランダム水平反転はありにした。学習と生成には GeForce RTX 3080 12GB OC モデルを用いた。なお、学習と生成ともに TF32 で計算しているため、再現するためには A100 や 30 シリーズ以降が必要である。すべての条件において画像を 1024 枚生成した。学習時間と生成時間は 12 時間以内であった。

定量的な比較結果を表 2 にまとめる。また、生成した画像のうち、著者が新作さに優れていると判断した画像を図 2 にそれぞれの比較条件でまとめる。

5.4 比較結果に対する考察

表 2 と図 2 を見る限り、DCGAN は顔アイコンをうまく生成できていないように考えられる。StyleGAN2 もよく見ると絵に粗さが見える。例えば、図 2d では髪の毛の縞模様が正しく表現されているが、図 2b の髪の毛を見ると波模様になっている事がわかる。拡散モデルではイラストレーターに類似した顔アイコンを FID の低さ、画像の質の面から生成できているように思える。しかし、提案手法では、NWI の高さから新作さ

(注4)：https://github.com/mseitzer/pytorch-fid

(a) DCGAN による例

(b) StyleGAN2 による例

(c) 拡散モデルによる例

(d) 提案手法による例

図 2: DCGAN や StyleGAN2、拡散モデル、提案手法により比較実験の中で生成された画像の例である。これらの画像は学習画像の中に完全に一致するものはない。すなわち、生成モデルによる創作である。

は高く、最も優れていると考えられる。図 2d は提案手法による例は特異である。一見普通の男性に見えるが、この画像と学習画像の中で一番 MAE が低いものは女性である。また、その女性と一致する顔のパーツは存在しない。提案手法は自身でオリジナルな作品を作り出したと言える。

本手法の限界について考える。例えば、学習画像にない顔のパーツを作ることはあまりない。また、これらのイラストからからスケッチなどの異なるドメインのイラストを創造することはできない。これらを解消するためには、転移学習を併用することが考えられる。また、現状では生成するものを利用者の望み通りに制御することはできない。このため、DALL-E 2 [8] や Imagen [9] のように text-to-image などの制御用 UI が必要であると考えられる。

以上の考察と NWI の高さから、拡散モデルによる生成画像が従来研究よりも特定のイラストレーターの新作さを高くする現象が発見し、さらに新作さを高める方法を提案できたと考えられる。

6. 品質と受容性に関する調査

（現在、調査中であるため、詳細は控える）

7. ま と め

本研究の目的は、「もし、有名なイラストレーターの画像数百枚とゲーム用パソコンを持ってさえいれば、そのイラストレーターの新作を個人で無数に作ることはできるのか」との問いに答えることである。本報告書の通り、その問いは技術的に可能であることが示された。

今後の課題は、この技術を活用することが倫理的に反さない行為であるか社会に問うことである。具体的には、技術デモや本原稿をプレプリントとして社会に公開することがどのような社会的影響を与えるのか、SNS 上の発言やイラストレーターへのインタビュー結果を自然言語処理などで分析し、調査することである。

謝　　辞

生成画像の公開に許可をくださった、みふねたかし氏に感謝の意を表します。技術デモに対するアドバイスをくださった兵頭亮哉氏にお礼を申し上げます。ただし、これらの方は、本研究に対する一切の責任はなく、本研究の責任すべては著者個人にあることを明確に述べます。

文　　献

[1] 本間英士，“作曲家・すぎやまこういち（４）　あの代表曲は５分でできた，” 産経新聞，pp.••–••，2017.

[2] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” ••, pp.••–••, June 2020.

[3] L.A. Gatys, A.S. Ecker, and M. Bethge, “Image style transfer using convolutional neural networks,” 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.2414–2423, 2016.

[4] A. Radford, L. Metz, and S. Chintala, “Unsupervised representation learning with deep convolutional generative adversarial networks,” arXiv preprint arXiv:1511.06434, pp.••–••, 2015.

[5] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, “Analyzing and improving the image quality of stylegan,” Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp.8110–8119, 2020.

[6] 文化庁，“著作物が自由に使える場合，” 2022.

[7] D. Picard, “Torch.manual_seed(3407) is all you need: On the influence of random seeds in deep learning architectures for computer vision,” CoRR, vol.abs/2109.08203, pp.••–••, 2021. https://arxiv.org/abs/2109.08203

[8] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, “Hierarchical text-conditional image generation with clip latents,” 2022. https://arxiv.org/abs/2204.06125

[9] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S.K.S. Ghasemipour, B.K. Ayan, S.S. Mahdavi, R.G. Lopes, T. Salimans, J. Ho, D.J. Fleet, and M. Norouzi, “Photorealistic text-to-image diffusion models with deep language understanding,” 2022. https://arxiv.org/abs/2205.11487