ファッションドメインにおける画像からのピクセルアート生成に向けて

入江 匠[1,a)]　横山 大作[1]

概要：ピクセルアートは，限られた解像度と色数で視覚情報を表現するデジタルアート形式であり，レトロゲームやデジタルアバター，NFT アートなど多様なコンテンツにおいて独自の価値を持つ．特にファッションドメインにおいては，現実のスタイリングをピクセルアートとして表現可能にすることが，メタバースや SNS でのアバター生成やデジタルファッションの表現手段として有用であり，その自動化はコンテンツ制作の効率化にも寄与する．しかし，単純なダウンサンプリングでは，面積の小さなアイテムの消失や衣服の複雑な模様の劣化が生じ，重要な特徴の保持が困難である．そこで本研究では，視覚的・意味的特徴を保ちながら，入力画像をピクセルアートとして効果的にデフォルメする手法の構築を目指す．具体的には，Stable Diffusion を用いた Score Distillation Sampling（SDS）による生成手法に加え，画像と生成結果のノルムに基づく SDS の早期停止を導入する．Fashionpedia データセットから抽出した人物全身画像を用いて，各手法のデフォルメの質と，視覚的・意味的整合性に関する比較評価を行う．

Towards Pixel Art Generation from Real Images in the Fashion Domain

1. はじめに

近年，メタバースや SNS におけるデジタルアバターの普及に伴い，実写画像をもとにスタイルやアイデンティティを視覚的に再構成する表現形式として，ピクセルアートの活用が注目されている．ピクセルアートは，限られた解像度と色数で視覚的特徴を抽象化する表現手法であり，レトロゲームや NFT アート，キャラクターデザインなどの文脈において独自の文化的・美的価値をもつ．なかでもファッションドメインにおいて，衣服の形や色，模様といったスタイリング要素を象徴的に伝達できる点は，アバター生成やデジタルファッションにおける応用可能性を広げており，その自動化はコンテンツ制作の効率化にも寄与する．

実写画像をピクセルアートに変換する最も単純な方法は，解像度を直接下げるダウンサンプリングである．しかしこの方法では，面積の小さなファッションアイテムの消失や，衣服の細かなパターンの喪失が生じやすく，視覚的・意味的な特徴が大きく損なわれる．特にファッションドメインにおいては，衣服の柄・シルエット・色といった要素はスタイリングを構成する本質的な特徴であり，これらの消失はコンテンツの価値を著しく低下させる．

こうした課題を克服すべく，近年では拡散モデル（Diffusion Model）を用いた高品質なピクセルアート生成の研究が進んでいる．なかでも SD-π XL[1] は，事前学習済みの text-to-image モデルである Stable Diffusion[2] と微分可能な生成器を用いて，Score Distillation Sampling（SDS）[3] により，実写画像からピクセルアートを生成する手法を提案している．この手法では，画像自体を最適化対象として扱い，テキストプロンプトとの意味的一致を促すことで，入力画像の意味的特徴を維持したデフォルメを実現している．

しかし，SDS による最適化が数千ステップと長く続く場合，生成されるピクセルアートは入力画像との視覚的な乖離が次第に大きくなり，スタイルを決定づける重要な特徴が失われるという課題がある．これは，テキストプロンプトとの整合性を過度に高めようとする最適化が，入力画像の特徴を過剰に変形してしまうことに起因する．また，SDS による最適化の停止判断には明確な基準が存在せず，品質が低下する前に最適なタイミングで停止することが必要である．

本研究では，SD-π XL における課題を踏まえ，SDS によ

[1] 明治大学大学院
Meiji University, Kawasaki, Kanagawa 214–8571, Japan
[a)] ce245033@meiji.ac.jp

る生成過程に対して早期停止戦略を導入することで，視覚的・意味的な破綻を防ぎつつ，より高品質なピクセルアートを生成する手法を提案する．具体的には，入力画像と生成画像の間の（1）ノルム距離に基づく視覚的変化の検出，(2）事前学習済み画像エンコーダによる埋め込み空間上の意味的類似度の変化の監視，という 2 軸から最適な停止タイミングを判断する手法を検討する．このアプローチにより，意味的・視覚的な特徴を保ちながら，実写画像を効果的にデフォルメし，高品質なピクセルアートを生成することを目指す．

2. 関連研究

2.1 意味的整合を考慮しない従来のピクセルアート生成

Nearest Neighbor 補完や Bicubic 補完などの単純なダウンサンプリングによるピクセルアート生成は，処理が軽量で実装も容易である．しかし，衣服の細かいパターンや，小物のような面積の小さな特徴が失われやすい．Gerstner らの提案 [4] では，SLIC を用いた多段階反復プロセスにより，実写画像を複数のスーパーピクセルに分割し，それらの色と形状を反復的に最適化することでピクセルアートに変換する．これにより，単純なダウンサンプリングと比べて，面積の小さな特徴が失われにくくなり，入力画像に忠実なピクセルアートが生成できる．しかし，意味的特徴の維持は考慮されておらず，重要な特徴の強調等の効果的なデフォルメは困難である．

2.2 学習が必要な実写画像からのピクセルアート生成

ピクセルアートの作成には高度な技術と多くの労力を要するため，実写画像とそれに対応するピクセルアート画像のペアデータを収集することは極めて困難である．このため，近年の学習ベースの手法では，ピクセルアート生成を教師なし学習の枠組みでスタイル変換タスクとして捉える研究が多く，特に GAN を用いたアプローチが一般的である．

Han らの提案 [5] では，ピクセル化と非ピクセル化という 2 つの変換タスクを，GAN を用いてそれぞれ異なるネットワークで双方向に学習することで，ペアデータを用いることなく，ピクセルアートを生成するモデルの訓練を可能にしている．この手法は，低解像度化，色数制限，エッジの強調といったピクセルアート特有のスタイルを模倣する点で一定の効果を示す．しかし，あくまでスタイル変換としてのアプローチにとどまるため，入力画像に含まれる意味的に重要な情報を効果的にデフォルメして保持・強調する仕組みは備えていない．

2.3 学習が不要な実写画像からのピクセルアート生成

実写画像とピクセルアートのペアデータの収集が困難であることから，学習を必要としないアプローチは有効な選択肢となる．[1] では，Stable Diffusion[2] と微分可能な生成器を組み合わせた SD-π XL という手法が提案されている．この手法は，まず実写画像をダウンサンプリングして得られた粗いピクセルアートを初期値として，生成器を構成する．生成器が出力する画像が，与えられたテキストプロンプトと整合するように，Score Distillation Sampling (SDS) [3] に基づいて生成器のパラメータが最適化される．SDS とは，あるテキスト条件のもとで望ましい画像に近づくよう，生成結果に対して勾配ベースで最適化を行う手法である．Stable Diffusion が生成画像を評価し，「どの方向に修正すればよりテキストと整合するか」を示す勾配（スコア）を算出することで，意味的整合性に基づいた更新が可能となる．このようなプロセスを通じて，通常のダウンサンプリングでは消失しやすい衣服の細かなディテールや小物類なども，テキストを介して重要性が強調され，視認性を保ったまま効果的にデフォルメされる．本研究では，こうした意味的特徴を保持しやすいという点から，SD-π XL の枠組みに基づくアプローチを採用する．

(a) 入力実写画像　　(b) 出力ピクセルアート

図 1: SD-π XL の課題

3. 手法

本研究では，SD-π XL の枠組みに基づいてピクセルアート生成を行い，その最適化プロセス（SDS）を途中で停止するという新たな操作を導入し，生成結果への影響を検討する．SD-π XL による生成過程においては，視覚的および意味的な特徴が適切にデフォルメされないケースが確認される．図 1 の上段は，視覚的特徴の保持に失敗した例である．実写画像では，青や緑，白色を基調とした花や葉の複雑な柄のワンピースを着用した女性が写っているが，生

成されたピクセルアートでは，それが黄色の単調な花柄へと変化している．花柄という意味的な特徴は維持されているものの，色彩やパターンが大きく乖離しており，視覚的特徴の保持という観点では不十分である．一方，図 1 の下段は，意味的特徴の保持に失敗した例である．実写画像にはサングラスをかけた男性が写っているが，生成されたピクセルアートでは，サングラスをかけているかどうかが判別しにくい．このように，視覚的には大きな乖離がないように見えても，意味的な特徴が適切に保たれていない場合がある．

視覚的・意味的な特徴の保持は，与えるテキストプロンプトに強く依存するため，最適化の過程において，必ずしも望ましいデフォルメができるとは限らない．しかし本研究では，あらかじめ定めたステップ数ではなく，視覚的または意味的な特徴が最も適切に保持されているタイミングで最適化を停止することを目指す．これは，最適な停止タイミングを検知できれば，将来的にはデフォルメがうまくいっていない場合に自動的に修正や再最適化を行うような，柔軟な制御が可能になると考えられるためである．本研究ではその第一段階として，視覚的・意味的な特徴の保持という観点から，最適な停止タイミングを評価・検出する方法の検討に取り組む．

3.1 視覚的一致に基づく停止

視覚的な特徴の保持に関する問題は，SD-π XL による最適化が進行するにつれて，テキストプロンプトとの整合性の向上という目的に引っ張られ，実写画像から視覚的に大きく乖離したデフォルメが生じることに起因する．この乖離を定量的に捉え，過剰なデフォルメを防ぐために，実写画像と生成ピクセルアートとの間の L2 ノルムを導入する．具体的には，実写画像 $\mathbf{x}_{\text{input}}$ と，最適化ステップ t における生成ピクセルアート $\mathbf{x}_t$ の間の L2 ノルムを $d_t = ||\mathbf{x}_t - \mathbf{x}_{\text{input}}||_2$ と定義し，以下の 2 つの停止基準を設ける．

- **固定閾値方式**: ノルム d_t が予め定めた閾値 ε を超えた時点で SDS を停止する．
- **増加率方式**: 初期ノルム d_0 に対する増加率 $r_t = \frac{d_t - d_0}{d_0}$ が閾値 ε を超えた時点で停止する．

固定閾値方式は単純だが，実写画像の内容やカラーパレット，ピクセルアートの解像度等によって，初期ノルムのスケールが異なる場合に対応しづらい．一方，増加率方式は初期値に対する相対的な変化量に着目するため，様々な条件においてより安定した停止判断が可能となる．これにより，視覚的特徴の乖離が大きくなる前に最適化を停止し，実写画像の視覚的な特徴を維持したピクセルアートの生成を目指す．

3.2 意味的一致に基づく停止

視覚的な特徴を維持することに加え，意味的な特徴を維持することも重要である．本研究では，事前学習済みの画像エンコーダを用いて，生成画像が入力画像の意味的な特徴をどれほど維持できているかを定量化する．まず，入力画像 $\mathbf{x}_{\text{input}}$ と現在のステップにおける生成画像 $\mathbf{x}_t$ を画像エンコーダに入力して，それぞれの埋め込み表現 $\mathbf{e}_{\text{input}}$ と $\mathbf{e}_t$ を得る．次に，この 2 つの埋め込み表現のコサイン類似度 $s_t = \cos(\mathbf{e}_{\text{input}}, \mathbf{e}_t)$ を計算する．この類似度 s_t が最大値に達した時点を，意味的な特徴が最も保持されているタイミングとみなし，SDS を停止する．エンコーダには以下の 2 種類を使用し，それぞれの特性を比較する．

- **CLIP**(openai/clip-vit-base-patch16[*1]): 最終層の CLS トークンを線形射影して得られるベクトル（テキストと画像の共通空間における埋め込み表現）を，画像全体の埋め込み表現として用いる．
- **DINOv2**(facebook/dinov2-base[*2]): 最終層の CLS トークンを画像全体の埋め込み表現として用いる．

これにより，意味的な特徴が保持されていない段階での生成を回避し，特徴の認識性に優れたピクセルアートを生成することを目指す．

4. 実験

4.1 データセット

実験に用いる実写画像として，Fashionpedia データセット [6] を用いる．このデータセットは，日常やイベントシーンにおけるファッション実写画像を集めたものであり，合計 48,825 枚の画像に対して，衣服ごとのセグメンテーションマスクと，詳細なカテゴリ・属性のアノテーションが付与されている．具体的には，46 のアパレルカテゴリ（例：shirt, pants, hat）および 294 のきめ細かな視覚属性（例：polo shirt, sweat pants, turtle neck）が定義されている．

このデータセットから，まず解像度が 1024 × 1024 の画像 3,878 枚を抽出した．これは，実験で使用する text-to-image モデルである SDXL [7] の標準解像度に合わせるためである．さらに，その中から人物が衣服を着用して写っている画像 868 枚を抽出し，可能な限り多様な衣服属性を含むように考慮した上で，最終的に 30 枚を実験用画像として選定した．

4.2 実験設定

視覚的・意味的一致に基づく停止戦略の有効性を検証するため，セクション 3 で提案した 4 つの停止戦略に加え，予め定めたステップ数で停止するという単純な停止戦略も含め，合計 5 つの異なる停止戦略の比較評価を行う．以下に各手法の概要と名称をまとめる．

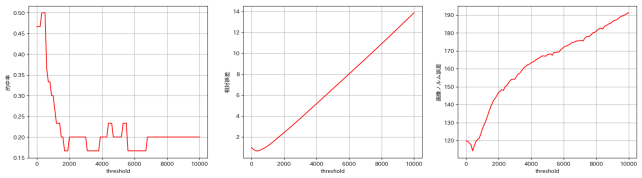
[*1] https://huggingface.co/openai/clip-vit-base-patch16
[*2] https://huggingface.co/facebook/dinov2-base

表 1: 定量評価

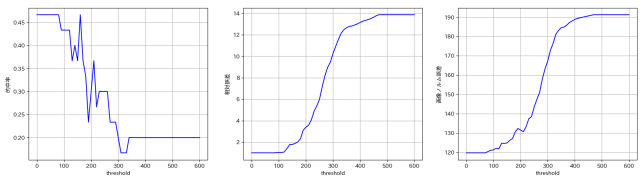
	的中率	相対中心誤差	画像ノルム誤差
固定ステップ法	0.500	0.676	114.224
ノルム閾値法	0.467	0.999	120.361
ノルム増加率法	0.467	0.707	118.678
意味的類似度最大法（CLIP）	0.400	5.169	137.308
意味的類似度最大法（DINOv2）	0.300	3.220	150.336

- **固定ステップ法**：最も単純な停止戦略であり，あらかじめ定めたステップ数に到達した時点で最適化を停止する．
- **ノルム閾値法**：実写画像と生成画像の L2 ノルム距離が，事前に定めた閾値を超えた時点で停止する．
- **ノルム増加率法**：上述の L2 ノルム距離の初期値に対する相対的な増加率が，事前に定めた閾値を超えた時点で停止する．画像ごとの L2 ノルム距離の初期値に対してスケール不変な制御が可能である．
- **意味的類似度最大法（CLIP）**：CLIP エンコーダを用いて，実写画像と生成画像の埋め込み表現を取得し，そのコサイン類似度が最大となった時点で停止する．
- **意味的類似度最大法（DINOv2）**：DINOv2 エンコーダを用いて，同様に埋め込み表現を取得し，そのコサイン類似度が最大となった時点で停止する．

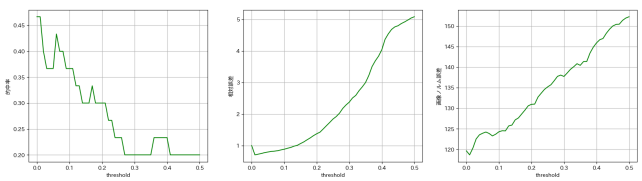
これらの手法を，SDS による最適化プロセスに適用し，ラベリングされた最適停止範囲との一致度をもとに，定量的に評価する．また，各停止タイミングで得られるピクセルアートについて，定性的な評価も行う．なお，出力画像のサイズは 64x64 に設定し，カラーパレットは各実写画像に k-means クラスタリングを適用することで，16 色を選定する．テキストプロンプトは，大規模言語モデルを用いて，実写画像のキャプションを生成し使用する．その他のパラメータは，SD-π XL のデフォルト値に準拠し，SDS は 10000 ステップ実行する．



(a) 固定ステップ法



(b) ノルム閾値法



(c) ノルム増加率法

図 2: 固定ステップおよび閾値を変化させたときの各手法における評価指標の推移

4.3 評価指標

30 枚の画像に対して，定量的および定性的に評価を行った．まず，1 名の被験者が各画像について，生成結果が良好なステップの範囲をラベリングした．ここでの良好とは，実写画像と比較して，視覚的および意味的な特徴が適切に保持されていると被験者が主観的に判断した範囲を指す．このラベリング結果を基準として，以下の評価指標により定量的な評価を行った．

- **的中率**：予測された停止ステップが，ラベリングされた範囲内に含まれている割合．値が高いほど，適切なタイミングで停止できていることを示す．
- **相対中心誤差**：予測された停止ステップとラベリング範囲の中心ステップとのステップ差を，ラベリング範囲の幅で正規化した値の平均値．値が小さいほど，理想的な停止タイミングに近いことを示す．ステップ差では，サンプル間のデフォルメの進行速度の違いを考慮できないと考え，本指標ではラベリング範囲の幅による正規化を行っている．
- **画像ノルム誤差**：予測された停止ステップにおける生成画像と，ラベリング範囲の中心ステップにおける生成画像の L2 ノルム距離の平均．生成画像と理想的な画像の視覚的な差異を定量的に表す指標であり，値が小さいほど再現性が高いことを示す．

5. 結果

5.1 定量評価

表 1 に，各手法における定量評価の結果を示す．図 2 は，固定ステップ法における停止ステップ，ノルム閾値法およびノルム増加率法における閾値を変化させた場合に，各評価指標がどのように変化するかを示している．表 1 の評価結果は，図 2 で得られた最も良好な性能を示す停止ステップおよび閾値を，各手法に適用した際のものである．

最も良好な性能を示したのは，固定ステップ法であり，最も高い的中率および最も低い誤差を記録した．ただし，

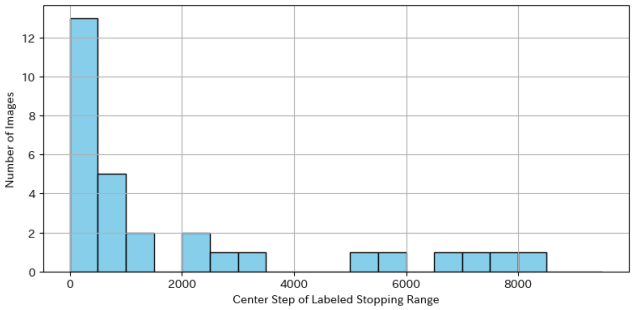


図 3: ラベリングした停止範囲の中心ステップの分布

図 3 に示すように，テストデータ 30 枚中の 6 割において，ラベリングされた最適停止範囲の中心が 1000 ステップ未満であった．このことから，今回用いたテストデータにおける早期停止傾向に強く依存した結果である可能性がある．従って，固定ステップ法が常に最適な手法であると断定することはできない．

次に良好な性能を示したのは，ノルム増加率法である．同様に視覚的一致に基づくノルム閾値法と比較すると，的中率は同じであるものの，相対中心誤差及び画像ノルム誤差において低い値を示した．これは，ノルム閾値法に比べて，ノルム増加率法の方が入力実写画像やカラーパレットなどの条件に依存しにくく，より汎化的に適切な停止タイミングを判断できたためと考えられる．ただし，ノルム閾値法とノルム増加率法においても，固定ステップ法と同様に，テストデータにおける早期停止傾向に依存した結果である可能性がある．

一方で，意味的一致に基づく停止手法は，視覚的一致に基づく手法に比べて，全ての指標で劣る結果となった．的中率が他の手法に比べて低く，誤差も大きいことから適切なタイミングでの停止が困難であったことが示唆される．これらの原因については，次節のケーススタディにおいて詳しく考察する．

5.2 ケーススタディ

図 4 に，テスト画像の 1 枚に対し，ピクセルアート生成過程における入力画像とピクセルアートの L2 ノルム，および CLIP と DINOv2 を用いた意味的な類似度をステップごとにプロットした結果を示す．緑および赤の点線は，それぞれ CLIP および DINOv2 による意味的類似度が最大値に達したステップを示しているが，いずれも灰色の帯で示したラベリングした停止範囲から大きく逸脱していることが分かる．図 5 は，図 4 の特徴的なステップにおいて生成されたピクセルアートである．(d), (e) はそれぞれ CLIP と DINOv2 による意味的な類似度の最大値に達したステップにおけるピクセルアートを示している．サングラスや，白と茶のチェック柄など一部のパーツは効果的にデフォルメできているが，ネックの形やジャケットの丈の長さなど誤ったデフォルメが行われているパーツも存在する．一方，今回のラベリングにおいては，重要なパーツが 1 つでも消失あるいは誤ったデフォルメがされた時点で停止するべきとする厳密な基準を用いた．このように，人手によるラベリングの停止基準と，CLIP や DINOv2 による停止基準が一致していないことが，意味的な類似度に基づく手法が他の手法に比べて大きく劣る原因の 1 つであると考えられる．本ケースは，30 枚のテスト画像の中でも典型的な例であり，意味的類似度に基づく停止手法が「画像の中の一部のパーツの特徴をよく捉えていても，適切な停止には至らない」ことを示している．このような傾向は，他の多くのテスト画像でも観察された．

この結果を踏まえた今後の課題として，次の 2 点が挙げられる．第一に画像全体ではなく，各パーツごとに意味的・視覚的特徴が保持されているかを個別に判断できる仕組みの導入が必要である．第二に，すべてのパーツが等しく重要であるわけではないという前提に立ち，パーツの重要度を定量的に評価し，重要なパーツの変化に着目した停止判断を行う手法の検討が求められる．画像全体として，視覚的・意味的特徴が維持されているかどうかの判断は非常に曖昧であり，人間であっても基準を設けなければ評価にばらつきが生じやすい．そのためどのパーツが重要か，そしてそのパーツの視覚的・意味的特徴が保持されているかを定量化することが，より信頼性の高い停止判断を実現するために必要である．

6. おわりに

本研究では，ファッションドメインにおける実写画像からのピクセルアート生成において，視覚的・意味的な特徴を保持しつつ効果的にデフォルメを行う手法を検討した．Stable Diffusion を用いた SDS によりピクセルアートを生成する SD-π XL をベースとし，その生成過程において複数の早期停止戦略を導入・評価した．

実験の結果，視覚的一致や意味的一致に基づく停止戦略が有効な手法となり得ることが示唆された一方で，テストデータの早期停止傾向に依存した結果も見られた．したがって，今後はデータ数を増加させ，より多様な条件下での評価を通じて，手法の有効性を精緻に検証する必要がある．

また，画像全体ではなく，画像内の各パーツごとに視覚的・意味的特徴の保持を個別に評価したり，パーツの重要度を定量的に考慮することが，より信頼性の高い停止判断に繋がる可能性があると示唆される．

参考文献

[1] Binninger, Alexandre, and Olga Sorkine-Hornung. ”SD-π XL: Generating Low-Resolution Quantized Imagery via Score Distillation.” SIGGRAPH Asia 2024 Confer-

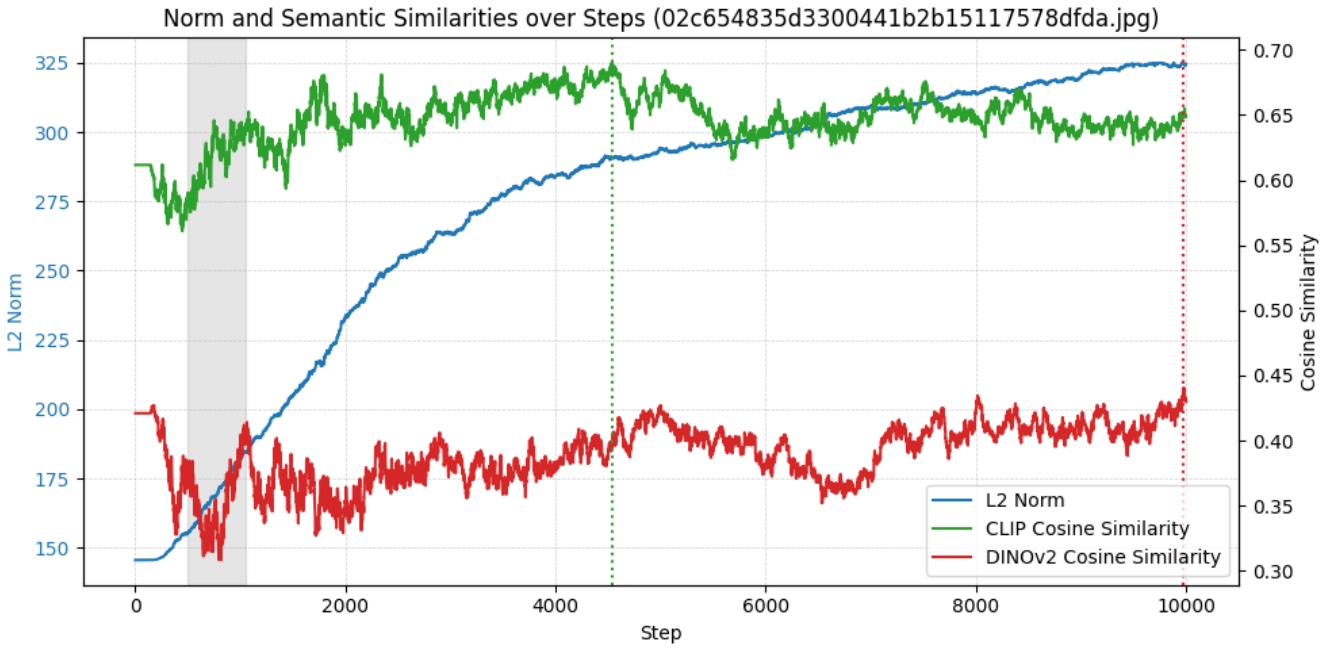


図 4: 1 枚のテスト画像における L2 ノルムおよび意味的類似度のステップごとの推移

(a) 入力実写画像　(b) 0 step　(c) 775 step　(d) 4535 step　(e) 9975 step

図 5: 図 4 と同じテスト画像におけるピクセルアート生成過程

ence Papers. 2024.
[2] Rombach, Robin, et al. ”High-resolution image synthesis with latent diffusion models.” Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.
[3] Poole, Ben, et al. ”Dreamfusion: Text-to-3d using 2d diffusion.” arXiv preprint arXiv:2209.14988 (2022).
[4] Gerstner, Timothy, et al. ”Pixelated image abstraction with integrated user constraints.” Computers & Graphics 37.5 (2013): 333-347.
[5] Han, Chu, et al. ”Deep unsupervised pixelization.” ACM Transactions on Graphics (TOG) 37.6 (2018): 1-11.
[6] Jia, Menglin, et al. ”Fashionpedia: Ontology, segmentation, and an attribute localization dataset.” European conference on computer vision. Cham: Springer International Publishing, 2020.
[7] Podell, Dustin, et al. ”Sdxl: Improving latent diffusion models for high-resolution image synthesis.” arXiv preprint arXiv:2307.01952 (2023).