拡散モデルを用いた 人物の後ろ姿からの正面服装画像生成

松雄 なずな[1,a)]　出口大輔[1,b)]　村瀬洋[1,c)]

概要

本発表では，拡散モデル Stable Diffusion [1] と ControlNet [2] を用いて，後ろ姿から正面画像を生成する “バックシャン拡散モデル” を提案する．近年，SNS の普及によって，服に関する情報を Instagram や TikTok などの投稿から手にいれる人が増えてきている．しかし，投稿されている写真は後ろ姿のみのことも多い．その為，画像生成技術を用いて後ろ姿から正面画像を生成する手法が求められる．そこで本発表では，拡散モデルの拡張機能 ControlNet から出力される制御特徴量を工夫して後ろ姿画像から正面画像を生成する手法を提案する．MVC データセット [3] での評価により，提案手法は服装をとらえた後ろ姿画像からの正面画像の生成に有効であることが確認できた．

1. はじめに

近年，Stable Diffusion [1] の登場に伴って画像生成技術が飛躍的な進化を遂げている．さらに，Stable Diffusion を拡張する枠組みである ControlNet [2] が提案され，特定の視点画像から任意視点画像の生成も可能になってきている．しかし，入力画像とは異なる視点の人物を生成しようとした場合，人物の服装の再現精度はまだ低い．一方，人の特徴が良く現れる正面向きの画像（以降，正面画像と呼ぶ）から後ろ姿の画像を再現する研究は行われているものの [4]，後ろ姿から正面画像を生成する研究は行われていない．しかしながら私たちは人の後ろ姿の服装からその正面から見た場合の服装を想像することができる．

このような背景から，本発表では ControlNet を拡張することで人物の後ろ姿から正面画像を生成する手法 “バックシャン拡散モデル” を検討する．具体的には，ControlNet を学習する際に正面画像が真の正面画像に類似する損失を考慮することで，より精度の高い正面画像生成を目指す．

[1] 名古屋大学
[a)] matsuo.nazuna.d8@s.mail.nagoya-u.ac.jp
[b)] ddeguchi@nagoya-u.jp
[c)] murase@nagoya-u.jp

2. 関連研究

2.1 拡散モデル（Diffusion Model）

Sohl-Dickstein らはマルコフ連鎖（遷移確率が現在の状態のみに依存し，それまでの履歴に依らない連鎖）を用いることで単純なガウス分布と複雑な分布を完全な形で対応させる過程 “ diffusion process” [5] を提案しており， Ho らはその手法を画像生成に応用する手法 [6] を発表した．拡散モデルは，元のデータにノイズを加えていく Forward Process（拡散過程）と，ノイズ分布の状態からノイズを除去することで画像データを作成する Reverse Process（逆拡散課程）の 2 つの過程で構成されている．

Forward Process は，入力画像データにランダムノイズを加えていき，最終的に入力画像データをガウス分布に変換する過程である．一方，Reverse Process は Forward Process の逆を辿り，ガウス分布から少しずつノイズを取り除いていきながら，画像を生成する過程である．この 2 つの過程を繰り返し行い，生成された画像が入力画像と類似するようにパラメータを学習することにより，ノイズから画像を生成することが可能となる．

2.2 Stable Diffusion

Stable Diffusion [1] は 2.1 項の拡散モデルを用いた代表的なテキストプロンプトを用いて画像生成を行う手法である．この手法は，拡散モデルの枠組みに基づき，Latent Seed（生成の元となるノイズ）と User Prompt（画像生成の指示文）の二つの入力を受け取り，512 × 512 画素の画像を出力する．

これまでの拡散モデルでは， 画像をピクセルの配列として直接扱い， ピクセル単位で拡散モデルを適用していた．しかし，この方法では知覚的にあまり重要でない画像の細部の表現にモデルの表現力が使われてしまったり，変数の次元が多いことから計算量が多くなるという問題があった．これらの問題を解決するために，VQGAN [7] と同様の仕組みを使い，入力画像を低次元の潜在表現へと変換してから拡散モデルを適用した． つまり， ノイズから画像

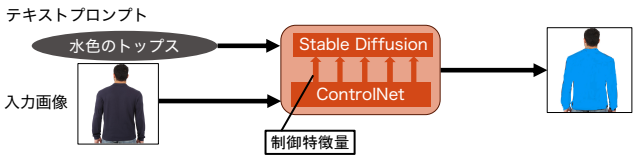

図 1 ControlNet を用いた際の画像生成

を生成する代わりに， まずノイズから潜在表現を生成し，それを画像へと戻すという 2 段階のプロセスを経て画像を生成している．このプロセスを経ることによってピクセル単位での計算よりもベクトルの次元を小さくすることができ， 計算量を減らすことに成功した．

また Stable Diffusion は， LAION-Aesthetics [8] と呼ばれる “ 美しい” 画像のみを集めたデータセットを用いて学習されている点も特徴的である． LAION-Aesthetics は大規模な画像キャプションデータセットである LAION [9] を元に構成されている．データセットの構築には，“ 美しさ” の人間に依る判定を模倣する事前学習済みモデルを用いている．

2.3 ControlNet

ControlNet [2] は， 2.2 項で述べた Stable Diffusion を拡張する手法の一つである．ControlNet では，まず学習済みの Stable Diffusion モデルのパラメータを固定し，ControlNet 部分の初期学習を 5 万ステップ以上行う．その後，Stable Diffusion の全てのパラメータを学習可能な状態にし，ControlNet 部分を含めてモデル全体の学習を行う．そうすることで画像の人物の姿勢や構図など特定の特徴を固定した状態で画像を生成するように Stable Diffusion モデルを制御することが可能になる．図 1 に示す通り，入力画像を ControlNet に，テキストプロンプトを Stable Diffusion に作用させながら学習を行う．また，ControlNet から Stable Diffusion にいくつかの制御特徴量が入力される．この制御特徴量によって画像の人物の姿勢や構図など特定の特徴を固定した状態で画像を生成することが可能となる．

3. バックシャン拡散モデル

本節では，ControlNet [2] の元の損失を工夫した “バックシャン損失” と，その損失を用いた拡散モデル “バックシャン拡散モデル” を提案する．以下では提案モデルで後ろ姿画像から正面姿画像を生成する生成過程と，学習過程の説明を行う．

3.1 バックシャン拡散モデルの生成過程

バックシャン拡散モデルを用いた画像生成は図 2 のように行われる．まず，後ろ姿画像とテキストプロンプト “Front side of the person” を用意する．そして，その後ろ

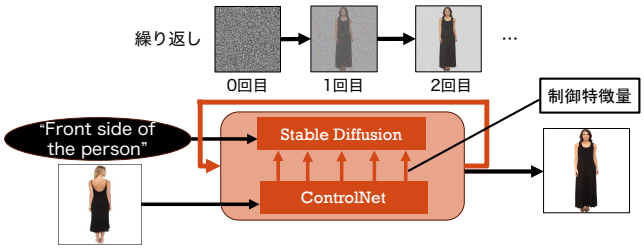

図 2 バックシャン拡散モデル生成過程

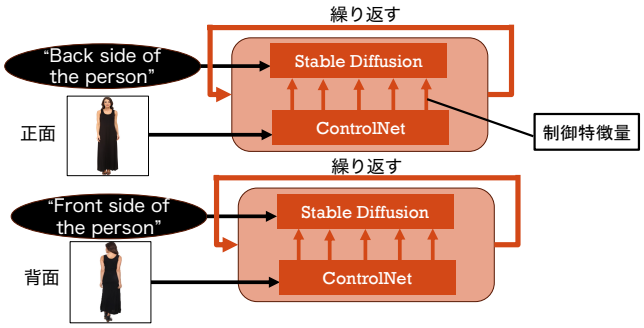

図 3 バックシャン拡散モデル学習過程

姿画像とテキストプロンプトを作用させながらバックシャン拡散モデルを繰り返し適用する．なお，各適用によってノイズのみの画像から段階的にノイズが取り除かれていく逆拡散の処理が行われる．その結果，目的画像である正面姿画像が生成される．

3.2 バックシャン拡散モデルの学習過程

バックシャン拡散モデルを用いた学習は図 3 のように行われる．同一服装を着用したペアとなる正面画像と後ろ姿を，その画像に対応するテキストプロンプトと共にそれぞれモデルに作用させながら，拡散と逆拡散を繰り返し，パラメータを更新していく．なお，パラメータを更新する際はバックシャン損失関数を考慮する．

続いて，モデルを繰り返し適用していく中のある一回，第 n 回の拡散と逆拡散の流れについて説明する．第 n 回の学習過程を示したのが図 4 である．正面画像とそれに対応する後ろ姿画像をそれぞれモデルに入力する．そして，正面画像にノイズを付加し，その画像を ControlNet に通す過程で生成される制御特徴量 β_f と，後ろ姿画像を ControlNet に通す過程で生成される制御特徴量 β_b の間のコサイン類似度を計算し，その値が大きくなるように交差エントロピーを用いた損失を用いて学習を行う．また，それに加えて元の ControlNet の損失である平均二乗誤差損失も計算する．以下では損失関数について詳しく説明する．

3.3 バックシャン損失

バックシャン損失は元の Stable Diffsuion の損失関数である平均二乗誤差損失と，コサイン類似度と交差エントロ

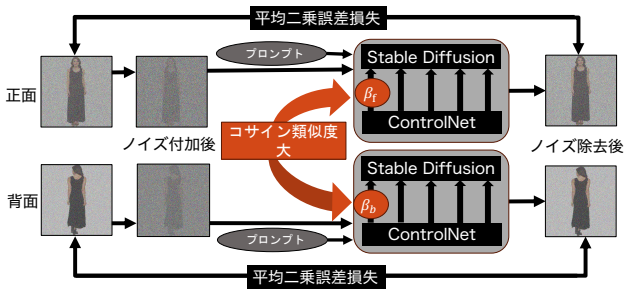

図 4 第 n 回目のバックシャン拡散モデル学習過程

ピーを用いた損失の和からなる．以下では平均二乗誤差損失について説明した後，コサイン類似度と交差エントロピーを用いた損失について説明する．

3.3.1 平均二乗誤差損失

まず，元の Stable Diffsuion の損失関数であった平均二乗誤差損失について説明する．平均二乗誤差損失は Stable Diffusion の拡散過程で用いられている損失関数である．Stable Diffusion を用いた画像生成は，

(1) 入力画像にノイズを段階的に追加（拡散過程）
(2) ノイズのみの画像に変換
(3) 画像からノイズを段階的に除去（逆拡散過程）

という流れで行われる．逆拡散過程のある段階でどのノイズを除去するかを予測する際に，その予測が拡散過程の対応する段階で実際に加えられたノイズとなるべく近くなるよう用いられるのが平均二乗誤差損失（MSE Loss）である．つまり，あるノイズ付加段階で実際に加えられたノイズ N_{gt}^{front} と除去段階で取り除くべきノイズ N_{pred}^{front} の平均二乗誤差を平均二乗誤差損失という．

3.3.2 コサイン類似度と交差エントロピーを用いた損失

続いてコサイン類似度と交差エントロピーを用いた損失について述べる．i 番目の画像を用いて学習する際のコサイン類似度と交差エントロピーを用いた損失 $\mathcal{L}_{CE}^{i}$ は次式で与えられる．

$$\mathcal{L}_{CE}^{i} = CE(D, l_i) = -\sum_{i} l_i \log(D) \tag{1}$$

$$D = -\sum_{k=0}^{batch} \frac{\beta_f^k \cdot \beta_b^{k\top}}{\sum_{l=0}^{batch} \left\|\beta_f^k\right\| \cdot \left\|\beta_b^l\right\|} \tag{2}$$

なお $batch$ はバッチサイズ，β_f^i，β_b^i は ControlNet から Stable Diffusion に出力される制御特徴量を表す．また，l_i は生成する画像のインデックスを表す 1-hot ベクトルである．具体的には，i 番目の要素のみが 1 で，他の要素が 0 のベクトルを表している．式 (2) により求めたコサイン類似度 D を用いて交差エントロピーを計算して損失とする．この損失を用いることによって，図 5 からも分かるように，インデックスが同じ制御特徴量のコサイン類似度が大きくなるように学習が進められる．これにより，後ろ姿画像と正面画像の対応関係が保持されるように学習が進み，より精度の高い正面画像生成を行う．また，ノイズを付加した画像からノイズを除去した後の画像と，ノイズを付加する前の画像間の平均二乗誤差損失が小さくなるように学習を進める．こうすることで後ろ姿画像と正面画像の対応関係が保持されるように学習を進めることが可能になる．

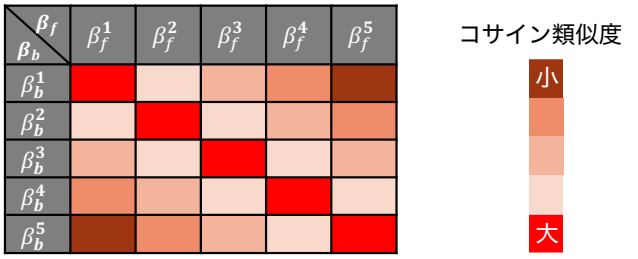

図 5 コサイン類似度に交差エントロピーを作用させた結果

4. 実験

訓練済みのバックシャン拡散モデルを用いて被験者実験を行なった．本節ではまず実験設定を示し，次に実験結果について述べる．

4.1 実験設定

本実験には，商品の着衣画像を様々な角度から撮影している MVC データセット [3] から目的の角度から正面と後ろ姿の画像の組を抽出して用いた．被験者実験の手順を以下に示す．

(1) 被験者に後述する注意事項を伝える
(2) 被験者に対して入力の後ろ姿画像，ベースラインと提案手法それぞれの出力を示す．画像は入力画像，ベースラインで生成した画像，提案手法で生成した画像の 3 枚 1 組を，全 20 組用意した．
(3) 図 6 に示す実験用紙で，入力画像に対してより自然な正面画像を 2 枚の画像から選ぶ．

なお，実験画像はランダムな表示順序で示した．回答は 1 人につき 1 度のみとし，調査結果の収集が完了するまで参加者同士の情報の共有を禁止した．また，注意事項として服装以外の身体特徴（顔や手，脚など）や背景は判断の際に考慮しないよう伝えた．評価は 10 分の制限時間を設けて行った．

4.2 実験結果

4.1 項で述べた方法に従い，5 人の被験者に対して被験者実験を行った結果，68%の画像において提案手法による生成結果は後ろ姿から想像できる正面画像に近いという結果が得られた．

5. 考察

バックシャン拡散モデルの有効性を証明するため，損失

図 6 実験用紙

図 7 生成例

関数を変更する前のベースラインと比較し，提案手法の有効性について考察する．図 7 より，ベースラインに比べて色や服のシルエットが保持できていることが確認できる．これは，損失関数を変更したことにより，入力された後ろ姿画像の特徴をより保持するように学習したからであるといえる．特に，ポロシャツが再現できたのは，ControlNet で制御特徴量で生成する際に画像のエッジを検出する Canny モデルを用いたからであるといえる．Canny モデルを用いたことで服のシルエットが忠実に再現された結果，襟が再現されたのだと考えられる．一方，図 7 の正解画像の胸にあるロゴなど細かな点の再現は不可能だったため，精度を上げるために損失関数をさらに工夫する必要がある．また，今回は顔も含めて学習を行ったが，目的は服装を再現することなので顔の情報は必要なく，服装の特徴を捉えるにあたってはノイズとなるためマスク必要であろう．

6. むすび

本発表では，本発表では，拡散モデル Stable Diffusion [1] を用いて，ControlNet の制御特徴量をコントロールする損失を追加することで後ろ姿画像から正面画像を生成する “バックシャン拡散モデル” を提案した．バックシャン拡散モデルを用いることで従来手法と比較し，色や服のシルエットの再現性において有効であることが確認できた．今後の課題として，精度向上のための顔のマスキング処理，などが挙げられる．

7. 謝辞

本研究の一部は JSPS 科研費 23H03474 による．本研究の一部は名古屋大学のスーパーコンピュータ「不老」の一般利用にて実施した．

参考文献

[1] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-resolution image synthesis with latent diffusion models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 10 684–10 695.

[2] L. Zhang, A. Rao, and M. Agrawala, “Adding conditional control to text-to-image diffusion models,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2023, pp. 3836–3847.

[3] T.-Y. C. Kuan-Hsien Liu and C.-S. Chen, “Mvc: A dataset for view-invariant clothing retrieval and attribute prediction,” June 2016, pp. 313–316.

[4] T.-C. W. I. K.-S. Johanna Karras, Aleksander Holynski, “Dreampose: Fashion image-to-video synthesis via stable diffusion,” 2023.

[5] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, “Deep unsupervised learning using nonequilibrium thermodynamics,” in *Proceedings of the 32nd International Conference on Machine Learning*. PMLR, 07–09 Jul 2015, pp. 2256–2265.

[6] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” in *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2020, pp. 6840–6851.

[7] P. Esser, R. Rombach, and B. Ommer, “Taming transformers for high-resolution image synthesis,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 12 873–12 883.

[8] C. Schuhmann, “Laion-aesthetics,” 2022.

[9] C. Schuhmann, R. Vencu, R. Beaumont, R. Kaczmarczyk, C. Mullis, A. Katta, T. Coombes, J. Jitsev, and A. Komatsuzaki, “Laion-400m: Open dataset of clip-filtered 400 million image-text pairs,” 2021.