拡散モデルを用いた後ろ姿からの正面画像生成の検討

松雄 なずな † 出口 大輔 † 村瀬 洋 †

† 名古屋大学

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，素材，柄といった詳細な部分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 [3]，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から興味のある正面画像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ような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．

2 バックシャン拡散モデル

本節では，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” の概要と ControlNet に追加する新しいバックシャン損失について述べる．

2.1 手法の概要

まず，Stable Diffusion の拡張手法の一つである ControlNet の概要について説明する．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．本発表では，この ControlNet の損失関数を改良することで後ろ姿からの正面画像生成の高精度化を図る．

2.2 バックシャン損失

後ろ姿と正面画像の対応関係を損失に加えた新しい損失 $\mathcal{L}$ は次式で与えられる．

$$\mathcal{L} = \mathcal{L}_0 + \alpha CE(d, l) \tag{1}$$

$$d = \sum_{i=0}^{N} \sum_{j=0}^{N} \frac{cos(f_i, b_i)}{cos(f_i, b_j)} (i \neq j) \tag{2}$$

ここで，$\mathcal{L}_0$ は従来の ControlNet の学習に利用される損失であり，$CE()$ は提案するバックシャン損失である．α は $\mathcal{L}_0$ と $CE()$ の割合を調整するハイパーパラメータであり，l は生成する画像のインデックスを示している．N は正面および後ろ姿の画像特徴量の要素数，f，b はそれぞれ正面と後ろ姿の画像特徴量を意味している．バックシャン損失の追加により，図 2 のような後ろ姿画像とそれを入力として生成した正面画像のコサイン類似度が大きくなるように学習される．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が可能である．

A preliminary study on front image generation from back by diffusion model.

† MATSUO Nazuna, DEGUCHI Daisuke, MURASE Hiroshi, Nagoya University

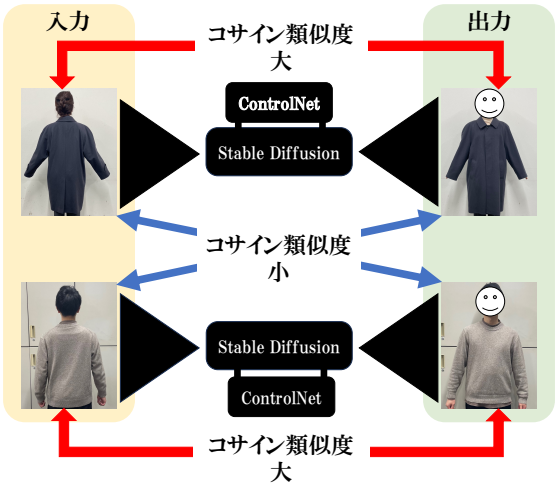

図 1　バックシャン損失の追加による学習効果

3　実験

3.1　実験方法

MVC データセット [4] から 30,293 組の画像を抽出し，学習データとして用いた．また，学習済み Stable Diffusion モデルとしては stable-diffusion-v1-5[*1]を利用し，ControlNet のモデルには sd-controlnet-canny[*2]を用いた．

3.2　実験結果

図 2 は，実際に撮影した後ろ姿画像を用いて従来の ControlNet によって生成された正面画像と，提案手法である “バックシャン拡散モデル” を用いて生成された正面画像を示している．左から順に入力となる後ろ姿画像，従来の ControlNet によって生成された正面画像，提案手法 “バックシャン拡散モデル” を用いて生成された正面画像，入力画像に対応する正面画像を示している．ControlNet によって生成された画像に比べ “バックシャン拡散モデル” では首元の詰まり具合が再現できていて、入力画像と類似した服装が生成できていることが確認できる．

[*1] https://huggingface.co/runwayml/stable-diffusion-v1-5

[*2] https://huggingface.co/lllyasviel/sd-controlnet-canny

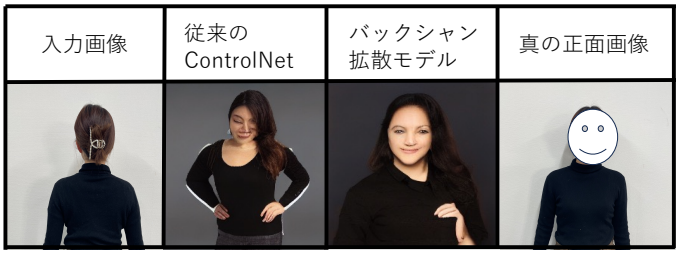

図 2　提案手法であるバックシャン拡散モデルを用いて正面画像を生成した結果

4　むすび

本発表では，従来の ControlNet に対して後ろ姿画像と正面画像の類似度を学習時に考慮することにより，精度良く正面画像が生成可能な “バックシャン拡散モデル” を提案した．従来の ControlNet では平均二乗誤差損失のみしか用いていなかったため，後ろ姿画像から正面画像を生成した際に服の素材など細かな点の再現が困難であった．これに対し，提案するバックシャン損失を導入することにより，服装等の細かな生成も可能であることを確認した．今後の課題として，衣服の同一性を効果的に再現可能な新しい損失の検討，正面画像生成システムの実装，などが挙げられる．

謝辞 本研究の一部は JSPS 科研費 23H03474 による．

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 *CVPR*, 2022, pp. 10 684–10 695.

[2] L. Zhang, A. Rao, and M. Agrawala, “Adding conditional control to text-to-image diffusion models,” in *ICCV*, 2023, pp. 3836–3847.

[3] J. Karras, A. Holynski, T.-C. Wang, and I. Kemelmacher-Shlizerman, “Dreampose: Fashion image-to-video synthesis via stable diffusion,” 2023.

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