4P-02

StyleGAN とマルチタスク学習を用いたテキストからの 3D 人体モデル作成

リュウ ホウ[†1]　澤田 隼[†2]　大村 英史[†3]　桂田浩一[†4]
東京理科大学 創域理工学研究科 情報計算科学専攻

1. はじめに

3D 人体モデルは映画，ビデオゲーム，AR/VR などの分野で広く使用されており，仮想世界を構築する際には欠かせないビジュアル要素である．しかし 3D 人体モデルを作成するにはデザイン能力，デッサン力，そして 3DCG 制作ソフトウェアの知識が必要であり，初心者にとっては難易度が高い．さらに，人体モデルの制作には単純な繰り返し作業が多く，モデルを作成するには大量の時間と労力が必要である．そのため，必要なスキルと知識を取得したとしても，3D 人体モデルを完成させる者は限られている．

この問題に対して，Hong らは 3D 人体モデルを自動的に生成する AvatarCLIP[1]を提案した．AvatarCLIP は入力されたテキストに基づいて 3D 人体モデルとアニメーションを生成する．しかし，AvatarCLIP で生成されたモデルはテクスチャの質が低く，毛髪の立体感が欠けているという問題があった．

そこで本研究では，StyleGAN[2]とマルチタスク学習を用いることでテクスチャの質を向上するとともに，パーティクルヘアのパラメータを出力することで立体的な毛髪の生成を可能にする手法を提案する．本稿では客観的および主観的な評価実験を行い，提案手法によってテキスト通りの 3D 人体モデルが生成できていることを確認する．

2. 前提知識

2.1 3D 人体モデル

本研究では 3D 人体モデルをメッシュ，テクスチャ，パーティクルヘアの 3 つの部分に分けている．メッシュとはオブジェクトの形状を定義する頂点，辺，面の集合である．テクスチャ画像はメッシュに貼り付けて物体の質感を表現する画像である．パーティクルヘアはパーティクルシステムを活用して毛髪を作成する手法である．メッシュ部分の生成には SMPL-X[3]を採用し，テクスチャ画像は StyleGAN によって生成した．以下で，SMPL-X と StyleGAN について説明する．

2.2 SMPL-X

SMPL-X は低次元パラメータで操作可能な人体モデルである．体型パラメータ，姿勢パラメータ，表情パラメータを入力するだけで表現力豊かな 3D モデルを作成できる．

本研究では SMPL-X モデルに 2 点の改良を加えた．(1)目，鼻，口，耳の形状を制御できるように，4 つの顔の形状パラメータを追加した．(2)身長と体重を入力して体型パラメータを取得できるレグレッサーを実装した．以上，本研究では，性別パラメータ，顔のパラメータ，身長と体重のパラメータを入力することにより，対応する SMPL-X モデルを生成する．これにより，パラメータ全体がより直感的で理解しやすくなった．

2.3 StyleGAN

StyleGAN は PG-GAN に基づいて段階的に画像を生成するモデルである．StyleGAN は優れた学習能力を持ち，学習させるデータセットによって，様々なジャンルの高品質な画像を生成できる．本研究では，2 万枚の SMPL-X の顔テクスチャを含むデータセットを用いて StyleGAN を訓練した．訓練済みの StyleGAN は高品質な SMPL-X の顔テクスチャを生成できる．

2.4 マルチタスク学習

マルチタスク学習とは，関連性を持つ複数の学習タスクを 1 つのモデルに同時に学習を行わせることで精度を向上させる手法である．複数の学習タスクを同時に学習させることで，異なるタスクから得られる知見や情報を相互に活用でき，各タスクのパフォーマンスが向上できる．また，マルチタスク学習はシングルタスク学習と比べ，一部の構造を共有することでメモリ使用量を削減し，推論時の計算を減少させ，推論速度を向上できるというメリットもある．

3. 提案手法

図 1 に示すように，提案手法のアルゴリズムは以下のとおりである．

(1)入力されたテキストを Bert Tokenizer で分割する．
(2)分割されたトークンを BERT モデルに入力し，Text Embedding を出力する．
(3)得られた Text Embedding は全結合層を通した上で，マルチタスク学習モデルの各タスクに入力する．
(4)マルチタスク学習モデルは Text Embedding から StyleGAN の潜在変数，髪型のインデックス，性別パラメータ，顔のパラメータ，体型のパラメータを同時に予測する．
(5)予測された StyleGAN の潜在変数を StyleGAN ジェネレーターに入力し，テクスチャ画像を生成する．
(6)予測された髪型のインデックスに基づき，髪型のデータ

Generating 3D Human Models from Text Using StyleGAN and Multi-Task Learning
†1 LIU FENG, Tokyo University of Science
†2 SAWADA SHUN, Tokyo University of Science
†3 OHMURA HIDEFUMI, Tokyo University of Science
†4 KATSURADA KOUICHI, Tokyo University of Science

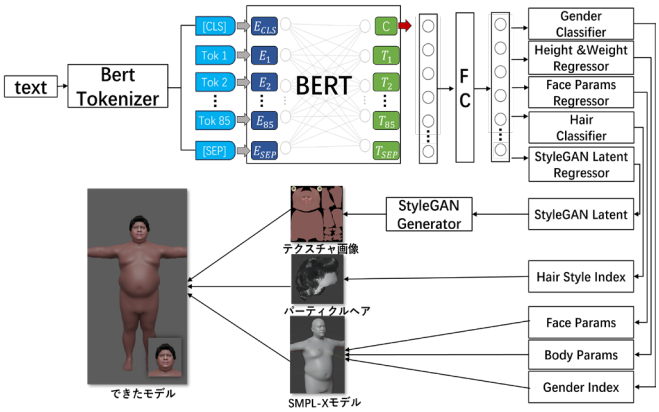


図 2 提案手法のアルゴリズム

表 1 マルチタスク&シングルタスク 客観的評価実験結果

	マルチタスク	シングルタスク
性別パラメータ正解率	**96.53%**	95.25%
顔パラメータ正解率	89.42%	**90.15%**
体型パラメータ正解率	**88.64%**	82.45%
平均推論時間	**68.21s**	102.50s

を読み込んで Blender に入力する．

(7)予測された性別パラメータ，体型パラメータと顔パラメータを使用して対応する SMPL-X モデルを生成し，Blender に入力する．

(8)テクスチャ画像，SMPL-X モデル，髪型を組み合わせ，3D 人体モデルを完成させる．

4. 実験

4.1 客観評価実験

客観評価実験の目的は，マルチタスクとシングルタスクの性能比較である．具体的には，1000 個のランダムなテキストを用いて 3D 人体モデルのパラメータを予測する．予測されたパラメータの正解率及び平均推論時間を指標として評価する．

表 1 に示すとおり，各パラメータの正解率に関しては，マルチタスクとシングルタスクの間で大きな差はない．一方で，平均推論時間では，マルチタスクの方が優れ，計算資源の利活用が適切に行われていることが分かった．

4.2 主観評価実験

主観評価実験は 2 つの実験から構成されている．実験 1 では，5 つのランダムなテキスト，および 5 つの小説に含まれる人物の外見を描写したテキストを使用して人体モデルを生成し，評価者に 5 段階で評価してもらう（1 点は全く一致しない，5 点はよく一致することを意味する）．実験 2 では，先行研究の AvatarCLIP と本研究で同じテキストを使用し，それぞれ人体モデルを生成させる．評価者はどちらが高品質な人体モデルを生成できるかを比較評価する．

合計 88 人の被験者がアンケートに回答した．実験結果は

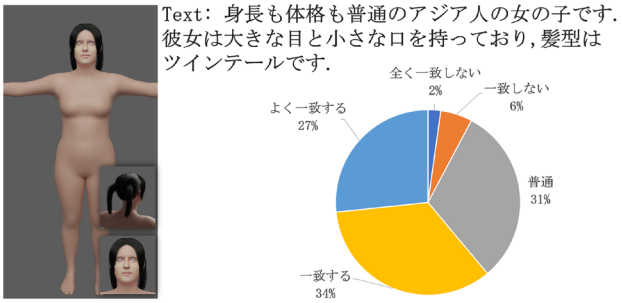
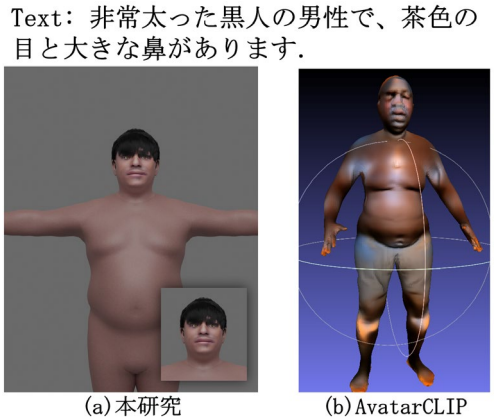


図 3 主観評価実験 1 の結果の一部



(a)本研究　(b)AvatarCLIP

図 1 主観評価実験 2 の結果の一部

以下の通りである．ページ数の制限により，結果の一部のみを示す．図 1 に示すように，実験 1 では，10 個のテキストの内，「一致する」または「よく一致する」を選んだ被験者は 73.3%であり，平均の評価は 5 段階中 4.01 であった．また，実験 2 で生成された人体モデルは，本研究が図 3（左），AvaterCLIP が図 3（右）のようなものであった．3 つのテキストから生成された人体モデルを被験者に評価させたところ，本手法で生成された人体モデルの方がクオリティが高いと評価した被験者は 84.83%であった．

実験結果より，本研究で生成された人体モデルは先行研究と比べて良い評価を得ていることが確認できた．

5. まとめ

本研究では StyleGAN とマルチタスク学習を用いることでテクスチャの質を向上するとともに，パーティクルヘアのパラメータを出力することで立体的な毛髪の生成を可能にする手法を提案した．客観評価実験および主観評価実験を通して，本手法の有効性が確認できた．今後は服装の自動生成等も検討したい．

参考文献

[1] Hong, F. et al. “Avatarclip: Zero-shot text-driven generation and animation of 3d avatars.” In arXiv preprint arXiv:2205.08535 (2022).
[2] Karras, T. et al. “Analyzing and improving the image quality of stylegan,” in Proc. IEEE/CVF conference on computer vision and pattern recognition, pp.8110-8119 (2020).
[3] Pavlakos, G. et al. “Expressive body capture: 3d hands, face, and body from a single image,” in Proc. IEEE/CVF conference on computer vision and pattern recognition, pp.10975-10985 (2019).