定期ゼミ

コンシューマー向け深度センサーを用いたスキャンによる3D顔再構成

2024年6月21日(金)

早稲田大学　基幹理工学研究科
電子物理システム学専攻 史研究室
石黒将太郎

アウトライン

1. 研究テーマ
2. 関連研究紹介
   - *Latent Diffusion Models*
   - *3DMM*
   - *DiffusionRig*
3. 提案モデル
4. 今後の計画

1. 研究テーマ

研究テーマ

研究目的

- 一般的なセンサーを用いてリアルなCGアバターを生成
  - フォトリアルなCGアバターを生成するには、専用システムが必要
  - 3Dメッシュ上の反射率や光源データがあった方が、顔のレンダリング結果がリアル
  - 既存手法はGANを用いた手法が多い

研究内容

- 拡散モデルを用いた3D顔再構成器の開発
  - 学習の安定性・生成結果の多様性に優れる拡散モデルを用いる
  - iPad Proに搭載されているTrue-Depthカメラ由来のデータから3D顔再構成を行う
  - 自然さ・アイデンティティを保ちながら、表情が操作された3D顔を出力

2. 関連研究紹介

R. Rombachらの研究[8]

High-resolution image synthesis with latent diffusion models

- 生成過程での計算コストの高さを改善するために、潜在空間でのデノイズを提案
  - ピクセル空間の正規分布ノイズから、100～1000段階に分けてノイズを除去することで生成
  - 人間には知覚できない高周波特徴をオートエンコーダーによって除去してから拡散モデルで生成
- 訓練済みのVAEは様々なタスクに応用可能
- VAEによるダウンサンプリングは以下の因子fに従う

$$f = \frac{H}{h} = \frac{W}{w} = 2^n (m \in N)$$

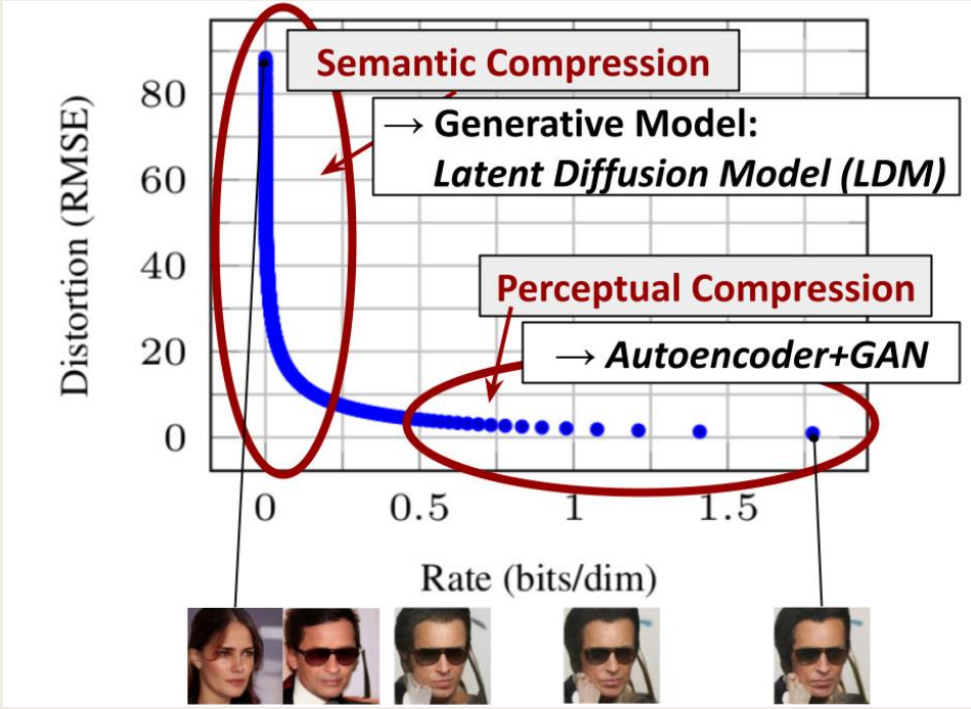

図. 生成過程の段階分け[8]

2. 先行研究紹介

[8] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-resolution image synthesis with latent diffusion models,” CoRR, vol. abs/2112.10752, 2021.

R. Rombachらの研究[8]

High-resolution image synthesis with latent diffusion models

- 5段階のダウンサンプリング因子fをVAEに適用した場合の評価指標FID、ISを比較
  - fが大きいほど圧縮率が高い(f=8の場合、1048×1048→131×131)
  - FID：特徴空間での分布距離を測定、IS：活性化マップの確立分布の差異を測定
- $f = 4, 8$の方がピクセル空間でのモデル(f=1)より高性能

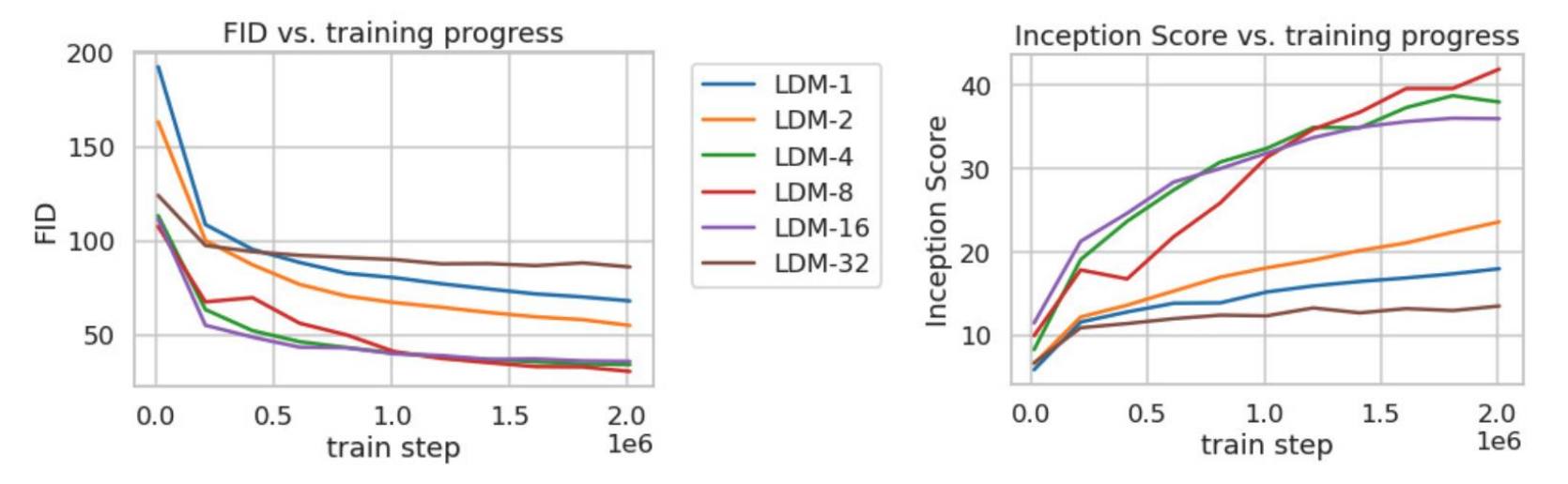

図. ダウンサンプリング因子fとFID、ISの関係[8]

[8] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-resolution image synthesis with latent diffusion models,” CoRR, vol. abs/2112.10752, 2021.

R. Rombachらの研究[8]

High-resolution image synthesis with latent diffusion models

- 5段階のダウンサンプリング因子fをVAEに適用した場合の評価指標FIDと生成速度を比較
  - CelebA-HQとImageNetを用いて学習、NVIDIA A100により生成
  - FID：特徴空間での分布距離を測定、Throughput：1秒毎の生成枚数(ノイズスケジューリングに依存)
- $f = 4, 8$を適用したモデルが最も生成速度と生成画像の精度のバランスが取れている

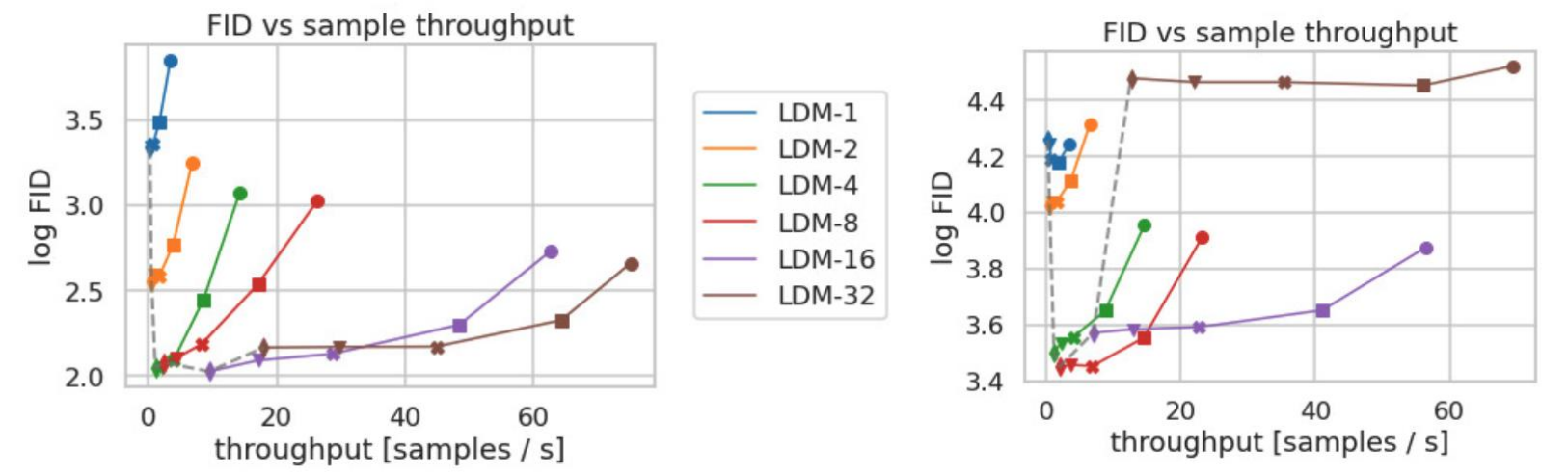

図. ダウンサンプリング因子fとFID、スループットの関係[8]

[8] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-resolution image synthesis with latent diffusion models,” CoRR, vol. abs/2112.10752, 2021.

V. Blanzらの研究[11]

3DMM

- 事前データセットから形状基底とテクスチャ基底を抽出

メリット

- PCAによる圧縮により、有限次元数で3D顔モデルを作成可能
- トポロジーー貫性を持つため、特定の顔領域の変化や、他人との顔置換、アクセサリーなどの後付けが得意
- パラメトリックな変化(表情や年齢、性別)が可能

デメリット

- 基底抽出データセットの影響を強く受ける
- 非線形・繊細な表現がしにくい
- 髪や瞳、歯などは3DMMの表現範囲外

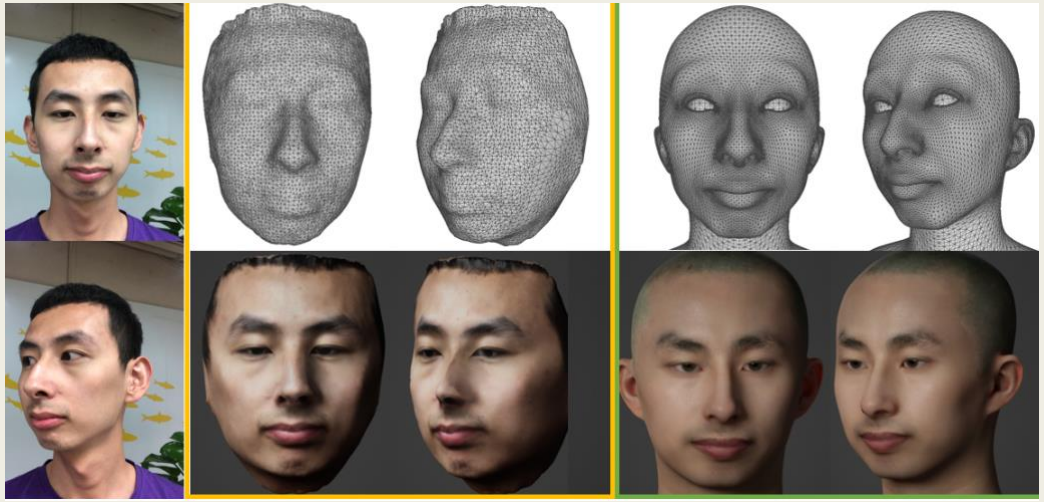

図. 3DMMを用いた場合と用いなかった場合[12]

[11] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3d faces," p. 187–194, 1999.
[12] X. Lin, Y. Chen, L. Bao, H. Zhang, S. Wang, X. Zhe, X. Jiang, J. Wang, D. Yu, and Z. Zhang, "High-fidelity 3d digital human creation from RGB-D selfies," CoRR, vol. abs/2010.05562, 2020.

Zheng Dingらの研究[13]

DiffusionRig: Learning personalized priors for facial appearance editing.

- 目的
  - 20枚ほどの同一人物ポートレート写真から人物特有の顔の特徴を学習
  - 顔の特徴やアイデンティティを保持しながら、表情・ライティング・顔の向きを編集
- 特徴
  - 大規模データセットから学習を行うstage1とターゲット人物の特徴を学ぶstage2に分ける
  - 外観を編集するために、パラメトリックな3D顔モデルであるFLAME [15]を拡散モデルの条件に使う

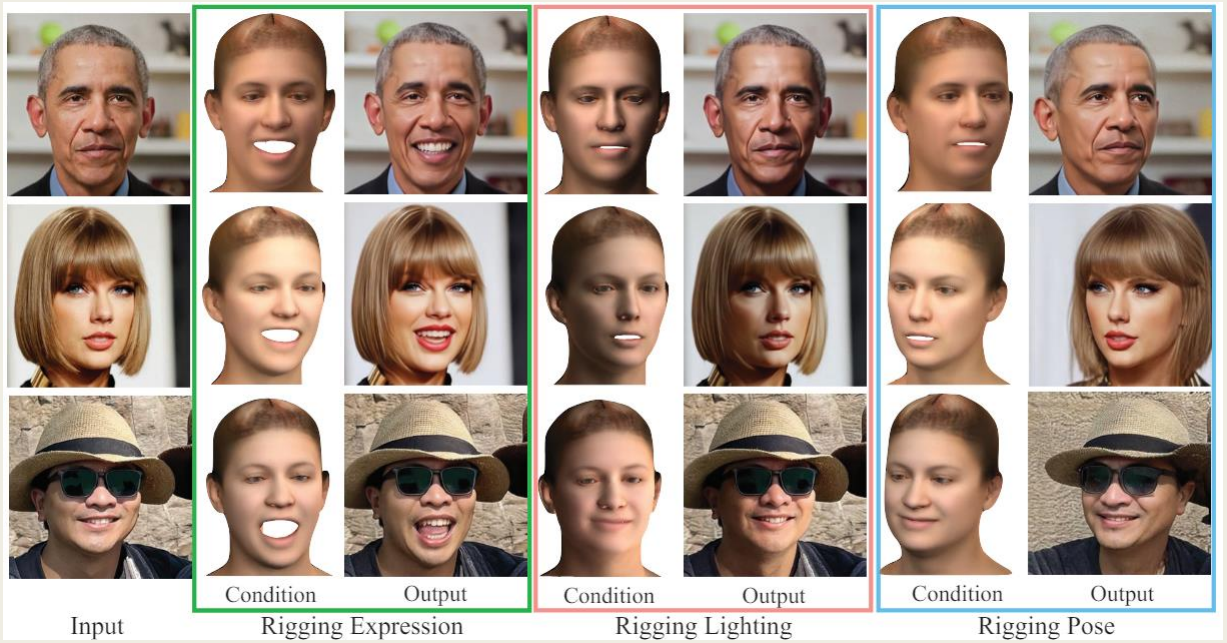

図. DiffusionRigによる編集例[13]

2. 先行研究紹介

[13] Z. X. L. J. Z. T. Zheng Ding, Cecilia Zhang and X. Zhang, “Diffusionrig: Learning personalized priors for facial appearance editing,” 2023.
[15] T. Li, T. Bolkart, M. J. Black, H. Li, and J. Romero, “Learning a model of facial shape and expression from 4d scans,” ACM Trans.Graph., vol. 36, no. 6, nov 2017.

Zheng Dingらの研究[13]

DiffusionRig: Learning personalized priors for facial appearance editing.

- アーキテクチャ
  - 3DMMの作成には学習済みDECAモデル[14]を使用
  - 顔の特徴やアイデンティティを保持しながら、表情・ライティング・顔の向きを編集
  - 3DMMが生成できない特徴のみを扱うEncoderを使用し、学習は大規模データセットのみで行う

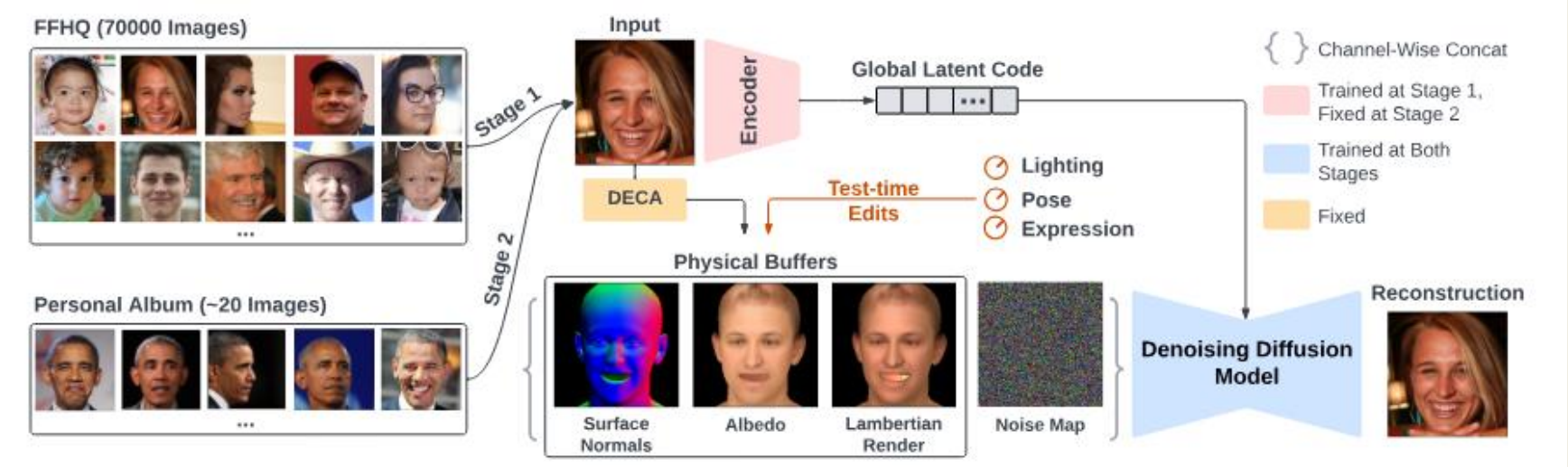

図. DiffusionRigのアーキテクチャ[13]

[13] Z. X. L. J. Z. T. Zheng Ding, Cecilia Zhang and X. Zhang, "Diffusionrig: Learning personalized priors for facial appearance editing," 2023.
[14] Y. Feng, H. Feng, M. J. Black, and T. Bolkart, "Learning an animatable detailed 3d face model from in-the-wild images," CoRR, vol. abs/2012.04012, 2020.

Zheng Dingらの研究[13]

DiffusionRig: Learning personalized priors for facial appearance editing.

- 出力画像の比較
  - ターゲット人物のアイデンティティを保持できている
  - トポロジーー貫性を保ちながら、物理的に基づいた方法で外観を編集できているため、不自然さが少ない
  - 制御性・解釈性に優れる
  - 髪や背景、メガネなども自然に出力できている

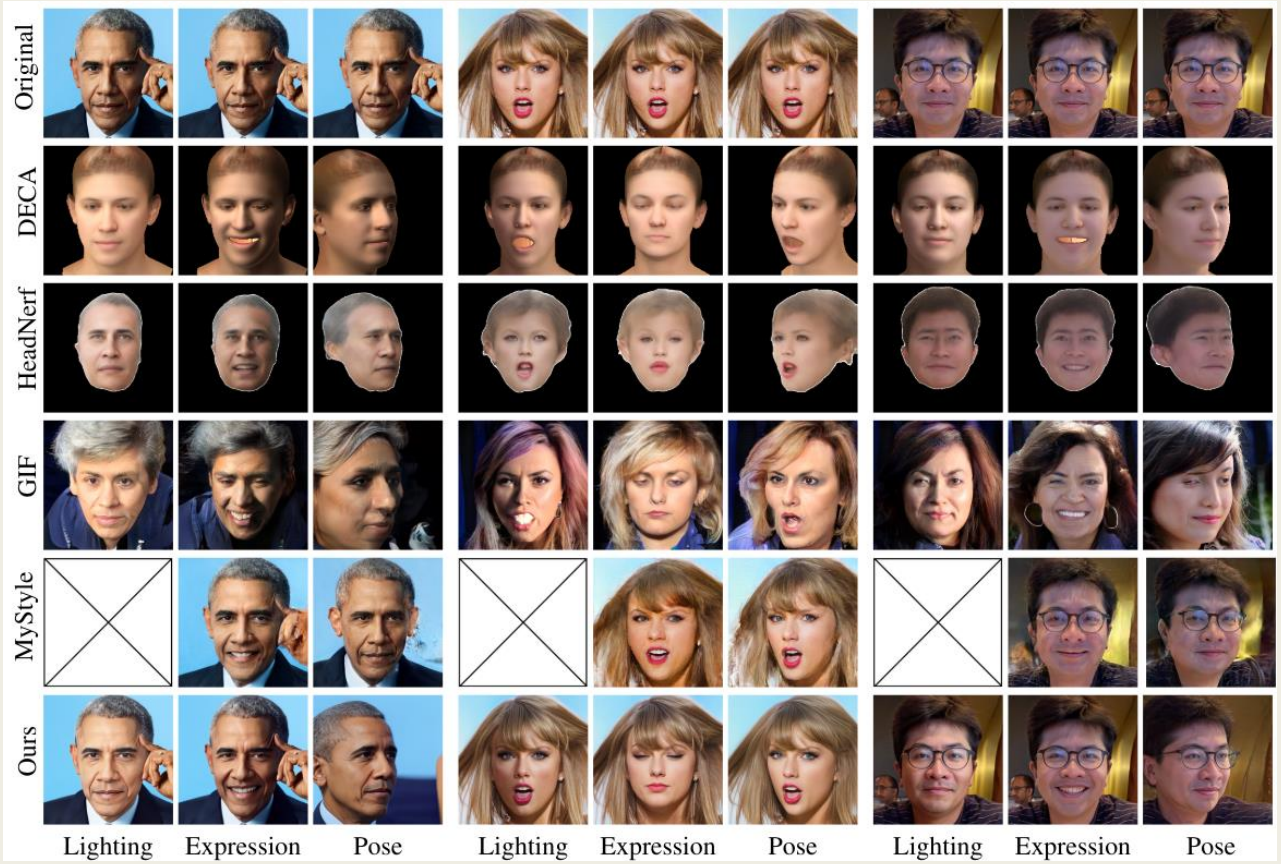

図. 各モデルの出力画像の比較[13]

2. 先行研究紹介

[13] Z. X. L. J. Z. T. Zheng Ding, Cecilia Zhang and X. Zhang, “Diffusionrig: Learning personalized priors for facial appearance editing,” 2023.
[15] T. Li, T. Bolkart, M. J. Black, H. Li, and J. Romero, “Learning a model of facial shape and expression from 4d scans,” ACM Trans.Graph., vol. 36, no. 6, nov 2017.

Zheng Dingらの研究[13]

DiffusionRig: Learning personalized priors for facial appearance editing.

- 2段階学習の効果
  - **Stage1**：一般的な顔の特徴を掴み、物理特性を画像にマッピングする方法を学習⇒制御性獲得
  - **Stage2**：ターゲット人物の顔の特徴・アイデンティティを学ぶ⇒解釈性獲得
- 3DMMとその他特徴の関係性
  - DECAを通さないエンコーダと3DMMから得た特徴量を入れ替えた場合、髪や背景、サングラスなどの特徴を移植できた
  - グローバル潜在変数を扱うエンコーダからの特徴量とDECA経由の特徴量の役割ははっきり分かれる

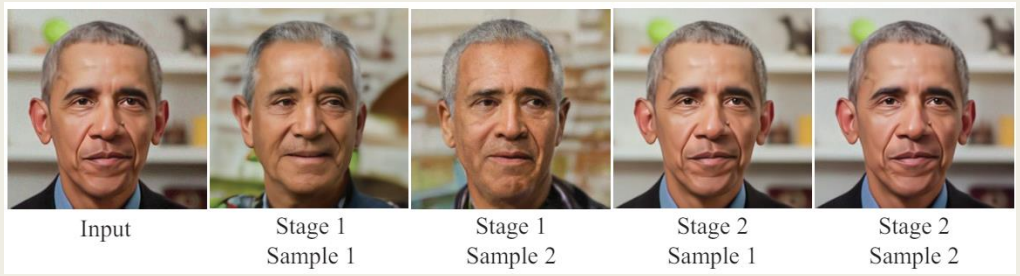

図. DiffusionRigの評価画像[13]

[13] Z. X. L. J. Z. T. Zheng Ding, Cecilia Zhang and X. Zhang, “Diffusionrig: Learning personalized priors for facial appearance editing,” 2023

3. 提案モデル

提案モデル

- 概要
  - iPad Proから得た3次元データを入力
  - 拡散モデルはRGB-Dを扱うため、既存stable-Diffusionのチャネル数を増やす
  - 表情エンコーダによるベクトルを拡散モデルに条件づけることで、自然な表情の変化を実現
- 課題
  - 3次元データから効率の良い特徴抽出ができるバックボーンの検討
  - 頭部完全モデルにおけるカメラ変数問題
  - トポロジー一貫性の欠如により、パラメトリックな表情変化ができない⇒離散的変化

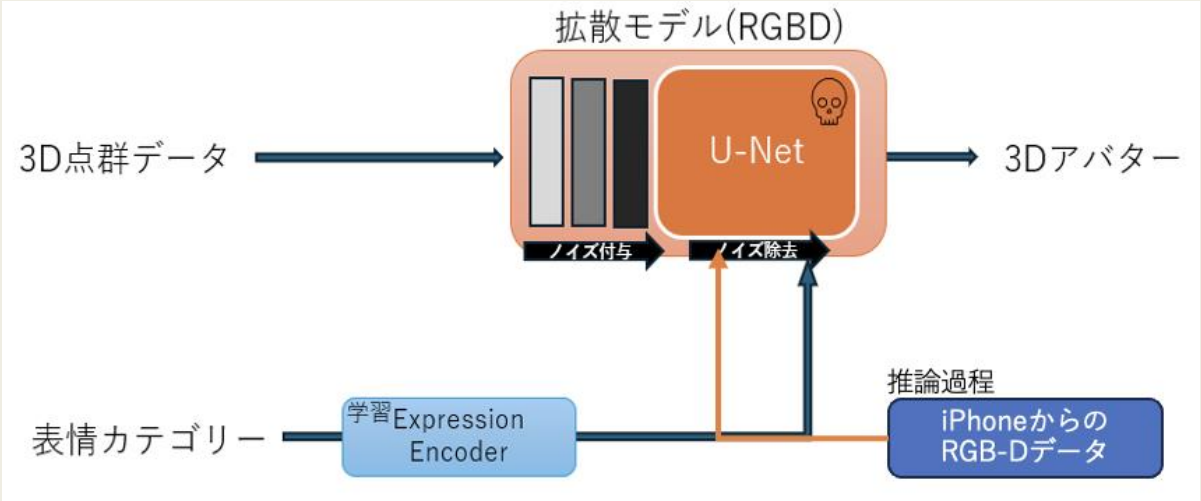

図.提案モデル

4. 今後の計画

今後の計画

- Stable-Diffusion・DiffusionRigのコード解析
  - 3次元データに適したVAEの設計、ノイズスケジューリングのコントロール方法
  - 3DMMなどの物理特性を含んだフレームワークの紐づけ方法
- 適切な3DMMフレームワークの選定
- トレーニングデータセット
  - .plyファイル or .objファイルと.jpgファイルの組み合わせ
- 評価指標
  - レンダリング結果を評価(FID・IS・SSIM・PSNR)
  - 形状とテクスチャを別々に評価
  - 表情の自然さ(EmoNet)