第3回定期ミーティング

2024年5月27日(水)

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・野口颯汰

アウトライン

- 論文紹介
  - DiffusionRig: Learning Personalized Priors for Facial Appearance Editing
  - High-Fidelity 3D Digital Human Head Creation from RGB-D Selfies
- 考察
  - 3DMM
  - データセット
- 参考文献

論文紹介

DiffusionRig: Learning Personalized Priors for Facial Appearance Editing(2023)

- **目的**
  - 20枚ほどの同一人物のポートレート写真から、その人物特有の顔の特徴を学習する
  - 顔の特徴やアイデンティティを保持しながら、表情やライティング、顔の向きなどを後から編集できる
- **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
  - 外観をリグするために、パラメトリックな3D顔モデルである3DMM(FLAME)を拡散モデルの条件に用いる

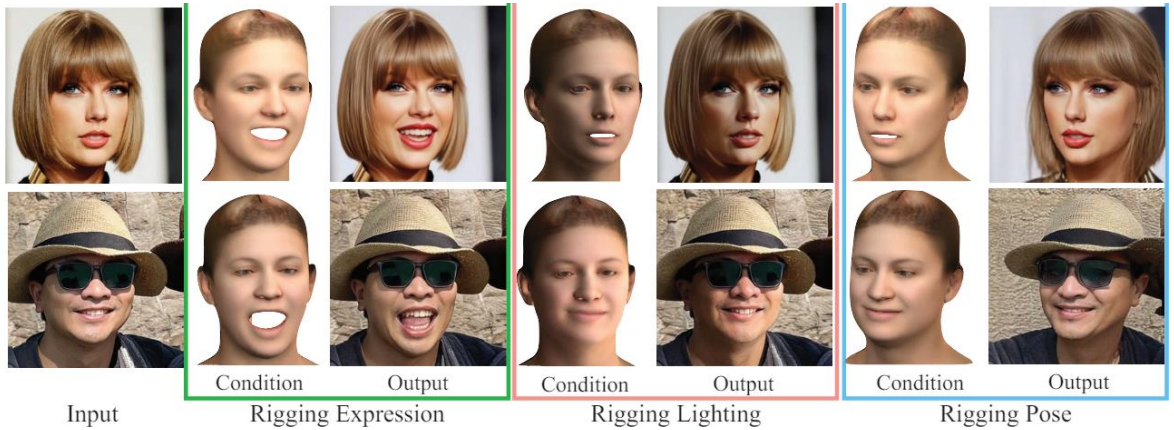


図. モデルの出力画像[1]

論文紹介

DiffusionRig: Learning Personalized Priors for Facial Appearance Editing(2023)

- **アーキテクチャ**
  - 3DMMの作成には学習済みDECAモデルを使用
  - タスクである外観のリグは3DMMのパラメータを通して行う
  - 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

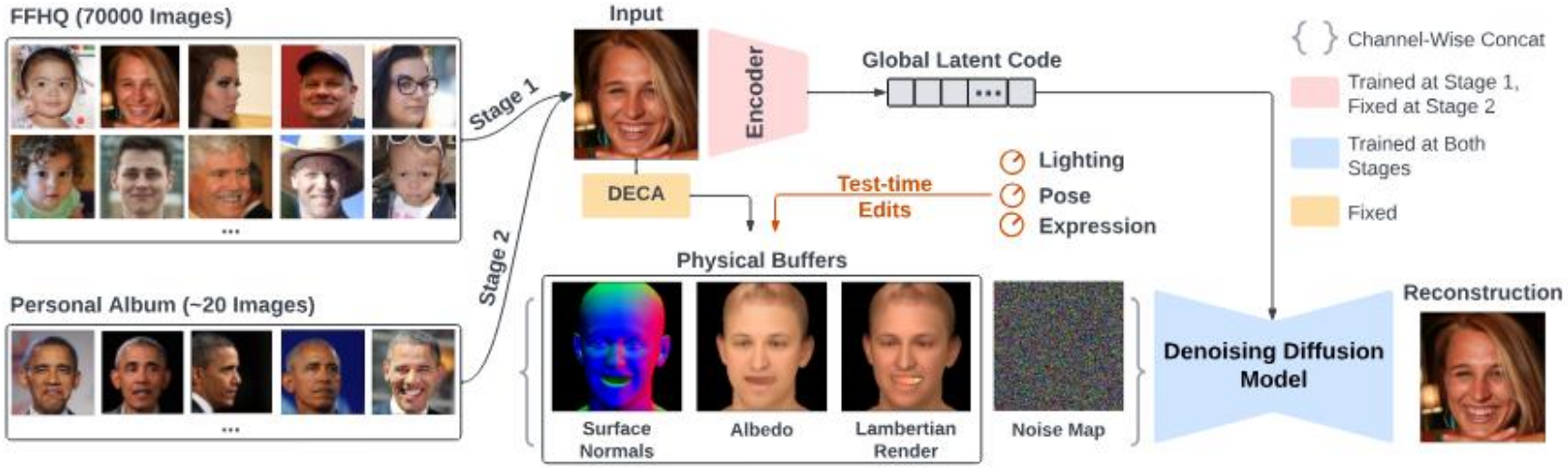


図. DiffusionRigのアーキテクチャ[1]

論文紹介

DiffusionRig: Learning Personalized Priors for Facial Appearance Editing(2023)

- **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
  - 物理的に基づいた方法で外観をリグ出来ているため、不自然さが少ない
  - 制御性・解釈性に優れる
  - 髪や背景、眼鏡なども自然に出力可能

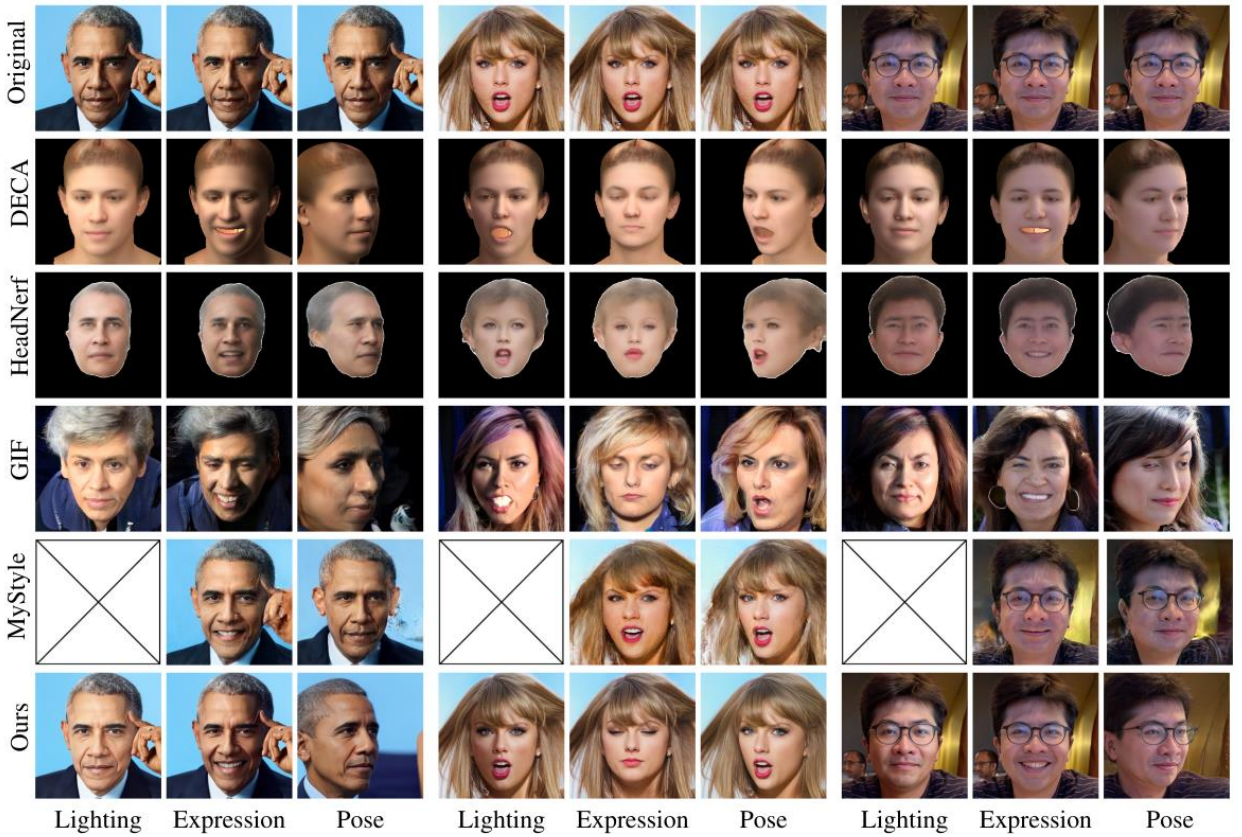


図. 各モデルの出力画像比較[1]

論文紹介

DiffusionRig: Learning Personalized Priors for Facial Appearance Editing(2023)

- **2段階学習の効果**
  - 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するstage1と、ターゲット人物の特徴を学ぶためのstage2を分けたことによって制御性と解釈性を両立

- **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ている

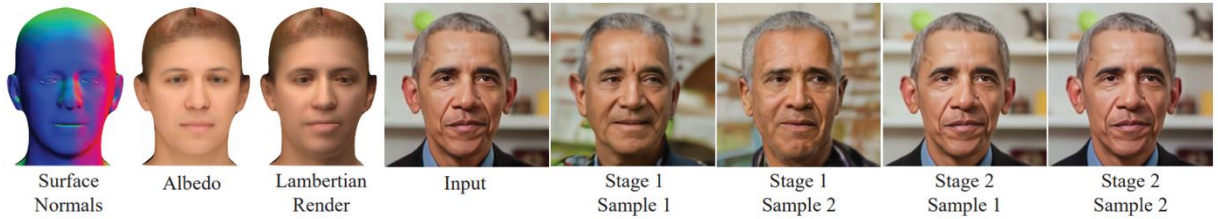


図. DiffusionRigの評価画像[1]

論文紹介

High-Fidelity 3D Digital Human Head Creation from RGB-D Selfies(2021)

- **目的**
  - 短いセルフィーRGB-Dビデオから30秒ほどで高品質な頭部の再構成を行う
  - セルフィーから得た個人のアイデンティティを保ちつつ、しわや毛穴などを高解像度で表現
- **特徴**
  - 攪拌の考え方を用いて3DMMを構築することで、3DMMは大きな表現力を持つ
  - 領域ごとで、異なる解像度のアルベド/法線マップをGANベースで精緻化する

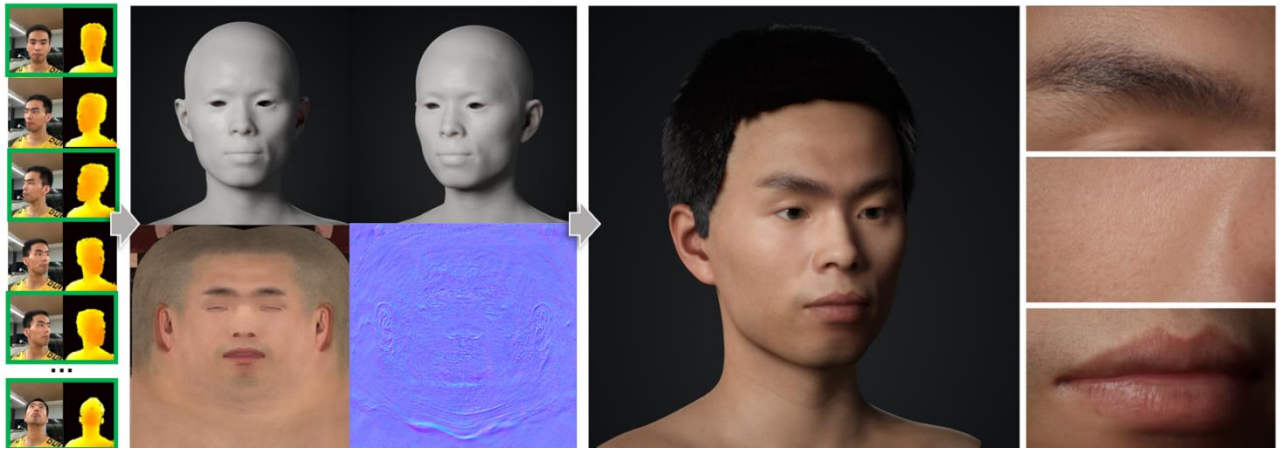
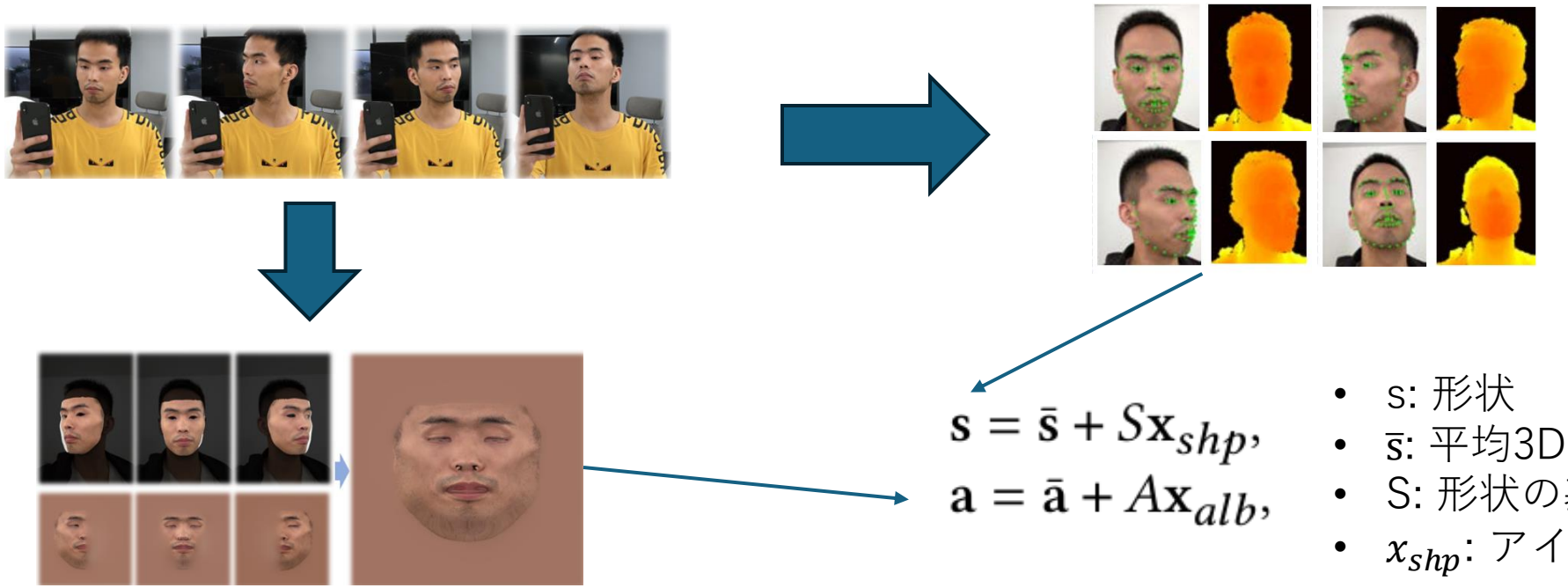


図. モデルの出力画像[2]

論文紹介

High-Fidelity 3D Digital Human Head Creation from RGB-D Selfies(2021)

- **アーキテクチャ(Initial Modeling)**
  - 撮影した動画フレームにMobileNetによりランドマークを付与し、4フレームに厳選
  - 4フレームのRGB-D画像・ランドマーク・テクスチャマップから3DMMを構築



初期3DMM

- s: 形状
- $\bar{s}$: 平均3D顔形状モデルベクトル
- S: 形状の基底
- x_{shp}: アイデンティティパラメータ

図. 初期3DMMを構築する過程[2]

論文紹介

High-Fidelity 3D Digital Human Head Creation from RGB-D Selfies(2021)

- **アーキテクチャ(Optimize)**
  - 最適化するべきパラメータは、$P = \{x_{shp}, x_{alb}, x_{light}, x_{pose}\}$
  - 推定パラメータと3DMM基底を用いて、レンダリングしたRGB-Dフレームを比較することで学習

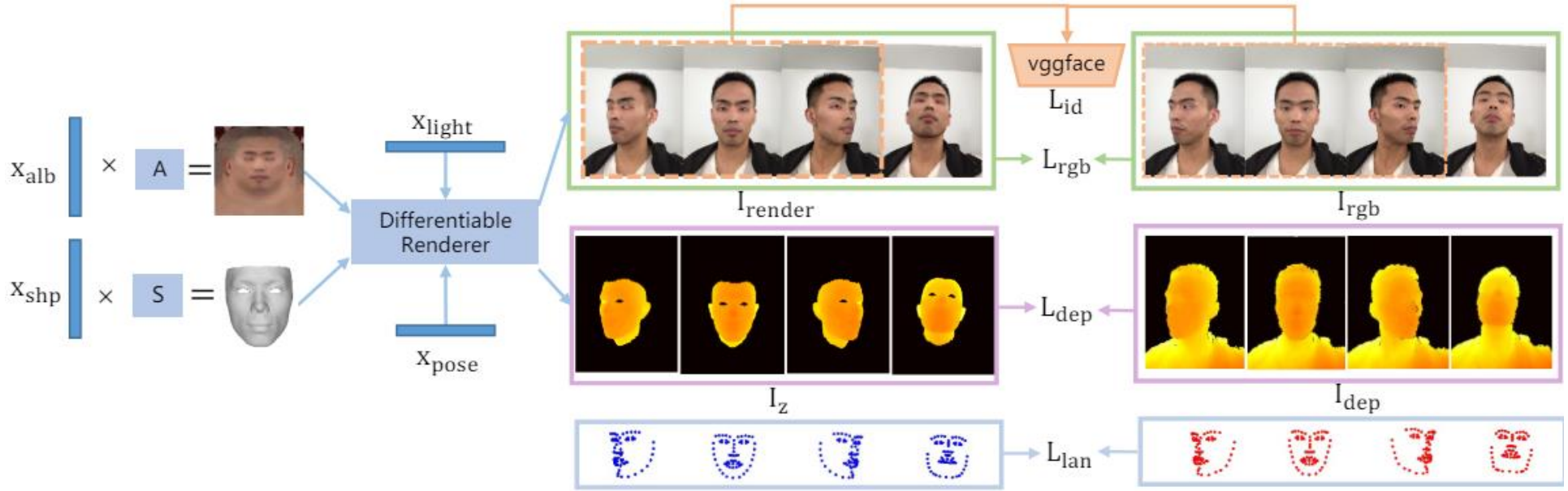


図. 3DMMの最適化[2]

論文紹介

High-Fidelity 3D Digital Human Head Creation from RGB-D Selfies(2021)

- **形状基底Sの設定(3DMMの拡張)**
  - ➢ 人間の顔は実は非対称性を持つが、従来の3DMMでは表現力に限界→**摂動操作によって3DMMの形状を拡張**
    1. 各モデルの目・鼻・口領域を回転摂動(~1°)を加えた他のモデルの同領域で置換(目は回転摂動なし)
    2. モデル自体を剛体変換摂動(回転・並進・スケーリング)させる
    3. モデル固有のローカル座標でのミラーリング

    →顔モデルを200から100000に拡張
  - ➢ 累積説明分散が99.9%になる範囲でPCAを適用し、顔モデルを有限基底数に落とし込む→$x_{shp} \in R^{500}$

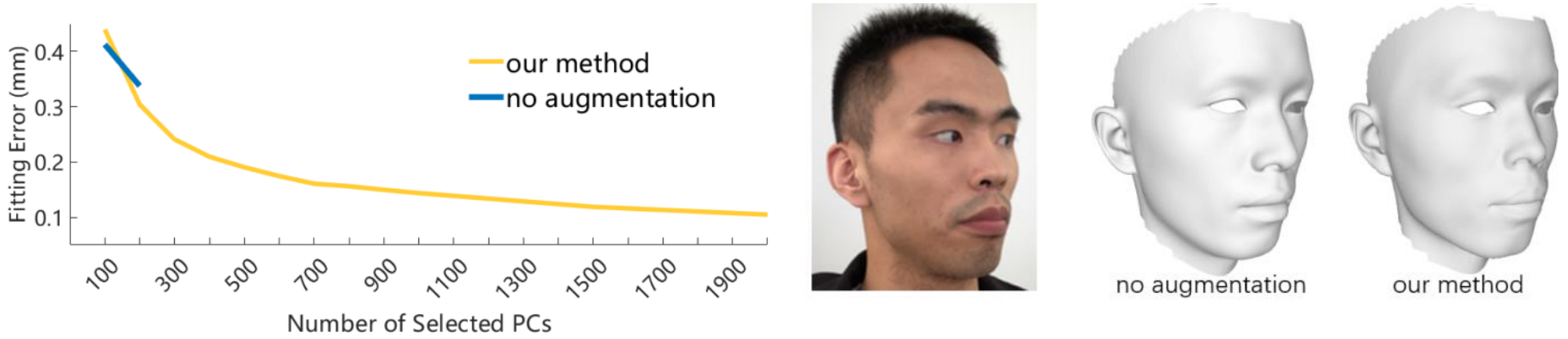


図. 3DMMの拡張による効果[2]

論文紹介

High-Fidelity 3D Digital Human Head Creation from RGB-D Selfies(2021)

- **形状基底Sの設定(3DMMの拡張)**
  - ➢ 人間の顔は実は非対称性を持つが、従来の3DMMでは表現力に限界→**摂動操作によって3DMMの形状を拡張**
    1. 各モデルの目・鼻・口領域を回転摂動(~1°)を加えた他のモデルの同領域で置換(目は回転摂動なし)
    2. モデル自体を剛体変換摂動(回転・並進・スケーリング)させる
    3. モデル固有のローカル座標でのミラーリング

    →顔モデルを200から100000に拡張
  - ➢ 累積説明分散が99.9%になる範囲でPCAを適用し、顔モデルを有限基底数に落とし込む→$x_{shp} \in R^{500}$

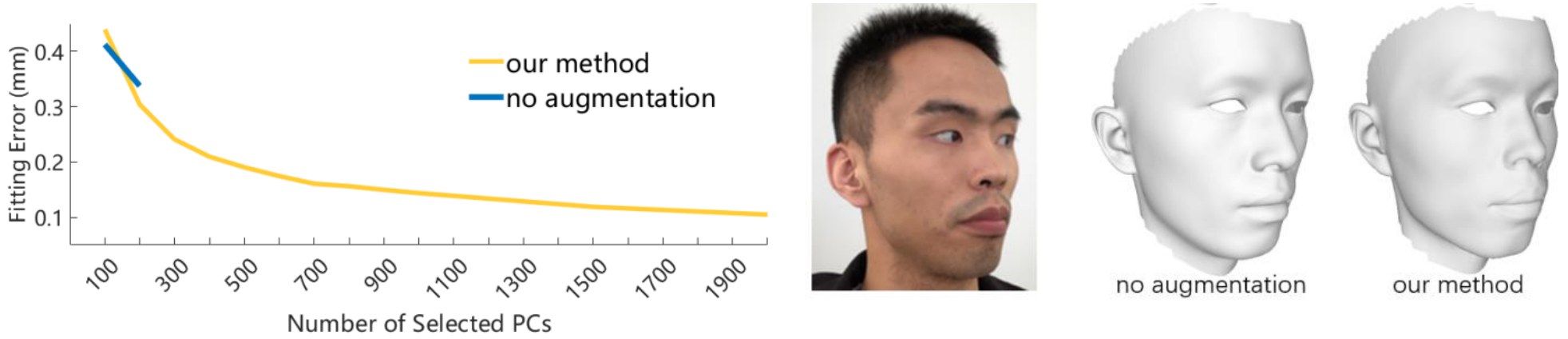


図. 3DMMの拡張による効果[2]

論文紹介

High-Fidelity 3D Digital Human Head Creation from RGB-D Selfies(2021)

- **形状基底Sの設定(3DMMの拡張)**
  - ➢ 基底数が増大すると、直接PCAを適用した方が表現力が高い

  →拡張過程に含まれる確率的アルゴリズムが高周波成分を切り捨て、低周波成分を強調するため

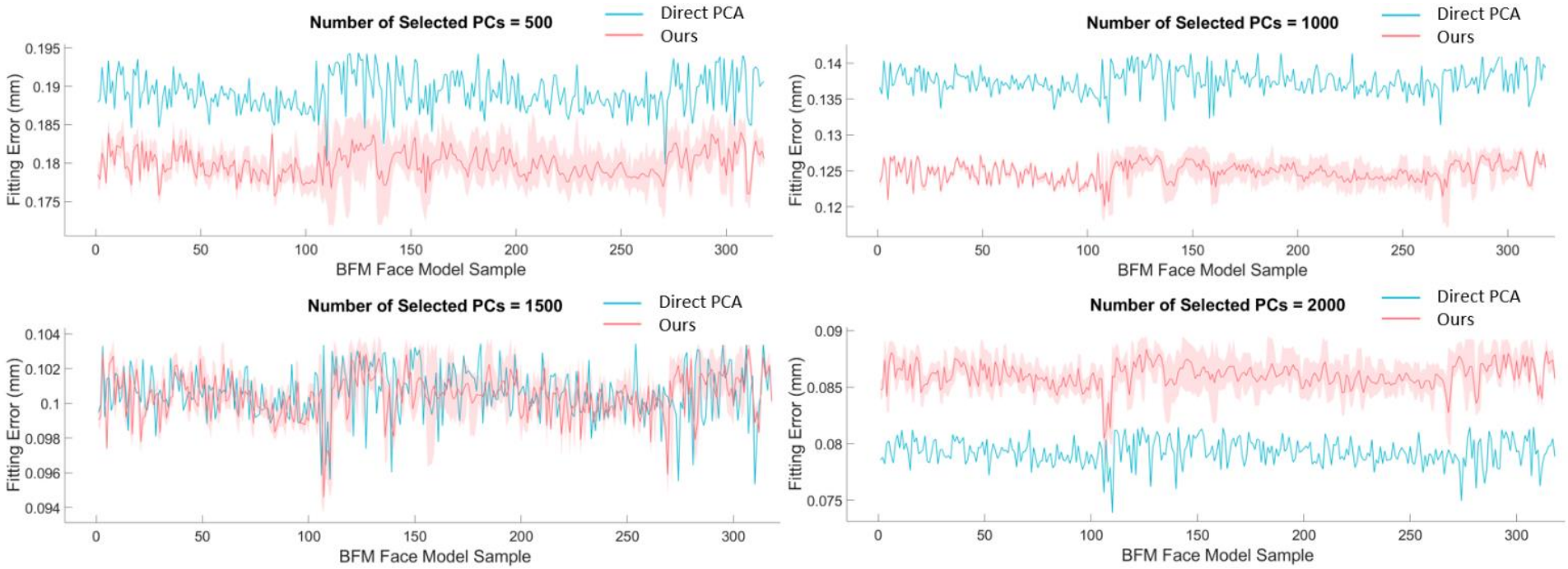


図. 3DMMの拡張による効果[2]

論文紹介

High-Fidelity 3D Digital Human Head Creation from RGB-D Selfies(2021)

- **アルベド/法線マップの高解像度化**
  - テクスチャマップからRegional Pyramid Basesを構築する
  - ピラミッド基底により、低解像度でパラメトリックフィッティングを行い、高解像度基底に同じパラメータを直接適用し、高解像度アルベド/法線マップを取得

  →表現力の高いアルベド基底によるテクスチャの過適合とジオメトリへの不適合を防止
  - 複数の解像度を使用することで、多くの構造情報を強調できる

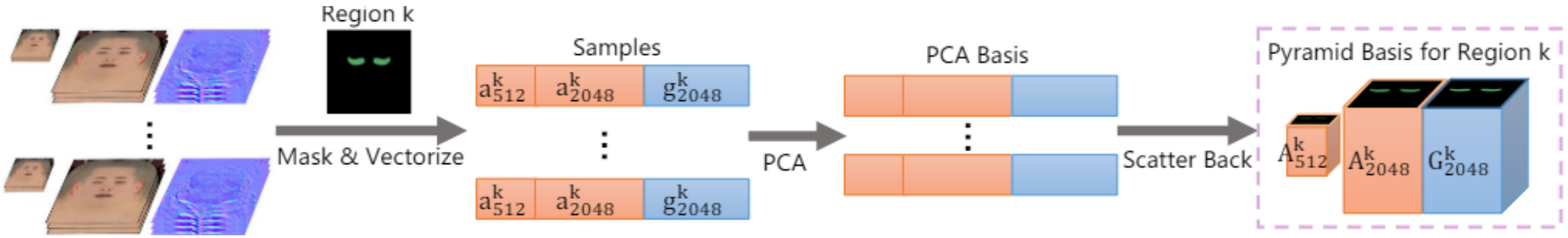


図. Regional Pyramid Basesの取得プロセス[2]

論文紹介

High-Fidelity 3D Digital Human Head Creation from RGB-D Selfies(2021)

- **アルベド/法線マップの高解像度化**
  - ピラミッド基底を用いて、初期3DMMのx_{alb}を最適化したい

$$L(\mathbf{x}_{alb}) = \|I_{fit}(\mathbf{x}_{alb}) - I_{init}\|_2 + \omega_{tv} f_{tv}(I_{fit}(\mathbf{x}_{alb})) + \omega_{alb}\|\mathbf{x}_{alb}\|_2^2,$$

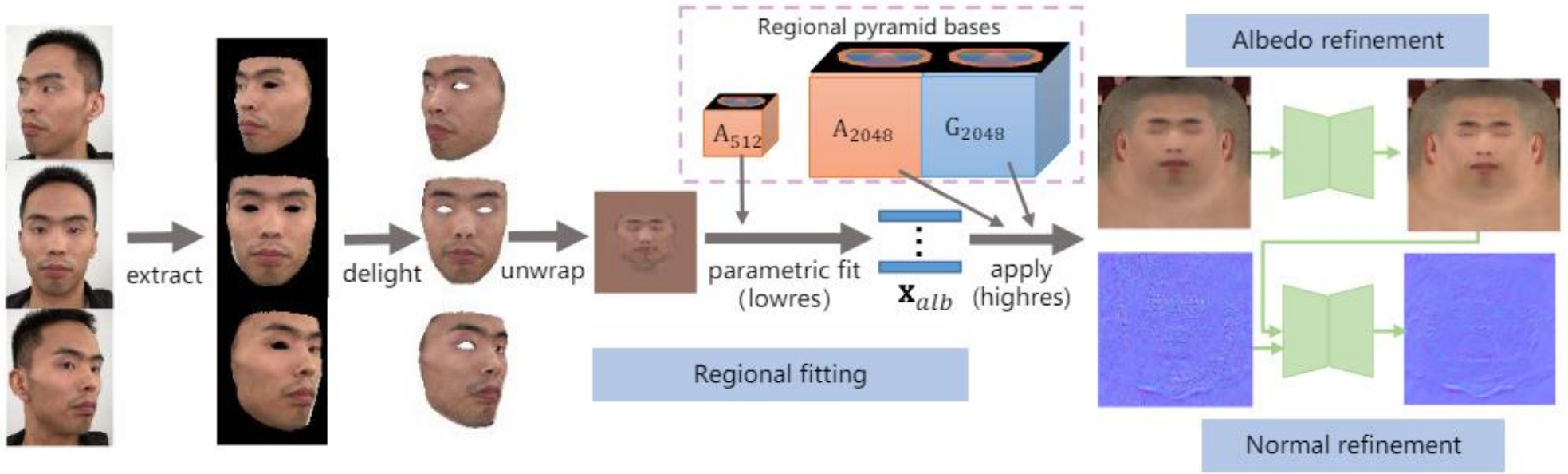


図. 高解像度アルベド/法線マップの取得プロセス[2]

論文紹介

High-Fidelity 3D Digital Human Head Creation from RGB-D Selfies(2021)

- **その他**
  - 頭部完全体モデルへ
    - 200ソースモデルを後頭部領域・中間領域に分け、直接PCAを適用
    - 生成した顔領域を条件付けしてから、完全な頭部3DMMパラメータを生成
  - ヘアスタイル
    - MobileNetの画像分類によって選ばれた30種類の髪モデルを頭部モデルに取り付ける
  - 歯・目
    - テンプレートモデルを使用
    - ランドマークによって、スケール・位置などを調整

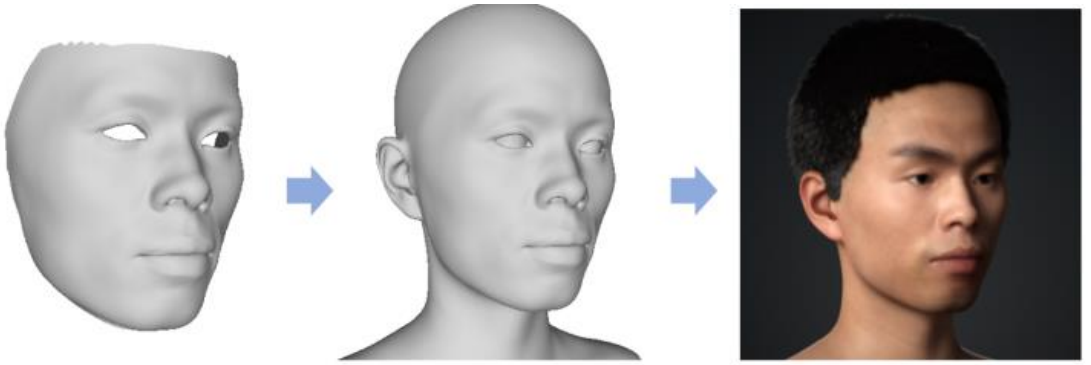


図. 最終出力結果の構成プロセス[2]

論文紹介

High-Fidelity 3D Digital Human Head Creation from RGB-D Selfies(2021)

- **ジオメトリの定量的/定性的評価**

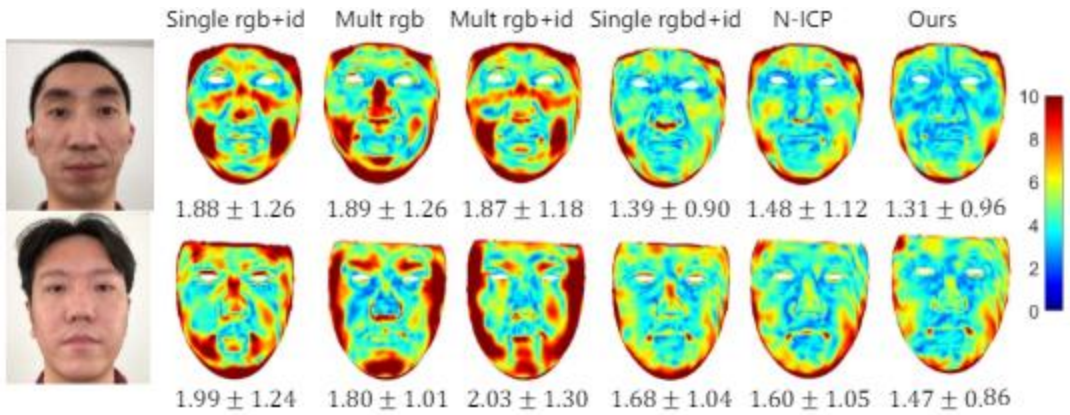


図. ジオメトリのエラー率の比較[2]

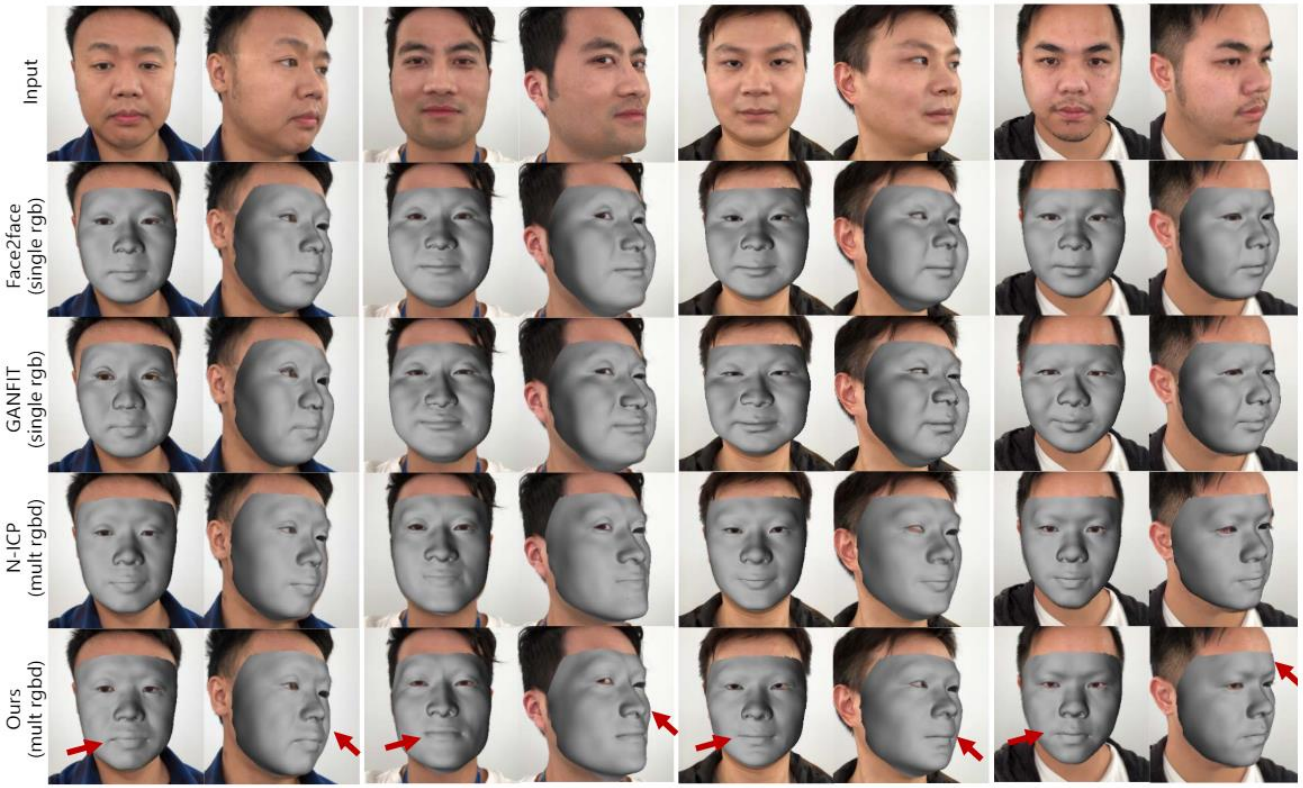


図. 他モデルとのジオメトリの比較[2]

論文紹介

High-Fidelity 3D Digital Human Head Creation from RGB-D Selfies(2021)

- **テクスチャの定性的評価**

図. 生成結果のテクスチャ[2]

入力画像では観察不可能な特徴を生成

領域ごとのx_{alb}最適化、GANによるマップ更新によってテクスチャが精緻化

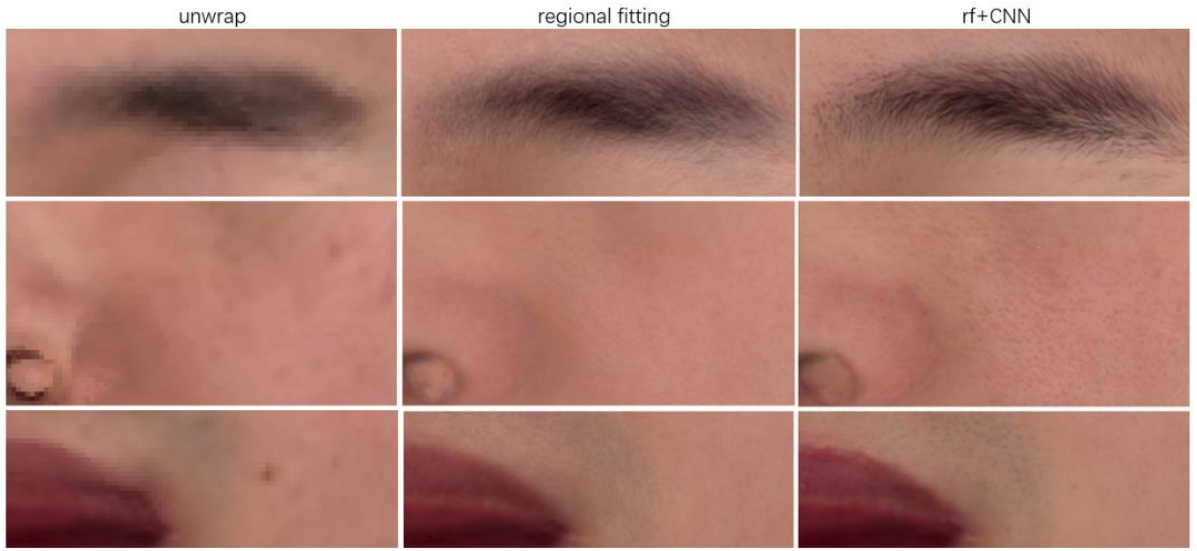


図. テクスチャ生成における各段階の影響[2]

論文紹介

High-Fidelity 3D Digital Human Head Creation from RGB-D Selfies(2021)

- **アイデンティティの定性的評価**
  - 36人のセルフィーから3D顔モデルを再構成し、そのレンダリング結果とセルフィー画像をデータセット(40000枚)に混ぜる
  - データセットに含まれる全ての画像に対して、レンダリング結果との特徴距離を計算し、小さい順にソート
  - レンダリング結果とセルフィー画像の特徴距離Xを計算し、データセットからの特徴距離と比較
  - Xが小さい順から何番目であるかをランキングし、36人の平均ランキングを計算
  - 特徴距離が小さいほど同一人物である可能性が高いため、よりアイデンティティを保持していると言える

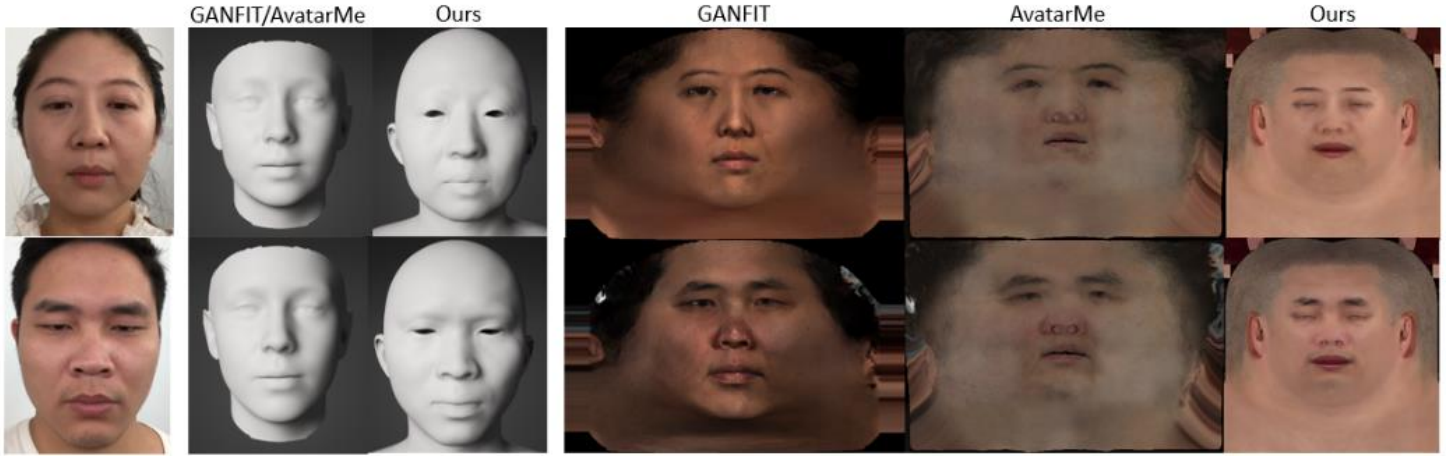


図. 他SOTAモデルとの比較[2]

表. 同一認識順位[2]

Method	Avg. ranking
AvatarMe [Lattas et al. 2020]	7.0%
GANFIT [Gecer et al. 2019]	5.1%
Ours	4.5%

論文紹介

High-Fidelity 3D Digital Human Head Creation from RGB-D Selfies(2021)

- **入力画像から直接再構成する手法との比較**
  - 入力写真から直接テクスチャや顔形状を抽出する手法は、入力に忠実だが、その他出力画像とのトポロジー的一貫性が生まれない
  - 入力された顔にある影や、化粧によるハイライトに強い影響を受け、欠陥が生じやすい
  - トポロジー的に一貫性がないため、アクセサリーをつけたり、アニメーション化することが難しい

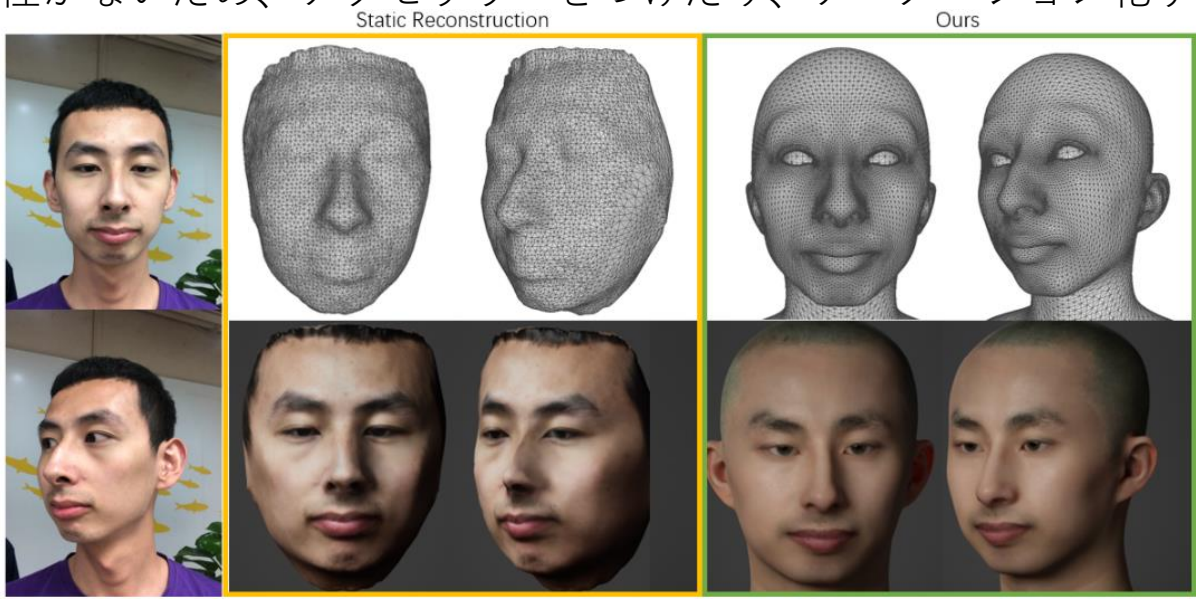


図. 他SOTAモデルとの比較[2]

考察

3DMM

- トポロジー的な一貫性のためには不可欠
- 目や鼻は後付けになってしまうため、3DMMを通さずに再構成した方がアイデンティティに忠実
- 髪形やアクセサリーの後付けは自分たちのモデルでは考えていない
- 形式的な基底に落とし込まないため、拡散モデルの多様性を最大限生かせるかも
- パラメトリックな表情管理ができないため、離散的な表情変化は難しい
- 入力に対するロバスト性を確保するために、DiffusionRigのように3DMMをガイドとして利用する

データセット

- 3Dオブジェクトファイル(.ply)を利用
- 歯や目を含む3Dデータセットがまだ見つけられていない
- 一眼レフカメラor深度センサーで取得したデータのどちらかを使用
- RGB-Dデータとの相性を考える必要アリ

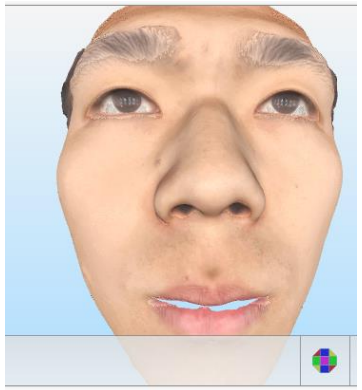


図. データセットサンプル[2]

モデル紹介　～アバター生成～

- 目的
  - コンシューマー向け深度センサーでスキャンした3D顔をアイデンティティを保持したまま表情などを変えたい
- 概要
  1. 3D点群データをRGBD画像に変換
  2. RGB-D画像を拡散モデルに入力
  3. ノイズ除去過程で表情カテゴリー情報を入力
  4. 推論過程でターゲットデータを初期ノイズに連結
- 問題点
  - 深度情報Dの特徴量をどの程度掴めるか
  - カメラ変数問題
  - すべての候補において入力顔画像のアイデンティティをどう残すか

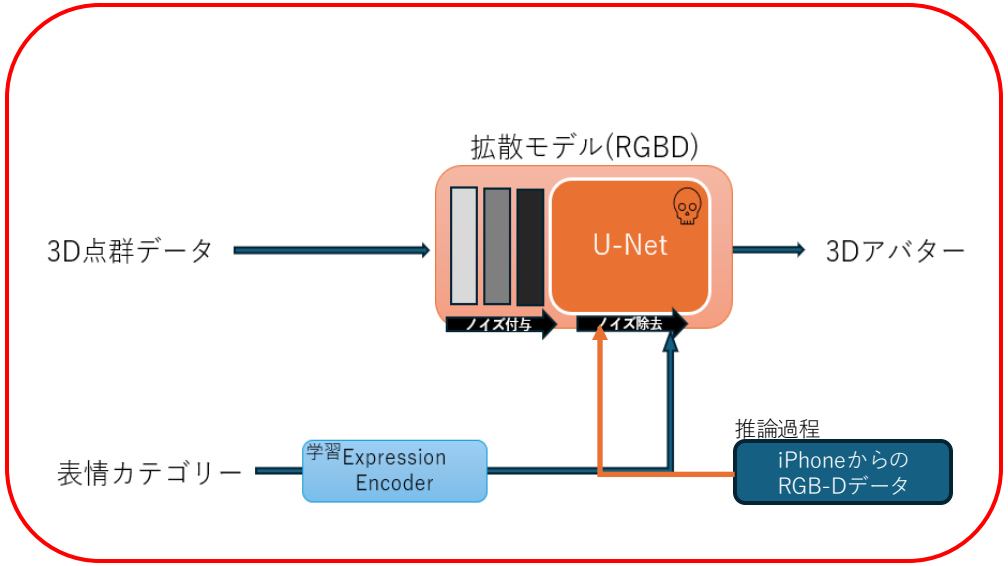


図. モデルのアーキテクチャ

参考文献

1. Zheng Ding, Xuaner Zhang, Zhihao Xia, Lars Jebe, Zhuowen Tu, Xiuming Zhang, “DiffusionRig: Learning Personalized Priors for Facial Appearance Editing”, https://openaccess.thecvf.com/content/CVPR2023/papers/Ding_DiffusionRig_Learning_Personalized_Priors_for_Facial_Appearance_Editing_CVPR_2023_paper.pdf , CVPR 2023, 2023-04-13
2. Linchao Bao, Xiangkai Lin, Yajing Chen, Haoxian Zhang, Sheng Wang, “High-Fidelity 3D Digital Human Head Creation from RGB-D Selfies”, https://arxiv.org/pdf/2010.05562, 2021-01-29

LDM3D: Latent Diffusion Model for 3D

Gabriela Ben Melech Stan, Diana Wofk, Scottie Fox, Alex Redden, Will Saxton, Jean Yu Intel, Estelle Aflalo, Shao-Yen Tseng, Fabio Nonato, Matthias Muller, Vasudev Lal

LDM3D
　拡散モデルを用いたテキストプロンプトからRGBD画像を出力するモデル

DepthFusion
　TouchDesignerを用いて没入型でインタラクティブな
　360° ビュー体験を作成するアプリケーション

→エンターテインメントやゲームから建築やデザインに応用可能

図1. “A table with a book”と入力した際の出力例[1]

背景：

- 近年のStable Diffusionをはじめとする拡散モデルによる生成AIの発展
- トランスフォーマーや拡散モデルを使用することによる単眼深度推定モデルの発展
  （ZoeDepth, DPT-Large depth estimation model ）

取り組んだ問題：

- 画像を拡散モデルで生成してその後深度推定するようなモデルの問題点
  - ✓ 深度推定モデルには大規模かつ多様なデータセットが必要
  - ✓ 生成された画像にはGround Truthが存在しない

  →深度推定モデルが拡散モデルの出力に適応するのは困難

深度情報(D)と色情報(RGB)を共同学習して画像生成に深度情報を関連付ける

学習の流れ[1/2]

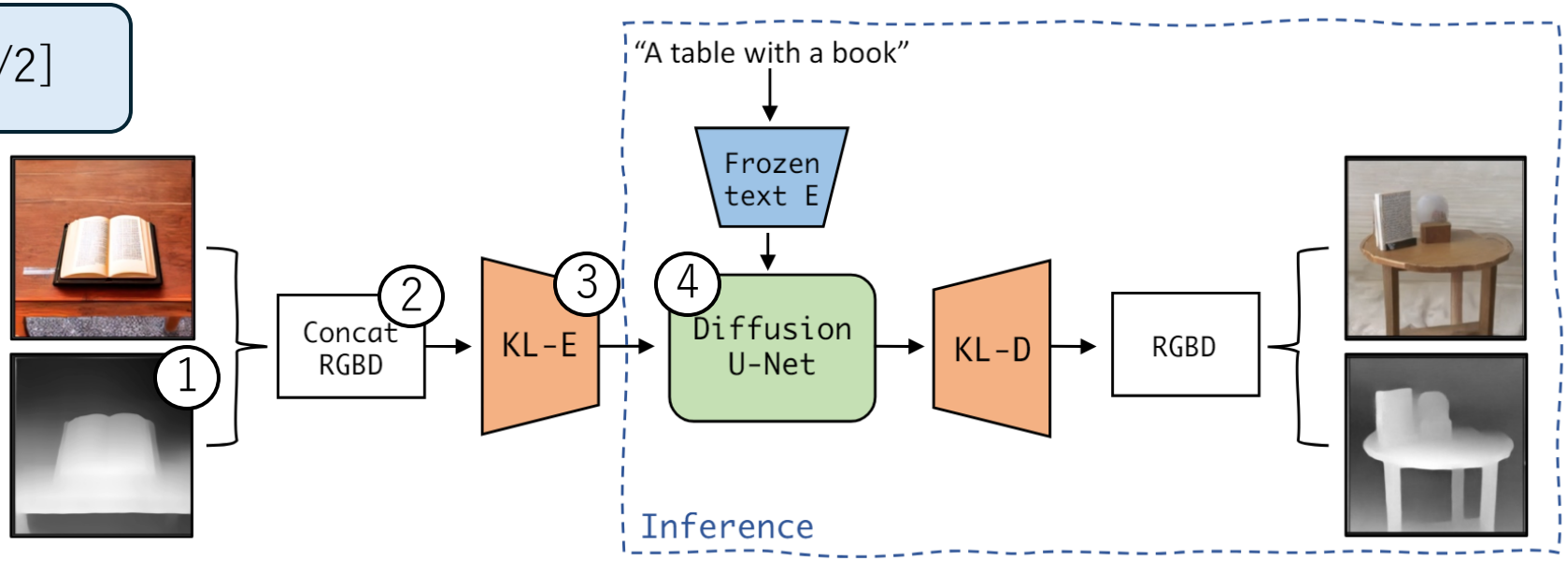


図2. 3DLDMの訓練時パイプライン[2]

①16ビット深度マップをRGB画像の形式に変換
②深度画像とRGB画像をチャネル方向に連結
③KL-AEで潜在空間マップに変換
④潜在表現にノイズが追加され、U-Netモデルによってノイズ除去

学習の流れ[2/2]

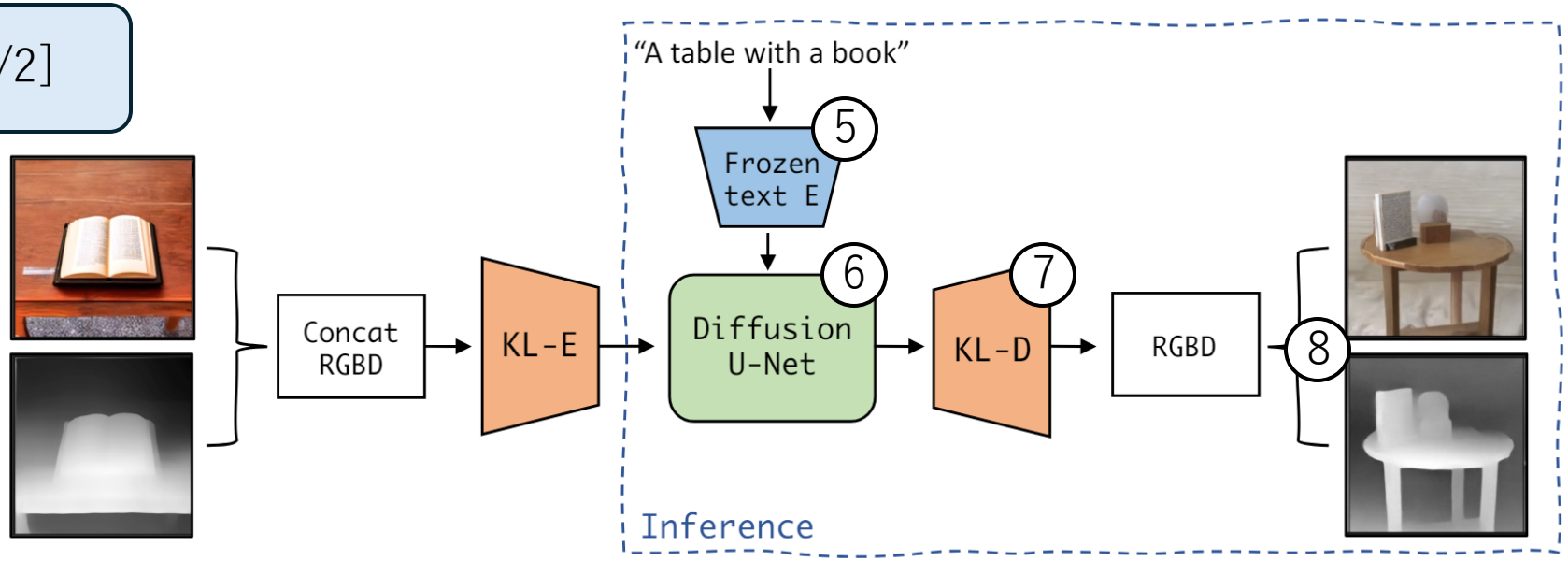


図2. 3DLDMの訓練時パイプライン[2]

⑤テキストをCLIPでエンコード
⑥クロスアテンションでU-Netに関連付ける
⑦KL-Dで潜在空間をRGBD画像を出力
⑧出力をRGB画像と16ビットグレースケールの深度マップに分離

モデルアーキテクチャ

Stable Diffusion：
Stable diffusion v1.4が元
2D畳み込み層で構成

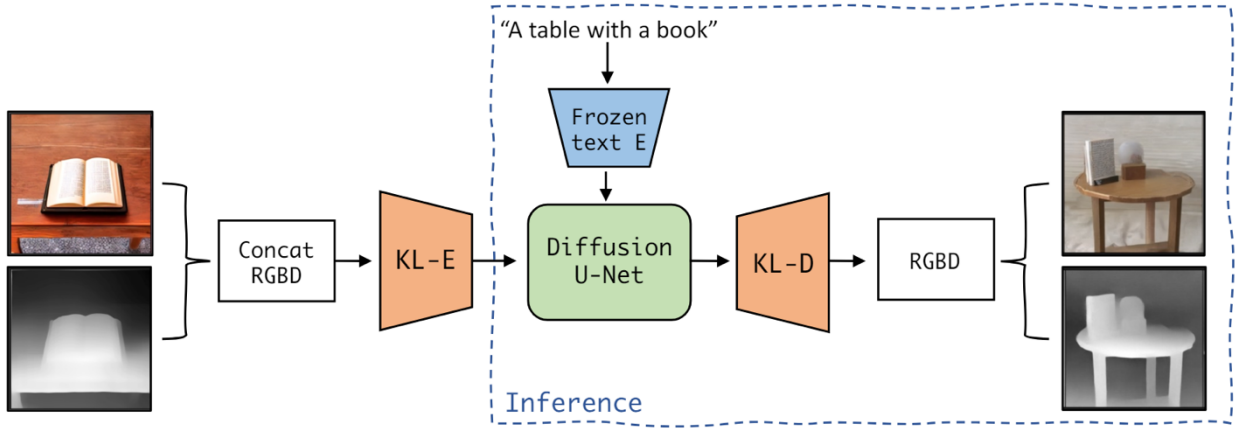


図2. 3DLDMの訓練時パイプライン[2]

KL-AE：
KLダイバージェンスを損失を含む
元となるモデルから最初と最後のConv2dレイヤーを修正
→連結されたRGB画像と深度マップからなる修正された入力形式を処理するために

学習

学習環境：

- Intel AIスーパーコンピューティングクラスター
  （Intel Xeonプロセッサー・Intel Habana Gaudi AIアクセラレーター）
- Nvidia A6000 GPU

表1. KL-AEの学習詳細[1]

訓練データセット	8233サンプル
バリデーションデータセット	2059サンプル
ダウンサンプリング率	8倍（元のStable Diffusionより判断）
オプティマイザ	Adam
学習率	10^{-5}
バッチサイズ	8
エポック	83
損失関数	知覚損失とパッチベースの敵対的な損失（元のStable Diffusionと同様）

表2. 拡散モデルの学習詳細[1]

潜在入力サイズ	$64 \times 64 \times 4$
オプティマイザ	Adam
学習率	10^{-5}
バッチサイズ	32
エポック	178
損失関数	元のStable Diffusionと同様

定性評価

COCO検証データセットからの512×512画像に対する比較

1列目：Stable Diffusion v1.4の出力(RGB画像)
2列目：LDM3Dの出力(RGB画像)
3列目：LDM3Dの出力(深度マップ)
4列目：DPT-Largeの深度マップ

上からそれぞれのキャプション

1. テーブルの上のピザのシートのクローズアップ
2. テーブルの上のレモンの写真
3. 髪に**ピンクのリボン**をつけた少女がブロッコリーを食べている
4. 男が馬に乗っている道路上
5. フォークの横に**黒いマフィン包装紙**に入ったマフィン
6. 水源から水を飲む白いシロクマが岩の隣にいる

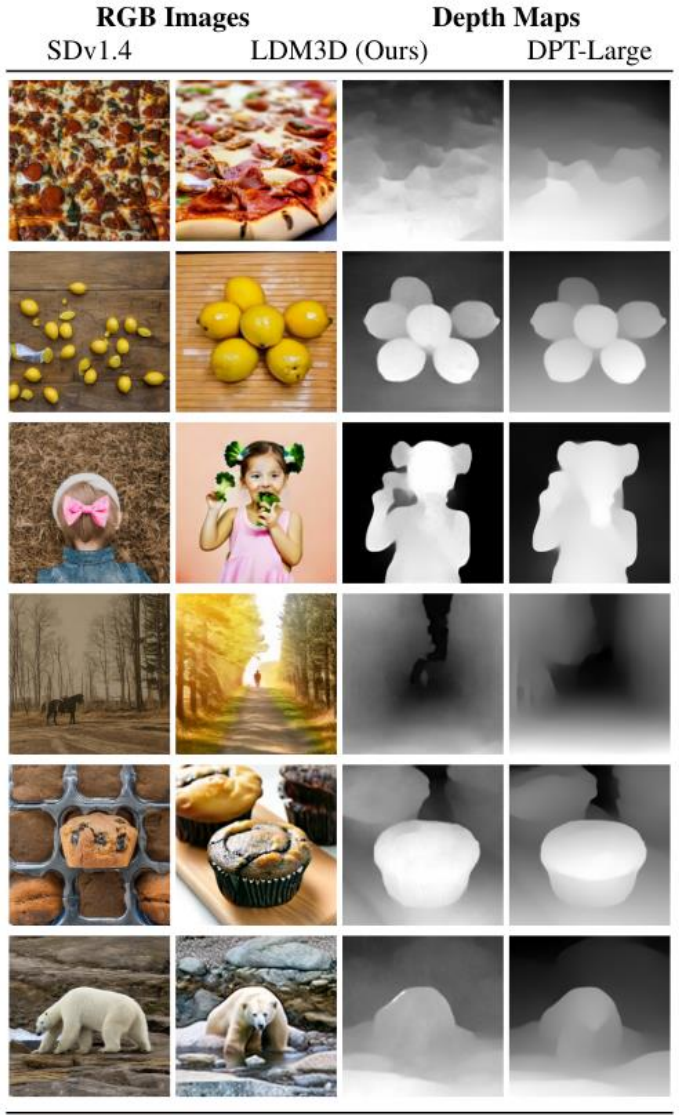


図3. 深度マップの質的比較[1]

定量評価

【評価指標】

FID：生成結果がどれくらいリアルを近いか
IS：生成された画像の品質と多様性を測定
CLIP：生成文と画像がどれくらい対応しているか

表3. テキストto画像の精度比較[1]

Method	FID↓	IS↑	CLIP↑
SD v1.4	28.08	**34.17**±0.76	26.13 ± 2.81
SD v1.5	**27.39**	34.02 ± 0.79	26.13 ± 2.79
LDM3D (ours)	27.82	28.79 ± 0.49	**26.61**±2.92

【推論詳細】

512×512サイズのMS-COCO [13]データセットでのテキスト条件付き画像合成
50 DDIM [27]ステップ

【考察】

FIDとCLIPについて同等のスコアを達成
ISは低下している→ISがFIDと比べてロバスト性が低い為

定性/定量評価(深度)

表4. 深度の精度比較[1]

	w.r.t. ZoeDepth-N	
	AbsRel	RMSE [m]
LDM3D	0.0911	0.334
DPT-Large	0.0779	0.297

【定性評価】

出力深度マップに対するGround Truthはないので
ZoeDepthという単眼深度推定モデルで行う

【評価指標】

絶対相対誤差（AbsRel）
平方平均誤差（RMSE）

【考察】

LDM３DとDPT-Largeは同程度の精度

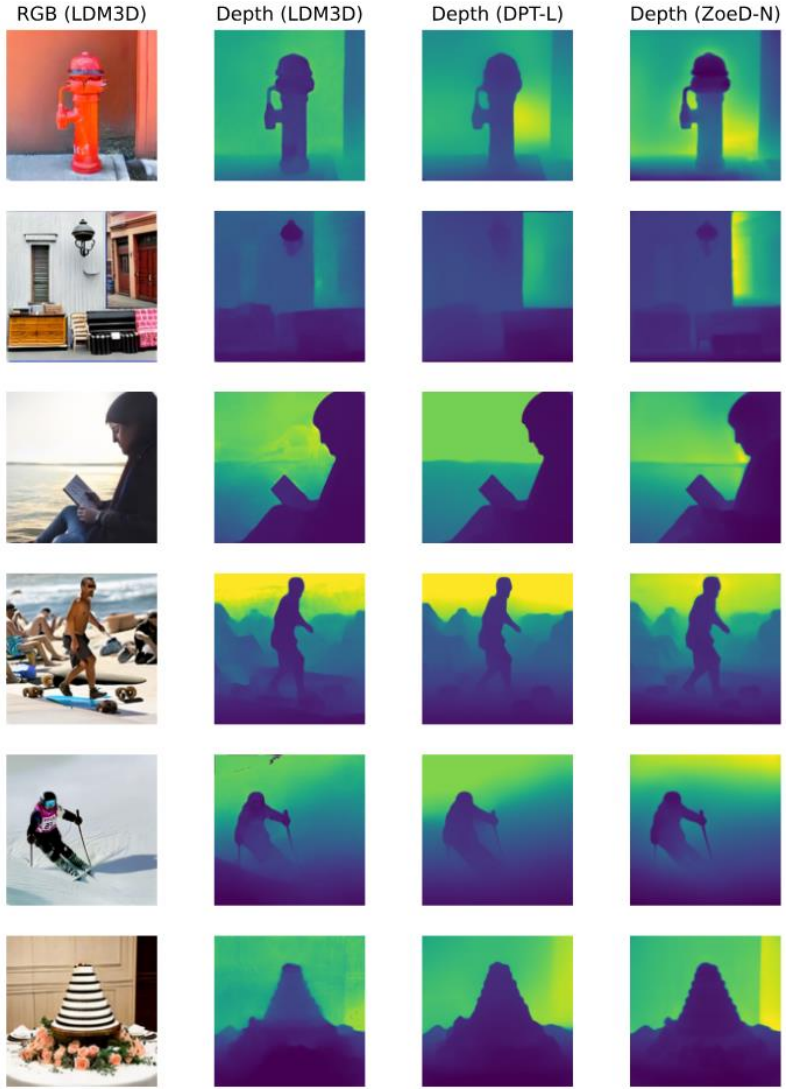


図4. 深度マップの比較[1]

KL-AE評価

【評価指標】

rFID：相対的なFID

Abs.Rel.：生成された深度マップと実在する深度マップの誤差

【考察】

精度の劣化が分かる

✓ 原因：

ピクセル空間でRGB画像に深度情報を組み込む際のデータ圧縮比の増加が原因

✓ 解決方法：

タスクに応じたAEへの調整

表5. KL-AEの精度比較[1]

Model	rFID	Abs.Rel.
pre-trained KL-AE, RGB	0.763	-
fine-tuned KL-AE, RGBD	1.966	0.179

MediaPipe

Pythonのパッケージ版で提供されている四つの機能
- Hands：手のランドマーク
- Pose：骨格取得
- Face Mesh：顔点座標
- Holistic：face_mesh+左右Hands+Pose

Face Mesh：
取得した顔特徴点は3次元空間上の座標として扱うことが可能

faceMase.setOptions：
- maxNumFases：認識する顔画像
- refineLandmarks：顔を詳細に認識するかどうか
  - ランドマークの出力数が、468から478箇所に増加

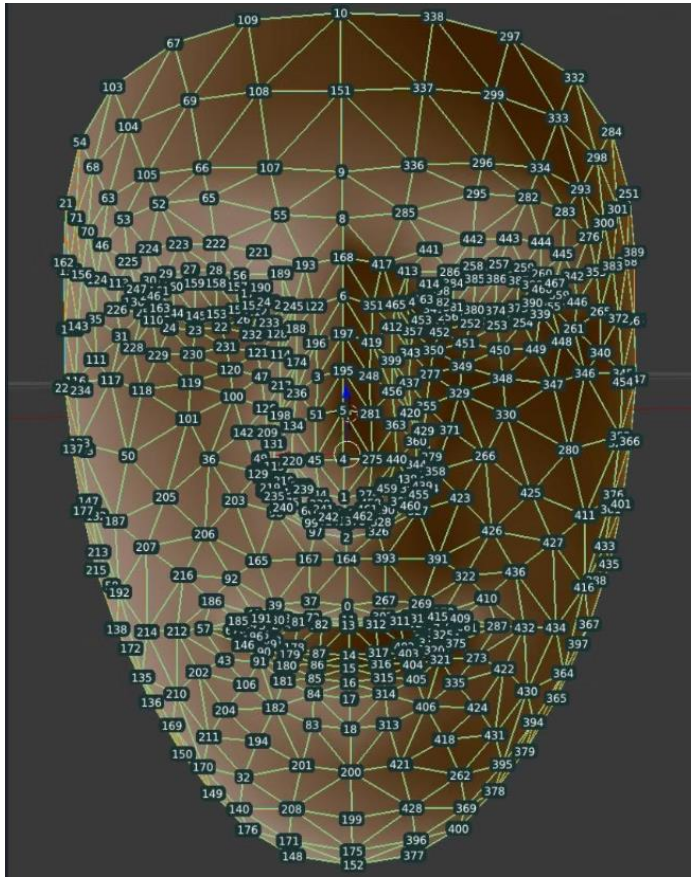


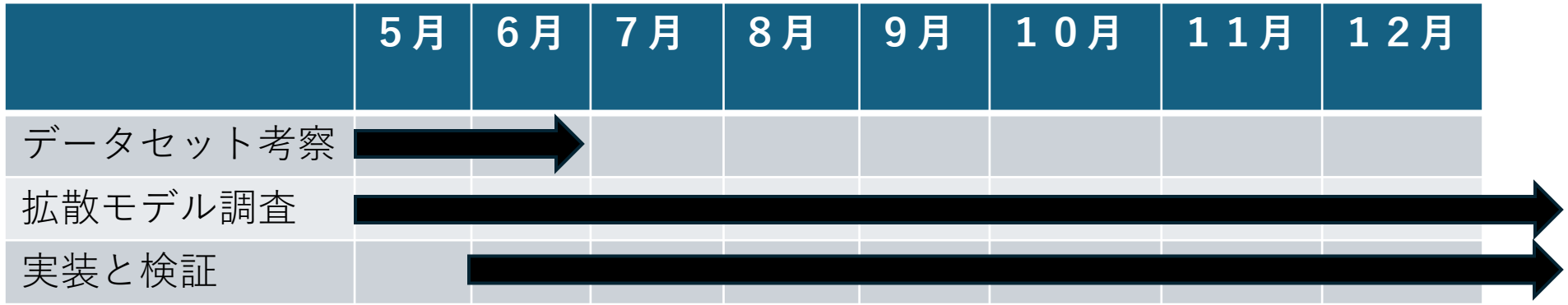
図5. 顔座標取得例[2]

考察

色情報は取得不可　→　別途データ前処理の必要
データセットが疎すぎる可能性　→　ARKitと比較
Pose推定を組み合わせるのに有効　→　Holisticの使用

今後の研究計画

表6.今後の研究計画



入力画像問題調査

- ✓ 3DMMを組み込む方法
- ✓ 既存拡散モデルの推論方法の調査

Stable Diffusion・3DLDMコード解析

①：2Dto2D Stable Diffusionに表情カテゴリーエンコーダを組み込む
②：3DLDMを実装
③：3DLDMを３D顔画像で学習(ファインチューニング？)
④：表情カテゴリーを組み込んだ3D表情変換拡散モデル実装

深度情報欠如問題

- ✓ RGBD画像に対する畳み込み
- ✓ GCN
- ✓ VAEの調整
- ✓ 点群toRGBD変換の補完