

# 定期ゼミ

3D ビューー一貫性を保った顔の表情変換に関する研究

2024年11月22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
  - *DECA*
  - *EMOCA*
  - *DiffusionRig*
3. 提案モデル DiffusionRig-EMO
4. 実装結果
5. 今後の方針

# 1. 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 × NeRFを用いた手法が多い

## 研究内容

### ● 拡散モデルを用いた3D顔生成器の開発

- 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
- 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
- DiffusionRig[5]を用いた表情編集は違和感あり

## 2. 関連研究紹介

## Learning an Animatable Detailed 3D Face Model from In-The-Wild Images

### ● できること

- 1枚の顔画像から、顔の特徴をモデリングし、既存の3Dモデル(FLAME)と組み合わせることで、精巧な3D顔形状を復元
- 高速で推論が可能: 120fps (w/Nvidia Quadro RTX 5000を使用)

### ● 特徴

- 入力画像は単一画像
- 個人の特徴(アイデンティティ)を有する顔モデルを復元
- 学習時に3D教師データが不要



図. DECAによる出力画像(左から1列目)[6]

## 2. 先行研究紹介

## Learning an Animatable Detailed 3D Face Model from In-The-Wild Images

### ● FLAMEの仕組み

六つのパラメーターによって3D顔モデルが構成される

identity shape  $\beta$   
facial expression  $\psi$   
pose parameter  $\theta$

3D顔モデルの形状を決定



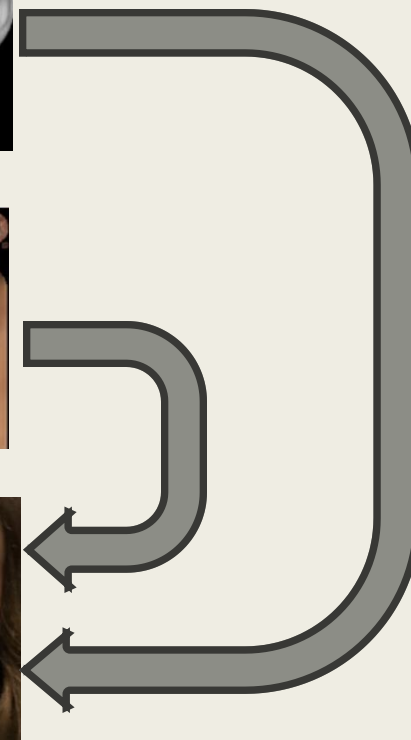
texture parameter  $\alpha$

表面のテクスチャ(albedo)を決定



lighting parameter  $l$   
camera parameter  $c$

レンダリングに使用



### 2. 先行研究紹介

## Learning an Animatable Detailed 3D Face Model from In-The-Wild Images

### ● DECAのアーキテクチャ

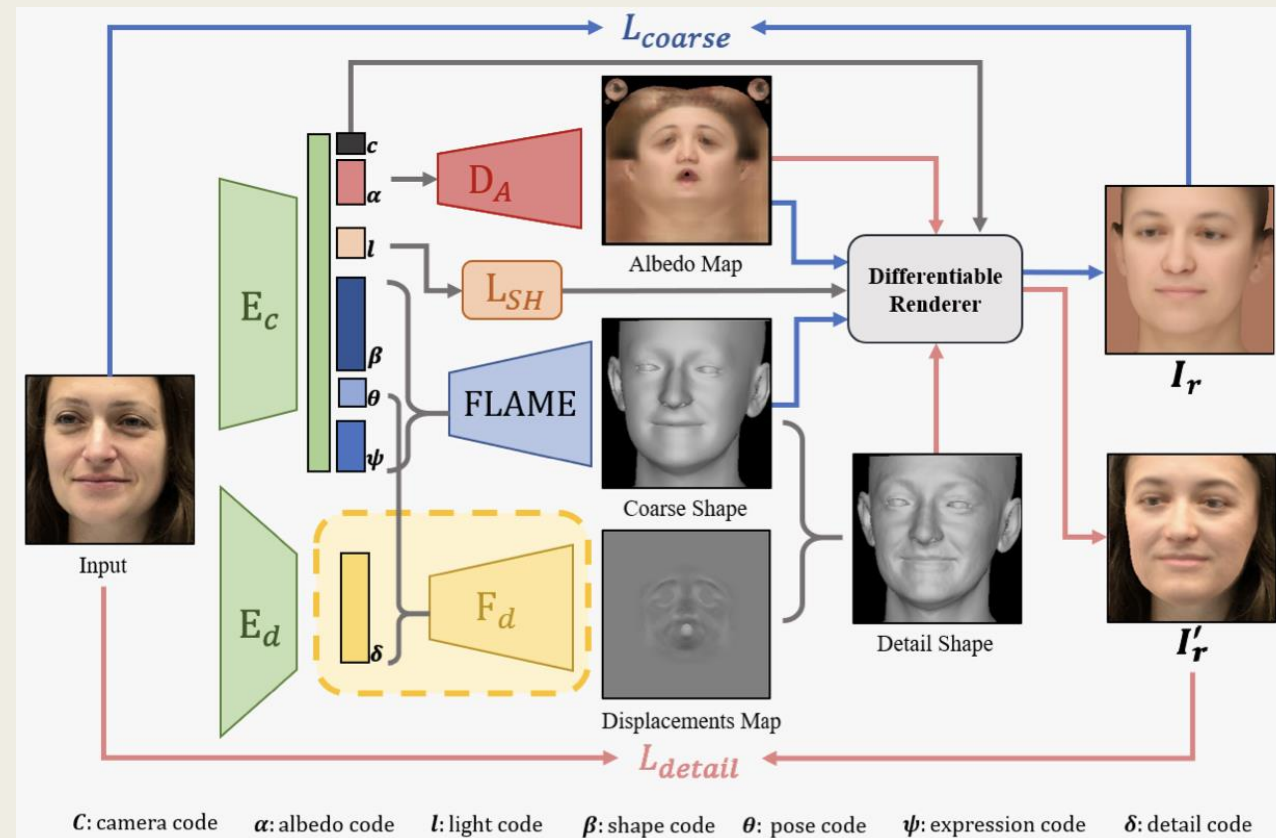


図. DECAのアーキテクチャ[6]

### 2. 先行研究紹介



## Learning an Animatable Detailed 3D Face Model from In-The-Wild Images

- 3段階の学習でネットワークを最適化

①Pre-training stage → ②Coarse stage → ③Detail stage

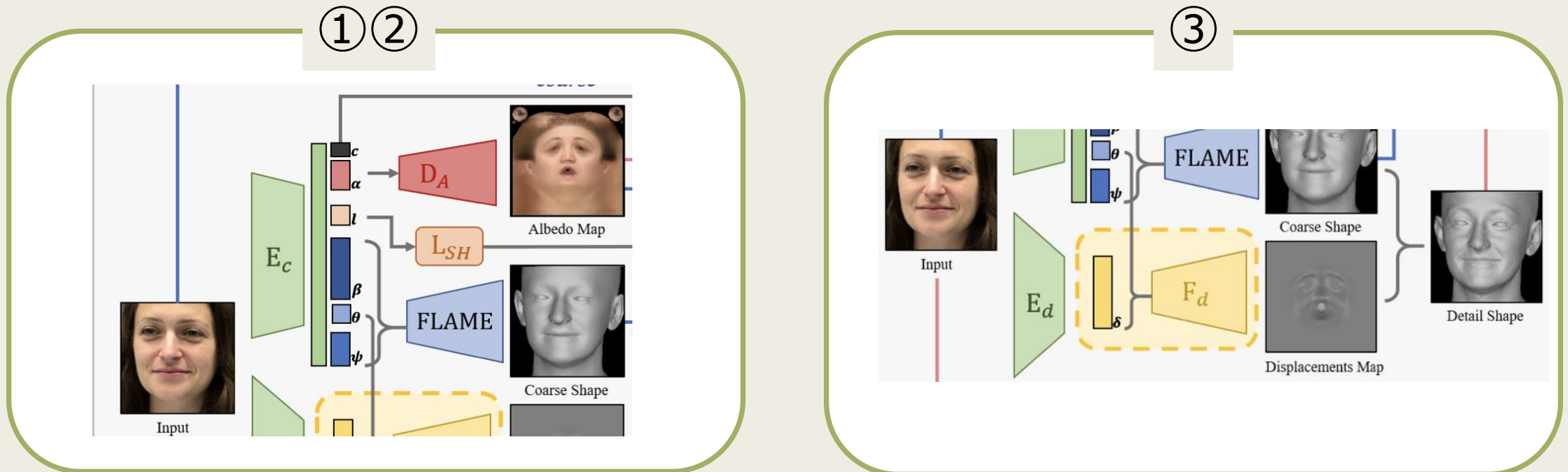


図. DECAのアーキテクチャ[6]

## 2. 先行研究紹介

# Learning an Animatable Detailed 3D Face Model from In-The-Wild Images

- 損失関数(Pre-training stage)

$$L_{pretrain} = L_{lmk} + L_{eye} + L_{reg} (+ L_{lip})$$

$L_{lmk}$  : ランドマーク座標誤差の最小化

$L_{eye}$  : 目の部分のランドマーク座標の差分を一致させる

$L_{lip}$  : 口の部分のランドマーク座標の差分を一致させる

$L_{reg}$  : shape, expressionパラメータに感じてL2正則化

## 2. 先行研究紹介

# Learning an Animatable Detailed 3D Face Model from In-The-Wild Images

- 損失関数(Coarse stage)

$$L_{Coarse} = L_{lmk} + L_{eye} + L_{pho} + L_{id} + L_{sc} + L_{reg}$$

$L_{pho}$  : 再構成損失...レンダリング後の画像とGT画像のL1距離を最小化

$L_{id}$  : 顔認識ネットワークを用いて顔の特徴量のコサイン距離を最小化

$L_{sc}$  : 同一人物の複数枚の画像からshapeパラメータ $\beta$ のみを交換し、  
 $L_{lmk} + L_{eye} + L_{pho} + L_{id} + L_{reg}$  を計算

$L_{reg}$  : shape  $\beta$ , expression  $\psi$ , texture  $\alpha$  パラメータに感じてL2正則化

## 2. 先行研究紹介

## Learning an Animatable Detailed 3D Face Model from In-The-Wild Images

- 損失関数(Detail stage)

$$L_{Detail} = L_{phoD} + L_{mrf} + L_{sym} + L_{dc} + L_{regD}$$

$L_{phoD}$  : 再構成損失...UV空間における顔領域のL1距離を最小化

$L_{mrf}$  : ID-MRFというinpaintingタスクで最初に用いられた損失  
VGG19によって抽出された特徴感パッチの差分を最小化

$L_{sym}$  : 見えていない部分へのロバスト性向上効果  
見えてない部分は左右対称とみなす

$L_{dc}$  : 同一人物の複数枚の画像から $\delta$ のみを交換し、 $L_{phoD} + L_{mrf} + L_{sym}$  を計算

## 2. 先行研究紹介

## Learning an Animatable Detailed 3D Face Model from In-The-Wild Images

### ● 定量評価

#### ①Now benchmark

100人の顔画像(2054枚)+3Dスキャンデータ

Method	Median (mm)	Mean (mm)	Std (mm)
3DMM-CNN [Tran et al. 2017]	1.84	2.33	2.05
PRNet [Feng et al. 2018b]	1.50	1.98	1.88
Deng et al.19 [2019]	1.23	1.54	1.29
RingNet [Sanyal et al. 2019]	1.21	1.54	1.31
3DDFA-V2 [Guo et al. 2020]	1.23	1.57	1.39
MGCNet [Shang et al. 2020]	1.31	1.87	2.63
DECA (ours)	<b>1.09</b>	<b>1.38</b>	<b>1.18</b>

#### ②Feng et al. benchmark

135人の顔画像(2000枚)+3Dスキャンデータ  
低解像度(LQ) & 高解像度(HQ)

Method	Median (mm)		Mean (mm)		Std (mm)	
	LQ	HQ	LQ	HQ	LQ	HQ
3DMM-CNN [Tran et al. 2017]	1.88	1.85	2.32	2.29	1.89	1.88
Extreme3D [Tran et al. 2018]	2.40	2.37	3.49	3.58	6.15	6.75
PRNet [Feng et al. 2018b]	1.79	1.59	2.38	2.06	2.19	1.79
RingNet [Sanyal et al. 2019]	1.63	1.59	2.08	2.02	1.79	1.69
3DDFA-V2 [Guo et al. 2020]	1.62	1.49	2.10	1.91	1.87	<b>1.64</b>
DECA (ours)	<b>1.48</b>	<b>1.45</b>	<b>1.91</b>	<b>1.89</b>	<b>1.66</b>	1.68

図. 3D Face Reconstruction Modelの定量的な比較結果[6]

## 2. 先行研究紹介

## Learning an Animatable Detailed 3D Face Model from In-The-Wild Images

- 定性評価



図. 生成顔形状の比較(Coarse)[6]

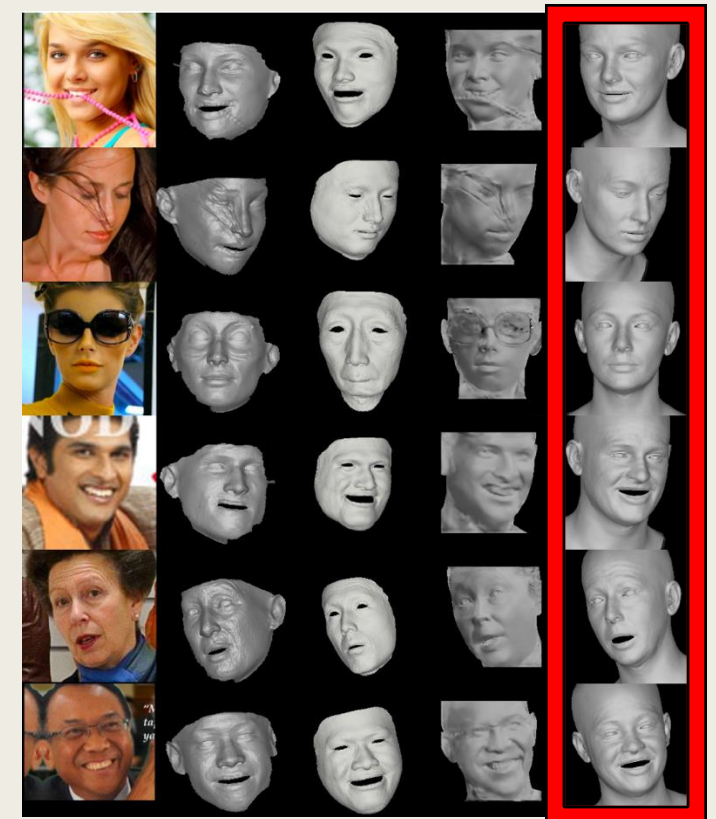


図. 生成顔形状の比較(Detail)[6]

## 2. 先行研究紹介



# EMOCA: Emotion Driven Monocular Face Capture and Animation

### ● 目的

- 入力画像からより表情豊かな3D顔モデルの生成を可能に
- 生成する3D顔形状が入力画像の感情と一致しない問題を、表情に関する損失感数を導入することで解決



図. DECA VS EMOCA[7]

## 2. 先行研究紹介

## EMOCA: Emotion Driven Monocular Face Capture and Animation

### ● 全体アーキテクチャ

- DECAの表情パラ $\psi$ のみ別の表情エンコーダで訓練し、感情認識用ネットワークからの特徴量で損失計算

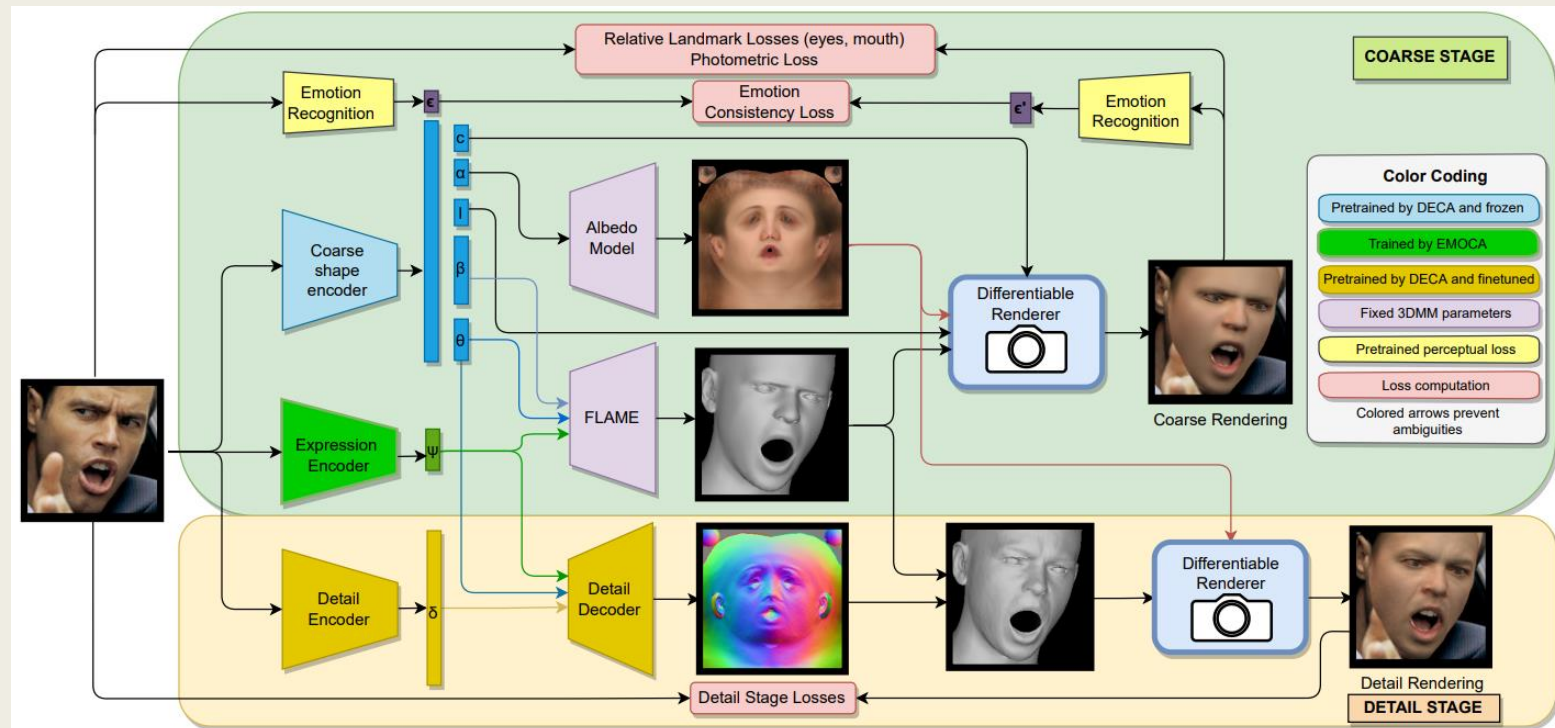


図. EMOCAのアーキテクチャ[7]

### 2. 先行研究紹介



# EMOCA: Emotion Driven Monocular Face Capture and Animation

- 損失関数

- Coarse stageでのみ変更

$$L_{coarse} = L_{emo} + L_{pho} + L_{eye} + L_{lmc} + L_{sc} + L_{id} + L_{reg}$$

- Emotion Consistency Loss

- 感情整合性損失は、入力画像の感情特徴 $\varepsilon_I = A(I)$ とレンダリング画像の感情特徴 $\varepsilon_{Re} = A(I_{Re})$ の間の差

$$L_{emo} = \|\varepsilon_I - \varepsilon_{Re}\|_2$$

- Emotion Recognition Network

- 感情離散ラベルやvalence/arousal予測するResNet-50ネットワーク
- 7つの感情離散ラベルとvalence/arousalがアノテーションされたAffectNet-DBで訓練

## 2. 先行研究紹介

## EMOCA: Emotion Driven Monocular Face Capture and Animation

- 定性評価



図. Detail形状での比較 vs DECA[7]

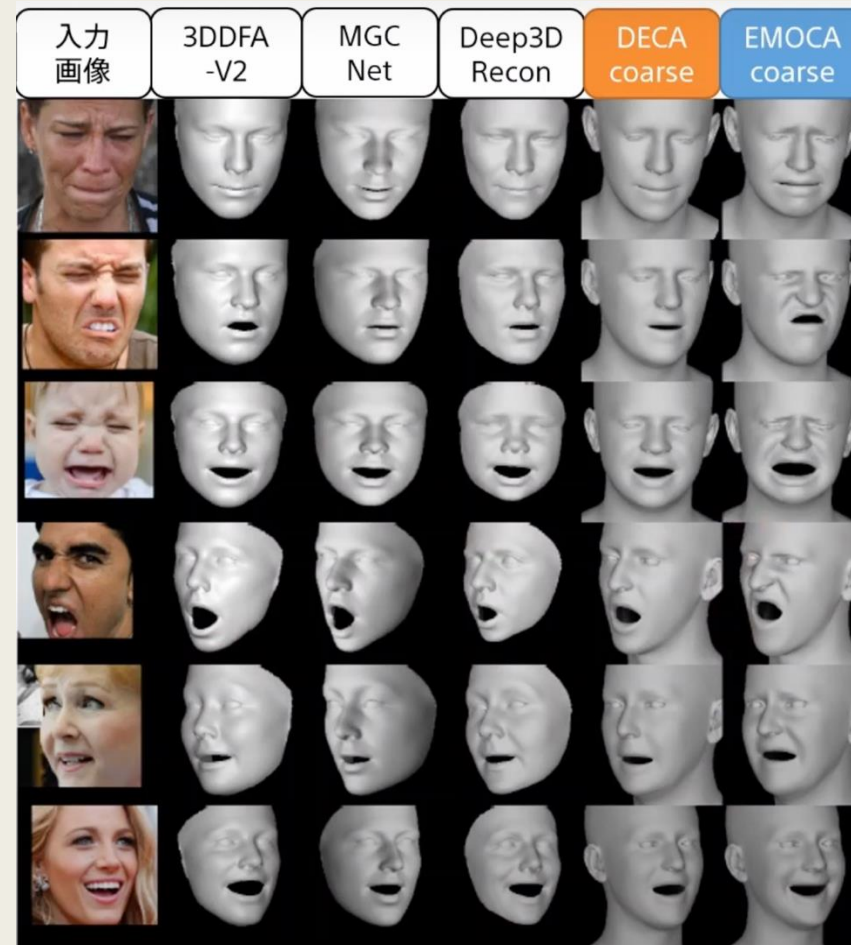


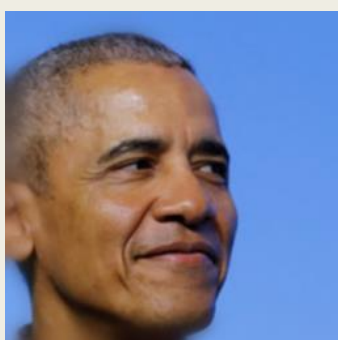
図. Coarse形状での比較 vs DECA[7]

## 2. 先行研究紹介

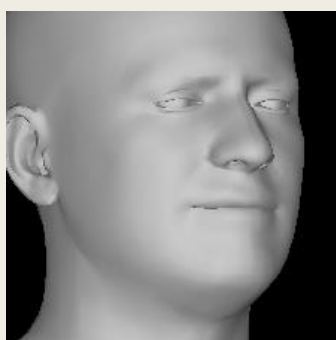
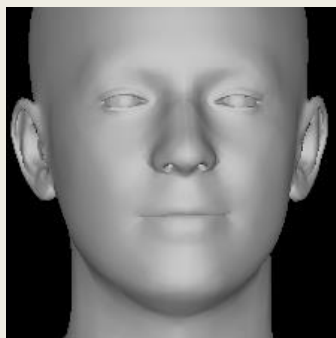
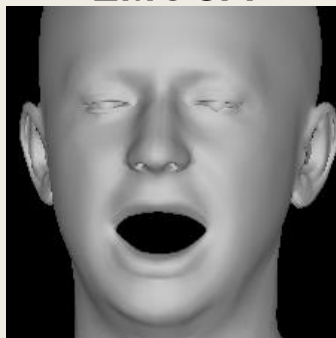
### 3. 実装結果

# EMOCA vs DECA

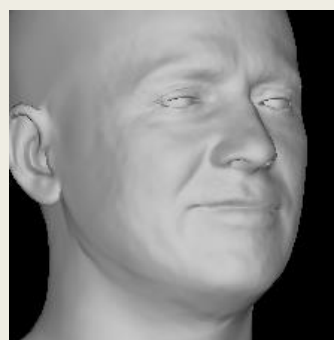
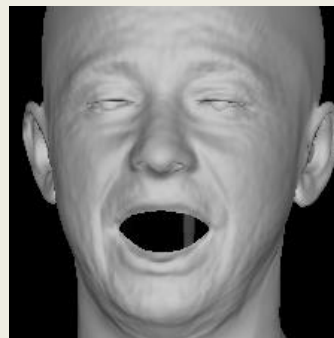
Input



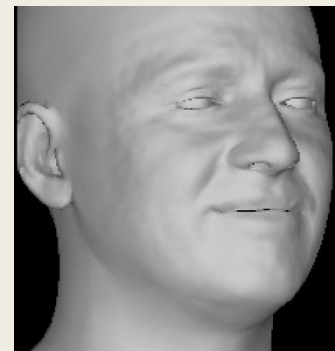
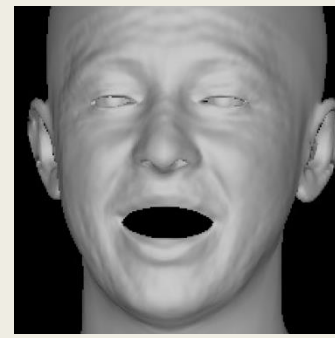
Geometry(coarse)  
EMOCA



Geometry(detail)  
EMOCA



Geometry(detail)  
DECA



Wild image(detail)  
EMOCA




















# DiffusionRigによる編集

	Source		target	生成画像	
exp					
lighting					
pose					

# DiffusionRigによる編集(表情変化)

	Source		target	生成画像	
Exp flown					
Exp smile					
Exp none					

## 4. 提案モデル



# 提案モデル DiffusionRig-EMO

## やりたいこと

1. DECAをEMOCAへ
2. 表情専用のエンコーダ追加

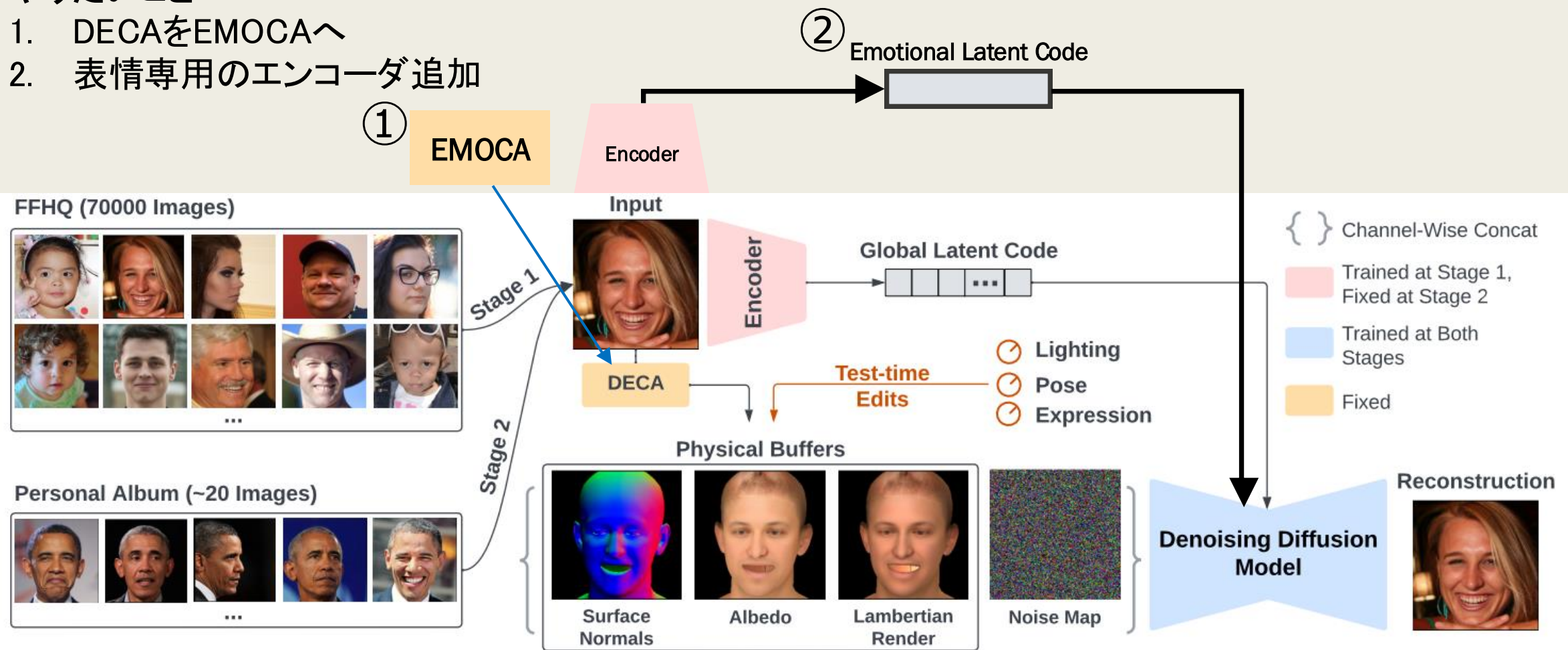


図. DiffusionRigをベースとしたDiffusionRig-EMO[5]



## 5. 今後の計画

# 今後の計画

## 課題

1. DECAをEMOCAへ
  - DECA周りのコード、pytorch3dライブラリが複雑
  - DECAとEMOCAの出力は同じであるが、レンダリング方法が少し異なる
2. 表情専用のエンコーダ追加
  - 入力画像と生成画像の感情特徴量抽出ネットワークの独自訓練が必要

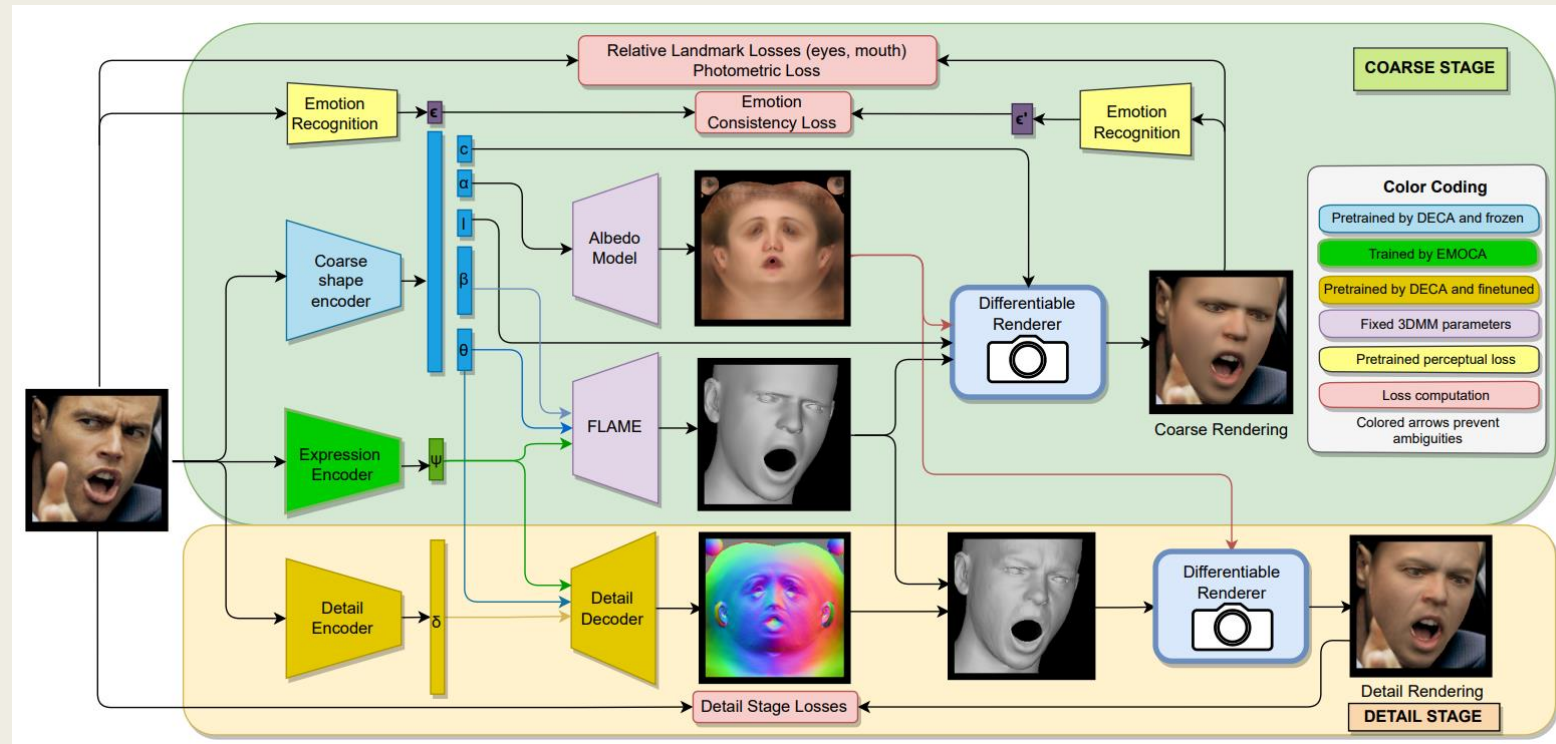


図. EMOCAのアーキテクチャ[7]

## 5. 今後の計画

## **6. appendix**

### 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

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

#### 2. 先行研究紹介

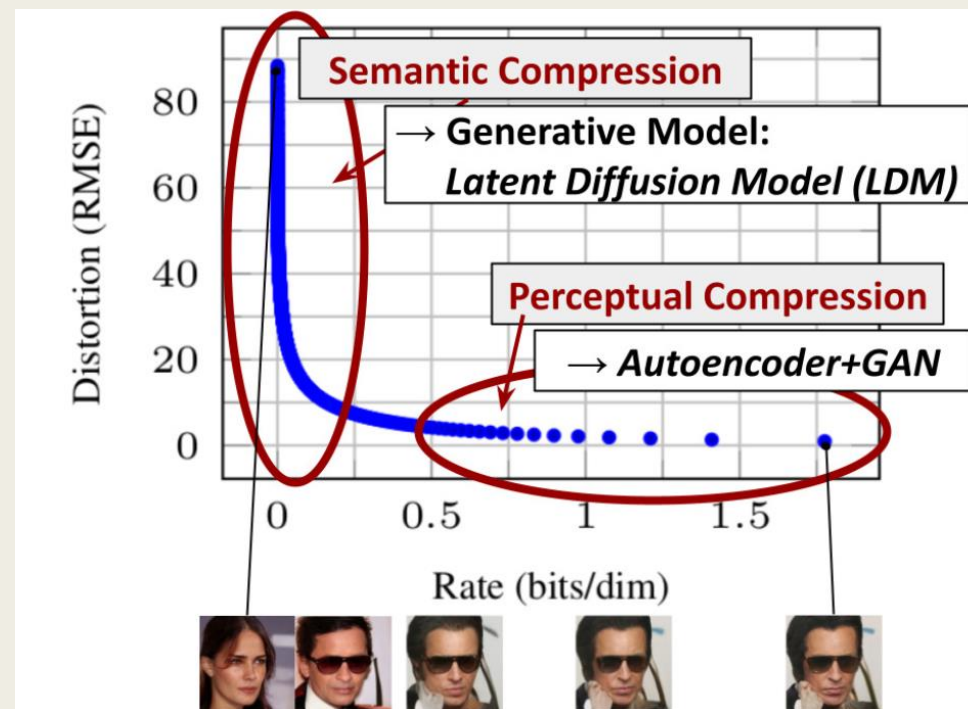


図. 生成過程の段階分け[3]

### 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、 $1048 \times 1048 \rightarrow 131 \times 131$ )
  - 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

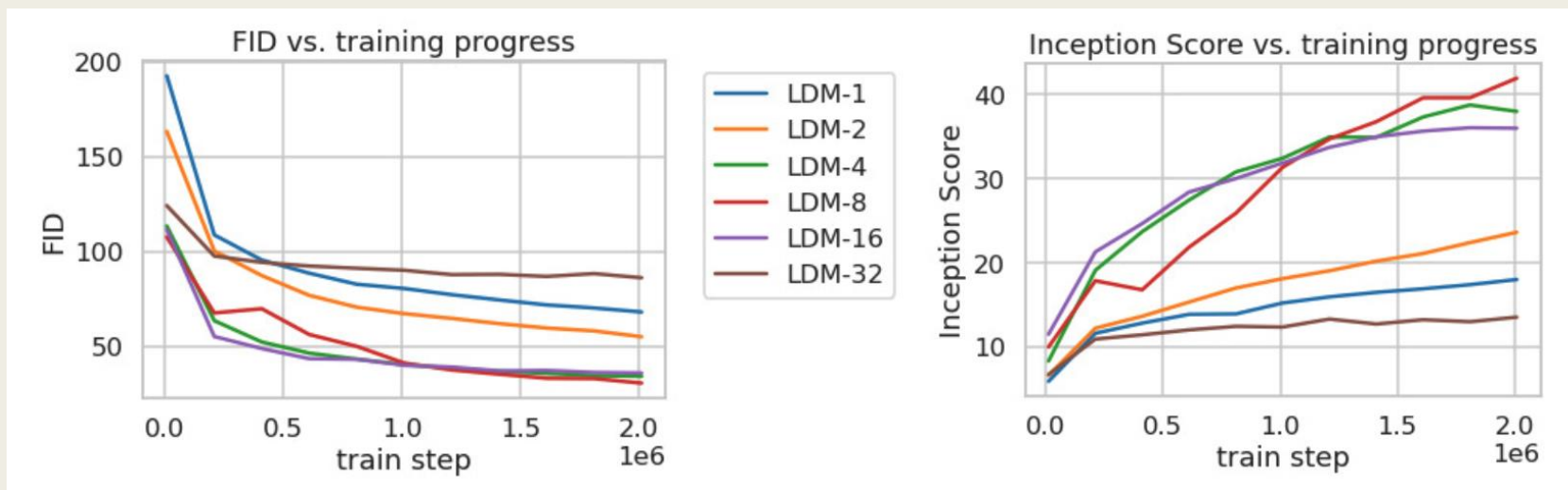


図. ダウンサンプリング因子 $f$ とFID、ISの関係[3]

### 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

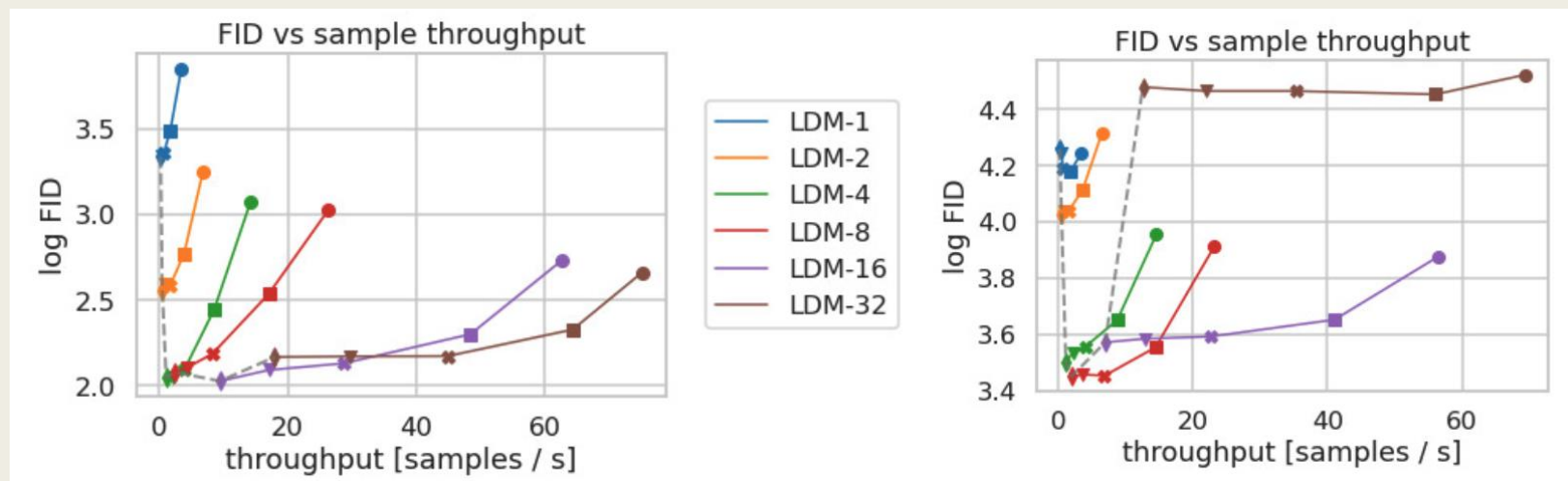


図. ダウンサンプリング因子 $f$ とFID、スループットの関係[3]



## V. Blanzらの研究

### 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
- トポロジー一貫性を持つため、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

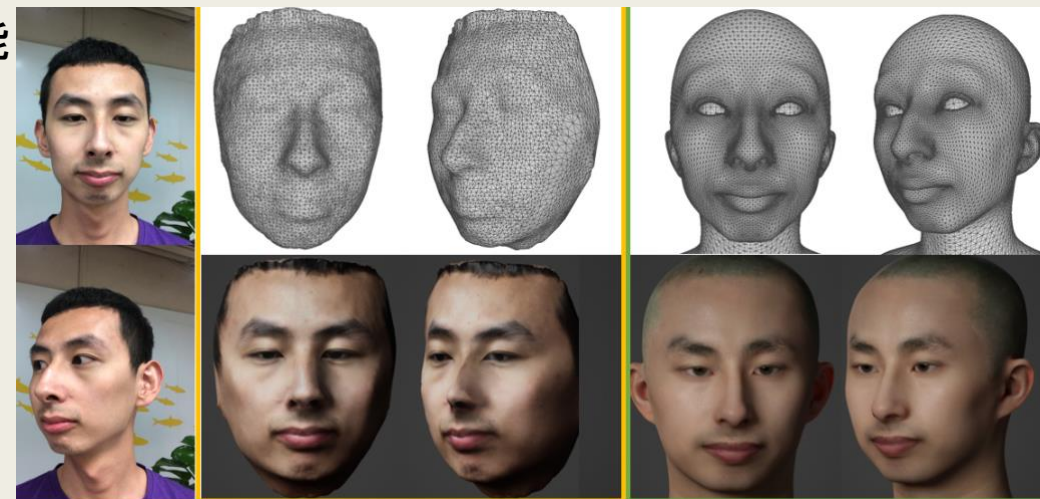


図. 3DMMを用いた場合と用いなかった場合[11]

## 2. 先行研究紹介

[10] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3d faces," p. 187–194, 1999.

[11] 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.

## 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

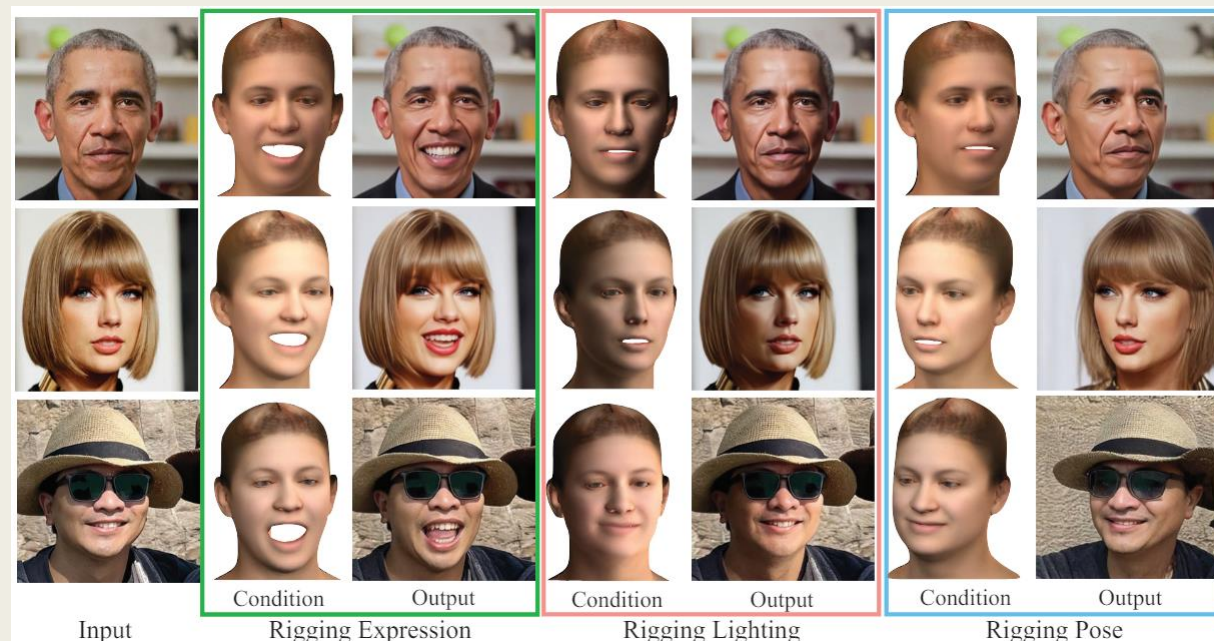


図. DiffusionRigによる編集例[5]

## 2. 先行研究紹介



## 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

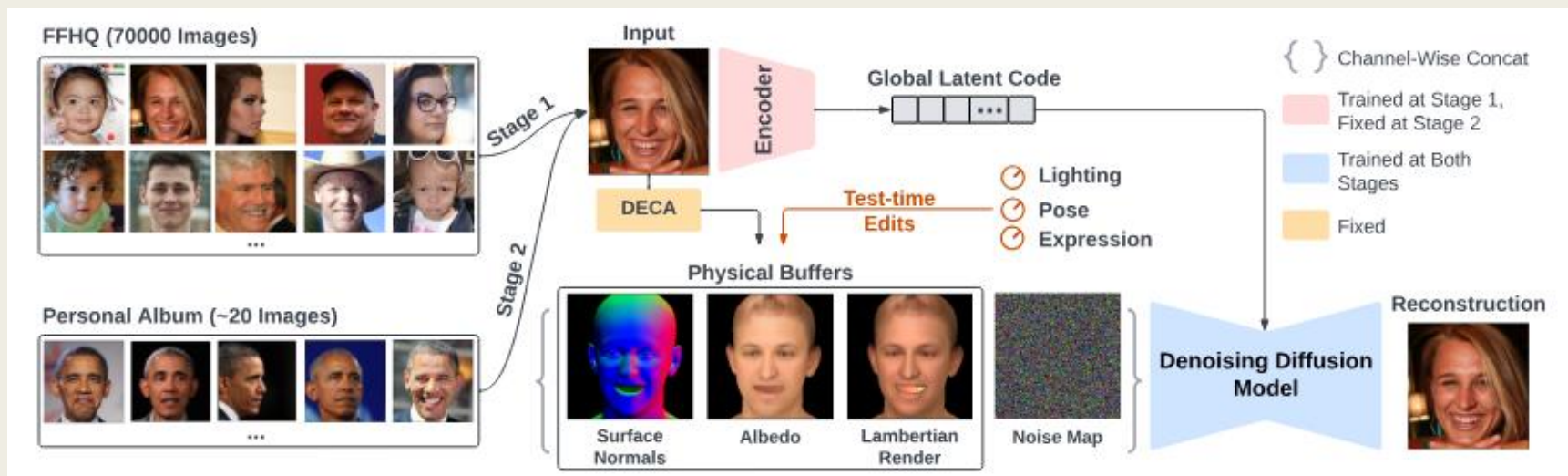


図. DiffusionRigのアーキテクチャ[5]

### 2. 先行研究紹介

- [5] 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.

## 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
- トポロジー一貫性を保ちながら、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

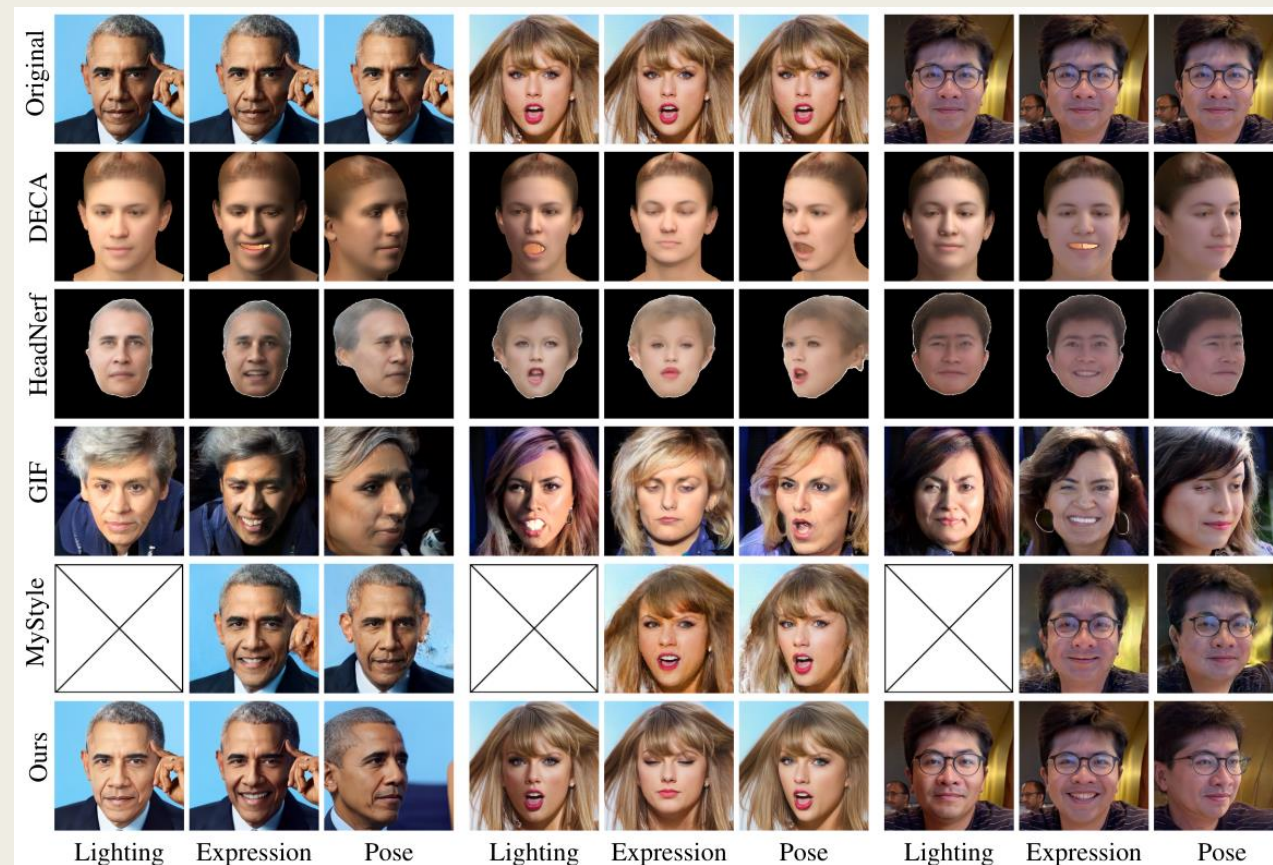


図. 各モデルの出力画像の比較[5]

## 2. 先行研究紹介

## 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

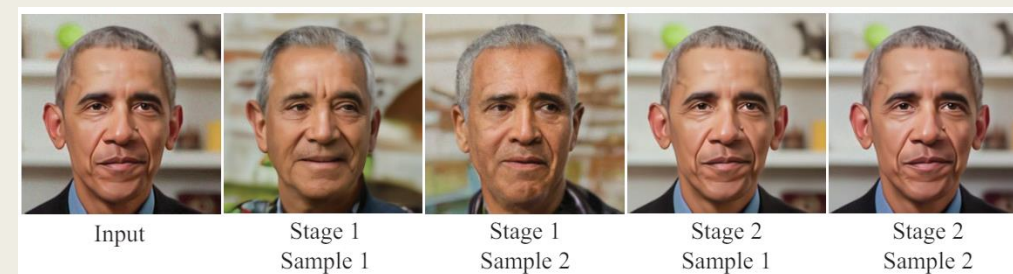


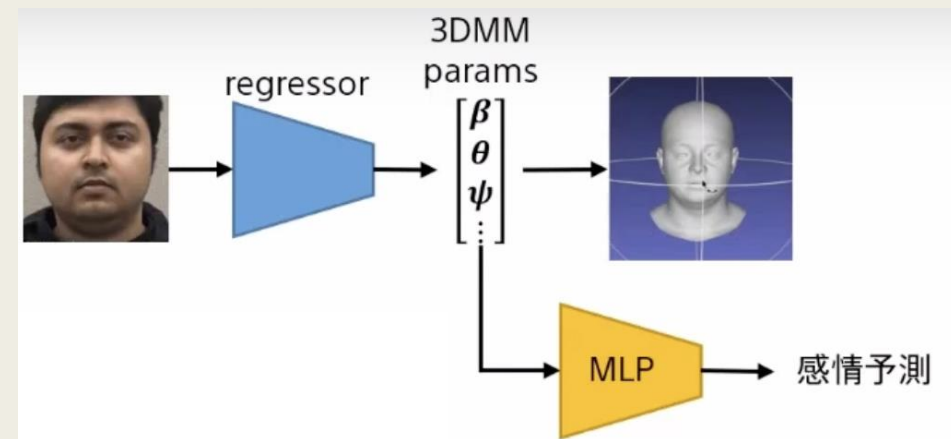
図. DiffusionRigの評価画像[5]



## EMOCA: Emotion Driven Monocular Face Capture and Animation

### ● 定量評価

- Emotion Recognitionによる定量評価
- 顔モデルを復元する際に用いる3DMMパラメータを入力とする感情予測NNを学習
- 感情予測が高精度  
⇒感情に関して詳細な特徴を表現できる3DMMパラを入力画像から抽出できている



		Valence (positive-negative)				Arousal (relaxed-intensive)				label
		V-PCC ↑	V-CCC ↑	V-RMSE ↓	V-SAGR ↑	A-PCC ↑	A-CCC ↑	A-RMSE ↓	A-SAGR ↑	E-ACC ↑
BFM	EmoNet [86]	0.75	0.73	0.32	0.80	0.68	0.65	<b>0.29</b>	0.78	0.68
	Deep3DFace [19]	0.75	0.73	0.33	0.80	0.66	0.65	0.31	0.78	0.65
	ExpNet [15]	0.45	0.42	0.43	0.73	0.39	0.36	0.38	0.64	0.46
	MGCNet [76]	0.71	0.69	0.35	0.80	0.59	0.58	0.34	0.77	0.60
	3DDFA_V2 [34]	0.63	0.62	0.39	0.75	0.53	0.50	0.34	0.73	0.52
FLAME	DECA [27]	0.70	0.69	0.36	0.76	0.59	0.58	0.33	0.74	0.59
	DECA w/ details [27]	0.70	0.69	0.37	0.77	0.59	0.57	0.33	0.77	0.58
	EMOCA (Ours)	<b>0.78</b>	<b>0.77</b>	<b>0.31</b>	<b>0.81</b>	0.69	0.68	0.30	0.81	0.68
		0.77	0.76	0.31	0.81	<b>0.70</b>	<b>0.69</b>	<b>0.29</b>	<b>0.83</b>	<b>0.69</b>

図.各生成モデル・感情認識ネットワークによる予測精度の比較[7]

### 2. 先行研究紹介

## EMOCA: Emotion Driven Monocular Face Capture and Animation

### ● Ablation Study

- 学習データの変更(FFHQ⇔Affect Net)による効果より、Emotion Consistency Lossによる効果が多い

EMOCA DS w/o EMO : DECAにExpression Encoderをつけただけ

EMOCA w/o EMO : AffectNetで学習したが、Emotion Consistency Lossなし

EMOCA DS : FFHQで学習

Model	V-PCC ↑	V-CCC ↑	V-RMSE ↓	V-SAGR ↑	A-PCC ↑	A-CCC ↑	A-RMSE ↓	A-SAGR ↑	E-ACC ↑
DECA [28]	0.70	0.69	0.36	0.76	0.59	0.58	0.33	0.74	0.59
EMOCA DS w/o Emo	0.70	0.69	0.37	0.78	0.61	0.58	0.32	0.79	0.60
EMOCA w/o Emo	0.68	0.66	0.36	0.74	0.59	0.58	0.32	0.77	0.59
EMOCA DS	0.77	0.76	<b>0.31</b>	<b>0.82</b>	<b>0.69</b>	0.67	<b>0.29</b>	0.79	<b>0.68</b>
EMOCA	<b>0.78</b>	<b>0.77</b>	<b>0.31</b>	0.81	<b>0.69</b>	<b>0.68</b>	0.30	<b>0.81</b>	<b>0.68</b>

図. Ablation studyの結果[7]