第10回定期ミーティング

2024/12/03

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

進捗報告

できたこと

- Pytorch3d問題の解決
- DiffuisionRigとEMOCAを共通の仮想環境で動かす
- EMOCAの潜在コードをDECAの微分可能レンダラーでレンダリング
- 推論過程において、ターゲット画像のEMOCA由来の表情潜在コードでソース画像を置換

これからの課題

1. EMOCA由来の潜在コードを用いたFFHQでの学習
2. グローバルエンコーダの調整
3. FFHQ + AffectNetのデータセットでの学習
4. DiffusionRigとDiffusionRig-EMOの性能差を表す定量的指標
5. DECAとEMOCAの生成時間の比較

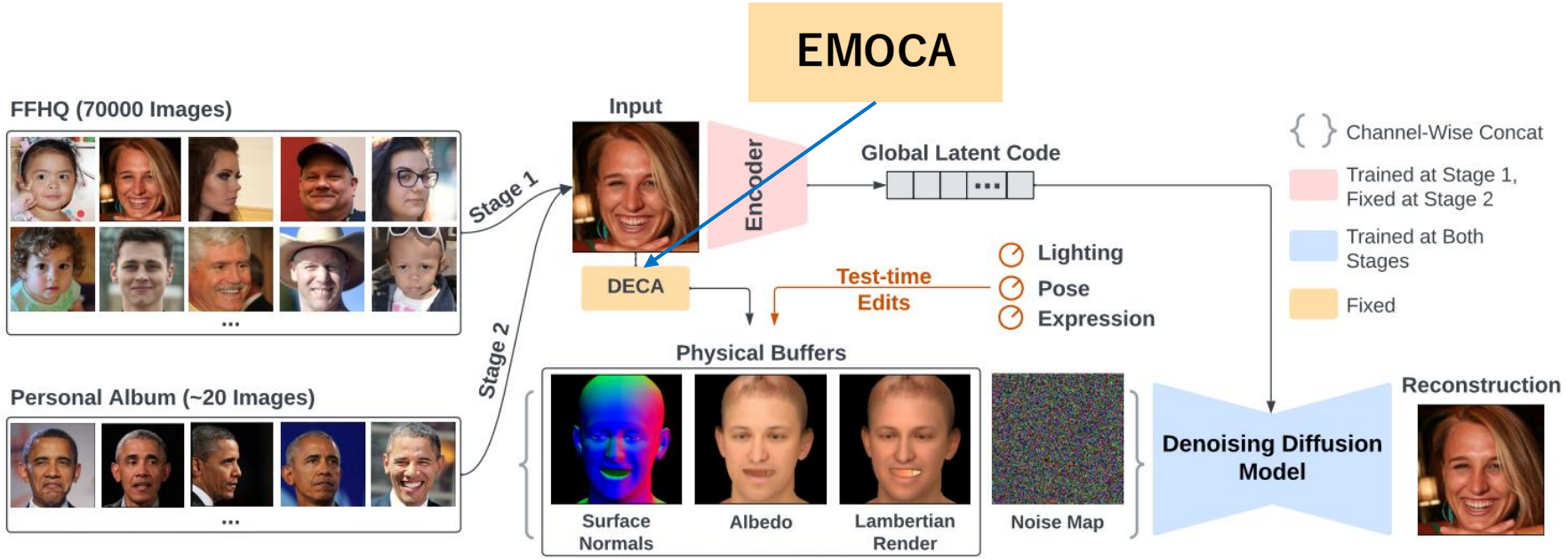
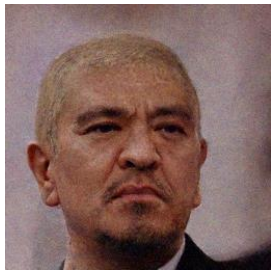


図. DiffusionRigのアーキテクチャ

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

Source				
Target				
Target:DECA Source:DECA				
Target:EMOCA Source:DECA				
Target:EMOCA Source:EMOCA				

研究の方向性

表情変換顔生成モデル(Diffsion)と2Dto3D顔形状を生成するモデルを
組み合わせることで、3Dの表情変換を軽量化かつ自然に行う

Step1：表情編集に特化したDDIMを訓練

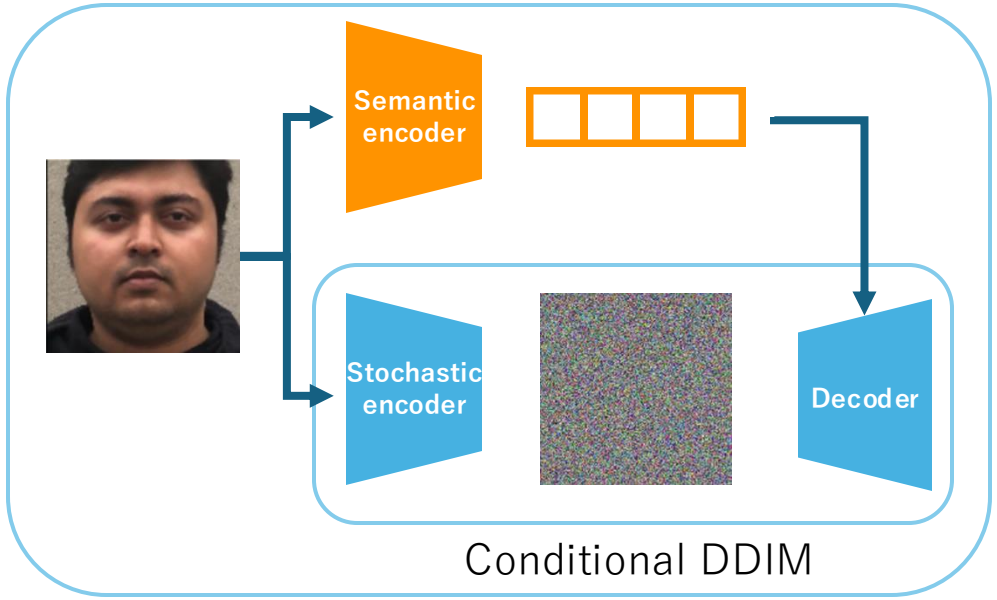


図. 表情特化DiffusionAutoencoders[1]

Step2：変換前後でβ変化しないように訓練

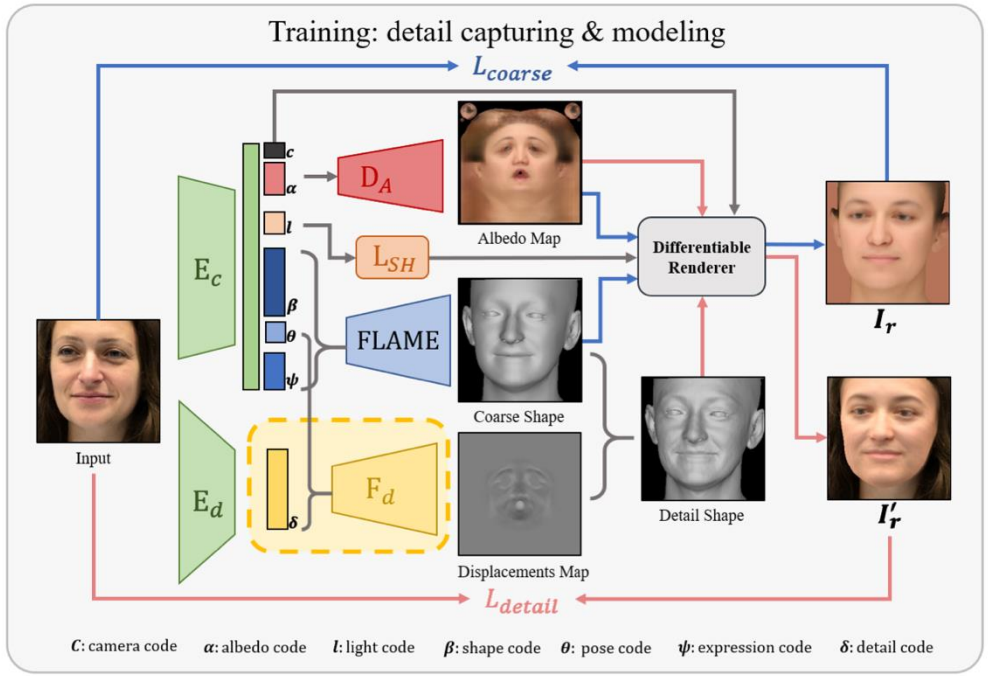


図. DECA[2]

[1] Konpat Preechakul et al. Diffusion Autoencoders: Toward a Meaningful and Decodable Representation
[2] YAO FENG et al. Learning an Animatable Detailed 3D Face Model from In-The-Wild Images

目次

1. 先行研究紹介

A）表情変換拡散モデル調査

B）AffectNet

2. 実装状況

A）DiffusionAutoencoders

B）DiffusionAutoencoders＋DECA

3. 今後の研究計画

目次

1. 先行研究紹介

A）表情変換拡散モデル調査

B）AffectNet

2. 実装状況

A）DiffusionAutoencoders

B）DiffusionAutoencoders＋DECA

3. 今後の研究計画

Photorealistic and Identity-Preserving Image-Based Emotion Manipulation with Latent Diffusion Models

Ioannis Pikoulis, Panagiotis P. Filntisis, Petros Maragos School of ECE,
National Technical University of Athens, Greece

取り組んだ問題と背景

- ✓ 画像間変換の中でもin-the-wildな顔画像に対する感情操作は先行事例が少なく未踏の領域
- ✓ 近年の拡散モデルの様々な発展により多くの応用例が考えられてきた

このモデルの特徴

- ✓ LDM(潜在拡散モデル)とCLIP潜在空間を用いたテキスト駆動型の操作を組み合わせることで in-the-wildな画像を用いた感情操作（Emotion Manipulation）を行う
- ✓ AffectNetデータセット上で広範な定性的および定量的評価を行い、画像の品質と現実性の点で優位性

VQGANの構成

VQGANのアーキテクチャ：

LDM（Latent Diffusion Model）の第一段階としてVQGAN をトレーニング

1.基本構成

1. アーキテクチャ: **U-Net [37]**
    1. ダウンサンプリングとアップサンプリングを行う**残差ブロック**を使用。
    2. 各空間解像度レベルにおける残差ブロック数は2。
2. 基本チャネル数: **128**。

2.入力画像と潜在表現

1. **入力画像サイズ**:128×128×3
2. **エンコーダの圧縮率**: f=4
    1. 圧縮後の潜在表現:32×32×3

VQGANの訓練

LDMトレーニングの設定

1.最適化と損失

1. **オプティマイザ**: Adam
2. **初期学習率**: 3.6×10^{-5}
3. **バッチサイズ**: 8
4. **エポック数**: 6
5. **再構築損失**:
   AffectNet検証セットで0.138
   (ℓ_1損失とLPIPS距離の合計)

2.モデルパラメータ

1. エンコーダとデコーダの構造は同一
2. デコーダではブロックの順序が逆になる
3. **学習可能パラメータ数**: 約6100万（61M）

3. **LDMのトレーニング設定**:
   1. **Classifier-free guidance** を使用。
   2. **ドロップアウト確率**: $P_{uncond} = 0.2$
      1. 条件なしサンプリング（unconditional sampling）のために、トレーニングデータの20%を無条件状態として処理。
   3. **DDPMのステップ数**: $T_{DDPM} = 1000$
      1. 逆方向プロセスにおけるノイズ除去のステップ数を1000に設定

CLIP-guided fine-tuning の概要：①

グローバルターゲット損失

CLIP空間において、生成された画像と与えられたターゲットテキストとのコサイン距離を最小化
主に画像とテキスト間の整合性を高める目的で使用

ローカル方向性損失

グローバル損失の課題（低い多様性や敵対的攻撃に対する脆弱性）を緩和するために設計
この損失は、CLIP空間において、参照画像と生成画像の埋め込み間の方向を、参照テキストとターゲットテキストの埋め込み間の方向と一致させることを誘導

$$L_{\text{dir}} = 1 - \cos\left(\text{CLIP}_{\text{img}}(x_{\text{gen}}) - \text{CLIP}_{\text{img}}(x_{\text{src}}),\ \text{CLIP}_{\text{text}}(y_{\text{trg}}) - \text{CLIP}_{\text{text}}(y_{\text{src}})\right)$$

CLIP-guided fine-tuning の概要：②

DiffusionCLIP を画像ベースの感情翻訳タスクに適応

1.アイデンティティ保存損失（identity preservation loss）：
ArcFace[9]を使用した事前学習済みIR-SE50モデルの埋め込み空間におけるコサイン距離で実装。
2.ピクセル空間の $\ell 2$ 損失：
ピクセル間の二乗距離。

$$L = \lambda_{\text{dir}} L_{\text{dir}} + \lambda_{\text{id}} L_{\text{id}}(x_{\text{gen}}, x_{\text{src}}) + \lambda_{\ell 2} \|x_{\text{gen}} - x_{\text{src}}\|_2^2$$

評価指標

評価指標	目的	計算手法/特徴	使用目的
PSNR	ピクセルレベルの類似度	平均二乗誤差 (MSE) を基に計算	再構築品質評価
SSIM	構造的類似性	輝度・コントラスト・構造の3要素	再構築品質評価
LPIPS	知覚的類似性	学習済みネットワークの特徴空間での距離	再構築品質評価
感情分類精度	感情転送性能	HSEmotionでターゲット感情との一致率を計算	感情操作の正確性評価
CSIM	被写体のアイデンティティ保持	CosFaceモデルでの特徴ベクトル間のコサイン類似度	被写体特徴の保持性能評価
ユーザースタディ	リアリズムと感情表現の主観的評価	ペア比較法・感情識別タスク	視覚的品質と感情表現の検証

ハイパーパラメーター

確率性パラメータ：η

全ての実験でη=0

η>0 の場合、多様性のある出力が得られるものの、元の画像との一貫性が低下するため

編集強度：t_0

ノイズ除去プロセスの途中で生成を開始する位置を指定。

値が高いほど、元の画像と比較して操作効果が顕著になる

t_0=500 がバランス良く感情転送効果とアイデンティティ保持を両立。

ステップ数：T_{DDIM}

T_{DDIM}=20 以上でサンプル品質が良好になるが、

過度なステップ数はLPIPSの増加やSSIMの低下を引き起こす場合もある。

無条件誘導スケール：γ

条件付きスコア と無条件スコア の比率を制御し、感情操作の強度を調整。

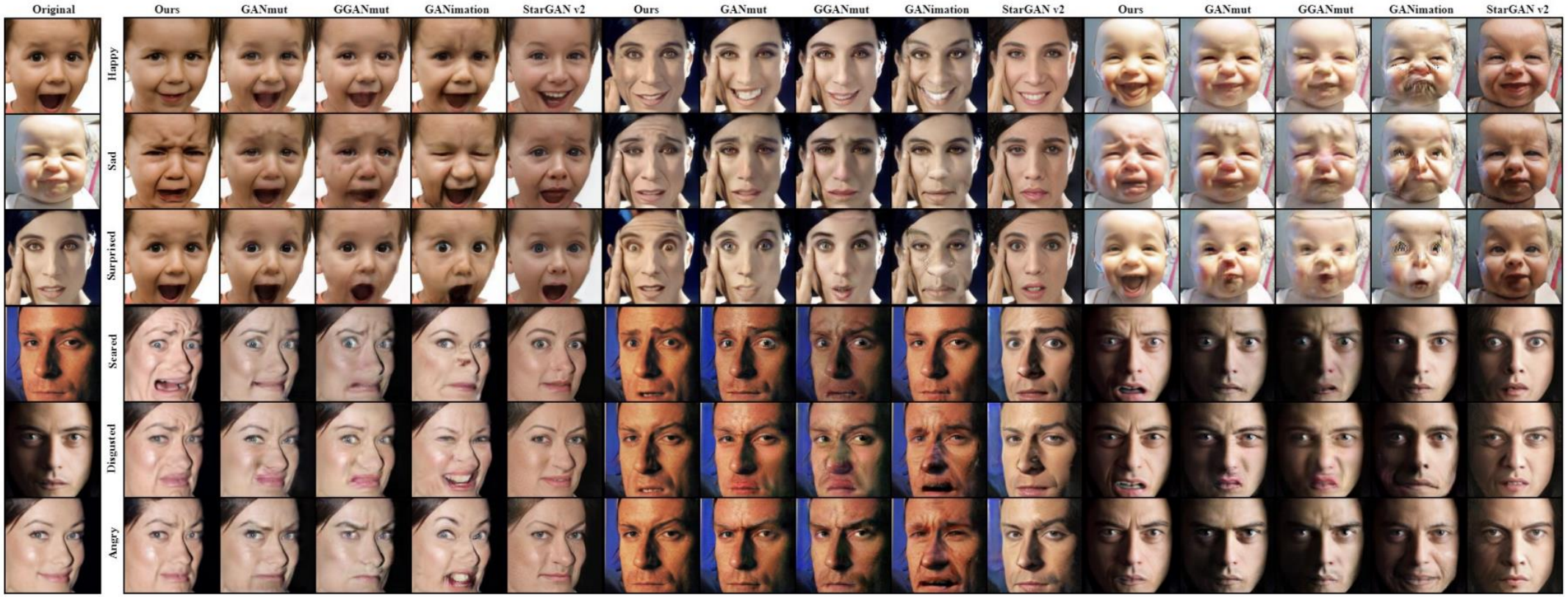
高いγ値が分類精度を向上させるが、画像品質（PSNR, SSIM）が若干低下

γ = 4.0,5.0で感情表現が明確になりつつ、顔の特徴の歪みが最小限

定量評価

Method \ y_{target}	Happy					Sad					Surprised				
	Accuracy↑	PSNR↑	SSIM↑	LPIPS↓	CSIM↑	Accuracy↑	PSNR↑	SSIM↑	LPIPS↓	CSIM↑	Accuracy↑	PSNR↑	SSIM↑	LPIPS↓	CSIM↑
Groundtruth [30]	0.758	–	–	–	–	0.638	–	–	–	–	0.606	–	–	–	–
GANimation [34]	0.645	24.07	0.816	0.099	0.547	0.212	24.52	0.830	0.097	0.582	0.360	23.82	0.817	0.101	0.559
StarGAN v2 [7]	**0.958**	17.46	0.659	0.165	0.441	0.569	18.30	0.712	0.149	0.593	0.761	17.99	0.678	0.160	0.503
GANmut [8]	0.879	21.42	0.809	0.106	0.663	0.888	23.85	**0.857**	0.094	0.755	0.829	22.43	0.810	0.112	0.675
GGANmut [8]	0.934	21.91	0.819	0.103	0.717	**0.986**	22.13	0.802	0.115	0.653	**0.970**	22.37	0.777	0.121	0.610
Ours	0.872	**25.80**	**0.841**	**0.090**	0.743	0.774	**25.71**	0.837	0.095	0.778	0.658	**25.54**	**0.838**	**0.094**	0.716
Ours w/ CGF	0.883	24.30	0.813	0.098	**0.744**	0.875	24.72	0.822	**0.093**	**0.794**	0.752	23.42	0.797	0.113	**0.721**
	Scared					Disgusted					Angry				
Groundtruth [30]	0.666	–	–	–	–	0.646	–	–	–	–	0.514	–	–	–	–
GANimation [34]	0.314	23.68	0.814	0.105	0.556	0.271	24.13	0.819	0.103	0.549	0.287	24.80	0.833	0.096	0.580
StarGAN v2 [7]	0.860	17.76	0.676	0.162	0.507	0.879	18.10	0.691	0.154	0.528	0.666	18.14	0.689	0.154	0.509
GANmut [8]	0.932	23.39	**0.841**	0.102	0.721	0.877	22.64	0.815	0.113	0.663	0.856	21.57	0.813	0.106	0.678
GGANmut [8]	**0.967**	20.13	0.763	0.135	0.556	**0.987**	21.68	0.772	0.128	0.562	**0.969**	21.77	0.767	0.133	0.590
Ours	0.764	**24.89**	0.824	0.103	**0.770**	0.676	**25.50**	**0.835**	0.100	0.707	0.710	**25.66**	**0.840**	**0.094**	0.714
Ours w/ CGF	0.764	24.83	0.826	**0.096**	0.764	0.450	24.87	0.822	**0.089**	**0.714**	0.886	24.18	0.796	0.106	**0.735**

定性評価


Original
Ours
GANmut
GGANmut
GANimation
StarGAN v2
Happy
Sad
Surprised
Scared
Disgusted
Angry

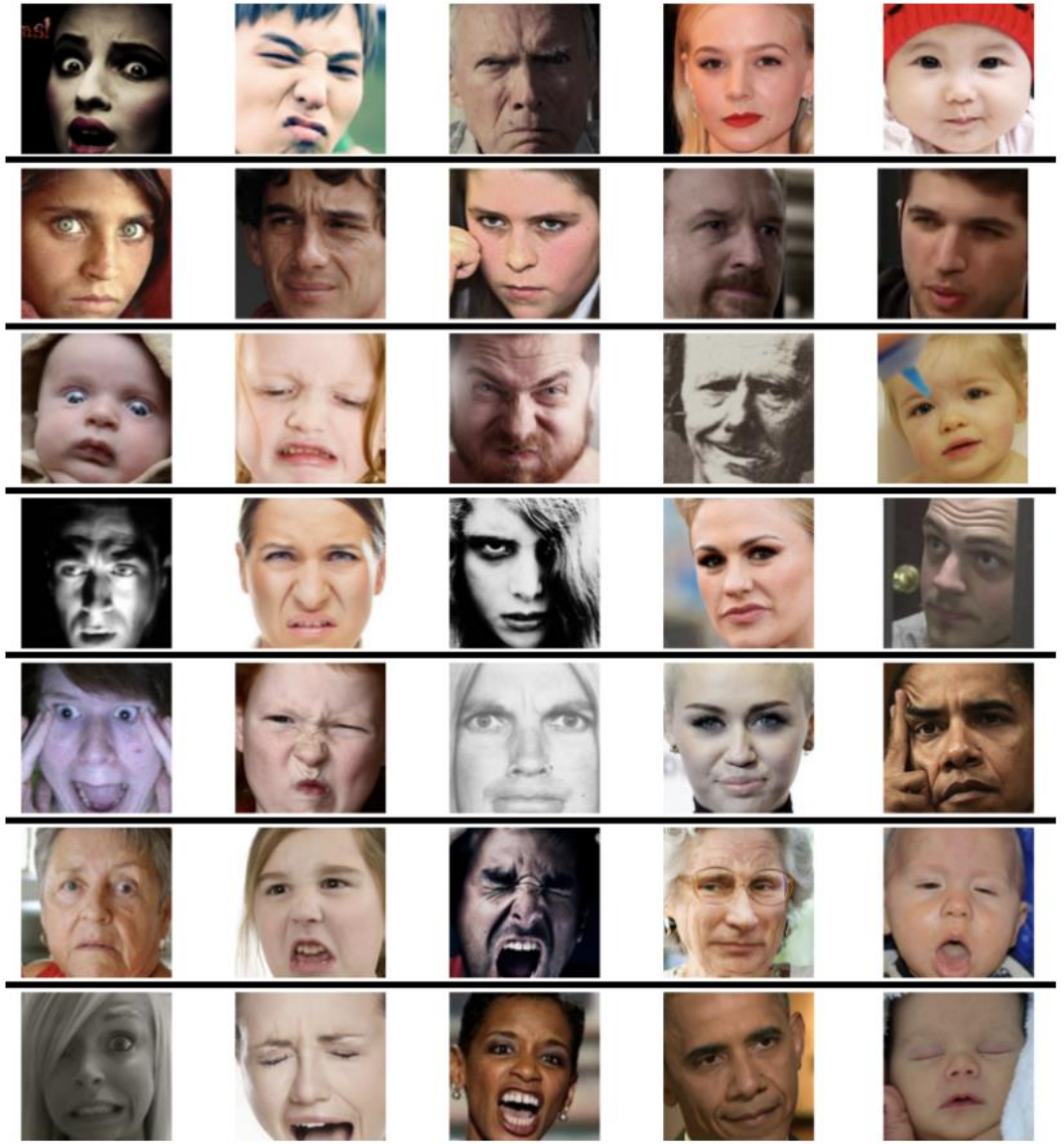
データセット紹介

AffectNet：
顔の表情に関するアノテーション付きデータセット(100万枚)

収集方法：
主要な検索エンジン3つを使用し、6つの異なる言語で1250の感情関連キーワードを用いてインターネット上から収集

ラベル：
valenceとarousal　&　8つの離散的な顔の表情

Neutral	75374
Happy	134915
Sad	25959
Surprise	14590
Fear	6878
Disgust	4303
Anger	25382
Contempt	4250
None	33588
Uncertain	12145
Non-Face	82915
Total	420299



[3] http://mohammadmahoor.com/affectnet/

目次

1. 先行研究紹介

A) DiffusionAutoencoders

B) AffectNet

2. 実装状況

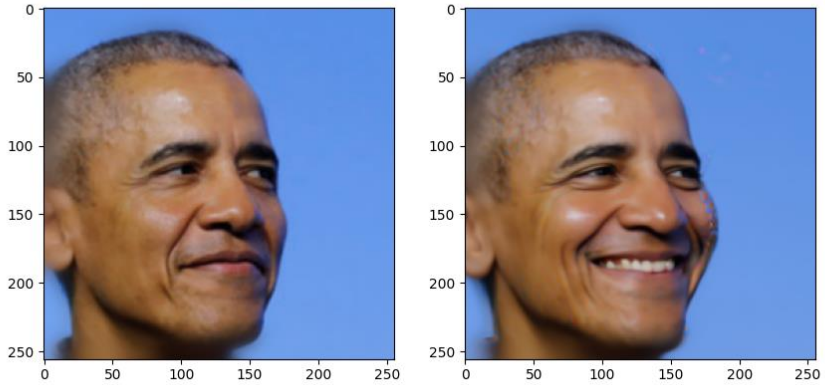
A) DiffusionAutoencoders

B) DiffusionAutoencoders＋DECA

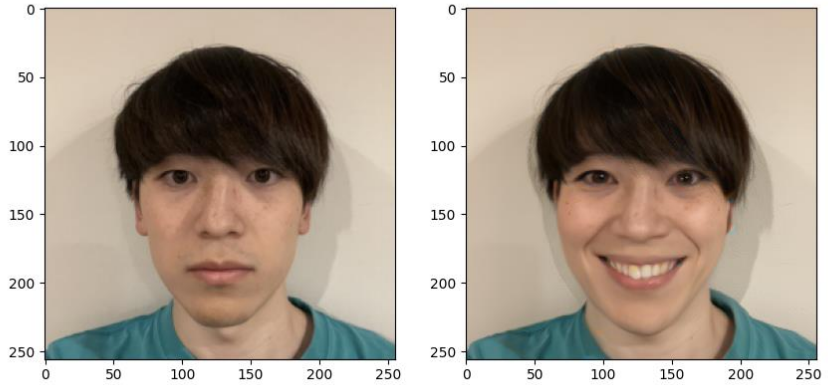
3. 今後の研究計画

DiffusionAutoencoderの出力結果

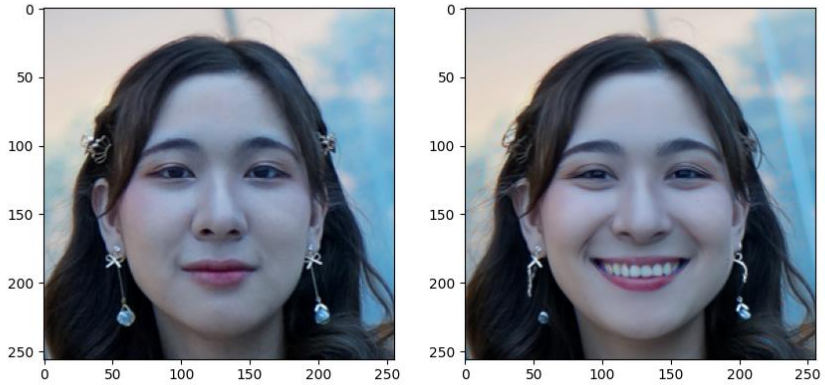
入力例①



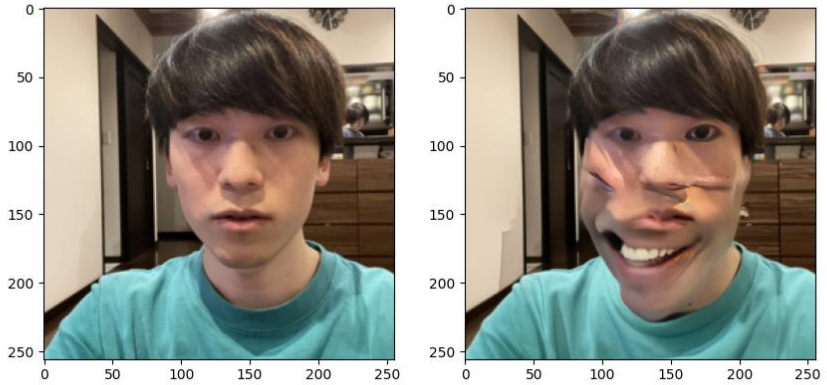
入力例②



入力例③



失敗例



実装結果

DiffusionAutoencoder + DECA の出力結果

入力画像

テクスチャマップ

出力
3D顔モデル

目次

1. 先行研究紹介

A) DiffusionAutoencoders
B) AffectNet

2. 実装状況

A) DiffusionAutoencoders
B) DiffusionAutoencoders＋DECA

3. 今後の研究計画

今後の研究計画

Step1：表情編集に特化したDDIMを訓練

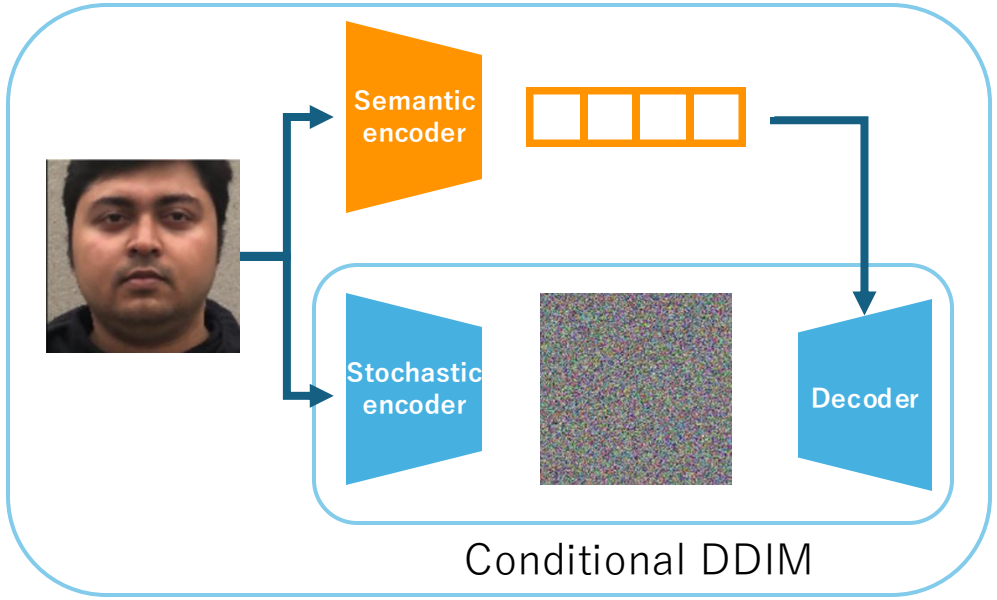


図. 表情特化DiffusionAutoencoders[1]

表情の条件付けに特化した画像変換生成

- 方法：
  AffectNetを用いて、表情の潜在コードを出力するエンコーダーを訓練
- メリット：
  AffectNetの8種類の感情に合わせて変換可能
  テクスチャ付きの3D顔画像が生成可能
- 懸念点：
  表情以外への適応が不可能に
  学習時間

[1] Konpat Preechakul et al. Diffusion Autoencoders: Toward a Meaningful and Decodable Representation

今後の研究計画

Step2：変換前後でβ変化しないように訓練

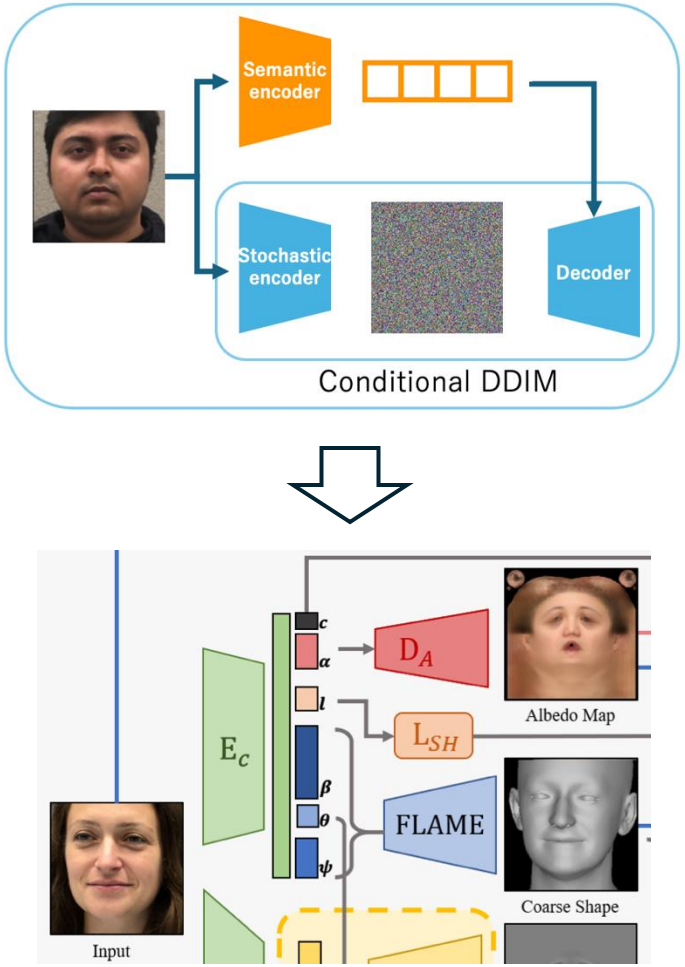


図. 提案モデル [1][2]

DECAの表情以外のパラメーターを用いて、
DECAに特化した表情変換を行う

- 方法：
  表情変換前後から生成されるFLAMEのパラメーターを比較し、感情を表すパラメーター以外が変化しないように調整

- メリット：
  FLAMEの表現に特化するのでより自然な3D顔モデルを生成可能

- 懸念点：
  拡散モデルへの損失関数追加方法
  学習時間
  評価手法

[1] Konpat Preechakul et al. Diffusion Autoencoders: Toward a Meaningful and Decodable Representation
[2] YAO FENG et al. Learning an Animatable Detailed 3D Face Model from In-The-Wild Images

今後の研究計画

今年度中に計画してる部分の実装を目指す