第13回定期ミーティング

2025/02/04

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

目次

HSEmotion: High-speed emotion recognition library

Andrey V. Savchenko
HSE University, Laboratory of Algorithms and Technologies for Network Analysis, Nizhny Novgorod, Russia

HSEmotionは、高速かつ高精度な感情認識を実現する

従来の顔表情認識（FER）の課題:

- 実験室環境で制御されたデータに基づくため、現実世界の多様な条件下での性能が低い。
- 感情データセットはデータ量が少なくノイズが多いため、モデルが偏りやすい。
- 高精度モデルは複雑で、高性能な計算機資源が必要となり、モバイルデバイスでの利用が困難。
- モデルのロバスト性が低く、多様な環境や条件に対する汎用性に欠ける。

HSEmotionの特徴:

- EfficientNetベースのCNNモデルを使用。
- 静止画と動画の両方に対応。
- 顔検出にMTCNNなどの外部ライブラリを使用。
- 8つの基本感情（怒り、軽蔑、嫌悪、恐怖、幸福、中立、悲しみ、驚き）の確率を出力。
- 感情特徴ベクトル（高次元の視覚的埋め込み）を抽出可能。
- Python3用のhsemotionパッケージを提供
- Androidデモアプリを提供

感情の覚醒度と感情の確率を出力することができる

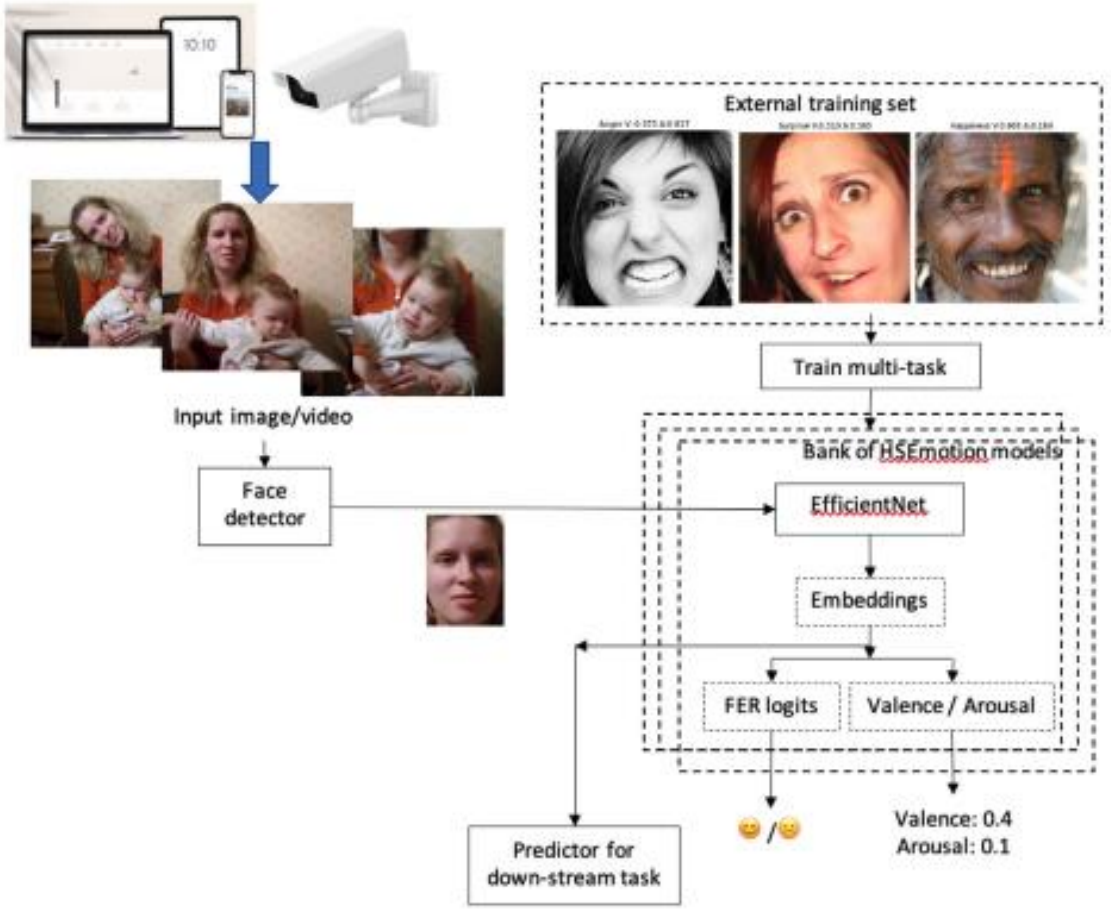


図1. HSEmotionツールを使用するためのパイプライン

- **入力:**
  静止画像または動画を入力
  動画の場合は、各フレームが個別に処理

- **顔検出:**
  MTCNNなどの外部ライブラリを用いて、
  入力画像から顔の領域を検出

- **HSEmotionモデル:**
  検出された顔領域をEfficientNetベースの
  事前学習済みCNNモデルに入力

- **出力:**
  Valence &Arousalまたは
  8つの基本感情の確率・感情分類を出力

AffectNetを用いたモデルなので、評価指標として使用可能

HSEmotionモデルのトレーニング

- **初期段階:**
  - EfficientNetをベースモデルとして使用。
  - VGGFace2データセットを用いて、顔検出器で切り取られた顔画像で顔識別タスクを実施し、EfficientNetを微調整。
- **感情認識のための微調整:**
  - AffectNetデータセットの静止画像を使用して、感情認識タスクのためにモデルを微調整

表. 使用できるモデル

モデル名	ベースモデル	入力画像サイズ	出力特徴量次元	最適化データセット	特徴
enet_b0_8_best_vgaf	EfficientNet-B0	224x224	1280	VGAF	1280次元の埋め込み
enet_b0_8_best_afew	EfficientNet-B0	224x224	1280	AFEW	1280次元の埋め込み
enet_b0_8_va_mtl	EfficientNet-B0	224x224	1280	-	8つの基本感情＋valence & arousal マルチタスク
enet_b2_8	EfficientNet-B2	260x260	1408	AffectNet	1408次元の埋め込み

引用：Andrey V. Savchenko, HSEmotion: High-speed emotion recognition library

Enet_b2_8モデルはAffentNetの検証セットでSOTAを達成(2022)

表. HSEmotionの各モデル精度

Model	Accuracy, %				F1-score	P_{MTL}	Inference time, ms
	AffectNet (8 classes)	AffectNet (7 classes)	AFEW	VGAF	LSD ABAW4	MTL ABAW4	
enet_b0_8_best_afew	60.90	64.71	**59.89**	66.80	59.32	1.110	59 ± 26
enet_b0_8_best_vgaf	61.33	64.57	55.14	68.29	59.72	1.123	
enet_b0_8_va_mtl	61.93	64.97	56.73	66.58	**60.94**	**1.276**	60 ± 32
enet_b2_8	**63.03**	**66.29**	57.78	**70.23**	52.06	1.147	191 ± 18

- EmotiW競技会の複数のサブチャレンジで最良の単一モデル として機能
- HSE-NNチームは、ABAW3競技会で以下の成績を獲得：
  - マルチタスク学習チャレンジ（MTL）：3位
  - Valence-Arousalタスク：4位
  - 表情認識タスク：4位
- ABAW4競技会では、LSDタスクで1位、MTLタスクで3位を獲得

引用：Andrey V. Savchenko, HSEmotion: High-speed emotion recognition library

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

Step2：変換前後でβ変化しないように訓練

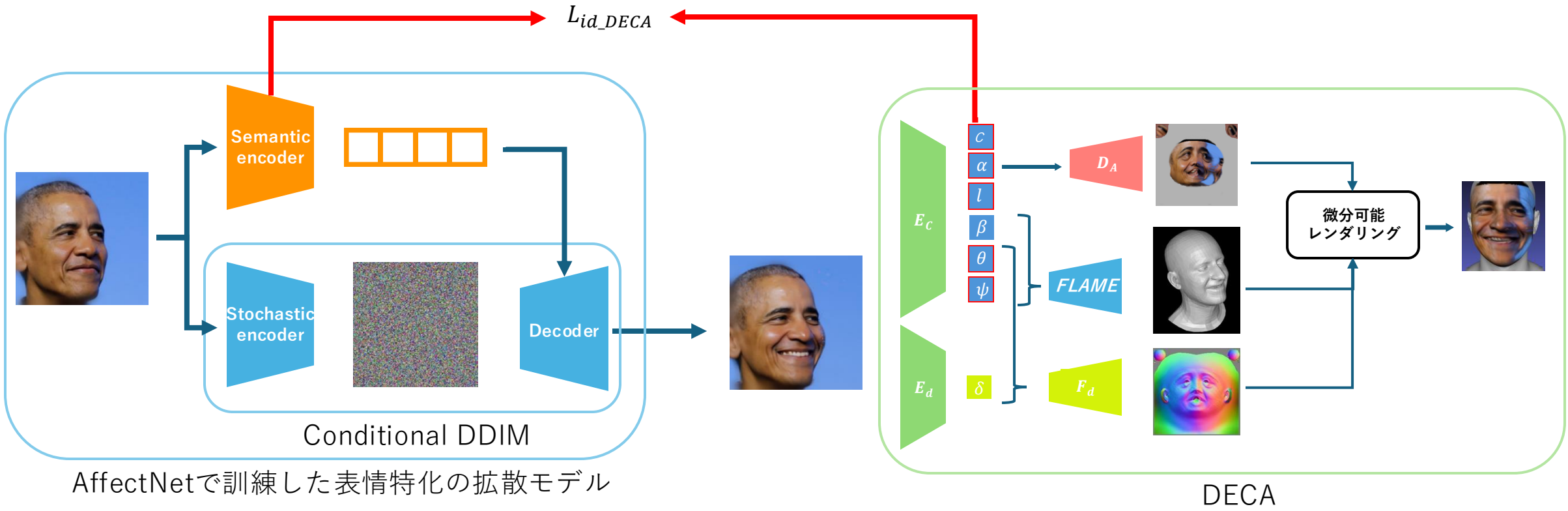


図. DECA[1]とDiffusionAutoencoders[2]を元にした提案モデル

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

研究の進捗
評価コード作成

感情分類精度・CSIMを用いて、拡散モデルでのIdentityの保持と感情分類精度を評価

表. 表情変換拡散モデルでの評価指標

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

- CSIMは全ての表情が全て満点レベル（前回実装）
- HSEmotionを今回実装

研究の進捗
感情分類精度 Part1

表情分類自体はAnger以外成功、今後、各位表情の確率を向上させる

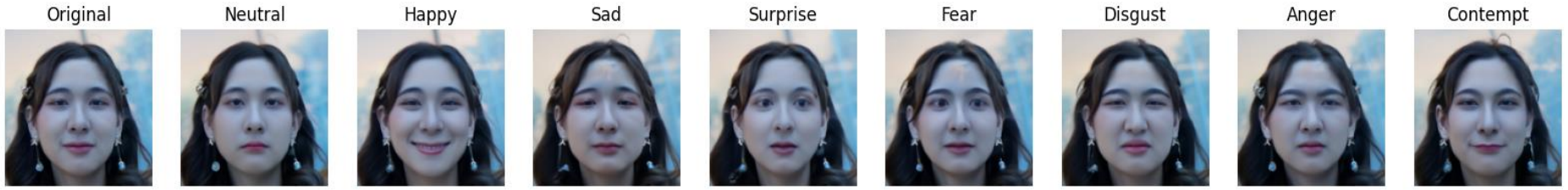


表.HSEmotionの各表情の確率

	Anger	Contempt	Disgust	Fear	Happiness	Neutral	Sadness	Surprise
Neutral	0.03864949	0.11584566	0.13711014	0.15667582	0.01892511	0.27473205	0.12030374	0.13775802
Happy	0.02480852	0.22372024	0.1931877	0.10860517	0.32003048	0.01372685	0.0387065	0.07721451
Sad	0.03075677	0.11294241	0.14292344	0.19185445	0.02270117	0.12398288	0.2730794	0.10175941
Surprise	0.02508629	0.11338066	0.11843958	0.24201417	0.04055903	0.115791	0.04722752	0.29750171
Fear	0.02500796	0.12272419	0.13342838	0.2508508	0.04425258	0.13663422	0.07321069	0.21389122
Disgust	0.04780658	0.12109236	0.22110176	0.13221651	0.03519486	0.1752103	0.14923671	0.11814097
Anger	0.07998699	0.11607216	0.17896989	0.15306064	0.02150113	0.19026837	0.1545323	0.10560852
Contempt	0.03644938	0.1692974	0.16875081	0.12296025	0.15637058	0.136866	0.0675307	0.14177483

表.HSEmotionの表情分類

Original Emotion	Neutral	Happiness	Sadness	Surprise	Fear	Disgust	Anger	Contempt
Predicted Emotion	Neutral	Happiness	Sadness	Surprise	Fear	Disgust	Neutral	Contempt

研究の進捗
感情分類精度 Part2

四つの表情で誤分類、今後、各位表情の確率を向上させる

表.HSEmotionの各表情の確率

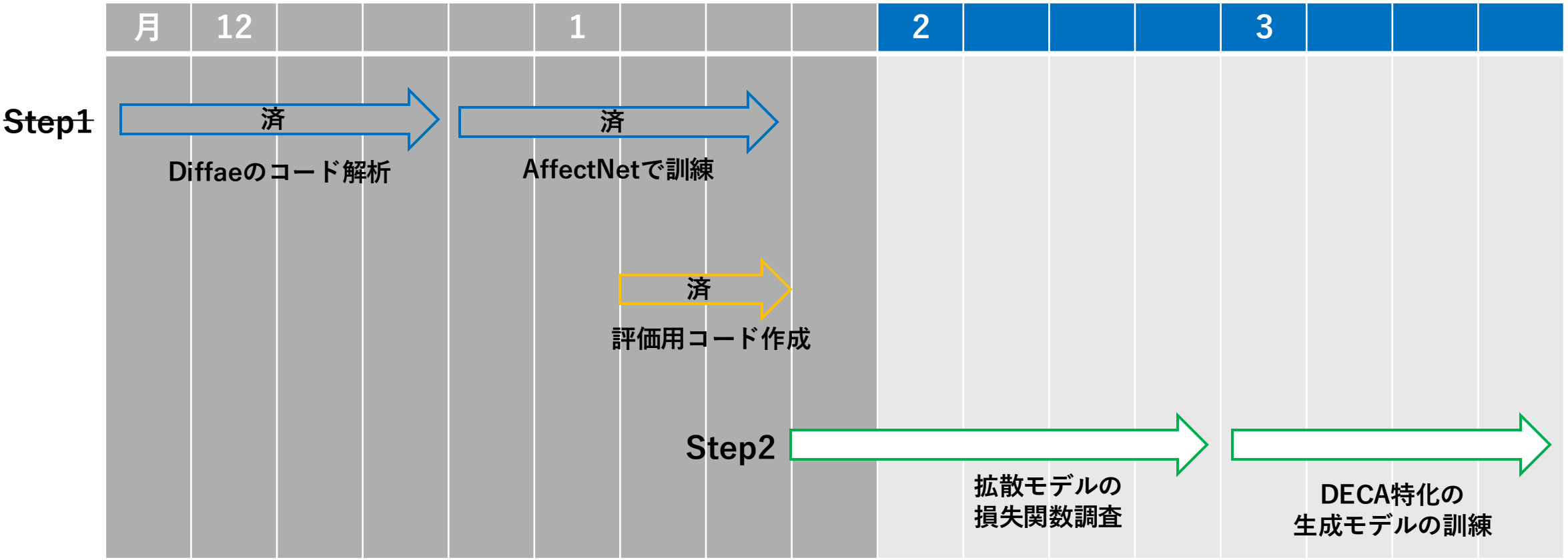
	Anger	Contempt	Disgust	Fear	Happiness	Neutral	Sadness	Surprise
Neutral	0.08983804	0.11835746	0.12076012	0.11081797	0.0100575	0.26437962	0.21224573	0.07354358
Happy	0.02257971	0.26237637	0.1955304	0.09387732	0.33390281	0.00605903	0.02859106	0.05708338
Sad	0.02389675	0.06402919	0.06746347	0.07549801	0.00375921	0.03047472	0.71220291	0.02267566
Surprise	0.02862763	0.0861389	0.12505601	0.41494909	0.02072958	0.06575909	0.12465041	0.13408922
Fear	0.02268051	0.09717619	0.11936814	0.39416319	0.02387075	0.05946692	0.18778381	0.09549052
Disgust	0.06517775	0.10123553	0.21719702	0.09206677	0.02068331	0.11065248	0.33751765	0.0554695
Anger	0.13064747	0.08380242	0.20833561	0.09529929	0.01045593	0.1069052	0.31875423	0.04579984
Contempt	0.04908623	0.21848573	0.18250702	0.10341771	0.22761634	0.07066401	0.07823687	0.06998617

表.HSEmotionの表情分類

Original Emotion	Neutral	Happiness	Sadness	Surprise	Fear	Disgust	Anger	Contempt
Predicted Emotion	Neutral	Happiness	Sadness	Fear	Fear	Sadness	Sadness	Happiness

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



目次

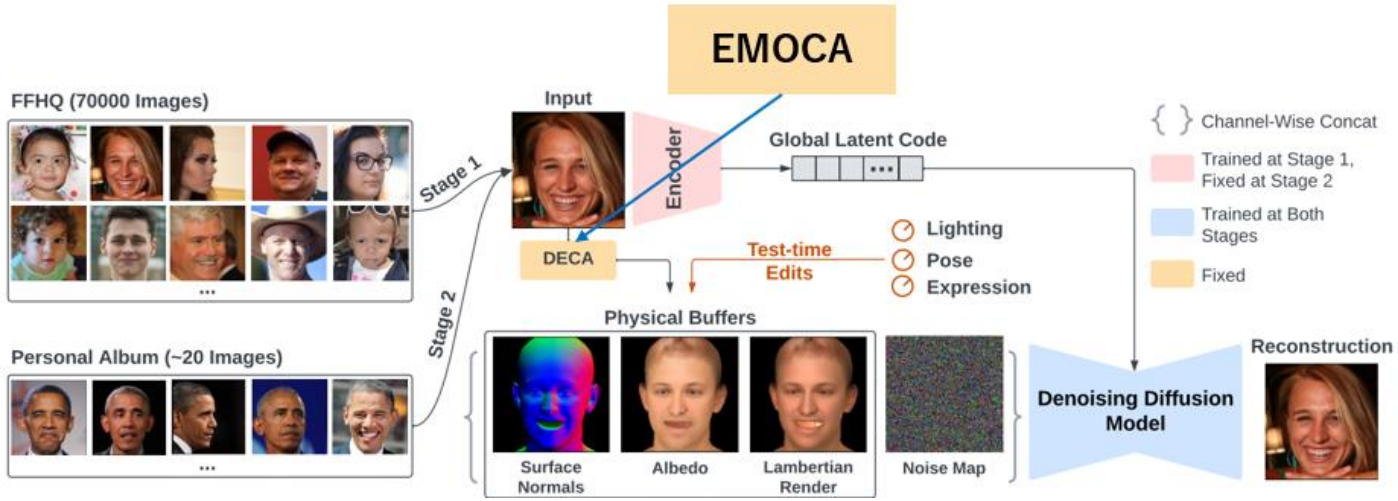
Stepの勘違いにより実験やり直し

DiffusionRigのstage1

使用GPU：V100
Image size：256×256
Batch size：32×1
Max step：50000

DiffusionRig-Emoのstage1

使用GPU：RTX 4090(×2)
Image size：256×256
Batch size：8×2
Max step：50000～100000



3/4までにやりたいことリスト

- ~~AffectNetデータセットでの学習(70000枚⇨280000枚)~~
- target画像と出力画像のvalence・arousalを得るために、**EmoNet**をローカルで実装
- 3種類の指標を使用した評価
- 感情認識用モデルの調査

拡散モデル特有の問題？

Source

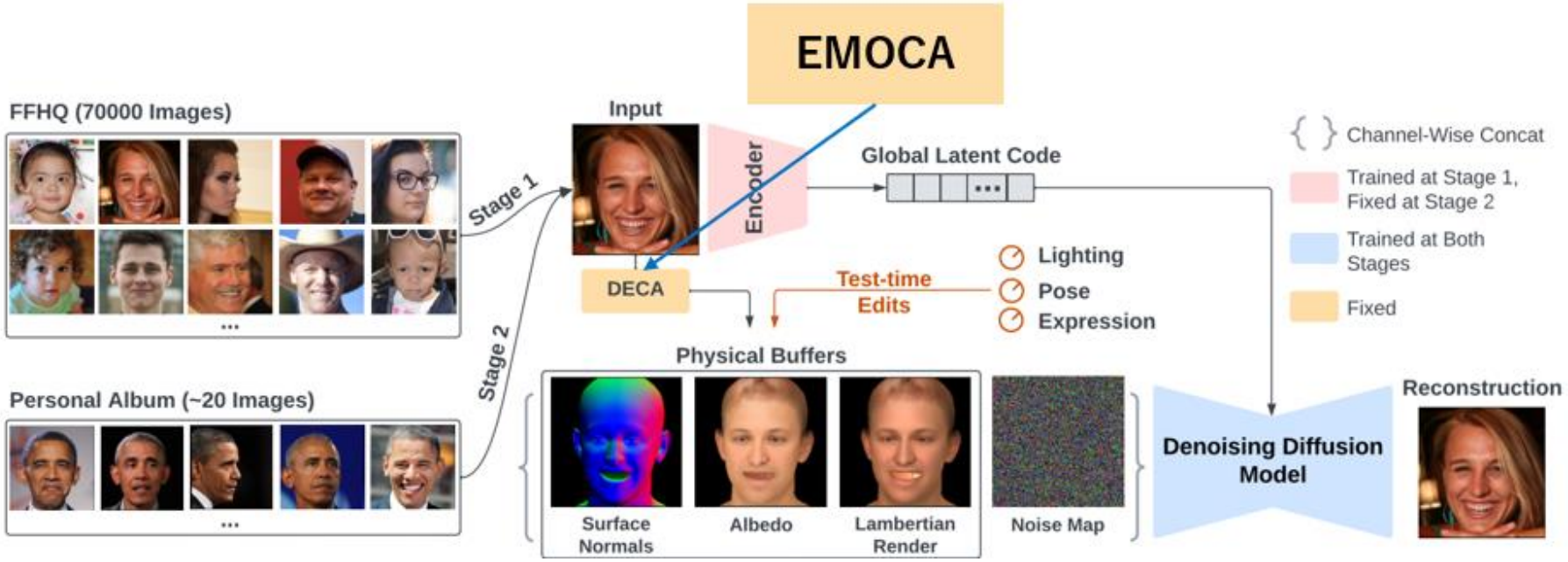
Target

失敗画像

正常画像

評価対象の表情変換モデル

Stage1 stage2	resnet	target expパラ	source expパラ	評価対象モデル
DECA or EMOCA	18 or 50	DECA or EMOCA	DECA or EMOCA	16種類



EmoNetによる評価のテスト

Source

Target

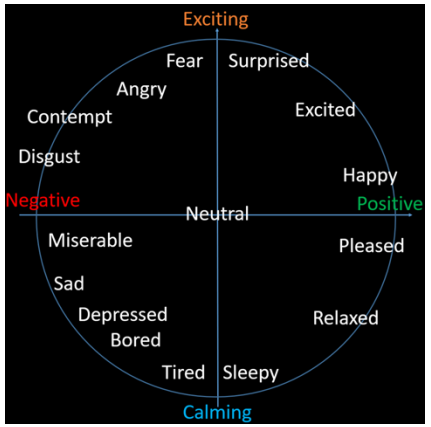
変換後

出力結果

Predicted Emotion Neutral
valence -0.276, arousal 0.214

↓

Predicted Emotion Happy
Valence 0.531, arousal 0.317



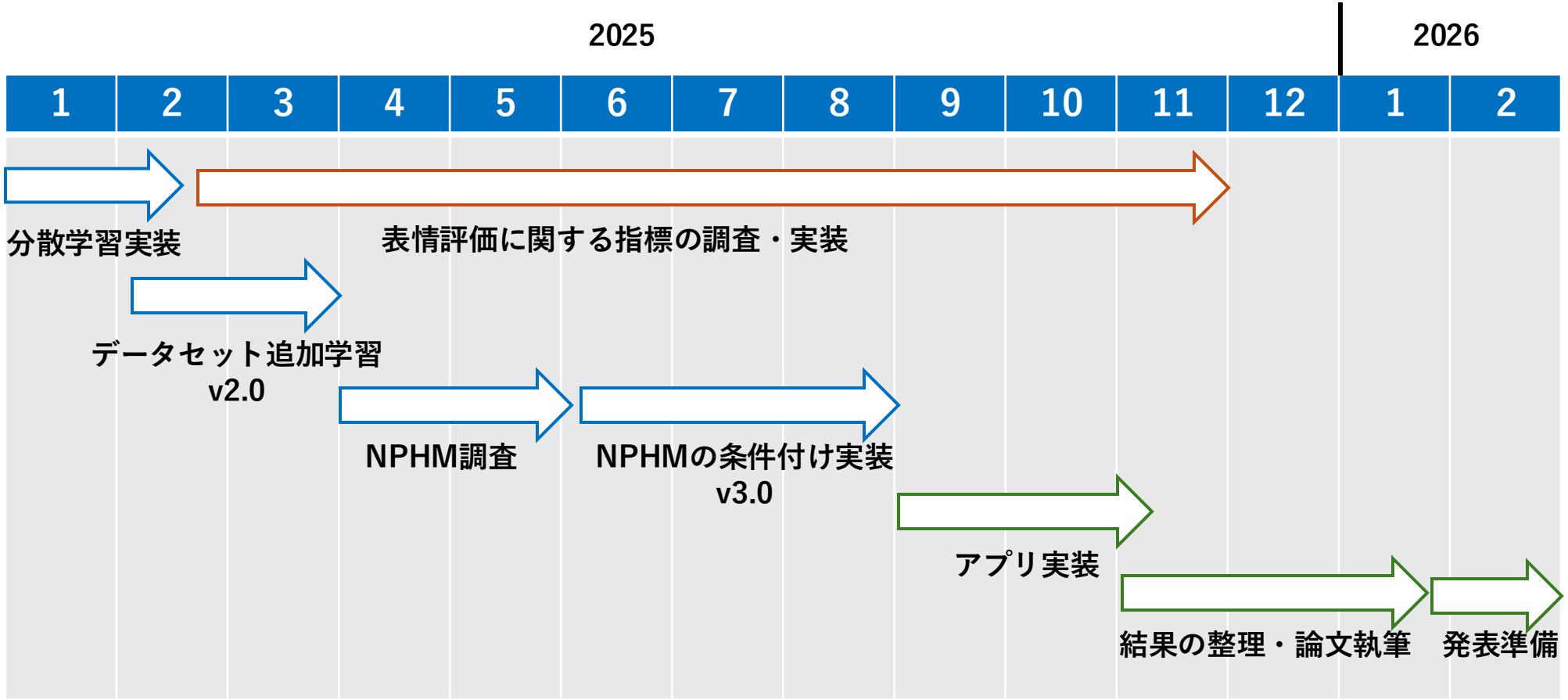
Predicted Emotion Happy
- valence 0.849 - arousal 0.130

Predicted Emotion Surprise
valence 0.255, arousal 0.829

↓

Predicted Emotion Surprise
valence 0.578, arousal 0.547

今後の研究計画



https://github.com/kdhht2334/awesome-SOTA-FER?tab=readme-ov-file#affect